# Challenging the Universal Representation of Deep Models for 3D Point Cloud Registration

David Bojanić[1]
david.bojanic@fer.hr

Kristijan Bartol[1]
kristijan.bartol@fer.hr

Josep Forest[2]
josep.forest@udg.edu

Stefan Gumhold[3]
stefan.gumhold@tu-dresden.de

Tomislav Petković[1]
tomislav.petkovic.jr@.hr

Tomislav Pribanić[1]
tomislav.pribanic@.hr

[1] Faculty of Electrical Engineering and Computing, University of Zagreb Zagreb, Croatia

[2] Computer Vision and Robotics Research Institute (VICOROB) University of Girona, Girona, Spain

[3] TU Dresden, Dresden, Germany

## Abstract

Learning universal representations across different applications domain is an open research problem. In fact, finding universal architecture within the same application but across different types of datasets is still unsolved problem too, especially in applications involving processing 3D point clouds. In this work we experimentally test several state-of-the-art learning-based methods for 3D point cloud registration against the proposed non-learning baseline registration method. The proposed method either outperforms or achieves comparable results w.r.t. learning based methods. In addition, we propose a dataset on which learning based methods have a hard time to generalize. Our proposed method and dataset, along with the provided experiments, can be used in further research in studying effective solutions for universal representations. Our source code is available at: github.com/DavidBoja/greedy-grid-search.

## 1 Introduction

3D point cloud registration is the problem of finding an optimal rotation and translation that aligns two overlapping 3D point clouds. It arises as a subtask in many different computer vision applications such as: 3D reconstruction [7, 28, 58], object recognition and categorization [22, 30, 59], shape retrieval [58], robot navigation [40] and is still a very researched area [9, 27].

The most recent advances have been inspired by the successes of deep learning, i.e., by the development of novel architectures and layers convenient for point cloud processing, such as PointNet [46] and 3D convolutions [29, 52]. Most of the learning-based approaches first extract point cloud features [11, 19, 64, 65], and then either apply RANSAC

for matching [4, 25, 76] or learn the whole registration pipeline end-to-end [3, 37, 67]. Current state-of-the-art methods [2, 4, 25, 44, 47, 63, 72] achieve remarkable performance on public benchmarks [20, 45, 66, 76], even on very difficult examples with an overlap smaller than 30% percent [25].

The main limitation of the state-of-the-art methods, which is typical for deep-learning based methods [53, 42], is that model performance drops on datasets that differ from the training data. Several recent methods [2, 5, 44, 47] address the generalizability issue and demonstrate significant performance retention between 3DMatch [76], KITTI [20], and ETH [45] datasets. To further test the generalization of the existing learning-based methods, we propose a novel FAUST-partial evaluation benchmark. The benchmark consists of overlapping partial views (>60%), based on the FAUST dataset [8] of 3D human scans, and is substantially different from other public benchmarks. We analyze the performance of the best methods under greater scrutiny by using common 3D registration performance metrics and several distance and angle thresholds.

The aim of this paper is to propose a simple, straightforward 3D registration method and use it as a baseline for comparison with the state-of-the-art, particularly to evaluate their generalization performance. We show competitive performance on KITTI and ETH datasets, when we compare against the models pretrained on 3DMatch. When we compare on a substantially different, FAUST-partial benchmark, we outperform state-of-the-art. This result suggests that the learning-based methods, although remarkable on many public datasets, are still not robust enough to be applied on any 3D data. On the other hand, our baseline method performs consistently, regardless of the data distribution.

The proposed method is based on the step-wise (grid) search over the possible rotations and translations. The point clouds are firstly voxelized. Then, the best transformation candidate is selected as the solution that has the maximum cross-correlation between the two voxel volumes; thus, we call our method *greedy* grid search.

In summary, we:

- Evaluate the generalization performance of the state-of-the-art methods under a common set of 3D registration metrics;

- Generate a specialized benchmark, called FAUST-partial, based on 3D human scans, which further challenges the generalization of learning-based methods.

- Propose a novel 3D registration baseline which selects the transformation candidate based on the maximum cross-correlation between the voxelized point clouds;

- Demonstrate comparable performance to the state-of-the-art 3D registration methods on public benchmarks and outperform them on FAUST-partial.

## 2 Related Work

**Classical registration.** The most popular classical registration method is the iterative closest point (ICP) algorithm. The algorithm selects a subset of points based on a criteria, calculates the optimal transformation between the clouds using SVD, and iterates until convergence. The original implementations used point-to-point [6] and point-to-plane [12] distances for finding the solution, but many other strategies have been proposed [13, 43, 59, 75]. GO-ICP [69] proposes a branch-and-bound scheme and proves the global optimality of the algorithm.

The 4-point congruent sets (4PCS) algorithm [1] and its variants [24, 41] are based on the idea that there exist sets of four coplanar points whose alignment corresponds to the alignment of the point clouds. To select the correspondences, RANSAC is used, and ICP is applied for refinement.

**Handcrafted features.** Methods based on handcrafted features first extract correspondences between the point clouds and then find the transformation using RANSAC. Similar to the image keypoint-based methods such as SIFT [39], 3D feature-based methods focus on keypoint detection [50, 53, 57] and their distinctive description [13, 23, 31, 49, 56, 58, 60, 61, 72]. A few recent methods [47, 73] retrieve correspondences without keypoint detection by considering all possible matches.

**Feature learning.** Instead of handcrafting distinctive features, keypoint detection and description can be learned. 3DMatch [76] transforms patches into volumetric voxel grids of truncated distance function (TDF) values and processes them through a 3D convolutional network [29, 52] to output local descriptors. Followed by 3DMatch and the popularity of deep learning, many other works propose to learn keypoint detection [4, 36, 70] and description [15, 17, 25, 34]. Most of these works are learned by optimizing some version of contrastive loss [14, 52] between the descriptors of matching and non-matching points, and then by applying RANSAC to select the final correspondences.

**End-to-end registration learning.** There are many recent approaches that learn not only feature description, but also the subsequent matching step, thus learning end-to-end. The first group of these methods [11, 19, 35, 54, 64, 65, 71, 74], pioneered by the deep closest point [64], follow the ICP idea by iteratively establishing soft correspondences and then applying weighted SVD to obtain the transformations. The remaining group of methods [3, 26, 37, 51, 67], represented by PointNetLK [3] use PointNet architecture [46] or similar global description strategy to regress the transformation based on the global feature vectors.

**Generalization to other datasets.** Several recent methods [2, 5, 44] attempt to generalize to datasets other than training. All of these methods demonstrate significant performance retention on novel datasets, for example, when evaluating 3DMatch-pretrained models on KITTI [2, 5]. On the other hand, the results reported on the ETH dataset only show that the computed descriptors have a high recall [2, 4, 44] (using the feature-matching-recall measurement), but never actually evaluate the 3D registration.

# 3 Method description

Let $\mathcal{X} \in \mathbb{R}^{N \times 3}$ be the *source* point cloud and $\mathcal{Y} \in \mathbb{R}^{M \times 3}$ the *target* point cloud. The goal of 3D registration is to find the rigid homogeneous transformation $\mathbf{T} \in \mathrm{SE}(3)$ that best aligns $\mathcal{X}$ to $\mathcal{Y}$. The rigid transformation $\mathbf{T}$ is composed of a rotation component $\mathbf{R} \in \mathrm{SO}(3)$ and a translation component $\mathbf{t} \in \mathbb{R}^3$.

The method pipeline is summarized in Fig. 1. We divide our method in 3 key steps: *preprocessing*, *processing* and *post-processing*. In this Section we present the general pipeline. The final estimation of the rigid transformation is provided in Eq. 7.

**R and t parametrization.** To find the correct rotation and translation, we perform a grid search over the rotation and translation space. The rotation space is sampled using 3 Euler angles $\alpha, \beta$ and $\gamma$ that rotate the source point cloud $\mathcal{X}$ around the $x, y$ and $z$ axes respectively. In a typical 3D scanning use case, two scanned point clouds that need to be registered should be fairly close in the rotation space since the scanning process is limited by a required overlap region. We uniformly sample each angle from the range $[-\gamma, \gamma]$ with an angle step of $S$ and
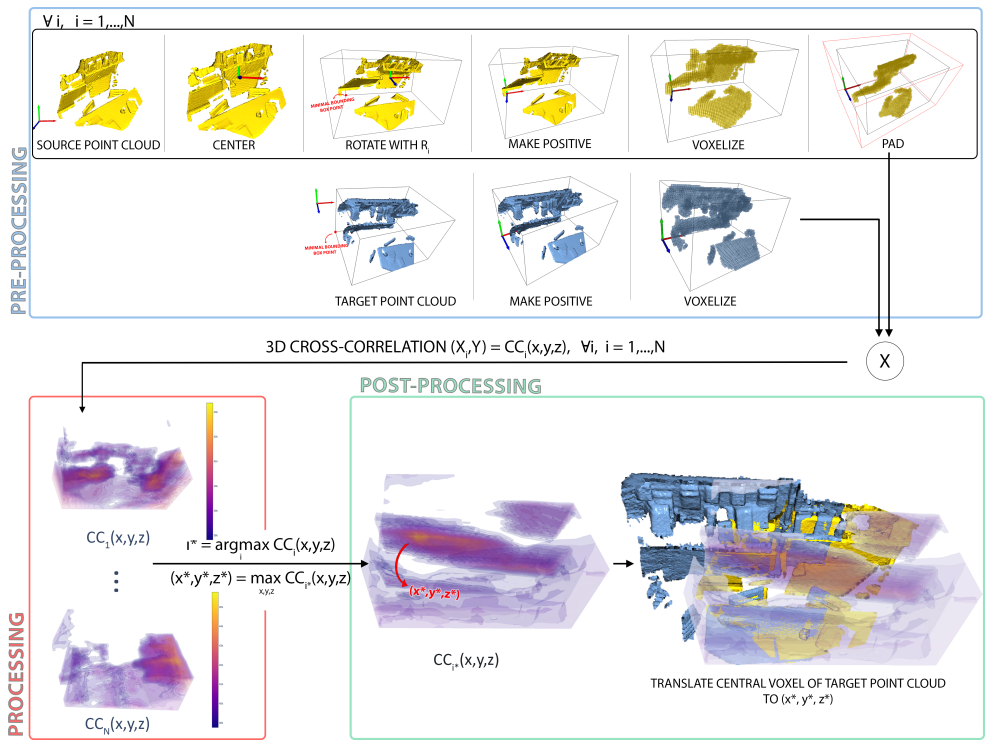
Figure 1: The proposed pipeline. The method is divided into 3 steps: *pre-processing*, *processing* and *post-processing*. Each step follows the previous one. The *pre-processing* step prepares the initial data and outputs $N$ voxelized source volumes and one target volume. The *processing* step performs the 3D cross-correlation over each source volume and the target volume. The cross-correlation volumes $CC_i(x,y,z)$ are heatmaps that should indicate higher (indicated in yellow on the volumes) or lower (indicated with purple on the volumes) matching between the source $\mathbf{X}_i$ and target $\mathbf{Y}$ volumes at the corresponding voxels. White spaces are present because the cross-correlations values are clipped so only the upper range is visible. Finally, the *post-processing* step finds the solution from the output volumes by finding the maximal cross-correlation from all the given volumes.

create a Cartesian product of all the possible combinations resulting in $N$ triplets. Finally, we convert these to rotation matrices $R_i, i = 1, \ldots, N$. Note that this step is only computed once, prior to registration. The translation space is inherently sampled by the voxelization process of the given point clouds. The possible translations hence correspond to the centers of the source point cloud voxels and are therefore dependent on the voxelization resolution. More details are provided in the following 4 sections.

**Pre-processing.** First, we center and rotate the source point cloud $\mathcal{X}$ around the origin using the precomputed rotation matrices $R_i$, obtaining $\mathcal{X}_i, i = 1, \ldots, N$. Next, we make all the point clouds coordinates positive by translating their minimal bounding box point into the origin. This step is only done to facilitate the voxelization process and can be completely omitted. We then voxelize each source $\mathcal{X}_i$ and target $\mathcal{Y}$ point clouds with a voxel resolution

of $VR$ cm. Instead of having a 3D grid with ones and zeros, we set a value of $PV$ (positive voxel) for the filled voxels (voxels where a point from the point cloud is present) and a value of $NV$ (negative voxel) for the empty ones (voxels where there are no points present from the point cloud). This results in $N$ source volumes $\mathbf{X}_i$ and one target volume $\mathbf{Y}$.

**Processing.** For each source volume $\mathbf{X}_i$, we perform a 3D cross-correlation with the target volume $\mathbf{Y}$. Essentially, the central voxel of the target volume is translated over each voxel of the source volume where the cross-correlation can be computed by multiplying the overlying voxel values of the two volumes and summing them together. This results in $N$ cross-correlation volumes $CC_i(x,y,z)$ with the same 3 dimensions as the source volume. The volumes can be thought of as discrete heatmaps where higher values should represent higher degrees of matching between the voxelized point clouds. Prior to the cross-correlation, each source volume is padded in order for the target volume to *slide* all over the source volume. We mark with $\mathbf{P} \in R^6$ the padding applied to each source volume $\mathbf{X}_i$, where the values represent the number of voxels padded to the left, right, top, bottom, front and back of the volume, respectively. We make use of the Fourier domain to accelerate the computation of the cross-correlation. Both volumes are first transformed into the Fourier space using the FFT algorithm [10], after which the cross-correlation simplifies to a matrix multiplication [43]. The output is then transformed back with an inverse FFT.

**Post-processing.** We finally estimate the rotation matrix $\hat{R}$ that aligns (rotation-wise) $\mathcal{X}$ to $\mathcal{Y}$ using one of the $N$ precomputed rotation matrices $R_i$. We select the matrix $R_i$ that corresponds to $\mathbf{X}_i$ with the maximal cross-correlation value from the $CC_i(x,y,z)$ volumes. More concretely, we use the

$$i^* = \underset{i}{\operatorname{argmax}} CC_i(x,y,z) \tag{1}$$

index to find the estimated rotation matrix $\hat{R} = R_{i^*}$. To estimate the translation we find the voxel with the maximal cross-correlation value from $CC_{i^*}$. Then, we translate the central voxel of the target volume to the just found voxel of the $CC_{i^*}$ volume. Since the $CC_{i^*}$ volume corresponds to the source $\mathbf{X}_{i^*}$ volume, we essentially translate the central voxel of the target volume to the voxel of the source volume with the maximal cross-correlation. More concretely, we find the index of the voxel with the maximal cross-correlation value with

$$(x^*, y^*, z^*) = \underset{x,y,z}{\operatorname{argmax}} CC_{i^*}(x,y,z). \tag{2}$$

Then, to translate the central voxel of the target volume to it, we use the translation:

$$t_{\text{est}} = \left( \underbrace{-\mathbf{Y}_{\text{cv}}}_{\substack{\text{target} \\ \text{volume} \\ \text{central} \\ \text{voxel}}} \underbrace{- \begin{bmatrix} \mathbf{P}[0] \\ \mathbf{P}[2] \\ \mathbf{P}[4] \end{bmatrix}}_{\substack{\text{padding} \\ \text{displacement}}} + \underbrace{\begin{bmatrix} x^* \\ y^* \\ z^* \end{bmatrix}}_{\substack{\text{max cc} \\ \text{voxel}}} + \underbrace{\begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}}_{\substack{\text{move to} \\ \text{center of} \\ \text{voxel}}} \right) \times \text{VR} \tag{3}$$

$$\underbrace{\phantom{-\mathbf{Y}_{\text{cv}}}}_{\substack{\text{move to} \\ \text{origin}}} \quad \underbrace{\phantom{xxxxxxxxxxxxxxxxxxxxxxxx}}_{\text{move to } (x^*,y^*,z^*)}$$

where each value is multiplied by the voxel resolution VR to transform from voxel indices to euclidean coordinates. The central voxel of the target volume can be computed as:

$$\mathbf{Y}_{\text{cv}} = [\lceil V_x/2 \rceil, \lceil V_y/2 \rceil, \lceil V_z/2 \rceil] \tag{4}$$

where $V_x, V_y, V_z$ are the number of voxels of $\mathbf{Y}$ along the 3 dimensions. Intuitively, the central voxel along a dimension is the middle voxel if the number of voxels is odd, and one on the left of the "middle point" if it's odd.

Following all of the steps above, the rigid registration can be summarized as:

$$\left(\hat{R}(\mathcal{X} - t_{\mathcal{X}}^{\text{CENTER}})\right) + t_{\mathcal{X}}^{\text{POSIT}} \sim \left(\mathcal{Y} + t_{\mathcal{Y}}^{\text{POSIT}}\right) - t_{\text{est}} \tag{5}$$

where $\sim$ indicates that the left and right part are aligned. The $t^{\text{CENTER}}$ translation moves the center of mass of the respective point cloud into the origin and $t^{\text{POSIT}}$ moves the minimal bounding box point into the origin. More concretely:

$$t_{\mathcal{X}}^{\text{CENTER}} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{X}[i,:] \in R^3, \quad t_{\mathcal{X}}^{\text{POSIT}} = -\begin{bmatrix} \min\mathcal{X}[:,1] \\ \min\mathcal{X}[:,2] \\ \min\mathcal{X}[:,3] \end{bmatrix} \in R^3, \quad t_{\mathcal{Y}}^{\text{POSIT}} = -\begin{bmatrix} \min\mathcal{Y}[:,1] \\ \min\mathcal{Y}[:,2] \\ \min\mathcal{Y}[:,3] \end{bmatrix} \in R^3 \tag{6}$$

where the $[:,:]$ notation indicates the row-wise and column-wise indexing, and min indicates the minimal element of an array.

Since the final rigid transformation needs to align $\mathcal{X}$ to $\mathcal{Y}$, Eq. 5 equation is further refined as:

$$\left(\hat{R}(\mathcal{X} - t_{\mathcal{X}}^{\text{CENTER}})\right) + t_{\mathcal{X}}^{\text{POSIT}} + t_{\text{est}} - t_{\mathcal{Y}}^{\text{POSIT}} \sim \mathcal{Y} \tag{7}$$

Therefore, the final rotation and translation estimations are:

$$\hat{R} = R_{i^*}, \qquad \hat{t} = -\hat{R}t_{\mathcal{X}}^{\text{CENTER}} + t_{\mathcal{X}}^{\text{POSIT}} + t_{\text{est}} - t_{\mathcal{Y}}^{\text{POSIT}} \tag{8}$$

**Refinement.** Since the rotation and translation spaces are discretized, the initial alignment is only a rough estimate. The upper bounds on the rotation and translation errors are $\frac{1}{2}\max_{i,j\,i\neq j}\arccos\left(\frac{\text{trace}(R_i^T R_j)-1}{2}\right)\frac{180}{\pi}$ degrees and $\frac{\text{VR}\sqrt{3}}{2}$ centimeters if the ground truth solution is located in the estimated discretized locations. For an angle step of $S = 15°$ and $VR = 6$cm the upper bound errors would be $7.5°$ and $5$cm. Hence, the rough initial alignment should provide a very good initialization for a fine registration algorithm. We use generalized ICP [53] to refine the initial solution since it provided the best results.

# 4   Experiments

We evaluate the generalization capabilities of state-of-the-art methods trained on 3DMatch [46] and compare them to our baseline. We use two established benchmarking datasets ETH [45] and KITTI [20] and create a novel FAUST-partial benchmark based on the FAUST dataset [8]. These datasets test the generalization abilities in terms of different sensor modalities (RGB-D, laser scanner), different environments (indoor, outdoor), resolution (6mm to 5cm) and completely different structure (from indoor objects to humans).

**Metrics.** Following [2, 4, 11, 16, 47] we evaluate the results using the Relative Rotation Error (RRE), the Relative Translation Error (RTE) and the Registration Recall (RR) measures. The Relative Rotation Error measures the relative angle in degrees between the ground-truth $R^*$ and estimated $\hat{R}$ rotation matrices:

$$RRE = \arccos\left(\frac{\text{trace}(\hat{\mathbf{R}}^T\mathbf{R}^*)-1}{2}\right)\frac{180}{\pi} \tag{9}$$

The Relative Translation Error measures the distance from the ground-truth $\mathbf{t}^*$ and estimated $\hat{\mathbf{t}}$ translation vectors:

$$RTE = \|\mathbf{t}^* - \hat{\mathbf{t}}\|_2 \tag{10}$$

The Registration Recall measures the fraction of successfully registered pairs of point clouds. A registration is deemed successful (or a true positive in terms of the recall measure) if its RRE and RTE are below predefined thresholds $\tau_r$ and $\tau_t$:

$$RR = \frac{1}{|\Omega|} \sum_{(i,j) \in \Omega} \mathbb{1}_{\{\text{RRE}(i,j) < \tau_r \ \wedge \ \text{RTE}(i,j) < \tau_t\}} \tag{11}$$

where $\Omega$ is the set of all the point cloud registration pairs $(i, j)$ in the dataset, $\mathbb{1}$ is an indicator function and $\text{RRE}(i, j), \text{RTE}(i, j)$ indicate the RRE and RTE for registration pairs $(i, j)$. The final RRE and RTE measurements are averaged only over the successfully registered pairs $(i, j)$ obtained from the RR.

**Parameters.** To fully define the baseline, we need to set the parameters of the euler angle range $\gamma = 90°$, angle step $S = 15°$, the positive voxel value $PV = 5$ and the negative voxel value $NV = -1$. The parameters $\gamma$ and $S$ then determine $N = 2028$. The padding $\mathbf{P}$ is determined for each source volume $\mathbf{X}_i$ so that the volume stays the same dimension after the cross-correlation. The only parameter we vary for each dataset is the voxel resolution $VR$ since the datasets vary greatly in their dimensions ranging from volumes of 150m × 85m × 10m for KITTI to 1.8m × 0.8m × 0.6m for FAUST-partial.

**KITTI.** The KITTI dataset [20] is comprised of 11 sequences of outdoor driving scenarios obtained by a LiDAR scanner. Compared to 3DMatch, the fragments are much larger, have lower resolution and a different structure. Following common practice [2, 4, 15, 25, 25, 47], we test our baseline on scenes 8 to 10 using pairs which are at least 10m away from each other. The ground truth transformation matrices are refined using ICP and the evaluation thresholds are set to $\tau_r = 5°$ and $\tau_t = 2$m. We set $VR = 75$cm.

As can be seen from Table 1, most state-of-the-art methods tend to generalize well onto the KITTI dataset. The fragments from the dataset are gravity aligned, which is reflected in lower RRE errors since most of the ground truth rotation comes from rotating around one ax. The fragments are also much bigger than those from 3DMatch (150m × 85m × 10m on average in volume compared to 3m$^3$ in 3DMatch) which is reflected in higher RTE errors. Surprisingly, DIP shows poor recall performance with only 51.71% aligned pairs. The KITTI dataset is much noisier than 3DMatch which might be affecting the local reference frame (LRF) alignment in DIP. The baseline achieves comparable results and only lags behind the best recall result from GeoTransformer by less than 1 percentage point (pp). However, it compensates by achieving the lowest RRE and RTE measurements.

| Method | RR(%) | RRE(°) | RTE (cm) |
|---|---|---|---|
| FCGF[†] [□] | 24.19 | 1.61 | 27.10 |
| D3Feat-rand[†] [□] | 18.47 | 1.58 | 37.80 |
| D3Feat-pred[†] [□] | 36.76 | 1.44 | 31.60 |
| SpinNet[†] [□] | 81.44 | 0.98 | 15.60 |
| DIP | 51.71 | 1.02 | 13.43 |
| GeoTransformer [□] | **90.63** | 0.54 | 154.27 |
| FCGF+PointDSC [□] | 66.13 | 0.88 | 107.71 |
| FCGF+YOHO-O [□] | 81.44 | 1.99 | 54.25 |
| FCGF+YOHO-C [□] | 82.16 | 1.38 | 39.30 |
| Baseline | 90.27 | **0.13** | **4.68** |

Table 1: Results on the KITTI dataset. All the methods are trained on the 3DMatch dataset. Results marked with † are taken from [□].

**ETH.** The ETH dataset [45] consists of 4 scenes mostly comprised of outdoor vegetation. Compared to 3DMatch, the fragments are larger, have lower resolution and have more

complex geometries. Following common practice [2, 21, 44] we use only point clouds with an overlap greater than 30%. We set the thresholds to $\tau_r = 5°$ and $\tau_t = 30$cm and the voxel resolution of the baseline to $VR = 60$cm. The stricter threshold for the RTE reflects the fact that the fragments from ETH are much smaller in volume than those from KITTI. As can be seen in Table 2, the deep learning methods generalize somewhat less than on the KITTI dataset. GeoTransformer and PointDSC achieve low recall results which can be explained by the high RTE error, also visible on the KITTI dataset. Hence, with a stricter RTE threshold, the recall drops significantly. The best performance is achieved by YOHO, followed by the baseline with 6.45 recall percentage points difference. However, the baseline achieves, again, the lowest RRE and RTE measures between all the methods.

| Method | RR(%) | RRE(°) | RTE (cm) |
|---|---|---|---|
| SpinNet [2] | 73.07 | 1.205 | 5.352 |
| DIP [12] | 62.41 | 1.940 | 14.716 |
| GeoTransformer [11] | 4.91 | 0.710 | 21.162 |
| FCGF+PointDSC [5] | 2.81 | 0.573 | 23.426 |
| FCGF+YOHO-O [44] | 79.94 | 2.167 | 16.112 |
| FCGF+YOHO-C [44] | **84.85** | 1.950 | 16.176 |
| Baseline | 78.40 | **0.355** | **1.706** |

Table 2: Results on ETH dataset. All the methods are trained on the 3DMatch dataset.

**FAUST-partial.** To further test the generalization capabilities of a 3D registration method, we create a new FAUST-partial benchmark based on the FAUST [8] dataset. The state-of-the-art methods that train on the 3DMatch dataset, test their generalization capabilities [2, 5, 11, 21, 44, 47] on the KITTI [20] and ETH [45] datasets that tend to be comprised of similar objects. We argue that because of that reason, using proper data augmentation whilst learning on 3DMatch, a method can increase its robustness and generalization on noisier and perturbed fragments from other datasets. However, when encountered with completely unseen data (such as 3D human scans), the methods have difficulty generalizing, even when having more than 60% overlap.

To create a 3D registration benchmark, we use the FAUST training dataset comprised of 100 human body scans. We divide each scan into multiple overlapping fragments that need to be aligned. The steps are illustrated in Fig. 2. To create the fragments, we begin by moving each scan so the $xz$-plane acts as the floor. We do this by moving the minimal bounding box point to the origin. Next, we create a regular icosahaedron centered at the center of mass of each scan. A regular icosahaedron contains 12 points that lie on a unit sphere around its center, each equidistant from its neighbors. We scale the icosahaedron to a 1.5m radius sphere so each scan fits inside it. These points are used as the viewpoints for creating the partial views (fragments), discarding the ones located below the $xz$-plane (below the floor). For each viewpoint, we use the hidden point removal [32] algorithm to create a partial point cloud. Next, for each two pairs of viewpoints $(i, j)$ we sample a rotation using 3 Euler angles from the range $[0°, 45°]$ and a random translation from the range $[-50$cm$, 50$cm$]$. We rotate and translate the partial point cloud obtained from viewpoint $i$ to finally get a benchmark registration pair. For the final benchmark we use every pair of viewpoints that have an overlap bigger than 60%, resulting in 1724 registration pairs.

For evaluating the baseline we set the voxel resolution to $VR = 6$cm and use the thresholds $\tau_r = 10°$ and $\tau_t = 3$cm. The strict threshold for RTE reflects the fact that fragments from FAUST-partial are much smaller in volume than all the other datasets. As can be seen from Table 3, deep learning methods are incapable of generalizing onto datasets that differ considerably from the 3DMatch-like datasets on which they were trained on. Additionally,
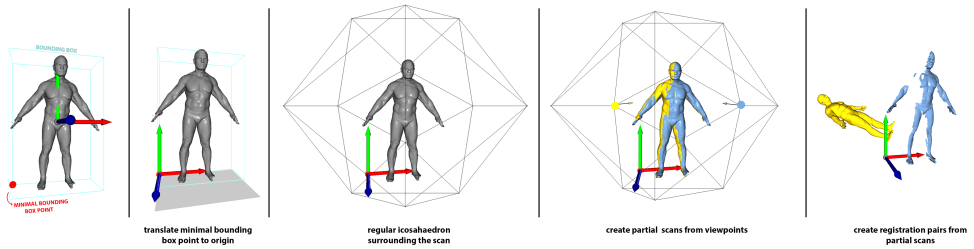
Figure 2: FAUST-partial dataset generation. For a given scan from the FAUST [8] dataset, we translate its minimal bounding box point to the origin. Next, we surround the scan with a regular icosahaedron. Each point of the icosahaedron acts as a viewpoint used to create a partial scan using the hidden point removal algorithm [32]. For two partial scans with an overalp bigger than 60%, we use a random rotation and translation to obtain a registration pair for the FAUST-partial benchmark.

| Method | RR(%) | RRE(°) | RTE(cm) |
|--------|-------|--------|---------|
| FPFH-8M [19] | 9.51 | 4.347 | 1.900 |
| SpinNet [1] | 42.46 | 3.105 | 1.670 |
| GeoTransformer [42] | 56.15 | 2.423 | 1.581 |
| FCGF+PointDSC [9] | 47.85 | 3.354 | 1.793 |
| FCGF+YOHO-O [53] | 18.91 | 4.489 | 1.852 |
| FCGF+YOHO-C [53] | 29.18 | 3.653 | 1.668 |
| DIP [44] | 54.81 | 4.058 | 2.052 |
| Baseline | **92.81** | **0.014** | **0.009** |

Table 3: Results on FAUST-partial dataset. All the methods are trained on the 3DMatch dataset. FPFH-8M is registered with RANSAC using 8 million iterations.

the FAUST-partial registration pairs have considerably bigger overlaps ($> 60\%$) than the 3DMatch benchmark. However, the best recall of 56.15% from the deep learning methods is achieved by GeoTransformer. Contrary, the baseline performs consistently, achieving a remarkable 92.81% recall and the lowest RRE and RTE measures.

# 5 Conclusion

The proposed classical approach provides a good 3D registration baseline. The method is simple but effective, which is demonstrated on the public benchmarks. Compared to the generalization performance of the state-of-the-art methods, the baseline is on-par. On the newly proposed, FAUST-partial benchmark, the competing methods are struggling to retain the results, or perform significantly worse, even though the generated overlap between the cloud pairs is reasonably high. Contrary to the deep learning methods, the baseline is simple and explainable and serves for detailed analyses. The effects of different strategies are clear and intuitive and provide insights into the registration process. Therefore, on the search of finding universal representations, designing a deep model mimicking the proposed baseline method is an interesting future avenue to pursue.

# References

[1] Dror Aiger, Niloy J. Mitra, and Daniel Cohen-Or. 4-points congruent sets for robust pairwise surface registration. *ACM Trans. Graph.*, 27(3):1–10, aug 2008. ISSN 0730-0301. doi: 10.1145/1360612.1360684. URL https://doi.org/10.1145/1360612.1360684.

[2] Sheng Ao, Qingyong Hu, Bo Yang, Andrew Markham, and Yulan Guo. Spinnet: Learning a general surface descriptor for 3d point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[3] Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivatsan, and Simon Lucey. Pointnetlk: Robust & efficient point cloud registration using pointnet. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7156–7165, 2019.

[4] 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. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6358–6366, 2020.

[5] Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li nad Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. PointDSC: Robust Point Cloud Registration using Deep Spatial Consistency. *CVPR*, 2021.

[6] Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14:239–256, 1992.

[7] G. Blais and M. D. Levine. Registering multiview range data to create 3d computer objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):820–824, 1995.

[8] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. Faust: Dataset and evaluation for 3D mesh registration. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Piscataway, NJ, USA, June 2014. IEEE.

[9] David Bojanić, Kristijan Bartol, Tomislav Petković, and Tomislav Pribanić. A review of rigid 3d registration methods. *Proceedings of 13th International Scientific – Professional Symposium TEXTILE SCIENCE ECONOMY*, 2020.

[10] E. O. Brigham and R. E. Morrow. The fast fourier transform. *IEEE Spectrum*, 4(12):63–70, 1967. doi: 10.1109/MSPEC.1967.5217220.

[11] Anh-Quan Cao, Gilles Puy, Alexandre Boulch, and Renaud Marlet. PCAM: Product of Cross-Attention Matrices for Rigid Registration of Point Clouds. In *International Conference on Computer Vision (ICCV)*, 2021.

[12] Yang Chen and Gérard G. Medioni. Object modelling by registration of multiple range images. *Image Vis. Comput.*, 10:145–155, 1992.

[13] D. Chetverikov, D. Svirko, D. Stepanov, and P. Krsek. The trimmed iterative closest point algorithm. In *2002 International Conference on Pattern Recognition*, volume 3, pages 545–548 vol.3, 2002. doi: 10.1109/ICPR.2002.1047997.

[14] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1, 2005. doi: 10.1109/CVPR.2005.202.

[15] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8957–8965, 2019. doi: 10.1109/ICCV.2019.00905.

[16] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *CVPR*, 2020.

[17] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppfnet: Global context aware local features for robust 3d point matching. 10 2018. doi: 10.1109/CVPR.2018.00028.

[18] Andrea Frome, Daniel Huber, Ravi Kolluri, Thomas Bülow, and Jitendra Malik. Recognizing objects in range data using regional point descriptors. volume 3, pages 224–237, 05 2004. ISBN 978-3-540-21982-8. doi: 10.1007/978-3-540-24672-5_18.

[19] Kexue Fu, Shaolei Liu, Xiaoyuan Luo, and Manning Wang. Robust point cloud registration framework based on deep graph matching. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8889–8898, 2021.

[20] 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, 2012. doi: 10.1109/CVPR.2012.6248074.

[21] Zan Gojcic, Caifa Zhou, Jan Dirk Wegner, and Andreas Wieser. The perfect match: 3d point cloud matching with smoothed densities. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5540–5549, 2019.

[22] Y. Guo, F. Sohel, M. Bennamoun, M. Lu, and J. Wan. Rotational projection statistics for 3d local surface description and object recognition. *International Journal of Computer Vision*, 105(1):63 – 86, Apr 2013. ISSN 1573-1405. doi: 10.1007/s11263-013-0627-y. URL http://dx.doi.org/10.1007/s11263-013-0627-y.

[23] Yulan Guo, Mohammed Bennamoun, Ferdous A. Sohel, Jianwei Wan, and Min Lu. 3d free form object recognition using rotational projection statistics. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 1–8, 2013. doi: 10.1109/WACV.2013.6474992.

[24] Jida Huang, Tsz-Ho Kwok, and Chi Zhou. V4PCS: Volumetric 4PCS Algorithm for Global Registration. *Journal of Mechanical Design*, 139(11), 10 2017. ISSN 1050-0472. doi: 10.1115/1.4037477. URL https://doi.org/10.1115/1.4037477. 111403.

[25] Shengyu Huang, Zan Gojcic, Mikhail (Misha) Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4265–4274, 2021.

[26] Xiaoshui Huang, Guofeng Mei, and Jian Zhang. Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11363–11371, 2020.

[27] Xiaoshui Huang, Guofeng Mei, Jian Zhang, and Rana Abbas. A comprehensive survey on point cloud registration. *ArXiv*, abs/2103.02690, 2021.

[28] D. Huber and M. Hebert. Fully automatic registration of multiple 3d data sets. *Image and Vision Computing*, 21:637–650, 07 2003. doi: 10.1016/S0262-8856(03)00060-X.

[29] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013. doi: 10.1109/TPAMI.2012.59.

[30] A. E. Johnson and M. 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.

[31] A.E. Johnson and M. 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. doi: 10.1109/34.765655.

[32] Sagi Katz, Ayellet Tal, and Ronen Basri. Direct visibility of point sets. volume 26, 07 2007. doi: 10.1145/1275808.1276407.

[33] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *ArXiv*, abs/1710.05468, 2017.

[34] Marc Khoury, Qian-Yi Zhou, and Vladlen Koltun. Learning compact geometric features. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 153–161, 2017.

[35] Jiahao Li, Changhao Zhang, Ziyao Xu, Hangning Zhou, and Chi Zhang. Iterative distance-aware similarity matrix convolution with mutual-supervised point elimination for efficient point cloud registration. In *European Conference on Computer Vision (ECCV)*, 2020.

[36] Jiaxin Li and Gim Hee Lee. Usip: Unsupervised stable interest point detection from 3d point clouds. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 361–370, 2019.

[37] Xueqian Li, Jhony Kaesemodel Pontes, and Simon Lucey. Pointnetlk revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12763–12772, June 2021.

[38] Y. Li, A. Dai, L. Guibas, and M. Niessner. Database-assisted object retrieval for real-time 3d reconstruction. *Computer Graphics Forum*, 34:435–446, 2015.

[39] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[40] M. Magnusson, A. Lilienthal, and T. Duckett. Scan registration for autonomous mining vehicles using 3d-ndt. *Journal of Field Robotics*, 24:803–827, 10 2007. doi: 10.1002/rob.20204.

[41] Mustafa Mohamad, David Rappaport, and Michael Greenspan. Generalized 4-points congruent sets for 3d registration. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 83–90, 2014. doi: 10.1109/3DV.2014.21.

[42] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. In *NIPS*, 2017.

[43] Artem Pavlov, Grigory Ovchinnikov, Dmitriy Derbyshev, Dzmitry Tsetserukou, and Ivan Oseledets. Aa-icp: Iterative closest point with anderson acceleration. pages 1–6, 05 2018. doi: 10.1109/ICRA.2018.8461063.

[44] Fabio Poiesi and Davide Boscaini. Distinctive 3D local deep descriptors. In *IEEE Proc. of Int'l Conference on Pattern Recognition*, Milan, IT, Jan 2021.

[45] François Pomerleau, Ming Liu, Francis Colas, and Roland Siegwart. Challenging data sets for point cloud registration algorithms. *The International Journal of Robotics Research*, 31:1705–1711, 12 2012. doi: 10.1177/0278364912458814.

[46] C. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2017.

[47] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kaiping Xu. Geometric transformer for fast and robust point cloud registration. *ArXiv*, abs/2202.06688, 2022.

[48] K. R. Rao, D. N. Kim, and J.-J. Hwang. *Fast Fourier Transform - Algorithms and Applications*. Springer Publishing Company, Incorporated, 1st edition, 2010. ISBN 1402066287.

[49] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE International Conference on Robotics and Automation*, pages 3212–3217, 2009. doi: 10.1109/ROBOT.2009.5152473.

[50] Samuele Salti, Federico Tombari, Riccardo Spezialetti, and Luigi Di Stefano. Learning a descriptor-specific 3d keypoint detector. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2318–2326, 2015. doi: 10.1109/ICCV.2015.267.

[51] Vinit Sarode, Xueqian Li, Hunter Goforth, Yasuhiro Aoki, Rangaprasad Arun Srivatsan, Simon Lucey, and Howie Choset. PCRNet: Point Cloud Registration Network using Pointnet Encoding. *ArXiv*, abs/1908.07906, 2019.

[52] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[53] Aleksandr Segal, Dirk Hähnel, and Sebastian Thrun. Generalized-icp. 06 2009. doi: 10.15607/RSS.2009.V.021.

[54] Martin Simon, Kai Fischer, Stefan Milz, Christian Witt, and Horst-Michael Groß. Stickypillars: Robust feature matching on point clouds using graph neural networks. *ArXiv*, abs/2002.03983, 2020.

[55] Ivan Sipiran and Benjamin Bustos. Harris 3d: A robust extension of the harris operator for interest point detection on 3d meshes. *The Visual Computer*, 27:963–976, 11 2011. doi: 10.1007/s00371-011-0610-y.

[56] Bastian Steder, Radu Bogdan Rusu, KurtKonolige, and Wolfram Burgard. Narf: 3d range image features for object recognition.

[57] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Proceedings of the Symposium on Geometry Processing*, SGP '09, page 1383–1392, Goslar, DEU, 2009. Eurographics Association.

[58] 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. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2014.06.015. URL https://www.sciencedirect.com/science/article/pii/S0924271614001701.

[59] F. Tombari, S. Salti, and L. Di Stefano. Unique signatures of histograms for local surface description. *Proc. ECCV*, 6313:356–369, 09 2010. doi: 10.1007/978-3-642-15558-1_26.

[60] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique shape context for 3d data description. In *Proceedings of the ACM Workshop on 3D Object Retrieval*, 3DOR '10, page 57–62, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450301602. doi: 10.1145/1877808.1877821. URL https://doi.org/10.1145/1877808.1877821.

[61] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique signatures of histograms for local surface description. In *Proceedings of the 11th European Conference on Computer Vision Conference on Computer Vision: Part III*, ECCV'10, page 356–369, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 364215557X.

[62] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. C3d: Generic features for video analysis. *ArXiv*, abs/1412.0767, 2014.

[63] Haiping Wang, Yuan Liu, Zhen Dong, Wenping Wang, and Bisheng Yang. You only hypothesize once: Point cloud registration with rotation-equivariant descriptors. *ACM Multimedia 2022*, 2022.

[64] Yue Wang and Justin M. Solomon. Deep closest point: Learning representations for point cloud registration. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3522–3531, 2019.

[65] Yue Wang and Justin M. Solomon. Prnet: Self-supervised learning for partial-to-partial registration. In *NeurIPS*, 2019.

[66] Zhirong Wu, S. Song, A. Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, Los Alamitos, CA, USA, jun 2015. IEEE Computer Society. doi: 10.1109/CVPR.2015.7298801. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298801.

[67] Hao Xu, Shuaicheng Liu, Guangfu Wang, Guanghui Liu, and Bing Zeng. Omnet: Learning overlapping mask for partial-to-partial point cloud registration. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3112–3121, 2021.

[68] J. Yang, Y. Xiao, and Z. Cao. Aligning 2.5d scene fragments with distinctive local geometric features and voting-based correspondences. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(3):714–729, 2019.

[69] Jiaolong Yang, Hongdong Li, and Yunde Jia. Go-ICP: Solving 3D Registration Efficiently and Globally Optimally. In *2013 IEEE International Conference on Computer Vision*, pages 1457–1464, 2013. doi: 10.1109/ICCV.2013.184.

[70] Zi Jian Yew and Gim Hee Lee. 3dfeat-net: Weakly supervised local 3d features for point cloud registration. *ArXiv*, abs/1807.09413, 2018.

[71] Zi Jian Yew and Gim Hee Lee. RPM-Net: Robust Point Matching Using Learned Features. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11821–11830, 2020.

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

[73] Hao Yu, Fu Li, Mahdi Saleh, Benjamin Busam, and Slobodan Ilic. Cofinet: Reliable coarse-to-fine correspondences for robust point cloud registration. In *NeurIPS*, 2021.

[74] Wentao Yuan, Benjamin Eckart, Kihwan Kim, V. Jampani, Dieter Fox, and Jan Kautz. DeepGMR: Learning Latent Gaussian Mixture Models for Registration. In *ECCV*, 2020.

[75] Pan Yue, Yang Bisheng, Liang Fuxun, and Dong Zhen. Iterative global similarity points: A robust coarse-to-fine integration solution for pairwise 3d point cloud registration. In *2018 International Conference on 3D Vision (3DV)*, 2018.

[76] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *CVPR*, 2017.

[77] Yu Zhong. Intrinsic shape signatures: A shape descriptor for 3d object recognition. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 689–696, 2009. doi: 10.1109/ICCVW.2009.5457637.