Deep Confidence Guided Distance for 3D Partial Shape Registration

Dvir Ginzburg,[1] **Dan Raviv** [1]

[1] Tel Aviv University
dvirginzburg@mail.tau.ac.il, darav@tauex.tau.ac.il

Abstract

We present a novel non-iterative learnable method for partial-to-partial 3D shape registration. The partial alignment task is extremely complex, as it jointly tries to match between points, and identify which points do not appear in the corresponding shape, causing the solution to be non-unique and ill-posed in most cases.

Until now, two main methodologies have been suggested to solve this problem: sample a subset of points that are likely to have correspondences, or perform soft alignment between the point clouds and try to avoid a match to an occluded part. These heuristics work when the partiality is mild or when the transformation is small but fails for severe occlusions, or when outliers are present. We present a unique approach named Confidence Guided Distance Network (CGD-net), where we fuse learnable similarity between point embeddings and spatial distance between point clouds, inducing an optimized solution for the overlapping points while ignoring parts that only appear in one of the shapes. The point feature generation is done by a self-supervised architecture that repels far points to have different embeddings, therefore succeeds to align partial views of shapes, even with excessive internal symmetries, or acute rotations. We compare our network to recently presented learning-based and axiomatic methods and report a fundamental boost in performance.

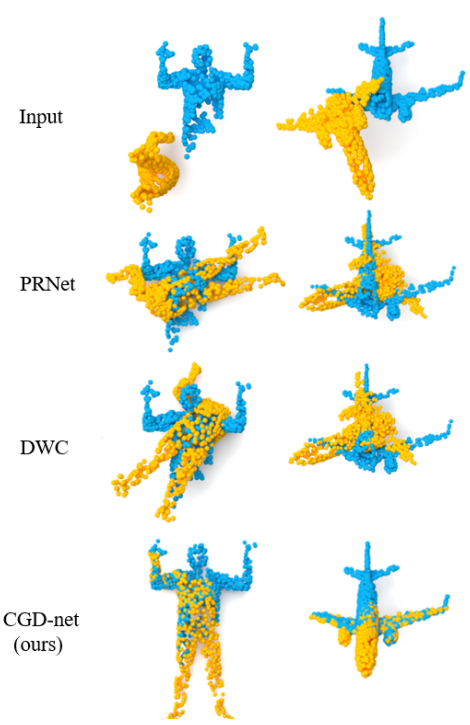

Figure 1: Partial-to-Partial rigid alignment in the full spectrum of the rotation group SO(3) (Blue - source shape, Yellow - target shape). While other state-of-the-art methods struggle to align under such acute rotations and severe partiality, CGD-net succeeds.

1 Introduction

Shape registration is essential for a wide range of applications in computer vision, being the back-bone mechanism in numerous tasks, such as medical imaging (Hajnal and Hill 2001), autonomous driving (Bresson et al. 2017), and robotics (Durrant-Whyte and Bailey 2006).

A canonical work in the field is the Iterative Closest Point (ICP) algorithm (Besl and McKay 1992), which aligns shapes in iterations until convergence, by matching points according to spatial proximity, and then, solving a least-squares problem for the global transformation (Kabsch 1976). ICP is extremely sensitive to noise and initial conditions and tends to get stuck in sub-optimal solutions (Zinsser, Schmidt, and Niemann 2003). To overcome ICP's drawbacks, Deep Closest Point (DCP) (Wang and Solomon 2019a) replaced the Euclidean nearest point step of ICP with a learnable per-point embedding network, followed by a high-dimensional feature-matching stage.

In the partial registration problem, the task is to align partially overlapping segments of a canonical shape. Recent methods have suggested choosing a subset of points as key points based on top-k heuristics and solving the optimal transformation between these essence points to address the partiality problem (Wang and Solomon 2019b; Yew and Lee 2020). A crucial problem with partial 3D shape registration arises from the internal symmetries within the objects. Many point clouds are characterized by co-occurring segments, like the legs of a chair or an airplane's wings. When

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

different co-occurring segments appear in the source and target shapes, these algorithms might still match them, resulting in wrong registrations. Until now, only a few methods have been suggested for the partiality problem in a way that addresses the internal symmetries and avoids the hard selection of corresponding points.

To solve the above issues and provide a reliable yet self-supervised measure for the similarity of partial shapes, we present Confidence Guided Distance (CGD). CGD is a learning-based measure that coalesces points' spatial and latent similarity to a unified score by factoring the Euclidean distance between points with the confidence of their match by a deep neural network. The architecture, named CGD-net, constrains pooled clusters to span the embedding space, causing distant segments to have different embeddings, thus overcoming the internal symmetries problem of previous algorithms.

We analyze the model's performance under popular point-cloud datasets and present superior results by a large margin compared to known state-of-the-art models in all datasets and configurations.

We summarize our key contributions as follows:

- Introduce Confidence Guided Distance, a learnable measure that melds the spatial and latent proximity of points, to overcome previous metrics disadvantages for partial shape correspondence.
- Present a new learnable paradigm for 3D partial shape registration that meets internal symmetries and severe partiality challenges.
- Offer a non-iterative network, both in training and test, granting the method robustness to initial conditions and high noise values.
- Report state of the art results on a variety of synthetic and real-world datasets, while being robust to noise, sparsity, and extreme rotations.

2 Related Work

Point cloud registration has been researched for more than 40 years (Besl and McKay 1992; Chen and Medioni 1992), offering new capabilities in a wide range of computer vision applications. Iterative Closest Point (ICP) (Besl and McKay 1992), one of the seminal works in the field presented an iterative solution to the problem, by alternating between matching points via a simple nearest-neighbor heuristic and calculating the transformation that best describes the above match. While being a cornerstone for further research, ICP is extremely sensitive to noise, outliers, and initial conditions, and thus prone to reach a local-minima. Many variants of ICP have been proposed to solve the caveats in the original solution (Kim et al. 2012; Zinsser, Schmidt, and Niemann 2003). Go-ICP (Yang et al. 2015) offers an outlier detector scheme, while Fitzgibbon (2003) formulates the registration error as the objective function of a non-linear optimization problem (the Levenberg–Marquardt algorithm), refining the alignment until convergence. Still, none of the above methods is suitable for the partial-to-partial framework, and can only handle full-to-full or partial-to-full settings. An important work to ours is ICP Registration Using Invariant Features (Sharp, Lee, and Wehe 2002) which offered a fusion of Euclidean and rotation invariant features as part of the ICP process. It was also among the first to identify that rotation invariant features are less descriptive than the euclidean counterparts by construction, requiring the algorithm to find other information extraction pipelines, as we offer here.

In recent years, many works have tried to solve the 3D rigid registration problem by harvesting Graph Neural Networks (GNNs) (Li et al. 2018) capabilities to create descriptive per-point features. Deep Closest Point (DCP) (Wang and Solomon 2019a) was among the first to propose a feature learning scheme, followed by a soft alignment of the two-point clouds instead of the spatial matching step presented in ICP. As DCP uses all points in the soft alignment step, it is sensitive to outliers or noise and does not work well for the partiality problem. Methods that follow DCP (Yuan et al. 2018; Aoki et al. 2019) have tried to enhance DCP's results using an iterative scheme similar to ICP. As with ICP, robustness to outliers and initial conditions remains an unsolved issue. PRNet (Wang and Solomon 2019b) chooses a subset of points and solves the alignment only for the limited set to overcome the problem of outliers and partiality. Deep Weighted Consensus (DWC) (Ginzburg and Raviv 2021) suggests a different selection strategy, where a sampling distribution is defined based on the confidence of each source point in its alignment. DWC offers multiple possible transformations in parallel, and chooses the best parameter based on the Chamfer Distance (Barrow et al. 1977a) between the source and the transformed target. DWC is irrelevant for partial-to-partial alignment, as choosing based on the minimal chamfer distance between partial samples leads to results that are far from the true transformation, as exemplified in figure 1.

The partial registration problem is the most relevant for real-world scenarios, thus became extremely researched in recent years (Choy, Dong, and Koltun 2020; Yew and Lee 2020). As mentioned, PRNet (Wang and Solomon 2019b) offered a hard sampling heuristic to choose the points subset the are present both in the source and target point clouds and compute the transformation only between the subsets. Other methods as (Huang et al. 2021; Dang, Wang, and Salzmann 2020) offered attention mechanisms that inspect both clouds jointly, weighting points with high probability to appear in both partial shapes. While surpassing previous methods, co-occur segments and high noise values were still a problem due to local feature generation and reliance on the initial state of the point clouds. Another interesting concept was introduced in (Yan et al. 2021; Yang, Yan, and Huang 2019) where instead of sampling relevant points, a decoder network generated a full shape from the partial scans, offering a method to overcome the partiality problem. While this is a novel line of work, it is less relevant for generalization tasks, where we train on one dataset and infer on another.

3 Deep Confidence Guided Distance

Given a transformation $\mathcal{T}$, a source shape $\mathcal{X}$, and a target $\mathcal{Y}$, the ability to define a distance measure $d(\mathcal{T}(\mathcal{X}), \mathcal{Y})$ that is invariant to outliers and non-overlapping segments is crucial for the success of partial-to-partial algorithms. Chamfer distance (CD) (Barrow et al. 1977a), a popular distance measure between 3D objects, computes the distance from each source point to its target shape nearest neighbor, and vice-versa. We denote $nn_{\mathcal{C}}(p)$ as the euclidean closest point to p in a point cloud $\mathcal{C}$, and $d_{nn}(p, \mathcal{C})$ to be that euclidean distance. Formally, the metric takes the following form:

$$CD(\mathcal{T}(\mathcal{X}), \mathcal{Y}) = \sum_{i=1}^{i=|\bar{\mathcal{X}}|} d_{nn}(\mathcal{T}(x_i), \mathcal{Y}) + \sum_{j=1}^{j=|\bar{\mathcal{Y}}|} d_{nn}(y_i, \mathcal{T}(\mathcal{X})) \quad (1)$$

$$\text{where} \quad d_{nn}(p, \mathcal{C}) = ||p - nn_{\mathcal{C}}(p)||_2^2 \quad \text{and} \quad nn_{\mathcal{C}}(p) = \underset{c_i \in \mathcal{C}}{\operatorname{argmin}} ||p - c_i||_2^2.$$

With $\bar{\mathcal{X}}$ being the set of points that uphold $d_{nn}(x_i, \mathcal{Y}) < 2d_s$ where d_s is the maximal sampling distance of the point clouds, as matches with higher d_{nn} are considered as outliers. Equation 1 fails for partial-to-partial alignment, or when the number of outliers is high. This is because CD is optimal in expectation, whereas for the partiality setting, one should strive to ignore segments that do not have a correspondence in the target point cloud. The problem of CD with partiality is discussed extensively in the literature (Barrow et al. 1977b) and exemplified in the results section (Sec. 5). Deep neural networks have the ability to represent points in contextualized latent spaces influenced by the shape topology and local geometry. Given such deep embeddings $\hat{h}_{x_i}, \hat{h}_{y_j}$ of points x_i, y_j, a natural choice for the latent similarity of the points is the cosine similarity:

$$\cos(x_i, y_j) = \frac{\hat{h}_{x_i} \cdot \hat{h}_{y_j}}{||\hat{h}_{x_i}||_2 \cdot ||\hat{h}_{y_j}||_2}. \quad (2)$$

For compactness, we denote by $\mathcal{P} \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{Y}|}$ the soft alignment mapping, where

$$\mathcal{P}_{i,j} = \cos(x_i, y_j). \quad (3)$$

Intuitively, given the per-point embeddings $\hat{h}_{x_i}$ and $\hat{h}_{y_j}$, $\mathcal{P}_{i,j}$ represents the latent similarity between x_i and y_j. We propose Confidence Guided Distance (CGD), a learning-based measure that fuses together the spatial and latent proximity, in a way that penalizes point clouds that are close under the Chamfer distance metric but are highly dissimilar in the latent space. Confidence Guided Distance is formulated as:

$$CGD(\mathcal{T}(\mathcal{X}), \mathcal{Y}) = \sum_{i=1}^{i=|\bar{\mathcal{X}}|} exp(-\gamma \mathcal{P}_{i,q}) d_{nn}(\mathcal{T}(x_i), \mathcal{Y}) + \sum_{j=1}^{j=|\bar{\mathcal{Y}}|} exp(-\gamma \mathcal{P}_{r,j}) d_{nn}(y_j, \mathcal{T}(\mathcal{X})) \quad (4)$$

$$\text{with } q = nn_{\mathcal{Y}}(\mathcal{T}(x_i)), r = nn_{\mathcal{T}(\mathcal{X})}(y_j).$$

where γ is a scalar that determines the weight given to the latent similarity and is conditioned by the shape's scale and sampling density. For high $\mathcal{P}_{ij}$, the combined measure is lower, while for matches with low or negative $\mathcal{P}_{ij}$ the distance increases.

In the ablation study (Sec. 5.5) we provide analysis on the importance of the CGD metric and the implications of using Chamfer Distance as the consensus metric instead.

4 Architecture

The following section outlines the architecture building blocks needed for the evaluation of the CGD measure.

4.1 Hierarchical Feature Extraction

Confidence Guided Distance Network (CGD-net) presents a hierarchical graph neural network, inspired by PointNet++ (Qi et al. 2017) for intrinsic per-point feature extraction. The decision to use a hierarchical network arises from the repulsion loss (Sec. 4.3) that pushes spatially far segments to have unique embeddings. As for the input point cloud representation, we adopt the rotation invariant (RI) features presented in DWC. Given rotation invariant features for the source and target shapes $\mathcal{X}_{RI}, \mathcal{Y}_{RI}$, the input topology is defined by the k nearest neighbors of each point. CGD-net uses EdgeConv (Wang et al. 2019) as the graph convolution operation, instead of PointNet++ suggested convolution, as PointNet operator is heavily biased toward spatial similarity, making it impractical for large rotations or high noise values. In EdgeConv, for a point x_i and its neighborhood $\mathcal{N}_{x_i}$, the output embedding of x_i is:

$$h_{x_i} = \max_{x_j \in \mathcal{N}_{x_i}} f_h([x_i || x_j]), \quad (5)$$

where f_h is a learnable function, $||$ is concatenation, and the $\max$ operation is applied on the feature dimension. $h_{\mathcal{X}}$ is then sampled by spatial FPS (Eldar et al. 1997) to guarantee equal representation to all input-segments in deeper levels of the network. The pooling step enlarges the receptive field of the network, which is used in order to repel faraway points. The up-pooling scheme and skip-connections are similar to the ones used in PointNet++. An illustration of the process is depicted in figure 2.

4.2 Soft Correspondence Sampling

CGD presents a novel metric for ranking possible transformations and choosing the most plausible mapping between the source and target point clouds. To create the initial possible maps, we present a unique sampling method that samples efficiently the candidate points. RANSAC (Fischler and

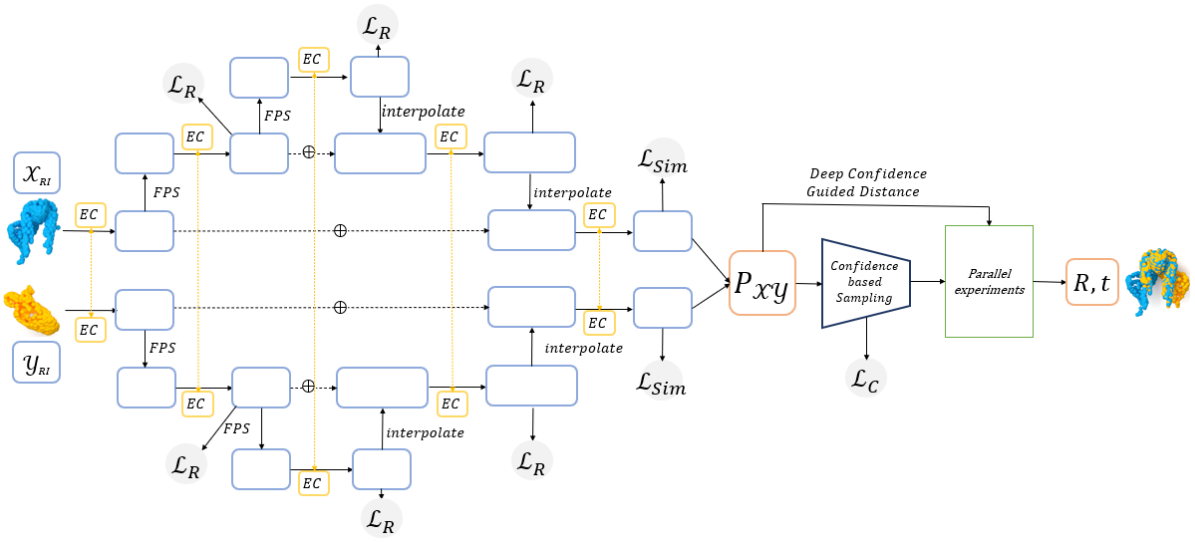

Figure 2: CGD-net is comprised of three steps: (a) Embed points via the hierarchical feature extraction network (Sec. 4.1). (b) Construct the confidence based sampling distribution of source points, formed from the soft correspondence matrix $\mathcal{P}_{\mathcal{X}\hat{\mathcal{X}}}$ (Sec. 4.2). (c) Conduct multiple parallel experiments, and choose the experiment achieving the lowest CGD (Sec. 3). RI are the rotation invariant features, EC stands for EdgeConv (Wang et al. 2019), and FPS (Eldar et al. 1997) refers to Euclidean farthest point sampling. $\mathcal{L}_{\mathcal{R}}, \mathcal{L}_{Sim}$ are the per-point repulsion and similarity loss functions, respectively, $\mathcal{L}_{\mathcal{C}}$ is the self-supervised contrastive loss (section 4.3). The plus sign stands for concatenation.

Bolles 1981), which samples randomly the candidate points, requires a massive amount of possible transformations to converge to the correct solution. To overcome this, we create a smart sampling algorithm, which samples points that have a higher probability to match the target point cloud, thus inducing a procedure with much fewer required candidate transformations.

The soft correspondence map $\mathcal{P} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ relates to the latent proximity between x_i and y_j and is defined by the cosine similarity between the point embeddings $\hat{h}_{x_i}, \hat{h}_{y_j}$ (Eq. 3). Source points with no matched target are characterized by rows in $\mathcal{P}$ with only low similarity values. This observation induces the *confidence score* $\mathcal{C}$ of a point x_i as the highest value in x_i's row in the column of the normalized soft-alignment matrix:

$$\mathcal{C}_{x_i} = \max_j \hat{\mathcal{P}}_{i,j} \quad \text{where } \hat{\mathcal{P}}_{*,j} = \frac{\mathcal{P}_{*,j}}{\sum_{i=1}^{|\mathcal{X}|} \mathcal{P}_{ij}}, \quad \mathcal{C} \in \mathbb{R}^{|\mathcal{X}|}. \quad (6)$$

$\mathcal{C}_{x_i}$ is high if and only if x_i is the highest correlated point to some point in $\mathcal{Y}$. $\mathcal{C}$ is normalized to be a valid probability function $\hat{\mathcal{C}}$ which defines the sampling probability distribution S over the source points that partake in the alignment. S is a categorical sampling distribution over the set of source points $X = \{x_i | x_i \in \mathcal{X}\}$ with the following probability mass function:

$$s_X(x_i) = S(X = x_i) = \hat{C}_{x_i} \quad (7)$$

CGD-net resembles RANSAC (Fischler and Bolles 1981) and samples multiple possible transformations. Unlike RANSAC and its follow-up works in the registration domain, CGD-net samples points based on the learned confidence measure of the source points in their alignment (Sec. 4.2). To do so, Q[1] points from $\mathcal{X}$ are sampled according to the distribution S, and are divided into $V = \lfloor Q/r \rfloor$ experiment groups. The experiment size r is constrained to $r \geq 3$ as three is the minimal number of co-linear points needed to find the transformation parameters based on the SVD solution (Kabsch 1976).

The output of the division step is V experiments, where for each experiment v the optimal transformation $\mathcal{T}^v$ parameters that best represent the alignment between the r sampled source points $\{x_1, x_2, \ldots, x_r\}$, and their corresponding points $\{\pi(x_1), \pi(x_2), \ldots, \pi(x_r)\}$ are calculated using the SVD algorithm (Kabsch 1976). The mapping π is set to be the maximum likelihood solution derived from $\mathcal{P}$. To identify the best $\mathcal{T}^v$, we use CGD (Sec. 3), and set:

$$\mathcal{T}^{opt} = \underset{\mathcal{T}^v}{\operatorname{argmin}} \, CGD(\mathcal{T}^v(\mathcal{X}), \mathcal{Y}). \quad (8)$$

The system's novelty relies upon the power to not only suggest a transformation but also assess its correctness in a learnable manner, allowing the consensus stage to rank different suggestions. We provide evidence for CGD-net sampling strategy in the ablation study (Sec. 2), where we show a substantial performance decrease when using RANSAC.

4.3 Self-Supervised Similarity Learning

CGD-net objective function is comprised of three different loss terms. The repulsion loss role is to assign each cluster

[1] Q is a hyper-parameter and was set to $\frac{|\mathcal{X}|}{10}$ in the experiments.

of points with a different embedding space, the similarity loss encourages close embeddings for neighbor points, and the contrastive loss motivates points that appear both in the source $\mathcal{X}$ and target $\mathcal{Y}$ to have the same embeddings.

A consequential flaw of most prior works is features locality, causing spatially far co-occurring segments to have similar latent representation, yielding wrong correspondences and bad transformation solutions. CGD-net constrains features of spatially far points to have dissimilar features through a metric loss on the sampled points after each FPS/interpolation step (Fig. 2). Given n_l pooled points, CGD-net repulsion loss takes the form of:

$$\mathcal{L}_{R_l} = \sum_{i=0}^{n_l-1} \sum_{j=0, j\neq i}^{n_l-1} d(x_i, x_j)(\cos(h^l_{x_i}, h^l_{x_j}))^\beta \tag{9}$$

where h^l_i is the embedding h of point x_i at layer l, $\cos(\cdot)$ refers to the cosine similarity, as in equation 2, and $d(x_i, x_j)$ is the Euclidean distance between the points. β is a hyperparameter inspired by the focal loss (Lin et al. 2017), whose purpose is to focus on points with higher metric similarity. The function of $d(x_i, x_j)$ is to have higher weight on distant pairs. $\mathcal{L}_{R_l}$ obliges the network to span the embeddings space and describe far points differently, even if they have locally similar geometric attributes. The unified repulsion loss is the sum of the repulsion terms, where each term is being multiplied by its index to adaptively constraint the features:

$$\mathcal{L}_R = \sum_{l=1}^{|L|} g(l)\mathcal{L}_{R_l} \tag{10}$$

where $|L|$ is the number of downsampling/interpolate layers, and $g(l)$ is an increasing function putting more weight to higher layers[2].

For h_{x_i}, the final embedding, an extension of the repulsion loss is used to drive close points to have similar embeddings. The similarity loss $\mathcal{L}_{Sim}$ strives to create a smooth embedding space, which in turn increase the probability of smooth correspondence maps. Given the neighborhood $\mathcal{N}_{x_i}$ of x_i, the similarity loss takes the form of:

$$\mathcal{L}_{Sim} = \sum_{i=0}^{|\mathcal{X}|-1} \sum_{j\notin\mathcal{N}_{x_i}} d(x_i, x_j)(\cos(h_{x_i}, h_{x_j}))^\beta + \tag{11}$$
$$\sum_{i=0}^{|\mathcal{X}|-1} \sum_{j\in\mathcal{N}_{x_i}} \frac{1}{\max(d(x_i, x_j), \epsilon)}(1-\cos(h_{x_i}, h_{x_j}))^\beta$$

where $1-\cos(h_{x_i}, h_{x_j})$ pushes close points to have a cosine similarity close to 1. The RHS is divided by d to emphasize the similarity between close points.

During **training**, the target shape is predefined as $\mathcal{Y} = \mathcal{T}(\mathcal{X})$ where $\mathcal{T}$ is a composition of rotation, translation and Gaussian noise. This is of course true only during training as during test different scans of the same object are used. Such target formulation defines the dense mapping during training to be $\pi(x_i) = y_i$. Our results imply that such setting is a good proxy for the real task. CGD-net follows DWC and uses a contrastive loss that strives for rotation and translation invariant features. The metric loss used is:

$$\mathcal{L}_C = \sum_{i=0}^{|\mathcal{X}|-1} \sum_{j\notin\mathcal{N}_{y_i}} \cos(h_{x_i}, h_{y_j}) + \tag{12}$$
$$\sum_{i=0}^{|\mathcal{X}|-1} \sum_{j\in\mathcal{N}_{y_i}} 1-\cos(h_{x_i}, h_{y_j}).$$

CGD-net unified loss is

$$\mathcal{L} = \lambda_R\mathcal{L}_R + \lambda_{Sim}\mathcal{L}_{Sim} + \lambda_C\mathcal{L}_C. \tag{13}$$

Ablation on the significance of the different loss functions appears in the ablation study 5.5.

5 Experimentation

CGD-net performance is compared to numerous learnable methods as DCP, PRNet, RPM-net, and DWC, as well as to axiomatic methods as ICP and Go-ICP. We evaluate the registration capability on multiple datasets as ModelNet40 (Wu et al. 2015), Stanford Bunny (Turk and Levoy 1994), 3DMATCH (Zeng et al. 2017) and FAUST scans (Bogo et al. 2014). The discrepancy between the predicted R, t and the ground truth transformations is measured using the root mean squared error (RMSE) metric. We measure the rotation difference using Euler angels and report the score in units of degrees. We train and evaluate CGD-net on a single RTX8000 GPU.

5.1 Implementation Details

The hierarchical feature extraction network consists of two downsampling modules, with a FPS factor of 0.5, 0.25 respectively, leaving 125 points in the bottleneck of the network. The per-point output feature after the last interpolation step is concatenated with the initial $\mathcal{X}_{RI}$ features, and passes through 3 more EC layers. The output feature dimension is 1024. The network consists of 8.9 million learnable parameters, resulting in an architecture of size $\sim$ 15.4MB in size. We use LeakyRelu activation function (Xu et al. 2015) with a negative slope of 0.2, and NormLayer normalization (Ba, Kiros, and Hinton 2016) for all the layers.

The initial learning rate is set to $5e^{-4}$ with a multiplicative scheduler of $\gamma = 0.9$ every 10 epochs. The entire training takes up to 50 epochs for the longest configuration. For the transformation solver (Kabsch 1976), we use the differentiable SVD layer provided by PyTorch (Paszke et al. 2019), where we take advantage of batch-parallelization and reshape the sampled points such that each subsampled set of k points acts as a new sample in a batch. This provides two orders of magnitude speed-up compared to naively evaluating the SVD per subsample, and, as we solve the least-squares problem for k points, instead of kl as previous methods do, our SVD solver is up to 10 times faster than other methods

[2] In practice we use $g(l) = l$

	Unseen point clouds		Unseen categories		Gaussian noise		FAUST	
Model	**RMSE**(R)	**RMSE**(t)	**RMSE**(R)	**RMSE**(t)	**RMSE**(R)	**RMSE**(t)	**RMSE**(R)	**RMSE**(t)
ICP	24.39	0.78	26.35	0.85	27.42	0.89	21.92	0.69
Go-ICP	22.54	0.82	23.82	0.86	24.55	0.85	19.17	0.73
DCP	12.36	0.67	13.86	0.61	14.73	0.65	13.38	0.64
PointNetLK	12.12	0.55	14.67	0.59	14.98	0.61	11.01	0.66
IT-Net	15.32	0.59	16.70	0.61	19.13	0.66	15.22	0.65
DWC	19.07	0.53	19.26	0.61	20.08	0.67	19.08	0.41
PRNet	5.93	0.32	6.01	0.38	6.93	0.41	6.32	0.34
RPM-net	4.93	0.1	5.21	0.21	5.09	0.25	11.35	0.22
CGD-net (ours)	**2.90**	**0.09**	**3.12**	**0.17**	**3.37**	**0.19**	**3.81**	**0.15**

Table 1: Partial-to-partial evaluations. CGD-net shows a considerable performance gain in all evaluation metrics on ModelNet40 (first three evaluations) and FAUST datasets, both for R and t.

using the SVD solver.

To create the partial view of a shape a 3D viewpoint is randomly sampled, defining a 2.5D partial shape with 60% points of the original point cloud. Such partiality leads to pairs with $< 40\%$ overlapping ratio. A translation t sampled from $[-0.5, 0.5]$ and rotation R bounded to $[0°, 60°]$ in each axis define the random augmentations applied separately on the source and target shapes. We choose these augmentations to ease comparison to previous methods that limit the transformation by these parameters. However, most real-world applications encounter more extreme rotations, as in the case of multi-view registration for example.

5.2 ModelNet40

ModelNet40 (Wu et al. 2015) consists of 12,411 synthetic CAD models[3] from various categories, such as planes, furniture, cars, etc. While ModelNet40 offers the per-point normal information, CGD-net uses only the spatial location as input information. We follow previous state-of-the-art methods and conduct 3 experiments on the ModelNet40 dataset: (1) Random train/test split - The official train-test split of ModelNet40, where 9843 samples are defined as train samples (80% of the dataset), and the rest 2,568 samples are the test set. The results for this test configuration are presented in table 1. The relative improvement compared to RPM-net, the second-best model is 2.03, where compared to DWC, the margin is higher than 16.07. We associate the bad performance of DWC on the partial-to-partial registration problem to Chamfer's inability to "ignore" points with no matched target, contrary to CGD. (2) Unseen categories - In this setting we train on the first 20 categories and test on the rest. This offers an evaluation of the generalization ability of the methods to unseen topologies. The trend is similar to the previous experiment, where CGD-net provides superior results, with sizeable margins of 2.09 R RMSE compared to the second-best model (RPM-net). (3) Random noise - Noise resilience is crucial for 3D registration systems, as 3D scans from real-world sensors are known to be unstable under various factors. The noise durability test is evaluated by adding random Gaussian noise to the point clouds from the distribution $\mathcal{N}(0, 0.05)$. Same as in the previous two test configurations, CGD-net has the best results, where the R RMSE difference is 16.71 compared to DWC and 1.72 compared to RPM-net. We ascribe this to our sampling scheme that selects an optimal Confidence Guided Distance solution from the experiment groups and results in outlier-free shape registration.

5.3 FUAST Scans

The FAUST scans dataset (Bogo et al. 2014) contains 300 real human scans, acquired by a real-world scanner (3dMD). The scans are partial by nature (occlusions), noisy, and contain a high number of outliers, as opposed to the ModelNet40 dataset. we do not train on FAUST scans and use the best-performing model [4] of each test case on ModelNet40 as the evaluated network. We do not train on real-world data as it is usually expensive to acquire and limited in size. Table 1 provides the results on FAUST scans for the partial-to-partial task by the different models. The results are consistent with previous tests, where CGD-net has the best results for all the evaluated metrics. We hold CGD-net repulsion loss (4.3) responsible for the performance boost, as human body models have many co-occurring segments (hands, palms, legs) which must have unique embeddings for a successful alignment.

5.4 Generalization on Different Modalities

We provide qualitative results on the Stanford 3D Scanning, and 3DMATCH datasets in figure 3. While we follow monumental works as DCP, PRNet, PointNetLK, or RPM-net and focus on 3D objects, we present visual results on 2.5D indoor scans from the 3DMATCH dataset. 3DMATCH samples are noisy, and vary dramatically from ModelNet in their statistical attributes. Nevertheless, a CGD-net trained on the

[3]Partial models that are globally symmetric around the axes have been filtered out.

[4]RNe (Li et al. 2010) is used as normal approximation method for RPM-net which requires normal-to-the surface.

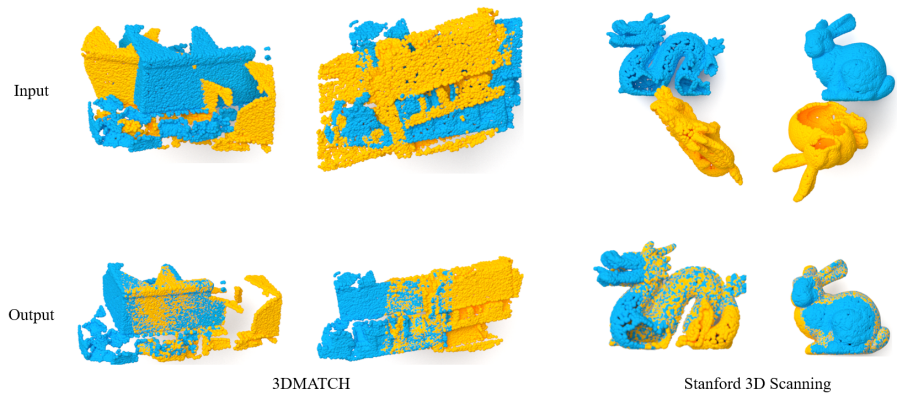

Figure 3: Generalization capability. Visual results of CGD-net on 3DMATCH (left) and Stanford 3D Scanning (right).

	ModelNet40	FAUST
Mode	**RMSE**(R)	**RMSE**(R)
(i) CGD metric	11.53	9.52
(ii) Repulsion loss	6.18	5.45
(iii) Similarity loss	5.39	5.13
(iv) Metric loss	6.79	6.14
(v) Hierarchical network	6.32	5.88
Full method	**2.90**	**3.81**

Table 2: Ablation study on ModelNet40. The "Mode" column states the switched off component.

synthetic ModelNet40 was able to extract correct registrations on this **real-world** dataset. The Stanford 3D Scanning dataset is also visually and statistically different from ModelNet40 dataset. Yet, CGD-net offers great abilities here as well.

5.5 Ablation Studies

The following contributions have been evaluated in the ablation studies: (i) CGD metric - Use Chamfer distance (Eq. 1) as the similarity metric for each experiment instead of CGD. (ii) Repulsion loss - Setting $\lambda_R = 0$ in equation 13, ignoring the hierarchical repulsion. (iii) Similarity loss - Setting $\lambda_{Sim} = 0$ in equation 13, ignoring the similarity loss. (iv) Metric loss - Replace equation 13 with a loss directly optimizing the transformation parameters:

$$\mathcal{L} = ||R^T_{\mathcal{XY}} R^{gt}_{\mathcal{XY}} - I||^2 + ||t_{\mathcal{XY}} - t^{gt}_{\mathcal{XY}}||^2. \quad (14)$$

(v) Hierarchical network - Use DGCNN (Wang et al. 2019) backbone instead of PointNet++. As a result, the repulsion loss is discarded as well.

The use of CGD metric provides the largest performance gain, improving the R RMSE by 8.63. This affirms our hypothesis that Chamfer distance as a metric to assess partial shapes similarity is irrelevant. The gain of using the repulsion loss is also substantial, providing 3.28 improvement in R RMSE. We associate it with the ability of CGD-net to identify co-occurring segments, creating unique embeddings, and resulting in improved alignments.

6 Limitations

Several design choices of CGD-net are starting points for future work. One limitation of CGD-net is the use of the dense alignment matrix $\mathcal{P}$, as the dimensions of such map are $|\mathcal{X}| \times |\mathcal{Y}|$. Using such a soft map is sometimes unfeasible, as in the case of depth sensors for autonomous driving, acquiring $\approx 100,000$ points per scene. Accordingly, one interesting future work is to produce a hierarchical soft alignment scheme, where $\mathcal{P}$ is evaluated only for sampled clusters, and propagated downwards to the dense point cloud.

7 Summary

In this work we examine the current state of partial-to-partial 3D rigid correspondence, particularly for large rotations, a high number of outliers, or when the input shapes have similar co-occurring segments, which are typical scenarios in real-world applications. We present CGD, a new learning-based measure that fuses latent and spatial information to evaluate the similarity of point clouds and solves previous methods' inability to approximate partial point-cloud similarity. CGD uses features extracted by CGD-net, a hierarchical architecture that samples cluster centers and constrains them to have distant embeddings via the presented repulsion loss, forcing remote co-occurring segments to have unique embeddings.

Acknowledgments

This work is partially funded by the Zimin Institute for Engineering Solutions Advancing Better Lives, the Israeli consortiums for soft robotics and autonomous driving, the Nicholas and Elizabeth Slezak Super Center for Cardiac Research and Biomedical Engineering at Tel Aviv University and TAU Science Data and AI Center.

References

Aoki, Y.; Goforth, H.; Srivatsan, R. A.; and Lucey, S. 2019. Pointnetlk: Robust & efficient point cloud registration using pointnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7163–7172.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Barrow, H. G.; Tenenbaum, J. M.; Bolles, R. C.; and Wolf, H. C. 1977a. Parametric correspondence and chamfer matching: Two new techniques for image matching. Technical report, SRI INTERNATIONAL MENLO PARK CA ARTIFICIAL INTELLIGENCE CENTER.

Barrow, H. G.; Tenenbaum, J. M.; Bolles, R. C.; and Wolf, H. C. 1977b. Parametric correspondence and chamfer matching: Two new techniques for image matching. Technical report, SRI INTERNATIONAL MENLO PARK CA ARTIFICIAL INTELLIGENCE CENTER.

Besl, P. J.; and McKay, N. D. 1992. Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, 586–606. International Society for Optics and Photonics.

Bogo, F.; Romero, J.; Loper, M.; and Black, M. J. 2014. FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3794–3801.

Bresson, G.; Alsayed, Z.; Yu, L.; and Glaser, S. 2017. Simultaneous localization and mapping: A survey of current trends in autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2(3): 194–220.

Chen, Y.; and Medioni, G. 1992. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3): 145–155.

Choy, C.; Dong, W.; and Koltun, V. 2020. Deep global registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2514–2523.

Dang, Z.; Wang, F.; and Salzmann, M. 2020. Learning 3D-3D Correspondences for One-shot Partial-to-partial Registration. *CoRR*, abs/2006.04523.

Durrant-Whyte, H.; and Bailey, T. 2006. Simultaneous localization and mapping: part I. *IEEE robotics & automation magazine*, 13(2): 99–110.

Eldar, Y.; Lindenbaum, M.; Porat, M.; and Zeevi, Y. 1997. The farthest point strategy for progressive image sampling. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 6 9: 1305–15.

Fischler, M. A.; and Bolles, R. C. 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM*, 24(6): 381–395.

Fitzgibbon, A. W. 2003. Robust registration of 2D and 3D point sets. *Image and vision computing*, 21(13-14): 1145–1153.

Ginzburg, D.; and Raviv, D. 2021. Deep Weighted Consensus: Dense correspondence confidence maps for 3D shape registration. arXiv:2105.02714.

Hajnal, J. V.; and Hill, D. L. 2001. *Medical image registration*. CRC press.

Huang, S.; Gojcic, Z.; Usvyatsov, M.; and Andreas Wieser, K. S. 2021. PREDATOR: Registration of 3D Point Clouds with Low Overlap. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Kabsch, W. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5): 922–923.

Kim, V. G.; Li, W.; Mitra, N. J.; DiVerdi, S.; and Funkhouser, T. 2012. Exploring collections of 3d models using fuzzy correspondences. *ACM Transactions on Graphics (TOG)*, 31(4): 1–11.

Li, B.; Schnabel, R.; Klein, R.; Cheng, Z.; Dang, G.; and Jin, S. 2010. Robust normal estimation for point clouds with sharp features. *Computers & Graphics*, 34(2): 94–106.

Li, R.; Wang, S.; Zhu, F.; and Huang, J. 2018. Adaptive graph convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Lin, T.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. *CoRR*, abs/1708.02002.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, 5099–5108.

Sharp, G. C.; Lee, S. W.; and Wehe, D. K. 2002. ICP registration using invariant features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1): 90–102.

Turk, G.; and Levoy, M. 1994. Zippered Polygon Meshes from Range Images. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '94, 311–318. New York, NY, USA: Association for Computing Machinery. ISBN 0897916670.

Wang, Y.; and Solomon, J. M. 2019a. Deep closest point: Learning representations for point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3523–3532.

Wang, Y.; and Solomon, J. M. 2019b. Prnet: Self-supervised learning for partial-to-partial registration. *arXiv preprint arXiv:1910.12240*.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5): 1–12.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1912–1920.

Xu, B.; Wang, N.; Chen, T.; and Li, M. 2015. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Yan, Z.; Yi, Z.; Hu, R.; Mitra, N.; Cohen-Or, D.; and Huang, H. 2021. Consistent Two-Flow Network for Tele-Registration of Point Clouds. *IEEE Transactions on Visualization and Computer Graphics*, 1–1.

Yang, J.; Li, H.; Campbell, D.; and Jia, Y. 2015. Go-ICP: A globally optimal solution to 3D ICP point-set registration. *IEEE transactions on pattern analysis and machine intelligence*, 38(11): 2241–2254.

Yang, Z.; Yan, S.; and Huang, Q. 2019. Extreme Relative Pose Network under Hybrid Representations. *CoRR*, abs/1912.11695.

Yew, Z. J.; and Lee, G. H. 2020. Rpm-net: Robust point matching using learned features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11824–11833.

Yuan, W.; Held, D.; Mertz, C.; and Hebert, M. 2018. itnet. *CoRR*, abs/1811.11209.

Zeng, A.; Song, S.; Nießner, M.; Fisher, M.; Xiao, J.; and Funkhouser, T. 2017. 3DMatch: Learning Local Geometric Descriptors from RGB-D Reconstructions. In *CVPR*.

Zinsser, T.; Schmidt, J.; and Niemann, H. 2003. A refined ICP algorithm for robust 3-D correspondence estimation. In *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, volume 2, II–695.