Unsupervised Point Cloud Registration with Self-Distillation

Christian Löwens[1,2]
christian.loewens@bosch.com
Thorben Funke[1]
thorben.funke@bosch.com
Andre Wagner[1]
andre.wagner2@bosch.com
Alexandru Paul Condurache[2,3]
alexandrupaul.condurache@bosch.com

[1] Bosch Research
[2] University of Lübeck
[3] Robert Bosch GmbH
Automated Driving

Abstract

Rigid point cloud registration is a fundamental problem and highly relevant in robotics and autonomous driving. Nowadays deep learning methods can be trained to match a pair of point clouds, given the transformation between them. However, this training is often not scalable due to the high cost of collecting ground truth poses. Therefore, we present a self-distillation approach to learn point cloud registration in an unsupervised fashion. Here, each sample is passed to a teacher network and an augmented view is passed to a student network. The teacher includes a trainable feature extractor and a learning-free robust solver such as RANSAC. The solver forces consistency among correspondences and optimizes for the unsupervised inlier ratio, eliminating the need for ground truth labels. Our approach simplifies the training procedure by removing the need for initial hand-crafted features or consecutive point cloud frames as seen in related methods. We show that our method not only surpasses them on the RGB-D benchmark 3DMatch but also generalizes well to automotive radar, where classical features adopted by others fail. The code is available at github.com/boschresearch/direg.

1 Introduction

The goal of rigid point cloud registration is to align two or more point clouds by finding the optimal rigid transformation. It is a fundamental task in fields such as 3D reconstruction [7, 13], augmented reality [18, 20] and autonomous navigation [27, 33]. Traditionally, these applications have relied on learning-free heuristics [7, 11, 23] or supervised deep learning approaches [8, 15, 32] that require ground truth poses during training. While these methods are effective, they often lack scalability across diverse scenarios. In the automotive context, ground truth surveys are expensive and limited in size due to their professional sensor setup [17]. However, crowdsourced data [21] can be obtained on a large scale from mass-produced vehicles, providing more diverse data as shown in Fig. 1. Unsupervised point cloud registration can utilize this and generate high-quality pseudo labels at a low cost.

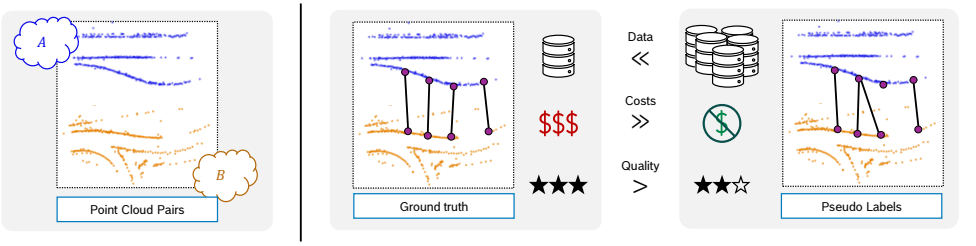
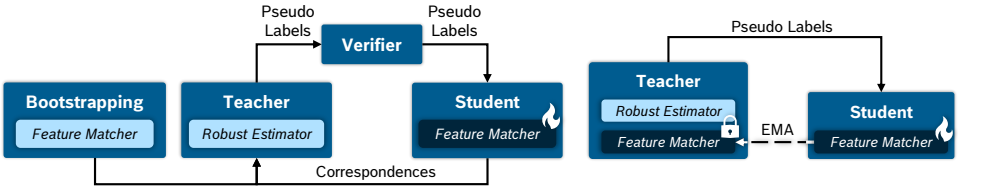


Figure 1: **Motivation for Unsupervised Point Cloud Registration.** Especially in the automotive context, the collection of ground truth poses is costly and limited in size. Crowd-sourced data from consumer-grade cars, on the other hand, contains orders of magnitude more unlabeled data. Using our approach, we can leverage this data and generate pseudo labels with a quality close to ground truth.

In recent years, some methods have emerged to overcome the need for ground truth [16, 24, 30]. Notably, Yang *et al*. [30] draw inspiration from student-teacher architectures and propose Self-supervised Geometric Perception (SGP). There, the student is a trainable feature matcher outputting putative correspondences and the teacher is a learning-free robust solver estimating rigid transformations (see Fig. 2a). The transformations are then used as pseudo labels to supervise the student for several epochs before new improved labels are generated again. Revisiting SGP, we adopt some of the research findings in self-distillation, where student and teacher are both parameterized feature extractors, leading to remarkable unsupervised performance in the image domain [5, 14]. For this, we update the teacher by an exponential moving average (EMA) of the student's parameters to provide continuously better pseudo labels on the fly. Furthermore, we propose a more simplified framework compared to SGP (see Fig. 2b) by eliminating the pseudo label verifier and the reliance on hand-crafted bootstrap features. This enhances the adaptability of our method to various modalities, whereas the previously mentioned SGP components may require careful adjustments or may not work at all. In this process, we converge on an architecture similar to the most recently published work Extend Your Own Correspondences (EYOC) [17]. However, EYOC focuses specifically on distant LiDAR point cloud registration, while we investigate the general unsupervised problem in diverse environments. Moreover, we show that the augmentation technique used in EYOC and others makes it difficult to robustly bootstrap the training process, and thus remove the augmentation for the teacher's input.

Our method surpasses the performance of both SGP and EYOC on 3DMatch and a radar dataset while having fewer hyperparameters requiring tuning. In summary, we propose a self-**di**stillation approach for **reg**istration (DiReg) with the following key contributions:

- We simplify unsupervised point cloud registration by removing the need for a pseudo label verifier, hand-crafted bootstrap features, and progressive datasets as seen in earlier work, which all need careful adjustments when used for data from various sensors.
- We demonstrate that the data augmentation commonly used for the teacher's input can impede robust bootstrapping in unsupervised settings and how to overcome it.
- We show that our approach generalizes well across modalities as evidenced by its performance on RGB-D and automotive radar point clouds.



(a) SGP [30] (b) DiReg (Ours)

Figure 2: We simplify the SGP algorithm by removing the verifier and its classical features used for bootstrapping. We also reinterpret its student-teacher analogy in view of self-distillation. While dark boxes indicate trainable methods, the teacher's feature extractor is not trained but instead updated using an exponential mean average (EMA).

2 Related Work

2.1 Point Cloud Registration

In point cloud registration, we align two point clouds either directly or estimate a set of correspondences first. Usually, those correspondences are computed by trained feature matchers and then passed to a robust estimator filtering out the outliers to predict the transformation.

Feature Matchers: Prior to learning-based methods, Fast Point Feature Histograms (FPFH) [23] uses hand-crafted features to capture the local geometry. Fully Convolutional Geometric Features (FCGF) [8] introduces sparse 3D convolutions to learn the feature extraction. GeoTransformer [22] integrates self- and cross-attention and estimates correspondences by computing the optimal transport, which PEAL [31] improves by modeling unidirectional attention from putative non-overlapping to overlapping superpoints. Recently, BUFFER [1] combines patch-wise and point-wise methods to improve generalizability. All learning-based methods need ground truth correspondences during training, which is usually expensive to obtain on a large scale.

Robust Estimators: Robust estimators optimize for the best transformation given the putative correspondences by the feature matcher. The Random Sample Consensus (RANSAC) [11] algorithm estimates transformations on random correspondence subsets and is widely utilized due to its robustness against a high percentage of outliers. The Iterative Closest Point (ICP) [2] method iteratively adjusts an initial transformation and the correspondence pairs. SC^2-PCR [6] searches for a spatial compatibility consensus among correspondences to better distinguish inliers and outliers. Recent methods like ∇RANSAC [28] allow fully-differentiable pipelines to train feature extractors in an end-to-end fashion.

Unsupervised Registration: Despite the prevalence of supervised feature extraction, some studies have demonstrated the feasibility of unsupervised training. SGP [30] draws an analogy to student-teacher models as illustrated in Fig. 2a. In this context, the student is a trainable feature extractor training on the same pseudo labels for a number of epochs until the teacher, which is RANSAC followed by ICP, generates new labels. For the initial pseudo labels, FPFH features are used. EYOC [17] extends this idea to automotive datasets, but its teacher incorporates a momentum encoder and SC^2-PCR as the learning-free solver. It utilizes consecutive frames from LiDAR sequences and progressively learns to register point clouds that are more distant from each other. Moreover, they spatially filter correspondences close to the ego vehicle. BYOC [10] exploits the fact that images and point clouds are cou-

pled in RGB-D data and trains a point cloud feature extractor with pseudo labels coming from a randomly initialized image feature extractor. UDPReg [19] models point clouds as GMMs and leverages consistency in feature and coordinate space as a self-supervisory signal, while RIENet [24] learns a neighborhood consensus between correspondences.

2.2 Self-Distillation

Several methods in the image domain demonstrate the effective use of the model's own predictions to enhance learning without extensive labeled data. The Mean Teacher model [26] relies on a student-teacher architecture, where the teacher is an EMA of the student's parameters, fostering consistency in predictions for semi-supervised learning. BYOL [14] uses a mean teacher to remove the need for labels completely. Caron *et al*. [5] propose self-distillation with no labels (DINO) by changing the non-contrastive loss in BYOL to a cross-entropy loss on pseudo-class labels. Xie *et al*. [29] introduce noise and data augmentation into the student's training, aiming to replicate the teacher's output, improving generalization.

3 Problem Formulation

Given two partially overlapping 3D point clouds $\mathcal{A} \subset \mathbb{R}^3$ and $\mathcal{B} \subset \mathbb{R}^3$, we want to find the optimal rigid transformation $T^* = \{R^* \in \mathrm{SO}(3), t^* \in \mathbb{R}^3\}$ for a set of ground truth correspondences $\mathcal{C}^*$ with a minimal mean-squared error (MSE):

$$T^* = \underset{T}{\arg\min}\ \mathrm{MSE}(\mathcal{C}^*, T) = \underset{T=\{R,t\}}{\arg\min} \sum_{(a,b)\in\mathcal{C}^*} ||Ra + t - b||_2^2 \tag{1}$$

Here, $a, b \in \mathbb{R}^3$ denote the coordinates of two corresponding points in $\mathcal{A}$ and $\mathcal{B}$. However, ground truth labels are usually not given and have to be estimated first. Depending on the dataset, both point clouds might hold k additional features such as color. For simplicity, we denote the combination of coordinates and features as $\mathcal{A} \subset \mathbb{R}^{3+k}$ and $\mathcal{B} \subset \mathbb{R}^{3+k}$. Feature matchers can learn a function m_θ to estimate the correspondences with $\mathcal{C} = m_\theta(\mathcal{A}, \mathcal{B})$, but usually require the ground truth pose T^* during the training process. We tackle the harder problem where no ground truth is available.

4 Method

We adopt a student-teacher architecture, where the teacher generates pseudo labels on the fly to train the student. This continuously improves the pseudo labels, in contrast to SGP, which generates new labels only after several epochs of training. Our teacher network is updated using an EMA of the student's parameters and therefore, shares the same architecture. Since we are in an unsupervised setting, this process is also termed self-distillation [5]. We incorporate a robust solver into our teacher to improve its estimation. In the final step, we apply a contrastive loss, where the positive pairs are determined by the teacher's estimated correspondences. Fig. 3 visualizes the training process.

4.1 Feature Matching

The feature matcher consists of the commonly used FCGF [8] feature extractor and a subsequent nearest neighbor search. This allows for a more accurate comparison of our method

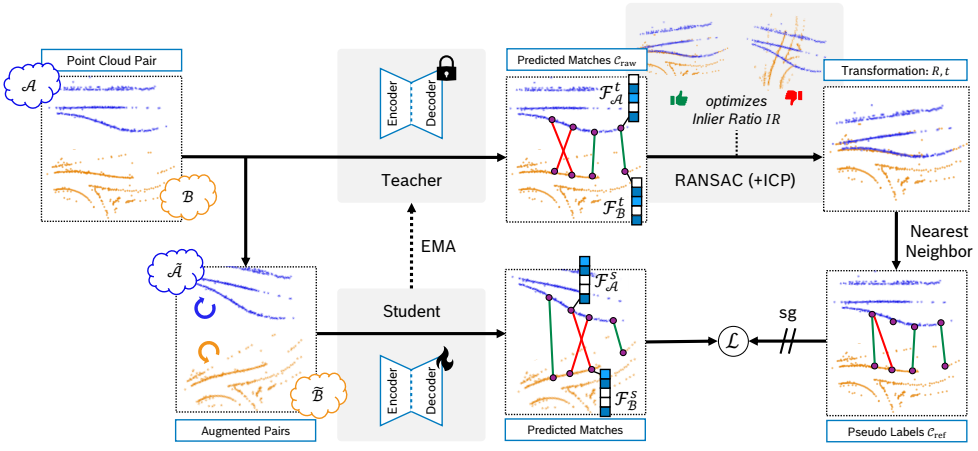


Figure 3: **Self-distillation for registration (DiReg).** Both point clouds are passed to the teacher, while the student receives the augmented views. The networks, FCGF [8] feature extractors, predict geometric features for all points in their pairs and we collect correspondences by searching for the nearest neighbors among the feature vectors of the teacher. Given those correspondences, RANSAC estimates a transformation to align both clouds. Next, we search for nearest neighbors in the coordinate space to get improved correspondences for supervising the student. sg denotes the stop-gradient operator to illustrate that we do not backpropagate through the teacher network. Best viewed on display.

with existing unsupervised approaches [17, 30] by eliminating any potential bias due to differences in the backbone network. In the forward pass, both featured point clouds $\mathcal{P} \in \{\mathcal{A}, \mathcal{B}\}$ are voxelized and passed to the teacher's feature extractor, a 3D variant of a ResUNet with sparse convolutions [8]. It predicts latent features $\mathcal{F}_{\mathcal{P}} \subset \mathbb{R}^{\ell}$ for all points in $\mathcal{P}$. The student receives an augmented view of both point clouds denoted as $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$. We discuss this design choice in more detail in section 4.4. The extracted features of the student and the teacher will be denoted as $\mathcal{F}^s_{\mathcal{P}}$ and $\mathcal{F}^t_{\mathcal{P}}$.

Then we search for all points in $\widetilde{\mathcal{A}}$ the corresponding points in $\widetilde{\mathcal{B}}$, where the distance between the features of our teacher $\mathcal{F}^t_{\mathcal{A}}$ and $\mathcal{F}^t_{\mathcal{B}}$ is minimal, resulting in the initially estimated correspondences $\mathcal{C}_{\text{raw}} = \text{NN}(\mathcal{F}^t_{\mathcal{A}}, \mathcal{F}^t_{\mathcal{B}})$.

4.2 Pseudo label generation

Directly using the correspondences $\mathcal{C}_{\text{raw}}$ from an untrained feature matcher as pseudo labels results in unsatisfactory performance [10]. Therefore, we improve the correspondence prediction by adopting RANSAC optionally followed by the ICP algorithm as proposed in SGP [30]. Both solvers optimize $\hat{T} = \{\hat{R}, \hat{t}\}$ with respect to the unsupervised inlier ratio (IR) defined as:

$$\text{IR}(\hat{T}) = \frac{1}{|\mathcal{C}_{\text{raw}}|} \sum_{(a,b)\in\mathcal{C}_{\text{raw}}} \mathbb{1}\left[||\hat{R}a + \hat{t} - b|| < \tau_1\right], \tag{2}$$

where $\mathbb{1}[\cdot]$ is the indicator function, $||\cdot||$ is the L_2 norm and τ_1 is the acceptable distance threshold. In contrast to the weighted Procrustes solver used in BYOC [10], RANSAC is

more robust to outliers and is therefore widely used. However, since the original RANSAC method is not differentiable, this choice prevents us from directly backpropagating a pose loss as in BYOC and limits us to using a correspondence-based loss with pseudo labels. To generate the new refined correspondences $\mathcal{C}_{\text{ref}}$ from the optimized transformation $\hat{T}$, we perform a nearest neighbor search in the coordinate space with $\mathcal{C}_{\text{ref}} = \text{NN}(\mathcal{A}\hat{R}^\top + \hat{t}, \mathcal{B})$ and keep only those with a distance below a second threshold τ_2.

Additionally, SGP proposed a verifier to remove samples with transformations, where the inlier ratio is below a certain threshold. While this slightly improves the training runtime, we consistently saw a decrease in performance (see Section 5.4) and therefore did not include it.

4.3 Unsupervised Training

Loss Function: We adapt the hardest-contrastive loss $\mathcal{L}$ [8] for training the student network. It combines the loss $\mathcal{L}^p$ for the positive pairs with the hardest negative losses $\mathcal{L}^n_{\mathcal{AB}}$ and $\mathcal{L}^n_{\mathcal{BA}}$ for the first and the second element in each pair.

$$\mathcal{L}(\mathcal{C}_{\text{ref}}, \mathcal{F}^s_{\mathcal{A}}, \mathcal{F}^s_{\mathcal{B}}) = \mathcal{L}^p + 1/2\left(\mathcal{L}^n_{\mathcal{AB}} + \mathcal{L}^n_{\mathcal{BA}}\right) \tag{3}$$

Thereby, we force the features for positive pairs to be close and for negative pairs to be distant. Since no ground truth labels are available, our positive pairs are the correspondences predicted by the teacher $\mathcal{C}_{\text{ref}}$, and negative pairs are determined accordingly. Note that we do not backpropagate gradients through the teacher.

Update-Strategy: We update the teacher's parameters θ^t_i at each step i with an exponential moving average (EMA) of the student's parameters θ^s as proposed by Mean Teacher [26]:

$$\theta^t_i = \alpha\theta^t_{i-1} + (1-\alpha)\theta^s_i, \tag{4}$$

where α follows a cosine schedule [14] from 0.9 to 1. Empirically, we get slightly better results compared to an architecture, where student and teacher share the same parameters, i.e. $\alpha = 0$. Moreover, this update strategy leads to a continuous improvement of our pseudo labels in contrast to SGP, where new labels are only generated after a complete training run.

4.4 Data Augmentation and Bootstrapping

We follow the augmentation for FCGF [8] and randomly rotate both point clouds to force the network to become rotation invariant. However, in our distillation process, it is important to overcome the bootstrap phase, where the randomly initialized teacher can hardly provide any beneficial pseudo labels. Here, the teacher is not yet trained to be invariant against rotations and thus performs weakly with augmented samples. To mitigate this, we draw inspiration from Noisy Student Training [29] and only apply the augmentation on the student's input. Surprisingly, keeping the data's original orientation substantially impacts the bootstrap phase as we demonstrate in our ablations (see Section 5.4). It makes the registration task much easier and accelerates the training. Nevertheless, we maintain the augmentation for the student so that our model learns the same invariance. Note that this approach is different from EYOC [17], which does the augmentation for the student and the teacher network.

SGP utilizes classical features from FPFH [23] for bootstrapping. These features are effective as initialization for RGB-D point clouds but not discriminative enough for radar (see Section 5.4). Moreover, FPFH processes only the 3D coordinates and cannot benefit from additional point cloud features such as color or radar cross-section. After removing the

augmentation for the teacher, we found that FPFH features are counterproductive. So, we omitted them and instead trained them with a teacher network that is randomly initialized.

5 Experiments

We evaluate our method on the widely used 3DMatch benchmark, a collection of RGB-D video data from indoor scenes. We also test on a proprietary automotive radar dataset collected for a mapping task. Specifically, it was designed to register point clouds from different drives, so each road must have been driven multiple times. This task is underrepresented in state-of-the-art point cloud registration, as the standard approach for automotive datasets [9, 12, 25] only registers different frames from the same drive.

5.1 Baselines

To provide a more insightful comparison of our approach with existing unsupervised methods, we adhere to the widely used FCGF [8] feature extractor as our backbone network. However, DiReg should be applicable to any other trainable feature matcher, such as GeoTransformer [22] or BUFFER [1]. We compare DiReg against the supervised setting [8], SGP [30], and EYOC [17], which all train an FCGF model.

For SGP, we partition the total number of epochs into 8 training runs for 3DMatch and 4 runs for the radar data. In each run, the model is further fine-tuned on the new pseudo labels. EYOC utilizes consecutive frames from point cloud sequences for its progressive dataset. Since this is not always available and for a better comparison against all other methods, which cannot access this additional data, we exclude the progressive dataset for our experiments. Additionally, we exclude its spatial filtering, as we have found it inapplicable to RGB-D indoor scenes and to radar scans, which have already undergone post-processing. Note that we use dataset-tuned parameters for its SC^2-PCR estimation.

5.2 Datasets and Implementation

3DMatch: We evaluate on the RGB-D dataset 3DMatch [32]. It contains 62 indoor scenes, where the point cloud pairs overlap by at least 30 %. We split the data according to the FCGF experiments [8] into 48 training, 6 validation, and 8 test scenes. We use the validation set to select the best-performing model in terms of the unsupervised feature match recall (FMR), except for the supervised model validated on the ground truth FMR. It is defined as the percentage of samples with an unsupervised $\mathrm{IR}(\hat{T}, \mathcal{C}_{\mathrm{ref}})$ or ground truth inlier ratio $\mathrm{IR}(T^*, \mathcal{C}_{\mathrm{ref}})$ (see Eq. 2) above 5 % [9]. All models are trained for 200 epochs and have seen only the training scenes. Further, we follow the experiment setup of SGP. Thus for all evaluated methods, we use a voxel size of 5 cm. For inference (and pseudo label generation of SGP and DiReg), we apply RANSAC with 10 k iterations followed by ICP and measure the registration recall (RR) by calling a registration as successful if the relative translation error (RTE) is below 30 cm and the relative rotation error (RRE) is below 15 °.

Automotive Radar: The dataset is designed so that each road is driven several times. It was collected from a few cars, each equipped with a long-range consumer-grade radar sensor. The driving was conducted on highways, city streets, and rural roads. Moving objects were filtered out. We split the data into approximate 250 k training, 30 k validation, and 30 k test pairs by their geographic location. Each point cloud contains several hundred points, along

Method	FMR (%) ↑	RR (%) ↑	IR (%) ↑
Supervised [8]	**93.5**	**92.0**	**24.3**
SGP [30]	91.3	90.8	22.4
EYOC [17]	62.5	76.8	10.6
DiReg (Ours), $\theta^t = \theta^s$	92.3	91.1	22.4
DiReg (Ours)	<u>92.7</u>	<u>91.6</u>	<u>24.1</u>

Table 1: **Registration results on 3DMatch.** $\theta^t = \theta^s$ means that the student and the teacher network share the same parameters, i.e. no momentum teacher is used.

Method	RR (%) ↑	RTE (cm) ↓	RRE (°) ↓
Supervised [8]	**96.6**	**0.355**	<u>0.160</u>
SGP [30]	90.2	0.596	0.219
EYOC [17]	91.7	0.487	0.166
DiReg (Ours), $\theta^t = \theta^s$	<u>96.1</u>	<u>0.390</u>	0.181
DiReg (Ours)	95.8	0.413	**0.156**

Table 2: **Registration results on the radar data.**

with additional features such as radar cross-section. Given the paucity of information in the z-axis, we remove it and learn 2D point cloud registration instead. We apply RANSAC with 5 k iterations without ICP during training and evaluation and use a voxel size of 50 cm. All methods are trained for 16 epochs. Analogous to 3DMatch, we report on the registration recall with 50*cm* and 1 ° as thresholds for RTE and RRE, respectively.

5.3 Results

3DMatch: The evaluation results are reported in Tab. 1. Our approach yields the best performance among all unsupervised methods on feature match recall, registration recall, and inlier ratio and is only surpassed by the supervised FCGF. A more rigorous version of ours, where the student and teacher share the same parameters (i.e. $\theta^t = \theta^s$) ranks third and can be considered as a more memory-friendly training alternative. SGP achieves comparable results with a marginal decline in performance. It is somewhat surprising that the EYOC method is not able to perform satisfactorily. Therefore, it seems plausible to be caused by the data augmentation applied to the student's and the teacher's input. It is important to note that EYOC was developed for distant point clouds. Consequently, its capabilities are not fully realized when applied indoors.

Automotive Radar: Tab. 2 presents the results for the radar dataset. Once again, our approach's performance is close to that of the supervised method, although the impact of our momentum teacher can be disregarded. In contrast to the previous experiment, EYOC performs reasonably well, which fits our observation that correct data augmentation is more crucial in 3DMatch. SGP exhibits suboptimal results. We attribute this to the use of learning-free FPFH features and their verifier as we show in ablations in Section 5.4. These results underline the generalizability of our method across modalities since no modality-specific hand-crafted features are needed.

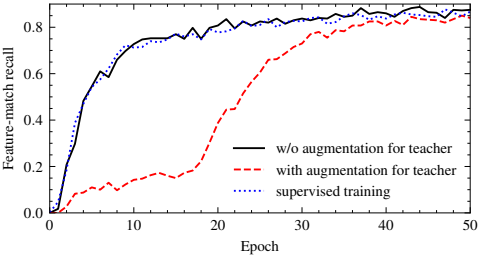


Figure 4: **Training with and without augmentation for teacher.** Student's feature-match recall on the 3DMatch validation set during training. While the training without augmentation follows the supervised trajectory, the training with augmentation needs more epochs to overcome the bootstrap phase.

Method	RR (%) ↑	RTE (cm) ↓	RRE (°) ↓
DiReg with ICP	95.3	**0.289**	**0.093**
DiReg w/o ICP	**95.7**	0.308	**0.093**
DiReg with Verifier	81.6	0.935	0.141
FPFH + RANSAC	3.7	52.405	62.909

Table 3: **Ablation study** of individual components on a subset of the radar data.

5.4 Ablations

Data Augmentation for Teacher: In Fig. 4, we demonstrate that the removal of data augmentation for the teacher's input significantly enhances the robustness of the training process. We saw, that keeping the augmentation at best results in a prolonged training process. At worst, the model even fails to converge. After a successful bootstrap phase, it might be advantageous to bring the teacher augmentation back again and then pass different views to the student and the teacher [5, 14]. We leave this to future work.

Importantance of individual components: Tab. 3 shows an ablation study on a subset containing one-third of the radar scans to evaluate the effect of individual components adopted from SGP. We saw that the additional ICP does not improve the final performance and hence omit it for our radar experiment. It is noteworthy that the pseudo label verifier exhibits suboptimal performance, which could be caused by the fact that it removes a substantial number of challenging samples. Furthermore, it may necessitate additional adjustments when utilized for novel datasets. We also see that a straightforward application of the learning-free FPFH features fails to work on radar and attribute this to the fact that these features are solely based on the 3D coordinates, with the additional radar-specific features not being considered. Both of these observations can be attributed to the suboptimal performance of SGP as seen in Tab. 2.

6 Conclusion

This study presents a novel self-distillation framework for point cloud registration. We show how to bootstrap student-teacher networks unsupervised without the need for initial hand-

crafted features, verifiers, or progressive datasets, while still reaching supervised performance on RGB-D and radar scans. We hope that future applications can benefit from this work by learning the registration task from large-scale crowdsourced data while needing no ground truth poses and fewer adjustments when dealing with a new modality.

Nevertheless, there is still room for improvement. One possibility might be a non-contrastive loss [14] to eliminate the need for negative pairs, as, without ground truth, the pairs are only estimates. Another research direction would be to integrate a differentiable version of RANSAC [3, 28] and then train in an end-to-end fashion.

References

[1] Sheng Ao, Qingyong Hu, Hanyun Wang, Kai Xu, and Yulan Guo. Buffer: Balancing accuracy, efficiency, and generalizability in point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1255–1264, 2023.

[2] Paul J Besl and Neil D McKay. Method for registration of 3-D shapes. In *Sensor fusion IV: Control Paradigms and Data Structures*, volume 1611, pages 586–606. Spie, 1992.

[3] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6684–6692, 2017.

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[6] Zhi Chen, Kun Sun, Fan Yang, and Wenbing Tao. SC2-PCR: A second order spatial compatibility for efficient and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13221–13231, 2022.

[7] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2015.

[8] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8958–8966, 2019.

[9] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppfnet: Global context aware local features for robust 3d point matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 195–205, 2018.

[10] Mohamed El Banani and Justin Johnson. Bootstrap your own correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6433–6442, 2021.

[11] 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.

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[13] 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.

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[15] Yoichiro Hisadome and Yusuke Matsui. Cascading feature extraction for fast point cloud registration. In *British Machine Vision Conference (BMVC)*, 2021.

[16] Dongrui Liu, Chuanchaun Chen, Changqing Xu, Robert C Qiu, and Lei Chu. Self-supervised point cloud registration with deep versatile descriptors for intelligent driving. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[17] Quan Liu, Hongzi Zhu, Zhenxi Wang, Yunsong Zhou, Shan Chang, and Minyi Guo. Extend your own correspondences: Unsupervised distant point cloud registration by progressive distance extension. *arXiv preprint arXiv:2403.03532*, 2024.

[18] Bilawal Mahmood and SangUk Han. 3d registration of indoor point clouds for augmented reality. In *ASCE International Conference on Computing in Civil Engineering 2019*, pages 1–8. American Society of Civil Engineers Reston, VA, 2019.

[19] Guofeng Mei, Hao Tang, Xiaoshui Huang, Weijie Wang, Juan Liu, Jian Zhang, Luc Van Gool, and Qiang Wu. Unsupervised deep probabilistic approach for partial point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13611–13620, 2023.

[20] Alexander Osipov, Mikhail Ostanin, and Alexandr Klimchik. Comparison of point cloud registration algorithms for mixed-reality cross-device global localization. *Information*, 14(3), 2023. ISSN 2078-2489. doi: 10.3390/info14030149. URL https://www.mdpi.com/2078-2489/14/3/149.

[21] Tong Qin, Haihui Huang, Ziqiang Wang, Tongqing Chen, and Wenchao Ding. Traffic flow-based crowdsourced mapping in complex urban scenario. *IEEE Robotics and Automation Letters*, 8(8):5077–5083, 2023. doi: 10.1109/LRA.2023.3291507.

[22] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11143–11152, 2022.

[23] 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. IEEE, 2009.

[24] Yaqi Shen, Le Hui, Haobo Jiang, Jin Xie, and Jian Yang. Reliable inlier evaluation for unsupervised point cloud registration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2198–2206, 2022.

[25] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020.

[26] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017.

[27] Rendong Wang, Youchun Xu, Miguel Angel Sotelo, Yulin Ma, Thompson Sarkodie-Gyan, Zhixiong Li, and Weihua Li. A robust registration method for autonomous driving pose estimation in urban dynamic environment using lidar. *Electronics*, 8 (1), 2019. ISSN 2079-9292. doi: 10.3390/electronics8010043. URL https://www.mdpi.com/2079-9292/8/1/43.

[28] Tong Wei, Yash Patel, Alexander Shekhovtsov, Jiri Matas, and Daniel Barath. Generalized differentiable ransac. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17649–17660, 2023.

[29] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

[30] Heng Yang, Wei Dong, Luca Carlone, and Vladlen Koltun. Self-supervised geometric perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14350–14361, 2021.

[31] 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.

[32] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 1802–1811, 2017.

[33] Ji Zhang and Sanjiv Singh. Loam: Lidar odometry and mapping in real-time. In *Robotics: Science and Systems*, volume 2, pages 1–9. Berkeley, CA, 2014.