Robotics: Science and Systems 2025
Los Angeles, CA, USA, June 21–25, 2025

LiDAR Registration with Visual Foundation Models

Niclas Vödisch[1,2], Giovanni Cioffi[2], Marco Cannici[2], Wolfram Burgard[3], and Davide Scaramuzza[2]
[1]University of Freiburg [2]University of Zurich [3]University of Technology Nuremberg

***Abstract*—LiDAR registration is a fundamental task in robotic mapping and localization. A critical component of aligning two point clouds is identifying robust point correspondences using point descriptors. This step becomes particularly challenging in scenarios involving domain shifts, seasonal changes, and variations in point cloud structures. These factors substantially impact both handcrafted and learning-based approaches. In this paper, we address these problems by proposing to use DINOv2 features, obtained from surround-view images, as point descriptors. We demonstrate that coupling these descriptors with traditional registration algorithms, such as RANSAC or ICP, facilitates robust 6DoF alignment of LiDAR scans with 3D maps, even when the map was recorded more than a year before. Although conceptually straightforward, our method substantially outperforms more complex baseline techniques. In contrast to previous learning-based point descriptors, our method does not require domain-specific retraining and is agnostic to the point cloud structure, effectively handling both sparse LiDAR scans and dense 3D maps. We show that leveraging the additional camera data enables our method to outperform the best baseline by $+24.8$ and $+17.3$ registration recall on the NCLT and Oxford Radar RobotCar datasets. We publicly release the registration benchmark and the code of our work on https://vfm-registration.cs.uni-freiburg.de.**

I. Introduction

Aligning two point clouds to compute their relative 3D transformation is a critical task in numerous robotic applications, including LiDAR odometry [30], loop closure registration [2], and map-based localization [19]. In this work, we specifically discuss map-based localization, which not only generalizes the other aforementioned tasks but is also critical for improving the efficiency and autonomy of mobile robots in environments where pre-existing map data is available.

Although place recognition [20, 21] or GNSS readings can provide an approximate initial estimate, their accuracy is generally insufficient for obtaining precise 3D poses relative to the map. In contrast, global point cloud registration [16, 33] enables accurate 3D localization but necessitates the identification of reliable point correspondences between the point clouds. These correspondences are typically established by iterating over all points in the source point cloud to identify the most similar counterparts in the target frame. Similarity is assessed using point descriptors, which are abstract feature representations of a point, e.g., encoding the geometry of its local environment. In scan-to-map registration, point descriptors must be as unique as possible since the number of potential combinations grows with $\mathcal{O}(m \cdot n)$, referring to the number of points in the scan and the map. An additional challenge arises from temporal changes in the environment, such as seasonal variations or ongoing construction, necessitating point descriptors that are robust to such changes for long-term applicability [11].

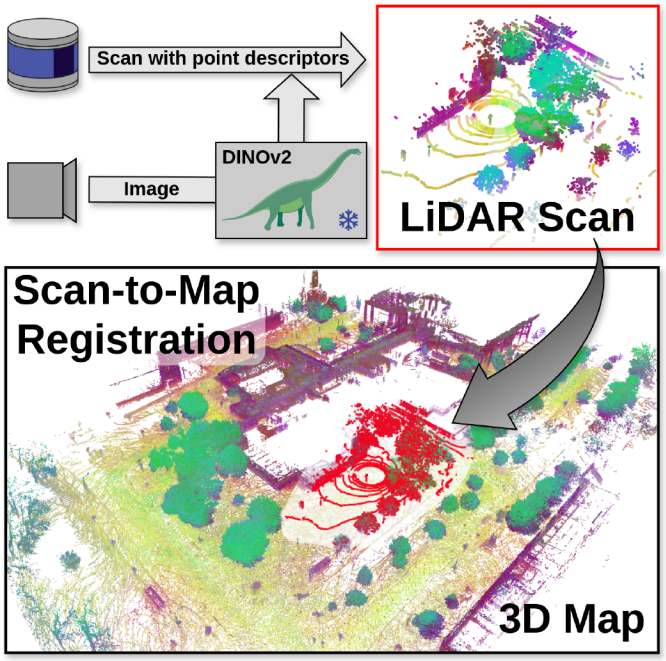

Fig. 1. Initialization-free registration of a LiDAR scan to a large-scale 3D map requires highly expressive point descriptors. We demonstrate that DINOv2 [25] features from surround-view images allow finding robust point correspondences, even with map data recorded more than a year before. In the map, the registered LiDAR scan is shown in red. The colors of the LiDAR scan (*top right*) and the map (*bottom*) are obtained using principal component analysis performed on the high-dimensional DINOv2 features.

Given the significance of this task, numerous point descriptors have been proposed, encompassing both traditional handcrafted [29] and learning-based [12, 27, 32] designs, primarily relying on 3D geometric features. While learning-based descriptors tend to exhibit greater expressiveness than handcrafted methods, they often fail to generalize effectively to out-of-training domains and different point cloud representations, such as between RGB-D data and LiDAR scans [28].

In this work, we address the task of long-term scan-to-map registration by leveraging the advances made by recent visual foundation models. Our main contribution is to demonstrate that using DINOv2 [25] features, obtained from surround-view images, as point descriptors allows finding highly robust point correspondences. We argue that using the additional vision modality does not pose a large burden as this combination is a common sensor setup in mobile robotics [6, 10, 11] and the camera is relatively inexpensive compared to the LiDAR.

The key idea behind our approach is to leverage the superior generalization capabilities of recent visual foundation models, like DINOv2, compared to networks operating in the 3D space. Furthermore, our approach is hence agnostic to the shape of a

point cloud, enabling correspondence search between sparse LiDAR scans and dense 3D voxel-maps. Using DINOv2-based point descriptors effectively allows for implicit semantic matching between points and, importantly, does not require re-training an in-domain descriptor network.

We make three claims: First, we demonstrate that coupling these descriptors with traditional registration algorithms, such as RANSAC [16] or ICP [30], facilitates robust 3D localization in a map that was recorded over a year before [11]. Second, although conceptually simple, our method substantially outperforms more complex baseline techniques. Third, our approach is robust to temporal changes in the environment that have occurred since the map was created.

We validate these claims through extensive experiments, showing that our approach outperforms the best baseline by $+24.8$ and $+17.3$ registration recall on the NCLT [11] and Oxford Radar RobotCar [6] datasets. To facilitate reproducibility and future research on long-term map registration, upon acceptance we will release our code along with instructions to re-create the evaluation scenes from our experiments. To the best of our knowledge, this work presents the first approach to combining visual foundation models with traditional LiDAR registration techniques.

II. Related Work

Point cloud registration has been extensively studied by the research community across a diverse range of applications. Previous works have addressed alignment of relatively small 3D objects [9], mid-size indoor scenes [27, 32], LiDAR odometry [14, 30], and scan registration to 3D maps [19]. Particularly the latter introduces further challenges, such as geometric and semantic discrepancies between the source and the target point clouds, arising from temporal changes in the environment since the creation of the reference map. While the majority of studies focus on object alignment or LiDAR odometry, only a few works [19, 21] explicitly target long-term scenarios. Typically, the point cloud registration problem is addressed in two stages: first, identifying point-to-point correspondences using point descriptors and, second, determining the six degrees of freedom (6DoF) transformation required to align the two point clouds. In the following paragraphs, we provide an overview of each of these steps.

Point Descriptors: Point descriptors refer to an abstract representation of a 3D point that can be used to search for point-to-point correspondences between two point clouds. In contrast to a naive nearest-neighbor search in the Euclidean space, e.g., performed by the iterative closest point (ICP) algorithm [7], point descriptors allow finding global correspondences. A classical, yet still commonly used [4, 31] descriptor is the FPFH [29] descriptor that captures the geometry around a point by computing local feature histograms based on the angles to its neighboring points. FPFH is designed to provide both translational and rotational invariance. In more recent years, many learning-based approaches have been proposed that employ deep neural networks for extracting point descriptors. These methods can generally be categorized into patch-based networks and fully convolutional networks. 3DMatch [32] pioneered the first category by learning local geometric patterns from volumetric patches around a point. While 3DMatch computes truncated distance function values from the patch, 3DSmoothNet [18] uses a smoothed density value representation to achieve rotation invariance. Both DIP [27] and GeDi [28] propose to follow a Siamese approach to train two neural networks with shared parameters and a contrastive loss on the patches. Unlike patch-based approaches, fully convolutional networks employ such a contrastive loss directly on the point level as first proposed by FCGF [12], which is commonly used by many registration algorithms [4]. While early convolutional descriptors were either computed for all points of a point cloud or a randomly sampled subset, later works such as D3Feat [3] include keypoint detection schemes. A particular challenge of operating on individual points instead of patches is to achieve density invariance. Therefore, GCL [23] focuses on low-overlap scenarios, e.g., required for early loop closure registration in LiDAR SLAM. Although learning-based descriptors have shown impressive performance, a substantial drawback is their poor generalization between different training and testing domains, e.g., RGB-D data versus LiDAR scans or aligning individual objects versus large-scale outdoor scenes.

The majority of point descriptors focus on encoding only geometric information neglecting semantic hints. Especially in the context of autonomous driving, a few works have proposed to include information from semantic segmentation. For instance, SAGE-ICP [14] extends the point correspondence search of ICP with a hard rejection scheme if the candidate points belong to different semantic classes. The transformer-based PADLoC [2] exploits panoptic information during the training phase to avoid wrong matches. Nonetheless, a major barrier to including semantic information is the lack of 3D segmentation networks that generalize well across domains requiring cost-intense retraining. In this work, we exploit the recent advances in the vision domain by proposing to extract point descriptors using a visual foundation model [25]. In contrast to the hard matching of discrete semantic classes, our method enables searching more soft correspondences while considering the scene semantics.

Point Cloud Registration: Algorithms for point cloud registration can be categorized into local and global registration schemes. Whereas local methods require an accurate initial guess, global registration often assumes given point correspondences based on the aforementioned point descriptors. Often, both types are combined in a coarse-to-fine manner to achieve global registration with the high performance of local approaches such as ICP [7] or NDT [8]. To obtain a sufficient coarse registration, a main requirement for global registration schemes is outlier rejection. The most popular traditional method for this task is still RANSAC [16], including its more recent variants [5]. However, the major drawbacks of RANSAC are slow convergence and low accuracy in the presence of large outlier rates, which are commonly faced in

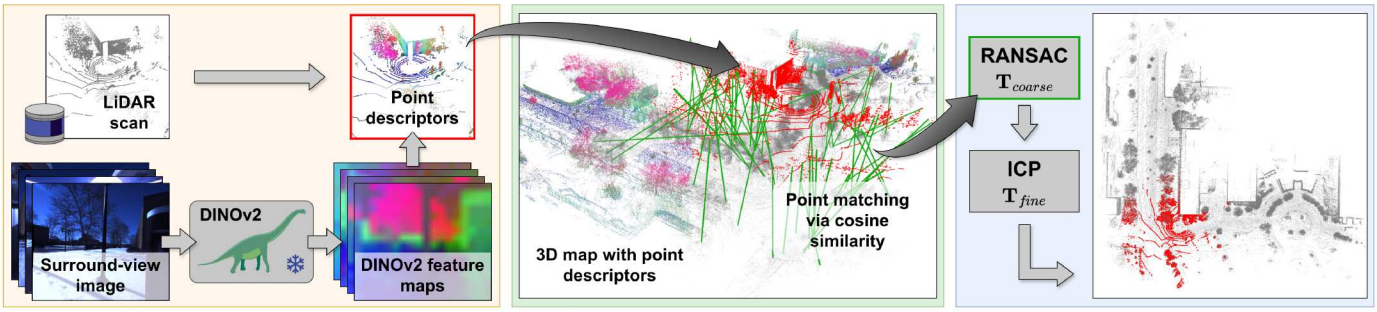

Fig. 2. Overview of our proposed approach for 6DoF point cloud registration. First, we extract DINOv2 [25] features from surround-view image data. These features are then attached to the point cloud as point descriptors via point-to-pixel projection. Second, we perform a point-wise similarity search using cosine similarity between the descriptors of the LiDAR scan and the descriptors of the voxelized 3D map. Finally, we use a traditional coarse-to-fine registration scheme with RANSAC [16] and point-to-point ICP [30] for obtaining a highly accurate pose estimate within the provided map frame.

point cloud registration. Fast global registration (FGR) [33] aims to overcome these problems by optimizing a robust objective function that is defined densely over the surfaces. TEASER [31] proposes a certifiable algorithm that decouples scale, rotation, and translation estimation. Similar to other fields, recent outlier rejection schemes attempt to improve their performance via deep learning. Both deep global registration (DGR) [13] and 3DRegNet [26] formulate outlier rejection as a point-based classification problem. PointDSC [4] extends this idea by including the spatial consistency between inlier correspondences when applying rigid Euclidean transformations. In this work, we demonstrate that coupling our proposed point descriptors with traditional registration algorithms, such as RANSAC or ICP, enables robust point cloud registration.

III. Technical Approach

In this section, we first formally define the problem addressed in this work. Then, we explain how to extract the point descriptors based on a visual foundation model. Finally, we elaborate on how we employ these descriptors for robust scan-to-map registration. We illustrate the separate steps in Fig. 2.

A. Problem Definition

In this work, we consider the following scenario: A voxelized 3D map $M \in \mathbb{R}^{n\times 3}$ is provided, where n denotes the number of 3D points stored in the map. At test-time, we receive a LiDAR scan $S \in \mathbb{R}^{m\times 3}$ composed of m 3D points. Furthermore, for both M and S, corresponding surround-view RGB camera data is available. The goal is to find the six degrees of freedom (6DoF) transform $\mathbf{T} \in \mathrm{SE}(3)$ that correctly registers the LiDAR scan to the map. We further assume a rough initial position $\hat{P} \in \mathbb{R}^3$ is given within approximately $100\,\mathrm{m}$ around the true position, reducing the size of the relevant part of the map to $k << n$ while preserving $k >> m$. Such an initial position could be obtained via place recognition [20, 22] or GNSS readings. Importantly, in long-term scan-to-map registration, the LiDAR scan can be recorded a considerable amount of time after the map was created, i.e., there might be a semantic and geometric discrepancy between the 3D map representation and the current state of the environment.

B. Point Descriptor Extraction

In this paragraph, we describe step 1) of Fig. 2. While we extract the point descriptors of the LiDAR scan S at test-time, we pre-compute the descriptors of the map M in an offline fashion. Commonly, 3D point-based mapping approaches rely on the concept of keyframes to frequently identify LiDAR scans that are eventually accumulated into a single voxelized point cloud, i.e., a map is formally composed of individual LiDAR scans $M = \bigcup_j M_j$. In the following, we hence use the general notation of a point cloud $C \in \mathbb{R}^{l\times 3}$ to refer to either S or M_j. Each point cloud can be associated with a surround-view RGB image taken at the same time as C. We denote this image as $I \in \mathbb{N}^{h\times w\times 3}$, where h and w represent its height and width, respectively. First, we feed I through a frozen DINOv2 [25] model to generate a dense 2D feature map $F \in \mathbb{R}^{h\times w\times d}$, e.g., $d = 384$ for the model type ViT-S/14. The core idea of using DINOv2 is to capture the semantics of the scene without an explicit assignment to discrete semantic classes [14] while leveraging its generalization capabilities across cameras, weather, and illumination conditions [20]. Second, we employ point-to-pixel projection via known extrinsic calibration parameters to convert each point $p \in C$ into pixel coordinates of I. Finally, we use the DINOv2 feature of the respective RGB pixel as the descriptor $\mathrm{desc}(\cdot)$ of point p. Formally,

$$F = \texttt{DINOv2}(I)\,, \tag{1}$$

$$\mathrm{desc}(p) = F\left[\Pi(p)\right]\,, \tag{2}$$

where $\Pi(\cdot) : \mathbb{R}^3 \rightarrow \mathbb{N}^2$ is the point-to-pixel projection function and $[\cdot] : \mathbb{N}^2 \rightarrow \mathbb{R}^d$ denotes the operator to access the DINOv2 feature of a given pixel. Consequently, we apply this step to all points in C to obtain C^D:

$$C^D = \{\mathrm{desc}(p)\,|\,\forall p \in C\} \tag{3}$$

We hence retrieve the descriptors $M^D \in \mathbb{R}^{k\times d}$ corresponding to the map M and, at test-time, the descriptors $S^D \in \mathbb{R}^{m\times d}$ for the current scan S.

C. Scan-to-Map Registration

As defined in Sec. III-A, the goal of scan-to-map registration is to find the 6DoF transform that correctly represents the robot pose with respect to the coordinate system of the map. In this paragraph, we describe the corresponding steps 2) and 3) of Fig. 2. First, we substantially downsample the LiDAR scan S with point descriptors S^D resulting in $\tilde{S} \in \mathbb{R}^{v \times 3}$ and $\tilde{S}^D \in \mathbb{R}^{v \times d}$ with $v \approx m/80$ to reduce the complexity of the subsequent steps. Second, we search for point correspondences between the scan and the map using an efficient similarity search [15]. For every point $p_s \in \tilde{S}$, we search for the point $p_m \in M$ that achieves the highest cosine similarity between the descriptors of both points.

$$p_m = \underset{p \in M}{\arg\max}\, \mathrm{sim}_{\cos}\left(\mathrm{desc}(p_s), \mathrm{desc}(p)\right) \tag{4}$$

$$= \underset{p \in M}{\arg\max}\, \frac{\mathrm{desc}(p_s) \cdot \mathrm{desc}(p)}{||\mathrm{desc}(p_s)||\,||\mathrm{desc}(p)||} \tag{5}$$

Finally, if the cosine similarity is greater than a threshold $\theta_{\cos} = 0.8$, we consider the pair (p_s, p_m) a valid point correspondence.

To achieve global registration within the map, we run 3-point RANSAC [16] on the set of all valid point correspondences, resulting in a coarse initial transform $\mathbf{T}_{coarse}$. For further refinement and accurate 6DoF registration, we employ classical point-to-point ICP [30] based on the 3D points of the original LiDAR scan S. That is, the DINOv2-based point descriptors are not used in this step. With ICP, we obtain the final 6DoF transform $\mathbf{T}_{fine}$ that aligns the scan S with the map M.

IV. Experiments

The main focus of this work is to enable LiDAR scan-to-map registration that is robust to the challenges arising in long-term scenarios. In our experiments, we demonstrate the effectiveness and capabilities of our approach to support our key claims: First, DINOv2 [25] features can serve as point descriptors that can be integrated into traditional registration algorithms to facilitate robust 3D registration in a map that was recorded more than a year before. Second, such a relatively straightforward approach outperforms more complex baseline techniques. Third, these descriptors are robust to temporal changes in the environment that have occurred since the map was created. We begin this section by describing the details of our experimental setup and defining the evaluation metrics used. Afterward, we provide results that support our claims and showcase the performance of our approach.

A. Experimental Setup

To verify our claims, we generate several scenes that present the problem formally defined in Sec. III-A. In particular, we identify the NCLT [11] and the Oxford Radar RobotCar [6] as two of the few long-term datasets containing both LiDAR and surround-view RGB images. Note that neither is part of the LVD-142M dataset [25] used to train DINOv2. The NCLT dataset comprises a total of 27 sessions recorded on a university campus spread out over 15 months, i.e., covering various seasons, illuminations, and environmental conditions. The Oxford Radar RobotCar dataset comprises recordings from 7 different days spread out over 1.5 weeks. It captures vehicle-centric urban data with weather conditions ranging from sunny to varying degrees of overcast. Both datasets contain global pose data that is consistent across recordings. For each dataset, we select the first of the provided recordings for mapping, while other recordings are used for registration. As shown in Tab. I, we carefully select these recordings; first, to maximize spread over the dataset to increase the probability of semantic and geometric changes and, second, to cover all available seasonal and weather conditions.

TABLE I
Recordings for Scene Generation

Recording date	Map.	Reg.	Time	Sky	Foliage	Snow
NCLT						
2012-01-08	✓		Midday	Partly cloudy	–	–
2012-02-12		✓	Midday	Sunny	–	✓
2012-03-17		✓	Morning	Sunny	–	–
2012-05-26		✓	Evening	Sunny	✓	–
2012-10-28		✓	Midday	Cloudy	–	–
2013-04-05		✓	Afternoon	Sunny	–	✓
Oxford Radar RobotCar						
2019-01-10	✓		11:46	Cloudy	–	–
2019-01-15		✓	13:06	Sunny	–	–
2019-01-17		✓	14:03	Partly cloudy	–	–
2019-01-18		✓	15:20	Cloudy	–	–

Overview of the recordings from the NCLT [11] and the Oxford Radar RobotCar [6] datasets used for mapping (*Map.*) and scan registration (*Reg.*).

To construct a scene, we sample a position ρ_i from the mapping route. If all registration recordings contain a point cloud associated with a pose in the vicinity of ρ_i, we use these point clouds to create $\{\mathcal{S}_1, \ldots, \mathcal{S}_r\}_i$ and $\{\mathcal{S}_1^D, \ldots, \mathcal{S}_r^D\}_i$, i.e., a set of LiDAR scans with corresponding DINOv2-based [25] point descriptors. As listed in Tab. I, we use $r = 5$ for NCLT and $r = 3$ for RobotCar. For mapping, we select the point clouds along the route within a $150\,\mathrm{m}$ radius around ρ_i and with a distance of $2\,\mathrm{m}$ between scans, simulating keyframes in LiDAR SLAM, and extract the point descriptors. That is, the maximum size of the map is $300\,\mathrm{m} \times 300\,\mathrm{m}$. Since the accuracy of the global poses provided in the datasets is insufficient for point cloud accumulation, we refine them with KISS-ICP [30]. Finally, we downsample the accumulated resulting point cloud to a voxel size of $0.25\,\mathrm{m}$, representing the 3D map M_i with point descriptors M_i^D. Note that we do not remove potentially dynamic objects, e.g., cars, from the map. To measure the registration error, we generate ground truth transformations $\{\mathbf{T}_1, \ldots, \mathbf{T}_r\}$ by running point-to-point ICP initialized with the pose from the dataset and manually verifying the correct registration. For both datasets, we construct 25 scenes, i.e., $i \in [1, 25]$, resulting in a sample size of 125 and 75 for NCLT and Oxford Radar RobotCar, respectively. For examples of the scenes, we refer to the qualitative results in Figs. 6 and 7. To facilitate the utilization of these scenes as a benchmark in future research, we provide a recreation script along with comprehensive instructions in our code release.

TABLE II
SCAN-TO-MAP REGISTRATION

Method		**NCLT**				**Oxford Radar RobotCar**			
Registration	Descriptor	RTE [m]	RRE [°]	RR [%]	ICP-RR [%]	RTE [m]	RRE [°]	RR [%]	ICP-RR [%]
RANSAC	FPFH [29]	47.51±46.54	59.95±62.21	0.00	37.60	32.56±40.18	36.87±57.25	0.00	40.00
TEASER++ [31]	FPFH [29]	60.15±34.94	121.67±48.95	0.00	0.00	69.51±41.98	129.01±49.91	0.00	0.00
PointDSC [4]	FPFH [29]	103.16±48.70	129.75±47.90	0.00	0.00	65.28±45.19	128.01±52.22	1.33	2.67
RANSAC	DIP [27]	42.94±45.26	49.86±56.08	0.00	40.00	24.52±32.96	45.08±63.06	0.00	40.00
RANSAC	GeDi [28]	48.93±45.12	66.75±65.59	0.00	32.00	23.50±31.15	41.26±62.85	0.00	44.00
RANSAC	FCGF [12]	22.74±33.53	31.20±48.44	0.80	64.00	20.34±32.29	37.92±62.81	1.33	49.33
PointDSC [4]	FCGF [12]	69.50±48.40	96.86±51.55	0.00	12.00	73.58±45.43	93.51±64.32	0.00	0.00
RANSAC	SpinNet [1]	33.70±42.47	41.00±55.14	0.00	54.40	23.39±35.86	30.57±56.48	0.00	56.00
RANSAC	GCL [23]	15.35±26.84	23.67±45.83	13.60	75.20	11.81±25.09	21.84±48.07	30.67	77.33
Our method:									
TEASER++ [31]	DINOv2 [25]	5.29±10.02	27.98±57.09	18.40	83.20	13.45±34.93	22.18±53.69	49.33	81.33
RANSAC	DINOv2 [25]	0.40±0.31	1.01±0.84	77.60	**100.00**	3.36±13.65	5.20±26.25	82.67	**94.67**
RANSAC + ICP	DINOv2 [25]	**0.01±0.02**	**0.03±0.06**	**100.00**	**100.00**	**3.12±13.63**	**4.52±26.21**	**94.67**	**94.67**

We report the mean and standard deviation of the relative translation error (RTE) and the relative rotation error (RRE) as defined in Sec. IV-B. The registration recall (RR) denotes the success rate, where success is defined as RTE $< 0.6\,\mathrm{m}$ and RRE $< 1.5°$. The recall after refinement with ICP is listed as ICP-RR. DIP and GeDi are trained on the 3DMatch dataset [32] (RGB-D data). FCGF, SpinNet, GCL, and PointDSC are trained on the KITTI dataset [17] (LiDAR scans).

B. Evaluation Metrics

Similar to prior work [1, 3, 4, 12], we use the following evaluation metrics: First, the relative translation error (RTE) measured in meters. Second, the relative rotation error (RRE) measured in degrees. Formally, the RTE and RRE are defined as:

$$\mathrm{RTE} = ||\hat{\mathbf{t}} - \mathbf{t}||_2 \tag{6}$$

$$\mathrm{RRE} = \arccos\left(\frac{\mathrm{tr}(\hat{\mathbf{R}}^\intercal \mathbf{R}) - 1}{2}\right), \tag{7}$$

where the transform $\mathbf{T} \in \mathrm{SE}(3)$ is decomposed into a translation $\mathbf{t}$ and rotation $\mathbf{R}$. The hat $\hat{(\cdot)}$ denotes the estimated transform. The operators $\mathrm{tr}(\cdot)$ and $(\cdot)^\intercal$ are the trace and transpose of a matrix. Third, we report the registration recall (RR) denoting the percentage of successful registrations, i.e., both the RTE and RRE are below a given threshold. While previous works have used different thresholds [1, 3, 4, 23], we adopt the criterion of Liu *et al.* [23] as it meets the expected error range of many baseline techniques in the more simple scan-to-scan registration tasks (see Sec. IV-C). That is, a registration is considered a success if RTE $< 0.6\,\mathrm{m}$ and RRE $< 1.5°$. Finally, we investigate whether the accuracy of the descriptor-based global registration is sufficient to initialize ICP. We measure its performance by recomputing the registration recall after ICP-based pose refinement (ICP-RR).

C. Scan-to-Map Registration

We compare our proposed approach to a variety of baselines that can be categorized as follows: (1) the popular handcrafted descriptor FPFH [29] coupled with three outlier rejection schemes, namely, RANSAC [16], TEASER++ [31], and PointDSC [4], which is a learning-based approach trained on the KITTI dataset [17]; (2) the learning-based descriptors DIP [27] and GeDi [28], which are trained on RGB-D data of the 3DMatch dataset [32] but claimed generalization to LiDAR scans [27, 28]; (3) the learning-based descriptors FCGF [12], SpinNet [1], and GCL [23], which are trained on the KITTI dataset [17]. For categories (2) and (3), we predominantly rely on the RANSAC algorithm for point cloud registration and follow prior work [3, 32] using 50,000 iterations without early stopping. For the baselines, we use the top 5,000 point correspondences. For SpinNet [1], we compute the descriptors only for 7,500 randomly sampled points due to GPU memory constraints (16 GB).

TABLE III
SCAN-TO-SCAN REGISTRATION

Method		**KITTI**		**NCLT**	
Registration	Descriptor	RTE [m]	RRE [°]	RTE [m]	RRE [°]
Handcrafted descriptor:					
ICP	*n/a*	11.00	5.86	10.91	6.32
RANSAC	FPFH [29]	0.87	1.13	1.35	3.55
TEASER++ [31]	FPFH [29]	0.04	0.10	0.54	3.24
PointDSC [4]	FPFH [29]	0.06	0.15	2.91	9.70
Descriptors trained on 3DMatch [32]:					
RANSAC	DIP [27]	0.16	0.22	0.80	2.25
RANSAC	GeDi [28]	0.50	0.73	1.71	4.25
Descriptors trained on KITTI [17]:					
RANSAC	FCGF [12]	0.10	0.15	0.45	1.36
PointDSC [4]	FCGF [12]	0.07	0.14	0.51	1.64
RANSAC	SpinNet [1]	0.09	0.15	0.46	1.28
RANSAC	GCL [23]	0.09	0.14	0.46	1.26
Our method:					
RANSAC	DINOv2 [25]	–	–	0.56	1.44

We underline the effect of a domain change on various point descriptors used as baselines in our experiments. While the descriptors trained on the KITTI dataset [17] perform well within the same domain, they suffer from degradation when tested on the NCLT dataset [11]. Note that PointDSC [4] is also trained on KITTI. Due to the lack of surround-view images, we do not employ our method on KITTI. In contrast to the primary use case of this work, previous studies mostly considered scan-to-scan registration.

A core advantage of employing DINOv2 [25] is exploiting its strong generalization capability across various semantic domains and decoupling the point descriptors from the density of a point cloud, e.g., RGB-D data versus LiDAR scans. Many learning-based descriptors suffer from such domain changes requiring in-domain training data and hindering their general applicability. To underline this effect and to establish the groundwork for the main experiment, we first report the performance of the baselines for the more simple scan-to-scan

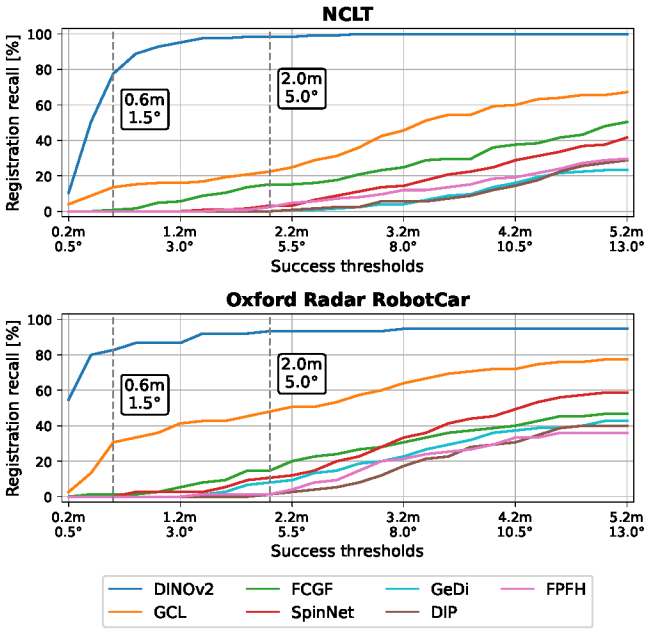

Fig. 3. We visualize the registration recall (RR) for a range of success thresholds obtained by linear inter/extrapolation of the thresholds used by GCL [23] (*left dashed line*) and SpinNet [1] (*right dashed line*). To perform scan-to-map registration, we couple RANSAC with the specified point descriptors.

registration task. This task is mainly considered by previous studies in the context of LiDAR odometry. Here, the two point clouds are highly similar in their geometry and a strong initial guess is available. We simulate the initial guess by perturbing the ground truth transform with noise sampled as:

$$\begin{aligned} \mathbf{t}_x, \mathbf{t}_y &\sim \mathcal{N}(0, 10)\,, \mathbf{t}_y \sim \mathcal{N}(0, 1)\,, \\ \mathbf{R}_x, \mathbf{R}_y &\sim \mathcal{N}(0, 2)\,, \mathbf{R}_z \sim \mathcal{N}(0, 10)\,, \end{aligned} \tag{8}$$

with $\mathbf{t}$ and $\mathbf{R}$ referring to the translational and rotational components measured in meters and degrees, respectively. In Tab. III, we report the average RTE and RRE on samples from the KITTI and NCLT datasets. For KITTI, we sample 125 pairs of two consecutive scans from sequence 08, which is not in the descriptors' training set [1, 12]. For NCLT, we use the scan of the map that is closest to the incoming LiDAR scan. Note that we do not employ our method on KITTI due to the lack of surround-view images. The key insights from Tab. III are as follows: (1) On KITTI, the error of the in-domain trained descriptors is substantially smaller than of the ones trained on 3DMatch; (2) We observe poor performance when the training and testing domains differ, i.e., from 3DMatch to KITTI or NCLT, and from KITTI to NCLT; (3) Most descriptors achieve a decent accuracy on NCLT for the scan-to-scan registration task.

In the main experiment, we support our claim that our proposed DINOv2-based point descriptors can be coupled with traditional registration schemes while outperforming previous baselines. We now consider scan-to-map registration using the extracted scenes as described in Sec. IV-A. We report results for the metrics defined in Sec. IV-B for both the NCLT dataset and the Oxford Radar RobotCar dataset in Tab. II. The most important observation is that our proposed method is the only one that achieves consistently low registration errors. For the ICP-based refinement, our method substantially outperforms the best baseline (GCL [23]) by 24.8 (NCLT) and 17.3 (RobotCar) percentage points, showing $100\,\%$ recall on NCLT. We hypothesize that the gap to $100\,\%$ on the Oxford Radar RobotCar is mainly caused by wrong point-to-pixel projections, e.g., due to erroneous extrinsic calibration, jerky ego-motion, and differences in the sensors' viewpoints. An indicator for this hypothesis is the observation that some points belonging to buildings are assigned tree-like descriptors if there is a tree in front of the wall. For completeness, we also report results when replacing RANSAC with TEASER++ [31], achieving higher registration recalls than all baseline methods. A further key insight is the large standard deviation of the errors of the baseline methods, whereas the results of our method rarely fluctuate. Finally, we note that the baselines yield almost no global registration meeting the success thresholds, resulting in $0.00\,\%$ registration recall. To further investigate this observation and to incorporate more relaxed thresholds as used in other studies, we recompute the registration recall (RR) for additional thresholds. In particular, we use linear inter/extrapolation of the thresholds based on GCL [23] ($0.6\,\mathrm{m}$, $1.5°$) and SpinNet [1] ($2.0\,\mathrm{m}$, $5.0°$). We visualize the registration recalls in Fig. 3 for all descriptors coupled with RANSAC. The recall of our DINOv2-based descriptors yields a high recall even for strict thresholds, eventually converging towards the recall after ICP refinement. The recall of the other baselines slowly increases for large success thresholds, failing to achieve accurate map-based localization.

We conclude this experiment by visualizing successful registrations in Figs. 6 and 7. Note that the colors of the 3D map are obtained via principal component analysis of the initial NCLT scene. Following previous work [25], we adopt the first three components as color channels in the RBG space. To identify the registered LiDAR scan within the 3D map, we do not employ this colorization to the scan but show it in red.

D. Robustness to Environmental Changes

This experiment is designed to support our third claim, i.e., the DINOv2-based descriptors are robust to temporal environmental changes resulting in outdated map data. While the previous experiment provides some insight due to the seasonal variations, e.g., snow and foliage in the NCLT dataset [11], we attempt to further amplify long-term changes. On the NCLT dataset, we carefully remove distinct objects from the 3D map following the steps visualized in Fig. 4. (1) Given an input map, (2) we initially query all tree-like points on the map by considering the similarity of the point descriptors with the descriptor of a point that is identified as part of a tree by a human supervisor. For simplicity, we adopt the Euclidean distance in the sRGB space after converting the point descriptors to the same PCA space as used in the visualization throughout this manuscript. If the distance is less than 50, we assume the point to be part of a tree. (3) Afterward, we use HDBSCAN [24] to cluster the tree-like points using their 3D coordinates. (4) Finally, for each cluster,

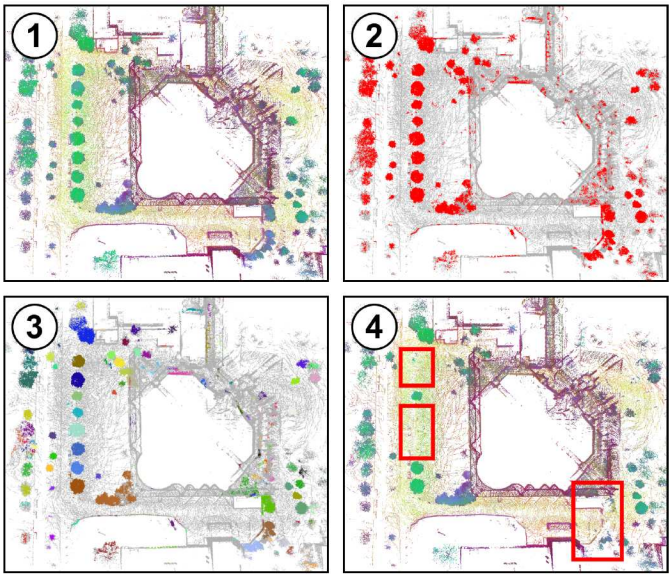

Fig. 4. To show the robustness of our proposed approach, we remove semantic entities from the 3D map. (1) The original map. (2) We identify tree-like points colored in red using the DINOv2-based descriptors. (3) We assign these points to separate clusters shown in different colors. (4) We randomly remove some clusters from the 3D map, highlighted by the red boxes.

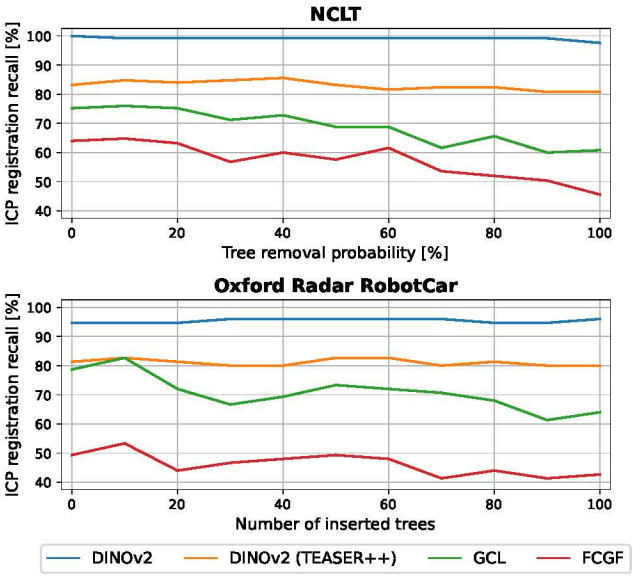

Fig. 5. We visualize the registration recall after ICP refinement (ICP-RR) for the removal and insertion of trees into the 3D map. Unlike the baseline methods, our proposed DINOv2-based point descriptor results in stable registration underlining its robustness.

we decide with a given probability whether to remove it from the map. Since the urban scenes of the Oxford Radar RobotCar dataset [6] contain fewer trees, we invert the idea. In particular, we insert up to 100 additional trees in the scenes that we previously extracted from an NCLT scene.

In Fig. 5, we report the registration recall after ICP refinement (ICP-RR) for both experiments using RANSAC and TEASER++ [31] with our proposed point descriptors. We further compute the scores for the two best-performing baselines (see Tab. II), i.e., GCL [23] and FCGF [12]. While the baseline descriptors suffer from some degradation due to increasing changes in the reference map, our DINOv2-based point descriptor results in stable registration throughout the experiments, supporting our claim of its robustness.

V. Discussion of the Limitations

In this section, we outline some limitations of our approach. As discussed in Sec. I, LiDAR-camera sensor configurations are a common setup on mobile robots. We further show in Sec. IV-C that the additional vision modality enables our method to substantially outperform LiDAR-only baselines. Nonetheless, fusing these modalities introduces new challenges such as accurate extrinsic calibration and time synchronization. A primary source of error arises from incorrect projection of DINOv2 features from image pixels to the wrong LiDAR points, which can lead to misalignments, such as a *tree* pixel being projected onto a building in the point cloud. Eventually, this results in an incorrect semantic description of a point. Other factors contributing to such errors include the presence of moving objects and varying fields of view of the sensors. Furthermore, limitations inherent to DINOv2, such as dependence on adequate illumination, may also affect performance. Lastly, we anticipate reduced performance when dealing with out-of-domain data, e.g., deployment in extraterrestrial environments like Mars. In most relevant scenarios though, this limitation is mitigated by the robust generalization capabilities of DINOv2.

VI. Conclusion

In this study, we demonstrated that features extracted from DINOv2 using surround-view image data can be reliably utilized to establish correspondences between point clouds. Through extensive experimentation, we showed that integrating these DINOv2-based point descriptors with traditional registration methods, such as RANSAC or ICP, substantially enhances scan-to-map registration, outperforming various handcrafted and learning-based baselines. We also verified the robustness of the descriptor against seasonal variations and long-term environmental changes. Future research will explore the potential application of visual foundation models directly to point cloud projections, eliminating the need for surround-view cameras and addressing challenges related to insufficient illumination and inaccuracies in point-to-pixel projection. Additionally, another direction for further research is explicitly combining the semantic richness of DINOv2 features with geometry-based features.

Acknowledgment

This work was partially supported by a fellowship of the German Academic Exchange Service (DAAD). Niclas Vödisch acknowledges travel support from the European Union's Horizon 2020 research and innovation program under ELISE grant agreement No. 951847 and from the ELSA Mobility Program within the project European Lighthouse On Safe And Secure AI (ELSA) under the grant agreement No. 101070617.

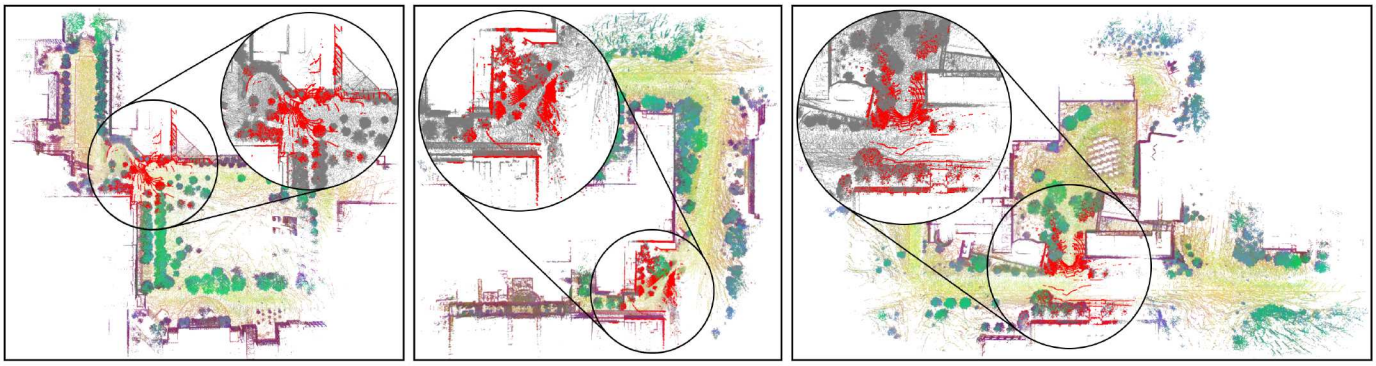

Fig. 6. Qualitative results on the NCLT dataset [11]. We colorize the 3D map by taking the first three components of performing principal component analysis (PCA) on the point descriptors of the map. The successfully registered LiDAR scan is shown in red.

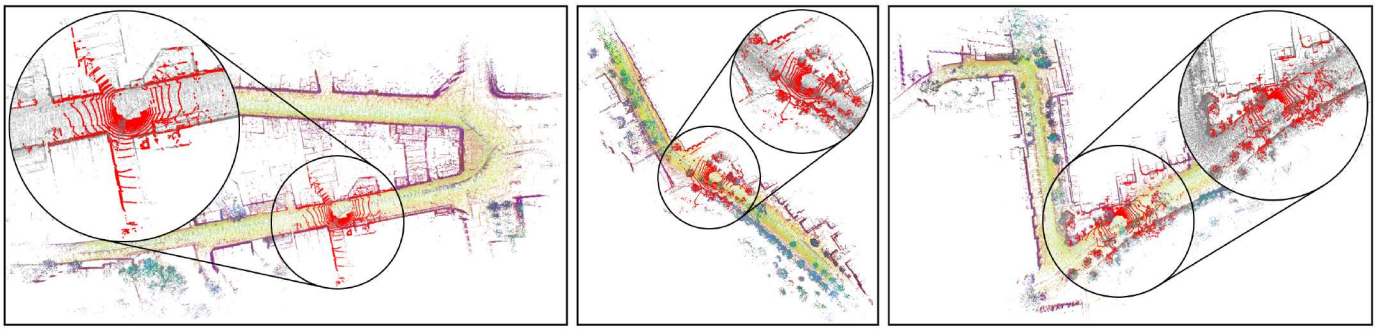

Fig. 7. Qualitative results on the Oxford Radar RobotCar dataset [6]. We colorize the 3D map by taking the first three components of performing principal component analysis (PCA) on the point descriptors of the map. The successfully registered LiDAR scan is shown in red.

REFERENCES

[1] Sheng Ao, Qingyong Hu, Bo Yang, Andrew Markham, and Yulan Guo. SpinNet: Learning a general surface descriptor for 3D point cloud registration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11748–11757, 2021.

[2] José Arce, Niclas Vödisch, Daniele Cattaneo, Wolfram Burgard, and Abhinav Valada. PADLoC: LiDAR-based deep loop closure detection and registration using panoptic attention. *IEEE Robotics and Automation Letters*, 8 (3):1319–1326, 2023.

[3] Xuyang Bai, Zixin Luo, Lei Zhou, Hongbo Fu, Long Quan, and Chiew-Lan Tai. D3Feat: Joint learning of dense detection and description of 3D local features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6358–6366, 2020.

[4] Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. PointDSC: Robust point cloud registration using deep spatial consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15854–15864, 2021.

[5] Daniel Barath, Jana Noskova, and Jiri Matas. Marginalizing sample consensus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8420–8432, 2022.

[6] Dan Barnes, Matthew Gadd, Paul Murcutt, Paul Newman, and Ingmar Posner. The Oxford Radar RobotCar Dataset: A radar extension to the Oxford RobotCar Dataset. In *IEEE International Conference on Robotics and Automation*, pages 6433–6438, 2020.

[7] Paul J. Besl and Neil D. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[8] Peter Biber and Wolfgang Strasser. The normal distributions transform: a new approach to laser scan matching. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 3, pages 2743–2748, 2003.

[9] Dirk Breitenreicher and Christoph Schnörr. Robust 3D object registration without explicit correspondence using geometric integration. *Machine Vision and Applications*, 21(5):601–611, 2010.

[10] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11618–11628, 2020.

[11] Nicholas Carlevaris-Bianco, Arash K Ushani, and Ryan M Eustice. University of Michigan North Campus long-term vision and lidar dataset. *International Journal of Robotics Research*, 35(9):1023–1035, 2016.

[12] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *International Conference on Computer Vision*, pages 8957–8965, 2019.

[13] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep

global registration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2511–2520, 2020.

[14] Jiaming Cui, Jiming Chen, and Liang Li. SAGE-ICP: Semantic information-assisted ICP. In *IEEE International Conference on Robotics and Automation*, pages 8537–8543, 2024.

[15] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

[16] Martin Fischler and Robert Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Readings in Computer Vision*, pages 726–740, 1987.

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2012.

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

[19] Ibrahim Hroob, Benedikt Mersch, Cyrill Stachniss, and Marc Hanheide. Generalizable stable points segmentation for 3D LiDAR scan-to-map long-term localization. *IEEE Robotics and Automation Letters*, 9(4):3546–3553, 2024.

[20] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. AnyLoc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 9(2):1286–1293, 2024.

[21] Giseop Kim, Byungjae Park, and Ayoung Kim. 1-day learning, 1-year localization: Long-term LiDAR localization using scan context image. *IEEE Robotics and Automation Letters*, 4(2):1948–1955, 2019.

[22] Giseop Kim, Sunwook Choi, and Ayoung Kim. Scan Context++: Structural place recognition robust to rotation and lateral variations in urban environments. *IEEE Transactions on Robotics*, 38(3):1856–1874, 2022.

[23] Quan Liu, Hongzi Zhu, Yunsong Zhou, Hongyang Li, Shan Chang, and Minyi Guo. Density-invariant features for distant point cloud registration. In *International Conference on Computer Vision*, pages 18169–18179, 2023.

[24] Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In *IEEE International Conference on Data Mining Workshops*, pages 33–42, 2017.

[25] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[26] G. Dias Pais, Srikumar Ramalingam, Venu Madhav Govindu, Jacinto C. Nascimento, Rama Chellappa, and Pedro Miraldo. 3DRegNet: A deep neural network for 3D point registration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7191–7201, 2020.

[27] Fabio Poiesi and Davide Boscaini. Distinctive 3D local deep descriptors. In *International Conference on Pattern Recognition*, pages 5720–5727, 2021.

[28] Fabio Poiesi and Davide Boscaini. Learning general and distinctive 3D local deep descriptors for point cloud registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3979–3985, 2023.

[29] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In *IEEE International Conference on Robotics and Automation*, pages 3212–3217, 2009.

[30] Ignacio Vizzo, Tiziano Guadagnino, Benedikt Mersch, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. KISS-ICP: In defense of point-to-point ICP – simple, accurate, and robust registration if done the right way. *IEEE Robotics and Automation Letters*, 8(2):1029–1036, 2023.

[31] Heng Yang, Jingnan Shi, and Luca Carlone. TEASER: Fast and certifiable point cloud registration. *IEEE Transactions on Robotics*, 37(2):314–333, 2021.

[32] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3DMatch: Learning local geometric descriptors from RGB-D reconstructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 199–208, 2017.

[33] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *European Conference on Computer Vision*, pages 766–782, 2016.