IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Point Cloud Registration with Object-Centric Alignment

BARE LUKA ŽAGAR[1], EKIM YURTSEVER[2], (MEMBER, IEEE), ARNE PETERS[1], ALOIS C. KNOLL.[1], (Senior Member, IEEE)

[1]Chair of Robotics, Artificial Intelligence and Real-Time Systems, Technical University of Munich, 85748 München, Germany
[2]College of Engineering, Center for Automotive Research, The Ohio State University, Columbus, OH 43212, USA

Corresponding author: Bare L. Žagar (e-mail: bare.luka.zagar@tum.de).

This result is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870133.

ABSTRACT Point cloud registration is a core task in 3D perception, which aims to align two point clouds. Moreover, the registration of point clouds with low overlap represents a harder challenge, where previous methods tend to fail. Recent deep learning-based approaches attempt to overcome this issue by learning to find overlapping regions in the whole scene. However, they still lack robustness and accuracy, and thus might not be suitable for real-world applications. Therefore, we present a novel registration pipeline that focuses on object-level alignment to provide a robust and accurate alignment of point clouds. By extracting and completing the missing points of the object of interest, a rough alignment can be achieved even for point clouds with low overlap captured from widely apart viewpoints. We provide a quantitative and qualitative evaluation on synthetic and real-world data captured with a Kinect v2. The proposed approach outperforms the current the current state-of-the-art methods by more than 29% w.r.t. the registration recall on the introduced synthetic dataset. We show that the overall performance and robustness increases due to the object-level alignment, while the baselines perform poorly as they take the entire scene into account.

INDEX TERMS Point Cloud Registration, Sensor Fusion, 3D Reconstruction, Deep Learning

I. INTRODUCTION

MULTI perception sensor setups with 3D depth sensors [1] and LIDARs [2] are becoming more and more prevalent for manufacturing. However, such multi sensor systems require accurate and robust extrinsic calibration in order to be usable. An increasing degree of automation in industrial manufacturing processes also raises a requirement for automated (re-)calibration, to keep the growing complexity in set-up and maintenance manageable.

While being a general computer vision problem, point cloud registration also provides a potential solution for extrinsic calibration of 3D sensors. The goal is to find the relative transformation between a point cloud pair with respect to a reference frame. Previous research was mainly based on traditional registration methods [4]–[6], which were combined, in most research works, with specific calibration objects [7]–[10] or markers [11]–[14]. Even though the target-based methods offer reliable and precise calibration, they are performed manually and require expert knowledge, which is not satisfactory for highly automated industrial processes. Later work [15] showed that automated and target-less calibration

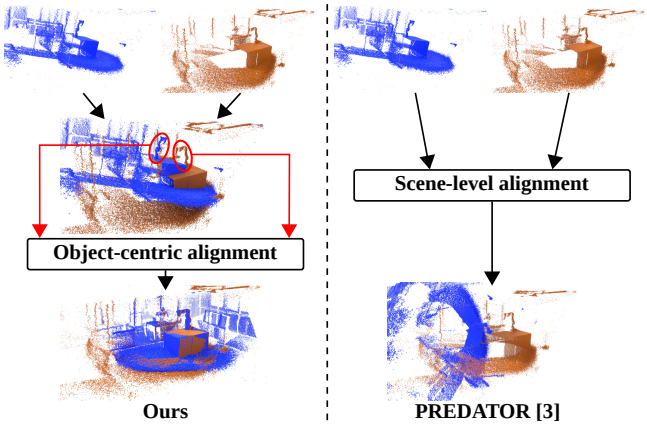

FIGURE 1: Right: The scene-level approach [3] focuses on the entire point clouds for finding correspondences and fails registration. **Left**: Our object-centric approach successfully registers the real-world point clouds. The proposed method finds and roughly aligns the object-of-interest in both point clouds. Then, this object-of-interest is used as a fulcrum point to fine-tune the scene-level matching process.

is possible but still relies on an approximate initial guess of the sensor placement. Moreover, traditional registration methods suffer from instability and lack robustness, if the input point clouds are captured from widely apart viewpoints and their overlap is relatively low.

However, these methods represent a bottleneck for reaching higher levels of autonomy in industrial processes where vision systems are vital. In such cases, learning-based techniques are used to overcome these issues. The tremendous success of deep learning for various 3D perception tasks [16]–[19] has resulted in the use of deep learning for point cloud registration as well. This can be seen in a number of approaches [3], [20]–[29] that appeared in the recent years. Despite learning-based approaches trying to mitigate the problems of previous registration methods, they require large amount of data, lack generalization and accuracy, and tend to fail when the test data distribution differs from the training data distribution.

Therefore, instead of learning low-level features on the entire point cloud, we could simply focus on an object of interest within the scene, as shown in figure 1. Moreover, a valid assumption for most relevant 3D computer vision applications is that there will always be at least one unique object, i.e. an object of interest in the scene. For example, vehicles in automated driving use cases, robot manipulators in industrial work cells, or furniture in domestic indoor scenes. Thus, by focusing on an object-centric alignment, we can overcome the problems of point clouds captured from different viewpoints, point clouds with low overlap, and the need for learning correspondences on the entire scene. Moreover, by applying this simple yet effective hypothesis, we can use any off-the-shelve methods and easily integrate them into our pipeline to adapt to any given requirements.

Hence, we propose a simple and modular registration pipeline for point cloud data to mitigate the limitations mentioned above. First, the object of interest is extracted from the input point cloud pair. The extracted points partially represent the object of interest, due to the self-occlusion of 3D sensors. Therefore, the next step in our registration pipeline predicts the missing points, which highly increases the similarity between the extracted point clouds. We leverage this similarity and perform a rough alignment of the completed point clouds of the object of interest. This provides a relatively good transformation estimation. Finally, we refine the alignment on the entire captured scene by using the roughly estimated transformation parameters as an initial guess.

Our main contributions can be summarized as follows:

- A novel registration pipeline based on object-level alignment
- The object extraction and completion modules that enable accurate and robust registration even for point clouds with low overlap
- Extensive experiments on a new synthetic dataset containing point clouds with low overlap captured from widely apart viewpoints, and qualitative evaluation on real point cloud data.

II. RELATED WORK

In this section, we provide an analysis of relevant related research. Furthermore, we extract the limitations for each subclass of the family of point cloud registration methods.

A. TRADITIONAL POINT CLOUD REGISTRATION

Point-based Registration Methods. The most known traditional optimization-based point cloud registration method is Iterative Closest Point (ICP), which was introduced by [30] and [31]. The core idea behind ICP is to iteratively search for correspondences and estimate the transformation between them, thus finding the optimal transformation between source and target point cloud. The main drawback of this method is that it heavily relies on a good initial pose estimate, which in case of a bad estimate can lead to convergence to a local-minima. To overcome this, [32], [33] and [34] use branch-and-bound to search for global optimal solution. Although these approaches may be effective with bad initial estimates, they still lack in terms of robustness in the case of point cloud pairs with low overlap. Additionally, global registration methods come with a high price in required computational effort, which makes them unusable for real-time applications. Finally, [35] introduces the estimation of the velocity of the rangefinder into the ICP algorithm to compensate for any kind of distortion caused by the movement of the sensor.

Handcrafted Descriptors. Contrary to optimization-based registration techniques, handcrafted descriptor-based approaches [36]–[39] try to extract relevant features from point cloud pairs, and thus find correspondences between them. Their advantage over most of the optimization-based methods is that handcrafted descriptors don't require an initial guess. However, some disadvantages of these methods are sensitivity due to noise and occlusions, which can result in wrong correspondences. Moreover, handcrafted feature extraction methods underperform when dealing with point cloud pairs with low overlap, because there might be fewer, or even none at all, matching correspondences in the two input point clouds.

B. LEARNING-BASED POINT CLOUD REGISTRATION

Feature Learning. The rapid advancement of data-driven deep learning approaches enabled the usage of these techniques for point cloud registration. Unlike the handcrafted feature extractors, feature learning approaches train deep neural networks on large training data sets for finding correspondences. 3DMatch [20] is one of the first feature learning point cloud registration approaches. It leverages volumetric data representation and 3D Convolutional Neural Networks (CNN) to learn 3D local descriptors for finding correspondences. The authors from [20] introduced the well-known real-world data registration benchmark under the same name as the method. In order to jointly capture local and global features, Deep Closest Point (DCP) [21] employs Dynamic Graph Convolutional Neural Network (DGCNN) [40] and leverages Transformer [41] to learn contextual information. Finally, an Singular Value Decomposition (SVD) module

produces the transformation matrix. A comprehensive survey of data-driven feature learning methods can be found in [42], including works up to 2021. More recent methods [22], [3] and [24] try to overcome the problem of registration of point clouds with low overlap. The approach in [22] enhances the quality of the correspondences, in a regime with low overlap, by using a graph-based self- and cross-attention network. PREDATOR [3] introduces a novel overlap-attention block that aims to focus more on the overlapping parts of point cloud pairs. [24] proposes to solve the registration of partially overlapping point clouds by learning overlapping masks to register those regions.

However, the main limitations of the above-mentioned methods are: 1) they need an immense amount of training data, and 2) if there is a relatively large gap between the training data and new scenes, then these methods suffer from a significant performance drop. 3) These methods still fail to accurately register extremely point cloud pairs with low overlap. On the contrary, our method focuses on finding correspondences on an object level. This object-centric strategy helps to address the aforementioned drawbacks. Moreover, our modular pipeline makes use of state-of-the-art methods and thus leveraging its strengths.

III. PROPOSED METHOD

Our novel point cloud registration method focuses on finding an object of interest in the input point cloud pairs for accurate and precise transformation estimation. Additionally, we make the assumption that a unique object of interest exists, specific for a particular use case, within any given scene. The proposed point cloud registration pipeline is modular, and thus enables the easy plug-and-play exchange of each module with other off-the-shelf methods or network models. Figure 2 shows our proposed calibration pipeline, which can be subdivided into four main modules:

1) We first extract the object of interest in the source and target point cloud (Sec. III-B).
2) The extracted points of the object of interest from both point clouds represent a partial point cloud representation and are completed within the point cloud completion module (Sec. III-C).
3) The completed point clouds of the object of interest from the input point cloud pair are roughly aligned using Principal Component Analysis (PCA) and ICP [30] (Sec. III-D).
4) The final step includes estimating the transformation parameters using ICP [30] (Sec. III-D). For the so-called initial guess, the estimated transformation parameters, from the previous step, are used.

The following subsections explain in detail the novel point cloud registration pipeline.

A. PROBLEM STATEMENT

Lets consider two input point clouds, source $\mathcal{P} = \{\mathbf{p}_1, ..., \mathbf{p}_i, ..., \mathbf{p}_M \mid \mathbf{p}_i \in \mathbb{R}^3\}$ and target point cloud $\mathcal{Q} = \{\mathbf{q}_1, ..., \mathbf{q}_i, ..., \mathbf{q}_N \mid \mathbf{q}_i \in \mathbb{R}^3\}$, where $M = N$ can be but is not necessary. Assume that the source and target point cloud have L point matches, where $0 < L < N$. The task of point cloud registration is to estimate the rigid transformation matrix $\mathbf{T}_{\mathcal{P}}^{\mathcal{Q}}$, which consists of a rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$, where $\mathbf{R} \in SO(3)$ and $\mathbf{t} \in \mathbb{R}^3$, by minimizing the least squares error:

$$E(\mathbf{R}, \mathbf{t}) = \frac{1}{L} \sum_{i=1}^{L} \|\mathbf{q}_i - (\mathbf{R}\mathbf{p}_i + \mathbf{t})\|^2. \tag{1}$$

The well-known ICP method tries to iteratively solve Eq. 1 by alternating in finding the right point matches, i.e. correspondences, and the optimal transformation matrix. Unfortunately, this approach is very sensitive to local optima and it fails to converge if the initial guess is poor. Therefore, we aim to provide a relatively well-aligned initial guess by focusing first on the object of interest in both the source and target point cloud.

B. OBJECT OF INTEREST EXTRACTION

As already mentioned, our approach finds first corresponding points on an object-level instead of searching for correspondences or features in the entire input point cloud sets like it is done by previous methods. Thus, the first step is to extract the object of interest point clouds $\mathcal{P}_S = \{\mathbf{p}_{S,1}, ..., \mathbf{p}_{S,i}, ..., \mathbf{p}_{S,J} \mid \mathbf{p}_{S,i} \in \mathbb{R}^3\}$ and $\mathcal{Q}_S = \{\mathbf{q}_{S,1}, ..., \mathbf{q}_{S,i}, ..., \mathbf{q}_{S,K} \mid \mathbf{q}_{S,i} \in \mathbb{R}^3\}$ from the source $\mathcal{P}$ and target point cloud $\mathcal{Q}$, where $\mathcal{P}_S \subset \mathcal{P}$, $\mathcal{Q}_S \subset \mathcal{Q}$, and $J = K$ can be but is not necessary. It can be described with:

$$\begin{aligned} \mathcal{P}_S &= f_e(\mathcal{P}), \\ \mathcal{Q}_S &= f_e(\mathcal{Q}), \end{aligned} \tag{2}$$

where f_e is a function that extracts the points of the object of interest from the input point cloud data. Since our approach is modular, the function for extracting the object of interest points can be implemented by any method which is able to distinguish the object of interest from the background. For example, a 3D object detection module, trained to detect the object of interest, can be used for this task.

In our experiments, we use the DGCNN [40] semantic segmentation network as f_e from equation 2 for extracting the object of interest points from the background. DGCNN is a lightweight graph-based network architecture leveraging edge convolution operations. The input point cloud is converted to a graph-based structure by using k nearest neighbours. Furthermore, their newly introduced edge convolution operation specifically combines global information with the local neighbourhood information. For more details, please refer to the original work.

C. POINT CLOUD COMPLETION

The extracted points $\mathcal{P}_S$ and $\mathcal{Q}_S$, of the same object of interest from both input point clouds, represent only partially the object of interest. This is due to obvious self-occlusion since the 3D sensor can only capture one side of an object.

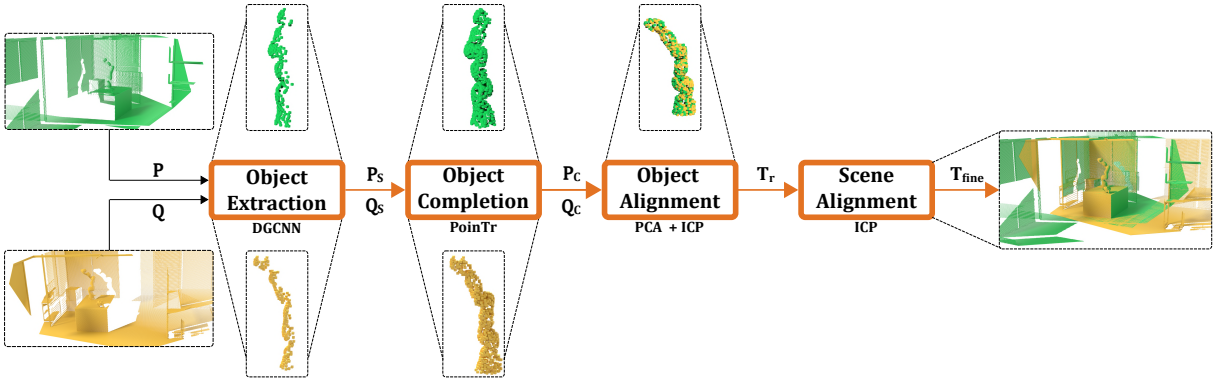

FIGURE 2: Our point cloud registration pipeline. The input point clouds are first passed through the object extraction module to retrieve the points from the object of interest. Then, the point cloud completion module infers the missing points of the object of interest. Object-Level alignment is applied on the completed object of interest point clouds. Afterward, the scene-level fine alignment module refines the point cloud registration on all the points.

Furthermore, we can assume that the two extracted point clouds only partially overlap which is caused by different viewpoints while capturing the input point clouds. The larger the translational and rotational difference between the two input point clouds is the smaller the expected overlap between them, and thus the harder the estimation problem. To tackle these issues we propose to predict the missing points of the extracted object of interest point clouds with:

$$\begin{aligned} \mathcal{P}_C &= f_c(\mathcal{P}_S), \\ \mathcal{Q}_C &= f_c(\mathcal{Q}_S), \end{aligned} \tag{3}$$

where f_c is a function that predicts the missing points of $\mathcal{P}_S$ and $\mathcal{Q}_S$. The complete point cloud representations of the extracted object of interests is denoted with $\mathcal{P}_C = \{\mathbf{p}_{C,1}, ..., \mathbf{p}_{C,i}, ..., \mathbf{p}_{C,U} \mid \mathbf{p}_{C,i} \in \mathbb{R}^3\}$ and $\mathcal{Q}_C = \{\mathbf{q}_{C,1}, ..., \mathbf{q}_{C,i}, ..., \mathbf{q}_{C,V} \mid \mathbf{q}_{C,i} \in \mathbb{R}^3\}$, where $U = V$ can be but is not necessary, and $\mathcal{P}_S \subset \mathcal{P}_C$ and $\mathcal{Q}_S \subset \mathcal{Q}_C$. The aim of completing the two extracted point clouds is to get a set of points that are similar w.r.t. their geometrical shape. We use this similarity between the two completed point clouds of the object of interest to roughly align them, but this is described in more detail in the following section.

To infer missing points from partial input point clouds, we use the PoinTr [43] network as f_c from equation 3. PoinTr is a transformer-based network architecture for the task of point cloud completion. To process the incomplete point cloud a lightweight DGCNN model is employed. However, to reduce the computational cost, the input point cloud is hierarchically downsampled using farthest point sampling (FPS). More details can be found in the original research article.

D. TRANSFORMATION PARAMETER ESTIMATION

As mentioned in Sec. III-A, the ICP algorithm is prone to errors if the initial alignment is inaccurate, which leads to a bad transformation matrix estimation, hence a bad registration. We solve this issue by providing a relatively good initial alignment of the input point clouds, by leveraging the similarity of the completed point clouds of the object of interest. We apply PCA [44], by following [45] and [46], and find the covariance matrices, $\mathbf{C}_{\mathcal{P}_C} \in \mathbb{R}^{3\times3}$ and $\mathbf{C}_{\mathcal{Q}_C} \in \mathbb{R}^{3\times3}$, of both completed point clouds:

$$\begin{aligned} \mathbf{C}_{\mathcal{P}_C} &= \frac{1}{U}\sum_{i=1}^{U}(\mathbf{p}_{C,i} - \bar{\mathbf{p}}_C)(\mathbf{p}_{C,i} - \bar{\mathbf{p}}_C)^T, \\ \mathbf{C}_{\mathcal{Q}_C} &= \frac{1}{V}\sum_{i=1}^{V}(\mathbf{q}_{C,i} - \bar{\mathbf{q}}_C)(\mathbf{q}_{C,i} - \bar{\mathbf{q}}_C)^T, \end{aligned} \tag{4}$$

where the centroids of the completed point clouds, $\bar{\mathbf{p}}_C \in \mathbb{R}^3$ and $\bar{\mathbf{q}}_C \in \mathbb{R}^3$, are calculated with:

$$\begin{aligned} \bar{\mathbf{p}}_C &= \frac{1}{U}\sum_{i=1}^{U}\mathbf{p}_{C,i}, \\ \bar{\mathbf{q}}_C &= \frac{1}{V}\sum_{i=1}^{V}\mathbf{q}_{C,i}. \end{aligned} \tag{5}$$

The point cloud reference system for each completed object of interest point cloud is defined by its principal components, i.e. feature vectors of the previously calculated covariance matrix, and with the centroid as its origin. By aligning the reference frame of the two completed point clouds, retrieved from the source and target input points, and applying ICP for further refinement, we obtain an object-level alignment $\mathbf{T}^{\mathcal{Q}_C}_{\mathcal{P}_C,\text{rough}}$.

Finally, we use the object-level alignment as the initial guess for the minimization problem of equation 1 and solve it using ICP on the entire input point clouds for finding $\mathbf{T}^{\mathcal{Q}}_{\mathcal{P},\text{fine}}$. Since we can provide an initial guess, which is already close to the optimal solution, the ICP algorithm converges and finds the optimal solution without getting stuck in local optima.

IV. EXPERIMENTS

In this section, we give a detailed description of the used dataset and give an overview of the implementation details and the used evaluation metrics. Then we provide an ablation study showing the contribution of our object-centric strategy. Finally, we compare the quantitative and qualitative

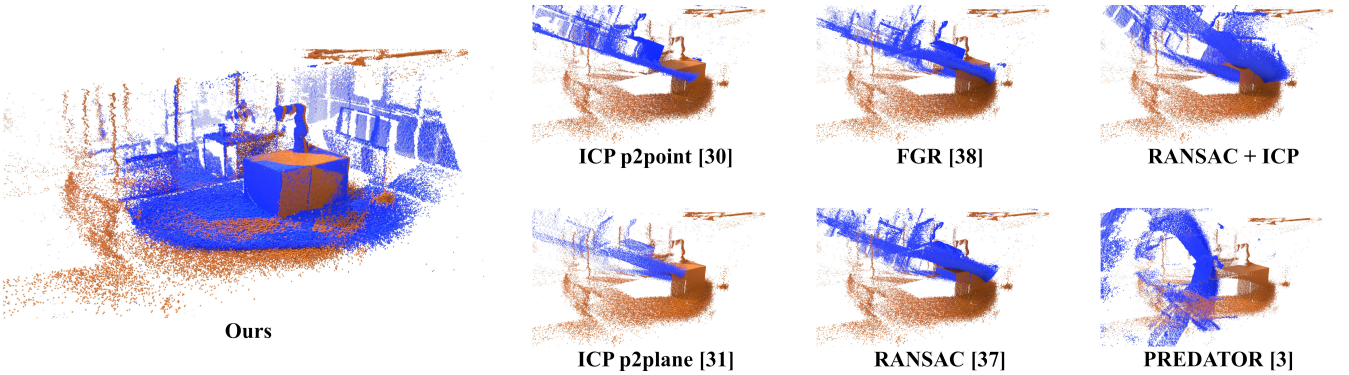

FIGURE 3: Qualitative comparison on real point cloud data captured with a Kinect v2. The traditional baseline methods underperform because searching for correspondences in the entire scene is prone to failure. On the other hand, only our method achieves satisfying registration due to the emphasis on object-centric alignment.

FIGURE 4: Kuka LBR iiwa collaborative robot mounted on a worktable inside our workcell dedicated for research purposes. We use multiple Kinect v2 depth sensors mounted in the corners of the workcell to get full coverage. The extrinsic calibration, i.e. point cloud registration, of the 3D vision system is performed with our proposed method, because it provides robust and accurate alignment due to the object-centric approach. Additionally, for the indoor real-world scenario, we assume that the movement of the object of interest, in our case the robot arm, cannot cause any significant positional discrepancy in the depth sensors. This assumption is valid because the time synchronization of multiple connected depth sensors is usually by magnitudes faster than the movement of any object within a scene.

performance of our method to traditional baselines, i.e. ICP [30] and [31], Fast Glogal Registration (FGR) [38], Random Sample Consensus (RANSAC) [37] and RANSAC followed by ICP, and, lastly, with one state-of-the-art feature learning-based method PREDATOR [3]. Firstly, we describe the used dataset.

A. DATASET

We assume that for most 3D computer vision applications that require point cloud registration as a necessary preprocessing step, a unique object of interest will be present within the captured scene. Hence, let us consider the scenario of an indoor industrial robotics workcell inside a manufacturing plant. We can assume that a robot manipulator will be present in all the captured scenes since it represents the main element for the operation of a robot workcell. Thus, we can consider the robot manipulator as our unique object of interest in any given robotic workcell use case. To the best of our knowledge, there are no open-source datasets that satisfy our task description.

TABLE 1: Synthetic dataset details.

SIR Dataset	**Train**	**Validation**	**Test**
Maps	1	1	1
Point cloud pairs	2250	250	250
Overlap ratio	**Mean**	**Max**	**Min**
	30.16($\pm$2.75)	36.32	21.48

Therefore, we introduce a new synthetic dataset, containing dense point clouds of an industrial workcell with a Kuka LBR iiwa inside it. This dataset is generated using Blender [47] by realistically recreating our real-world lab robotic workcell, as can be seen in figure 4 and its real-world 3D scan taken with a Kinect v2 in figure 3, and contains 2750 scan pairs with randomly sampled robot joint states for each scan. Additionally, the scans were taken from random poses within the workcell with the condition that the robot arm is inside the field of view. Our synthetic dataset can be used to train for semantic segmentation, point cloud completion, and point cloud registration tasks. Therefore, we give the ground truth point-wise labels containing either the background class or the robot arm class. In addition to that, we provide the complete robot arm point cloud for each scan as ground truth data, in order to be able to train a point cloud completion network. The split into subsets for training, validation, and testing can be seen in table 1. We follow [3] to calculate the overlap ratio between the point cloud pairs, which is reported in table 1 as well. The overlap ratio tells us how many points of the perfectly aligned source and target point cloud lie within a threshold distance. The lower the

overlap ratio between two input point clouds is, the fewer potential correspondences exist, thus making the registration problem harder. The mean overlap ratio of the point cloud pairs from our introduced synthetic dataset is relatively low at around 30%. If we compare our dataset with the well-known 3DMatch dataset [20], where only scan pairs with an overlap $> 30\%$ are considered, we can see that our dataset represents a harder challenge for registration.Specifically, point cloud pairs with an overlapping region $< 30\%$ are considered to be low overlapping [3] and thus various methods show a rapid decrease in performance.

B. IMPLEMENTATION DETAILS

TABLE 2: Quantitative evaluation of semantic segmentation of DGCNN [40] on the synthetic dataset. We use the intersection over union metric to show the point-wise semantic segmentation performance on our synthetic dataset.

DGCNN [40]	**Train**	**Validation**	**Test**
IoU ↑	0.914	0.902	0.898

Our method is implemented in the programming language Python using the well-known machine learning framework PyTorch [48] and Open3D [49] for 3D data processing and visualization. As mentioned previously, we used Blender [47] to generate our synthetic dataset. For the object of interest extraction module, we use the DGCNN [40], trained on our dataset by following their recommendations regarding the hyperparameters. We report in table 2 the performance of DGCNN [40] on our synthetic dataset based on the Intersection over Union (IoU) metric. The point cloud completion is obtained by using the PoinTr network [43], again following their hyperparameter recommendations. We trained PoinTr on our dataset, with the addition of generating the corresponding ground truth, i.e. the complete robot arm point cloud representation for each scan, with Blender. Table 3 shows the performance of PoinTr [43] on our introduced dataset. We set the threshold for ICP in the rough and fine alignment steps to 0.01 and 0.1, respectively. The experiments were conducted on our workstation PC with an AMD Ryzen Threadripper 2950X (16-Core) and an NVIDIA GeForce RTX 3090 GPU.

TABLE 3: Quantitative evaluation of point cloud completion of PoinTr [43] on the synthetic dataset. We use the L1 Chamfer Distance to show the performance on our synthetic dataset.

PoinTr [43]	**Train**	**Validation**	**Test**
CD - $\mathrm{l}_1(\times 1000) \downarrow$	14.712	18.377	18.754

For the traditional baselines, we use the implementation provided in the Open3D library [49], while for the feature learning-based method we use their publicly available open-source implementation. For fairness, we trained the feature learning-based method, PREDATOR [3], on our introduced synthetic data set by using their recommended hyperparameter settings.

C. EVALUATION METRICS

We follow [3], [50], [22], and evaluate the point cloud registration performance w.r.t. the relative translation (RTE) and rotation error (RRE) calculated by:

$$\begin{aligned} \mathbf{RTE} &= \|\hat{\mathbf{t}} - \mathbf{t}_{gt}\|_2, \\ \mathbf{RRE} &= \arccos \frac{Tr(\hat{\mathbf{R}}\mathbf{R}_{gt})}{2}, \end{aligned} \tag{6}$$

where $\hat{\mathbf{R}}$ and $\hat{\mathbf{t}}$ are the estimated, and, $\mathbf{t}_{gt}$ and $\mathbf{R}_{gt}$ the ground truth transformation parameters. Based on RTE and RRE, we calculate the mean translation (MTE) and rotation error (MRE), in order to evaluate the performance of the compared methods. Additionally, we also calculate the registration recall rate (RR), which gives a quantitative measure of the registration success ratio. A registration is considered successful if the relative translation and rotation error is below a certain threshold. For our use case, we consider the following thresholds: $RTE < 0.05m$ and $RRE < 5°$.

D. ABLATION STUDY

TABLE 4: Ablation study of the proposed registration pipeline performed on our synthetic test set. The threshold for the relative translation and rotation error were set to $0.05m$ and $5\,\mathrm{deg}$.

Method	MTE $[m] \downarrow$	MRE [deg] ↓	RR [%] ↑
A	5.85^*	106.61^*	0
B	3.92^*	86.41^*	0
C	0.056	1.418	8.84
D	0.0073	0.223	20.88
Full	0.0065	0.211	86.35

* We report the overall MTE and MRE if no point cloud pair is successfully registered

We ablate our proposed registration pipeline to prove the robustness of our object-centric alignment strategy, as shown in table 4. The ablation study is conducted on the introduced synthetic data set. First, we evaluate the performance of only using the object extraction module (model A) and apply ICP on the partial point clouds. By adding the object completion module (model B) and applying ICP on it, we can observe how the performance improved. By using the object extraction, object completion, and object-level alignment module (model C), we see a clear increase in performance, but still, the overall registration recall remains relatively low for the selected thresholds. Model D represents all the modules except the object completion module. Here, we intend to emphasize the importance of the object completion module. Additionally, model D simulates a scenario where the registration pipeline fails to correctly complete the object of interest, which is caused by either poor point-wise extraction of the object extraction module or a poor reconstruction of the extracted points of the object completion module. However, if we compare model D with the entire pipeline

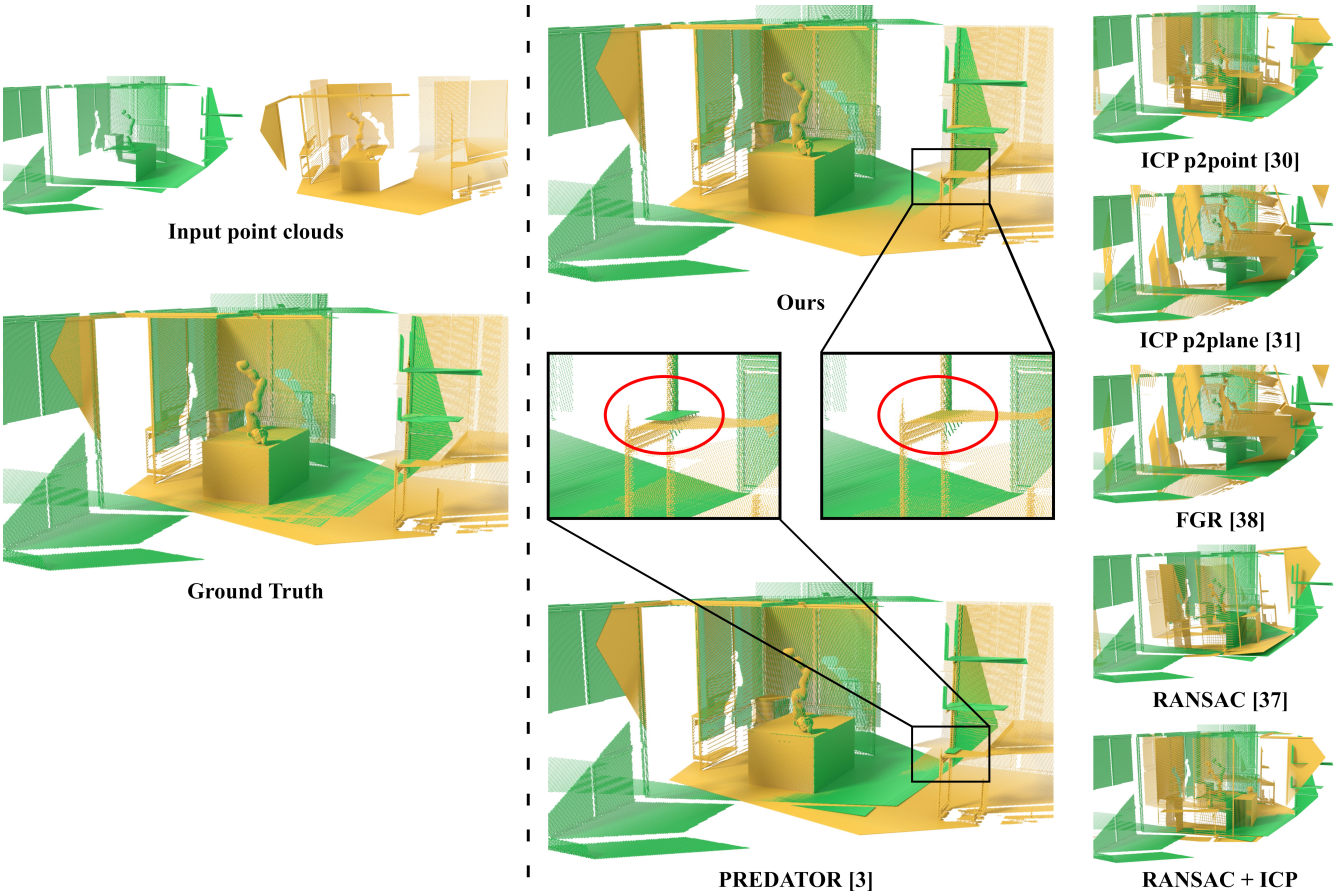

FIGURE 5: Qualitative comparisons on our synthetic dataset. The traditional methods fail because of the large viewpoint difference of the point clouds, and due to the low overlap. Even though our method and the learning-based baseline appear to be visually identical at the scene-level, it is clear that our method performs better for the points measured on the shelf, which can be seen in the zoomed in region. Our approach has a better registration quality because we use an object of interest as a fulcrum point for rough alignment. This rough alignment serves as a good starting point for the fine alignment step at the scene level and makes the search for correct correspondences easier.

(model Full), one can see the importance of the object completion module, which adds to the overall robustness of the registration pipeline by improving the registration recall by more than 66%. The robustness increases because the object completion module minimizes the discrepancy of the geometrical shapes between the two extracted object-of-interest point clouds significantly. Moreover, this step enables a rough alignment which serves as the initial guess for the fine alignment step.

E. QUANTITATIVE RESULTS

Traditional methods are not robust against point cloud pairs with low overlap, because they can't find enough relevant correspondences, which results in poor performance. Only the method, where RANSAC is used together with ICP, manages to register a few point cloud pairs very accurately, but still, due to a very low registration recall, this approach remains unusable for real-world applications. On the other hand, PREDATOR manages to generalize well over the test set, showing the robustness of a learning-based method specifically designed for point cloud pairs with low overlap.

TABLE 5: Evaluation results on our synthetic test set. The threshold for the relative translation and rotation error were set to $0.05m$ and $5\,\mathrm{deg}$.

Method	MTE $[m]\downarrow$	MRE $[\mathrm{deg}]\downarrow$	RR $[\%]\uparrow$	A.R. $[s]\downarrow$
ICP [1] [30]	5.81*	107.21*	0	48.47
ICP [2] [31]	5.44*	106.31*	0	38.17
FGR [38]	6.72*	137.92*	0	13.24
RANSAC [37]	5.82*	112.23*	0	9.27
RANSAC + ICP	0.0085	0.013	4.42	9.80
PREDATOR [3]	0.0288	0.706	57.03	0.135
Ours	0.0065	0.2111	86.35	4.02

* We report the overall MTE and MRE if no point cloud pair is successfully registered
[1] point to point
[2] point to plane

However, our proposed method performs similar or more accurately, w.r.t. the MTE and MRE, and is more robust in terms of registration recall compared to the baseline methods, as shown in 5. The reason for achieving such robustness is because our method focuses first on finding an object of interest and predicts the missing points to generate a similar shape. On the other hand, high accuracy is obtained with the combination of the object-level and scene alignment. The

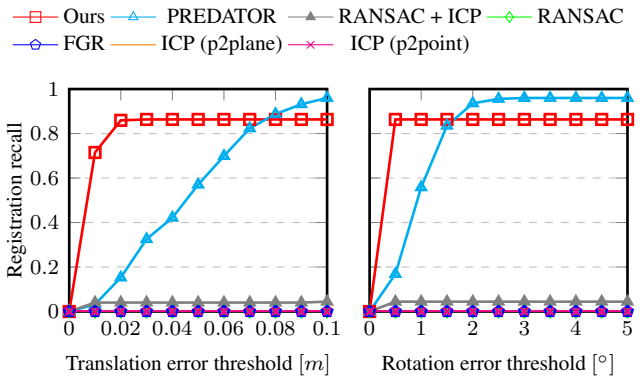

FIGURE 6: Registration recall with different translation and rotation error thresholds on the introduced synthetic dataset. Our approach outperforms the baselines in the low threshold region by a large margin, because of the object-centric alignment strategy followed by a scene-level fine alignment step.

alignment on the entire scene begins with a good initial guess, retrieved from the object-level alignment, which guarantees convergences to an optimal solution for most cases.

Furthermore, we evaluate the registration recall with different translation and rotation error thresholds, as shown in figure 6. Comparing the results, our method shows superiority for lower threshold values, which indicates that it is highly reliable for applications with a strict requirement on accuracy, such as industrial robotic workcell use cases. The better performance of our method can be attributed to the effective strategy of our novel registration pipeline. Instead of finding low-level correspondences in the entire scene, we first focus on roughly aligning an object of interest to provide a good starting point for the scene level fine alignment.

F. QUALITATIVE RESULTS

The qualitative comparison of the baseline methods and the introduced registration approach on the synthetic test data is shown in figure 5. As expected, the traditional methods suffer from instability and fail to achieve satisfying alignment, because of the widely apart viewpoints the point clouds were captured from, and due to the low overlap. On the other hand, PREDATOR manages to handle such input pairs and successfully registers them. However, by comparing the highlighted part of PREDATOR and our method, it is clearly visible that the learning-based baseline falls short of accurately aligning the input point cloud pair. The reason for this might be that the learning-based method probably requires larger amount of training data in order to learn more fine-grained correspondences.

Finally, a qualitative comparison on real-world point cloud data, is displayed in figure 3. The scenes were captured with two Kinect v2 which were mounted in the corners of our workcell, as shown in figure 4. Again, the traditional baselines fail for the same reasons to successfully register the point clouds with low overlap. Surprisingly, the learning-based method, PREDATOR, fails as well to align the real point cloud pair. This is very likely due to the difference in data distribution between the training set and the real point cloud data. However, only our approach successfully registers the real input point clouds, which can be attributed again to the effective design choice of the proposed registration pipeline, by first putting the focus on an object-level instead of on the entire scene.

V. DISCUSSION AND FUTURE WORK

The design of our point cloud registration pipeline enables two properties: 1) scalability and 2) simplicity. Each module within our point cloud registration pipeline can be exchanged with any other off-the-shelf method and adapted accordingly. Therefore, our proposed method can be easily extended for other applications where an accurate and robust registration of challenging point cloud pairs is required, e.g. automated driving, 3D indoor mapping, multi-agent slam, and others. Moreover, our approach opens up a number of directions for further research. It would be interesting to see how our method could be used for cross-source point cloud registration, where different densities of the input point clouds present a difficult challenge for current methods. Finally, publicly available datasets, such as the 3DMatch [20], lack the ground truth information about the completed point cloud for potential objects of interest within the scene. Therefore, to fill this gap, we believe that our synthetic dataset will help further research in this particular direction. Finally, the conducted experiments on synthetic and on real-world data showcase the robustness and accuracy of our object-centric alignment strategy.

VI. CONCLUSION

In this work, we introduced a simple and modular approach for robust and accurate registration of point clouds with low overlap. The main idea behind this novel registration pipeline was to put focus on an object of interest in the input point cloud pair and use it as a fulcrum point. Inferring the missing points of the object of interest created a geometrically similar shape of it in both the input point clouds, which then helped to roughly align them. This rough alignment provided a good and robust initial guess for the scene-level fine alignment, and thus ensured convergence to an optimal solution. Moreover, we showed that the introduced approach outperforms other baselines on our synthetic dataset, and our method proved to be robust on even noisy real-world data while the compared baselines failed.

ACKNOWLEDGMENT

This work was funded in part by the United States Department of Transportation under award number 69A3551747111 for Mobility21: the National University Transportation Center for Improving Mobility. Any findings, conclusions, or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the United States Department of Transportation, Carnegie Mellon University, or The Ohio State University.

REFERENCES

[1] K. P. Cop, A. Peters, B. L. Žagar, D. Hettegger, and A. C. Knoll, "New metrics for industrial depth sensors evaluation for precise robotic applications*," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 5350–5356.

[2] J. Lambert, A. Carballo, A. M. Cano, P. Narksri, D. Wong, E. Takeuchi, and K. Takeda, "Performance analysis of 10 models of 3d lidars for automated driving," *IEEE Access*, vol. 8, pp. 131 699–131 722, 2020.

[3] S. Huang, Z. Gojcic, M. Usvyatsov, A. Wieser, and K. Schindler, "Predator: Registration of 3d point clouds with low overlap," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4267–4276.

[4] M. He, H. Zhao, F. Davoine, J. Cui, and H. Zha, "Pairwise lidar calibration using multi-type 3d geometric features in natural scene," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 1828–1835.

[5] J. Jiao, Y. Yu, Q. Liao, H. Ye, R. Fan, and M. Liu, "Automatic calibration of multiple 3d lidars in urban environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 15–20.

[6] J. Jiao, Q. Liao, Y. Zhu, T. Liu, Y. Yu, R. Fan, L. Wang, and M. Liu, "A novel dual-lidar calibration algorithm using planar surfaces," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, 2019, pp. 1499–1504.

[7] R. S. Yang, Y. H. Chan, R. Gong, M. Nguyen, A. G. Strozzi, P. Delmas, G. Gimel'farb, and R. Ababou, "Multi-kinect scene reconstruction: Calibration and depth inconsistencies," in *2013 28th International Conference on Image and Vision Computing New Zealand (IVCNZ 2013)*, 2013, pp. 47–52.

[8] M. Ruan and D. Huber, "Calibration of 3d sensors using a spherical target," in *2014 2nd International Conference on 3D Vision*, vol. 1, 2014, pp. 187–193.

[9] P.-C. Su, J. Shen, W. Xu, S.-C. S. Cheung, and Y. Luo, "A fast and robust extrinsic calibration for rgb-d camera networks," *Sensors*, vol. 18, no. 1, 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/1/235

[10] N. Heide, T. Emter, and J. Petereit, "Calibration of multiple 3d lidar sensors to a common vehicle frame," in *ISR 2018; 50th International Symposium on Robotics*, 2018, pp. 1–8.

[11] C. Chen, B. Yang, S. Song, M. Tian, J. Li, W. Dai, and L. Fang, "Calibrate multiple consumer rgb-d cameras for low-cost and efficient 3d indoor mapping," *Remote Sensing*, vol. 10, no. 2, 2018. [Online]. Available: https://www.mdpi.com/2072-4292/10/2/328

[12] J. Chaochuan, Y. Ting, W. Chuanjiang, F. Binghui, and H. Fugui, "An extrinsic calibration method for multiple RGB-d cameras in a limited field of view," *Measurement Science and Technology*, vol. 31, no. 4, p. 045901, jan 2020. [Online]. Available: https://doi.org/10.1088/1361-6501/ab48b3

[13] G. Chen, G. Cui, Z. Jin, F. Wu, and X. Chen, "Accurate intrinsic and extrinsic calibration of rgb-d cameras with gp-based depth correction," *IEEE Sensors Journal*, vol. 19, no. 7, pp. 2685–2694, 2019.

[14] F. Basso, E. Menegatti, and A. Pretto, "Robust intrinsic and extrinsic calibration of rgb-d cameras," *IEEE Transactions on Robotics*, vol. 34, no. 5, pp. 1315–1332, 2018.

[15] A. Peters, A. Schmidt, and A. C. Knoll, "Extrinsic calibration of an eye-in-hand 2d lidar sensor in unstructured environments using icp," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 929–936, 2020.

[16] F. Shao, L. Chen, J. Shao, W. Ji, S. Xiao, L. Ye, Y. Zhuang, and J. Xiao, "Deep learning for weakly-supervised object detection and object localization: A survey," *arXiv preprint arXiv:2105.12694*, 2021.

[17] Y. Zhao, X. Zhang, and X. Huang, "A technical survey and evaluation of traditional point cloud clustering methods for lidar panoptic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 2464–2473.

[18] L. Roldao, R. de Charette, and A. Verroust-Blondet, "3d semantic scene completion: a survey," *arXiv preprint arXiv:2103.07466*, 2021.

[19] M. P. Muresan, S. Nedevschi, and I. Giosan, "Real-time object detection using a sparse 4-layer lidar," in *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2017, pp. 317–322.

[20] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser, "3dmatch: Learning local geometric descriptors from rgb-d reconstructions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1802–1811.

[21] Y. Wang and J. M. Solomon, "Deep closest point: Learning representations for point cloud registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3523–3532.

[22] E. Arnold, S. Mozaffari, and M. Dianati, "Fast and robust registration of partially overlapping point clouds," *IEEE Robotics and Automation Letters*, 2021.

[23] B. Wu, J. Ma, G. Chen, and P. An, "Feature interactive representation for point cloud registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 5530–5539.

[24] H. Xu, S. Liu, G. Wang, G. Liu, and B. Zeng, "Omnet: Learning overlapping mask for partial-to-partial point cloud registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 3132–3141.

[25] S. Ao, Q. Hu, B. Yang, A. Markham, and Y. Guo, "Spinnet: Learning a general surface descriptor for 3d point cloud registration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 753–11 762.

[26] Z. J. Yew and G. H. Lee, "Rpm-net: Robust point matching using learned features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 824–11 833.

[27] C. Choy, W. Dong, and V. Koltun, "Deep global registration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2514–2523.

[28] H. Yu, F. Li, M. Saleh, B. Busam, and S. Ilic, "Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[29] F. Lu, G. Chen, Y. Liu, L. Zhang, S. Qu, S. Liu, and R. Gu, "Hregnet: A hierarchical network for large-scale outdoor lidar point cloud registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 014–16 023.

[30] P. Besl and H. McKay, "A method for registration of 3-d shapes. ieee trans pattern anal mach intell," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 14, pp. 239–256, 03 1992.

[31] Y. Chen and G. Medioni, "Object modeling by registration of multiple range images," in *Proceedings. 1991 IEEE International Conference on Robotics and Automation*, 1991, pp. 2724–2729 vol.3.

[32] C. Olsson, F. Kahl, and M. Oskarsson, "Branch-and-bound methods for euclidean registration problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 783–794, 2009.

[33] J. Yang, H. li, D. Campbell, and Y. Jia, "Go-icp: A globally optimal solution to 3d icp point-set registration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 1–1, 12 2015.

[34] D. Campbell and L. Petersson, "Gogma: Globally-optimal gaussian mixture alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[35] S. Hong, H. Ko, and J. Kim, "Vicp: Velocity updating iterative closest point algorithm," in *2010 IEEE International Conference on Robotics and Automation*, 2010, pp. 1893–1898.

[36] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *2009 IEEE International Conference on Robotics and Automation*, 2009, pp. 3212–3217.

[37] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," vol. 24, no. 6, 1981.

[38] Q.-Y. Zhou, J. Park, and V. Koltun, "Fast global registration," vol. 9906, 10 2016.

[39] H. Yang, J. Shi, and L. Carlone, "Teaser: Fast and certifiable point cloud registration," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 314–333, 2020.

[40] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *Acm Transactions On Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[42] X. Huang, G. Mei, J. Zhang, and R. Abbas, "A comprehensive survey on point cloud registration," *arXiv preprint arXiv:2103.02690*, 2021.

[43] X. Yu, Y. Rao, Z. Wang, Z. Liu, J. Lu, and J. Zhou, "Pointr: Diverse point cloud completion with geometry-aware transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 498–12 507.

[44] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers & Geosciences*, vol. 19, pp. 303–342, 1993.

[45] B. Bellekens, V. Spruyt, R. Berkvens, and M. Weyn, "A survey of rigid 3d pointcloud registration algorithms," 08 2014.

[46] C. Yuan, X. Yu, and Z. Luo, "3d point cloud matching based on principal component analysis and iterative closest point algorithm," in *2016 Inter-*

national Conference on Audio, Language and Image Processing (ICALIP), 2016, pp. 404–408.

[47] B. O. Community, *Blender - a 3D modelling and rendering package*, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. [Online]. Available: http://www.blender.org

[48] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[49] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," *arXiv:1801.09847*, 2018.

[50] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser, "3dmatch: Learning local geometric descriptors from rgb-d reconstructions," in *CVPR*, 2017.

BARE LUKA ŽAGAR is a Ph.D. candidate currently working in the Chair of Robotics, Artificial Intelligence and Real-time Systems at the Technical University of Munich (TUM). He completed his B.Sc. and M.Sc. in Control Engineering and Automation at the University of Zagreb (Faculty of Electrical Engineering and Computing). He wrote his Master thesis on the topic of.

Before starting at TUM, he worked for more than two years in the industry as a robotics engineer, which gave him wide practical experience and knowledge in the field of robotics. His further research interests include computer vision for robotic navigation and manipulation, motion planning, 3d reconstruction and human-robot interaction.

EKIM YURTSEVER (Member, IEEE) received the B.S. and M.S. degrees from Istanbul Technical University, in 2012 and 2014, respectively, and the Ph.D. degree in information science from Nagoya University, Japan, in 2019.

Since 2019, he has been with the Department of Electrical and Computer Engineering, The Ohio State University, as a Research Associate. Currently, he is working on bridging the gap between the advances in the machine learning field and the intelligent vehicle domain. His research interests include artificial intelligence, machine learning, computer vision, reinforcement learning, intelligent transportation systems, and automated driving systems.

ARNE PETERS studied computational visualistics at the University of Koblenz from 2010 to 2016. He wrote his Master's thesis on "Detection of Unknown, Convex Objects in multi-view RGB-D Data" at KUKA robotics in Augsburg. After reaching his M.Sc. degree he joined the Chair of Robotics, Artificial Intelligence and Real-time Systems at the Technical University of Munich (TUM), where he is currently researching on computer vision for robotic application.

While his works include navigation and manipulation, motion planning, 3d reconstruction and human-robot interaction, his major focus lies on the developing self-calibrating systems, aiming to bring robots one step closer to full autonomy.

ALOIS C. KNOLL (Senior Member, IEEE) received the M.Sc. degree in electrical / communications engineering from the University of Stuttgart, Stuttgart, Germany, in 1985, and the Ph.D. degree (summa cum laude) in computer science from the Technical University of Berlin (TU Berlin), Berlin, Germany, in 1988.

He was with the Faculty of the Computer Science Department, TU Berlin, until 1993. He joined Bielefeld University, Bielefeld, Germany, as a Full Professor, where he has served as the Director for the Technical Informatics Research Group, until 2001. Since 2001, he has been a Professor with the Department of Informatics, Technical University of Munich (TUM), Munich. He was also on the Board of Directors of the Central Institute of Medical Technology, TUM (IMETUM). From 2004 to 2006, he was an Executive Director of the Institute of Computer Science, TUM. His research interests include cognitive, medical robotics, mult-iagent systems, data fusion, adaptive systems, multimedia information retrieval, model-driven development of embedded systems with applications to automotive software and electric transportation, and simulation systems for robotics and traffic. He was a member of the EU's Highest Advisory Board on Information Technology, Information Society Technology Advisory Group (ISTAG), and its subgroup on Future and Emerging Technologies (FET), from 2007 to 2009.