2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

HEGN: Hierarchical Equivariant Graph Neural Network for 9DoF Point Cloud Registration

Adam Misik[1,2], Driton Salihu[2], Xin Su[2], Heike Brock[1], and Eckehard Steinbach[2]

***Abstract*— Given its wide application in robotics, point cloud registration is a widely researched topic. Conventional methods aim to find a rotation and translation that align two point clouds in 6 degrees of freedom (DoF). However, certain tasks in robotics, such as category-level pose estimation, involve non-uniformly scaled point clouds, requiring a 9DoF transform for accurate alignment. We propose HEGN, a novel equivariant graph neural network for 9DoF point cloud registration. HEGN utilizes equivariance to rotation, translation, and scaling to estimate the transformation without relying on point correspondences. Based on graph representations for both point clouds, we extract equivariant node features aggregated in their local, cross-, and global context. In addition, we introduce a novel node pooling mechanism that leverages the cross-context importance of nodes to pool the graph representation. By repeating the feature extraction and node pooling, we obtain a graph hierarchy. Finally, we determine rotation and translation by aligning equivariant features aggregated over the graph hierarchy. To estimate scaling, we leverage scale information in the vector norm of the equivariant features. We evaluate the effectiveness of HEGN through experiments with the synthetic ModelNet40 dataset and the real-world ScanObjectNN dataset. The results show the superior performance of HEGN in 9DoF point cloud registration and its competitive performance in conventional 6DoF point cloud registration.**

I. INTRODUCTION

Point cloud registration is central to robotics by facilitating accurate perception, mapping, and localization for autonomous systems. Methods for point cloud registration have been used in robotic applications such as Simultaneous Localization and Mapping (SLAM) [1], camera localization [2], or object pose estimation [3]. While non-learning-based approaches are currently well-established in the robotics landscape [3], [4], learning-based methods are becoming increasingly popular [5], [6]. Among learning-based approaches, techniques that exploit attention mechanisms have gained significant traction due to their ability to encode contextual information and enable efficient correspondence estimation [7], [8], [9].

Conventional point cloud registration methods [10], [11], [12] aim to find a 6DoF transformation that aligns a source and target point cloud. The transformation commonly consists of a 3DoF rotation and a 3DoF translation. However, several tasks related to the robotics domain deal with non-uniformly scaled source and target point clouds.

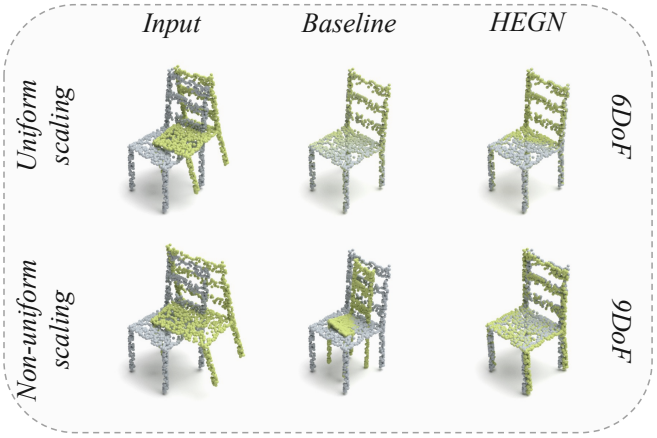


Fig. 1. Given uniform scaling, baseline methods and HEGN accurately align two point clouds. However, if both point clouds are non-uniformly scaled, baseline methods fail to align the two point clouds accurately, while HEGN performs a successful alignment.

These include (a) category-level object pose estimation problems [13], [14], [15], [16], (b) scan-to-CAD indoor reconstruction problems [17], [18], [19], [20], (c) loop closure in object SLAM problems [21], [22], [23], and (d) cross-sensor point cloud registration [24]. To achieve an accurate registration for such cases, additionally, 3DoF scaling must be estimated. This extends the search space to 9DoF for estimating the transformation between source and target point clouds.

This paper addresses the 9DoF point cloud registration problem by capitalizing on recent advances in equivariant point cloud analysis. To estimate a 9DoF transformation, we leverage global features equivariant to the geometric group of 3DoF rotation, 3DoF translation, and 3DoF scaling. First, we map source and target point clouds to their graph representation, from which we extract equivariant node features aggregated in their local, cross, and global contexts. Then, we pool the graph using a novel pooling mechanism that considers the cross-contextual importance of node features. We repeat both feature extraction and node pooling M times to obtain a graph hierarchy. Finally, we find rotation and translation by aligning equivariant global features aggregated over the graph hierarchy while we estimate scaling from the vector norms of the aggregated features. Through extensive experiments conducted in synthetic and real-world environments, we demonstrate the effectiveness of the proposed method for point cloud registration in 6DoF and 9DoF (Fig. 1).

[1] Siemens Technology, Munich, Germany
[2] Technical University of Munich, School of Computation, Information, and Technology, Chair of Media Technology, Munich Institute of Robotics and Machine Intelligence (MIRMI), Munich, Germany
This work is partially funded by Germany's BMWK under the grant UHSS4Lightweight-03LB3070B.

We list the following main contributions:

- We leverage equivariance to rotation, translation, and scaling to find a 9DoF transformation from the feature space without relying on point correspondences.
- We present HEGN, a novel hierarchical equivariant graph neural network designed to extract equivariant global features. HEGN obtains its hierarchical structure through a novel node pooling mechanism that preserves equivariance and enables registration from feature alignment.
- We showcase the effectiveness of the proposed HEGN on the synthetic ModelNet40 and real-world ScanObjectNN datasets.

The rest of the paper is organized as follows. Section II provides an overview of the related works in point cloud registration and equivariant point cloud analysis. Section III presents a formal definition of the 9DoF point cloud registration problem. We explain the modules of the proposed HEGN in Section IV, and evaluate its performance in Section V. Finally, we summarize the findings and conclusions of our method in Section VI.

II. RELATED WORK

A. Point Cloud Registration

Point cloud registration research can be categorized into two main approaches: local and global methods. Local methods are efficient when provided with an initial transformation, while global methods excel in scenarios with high outlier presence. Among the local methods, ICP is a prominent method [25]. ICP iteratively improves the registration by minimizing the distance between corresponding points in the source and target point clouds. In the group of global methods, RANSAC has been deployed in combination with feature matching schemes tackling point cloud registration [11]. Due to its outlier rejection mechanism, RANSAC is a robust algorithm that produces accurate registration results even in high-noise scenarios.

A large body of related work introduces learning-based methods for the point cloud registration task [9], [26]. IDAM proposes an iterative distance-aware similarity matrix convolution module to match points based on joint information of geometric features and an Euclidean offset [27]. As a global learning-based registration method, DeepGMR finds point-to-component correspondences with a PointNet backbone and fits Gaussian Mixture Models (GMM) to the correspondences to obtain the transformation by GMM parameter matching [12]. DeepUME builds upon the Universal Manifold Embedding and proposes a deep neural network leveraging the embedding for registration [28]. Recently, RIENet proposed a reliable inlier evaluation method for unsupervised point cloud registration, utilizing neighborhood consensus to distinguish effective inliers by capturing discriminative geometric differences between the source and pseudo target neighborhoods [29]. In contrast to the discussed methods, HEGN leverages equivariance to rotation, translation, and scaling, which allows the estimation of a 9DoF transform directly from feature space. We compare the performance of the proposed HEGN with the aforementioned approaches and show how the proposed HEGN outperforms these methods in 9DoF point cloud registration.

B. Equivariant Point Cloud Analysis

Equivariant point cloud analysis aims to leverage the inherent symmetries and transformations in point cloud data. This category of methods has shown improvements to tasks such as point cloud classification, segmentation, and completion [30], [31], [32]. Equivariant point cloud analysis has also proven beneficial for numerous robotics applications [33], [34], [35]. Vector Neurons (VN) [30] is a popular framework for designing equivariant networks. VN contain equivariance by extending neurons from 1D scalars to 3D vectors. This enables a straightforward mapping of SO(3) actions to latent spaces across common neural operations such as linear layers, nonlinearities, pooling, and normalizations. Recently, learning-based equivariant methods based on the VN framework have been applied to point cloud registration. Zhu et al. [36] propose an SO(3)-equivariant implicit learning strategy to achieve correspondence-free 3DoF point cloud registration from VN features. Lin et al. [37] follow up on this strategy and propose an SE(3)-equivariant global-local model for 6DoF point cloud registration. While we similarly leverage VN features, we extend point cloud registration to 9DoF. Moreover, the implicit learning proposed in both approaches introduces additional computational complexity, which decreases its applicability in robotics. In contrast, we propose an explicit learning strategy to address this potential issue.

III. PROBLEM FORMULATION

Given a source point cloud $X = \{x_i \in \mathbb{R}^3 | i = 1, ..., N\}$ and a target point cloud $Y = \{y_i \in \mathbb{R}^3 | i = 1, ..., N\}$, the goal of point cloud registration is to find a transformation T that minimizes the distance between transformed point cloud $T(X) = \{T(x_i) \in \mathbb{R}^3 | i = 1, ..., N\}$ and Y. When considering a registration in 9DoF, the transformed point cloud $T(X)$ is given by:

$$T(X) = SRX + t, \tag{1}$$

where $T = \{R, t, S\}$. Here, T consists of a 3×3 rotation matrix $R \in SO(3)$, translation vector $t \in \mathbb{R}^3$, and diagonal scaling matrix $S \in \mathbb{R}^3_{diag \geq 0}$. Commonly, the transformation that minimizes the distance between $T(X)$ and Y is found by solving a least-squares optimization problem:

$$\min_T \sum_{i=0}^{N} || \, T(x_i) - y_i \, ||_2^2 . \tag{2}$$

To solve Eq. 2 directly from the feature space of X and Y instead of point-to-point correspondences, the feature mapping $f(\cdot)$ needs to exhibit equivariance to the 9DoF transform $T = \{R, t, S\}$:

$$f(SRX + t) = SRf(X) + t. \tag{3}$$

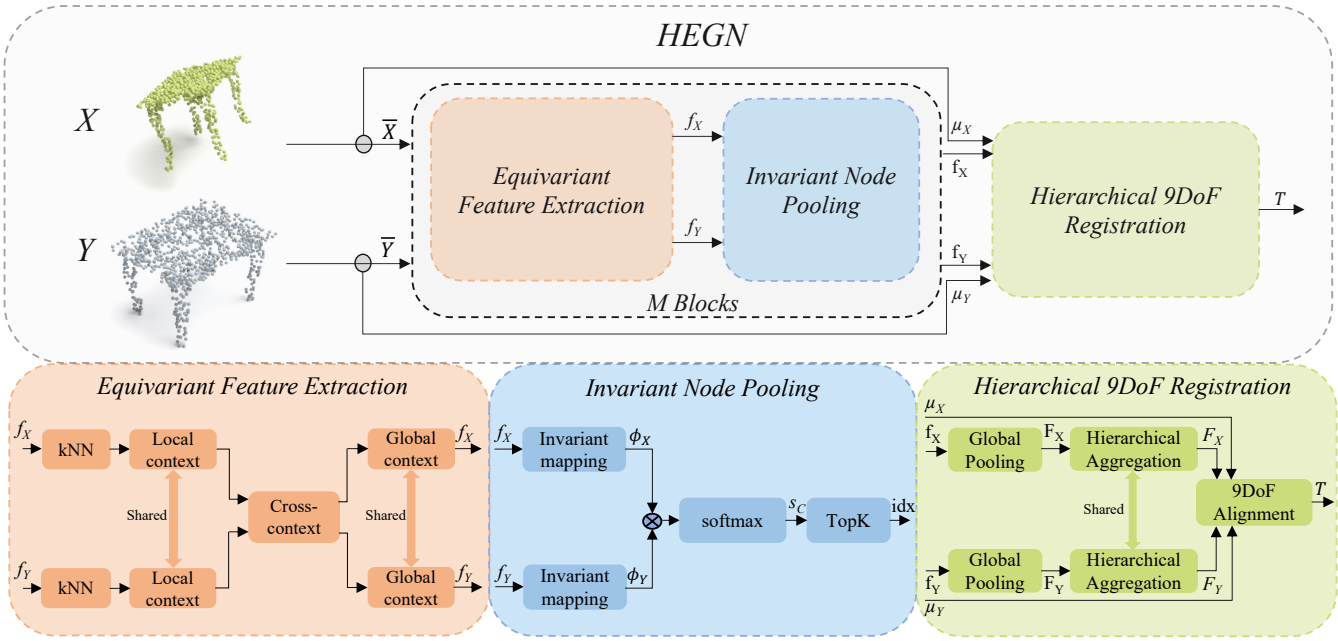


Fig. 2. Overview of the proposed HEGN. First, we achieve translation equivariance by subtracting the centroid from both point clouds. Next, we convert the point clouds into their graph representations using a k-nearest neighbor search. From these graph representations, we extract equivariant features aggregated in their local, cross, and global contexts (orange). Then, we pool the graph and its features using our proposed invariant node pooling (blue). We repeat both procedures *M* times to obtain hierarchical graph representations for both point clouds. Finally, we aggregate the hierarchical graph representations to derive a 9DoF transformation from the equivariant feature space (green).

IV. PROPOSED METHOD

A. Overview

To address the 9DoF point cloud registration problem, we use recent progress in equivariant point cloud analysis. We address 9DoF point cloud registration by aligning equivariant global features. First, we subtract the respective centroid μ from X and Y to obtain centered point clouds $\overline{X}$ and $\overline{Y}$. We preserve the centroids for the registration step, ensuring translation equivariance through the whole architecture as proposed in [37], [38]. Based on a graph representation of $\overline{X}$ and $\overline{Y}$, the encoder extracts node features equivariant to rotation and scaling (Section IV-B). After feature extraction, we pool the graph nodes in an invariant pooling operation (Section IV-C). To obtain a graph hierarchy, we repeat feature extraction and node pooling *M* times. We then map the node features to global features and use them in a hierarchical 9DoF registration (Section IV-D). The overall architecture of the proposed HEGN is displayed in Fig. 2.

B. Equivariant Feature Extraction

To construct the initial graph representation, we perform a k-nearest neighbors search for each point in $\overline{X}$ and $\overline{Y}$. We use a shared VN-DGCNN layer [30] with mean pooling over the k neighborhood to map the spatial dimension of $\overline{X}$ and $\overline{Y}$ to the feature space. This step generates the node features $f_X \in \mathbb{R}^{N\times C\times 3}$ and $f_Y \in \mathbb{R}^{N\times C\times 3}$, where N denotes the number of points and C denotes the feature dimension. Given the characteristics of VN, the extracted node features are equivariant to rotations [39]. Furthermore, scale equivariance is given through the vector nature of the features [38], [40]. After the initial graph creation and feature extraction, we update the graph dynamically after each feature extraction and node pooling block (kNN in Fig. 2).

To aggregate the local context for each node n_i, we utilize a VN-DGCNN layer over the feature f_X^i of n_i and the node features f_X^j of its k-nearest neighborhood $\mathcal{N}_i$ as:

$$f_X^i = \frac{1}{|\,\mathcal{N}_i\,|}\sum_{(i,j)\in\mathcal{N}_i} \text{VN-MLP}((f_X^j - f_X^i)\oplus f_X^i), \quad (4)$$

where VN-MLP denotes a combination of a VN-Linear and VN-LeakyReLU as proposed in [30], and $(f_X^j - f_X^i)\oplus f_X^i$ denotes a concatenation of the feature of n_i and the node features of its k-nearest neighborhood $\mathcal{N}_i$.

As shown in related work [7], [8], capturing cross-contextual information between $\overline{X}$ and $\overline{Y}$ can benefit the point cloud registration problem. To capture such context in our setting, we utilize cross-attention between f_X and f_Y using a VN-Transformer [40]. For each n_i in $\overline{X}$, we compute its query Q_X^i as:

$$Q_X^i = \text{ChNorm}(\text{VN-MLP}(f_X^i)), \quad (5)$$

where ChNorm is a vector channel normalization that excludes the scale factor, as discussed in [38], [40]. The key K_Y^j, and value V_Y^j are derived from the node features f_Y^j in the k-nearest neighborhood $\mathcal{N}_i$ of n_i in $\overline{Y}$ as:

$$K_Y^j = \text{ChNorm}(\text{VN-MLP}((f_Y^j - f_Y^i)\oplus f_Y^i)), \quad (6)$$

$$V_Y^j = \text{VN-MLP}((f_Y^j - f_Y^i)\oplus f_Y^i). \quad (7)$$

The cross-attention score a_X^j for n_i in $\overline{X}$ is then defined as follows:

$$a_X^j = \text{softmax}(\frac{Q_X^i K_Y^{j,T}}{\sqrt{3C}}). \tag{8}$$

Given the excluded scale from the vector channel normalization and rotation invariance from the query and key dot product, the obtained cross-attention score a_X^j is both rotation- and scale-invariant. This step ensures that the cross-attention stays equivariant w.r.t. to rotation and scale [40]. We then obtain the feature f_X^i for n_i aggregated with cross-contextual information as follows:

$$f_X^i = f_X^i + \sum_{(i,j)\in \mathcal{N}_i} a_X^j V_Y^j. \tag{9}$$

To improve the expressiveness of the extracted node features, we use global context aggregation. We achieve this type of aggregation by averaging f_X along the node dimension to obtain global features $F_X \in \mathbb{R}^{C\times 3}$ for $\overline{X}$. We then update f_X through global context aggregation as follows:

$$f_X = \text{VN-MLP}(f_X \oplus F_X). \tag{10}$$

We apply the same local, cross, and global context aggregation with shared layers to derive f_Y.

C. Invariant Node Pooling

To increase efficiency, methods relying on graph neural networks propose to use the farthest point sampling to reduce the graph size [38], [40]. However, such a sampling procedure is associated with high computational costs, thus decreasing applicability to time-sensitive robotics applications. Furthermore, contextual information in the graph is not considered. We propose a sampling-free graph pooling mechanism that captures contextual information to account for such deficiencies. We map the rotation and scale equivariant node features $f_X \in \mathbb{R}^{N\times C\times 3}$ to their rotation and scale invariant representation $\phi_X \in \mathbb{R}^{N\times C}$ as in [38]:

$$\phi_X = \left\langle f_X, \frac{\overline{f_X}}{\|\overline{f_X}\|} \right\rangle, \tag{11}$$

where $\overline{f_X}$ is the vector feature averaged along the dimension C. We find ϕ_Y analogous to ϕ_X. We aim to derive node-specific importance scores based on invariant node features. To ensure that cross-contextual information is considered in the node importance calculation, we correlate ϕ_X and ϕ_Y to obtain invariant importance scores $s_C \in \mathbb{R}^N$ as:

$$s_C = \text{softmax}(\phi_X \phi_Y^T). \tag{12}$$

By deriving importance scores from cross-contextual information, we ensure that the node pooling chooses nodes based on their shared importance in both point clouds. Given the rotation and scale invariance of ϕ_X and ϕ_Y, the learned node importance scores are also invariant, ensuring consistent importance of nodes under arbitrary rotations and scaling. We utilize s_C to pool the nodes in both $\overline{X}$ and $\overline{Y}$ in a TopK-ranking to find a set of K node indices idx:

$$\text{idx} = \text{rank}(s_C, K). \tag{13}$$

We use idx to select the top-ranking nodes and define the pooled graph features as $f_X^{m+1} = f_X^m(\text{idx})$ and $f_Y^{m+1} = f_Y^m(\text{idx})$, where m denotes the graph hierarchy level. Both f_X^{m+1} and f_Y^{m+1} are then used as input for the next block of feature extraction and node pooling.

D. Hierarchical 9DoF Registration

After M blocks of feature extraction and node pooling, we obtain M equivariant node features $\text{f}_\text{X} = \{f_X^1, ..., f_X^M\}$, representing the graph hierarchy. We perform mean pooling along the node dimension of the node features f_X to obtain global equivariant features $\text{F}_\text{X} = \{F_X^1, ..., F_X^M\}$. Then, we perform a hierarchical aggregation using a VN-MLP as:

$$F_X = \text{VN-MLP}(F_X^1 \oplus ... \oplus F_X^M), \tag{14}$$

to obtain a global equivariant feature F_X for $\overline{X}$. F_Y is obtained in the same way. We employ singular value decomposition between F_X and F_Y to find a rotation R that aligns $\overline{X}$ and $\overline{Y}$ without relying on point correspondences. We then use the preserved centroids to find t as $t = \mu_Y - R\mu_X$. To find S, we make use of the scale information that vector norms hold and derive S as:

$$S = \frac{S_Y}{S_X} = \frac{||\, F_Y \,||}{||\, F_X \,||}, \tag{15}$$

where S_X and S_Y denote the 3DoF scales of X and Y. Given the scale equivariance of global feature vectors, the scale of both X and Y can be efficiently extracted.

E. Loss Function

To train HEGN, we introduce a two-part loss term. The first component is the registration loss L_R, which minimizes the mean squared error between the predicted (p) and ground truth (g) 9DoF transformation elements, defined as:

$$L_R = ||\, R_g^T R_p - \mathbb{1}_{3\times 3} \,||_2^2 + ||\, t_g - t_p \,||_2^2 + ||\, S_g - S_p \,||_2^2, \tag{16}$$

where $\mathbb{1}_{3\times 3}$ denotes a 3×3 identity matrix. We also use the Chamfer Distance (CD) [41] to measure the dissimilarity between the points of $T(X)$ and Y:

$$L_{CD} = CD((T(X), Y). \tag{17}$$

The final training loss is defined as $L = L_R + L_{CD}$.

V. EXPERIMENTAL EVALUATION

A. Implementation Details

We use $M = 4$ blocks, with $k = 20$ for the k nearest neighbor search in the first two blocks and $k = 16$ for the last two blocks. For the TopK pooling, we set $K = \frac{N}{4}$ for the first two blocks, and $K = \frac{N}{2}$ for the last two blocks. The dimension of the global descriptors used for 9DoF alignment is set to 512×3. We train HEGN for 100 epochs and a batch size of 32. For optimization, we choose the ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. We set the learning rate to 10^{-3}, and use a cosine annealing scheduler. We train on two NVIDIA GeForce GTX 1080 Ti GPUs. In our experiments, we compare HEGN to non-learning (ICP, FPFH+RANSAC) and learning-based methods (IDAM, DeepGMR, DeepUME,

RIENet). To estimate S for these methods and enable a comparison to HEGN, we fit 3D-oriented bounding boxes O_X and O_Y to X and Y. We then obtain S by dividing over the eight bounding box corners of Y and X, similarly as in [18]:

$$S = \frac{1}{8} \sum_{i=1}^{8} \frac{\mid O_Y^i \mid}{\mid O_X^i \mid}, \tag{18}$$

We additionally assess HEGN using this scale estimation method to thoroughly compare with other methods. By performing global feature alignment, we determine the rotation and translation. Subsequently, we apply the obtained 6DoF transformation to transform X and calculate the scale according to Eq. 18. For a fair comparison, we exclude the scale error term from L_R in this scale estimation approach. We denote this version of HEGN as HEGN (OBB), while HEGN (VN) denotes the version utilizing vector norms.

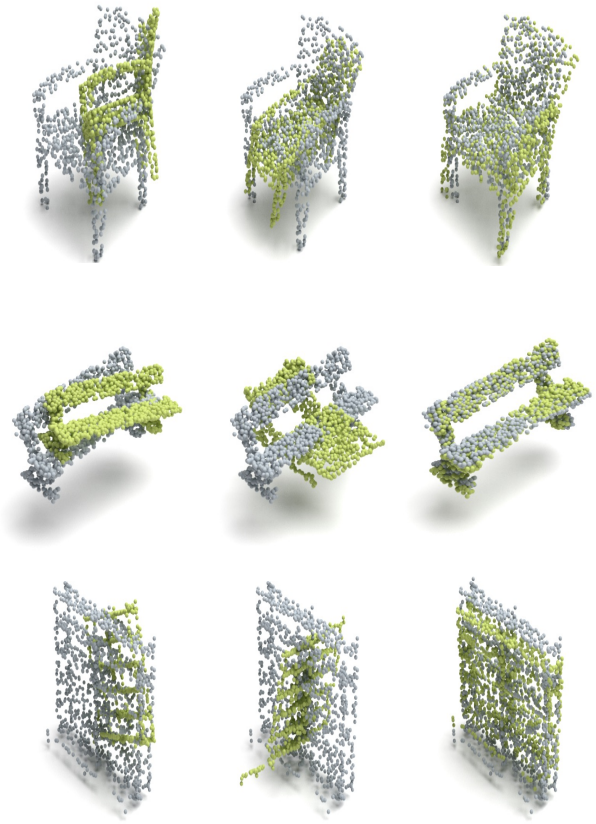
B. Synthetic Data

In our first set of experiments, we evaluate HEGN using the synthetic ModelNet40 dataset [42]. We use the processed data from [43] and downsample all point clouds to 1024 points. We use 5,112 samples for training, 1,202 samples for validation, and 1,266 for testing, as in [44]. We evaluate both 6DoF and 9DoF point cloud registration. To generate the relative poses between the source and target point clouds, we randomly sample elements of the rotation matrix R from the range of [0°, 180°], elements of the scaling S from the range of [0.5, 1.5], and the translation t from the range of [-0.5, 0.5]. For the 6DoF point cloud registration experiments, we transform X by the generated R and t. For the 9DoF experiments, we additionally scale X by S. To further simulate real-world conditions, we jitter both point clouds with Gaussian noise sampled from $\mathcal{N}(0, 0.01)$, clipped to a range [-0.05, 0.05] [12]. Furthermore, we test on unseen model categories to test the generalization capabilities of learning-based methods. We rely on global metrics to evaluate the registration performance and compute the root mean square error (RMSE) and CD between $T(X)$ and Y. We also assess runtime performance by calculating the average number of registered point clouds per second (FPS).

Table I lists the quantitative results for 6DoF and 9DoF point cloud registration. HEGN achieves competitive performance with other methods in the conventional 6DoF setting. In the 9DoF setting, HEGN outperforms both local and global methods regarding CD and RMSE. These results can be linked to HEGN's ability to incorporate 9DoF information into the global features utilized for feature alignment and scale estimation. When comparing HEGN (VN) and HEGN (OBB), we observe that using vector norms of global features outperforms the fitting of bounding boxes. We attribute this result to the differentiability of the vector norm scale estimation approach. Regarding runtime performance, HEGN has a lower FPS than other learning-based methods. However, we argue that the achieved runtime meets the requirements for applications such as SLAM and pose estimation. Moreover, in the 9DoF setting, the scale estimation does not significantly decrease FPS compared to the 6DoF setting, as is the case for other methods. Interestingly, the FPS of RANSAC increases for the 9DoF setting, which we explain by RANSAC's inability to find inliers, breaking the iterative refinement. Fig. 3 visualizes qualitative results for the 9DoF point cloud registration. We compare HEGN to DeepGMR. The figure demonstrates that additional scaling introduces a distorted geometry of X, resulting in unsuccessful registrations for DeepGMR. However, by leveraging rotation and scale equivariance, HEGN achieves an accurate 9DOF transformation, precisely aligning the source and target point clouds.

C. Real-world Data

To further demonstrate the effectiveness of HEGN, we evaluate its performance on a real-world benchmark. For this, we utilize the ScanObjectNN dataset [45]. The ScanObjectNN dataset consists of object-level indoor scans and presents challenges such as additional background clutter, uneven point distributions, and occlusions. We follow [46], [47] and test the synthetic-to-real generalization capabilities of learning-based methods using ModelNet40 pretraining. We resample all scans to 1024 points and generate source



(a) Input (b) DeepGMR (c) HEGN

Fig. 3. Qualitative results for 9DoF point cloud registration on ModelNet40. Green illustrates the source point cloud X, while grey illustrates the target point cloud Y.

TABLE I

RESULTS FOR 6DOF AND 9DOF POINT CLOUD REGISTRATION ON THE MODELNET40 DATASET. CD IS SCALED BY 10^2.

Setting	Method	CD (↓)	RMSE (↓)	FPS (↑)
6DoF	ICP	22.81	0.89	**128.49**
	FPFH+RANSAC	0.43	0.14	8.60
	IDAM	0.81	0.07	51.70
	DeepGMR	0.10	0.02	95.44
	DeepUME	**0.02**	**0.01**	52.21
	RIENet	6.05	0.49	28.06
	HEGN	0.09	0.02	37.11
9DoF	ICP	75.36	1.09	**93.49**
	FPFH+RANSAC	69.59	0.76	9.80
	IDAM	36.18	0.82	46.08
	DeepGMR	24.06	0.39	64.41
	DeepUME	19.25	0.59	47.32
	RIENet	38.01	0.79	26.55
	HEGN (OBB)	15.34	0.34	33.03
	HEGN (VN)	**2.88**	**0.18**	35.73

TABLE II

RESULTS FOR 6DOF AND 9DOF POINT CLOUD REGISTRATION ON THE REAL-WORLD SCANOBJECTNN DATASET. CD IS SCALED BY 10^2.

Setting	Method	CD (↓)	RMSE (↓)	FPS (↑)
6DoF	ICP	22.36	0.87	**123.50**
	FPFH+RANSAC	0.13	0.01	2.77
	IDAM	**0.00**	**0.00**	52.44
	DeepGMR	**0.00**	**0.00**	100.20
	DeepUME	**0.00**	**0.00**	53.42
	RIENet	2.73	0.20	28.99
	HEGN	0.01	0.01	36.77
9DoF	ICP	134.63	1.24	**88.22**
	FPFH+RANSAC	92.20	0.68	9.67
	IDAM	72.04	0.82	46.29
	DeepGMR	43.26	0.45	66.30
	DeepUME	42.29	0.59	48.43
	RIENet	81.88	0.73	27.04
	HEGN (OBB)	57.39	0.44	32.15
	HEGN (VN)	**3.62**	**0.19**	37.46

and target point clouds for ModelNet40. We employ the same evaluation metrics as the synthetic benchmark (CD, RMSE, FPS).

Table II summarizes the quantitative results for the real-world experiments. In the 6DoF setting, most methods generalize well and yield highly accurate results. Similarly, as in the synthetic benchmark, HEGN significantly outperforms other methods in terms of CD and RMSE in the 9DoF setting. These results indicate that HEGN (a) performs well in real-world conditions, and (b) generalizes well to real-world conditions when trained on synthetic data. Furthermore, we conclude that registration based on global feature alignment is robust in the presence of noise common to real-world scanning scenarios. Fig. 4 illustrates qualitative results for 9DoF point cloud registration for a table scan. Given an uneven point distribution and sensor noise, HEGE finds an accurate transformation that aligns X and Y.

D. Ablation Study

To evaluate the impact of each module on the overall performance of HEGN, we perform an ablation study with four settings. We compare the performance when omitting (a) the local context aggregation, (b) the cross-context aggregation, (c) the global context, and (d) the hierarchical feature aggregation. In the last setting, we perform registration using the global features extracted after the last block. We retrain HEGN for all settings on ModelNet40 for the 9DoF registration with vector norm-based scale estimation. We obtain similar performances for the first two settings. These results highlight the importance of leveraging local and cross-contextual information in feature extraction for point cloud registration problems, as highlighted in [7]. Albeit higher FPS, the second setting yields lower registration accuracy than the original setting, emphasizing cross-context aggregation despite increased computational costs.

TABLE III

ABLATION STUDY OF HEGN FOR 9DOF POINT CLOUD REGISTRATION ON THE MODELNET40 DATASET. CD IS SCALED BY 10^2.

Setting	CD (↓)	RMSE (↓)	FPS (↑)
w/o local context	4.30	0.22	38.75
w/o cross-context	4.05	0.21	**61.89**
w/o global context	3.28	0.19	37.50
w/o hierarchical aggregation	3.06	0.19	37.74
HEGN (VN)	**2.88**	**0.18**	35.73

(a) Input (b) DeepGMR (c) HEGN

Fig. 4. Qualitative results for 9DoF point cloud registration on ScanObjectNN. Green illustrates the source point cloud X, while grey illustrates the target point cloud Y.

VI. CONCLUSION

This paper presents HEGN, a novel equivariant graph neural network for 9DoF point cloud registration. By leveraging equivariance to rotations, translations, and scaling, HEGN effectively estimates a transformation that facilitates the alignment of point clouds in 9DoF. We demonstrate the efficacy of HEGN through comprehensive experiments for 9DoF point cloud registration on both synthetic and real-world benchmarks. We achieve competitive results in the conventional 6DoF setting, highlighting its versatility and robustness. Furthermore, HEGN generalizes well to real-world settings when trained on synthetic data. In the future, we anticipate that HEGN will find practical applications in various robotics tasks, such as pose estimation, indoor 3D reconstruction, and SLAM.

REFERENCES

[1] T. Hitchcox and J. R. Forbes, "A point cloud registration pipeline using gaussian process regression for bathymetric slam," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 4615–4622.

[2] W. Junyi and Q. Yue, "Camera relocalization using deep point cloud generation and hand-crafted feature refinement," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5891–5897.

[3] M. Gentner, P. K. Murali, and M. Kaboli, "GMCR: graph-based maximum consensus estimation for point cloud registration," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 4967–4974.

[4] I. Vizzo, T. Guadagnino, B. Mersch, L. Wiesmann, J. Behley, and C. 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*, vol. 8, no. 2, pp. 1029–1036, 2023.

[5] L. Wiesmann, T. Guadagnino, I. Vizzo, G. Grisetti, J. Behley, and C. Stachniss, "Dcpcr: Deep compressed point cloud registration in large-scale outdoor environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6327–6334, 2022.

[6] D. Cattaneo, M. Vaghi, and A. Valada, "Lcdnet: Deep loop closure detection and point cloud registration for lidar slam," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2074–2093, 2022.

[7] 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 (CVPR)*, 2021, pp. 4267–4276.

[8] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, S. Ilic, D. Hu, and K. Xu, "Geotransformer: Fast and robust point cloud registration with geometric transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[9] G. Chen, M. Wang, Q. Zhang, L. Yuan, T. Liu, and Y. Yue, "Deep interactive full transformer framework for point cloud registration," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 2825–2832.

[10] Q.-Y. Zhou, J. Park, and V. Koltun, "Fast global registration," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 766–782.

[11] H. Lei, G. Jiang, and L. Quan, "Fast descriptors and correspondence propagation for robust global point cloud registration," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3614–3623, 2017.

[12] W. Yuan, B. Eckart, K. Kim, V. Jampani, D. Fox, and J. Kautz, "Deepgmr: Learning latent gaussian mixture models for registration," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 733–750.

[13] X. Liu, G. Wang, Y. Li, and X. Ji, "Catre: Iterative point clouds alignment for category-level object pose refinement," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 499–516.

[14] Y. Weng, H. Wang, Q. Zhou, Y. Qin, Y. Duan, Q. Fan, B. Chen, H. Su, and L. J. Guibas, "Captra: Category-level pose tracking for rigid and articulated objects from point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 13 209–13 218.

[15] Q. Feng and N. Atanasov, "Fully convolutional geometric features for category-level object alignment," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8492–8498.

[16] X. Deng, J. Geng, T. Bretl, Y. Xiang, and D. Fox, "icaps: Iterative category-level object pose and shape estimation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1784–1791, 2022.

[17] A. Avetisyan, A. Dai, and M. Nießner, "End-to-end cad model retrieval and 9dof alignment in 3d scans," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2551–2560.

[18] D. Salihu and E. Steinbach, "Sgpcr: Spherical gaussian point cloud representation and its application to object registration and retrieval," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 572–581.

[19] A. Misik, D. Salihu, H. Brock, and E. Steinbach, "Cocca: Point cloud completion through cad cross-attention," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 580–584.

[20] D. Salihu, A. Misik, M. Hofbauer, and E. Steinbach, "S2cmaf: Multi-method assessment fusion for scan-to-cad methods," in *2022 IEEE International Symposium on Multimedia (ISM)*. IEEE Computer Society, 2022, pp. 129–136.

[21] J. McCormac, R. Clark, M. Bloesch, A. Davison, and S. Leutenegger, "Fusion++: Volumetric object-level slam," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 32–41.

[22] S. Lin, J. Wang, M. Xu, H. Zhao, and Z. Chen, "Topology aware object-level semantic mapping towards more robust loop closure," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7041–7048, 2021.

[23] Z. Liao, Y. Hu, J. Zhang, X. Qi, X. Zhang, and W. Wang, "So-slam: Semantic object slam with scale proportional and symmetrical texture constraints," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4008–4015, 2022.

[24] X. Huang, J. Zhang, Q. Wu, L. Fan, and C. Yuan, "A coarse-to-fine algorithm for matching and registration in 3d cross-source point clouds," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2965–2977, 2018.

[25] S. T. S. H. S. D. B. Arun, "Least-square fitting of two 3-D point sets," *IEEE Pattern Analysis and Machine Intelligence*, vol. 9, 1987.

[26] P. Yin, S. Yuan, H. Cao, X. Ji, S. Zhang, and L. Xie, "Segregator: Global point cloud registration with semantic and geometric cues," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 2848–2854.

[27] J. Li, C. Zhang, Z. Xu, H. Zhou, and C. Zhang, "Iterative distance-aware similarity matrix convolution with mutual-supervised point elimination for efficient point cloud registration," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 378–394.

[28] N. Lang and J. M. Francos, "Deepume: Learning the universal manifold embedding for robust point cloud registration," in *British Machine Vision Conference (BMVC)*, 2021.

[29] Y. Shen, L. Hui, H. Jiang, J. Xie, and J. Yang, "Reliable inlier evaluation for unsupervised point cloud registration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2198–2206.

[30] C. Deng, O. Litany, Y. Duan, A. Poulenard, A. Tagliasacchi, and L. J. Guibas, "Vector neurons: A general framework for so (3)-equivariant networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12 200–12 209.

[31] H. Wu and Y. Miao, "So (3) rotation equivariant point cloud completion using attention-based vector neurons," in *2022 International Conference on 3D Vision (3DV)*, 2022, pp. 280–290.

[32] H. Chen, S. Liu, W. Chen, H. Li, and R. Hill, "Equivariant point network for 3d point cloud analysis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021, pp. 14 514–14 523.

[33] A. Simeonov, Y. Du, A. Tagliasacchi, J. B. Tenenbaum, A. Rodriguez, P. Agrawal, and V. Sitzmann, "Neural descriptor fields: Se (3)-equivariant object representations for manipulation," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 6394–6400.

[34] D. Wang, R. Walters, X. Zhu, and R. Platt, "Equivariant q learning in spatial action spaces," in *Conference on Robot Learning (CoRL)*, 2022, pp. 1713–1723.

[35] H. Huang, D. Wang, X. Zhu, R. Walters, and R. Platt, "Edge grasp network: A graph-based se (3)-invariant approach to grasp detection," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3882–3888.

[36] M. Zhu, M. Ghaffari, and H. Peng, "Correspondence-free point cloud registration with so (3)-equivariant implicit shape representations," in *Conference on Robot Learning (CoRL)*, 2022, pp. 1412–1422.

[37] C.-W. Lin, T.-I. Chen, H.-Y. Lee, W.-C. Chen, and W. H. Hsu, "Coarse-to-fine point cloud registration with se (3)-equivariant representations," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 2833–2840.

[38] Y. Chen, B. Fernando, H. Bilen, M. Nießner, and E. Gavves, "3d equivariant graph implicit functions," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 485–502.

[39] 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*, vol. 38, no. 5, pp. 1–12, 2019.

[40] J. Lei, C. Deng, K. Schmeckpeper, L. Guibas, and K. Daniilidis, "Efem: Equivariant neural field expectation maximization for 3d object segmentation without scene supervision," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4902–4912.

[41] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, "Parametric correspondence and chamfer matching: Two new techniques for image matching," in *Proceedings: Image Understanding Workshop*. Science Applications, Inc, 1977, pp. 21–27.

[42] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1912–1920.

[43] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 652–660.

[44] 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 (CVPR)*, 2020, pp. 11 824–11 833.

[45] M. A. Uy, Q.-H. Pham, B.-S. Hua, D. T. Nguyen, and S.-K. Yeung, "Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[46] T. Min, C. Song, E. Kim, and I. Shim, "Distinctiveness oriented positional equilibrium for point cloud registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 5490–5498.

[47] D. Bauer, T. Patten, and M. Vincze, "Reagent: Point cloud registration using imitation and reinforcement learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 586–14 594.