Boosting Fine-grained Feature Fusion in 3D Point Cloud Registration

Huaiyuan Yu, Haijiang Zhu, Jian Cheng, and Ning An

***Abstract*—Existing point cloud registration methods have achieved significant progress through transformer architecture. However, these methods often overlook the fine-grained structural information in local features, which limits their performance in complex scenes. To address this issue, we propose a fine-grained module that enhances element-wise feature interaction. This approach provides finer-grained feature information and improves the accuracy of point cloud registration. First, a multi-scale hierarchical feature fusion module is designed to capture fine-grained feature. Second, this module is integrated into the REGTR (Registration Transformer) backbone to enhance feature correlation. Furthermore, we propose an accelerated registration strategy that balances efficiency and accuracy by enhancing the contribution of high-probability overlapping features. Comprehensive experiments on indoor and outdoor benchmarks demonstrate the effectiveness of our method. Compared to the REGTR baseline, our method achieves relative error reductions of 17.6% and 8.9% on 3DMatch and ModelNet40, respectively, while maintaining competitive computational efficiency. Furthermore, consistent performance improvements on the MCD (Multi-Campus Dataset) further validate the robustness of our method across diverse scenes.**

Link to graphical and video abstracts, and to code:
https://latamt.ieeer9.org/index.php/transactions/article/view/10157

***Index Terms*—3D point cloud, Point cloud registration, Granular feature.**

I. Introduction

With the widespread application of point clouds in fields such as 3D reconstruction [1], simultaneous localization and mapping (SLAM) [2], and robot navigation, point cloud registration has become a crucial technology [3]. The objective of point cloud registration is to estimate the spatial transformation between two point clouds. This is typically achieved by solving for the rigid transformation matrix based on correspondences between the point clouds.

In recent years, numerous learning-based registration methods have emerged, along with comprehensive review articles. Huang [4] provided a comprehensive overview of point cloud registration, including traditional optimization and deep learning methods. Zhang [5] presented a detailed survey and taxonomy, classifying methods into supervised and unsupervised categories. Compared to supervised approaches, unsupervised registration is based on the geometric properties of point clouds rather than external labels. Lyu [6] introduced the IBTaxon classification, which categorizes the registration methods into one-stage and two-stage approaches. The authors emphasized that balancing accuracy, speed, and robustness is more important than optimizing any single metric.

Mainstream point cloud registration methods include ICP (Iterative Closest Point) [7], feature-based methods [8], learning-based methods [9], and statistical distribution models [10], [11]. Most of these approaches follow a two-stage pipeline: first, estimating accurate correspondences, and then optimizing the registration result.

Recent work such as DGR (Deep Global Registration) [12] proposed a 6-dimensional convolutional network for correspondence estimation and a robust gradient-based optimizer for registration refinement. However, DGR struggles with incomplete or partially overlapping point clouds and suffers from long computation times due to its refinement process and RANSAC (Random Sample Consensus) safeguard. To further improve registration accuracy and robustness, Yew proposed REGTR (Registration Transformer) [13], an end-to-end network based on transformer architecture. REGTR predicts the correspondences of feature points in the overlapping region and estimates the transformation result without post-processing.

In this work, we improve REGTR in both feature extraction and transformation estimation. First, we design a feature fusion module to capture fine-grained features through element-wise interaction. This module is integrated into the feature extraction backbone of REGTR to enhance feature correlation. Second, we propose an accelerated registration strategy to balance registration accuracy and efficiency. We conducted experiments on indoor and outdoor scenes to evaluate the effectiveness of the proposed optimization scheme in improving registration accuracy. For example, Fig. 1 demonstrates the performance of our method in point cloud registration on the 3DMatch [14] and MCD (Multi-Campus Dataset) [15] datasets. Specifically, the first row of Fig. 1 presents an indoor kitchen scene from 3DMatch, while the second row shows an outdoor scene from MCD. The experimental results on indoor and outdoor datasets confirm the effectiveness of our method.

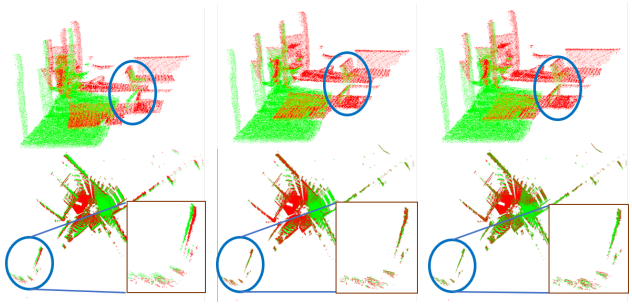
II. Related Work

Traditional point cloud registration typically estimates the transformation between the source and target point clouds

The associate editor coordinating the review of this manuscript and approving it for publication was Giner Alor-Hernández (*Corresponding author: Huaiyuan Yu*).

Huaiyuan Yu, and H. Zhu are with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China (e-mails: 2024400285@buct.edu.cn, and zhuhj@mail.buct.edu.cn).

J. Cheng, and N. An are with the Research Institute of Mine Artificial Intelligence and the State Key Laboratory of Intelligent Coal Mining and Strata Control, Chinese Institute of Coal Science, Beijing, China (e-mails: jiancheng@tsinghua.org.cn, and ning.an.010@foxmail.com).



(a) REGTR (b) Ours (c) Ground truth

Fig. 1. Registration results on the indoor and outdoor scenes. The point clouds in the first row are from 3DMatch, and those in the second row are from MCD. (a) REGTR, (b) Ours, (c) Ground truth.

through iterative optimization [16]. In this strategy, the transformation matrix between two point clouds is iteratively refined until specific termination criteria are satisfied. However, iterative optimization-based registration methods are sensitive to noise and initial values [17].

In recent years, deep learning has significantly improved the robustness of point cloud registration [18]. Many studies leverage deep learning to extract learning-based features, such as FCGF (Fully Convolutional Geometric Features) [19], KPConv (Kernel Point Convolution) [20], PointNet [21], and DIP (Distinctive 3D Local Deep Descriptors) [22]. Furthermore, CoFF [23] (Cross-modal Feature Fusion) takes images as auxiliary input to enhance the learned 3D features. These learning-based features are then used to establish correspondences based on similarity, which constrain the point cloud registration. However, due to the lack of perception of spatial positions, the accuracy of feature-based registration is easily affected by 3D points outside the overlapping region when registering partially overlapping point clouds.

In addition, end-to-end point cloud registration has also gained attention [24]. Yang [25] and Huang [26] reformulate the registration as optimizing feature descriptor distances instead of 3D reprojection errors. CoFiNet [27] (Coarse-to-Fine Network) extracts correspondences from coarse to fine granularity without the need for salient point detection. GeoTransformer [28] (Geometric Transformer) learns rigid transformation-invariant geometric features for robust superpoint matching. GPI-Net [29] (Gestalt-guided Parallel Interaction Network) leverages Gestalt principles to facilitate complementary interaction between local and global information. BUFFER [30] combines point-wise and patch-wise feature extractors to balance registration accuracy and efficiency. PointDifformer [31] (Point Cloud Diffusion Transformer) enhances feature representation by leveraging Graph Neural Partial Differential Equations and exhibits robustness to noise in point cloud registration. REGTR combines correspondence estimation and overlap prediction into an end-to-end registration model, allowing information interaction between tasks during training [32]. The limitation of end-to-end approaches lies in their substantial demand for training data and computational resources.

REGTR achieves promising performance in registering partially overlapping point clouds. To address the lack of global information in incomplete point clouds, REGTR uses KPConv to extract features and leverages transformer layers to enhance feature representation. The KPConv achieves 3D feature extraction by 3D convolution kernels, which effectively capture geometric structure and salient points. Similarly, D3Feat [33] uses KPConv and further predicts confidence scores for matching features, which help select high-confidence correspondences for registration.

However, KPConv faces challenges when processing incomplete point clouds with vacant areas. For these incomplete point clouds, where information loss occurs due to object occlusion [34], improving the information representation of features is critical for high-precision point cloud registration. This requires the feature extraction module to capture fine-grained features.

Taking inspiration from Res2Net [35], which has been widely applied in 2D vision tasks, we propose an element-wise feature fusion module to enhance feature representation. This module is integrated into the REGTR backbone network to facilitate the prediction of correspondences and overlap, which is critical for improving point cloud registration accuracy. Finally, we develop an efficient registration strategy to achieve the balance between accuracy and computational efficiency.

III. Methods

Using REGTR as the base framework, we construct a fine-grained 3D feature representation and design a fine-grained multi-scale module to enhance the network's feature learning capacity. In addition, we propose an accelerated registration strategy to improve computational efficiency. Fig. 2 shows the pipeline of our method for point cloud registration.

A. Fine-grained 3D Point Feature Descriptor

REGTR uses 3D kernel point convolution (KPConv) [20] for feature extraction. For incomplete point clouds, the network needs to extract features that have a better ability to represent information, thereby reducing the interference caused by vacant areas. To enhance the representation of information in incomplete point clouds, we improve the granularity of features without increasing network depth. This approach prevents the potential reduction in the number of extracted features caused by deepening the network, which could compromise the accuracy of point cloud registration.

In this work, consistent with REGTR, we employ the rigid-kernel variant of KPConv to extract features. Fig. 3 uses 2D points to illustrate the feature extraction mechanism of kernel point convolution (KPConv). In Fig. 3, the gray dots represent the initial X feature points, and each feature point contains a feature vector of dimension D_{in}. The black dots are the 2D point convolution kernels used by KPConv, composed of K kernel points. Each kernel point is associated with a convolution matrix W_k, which transforms the feature vector from D_{in} to D_{out} dimensions. After the initial features pass through the KPConv, N new features are obtained, represented

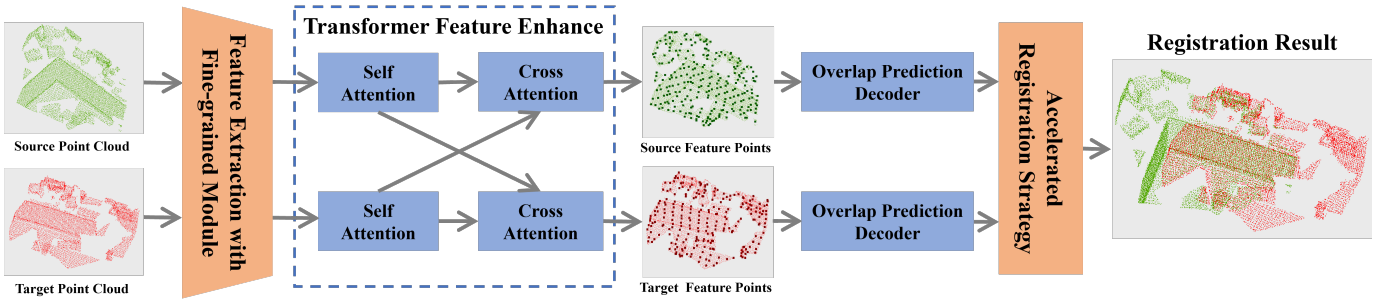


Fig. 2. The overall pipeline of our method for point cloud registration. The orange blocks represent the major optimizations made in this paper compared to the REGTR baseline.

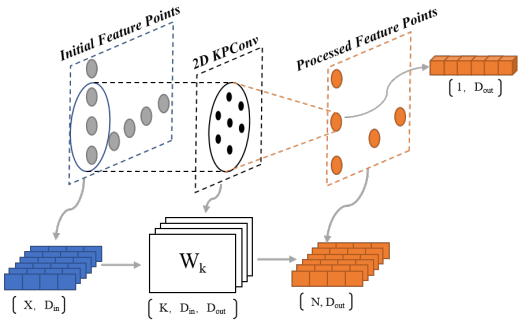


Fig. 3. The feature extraction mechanism of kernel point convolution (KPConv).

by orange dots. Each orange dot contains a feature vector of dimension D_{out}.

Inspired by Res2Net, we draw on its concept to extract more fine-grained features through element-wise information fusion. In Res2Net, fine-grained features are obtained by performing multi-channel information fusion on the 2D feature maps. For a salient point in the image, its corresponding feature vector is composed of the values at the same point location across the multi-channel feature maps. The multi-scale channel processing of Res2Net can be equivalent to the information fusion at the element level of a feature vector. Based on this, we implement element-wise information fusion in the features extracted by KPConv, which enhances the feature representation capability. Subsequently, we apply this strategy to 3D point clouds and construct a fine-grained multi-scale module.

B. Fine-grained Multi-scale Module

The fine-grained multi-scale module consists of the linear convolution layer and the normalization layer, and is used to perform element-wise information fusion for each individual feature vector. The module takes N feature vectors as input, where N is the number of 3D feature points, and the output of this module is fine-grained feature vectors obtained by element-wise interaction. Fig. 4 illustrates the structure of this module for 3D point clouds. The input feature vector of 3D salient point is divided into ***scale*** sub-vectors of equal length in the dimension. The sub-vectors are denoted as d_i (i = 1, 2,..., ***scale***).

Except for d_1, each d_i has a corresponding linear convolution layer $L_i(d_i)$ (i = 2,..., ***scale***). The output of the linear convolution layer L_i is denoted as $d_i^{'}$, while d_1 directly serves as $d_1^{'}$. For d_2, it is processed by L_2 to generate $d_2^{'}$. When 2 $< i <=$ ***scale***, d_i is added with $d_{i-1}^{'}$ and the fused result is processed by L_i. Each $d_i^{'}$ is then concatenated to form $d^{'}$, which is further processed by a linear convolution layer to generate the output feature. The formulation of $d_i^{'}$ can be written as (1).

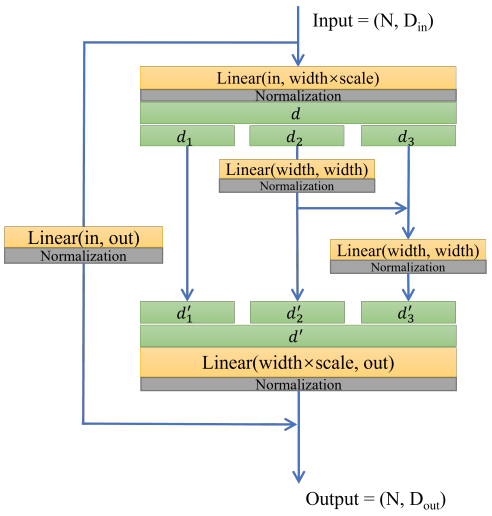


Fig. 4. Fine-grained multi-scale module for 3D feature points. Taking ***scale***=3 as an example, the input feature vector is divided into ***scale*** equal-length sub-vectors, which then undergo element-wise information interaction and fusion. The green block represents a sub-vector, the orange block represents a linear convolution layer, and the gray block corresponds to a normalization layer. The values in parentheses indicate the input and output feature vector dimensions for each layer.

$$d_i^{'} = \begin{cases} d_i & (i = 1) \\ L_i(d_i) & (i = 2) \\ L_i(d_i + d_{i-1}^{'}) & (2 < i <= scale) \end{cases} \tag{1}$$

In REGTR, the feature extraction backbone consists of several ResNet Blocks that incorporate KPConv and dimension-altering convolution layer. In this paper, the fine-grained multi-scale module is integrated into the REGTR backbone to construct the optimized feature extraction network, as shown in Fig. 2. Specifically, as depicted in Fig. 5, the proposed

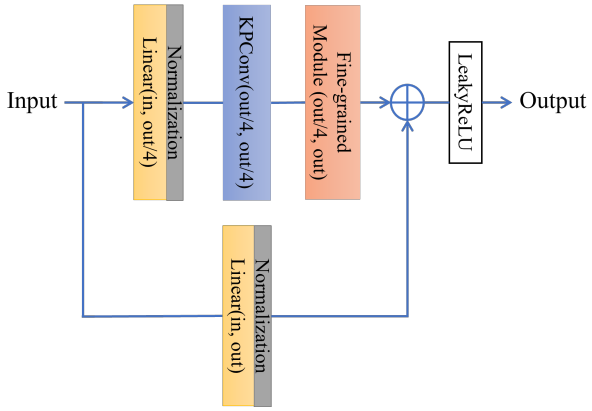


Fig. 5. ResNet Block architecture integrating KPConv with the proposed fine-grained multi-scale module. The values in parentheses within each block indicate the input and output feature vector dimensions for each layer.

module is inserted directly after the KPConv module within the ResNet Block structures, while other dimension-altering convolution layers are removed. This design allows the fine-grained module to simultaneously process feature vectors and perform dimensional expansion. In the shortcut path, the input feature vector passes through a linear convolution layer and a normalization layer to adjust its dimension, so that it can be added and fused with the output of the fine-grained module. The fused result is then processed by a LeakyReLU layer to generate the final output.

C. Accelerated Registration Strategy

To balance computational efficiency and registration accuracy in point cloud registration, we propose an accelerated registration strategy. In REGTR, after the backbone network extracts point cloud features, a multi-layer transformer based on self-attention and cross-attention mechanisms enhances these features. The decoder then uses the enhanced features to predict their correspondences, as well as the overlapping region between the source and target point clouds, along with the probability of each feature point being within this region. Subsequently, REGTR filters out interference from feature points outside the overlapping region using this overlap probability and only uses the points within the overlapping region for registration.

Building on this foundation, we follow the REGTR's overlap prediction and further design a threshold-based filtering to accelerate registration, where high-confidence overlapping points are preserved based on the filtering threshold. The core idea is to assign higher weight coefficients to feature points with greater overlap probabilities during registration while filtering out less relevant feature points. Using these weight coefficients, a weighted Kabsch algorithm is applied to estimate the transformation between the source and target point clouds. By prioritizing high-confidence overlapping points, this approach achieves a balance between efficiency and accuracy. The main steps of the accelerated registration strategy are summarized as follows:

- Calculating the probability of each feature point being within the overlapping region.
- Sorting the 3D feature points in descending order of their overlap probability.
- Setting a filtering threshold to partition the sorted feature points into two groups: feature points with overlap probabilities exceeding the threshold retain their probability values as weight coefficients, while those below the threshold have their weight coefficients set to 0.
- Using the feature points with non-zero weight coefficients to perform a weighted Kabsch algorithm for estimating the transformation between the source and target point clouds.

IV. Experiments

A. Dataset and Experiment Details

This work evaluates the proposed method on indoor and outdoor scenes. 3DMatch [14], 3DLoMatch and ModelNet40 [36] are used as indoor datasets. The 3DMatch and 3DLoMatch include incomplete point clouds of various indoor scenes, such as living room and kitchen, while the ModelNet40 contains incomplete point clouds of various objects, such as vase and table. MCD [15] is used as an outdoor scene dataset, consisting of three large-scale maps from different campuses and providing semantic labels for the point clouds. Compared to indoor scenes, the point clouds from the MCD outdoor scene are relatively sparse.

The hardware environment used in the experiments includes an NVIDIA RTX A6000 GPU with 48 GB and an Intel Core i9-12900K. The software environment includes Ubuntu 20.04.6 LTS, Python 3.9, Pytorch 1.12.1, and CUDA 11.6. The training parameters are set as follows: AdamW optimizer, initial learning rate 0.0001, weight decay 0.0001, gradient clip 0.1, no pretrained weights. For the 3DMatch dataset, we train for 40 epochs with a batch size of 2. For the 3DLoMatch dataset, we directly use the model weights trained on 3DMatch without additional training. For the ModelNet40 dataset, we train for 400 epochs with a batch size of 4. For the MCD dataset, we employ a batch size of 1 and train for 80 epochs.

B. Evaluation Metrics

In this paper, the absolute error between the estimated pose and the ground truth is used as a metric to evaluate the accuracy of point cloud registration. The absolute pose error (APE) is defined by (2).

$$APE = \left\| log(T_{gt,i}^{-1} T_{est,i})^{\vee} \right\|_2^2 \tag{2}$$

Let the rotation matrix R and the translation vector t be T=[R, t| 0, 1]. The $T_{gt,i}$ and $T_{est,i}$ represent the ground truth and the estimation transformation, respectively. The symbol $\vee$ denotes a transformation that maps the matrix results of $log(\bullet)$ to a vector, which is used to calculate the error between the registration result and the ground truth. If $T_{gt,i}$ is equal to $T_{est,i}$, then $T_{gt,i}^{-1} T_{est,i}$ will be the identity matrix, and the result of applying the log operation to the identity matrix will be zero. To comprehensively compare and assess the overall

TABLE I
THE RMSE OF DIFFERENT REGISTRATION METHODS ON 3DMATCH. THE SMALLEST RMSE VALUES ARE HIGHLIGHTED IN BOLD TO INDICATE SUPERIOR REGISTRATION PERFORMANCE

Scenes	Method					
	D3Feat	DGR	REGTR	Predator	RoREG	Ours
kitchen	0.709	0.686	0.285	0.394	0.275	**0.170**
jan1	1.009	0.921	0.692	0.845	**0.513**	0.603
seq30	1.564	1.556	1.013	1.198	1.044	**1.006**
scan3	0.782	0.993	0.259	0.302	0.231	**0.137**
hotel1	1.412	1.213	0.768	0.812	**0.552**	0.679
hotel3	1.513	1.568	0.473	1.068	0.504	**0.342**
study-room	1.230	1.264	1.079	0.939	**0.509**	0.822
erika	0.822	1.026	0.212	1.452	0.262	**0.182**

accuracy of point cloud registration, we use the root mean square error (RMSE) to quantify the overall error between multiple poses and ground truth. The RMSE is defined as follows, where N is the number of registration results:

$$RMSE \equiv APE_{all} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|log(T_{gt,i}^{-1}T_{est,i})^{\vee}\right\|_2^2} \quad (3)$$

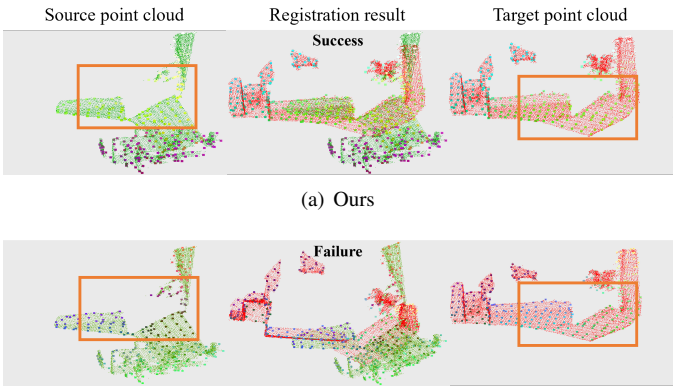
Notably, APE and RMSE are unitless, as they are derived from the transformation matrix T, which includes the rotation matrix R and the translation vector t. For both APE and RMSE, smaller values indicate better registration accuracy.

C. Indoor Scene Experiments

In indoor registration experiments, we evaluate the proposed method on the 3DMatch, 3DLoMatch, and ModelNet40 datasets. Consistent with the configurations in Predator [37] and REGTR [13], the point cloud pairs in 3DMatch exhibit >30% overlap, while those in 3DLoMatch range between 10% and 30%. In this section, the proposed fine-grained module is configured with a ***scale*** value of 8. An ablation experiment on ***scale*** selection will be introduced in the next section.

To intuitively compare the differences in point cloud feature acquisition between our method and the REGTR baseline, we perform feature visualization on the 3DMatch dataset to demonstrate the effect of the proposed fine-grained multi-scale module, as shown in Fig. 6. Specifically, we map the high-dimensional feature vectors to the RGB color space through Principal Component Analysis (PCA). This approach ensures that semantically similar feature points in the source and target point clouds exhibit similar colors, allowing for a visual comparison of feature semantics. During this process, the feature vectors are extracted by different methods and processed by the same transformer layers. As shown in Fig. 6, the feature points derived from our method in the overlapping regions of the source and target point clouds exhibit high similarity, which is critical for identifying the overlapping region and improving the registration accuracy.

Table I presents the RMSE of different registration methods on various indoor scenes of the 3DMatch dataset. The results show that our method achieves the smallest RMSE in most test scenes, while D3Feat [33] and DGR [12] exhibit comparatively



(a) Ours

(b) REGTR

Fig. 6. The feature visualization results of point cloud registration. In this figure, the left side of each subplot shows the source point cloud, the right side shows the target point cloud, and the middle shows the registration result of the two point clouds. Green and red represent the source and target point clouds, respectively. Colored points denote feature points, with the color of each point determined by its feature vector. Points with similar colors indicate higher feature vector similarity.

TABLE II
THE RUNTIME (MILLISECONDS) OF DIFFERENT REGISTRATION METHODS ON VARIOUS SCENES OF THE 3DMATCH

Scenes	Method			
	REGTR	Predator	RoREG	Ours
kitchen	57.0	330.5	279.2	62.6
jan1	53.8	154.0	353.6	59.2
seq30	55.4	318.4	314.2	57.3
scan3	65.1	368.4	348.5	71.5
hotel1	62.7	380.9	428.3	69.0
hotel3	62.7	380.9	492.2	61.7
studyroom	68.2	309.2	393.7	69.1
erika	60.9	329.9	435.2	66.9
AvgTime	60.7	321.5	380.6	64.6

high errors. RoReg [38] achieves the best accuracy in some test scenes but requires point cloud preprocessing and incurs long computational latency, making it difficult to meet real-time requirements in practical applications. Table II shows the computational efficiency of various methods on 3DMatch, demonstrating that our method maintains competitive computational efficiency while achieving higher registration accuracy. Experiments on 3DLoMatch, presented in Table III, show that our method maintains high-precision registration even

TABLE III
THE RMSE OF DIFFERENT METHODS ON 3DLOMATCH. THE SMALLEST RMSE VALUES ARE HIGHLIGHTED IN BOLD TO INDICATE SUPERIOR REGISTRATION PERFORMANCE

Methods	Scenes							
	kitchen	jan1	seq30	scan3	hotel1	hotel3	study-room	erika
Predator	2.303	4.150	2.667	3.029	2.805	2.289	**2.731**	3.728
REGTR	2.238	4.107	2.367	2.924	2.543	1.886	3.171	3.514
Ours	**2.081**	**3.817**	**2.246**	**2.893**	**2.303**	**1.610**	2.755	**2.537**

TABLE IV
THE REGISTRATION RECALL OF DIFFERENT REGISTRATION METHODS ON VARIOUS SCENES FROM THE 3DMATCH AND 3DLOMATCH

Methods	kitchen	jan1	seq30	scan3	hotel1	hotel3	study-room	erika
3DMatch Registration Recall (%)								
REGTR	98.2	85.8	**76.7**	96.7	**94.9**	**100.0**	**84.6**	**95.6**
Ours	**98.4**	**90.6**	76.1	**97.8**	**94.9**	**100.0**	**84.6**	94.3
3DLoMatch Registration Recall (%)								
REGTR	56.2	51.4	63.1	63.2	53.3	65.9	45.8	58.0
Ours	**67.3**	**59.9**	**68.9**	**72.7**	**61.3**	**68.3**	**51.3**	**68.1**

TABLE V
THE RMSE AND COMPUTATIONAL RUNTIME OF DIFFERENT REGISTRATION METHODS ON MODELNET40. THE SMALLEST RMSE VALUES ARE HIGHLIGHTED IN BOLD TO INDICATE SUPERIOR REGISTRATION PERFORMANCE

	REGTR	Predator	Ours
RMSE	0.0895	0.1140	**0.0815**
Run Time(ms)	15.8	101.8	19.7

under low-overlap point cloud in indoor scene. We also use the standard benchmark metric, Registration Recall (RR), for a comprehensive evaluation. The experimental configuration follows that of REGTR, and the results in Table IV confirm the robustness of our method in registration.

In addition, we conducted registration experiments on the ModelNet40 dataset. Since other methods were not trained on ModelNet40, we only compared REGTR, Predator, and our method on this dataset. The point cloud registration accuracy of these methods on ModelNet40 is presented in Table V.

Extensive experiments on 3DMatch, 3DLoMatch, and ModelNet40 confirm that our method achieves high-precision point cloud registration. Compared to the REGTR baseline, our method reduces the average RMSE by 17.6% on 3DMatch and 8.9% on ModelNet40, respectively. Although the fine-grained multi-scale module introduces additional computational overhead, the proposed accelerated registration strategy ensures that our method maintains highly competitive computational efficiency. In the 3DMatch experiments, D3Feat and DGR exhibit relatively high RMSE values, likely due to interference from points outside the overlapping region, which affects these feature-based methods.

D. Outdoor Scene Experiments

For outdoor scene experiments, we evaluate the performance of our method and the REGTR baseline on the MCD [15] dataset. MCD includes three diverse large-scale outdoor scenes and provides corresponding semantic labels for the point clouds. By leveraging these semantic labels, we evaluate the performance of our method in point cloud registration using semantic information. The results of different methods on the MCD dataset are presented in Table VI.

In the experiments, three outdoor scenes (TUHH, NTU, and KTH) are used to train and evaluate our method and REGTR. Within the MCD dataset, different point cloud sequences were

TABLE VI
THE RMSE OF DIFFERENT REGISTRATION METHODS ON MCD. THE SMALLEST RMSE VALUES ARE HIGHLIGHTED IN BOLD TO INDICATE SUPERIOR PERFORMANCE. OURS-SEMANTIC DENOTES THE TRAINING AND TESTING OF THE PROPOSED METHOD USING POINT CLOUDS WITH CORRESPONDING SEMANTIC LABELS

Methods	Scenes			
	TUHH04	KTH04	NTU10	Semantic-TUHH08
REGTR	2.467	2.086	9.568	0.277
Ours	**2.400**	**1.840**	**9.446**	0.271
Ours-Semantic	-	-	-	**0.250**

TABLE VII
PERFORMANCE OF OUR METHOD WITH DIFFERENT ***Scales***. THE AVERAGE ROOT MEAN SQUARE ERROR AND AVERAGE COMPUTATIONAL RUNTIME (MILLISECONDS) ARE USED TO EVALUATE REGISTRATION ACCURACY AND COMPUTATIONAL EFFICIENCY, RESPECTIVELY

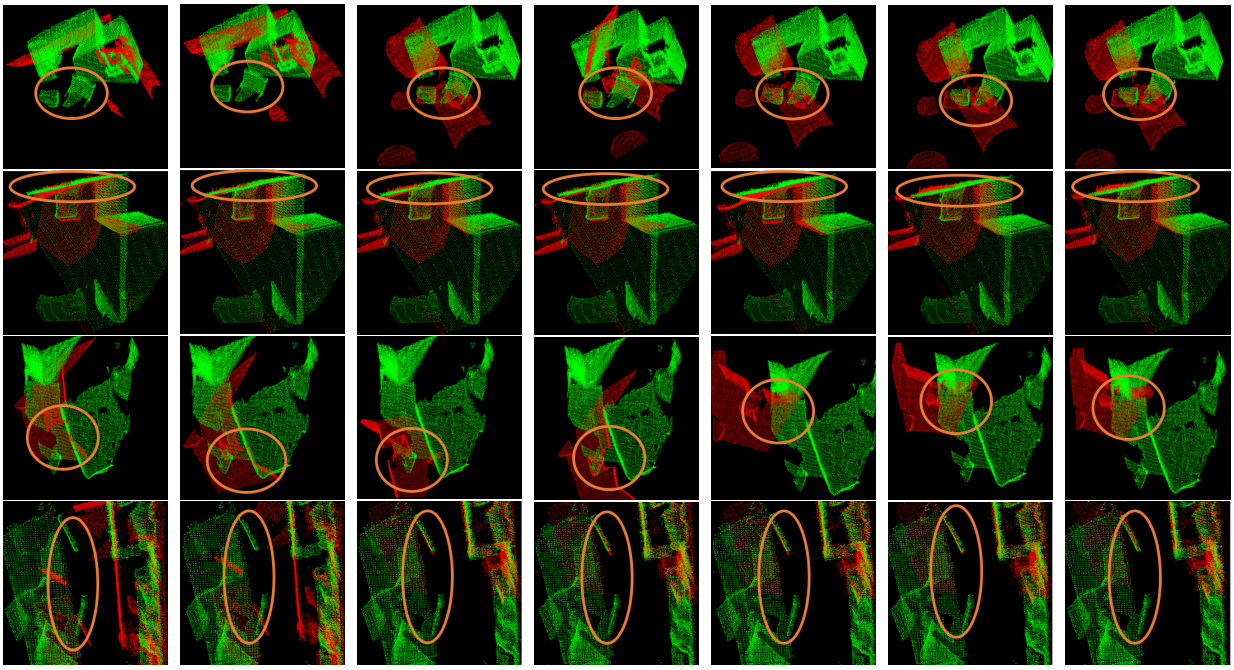
Methods	3DMatch		ModelNet40	
	AvgRMSE	AvgTime	AvgRMSE	AvgTime
REGTR	0.598 (0%)	60.7	0.0895 (0%)	15.8
10scale	0.520 (-13.0%)	76.5	0.0990 (10.6%)	20.8
8scale	0.493 (-17.6%)	64.6	0.0815 (-8.9%)	19.7
6scale	0.504 (-15.7%)	63.9	0.0842 (-5.9%)	19.1
4scale	0.560 (-6.4%)	63.2	0.0876 (-2.1%)	18.5

captured for each scene, based on variations in data acquisition timing and paths. We preprocess the MCD data to prepare them for training registration methods. Taking the TUHH scene as an example, we split the TUHH09 sequence into training and validation sets at an 8:2 ratio for model training, while the TUHH04 sequence is used as test data to evaluate the point cloud registration accuracy. We apply this dataset preparation process to KTH and NTU, generating scene-specific training and testing datasets. The experimental results for different outdoor scenes are presented in Table VI. As shown by these results, our method achieves superior registration accuracy compared to the REGTR baseline, demonstrating its robustness under diverse outdoor environmental conditions.

In addition, we use the semantic labels provided by MCD to test the scalability of our method. To incorporate semantic information, we first construct four-dimensional vectors by concatenating 3D point coordinates with discrete semantic labels, and then use MLP (Multi-Layer Perceptron) to fuse this semantic information with the fine-grained features extracted from the point clouds by our method. As shown in Table VI, our method achieves a significant improvement in registration accuracy after incorporating semantic information. Although the semantic integration method used in this experiment is simple, it confirms the scalability of our method in semantic-augmented registration.

E. Ablation Study

1) Fine-grained module ablation study: Ablation studies are systematically conducted to quantify the contribution of multi-scale configurations in the proposed fine-grained module



(a) D3Feat (b) DGR (c) REGTR (d) Predator (e) RoREG (f) Ours (g) GT

Fig. 7. The registration visualized results on 3DMatch. (a) D3Feat [33], (b) DGR [12], (c) REGTR [13], (d) Predator [37], (e) RoREG [38], (f) Ours, and (g) Ground truth.

to registration accuracy. Specifically, we evaluate how different ***scale*** settings affect the RMSE of point cloud registration. Table VII presents the performance of the proposed module on the 3DMatch and ModelNet40 datasets under various ***scale*** settings. In these ablation experiments, the average RMSE is used as the evaluation metric. As ***scale*** value increases, the 3D feature descriptor is divided into finer segments, enabling more intensive information interaction within the descriptor. Consequently, the proposed module can obtain more fine-grained features, which helps improve the accuracy of point cloud registration.

From the results in Table VII, the fine-grained module with a ***scale*** of 8 achieves a favorable balance between accuracy and efficiency. Its runtime on 3DMatch is 6.4% higher than REGTR, but the average registration error is 17.6% lower. However, the performance of the fine-grained module with a ***scale*** of 10 is weaker on different datasets compared to the module with lower ***scales*** values. The reason is that a large ***scale*** increases network complexity, potentially leading to overfitting and performance degradation. Therefore, ***scale*** should be selected based on the specific requirements for accuracy and efficiency in point cloud registration.

2) Accelerated registration strategy ablation study: A computational efficiency ablation study is conducted to analyze the impact of our speed optimization strategy. In this evaluation, the proposed fine-grained multi-scale module is disabled to solely assess the effect of filtering thresholds on the results.

The filtering threshold in the acceleration strategy is used to filter feature points with low probability in the overlapping region while ensuring that the remaining feature points satisfy the demands for point cloud registration. We evaluate the impact of different filtering thresholds on registration recall using the kitchen and jan1 scenes from 3DMatch, with results presented in Table VIII. As shown in Table VIII, when the threshold is set to 0.9, the registration recall experiences a significant decrease, while other threshold values have only a minimal influence. In our experiment, the filtering threshold is set to 0.85. We further evaluate the impact of this filtering threshold on registration accuracy and efficiency on the 3DMatch and ModelNet40 datasets, with results presented in Table IX and Table X. For this evaluation, we use the RMSE and the computational time required for transformation estimation as metrics. Since the proposed accelerated registration strategy is applied to the transformation estimation stage in REGTR, the experiments focus on comparing the computational speed of this stage. The results in Table IX and Table X demonstrate that our speed optimization strategy effectively reduces the computation time for transformation

TABLE VIII
THE REGISTRATION RECALL OF THE ACCELERATED REGISTRATION STRATEGY UNDER DIFFERENT FILTERING THRESHOLDS ON THE KITCHEN AND JAN1 SCENES FROM THE 3DMATCH

threshold	th=0.9	th=0.8	th=0.7	th=0.6	th=0.5	REGTR (th=0.0)
3DMatch Registration Recall (%)						
kitchen	97.8	98.4	98.4	98.4	98.4	98.4
jan1	84.0	85.5	85.8	85.8	85.8	85.8
3DLoMatch Registration Recall (%)						
kitchen	52.6	56.3	56.8	56.8	56.8	56.8
jan1	45.7	51.2	51.4	51.4	51.4	51.4

TABLE IX
PERFORMANCE COMPARISON BETWEEN THE ACCELERATED OPTIMIZED METHOD AND THE BASELINE METHOD ON 3DMATCH. THE FIFTH COLUMN SHOWS THE TIME REDUCTION RATE

Scenes	REGTR		Speed-REGTR	
	RMSE	Pose Estimation Time(ms)	RMSE	Pose Estimation Time(ms)
kitchen	0.285	21.5	0.291	13.6(-36.7%)
jan1	0.692	12.9	0.690	6.5(-49.6%)
seq30	1.013	13.9	0.988	6.6(-52.5%)
scan3	0.259	13.8	0.252	5.7(-58.7%)
hotel1	0.768	11.1	0.770	5.6(-49.6%)
hotel3	0.473	15.7	0.490	8.7(-44.6%)
study-room	1.079	12.9	0.904	6.6(-48.8%)
erika	0.212	13.3	0.198	7.6(-42.9%)

TABLE X
PERFORMANCE COMPARISON BETWEEN THE ACCELERATED OPTIMIZED METHOD AND THE BASELINE METHOD ON MODELNET40. THE THIRD LINE SHOWS THE TIME REDUCTION RATE

	REGTR	Speed-REGTR
RMSE	0.0895	0.0895
Mean Error	0.0420	0.0419
Pose Estimation Time(ms)	15.8	10.4(-34.2%)

estimation while maintaining registration accuracy comparable to that of REGTR.

F. Visualization Results

For qualitative experiments, we provide a visual comparison of the registration performance across different methods, as shown in Fig. 7. The source and target point clouds are represented by green and red points, respectively, while orange ellipses highlight regions with significant alignment discrepancies between the registration results and the ground truth. From the results in the first two rows, our method achieves the best registration performance. In the last three rows, D3Feat and DGR exhibit relatively large registration errors, while our method exhibits almost the same registration accuracy as REGTR.

V. CONCLUSIONS

Inspired by Res2Net, this paper constructs a fine-grained multi-scale module for 3D point clouds and integrates it into the backbone network of REGTR, enhancing feature correlation. Comprehensive experiments on indoor and outdoor datasets show that our method achieves better registration accuracy than REGTR. In addition, we propose an accelerated registration strategy to balance accuracy and efficiency of point cloud registration. Its effectiveness is verified through experimental results on the 3DMatch and ModelNet40 datasets. However, despite the proposed accelerated registration strategy, the introduced fine-grained multi-scale module increases the model's computational overhead. Compared to the REGTR baseline, our approach achieves higher accuracy at the expense of a small amount of computational efficiency, which may present challenges for resource-constrained or real-time applications. Future work will thus focus on model lightweighting to improve practicality. Furthermore, we have conducted point cloud registration experiments with semantic information integration on the MCD outdoor dataset, showing promising performance. However, in these experiments, semantic information was integrated by simply concatenating discrete labels with 3D coordinates. Therefore, it is also necessary to explore more sophisticated semantic fusion schemes in the future.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant No.52574193, Grant No. 92367111, Grant U24A20265, and the Key Science and Technology Innovation Project of CCTEG under Grants 2024-TD-ZD016-01 and 2024-TD-MS017.

REFERENCES

[1] A. de Carvalho Santana and A. Silva Macedo, "Use of lidar sensors for non-contact, real-time measurement of ore mass on belt conveyors," *IEEE Latin America Transactions*, vol. 22, no. 1, pp. 63–70, 2024, doi:10.1109/TLA.2024.10375737.

[2] J. Gaia, E. Orosco, F. Rossomando, and C. Soria, "Mapping the landscape of slam research: A review," *IEEE Latin America Transactions*, vol. 21, no. 12, pp. 1313–1336, 2023, doi:10.1109/TLA.2023.10305240.

[3] M. Yuan, X. Huang, K. Fu, Z. Li, and M. Wang, "Boosting 3d point cloud registration by transferring multi-modality knowledge," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 11 734–11 741, doi:10.1109/ICRA48891.2023.10161411.

[4] X. Huang, G. Mei, J. Zhang, and R. Abbas, "A comprehensive survey on point cloud registration," *ArXiv*, vol. abs/2103.02690, 2021, doi:arXiv:2103.02690.

[5] Y.-X. Zhang, J. Gui, X. Cong, X. Gong, and W. Tao, "A comprehensive survey and taxonomy on point cloud registration based on deep learning," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024, doi:10.24963/ijcai.2024/922.

[6] M. Lyu, J. Yang, Z. Qi, R. Xu, and J. Liu, "Rigid pairwise 3d point cloud registration: A survey," *Pattern Recognition*, vol. 151, p. 110408, 2024, doi:10.1016/j.patcog.2024.110408.

[7] S. Granger and X. Pennec, "Multi-scale em-icp: A fast and robust approach for surface registration," in *7th European Conference on Computer Vision, Copenhagen, Denmark, May 28-31, 2002, Proceedings, Part IV*, 2002, pp. 418–432, doi:10.1007/3-540-47979-1_28.

[8] S. Salti, F. Tombari, and L. D. Stefano, "Shot: Unique signatures of histograms for surface and texture description," *Computer Vision and Image Understanding*, vol. 125, pp. 251–264, 2014, doi:10.1016/j.cviu.2014.04.011.

[9] A. Hatem, Y. Qian, and Y. Wang, "Point-tta: Test-time adaptation for point cloud registration using multitask meta-auxiliary learning," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 16 448–16 458, doi:10.1109/ICCV51070.2023.01512.

[10] P. Vial, M. Malagón, R. Segura, N. Palomeras, and M. Carreras, "Gmm registration: a probabilistic scan matching approach for sonar-based auv navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 1033–1039, doi:10.1109/ICRA48891.2023.10160697.

[11] J. Zhang, F. Xie, L. Sun, P. Zhang, Z. Zhang, J. Chen, F. Chen, and M. Yi, "Multi-view point cloud registration based on improved ndt algorithm and odm optimization method," *IEEE Robotics and Automation Letters*, vol. 9, pp. 6816–6823, 2024, doi:10.1109/LRA.2024.3408086.

[12] C. B. Choy, W. Dong, and V. Koltun, "Deep global registration," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2511–2520, doi:10.1109/CVPR42600.2020.00259.

[13] Z. J. Yew and G. H. Lee, "Regtr: End-to-end point cloud correspondences with transformers," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6667–6676, doi:10.1109/CVPR52688.2022.00656.

[14] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser, "3dmatch: Learning local geometric descriptors from rgb-d reconstructions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 199–208, doi:10.1109/CVPR.2017.29.

[15] T.-M. Nguyen, S. Yuan, T. H. Nguyen, P. Yin, H. Cao, L. Xie, M. Wozniak, P. Jensfelt, M. Thiel, J. Ziegenbein, and N. Blunder, "Mcd: Diverse large-scale multi-campus dataset for robot perception," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 22 304–22 313, doi:10.1109/CVPR52733.2024.02105.

[16] X. Zhang, J. Yang, S. Zhang, and Y. Zhang, "3d registration with maximal cliques," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 17 745–17 754, doi:10.1109/CVPR52729.2023.01702.

[17] X. Huang, S. Li, Y. Zuo, Y. Fang, J. Zhang, and X. Zhao, "Unsupervised point cloud registration by learning unified gaussian mixture models," *IEEE Robotics and Automation Letters*, vol. 7, pp. 7028–7035, 2022, doi:10.1109/LRA.2022.3180443.

[18] G. Mei, H. Tang, X. Huang, W. Wang, J. Liu, J. Zhang, L. Van Gool, and Q. Wu, "Unsupervised deep probabilistic approach for partial point cloud registration," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 13 611–13 620, doi:10.1109/CVPR52729.2023.01308.

[19] C. Choy, J. Park, and V. Koltun, "Fully convolutional geometric features," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 8957–8965, doi:10.1109/ICCV.2019.00905.

[20] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6410–6419, doi:10.1109/ICCV.2019.00651.

[21] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 77–85, doi:10.1109/CVPR.2017.16.

[22] F. Poiesi and D. Boscaini, "Distinctive 3d local deep descriptors," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 5720–5727, doi:10.1109/ICPR48806.2021.9411978.

[23] Z. Wang, S. Huang, J. A. Butt, Y. Cai, M. Varga, and A. Wieser, "Cross-modal feature fusion for robust point cloud registration with ambiguous geometry," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 227, pp. 31–47, 2025, doi:10.1016/j.isprsjprs.2025.05.012.

[24] H. Chen, B. Chen, Z. Zhao, and B. Song, "Point cloud registration based on learning gaussian mixture models with global-weighted local representations," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023, doi:10.1109/LGRS.2023.3256005.

[25] Z. Yang, J. Z. Pan, L. Luo, X. Zhou, K. Grauman, and Q. Huang, "Extreme relative pose estimation for rgb-d scans via scene completion," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4526–4535, doi:10.1109/CVPR.2019.00466.

[26] X. Huang, G. Mei, and J. Zhang, "Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 363–11 371, doi:10.1109/CVPR42600.2020.01138.

[27] 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, pp. 23 872–23 884, 2021, doi:10.1016/j.neucom.2024.128763.

[28] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, and K. Xu, "Geometric transformer for fast and robust point cloud registration," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 133–11 142, doi:10.1109/CVPR52688.2022.01086.

[29] W. Gu, M. Han, L. Xue, H. Dong, C. Yang, R. Chen, and L. Wei, "Gpi-net: gestalt-guided parallel interaction network via orthogonal geometric consistency for robust point cloud registration," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, ser. IJCAI '25, 2025, doi:10.24963/ijcai.2025/118.

[30] S. Ao, Q. Hu, H. Wang, K. Xu, and Y. Guo, "Buffer: Balancing accuracy, efficiency, and generalizability in point cloud registration," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 1255–1264, doi:10.1109/CVPR52729.2023.00127.

[31] R. She, Q. Kang, S. Wang, W. P. Tay, K. Zhao, Y. Song, T. Geng, Y. Xu, D. N. Navarro, and A. Hartmannsgruber, "Pointdifformer: Robust point cloud registration with neural diffusion and transformer," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024, doi:10.1109/TGRS.2024.3351286.

[32] G. Chen, M. Wang, L. Yuan, Y. Yang, and Y. Yue, "Rethinking point cloud registration as masking and reconstruction," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 17 671–17 681, doi:10.1109/ICCV51070.2023.01624.

[33] X. Bai, Z. Luo, L. Zhou, H. Fu, L. Quan, and C.-L. Tai, "D3feat: Joint learning of dense detection and description of 3d local features," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6358–6366, doi:10.1109/CVPR42600.2020.00639.

[34] Z. Qiao, Z. Yu, H. Yin, and S. Shen, "Pyramid semantic graph-based global point cloud registration with low overlap," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 11 202–11 209, doi:10.1109/IROS55552.2023.10341394.

[35] S. Gao, M.-M. Cheng, K. Zhao, X. Zhang, M.-H. Yang, and P. H. S. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 652–662, 2019, doi:10.1109/TPAMI.2019.2938758.

[36] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1912–1920, doi:10.1109/CVPR.2015.7298801.

[37] S. Huang, Z. Gojcic, M. Usvyatsov, A. Wieser, and K. Schindler, "Predator: Registration of 3d point clouds with low overlap," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4265–4274, doi:10.1109/CVPR46437.2021.00425.

[38] H. Wang, Y. Liu, Q. Hu, B. Wang, J. Chen, Z. Dong, Y. Guo, W. Wang, and B. Yang, "Roreg: Pairwise point cloud registration with oriented descriptors and local rotations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 10 376–10 393, 2023, doi:10.1109/TPAMI.2023.3244951.

Huaiyuan Yu received the B.S. degree in Automation from Beijing University of Chemical Technology, China, in 2020, and the M.S. degree in Control Science and Engineering from the same university in 2023. Currently, he is pursuing a Ph.D. degree in Control Science and Engineering at Beijing University of Chemical Technology. His research interests include computer vision, point cloud registration and 3D reconstruction.

Haijiang Zhu received the Ph.D. degree in pattern recognition and intelligent system from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2004. From 2006 to 2007, he was a visiting scholar at the Faculty of Engineering, Iwate University, Japan. Currently, he is professor and Ph.D. supervisor in the College of Information Science and Technology at Beijing University of Chemical Technology. His research interests include image processing and computer vision.

Jian Cheng received the B.Sc. degree in Automation, the M.Sc. degree in Control Theory and Control Engineering, and the Ph.D. degree in Communication and Information System from the China University of Mining and Technology, Xuzhou, China, in 1997, 2003, and 2008 respectively. He has been a postdoctoral fellow at Tsinghua University and University of Birmingham from 2009 to 2013. He is currently a Professor and the Chief Scientist with the Research Institute of Mine Artificial Intelligence, Chinese Institute of Coal Science, Beijing, China. His current research interests include machine learning and pattern recognition, data mining and big data, as well as imbalance learning and image processing and their applications in industrial fields.

Ning An received the Ph.D. degree in Control Theory and Control Engineering from the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, in 2017, and the B.S. degree in Automation from China University of Mining and Technology, in 2011. He is currently a Professor at Institute of Mining Artificial Intelligence, Chinese Institute of Coal Science. His research interests include 3D perception for intelligent robot, 3D reconstruction of large-scale scenes, and multi-modal artificial intelligence systems.