Gao JJ, Dong QJ, Wang RA *et al.* OAAFormer: Robust and efficient point cloud registration through overlapping-aware attention in Transformer. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 39(4): 755-770 July 2024. DOI: 10.1007/s11390-024-4165-6

OAAFormer: Robust and Efficient Point Cloud Registration Through Overlapping-Aware Attention in Transformer

Jun-Jie Gao[1] (高俊杰), Qiu-Jie Dong[1] (董秋杰), Rui-An Wang[1] (王瑞安), Shuang-Min Chen[2] (陈双敏)
Shi-Qing Xin[1,*] (辛士庆), Chang-He Tu[1] (屠长河), and Wenping Wang[3] (王文平), *Fellow, ACM, IEEE*

[1] *School of Computer Science and Technology, Shandong University, Qingdao 266237, China*

[2] *School of Information and Technology, Qingdao University of Science and Technology, Qingdao 266061, China*

[3] *College of Engineering, Texas A&M University, Texas, TX 77843, U.S.A.*

E-mail: 202020631@mail.sdu.edu.cn; 202020639@mail.sdu.edu.cn; 202135163@mail.sdu.edu.cn; chenshuangmin@qust.edu.cn
xinshiqing@sdu.edu.cn; chtu@sdu.edu.cn; wenping@tamu.edu

Received February 1, 2024; accepted June 13, 2024.

Abstract In the domain of point cloud registration, the coarse-to-fine feature matching paradigm has received significant attention due to its impressive performance. This paradigm involves a two-step process: first, the extraction of multi-level features, and subsequently, the propagation of correspondences from coarse to fine levels. However, this approach faces two notable limitations. Firstly, the use of the Dual Softmax operation may promote one-to-one correspondences between superpoints, inadvertently excluding valuable correspondences. Secondly, it is crucial to closely examine the overlapping areas between point clouds, as only correspondences within these regions decisively determine the actual transformation. Considering these issues, we propose OAAFormer to enhance correspondence quality. On the one hand, we introduce a soft matching mechanism to facilitate the propagation of potentially valuable correspondences from coarse to fine levels. On the other hand, we integrate an overlapping region detection module to minimize mismatches to the greatest extent possible. Furthermore, we introduce a region-wise attention module with linear complexity during the fine-level matching phase, designed to enhance the discriminative capabilities of the extracted features. Tests on the challenging 3DLoMatch benchmark demonstrate that our approach leads to a substantial increase of about 7% in the inlier ratio, as well as an enhancement of 2%–4% in registration recall. Finally, to accelerate the prediction process, we replace the Conventional Random Sample Consensus (RANSAC) algorithm with the selection of a limited yet representative set of high-confidence correspondences, resulting in a 100 times speedup while still maintaining comparable registration performance.

Keywords point cloud registration, coarse-to-fine, overlapping region, feature matching, Transformer

1 Introduction

The task of point cloud registration involves determining a rigid transformation that aligns one point cloud with another. This challenge is of fundamental importance in the fields of computer vision and robotics and has wide-ranging applications, including 3D reconstruction, SLAM (simultaneous location and mapping), and autonomous driving. A common approach to this task involves two key stages: point feature matching and globally consistent refinement. During the point feature matching phase, the goal is to generate a set of initial correspondences with a high inlier ratio, ideally including as many true correspondences as possible while minimizing false ones. However, achieving this objective is a formidable chal-

Regular Paper

Special Section of CVM 2024

This work is supported by the National Natural Science Foundation of China under Grant Nos. 62272277, U23A20312, and 62072284, the National Key Technology Research and Development Program of the Ministry of Science and Technology of China under Grant No. 2022YFB3303200, and the Natural Science Foundation of Shandong Province of China under Grant No. ZR2020MF036.

*Corresponding Author

lenge due to inherent noise and disparities in the input point clouds, as well as the possibility of partial overlap between them. In the globally consistent refinement step, the focus shifts to rapidly identifying a subset of correspondences capable of consistently encoding the actual transformation through further refinement.

While a substantial body of literature[1, 2] has focused on the extraction of discriminative features to enhance correspondence quality, the inherent sparsity and disparities in point clouds, along with potential partial overlap, present persistent challenges. Recently, the coarse-to-fine matching paradigm[3, 4] has garnered significant attention for its impressive performance. This paradigm begins by downsampling the input point cloud into superpoints and establishing correspondences between these superpoints, where each superpoint inherently represents a point patch. Subsequently, sparse correspondences are propagated to encompass more points, resulting in the generation of dense correspondences.

However, accurately matching a superpoint from one scan to another can be challenging, as the corresponding point patches may not exhibit perfect alignment. As illustrated in Fig.1, suppose we have two input point clouds P and Q. The superpoint (patch) A is overlapped with B, C, and D simultaneously. Yet, the use of the Dual Softmax operation[5] within the coarse-to-fine paradigm has the potential to enforce one-to-one correspondences between superpoints, unintentionally excluding valuable correspondences. This represents the first limitation of the coarse-to-fine paradigm. On the other hand, it is crucial to examine the overlapping regions between point clouds, as only correspondences within these areas decisively determine the actual transformation. Consequently, there is a pressing need to enhance the discriminability of the features extracted from points within these overlapping regions to improve the overall performance of the coarse-to-fine paradigm.

Motivated by these considerations, we propose a robust matching network, named OAAFormer, with the explicit objective of augmenting the performance of the coarse-to-fine matching paradigm. This augmentation is achieved through the systematic integration of a suite of strategies meticulously designed to elevate the quality of correspondences. Firstly, OAAFormer employs a sophisticated soft matching mechanism, with the explicit purpose of seamlessly propagating potentially valuable correspondences from the coarse to the fine levels of the matching process. Secondly, OAAFormer incorporates an intricately designed overlapping region detection module, strategically engineered to minimize the probability of mismatches. Thirdly, it introduces a region-wise attention module characterized by linear computational

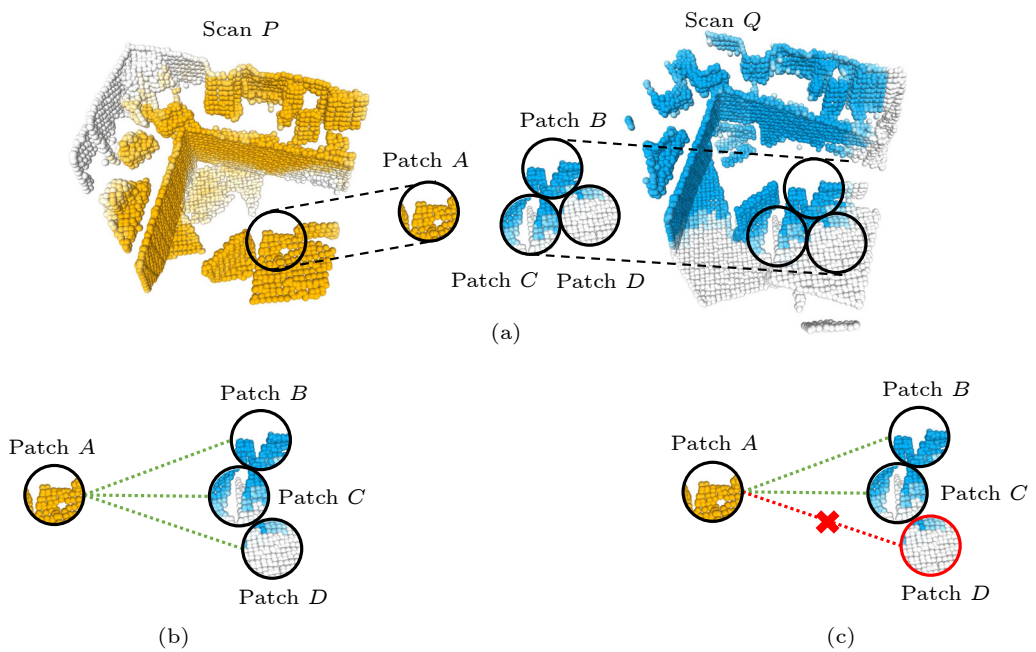

Fig.1. Illustration of coarse-level matching. (a) Patch A is overlapped with B, C, and D simultaneously. (b) Soft matching module for one-to-many matching. (c) Overlap detection module for eliminating mismatches outside predicted overlapping regions. Color intensity (yellow or blue) represents overlap scores.

complexity, meticulously designed to enhance the discriminative capabilities of the extracted features during the fine-level matching phase. Empirical validation underscores the efficacy of these strategies. For instance, tests on the exacting 3DLoMatch benchmark show that our approach yields a substantial increase of approximately 7% in the inlier ratio, as well as a discernible enhancement of 2%–4% in registration recall. Furthermore, we replace the conventional RANSAC algorithm[6] with the selection of a limited yet representative set of high-confidence correspondences for accelerating the prediction process.

In summary, the main contributions of this work are as follows.

• We propose a soft matching mechanism and an overlapping detection module to facilitate the propagation of potentially valuable correspondences from coarse to fine levels, which finally results in a substantial increase in inlier ratio and registration recall.

• We introduce a region-wise attention module with linear complexity during the fine-level matching phase, designed to enhance the discriminative capabilities of the extracted features.

• Through the replacement of the inefficient RANSAC algorithm with a more intelligent mechanism for selecting high-confidence correspondences, we achieve a remarkable 100 times acceleration in the prediction process.

2 Related Work

2.1 Point Cloud Registration

The construction of feature descriptors with specific characteristics proves to be an effective means of encoding the curvature of the underlying surface, providing valuable information for the alignment of point clouds. In previous researches, a multitude of traditional methods[7, 8] have relied on handcrafted features to craft such descriptors. With the proliferation of deep learning techniques, various learning-based descriptors[9–16] have been introduced to enhance the expressiveness of these feature descriptors. However, the task of identifying valuable correspondences between points based solely on geometric descriptors remains a challenging one, primarily due to the presence of various defects in the input point clouds, including noise, disparities, and partial overlapping. Consequently, approaches such as the Conventional Random Sample Consensus (RANSAC) algorithm[6, 17] or meticulously designed neural networks[18, 19] are frequently employed to address this challenge. These methods aim to eliminate mismatches, even when dealing with points possessing similar features, ultimately resulting in a more robust and accurate registration outcome.

Additionally, a variety of keypoint detectors tailored for rigid registration tasks have emerged. For instance, D3Feat[1] introduces a keypoint selection strategy that overcomes the inherent density variations of 3D point clouds. However, this approach does not fully account for overlapping areas and exhibits reduced robustness in scenarios with low overlap. Another noteworthy method, Predator[2], develops an overlap-attention block for early information exchange between the latent encodings of the two point clouds. Keypoints are selected based on both saliency and overlap scores. While Predator[2] demonstrates substantial improvements over existing methods across indoor and outdoor benchmarks, challenges persist in extracting a set of repeatable keypoints.

Recently, the coarse-to-fine paradigm has garnered attention for enhancing the quality of correspondences, not only in 2D image matching[20] but also in the domain of point cloud registration[3, 4]. For instance, CoFiNet[3] incorporates an optimal transport[21] matching layer to establish correspondences between mutually nearest patches and subsequently refines these correspondences at the fine-level stage. In a similar vein, GeoTrans[4] introduces a self-attention mechanism to learn geometric features, thereby improving the matching accuracy between superpoints based on whether their neighboring patches overlap.

In this paper, we further enhance the coarse-to-fine mechanism through a set of strategies, including a soft matching mechanism that streamlines the propagation of potentially valuable correspondences from the coarse to fine levels and a region-wise attention module characterized by linear complexity during the fine-level matching phase.

2.2 Efficient Transformer

In the standard Transformer model[22], the memory cost exhibits a quadratic increase due to matrix multiplication, which has become a bottleneck when handling long sequences. Recently, several efficient Transformer variants[23–25] have been introduced. For example, the Linear Transformer[23] reformulates self-attention as a linear dot product of kernel feature

maps and exploits the associativity property of matrix products to reduce computational complexity. BigBird[24] combines local and global attention mechanisms at specific positions and introduces random attention for selected token pairs. FastFormer[25] employs an additive attention mechanism to model global contexts, achieving effective context modeling with linear complexity. Inspired by these advancements, we propose a region-wise attention module with linear complexity during the fine-level matching phase, meticulously designed to enhance the discriminative capabilities of the extracted features for points within overlapping areas.

3 Method

3.1 Pipeline

Suppose that we have a source point cloud $P = \{\boldsymbol{p}_i \in \mathbb{R}^3 \mid i = 1, \ldots, N\}$ and a target point cloud $Q = \{\boldsymbol{q}_i \in \mathbb{R}^3 \mid i = 1, \ldots, M\}$. The objective of rigid registration is to estimate the unknown rigid transformation $\boldsymbol{T} = (\boldsymbol{R}, \boldsymbol{t})$, where $\boldsymbol{R}$ represents a rotation matrix and $\boldsymbol{t}$ represents a translation vector. Let $C^* = \{\boldsymbol{p}_{i_k} \mapsto \boldsymbol{q}_{j_k}, k = 1, \ldots, K\}$ denotes the set of ground-truth correspondences between P and Q. The true transformation $\boldsymbol{T}$ should accurately map each $\boldsymbol{p}_{i_k} \in P$ to $\boldsymbol{q}_{j_k} \in Q$, meaning that it should minimize the difference vector $\boldsymbol{R} \cdot \boldsymbol{p}_{i_k} + \boldsymbol{t} - \boldsymbol{q}_{j_k}$ to be nearly zero. In real-world scenarios, where the elusive optimal correspondences set C^* is challenging to obtain, the prevalent approach involves extracting a subset of correspondences that are deemed reasonably reliable between two point clouds. Subsequently, the estimation of the transformation matrix relies on the consistency of these correspondences.

As shown in Fig.2, our algorithmic pipeline includes four main modules.

1) In the feature extraction module, we utilize KPConv[13] as the backbone to downsample the point cloud and extract multi-level features. Subsequently, we select sample points from both the first and last levels for the subsequent matching process.

2) In the coarse-level matching module, we utilize GeoTrans[4] to generate the geometric features of the superpoints. Additionally, we estimate the overlap region using a dedicated detection module specifically designed for this purpose. Subsequently, we introduce a soft matching mechanism to extract valuable correspondences at the patch level, followed by a filtering step to remove potential mismatches (refer to Subsection 3.2).

3) In the fine-level matching module, we introduce a region-wise attention module characterized by linear complexity. This module is designed to enhance the discriminative capabilities of the extracted features (refer to Subsection 3.3).

4) In the efficient registration module, we propose an efficient seeding mechanism for the identification of high-confidence correspondences, aiming to expe-

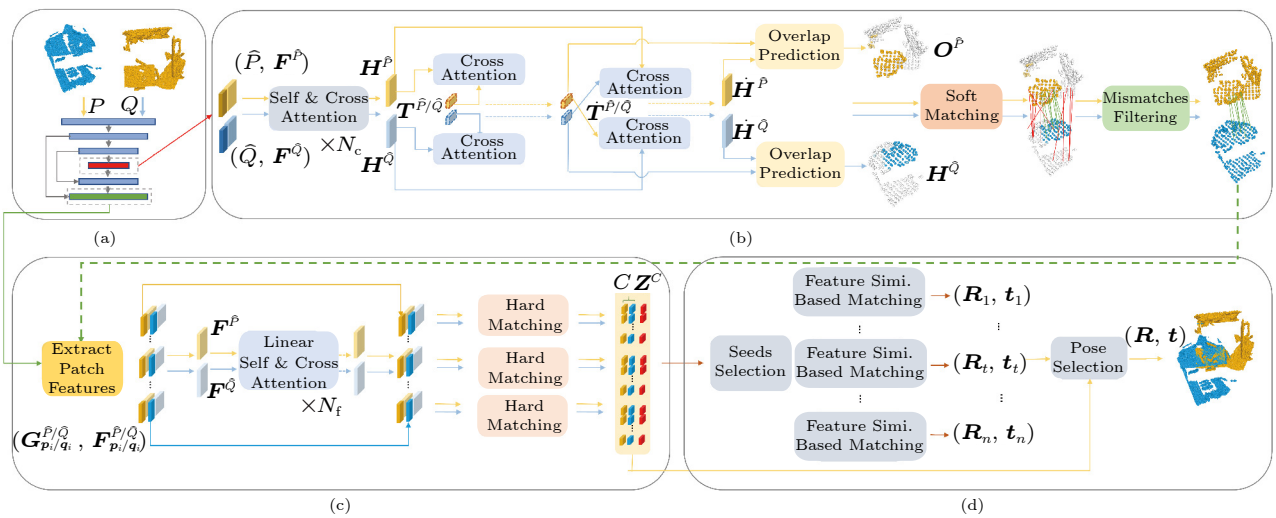

Fig.2. Algorithmic pipeline of our proposed OAAFormer. (a) Feature extraction. The feature extraction module downsamples the point cloud and extracts multi-level features. (b) Coarse-level matching. The coarse-level matching module employs a soft matching mechanism to establish one-to-many correspondences, while simultaneously utilizing an overlap detection module to eliminate mismatches outside the predicted overlapping regions. (c) Fine-level matching. The fine-level matching module employs a linear attention mechanism to enhance feature matching capabilities and utilizes a hard matching module to reduce mismatches, resulting in a set of correspondences along with their associated confidence scores. (d) Efficient registration. The efficient registration module, based on feature similarity, is utilized to accelerate registration. Simi. means similarity.

dite the process (refer to Subsection 3.4).

3.2 Coarse-Level Matching

3.2.1 Intra- and Inter- Consistency Enhancements

In the coarse-level matching phase, considering superpoints $\hat{P}$ and $\hat{Q}$ with associated features $\boldsymbol{F}^{\hat{P}} \in \mathbb{R}^{|\hat{P}| \times d_t}$ and $\boldsymbol{F}^{\hat{Q}} \in \mathbb{R}^{|\hat{Q}| \times d_t}$, respectively, we alternately apply the self-attention layer within each point cloud and the cross-attention layer between point clouds N_c times to enhance the consistency. It is worth noting that we utilize the geometry-aware self-attention mechanism[14] instead of the vanilla self-attention[22], as the former is better suited for capturing long-range contextual information.

3.2.2 Overlap Region Detection Module

To enhance the distinction between overlapping and non-overlapping regions, we introduce a token-based attention mechanism. Specifically, we employ a feature token, denoted as $\boldsymbol{T}^{\hat{P}}$, to encapsulate information related to the overlapping region. The initialization of $\boldsymbol{T}^{\hat{P}}$ is accomplished through a max-pooling operation applied to the augmented feature set $\boldsymbol{H}^{\hat{P}}$. Subsequently, we employ a cross-attention operation to update the token $\boldsymbol{T}^{\hat{P}}$, resulting in $\dot{\boldsymbol{T}}^{\hat{P}}$. This updated token is instrumental in distinguishing between overlapping and non-overlapping regions. In the implementation, the query originates from the initialized token $\boldsymbol{T}^{\hat{P}}$, while both keys and values are derived from the feature set $\boldsymbol{H}^{\hat{P}}$. Finally, the tokens obtained, $\dot{\boldsymbol{T}}^{\hat{P}}$ and $\dot{\boldsymbol{T}}^{\hat{Q}}$, serve as guiding elements for updating the original features $\boldsymbol{H}^{\hat{P}}$ and $\boldsymbol{H}^{\hat{Q}}$ through an additional cross-attention operation, respectively. This is formally represented as:

$$\dot{\boldsymbol{T}}^{\hat{P}} = \text{VanillaTransformer}(a = \boldsymbol{T}^{\hat{P}}, b = \boldsymbol{H}^{\hat{P}}, \text{c} = \boldsymbol{H}^{\hat{P}}),$$

where a, b, and c stand for the query, key, and value, respectively. $\dot{\boldsymbol{T}}^{\hat{Q}}$ is computed in the same way.

Subsequently, the obtained tokens $\dot{\boldsymbol{T}}^{\hat{P}}$ and $\dot{\boldsymbol{T}}^{\hat{Q}}$ are used as guide items to update the original features $\boldsymbol{H}^{\hat{P}}$ and $\boldsymbol{H}^{\hat{Q}}$ through another cross-attention operation:

$$\dot{\boldsymbol{H}}^{\hat{P}} = \text{VanillaTransformer}(a = \boldsymbol{H}^{\hat{P}}, b = \boldsymbol{T}^{\hat{Q}}, c = \boldsymbol{T}^{\hat{Q}}),$$

and $\dot{\boldsymbol{H}}^{\hat{Q}}$ is computed in the same way. During this process, $\boldsymbol{H}^{\hat{P}}$ and $\boldsymbol{H}^{\hat{Q}}$ are updated to $\dot{\boldsymbol{H}}^{\hat{P}}$ and $\dot{\boldsymbol{H}}^{\hat{Q}}$, respectively, such that they are aware of the overlapping region between $\hat{P}$ and $\hat{Q}$. The overlapping-aware mechanism is highly advantageous as it enhances the ability to effectively discriminate between the overlapping region and the non-overlapping region.

To further identify the location of the overlapping regions, we introduce an additional module that assigns a probability score indicating the likelihood that a point is within the overlapping region. Specifically, we project the decoded tokens $\dot{\boldsymbol{T}}^{\hat{P}}$ and $\dot{\boldsymbol{T}}^{\hat{Q}}$ through matrix multiplication and the sigmoid function to create the weight mapping. The weight map $\boldsymbol{w}^{\hat{P}}$ is employed to enhance the overlap information within the features. Subsequently, a linear projection operator $\boldsymbol{W}^{\text{O}} \in \mathbb{R}^{d_t \times 1}$, and a sigmoid function are applied to obtain the overlapping confidence:

$$\boldsymbol{w}^{\hat{P}} = \text{sigmoid}((\dot{\boldsymbol{H}}^{\hat{P}})^{\text{T}} \dot{\boldsymbol{T}}^{\hat{P}}),$$

$$\boldsymbol{O}^{\hat{P}} = \text{sigmoid}((\boldsymbol{w}^{\hat{P}} \odot \dot{\boldsymbol{H}}^{\hat{P}} + \dot{\boldsymbol{H}}^{\hat{P}}) \boldsymbol{W}^{\text{O}}),$$

where $\boldsymbol{O}^{\hat{Q}}$ is then computed in the same way. To this end, we consider the points whose confidences are greater than a threshold θ_o to be within the overlap region.

3.2.3 Soft-Matching Module

For the output features $\dot{\boldsymbol{H}}^{\hat{P}}$ and $\dot{\boldsymbol{H}}^{\hat{Q}}$ generated by the overlapping region detection module, we first normalize them to the unit hypersphere. Subsequently, we calculate the similarity matrix $\boldsymbol{M} \in \mathbb{R}^{|\hat{P}| \times |\hat{Q}|}$, where each element is defined as $\boldsymbol{m}_{i,j} = \exp\left(-\left\|\boldsymbol{h}_i^{\hat{P}} - \boldsymbol{h}_j^{\hat{Q}}\right\|_2^2\right)$.

Accordingly, we apply the softmax operation to the similarity matrix $\boldsymbol{S}$ on two dimensions separately to allow one-to-many matching. Next, we extract purified correspondences by applying a threshold θ_{m}.

$$\begin{aligned} \boldsymbol{M}_k &= \text{softmax}(\boldsymbol{M}(i, \cdot))_j, \\ \hat{C}_k &= \{(\hat{\boldsymbol{p}}_i, \hat{\boldsymbol{q}}_j) | \boldsymbol{S}_k(i, j) \geqslant \theta_m \| \}, \end{aligned} \tag{1}$$

where $k \in \{0, 1\}$, $\boldsymbol{M}_0$ and $\boldsymbol{M}_1$ are the matching probability matrix obtained by softmax operation along the first dimension and the zeroth dimension, respectively, and $\hat{C}_0$ and $\hat{C}_1$ are the corresponding coarse-level correspondences proposals. Compared with the commonly used top-k selection strategy that needs to specify the number of matches, our strategy of using a tolerance can ensure that the number of selected correspondences is adaptive to the overlapping rate.

It is important to acknowledge that while the previously mentioned strategy generates a larger number

of potentially beneficial correspondences, it may lead to a low inlier ratio. To enhance this inlier ratio, we introduce a procedure where, for each superpoint in the source point cloud, we initially identify the most closely matched target superpoint based on $\boldsymbol{S}$, as well as the k-nearest neighbors of the target superpoint. Out of these $k+1$ correspondences, only those that satisfy the condition defined in (1) are retained. Similarly, for each superpoint in the target point cloud, this process is repeated until a pruned correspondence set $\hat{C}_k$ is obtained. Finally, we further filter out mismatches outside predicted overlap regions.

3.3 Fine-Level Matching

3.3.1 *Linear Transformer*

Linear Transformer[23] is proposed to reduce the computation complexity by substituting the exponential kernel used in the original attention layer[22] with an alternative kernel function:

$$\mathrm{sim}(a,b) = \phi(a) \times \phi(b)^{\mathrm{T}},$$

where $\phi(\cdot) = \mathrm{elu}(\cdot) + 1$. Utilizing the associativity property of matrix products, the multiplication between $\phi(a)^{\mathrm{T}}$ and c can be carried out first. Since $d_{\mathrm{t}} \ll |P|$, the computation cost is reduced to $O(d_{\mathrm{t}})$. Here a, b, and c stand for the query, key, and value, respectively.

Thanks to our overlap region detection module, we perform linear attention operations to improve feature discrimination only for points within the overlap region and not for all dense points. This reduces the impact of points in the non-overlapping region on the one hand, and reduces the cost of calculation on the other hand. To be specific, we only focus on the points $\bar{P}$ within patch $\{\boldsymbol{G}^{\tilde{P}}_{\boldsymbol{p}_i} | \boldsymbol{p}_i \in \boldsymbol{O}^{\hat{P}}\}$ instead of all dense points $\tilde{P}$, and the relevant features note as $\boldsymbol{F}^{\bar{P}}$. We perform the same operation to get overlapping region points $\bar{Q}$ and relevant features $\boldsymbol{F}^{\bar{Q}}$.

Next, we adopt the Linear Transformer[23] to perform the self- and cross-attention to collect the global information through intra- and inter-relationship between features $\boldsymbol{F}^{\bar{P}}$ and $\boldsymbol{F}^{\bar{Q}}$. The self-attention layer updates its message by:

$$\boldsymbol{Z}^{\bar{P}} = \mathrm{LinearTransformer}(a = \boldsymbol{F}^{\bar{P}}, b = \boldsymbol{F}^{\bar{P}}, c = \boldsymbol{F}^{\bar{P}}),$$

and for $\boldsymbol{Z}^{\bar{Q}}, a = \boldsymbol{F}^{\bar{Q}}, b = \boldsymbol{F}^{\bar{Q}}, c = \boldsymbol{F}^{\bar{Q}}$. The cross-attention layer updates messages with information collected from the inter-relationship between two frame features:

$$\boldsymbol{Z}^{\bar{P}} = \mathrm{LinearTransformer}(a = \boldsymbol{F}^{\bar{P}}, b = \boldsymbol{F}^{\bar{Q}}, c = \boldsymbol{F}^{\bar{Q}}),$$

and for $\boldsymbol{Z}^{\bar{Q}}, a = \boldsymbol{F}^{\bar{Q}}, b = \boldsymbol{F}^{\bar{P}}, c = \boldsymbol{F}^{\bar{P}}$.

3.3.2 *Relative Position Embedding*

Unlike the previous work, which either chooses to reduce the point cloud resolution[26, 27] to decrease the computing overhead of the Transformer or only aims to enhance the feature representation capability of superpoints[4, 27], our approach introduces a Linear Transformer[23] to augment the fine-level features. To improve the rotation invariance of the features, inspired by the work of Lepard[26], we integrate rotation-invariant information by adding rotation position embeddings[28] to the inputs at each Transformer layer. This helps mitigate limitations on rotation datasets.

3.3.3 *Hard-Matching Module*

Through the aforementioned operations, we obtain a series of one-to-many superpoint correspondences situated in overlapping regions. The associated patches may have a low overlap rate, inevitably leading to a large number of dense point mismatches. Therefore, different from the soft matching strategy in Subsection 3.2, adopting a stricter matching strategy to suppress mismatches at the fine level is the key to obtaining robust registration. Hence, we employ the point matching module[4], which operates in conjunction with the optimal transmission strategy[21], to extract dense correspondences. The resultant correspondence set is denoted as C. Additionally, the confidence score of C is denoted as $\boldsymbol{Z}^C$.

3.4 Feature Similarity Based Efficient Registration

In robust pose estimators such as RANSAC[21], a large number of iterations is typically required to guarantee accuracy, leading to inefficiency. Considering the high inlier ratio of OAAFormer, we have designed an efficient estimator to achieve comparable performance while significantly reducing the computational cost. This design is motivated by the crucial observation that a well-distributed set of correspondences, which are more similar in the feature space, is beneficial for transform estimation.

3.4.1 Global Sampling Strategy

In order to obtain the global sampling distribution, we employ the spectral matching technique[29] to select reliable seeds. Correspondences with a local maximum confidence score within their neighborhood with radius R are then chosen. The number of seed points N_s is determined by the proportion of the whole correspondences $|C|$. For each seed, we select its k-nearest neighbors in $\boldsymbol{Z}^C$ to expand into a consensus set. The total consensus sets can be noted as: $\boldsymbol{S} \in \mathbb{R}^{N_s \times k}$.

3.4.2 Feature Similarity Compatibility

We conduct further analysis on the feature similarity of correspondences within each consensus set. The intra-difference of each correspondence in a consensus set is denoted as $\boldsymbol{D}^F \in \mathbb{R}^{k\times 1}$, and subsequently normalized as: $\boldsymbol{D}^F = 1 - \boldsymbol{D}^F / \max(\boldsymbol{D}^F)$. Moreover, we employ a sigmoid operation to expand the inter-difference of correspondences as follows:

$$\boldsymbol{D}^F = \text{sigmoid}((\boldsymbol{D}^F - \text{mean}(\boldsymbol{D}^F)) \times \sigma_s),$$

where σ_s is a parameter controlling the sensitivity to differences in features. Simultaneously, $\boldsymbol{D}^F$ serves as a feature similarity score. The closer the correspondence features are, the closer the score is to 1; otherwise, it approaches 0. Subsequently, we compute the compatibility matrix of this consensus set, denoted as $\boldsymbol{M} \in \mathbb{R}^{k\times k}$, where each element of $\boldsymbol{M}$ represents the minimum value of the two correspondence scores.

3.4.3 Hypothesis Selection

The association of each correspondence with the leading eigenvector is adopted as the weight for this correspondence and can be solved by the power iteration algorithm. Then we use the weighted SVD on the consensus set to generate an estimation $(\boldsymbol{R}_i, \boldsymbol{t}_i)$ for each seed. Finally, we choose the transformation that allows the most correspondences in C:

$$\boldsymbol{R}, \boldsymbol{t} = \max_{\boldsymbol{R}_i,\ \boldsymbol{t}_i} \sum_{(\tilde{\boldsymbol{p}}_j,\ \tilde{\boldsymbol{q}}_j)\in C} \left[\!\left[\|\boldsymbol{R}_i \cdot \tilde{\boldsymbol{p}}_j + \boldsymbol{t}_i - \tilde{\boldsymbol{q}}_j\|_2^2 < \tau_a \right]\!\right],$$

where $[\![\cdot]\!]$ is the Iverson bracket and τ_a is the acceptance radius.

3.5 Loss Function

The final loss consists of the coarse-fine-level loss and the overlap loss: $L = L_\text{c} + L_\text{f} + 0.5 \times L_\text{o}$. As with GeoTrans[4], we use overlap-aware circle loss L_c and negative log-likelihood loss L_f for coarse and fine level features, respectively. This also benefits us in allowing features to be closer between superpoints/patches with higher overlap ratios in coarse-level matching, rather than strictly limiting one-to-one matching. At the fine level, stricter supervise can also help eliminate mismatches. Here, the overlap region estimation is regarded as a binary classification task, and the overlap loss $L_\text{o} = \left(L_\text{o}^{\hat{P}} + L_\text{o}^{\hat{Q}}\right)/2$ is defined as:

$$L_\text{o}^{\hat{P}} = \frac{1}{|\hat{P}|} \sum_{i=1}^{|\hat{P}|} \bar{\boldsymbol{o}}_{\hat{\boldsymbol{p}}_i} \log\left(\boldsymbol{o}_{\hat{\boldsymbol{p}}_i}\right) + \left(1 - \bar{\boldsymbol{o}}_{\hat{\boldsymbol{p}}_i}\right) \log\left(1 - \boldsymbol{o}_{\hat{\boldsymbol{p}}_i}\right),$$

where $\boldsymbol{o}_{\hat{\boldsymbol{p}}_i}$ represents the predicted overlap score, and the ground truth label $\bar{\boldsymbol{o}}_{\hat{\boldsymbol{p}}_i}$ of superpoint $\hat{\boldsymbol{p}}_i$ is defined based on whether it is in the ground-truth coarse matches set C^*.

$$\bar{\boldsymbol{o}}_{\hat{\boldsymbol{p}}_i} = \begin{cases} 1, & \text{if } i \in C^*(x, \cdot), \\ 0, & \text{otherwise.} \end{cases}$$

The reverse loss $L_\text{o}^{\hat{Q}}$ and ground truth label $\bar{\boldsymbol{o}}_{\hat{\boldsymbol{p}}_i}$ are computed in the same way.

4 Experiments

In this section, we evaluate OAAFormer on indoor 3DMatch/3DLoMatch benchmarks (Subsection 4.1), the outdoor KITTI odometry benchmark (Subsection 4.2), and synthetic ModelNet/ModelLoNet benchmarks (Subsection 4.3). For the coarse-level matching module, we repeatedly alternate between the geometric self-attention module[4] and the vanilla cross-attention module[22] by setting $N_\text{c} = 3$ and then pass through the overlap region detection module. Regarding the threshold θ_m, we observe that $\theta_\text{m} = 0.05$ is safe to limit the number of superpoint matches to be within the range of $[256, 512]$. For k-nearest neighbors, we find that $k = 3$ achieves the best results. For fine-level matching, we also interleave the linear self-/cross-attention module by setting $N_\text{f} = 3$ to enhance feature discrimination. In the proposed efficient registration module, $\sigma_\text{s} = 10$ is used to enhance the distinctiveness of correspondences, with $k = 20$ for establishing the minimum consensus set, and the number of seeds N_s set to 30% of the total sampled correspondence count.

4.1 Indoor Benchmark: 3DMatch

4.1.1 Dataset and Metrics

Dataset. 3DMatch[9] is a collection of 62 scenes, of

which we employ 46 scenes for training, 8 for validation, and 8 for testing. We utilize the training data preprocessed by [2] and conduct evaluations on both the 3DMatch and 3DLoMatch benchmarks. The former features a 30% overlap, while the latter exhibits low overlap in the range of 10% to 30%. To assess robustness to arbitrary rotations, we follow [14] to create rotated benchmarks, where full-range rotations are independently applied to the two frames of each point cloud pair.

Metrics. We follow [3, 4] to employ three metrics for evaluation: 1) inlier ratio (IR), which computes the ratio of putative correspondences with a residual distance smaller than a threshold (i.e., 0.1 m) under the ground-truth transformation; 2) feature matching recall (FMR), which calculates the fraction of point cloud pairs with an IR exceeding a threshold (i.e., 5%); and 3) registration recall (RR), which quantifies the fraction of point cloud pairs that are accurately registered (i.e., with a root mean square error, RMSE < 0.2 m).

4.1.2 Correspondence Results

We begin by comparing the results of OAAFormer with the recent state-of-the-art methods in Table 1, and then proceed to analyze the impact of varying the number of correspondences in Tables 2–5. Notably, our method excels in terms of FMR, outperforming all baselines significantly, particularly in the case of 3DLoMatch. This implies a substantial increase in the likelihood of achieving correct registration with our robust pose estimator in low-overlap scenarios, where we consistently find more than 5% inliers. Furthermore, for IR, our approach exhibits even more substantial improvements, surpassing all benchmarks by over 10% on 3DMatch and more than 7% on 3DLoMatch. It is worth mentioning that our method maintains a stable performance even when the number of correspondences varies. Additionally, due to our incorporation of rotational invariance position information during fine-level matching, we perform admirably on the rotated benchmarks.

4.1.3 Registration Results

As Table 1 shows, the primary metric related to the ultimate objective of point cloud registration is RR. In all the tables of the paper, the boldfaced numbers are the best results, and the underlined ones are the second best. For this metric, we compute the

Table 1. Evaluation Results on 3DMatch and 3DLoMatch

Method	FMR (%)				IR (%)				RR (%)			
	3DMatch		3DLoMatch		3DMatch		3DLoMatch		3DMatch		3DLoMatch	
	Original	Rotated	Original	Rotated	Original	Rotated	Original	Rotated	Origin	Rotated	Original	Rotated
SpinNet[12]	97.4	97.4	75.5	75.2	48.5	48.7	25.7	25.7	88.8	93.2	58.2	61.8
Predator[2]	96.6	96.2	78.6	73.7	58.0	52.8	26.7	22.4	89.0	92.0	59.8	58.6
CoFiNet[3]	98.1	97.4	83.1	78.6	49.8	46.8	24.4	21.5	89.3	92.0	67.5	62.5
YOHO[14]	98.2	97.8	79.4	77.8	64.4	64.1	25.9	23.2	90.8	92.5	65.2	66.8
RIGA[16]	97.9	98.2	85.1	84.5	68.4	68.5	32.1	32.1	89.3	93.0	65.1	66.9
Lepard[26]	98.0	97.4	83.1	79.5	58.6	53.7	28.4	24.4	91.7	84.9	62.5	49.0
GeoTrans[4]	97.9	97.8	88.3	85.8	71.9	68.2	43.5	40.0	92.0	92.0	75.0	71.8
Ours	**98.6**	**98.2**	**89.8**	**89.5**	**82.9**	**79.6**	**50.1**	**48.2**	**94.2**	**93.8**	**77.2**	**76.0**

Table 2. Evaluation Results on 3DMatch (Original) with a Varying Number of Correspondences

Method	FMR (%)					IR (%)					RR (%)				
	5 000	2 500	1 000	500	250	5 000	2 500	1 000	500	250	5 000	2 500	1 000	500	250
PMatch[10]	95.0	94.3	92.9	90.1	82.9	36.0	32.5	26.4	21.5	16.4	78.4	76.2	71.4	67.6	50.8
FCGF[11]	97.4	97.3	97.0	96.7	96.6	56.8	54.1	48.7	42.5	34.1	85.1	84.7	83.3	81.6	71.4
D3Feat[1]	95.6	95.4	94.5	94.1	93.1	39.0	38.8	40.4	41.5	41.8	81.6	84.5	83.4	82.4	77.9
SpinNet[12]	97.6	97.2	96.8	95.5	94.3	47.5	44.7	39.4	33.9	27.6	88.6	86.6	85.5	83.5	70.2
Predator[2]	96.6	96.6	96.5	96.3	96.5	58.0	58.4	57.1	54.1	49.3	89.0	89.9	90.6	88.5	86.6
YOHO[14]	98.2	97.6	97.5	97.7	96.0	64.4	60.7	55.7	46.4	41.2	90.8	90.3	89.1	88.6	84.5
CoFiNet[3]	98.1	98.3	98.1	98.2	98.3	49.8	51.2	51.9	52.2	52.2	89.3	88.9	88.4	87.4	87.0
GeoTrans[4]	97.9	97.9	97.9	97.9	97.6	71.9	75.2	76.0	82.2	85.1	92.0	91.8	91.8	91.4	91.2
Ours	**98.6**	**98.6**	**98.5**	**98.5**	**98.2**	**82.9**	**83.1**	**83.3**	**85.5**	**86.1**	**94.2**	**94.2**	**93.8**	**93.2**	**93.0**

Table 3. Evaluation Results on 3DLoMatch (Original) with a Varying Number of Correspondences

Method	FMR (%)					IR (%)					RR (%)				
	50 00	2 500	1 000	500	250	5 000	2 500	1 000	500	250	5 000	2 500	1 000	500	250
PMatch[10]	63.6	61.7	53.6	45.2	34.2	11.4	10.1	8.0	6.4	4.8	33.0	29.0	23.3	17.0	11.0
FCGF[11]	76.6	75.4	74.2	71.7	67.3	21.4	20.0	17.2	14.8	11.6	40.1	41.7	38.2	35.4	26.8
D3Feat[1]	67.3	66.7	67.0	66.7	66.5	13.2	13.1	14.0	14.6	15.0	37.2	42.7	46.9	43.8	39.1
SpinNet[12]	75.3	74.9	72.5	70.0	63.6	20.5	19.0	16.3	13.8	11.1	59.8	54.9	48.3	39.8	26.8
Predator[2]	78.6	77.4	76.3	75.7	75.3	26.7	28.1	28.3	27.5	25.8	59.8	61.2	62.4	60.8	58.1
YOHO[14]	79.4	78.1	76.3	73.8	69.1	25.9	23.3	22.6	18.2	15.0	65.2	65.5	63.2	56.5	48.0
CoFiNet[3]	83.1	83.5	83.3	83.1	82.6	24.4	25.9	26.7	26.8	26.9	67.5	66.2	64.2	63.1	61.0
GeoTrans[4]	88.3	88.6	88.8	88.6	88.3	43.5	45.3	46.2	52.9	57.7	75.0	74.8	74.2	74.1	73.5
Ours	**89.8**	**89.9**	**90.1**	**90.1**	**89.9**	**50.1**	**52.4**	**55.6**	**58.6**	**60.1**	**77.2**	**77.2**	**77.0**	**76.8**	**76.4**

Table 4. Evaluation Results on 3DMatch (Rotated) with a Varying Number of Correspondences

Method	FMR (%)					IR (%)					RR (%)				
	5 000	2 500	1 000	500	250	5 000	2 500	1 000	500	250	5 000	2 500	1 000	500	250
SpinNet[12]	97.4	97.4	96.7	96.5	94.1	48.7	46.0	40.6	35.1	29.0	93.2	93.2	91.1	87.4	77.0
Predator[2]	96.2	96.2	96.6	96.0	96.0	52.8	53.4	52.5	50.0	45.6	92.0	92.8	92.0	92.2	89.5
YOHO[14]	97.8	97.8	97.4	97.6	96.4	64.1	60.4	53.5	46.3	36.9	92.5	92.3	92.4	90.2	87.4
CoFiNet[3]	97.4	97.4	97.2	97.2	97.3	46.8	48.2	49.0	49.3	49.3	92.0	91.4	91.0	90.3	89.6
GeoTrans[4]	97.8	97.9	98.1	97.7	97.3	68.2	72.5	73.3	79.5	82.3	92.0	91.9	91.8	91.5	91.4
Ours	**98.2**	**98.2**	**98.2**	**98.1**	**98.1**	**82.9**	**82.9**	**83.3**	**83.3**	**83.5**	**93.8**	**93.8**	**93.6**	**93.6**	**93.2**

Table 5. Evaluation Results on 3DLoMatch (Rotated) with a Varying Number of Correspondences

Method	FMR (%)					IR (%)					RR (%)				
	5 000	2 500	1 000	500	250	5 000	2 500	1 000	500	250	5 000	2 500	1 000	500	250
SpinNet[12]	75.4	74.9	72.6	69.2	61.8	25.7	23.9	20.8	17.9	15.6	61.8	59.1	53.1	44.1	30.7
Predator[2]	73.7	74.2	75.0	74.8	73.5	22.4	23.5	23.0	23.2	21.6	58.6	59.5	60.4	58.6	55.8
YOHO[14]	77.8	77.8	76.3	73.9	67.3	23.2	23.2	19.2	15.7	12.1	66.8	67.1	64.5	58.2	44.8
CoFiNet[3]	78.6	78.8	79.2	78.9	79.2	21.5	22.8	23.6	23.8	23.8	62.5	60.9	60.9	59.9	56.5
GeoTrans[4]	85.8	85.7	86.5	86.6	86.1	40.0	40.3	42.7	49.5	54.1	71.8	72.0	72.0	71.6	70.9
Ours	**89.8**	**89.6**	**89.6**	**89.4**	**89.2**	**48.2**	**48.5**	**50.4**	**52.3**	**54.6**	**76.0**	**75.4**	**75.4**	**75.3**	**74.9**

transformation using RANSAC[6] with 50k iterations. OAAFormer excels in terms of RR, outperforming the competition with significant margins. Specifically, we achieve improvements of 2.2% on both the original and rotated benchmarks for 3DMatch. On the original and rotated benchmarks of 3DLoMatch, we achieve significant improvements of 2.2% and 4.2%, respectively.

Additionally, we report RR under different numbers of correspondences in Tables 2–5. It is evident that the performance of our method is remarkably stable, eliminating the need for extensive correspondence sampling as seen in previous methods aimed at performance improvement. It is worth noting that, due to its greater focus on overlapping regions and exploration of more potentially useful correspondences, our method exhibits superior performance on datasets with low overlap and large rotation.

We subsequently report the RR using RANSAC-free estimators in Table 6. We begin with weighted SVD over correspondences to solve for the alignment transformation. Thanks to high values of FMR and IR, OAAFormer achieves RR of 88.4% and 62.1% on 3DMatch and 3DLoMatch, respectively, while the results of the baseline methods deteriorate significantly. This can be explained by the fact that, on the one hand, the coarse-to-fine mechanism constrains the correspondences to specific patches rather than the global domain. On the other hand, our model further narrows down the correspondences to the overlapping region and enhances the discriminative capabilities of fine-level features.

Subsequently, we employ the local-to-global registration module (LGR)[4] and our proposed feature similarity registration (FSR) in Subsection 3.3 separately to compute the transformation. In comparison with LGR, the FSR maintains a similar time cost but significantly improves the sampling distribution, mak-

Table 6. Registration Results w/o RANSAC on Original 3DMatch and 3DLoMatch

Method	Estimator	#Samples	3DMatch	3DLoMatch
SpinNet[12]	RANSAC-50k	5 000	88.6	59.8
Predator[2]	RANSAC-50k	5 000	89.0	59.8
CoFiNet[3]	RANSAC-50k	5 000	89.3	67.5
GeoTrans[4]	RANSAC-50k	5 000	92.0	75.0
Ours	RANSAC-50k	5 000	**94.2**	**77.2**
SpinNet[12]	Weighted SVD	250	34.0	2.5
Predator[2]	Weighted SVD	250	50.0	6.4
CoFiNet[3]	Weighted SVD	250	64.6	21.6
GeoTrans[4]	Weighted SVD	250	86.5	59.9
Ours	Weighted SVD	250	**88.4**	**62.1**
CoFiNet[3]	LGR	5 000	85.5	63.2
GeoTrans[4]	LGR	5 000	91.2	73.4
Ours	LGR	5 000	**93.2**	**76.2**
CoFiNet[3]	FSR	5 000	85.8	64.2
GeoTrans[4]	FSR	5 000	91.5	73.8
Ours	FSR	5 000	**93.4**	**76.8**

Note: The time consumption for the four pose estimators is 2.344 s, 0.008 s, 0.019 s, and 0.022 s, respectively.

ing it more effective for transformation estimation and yielding higher RR. This efficient estimator delivers performance on par with the robust pose estimator such as RANSAC but with significantly lower time costs, offering over 100 times acceleration.

4.1.4 *Robustness to Noise*

To test the robustness of our method to noise, we add Gaussian noise at different levels on 3DLoMatch. Here, we primarily compare our method with Predator[2] and GeoTrans[4], where the former is a detector-based sparse matching method and the latter is a dense matching method. As shown in Fig.3, the RR of different methods under various noise levels is depicted. With the increase in noise level, the performance of all methods declines to varying degrees. Predator[2] is severely affected due to limitations in the accuracy of score estimation and the repeatability of sampled keypoints. This significantly impacts its registration success rate when noise-induced degradation in feature matching capability occurs. In contrast, compared with GeoTrans[4], our method demonstrates stronger robustness against high noise levels. This is attributed to our proposed soft matching mechanism, which exploits more potentially valuable correspondences and effectively eliminates mismatches through an overlapping detection module. Additionally, a linear attention module enhances the feature representation capability. Therefore, our method exhibits superior robustness even under high noise levels.

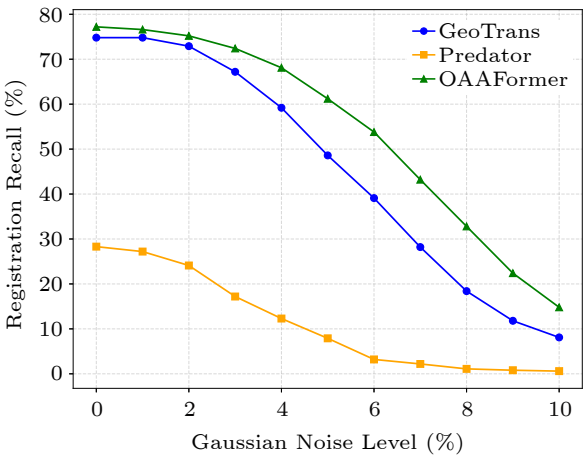

Fig.3. Registration results with various noise levels on 3DLoMatch.

4.1.5 *Ablation Studies*

To gain a more comprehensive understanding of the individual modules within our method, we conduct a series of ablation studies. Following the methodology outlined in [4], we introduce the metric, patch inlier ratio (PIR), to measure the fraction of patch matches with actual overlap. Additionally, we introduce another metric, patch overlap precision (POP), to assess the precision of patches within the actual overlap. It is worth noting that the metrics FMR and IR are reported with correspondences, while RANSAC[6] is employed for the registration process.

To investigate the effectiveness of the overlap detection module (ODM), we compare it with the MLP-directly[2] module (MLP) in Table 7. Leveraging the attention mechanism, our module has the capability to model the global overlap position, allowing for better perception of the overlap region. With a well-designed re-weighted prediction module, we obtain more

Table 7. Ablation Study of Overlap Detection Module

Method	3DMatch					3DLoMatch				
	POP	PIR	FMR	IR	RR	POP	PIR	FMR	IR	RR
MLP[2]	89.6	84.2	98.2	73.4	92.5	84.5	53.4	88.5	45.2	75.5
ODM	**93.5**	**85.6**	**98.6**	**82.9**	**94.2**	**88.1**	**54.2**	**89.8**	**50.1**	**77.2**

Note: The values of all metrics in this table are in percentage.

accurate detection results for the overlap region. As accurate overlap estimation is pivotal for eliminating mismatches, our proposed module outperforms alternatives across all metrics.

Moving forward, to explore the interactions between the soft-matching module (SMM), overlapping detection module (ODM), and linear transformer module (LTM), we conduct relevant ablation experiments in Table 8. When all modules are removed, OAAFormer reverts to GeoTrans[4] and serves as the baseline. In general, when we replace the strict matching mechanism of the original implementation with SMM, due to the introduction of a one-to-many matching paradigm, while introducing a prior for local-to-local matching, some mismatches are inevitably introduced, resulting in a decline in all metrics. The introduction of ODM and LTM, on the other hand, enhances the accuracy of coarse- and fine-level matching, respectively, and outperforms the original implementation. When all three modules are introduced simultaneously, SMM mines more potential patch matches, ODM eliminates mismatches distributed outside the estimated overlapping regions, and LTM makes the dense features of the overlapping region more discriminative, achieving the best performance.

To better elucidate the impact of each module, we present qualitative results of coarse-/fine-level correspondences under different module ablations. Fig.4(a) showcases the outcomes of GeoTrans[4], which extracts a fixed number of coarse and dense correspondences. Fig.4(b) illustrates the outcomes when solely the SMM module is incorporated. Due to the introduction of one-to-many matching and adaptive threshold settings, more matches are established at the coarse-level matching stage, inevitably introducing some outliers that propagate to the fine-level matching step. However, more inliers are fortunately discerned at this stage. Fig.4(c) demonstrates that introducing the ODM module can preserve inliers while predominantly eliminating outliers introduced by the SMM module. In comparison with GeoTrans[4], which samples a fixed number of coarse correspondences, we can adaptively sample fewer correspondences in low-

Table 8. Ablation Study of Main Modules

SMM	ODM	LTM	3DMatch				3DLoMatch			
			PIR (%)	FMR (%)	IR (%)	RR (%)	PIR (%)	FMR (%)	IR (%)	RR (%)
×	×	×	86.1	97.9	71.9	92.0	54.9	88.3	43.5	75.0
√	×	×	82.7	97.4	68.0	91.3	46.4	86.1	38.1	73.5
×	√	×	**86.4**	98.1	73.6	92.7	**55.3**	88.7	44.8	75.5
×	×	√	86.1	98.4	79.2	93.4	54.9	89.3	46.4	75.8
√	√	√	85.6	**98.6**	**82.9**	**94.2**	54.4	**89.8**	**50.1**	**77.2**

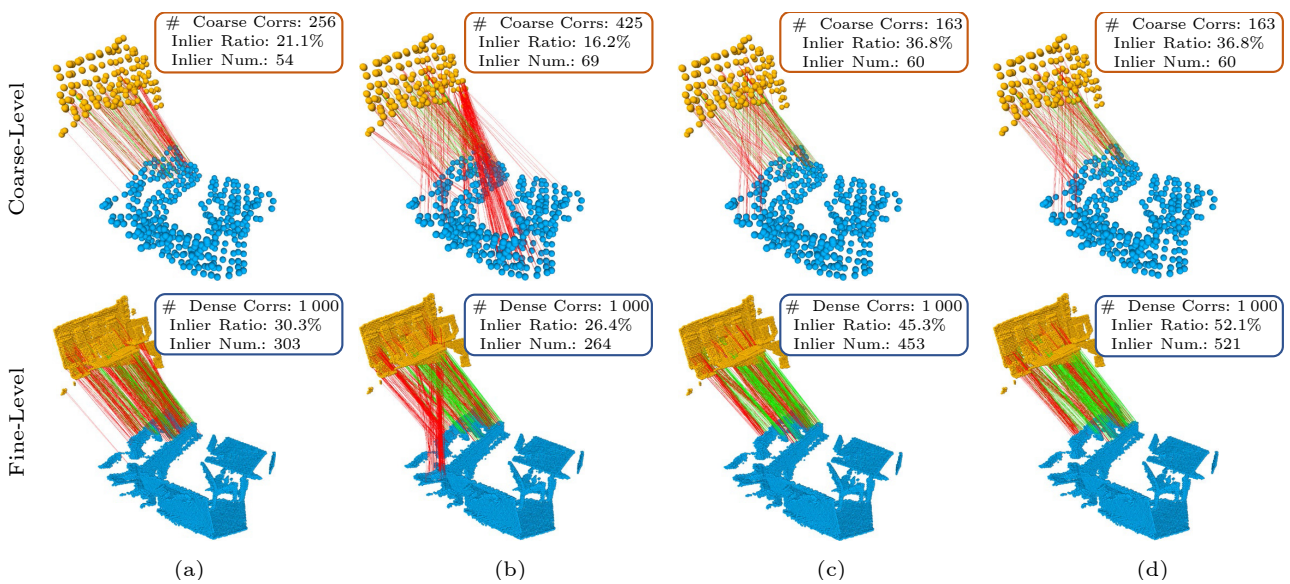

Fig.4. Qualitative results of coarse-/fine-level correspondences under different module ablations. (a) Results of GeoTrans[4]. (b) Results with SMM. (c) Results with SMM+ODM. (d) Results with SMM+ODM+LTM. Green/red lines indicate inliers/outliers. # Coarse Corrs: the number of coarse correspondense.

overlap scenes, thereby enhancing PIR and diminishing unnecessary interference in the fine-level matching stage. Fig.4(d) illustrates that introducing the LTM module notably enhances the feature matching capability in the overlapping regions, thereby further refining the matching capability at the fine-level matching step.

In addition, we replace the relative position embedding[28] with the absolute position embedding[22] in the linear attention module and conduct relevant ablation experiments. As shown in Table 9, in the context of the rotated version benchmark within the 3DMatch/3DLoMatch dataset, it is evident that the inclusion of relative position embedding resulted in superior performance. This observation suggests that incorporating relative positional information not only assists the neural network in effectively modeling distant spatial relationships but also enhances the network's capacity to discriminate between features within regions that are otherwise similar. Furthermore, it contributes to the augmentation of feature rotation invariance, thereby strengthening the network's robustness in handling variations in rotational transformations.

Table 9. Ablation Study of Position Embedding

Method	3DMatch (Rotated)			3DLoMatch (Rotated)		
	FMR (%)	IR (%)	RR (%)	FMR (%)	IR (%)	RR (%)
Absolute[22]	98.0	80.2	93.2	88.4	43.8	75.2
Relative[28]	**98.2**	**82.9**	**93.8**	**89.8**	**48.2**	**76.0**

4.1.6 *Qualitative Results*

Fig.5 offers a visualization of the overlap region prediction in the coarse level and the dense correspondence results in the fine level. The overlapping region detection module excels in perceiving the global position, and the interaction module aids in determining whether superpoints are situated within the overlap region. Moreover, the linear transformer module with the relative position embedding strategy enhances the discriminative ability for dense correspondences, resulting in more reliable correspondences.

A gallery of registration and matching comparison results with state-of-the-art methods is shown in Fig.6. It is evident that our method can establish more accurate correspondences across a broader spectrum of domains, yielding robust registration outcomes.

4.2 Outdoor Benchmark: KITTI

4.2.1 *Dataset and Metrics*

Dataset. The KITTI odometry dataset[30] comprises 11 sequences of LiDAR-scanned outdoor driving scenarios. For training, we adhere to the setup of [2, 4], utilizing sequences 0–5, while sequences 6–7 are reserved for validation, and sequences 8–10 are designated for testing. In line with the approach described in [2], we refine the ground-truth poses using ICP, and restrict the evaluation to point cloud pairs that are within a maximum distance of 10 meters.

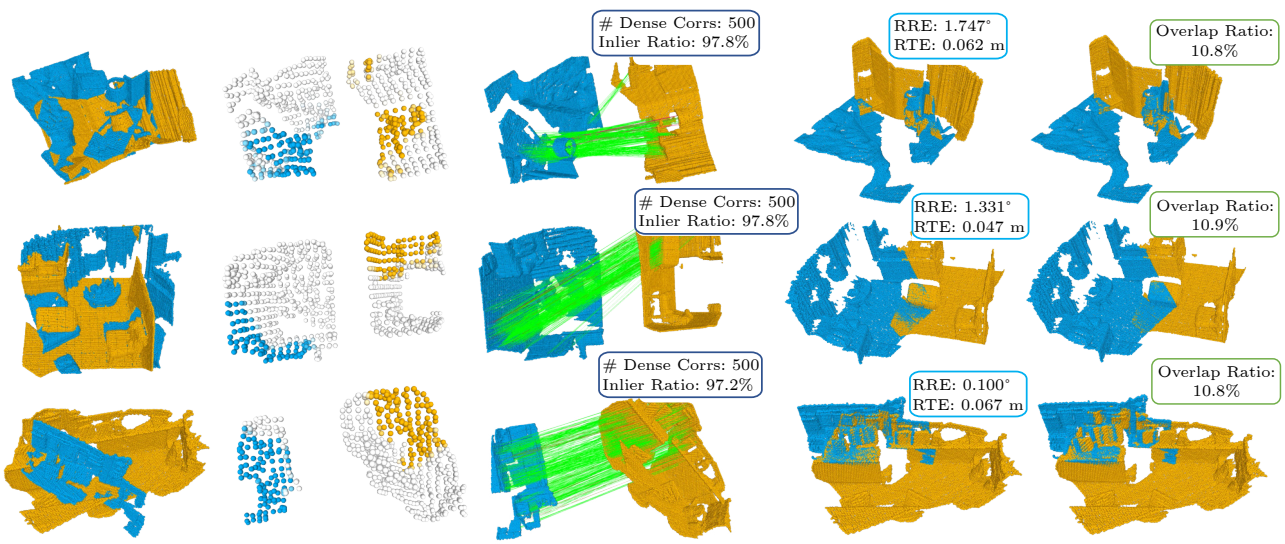

Fig.5. Qualitative results on 3DLoMatch. (a) Input point cloud. (b) Visualization of predicted overlap region. (c) Correspondence results. (d) Registration results. Green/red lines indicate inliers/outliers.

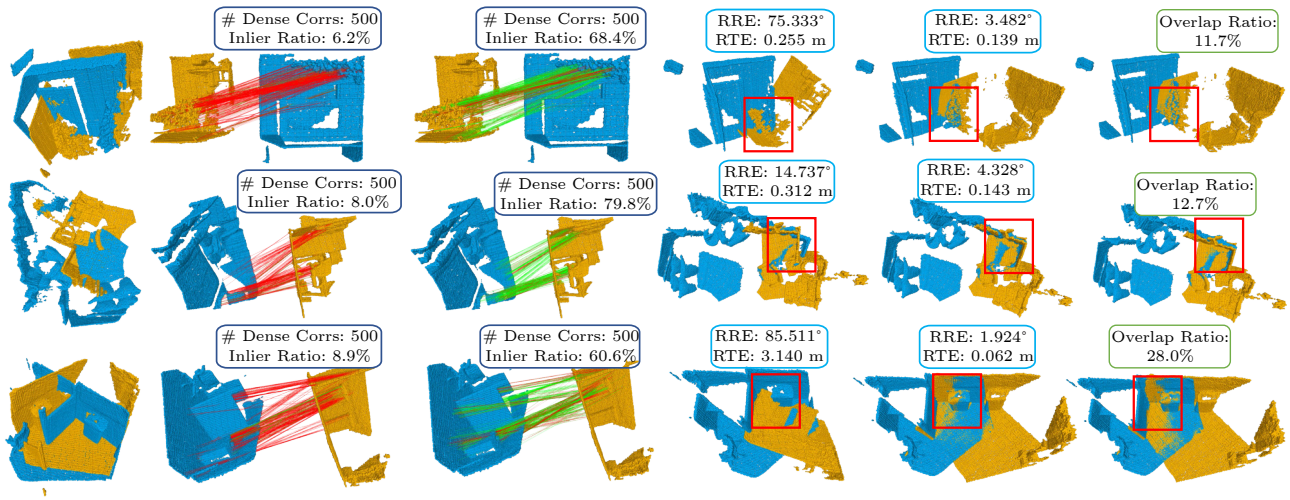

Fig.6. Qualitative comparison results on 3DLoMatch. GeoTrans[4] serves as the baseline. (a) Input point cloud. (b) Correspondence results of GeoTrans[4] (c) Correspondence results of OAAFormer. (d) Registration results of GeoTrans[4]. (e) Registration results of OAAFormer. (f) Ground-truth. Green/red lines indicate inliers/outliers.

Metrics. We adhere to the evaluation metrics established by [2, 4], which include the following: 1) relative rotation error (RRE), which quantifies the geodesic distance between the estimated and ground-truth rotation matrices; 2) relative translation error (RTE), which calculates the Euclidean distance between the estimated and ground-truth translation vectors; 3) registration recall (RR), which measures the fraction of point cloud pairs for which both RRE and RTE fall below specific thresholds, typically set as RRE $< 5°$ and RTE < 2 meters.

4.2.2 Registration Results

In Table 10 (rows 2–9), we compare OAAFormer with recent state-of-the-art methods, employing RANSAC as the pose estimator: 3DFeat-Net[31], FCGF[11], D3Feat[1], SpinNet[12], Predator[2], CoFiNet[3], and GeoTrans[4]. Our method performs comparably to these methods on RR but outperforms the baseline by approximately 0.7 cm in terms of RTE and 0.03° in RRE. We also compare our method with four RANSAC-free methods in Table 10 (rows 10–14): FMR[32], DGR[18], HRegNet[33], and GeoTrans (with LGR)[4]. Our method outperforms all the baselines significantly. Furthermore, when using FSR as a pose estimator, our method surpasses all the RANSAC-based methods.

Table 10. Registration Results on KITTI Odometry

Method	RTE (cm)	RRE (deg)	RR (%)
3DFeat-Net[31]	25.9	0.25	96.0
FCGF[11]	9.5	0.30	96.6
D3Feat[1]	7.2	0.30	**99.8**
SpinNet[12]	9.9	0.47	99.1
Predator[2]	6.8	0.27	**99.8**
CoFiNet[3]	8.2	0.41	**99.8**
GeoTrans[4]	7.4	0.27	**99.8**
Ours (RANSAC)	**6.6**	**0.24**	**99.8**
FMR[32]	~66	1.49	90.6
DGR[18]	~32	0.37	98.7
HRegNet[33]	~12	0.29	99.7
GeoTrans (LGR)[4]	6.8	0.24	**99.8**
Ours (FSR)	**6.0**	**0.21**	**99.8**

4.3 Synthetic Benchmark: ModelNet

4.3.1 Dataset and Metrics

Dataset. ModelNet comprises 12 311 CAD models of synthetic objects spanning 40 distinct categories. We adhere to the practice of employing 5 112 samples for training, 1 202 samples for validation, and 1 266 samples for testing. Similar to [2, 4], we conduct evaluations under two partial overlap scenarios: ModelNet, characterized by an average pairwise overlap of 73.5%, and ModelLoNet, exhibiting a lower average overlap of 53.6%.

Metrics. We adhere to the methodology outlined in [2, 4] for performance evaluation, employing three key metrics: 1) RRE, 2) RTE (with definitions consistent with those in Subsection 4.2), and 3) Chamfer distance (CD), which quantifies the chamfer distance between two registered scans.

4.3.2 Registration Results

In Table 11, we conduct a comparative analysis of

Table 11. Registration Results on ModelNet and ModelLoNet

Method	ModelNet			ModelLoNet		
	RRE (deg)	RTE (cm)	CD (cm)	RRE (deg)	RTE (cm)	CD (cm)
Predator[2]	1.739	0.019	0.000 89	5.235	0.132	0.008 3
Ours (RANSAC)	**1.484**	**0.016**	**0.000 81**	**4.143**	**0.091**	**0.004 4**
PointNetLK[34]	29.725	0.297	0.023 50	48.567	0.507	0.036 7
OMNet[35]	2.947	0.032	0.001 50	6.517	0.129	0.007 4
DCP-v2[36]	11.975	0.171	0.011 70	16.501	0.300	0.026 8
RPM-Net[37]	1.712	0.018	0.000 85	7.342	0.124	0.005 0
RegTR[27]	1.473	0.014	0.000 78	3.930	0.087	0.003 7
Ours (FSR)	**1.366**	**0.012**	**0.000 74**	**3.884**	**0.074**	**0.003 2**

OAAFormer against state-of-the-art RANSAC-based methods (rows 3 and 4) and RANSAC-free methods (rows 5–10). Notably, a few RANSAC-free methods are optimized primarily for ModelNet, and these models exhibit rapid performance deterioration in real-world scenarios. In contrast, OAAFormer demonstrates a substantial performance advantage over all baseline methods across all metrics, whether in the context of high overlap (ModelNet) or low overlap (ModelLoNet) scenarios.

5 Conclusions

In this paper, we enhanced the coarse-to-fine matching mechanism through a series of strategies. The key enhancements include: 1) the development of a soft matching module to preserve valuable correspondences among superpoints, 2) the introduction of an overlapping region detection module for the elimination of mismatches, and 3) the incorporation of a region-wise attention module with linear complexity to bolster the discriminative capabilities of the extracted features. Furthermore, we proposed a technique to accelerate the prediction process by carefully selecting limited but representative correspondences with high-confidence. Compared with RANSAC, our method achieved a 100 times acceleration. Additionally, we conducted extensive experiments on multiple benchmarks. Notably, on the challenging 3DLoMatch benchmark, our method improved the inlier ratio by 7% and the registration recall by 2%–4%. This fully demonstrates the superior performance and robustness of our method.

Conflict of Interest The authors declare that they have no conflict of interest.

References

[1] Bai X Y, Luo Z X, Zhou L, Fu H B, Quan L, Tai C L. D3Feat: Joint learning of dense detection and description of 3D local features. In *Proc. the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp.6358–6366. DOI: 10.1109/CVPR42600.2020.00639.

[2] Huang S Y, Gojcic Z, Usvyatsov M, Wieser A, Schindler K. PREDATOR: Registration of 3D point clouds with low overlap. In *Proc. the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp.4265–4274. DOI: 10.1109/CVPR46437.2021.00425.

[3] Yu H, Li F, Saleh M, Busam B, Ilic S. CoFiNet: Reliable coarse-to-fine correspondences for robust point cloud registration. In *Proc. the 35th Conference on Neural Information Processing Systems*, Dec. 2021, pp.23872–23884.

[4] Qin Z, Yu H, Wang C J, Guo Y L, Peng Y X, Xu K. Geometric transformer for fast and robust point cloud registration. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp.11133–11142. DOI: 10.1109/CVPR52688.2022.01086.

[5] Cheng X, Lin H Z, Wu X Y, Yang F, Shen D. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. arXiv: 2109.04290, 2021. https://doi.org/10.48550/arXiv.2109.04290, Jun. 2024.

[6] Fischler M A, Bolles R C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981, 24(6): 381–395. DOI: 10.1145/358669.358692.

[7] Johnson A E, Hebert M. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1999, 21(5): 433–449. DOI: 10.1109/34.765655.

[8] Rusu R B, Blodow N, Beetz M. Fast point feature histograms (FPFH) for 3D registration. In *Proc. the 2009 IEEE International Conference on Robotics and Automation*, May 2009, pp.3212–3217. DOI: 10.1109/ROBOT.2009.5152473.

[9] Zeng A, Song, S R, Nießner M, Fisher M, Xiao J X, Funkhouser T. 3DMatch: Learning local geometric descriptors from RGB-D reconstructions. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Jul. 2017, pp.199–208. DOI: 10.1109/CVPR.2017.29.

[10] Gojcic Z, Zhou C F, Wegner J D, Wieser A. The perfect match: 3D point cloud matching with smoothed densities. In *Proc. the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2019, pp.5540–5549.

DOI: 10.1109/CVPR.2019.00569.

[11] Choy C, Park J, Koltun V. Fully convolutional geometric features. In *Proc. the 2019 IEEE/CVF International Conference on Computer Vision*, Oct. 2019, pp.8957–8965. DOI: 10.1109/ICCV.2019.00905.

[12] Ao S, Hu Q Y, Yang B, Markham A, Guo Y L. SpinNet: Learning a general surface descriptor for 3D point cloud registration. In *Proc. the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2021, pp.11748–11757. DOI: 10.1109/CVPR46437.2021.01158.

[13] Thomas H, Qi C R, Deschaud J E, Marcotegui B, Goulette F, Guibas L. KPConv: Flexible and deformable convolution for point clouds. In *Proc. the 2019 IEEE/CVF International Conference on Computer Vision*, Oct. 27–Nov. 2, 2019, pp.6410–6419. DOI: 10.1109/ICCV.2019.00651.

[14] Wang H P, Liu Y, Dong Z, Wang W P. You only hypothesize once: Point cloud registration with rotation-equivariant descriptors. In *Proc. the 30th ACM International Conference on Multimedia*, Oct. 2022, pp.1630–1641. DOI: 10.1145/3503161.3548023.

[15] Wang H P, Liu Y, Hu Q Y, Wang B, Chen J G, Dong Z, Guo Y L, Wang W P, Yang B S. RoReg: Pairwise point cloud registration with oriented descriptors and local rotations. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2023, 45(8): 10376–10393. DOI: 10.1109/TPAMI.2023.3244951.

[16] Yu H, Hou J, Qin Z, Saleh M, Shugurov I, Wang K, Busam B, Ilic S. RIGA: Rotation-invariant and globally-aware descriptors for point cloud registration. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2024, 46(5): 3796–3812. DOI: 10.1109/TPAMI.2023.3349199.

[17] Myatt D R, Torr P H S, Nasuto S J, Bishop J M, Craddock R. NAPSAC: High noise, high dimensional robust estimation—It's in the bag. In *Proc. the British Machine Vision Conference*, Sept. 2022, pp.1–10. DOI: 10.5244/C.16.44.

[18] Choy C, Dong W, Koltun V. Deep global registration. In *Proc. the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp.2511–2520. DOI: 10.1109/CVPR42600.2020.00259.

[19] Bai X Y, Luo Z X, Zhou L, Chen H K, Li L, Hu Z Y, Fu H B, Tai C L. PointDSC: Robust point cloud registration using deep spatial consistency. In *Proc. the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2021, pp.15854–15864. DOI: 10.1109/CVPR46437.2021.01560.

[20] Zhou Q J, Sattler T, Leal-Taixé L. Patch2Pix: Epipolar-guided pixel-level correspondences. In *Proc. the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2021, pp.4667–4676. DOI: 10.1109/CVPR46437.2021.00464.

[21] Peyré G, Cuturi M . Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 2019, 11(5/6): 355–607. DOI: 10.1561/2200000073.

[22] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I. Attention is all you need. In *Proc. the 31st International Conference on Neural Information Processing Systems*, Dec. 2017, pp.6000–6010.

[23] Katharopoulos A, Vyas A, Pappas N, Fleuret F. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proc. the 37th International Conference on Machine Learning*, Jul. 2020, pp.5156–5165.

[24] Zaheer M, Guruganesh G, Dubey A, Ainslie J, Alberti C, Ontanon S, Pham P, Ravula A, Q. F. Wang, Yang L, Ahmed A. Big bird: Transformers for longer sequences. In *Proc. the 34th International Conference on Neural Information Processing Systems*, Dec. 2020, Article No. 1450.

[25] Wu C H, Wu F Z, Qi T, Huang Y F, Xie X. Fastformer: Additive attention can be all you need. arXiv: 2108.09084, 2021. https://doi.org/10.48550/arXiv.2108.09084, Jun. 2024.

[26] Li Y, Harada T. Lepard: Learning partial point cloud matching in rigid and deformable scenes. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp.5544–5554. DOI: 10.1109/CVPR52688.2022.00547.

[27] Yew Z J, Lee G H. REGTR: End-to-end point cloud correspondences with transformers. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp.6667–6676. DOI: 10.1109/CVPR52688.2022.00656.

[28] Su J L, Lu Y, Pan S F, Murtadha A, Wen B, Liu Y F. RoFormer: Enhanced transformer with rotary position embedding. arXiv: 2104.09864, 2021. https://doi.org/10.48550/arXiv.2104.09864, Jun. 2024.

[29] Leordeanu M, Hebert M. A spectral technique for correspondence problems using pairwise constraints. In *Proc. the 10th IEEE International Conference on Computer Vision*, Oct. 2005, pp.1482–1489. DOI: 10.1109/ICCV.2005.20.

[30] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proc. the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp.3354–3361. DOI: 10.1109/CVPR.2012.6248074.

[31] Yew Z J, Lee G H. 3DFeat-Net: Weakly supervised local 3D features for point cloud registration. In *Proc. the 15th European Conference on Computer Vision*, Sept. 2018. pp.630–646. DOI: 10.1007/978-3-030-01267-0_37.

[32] Huang X S, Mei G F, Zhang J. Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences. In *Proc. the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp.11363–11371. DOI: 10.1109/CVPR42600.2020.01138.

[33] Pais G D, Ramalingam S, Govindu V M, Nascimento J C, Chellappa R, Miraldo P. 3DRegNet: A deep neural network for 3D point registration. In *Proc. the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp.7191–7201. DOI: 10.1109/CVPR42600.2020.00722.

[34] Li X Q, Pontes J K, Lucey S. PointNetLK revisited. In *Proc. the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2021, pp.12758–12767. DOI: 10.1109/CVPR46437.2021.01257.

[35] Xu H, Liu S C, Wang G F, Liu G H, Zeng B. OMNet:

Learning overlapping mask for partial-to-partial point cloud registration. In *Proc. the 2021 IEEE/CVF International Conference on Computer Vision*, Oct. 2021, pp.3112–3121. DOI: 10.1109/ICCV48922.2021.00312.

[36] Wang Y, Solomon J. Deep closest point: Learning representations for point cloud registration. In *Proc. the 2019 IEEE/CVF International Conference on Computer Vision*, Oct. 27–Nov. 2, 2019, pp.3522–3531. DOI: 10.1109/ICCV.2019.00362.

[37] Yew Z J, Lee G H. RPM-Net: Robust point matching using learned features. In *Proc. the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp.11821–11830. DOI: 10.1109/CVPR42600.2020.01184.

Jun-Jie Gao received his M.S. degree in electronic science and technology from Shandong Normal University, Jinan, in 2020. Currently, he is pursuing his Ph.D. degree in the School of Computer Science and Technology, Shandong University, Qingdao, majoring in computer science and technology. His research interests include computer vision, computer graphics, and deep learning.

Qiu-Jie Dong received his M.S. degree from the Fujian Institute of Research on the Structure of Matter, Chinese Academy of Sciences, Fuzhou, in 2020. Now, he is pursuing his Ph.D. degree in the School of Computer Science and Technology, Shandong University, Qingdao, majoring in computer technology. His research interests include computer graphics and deep learning.

Rui-An Wang received his B.S. degree from the School of Computer and Software, Hohai University, Xuzhou, in 2021. Now, he is pursuing his M.S. degree in the School of Computer Science and Technology, Shandong University, Qingdao, majoring in computer technology. His research interests include computer graphics and deep learning.

Shuang-Min Chen received her Ph.D. degree from Ningbo University, Ningbo, in 2018. She is currently a lecturer at the School of Information and Technology, Qingdao University of Science and Technology. Her research interests include computer graphics and computational geometry.

Shi-Qing Xin is a professor in the School of Computer Science and Technology, Shandong University, Qingdao. He received his Ph.D. at Zhejiang University, Hangzhou, in 2009. After that, he worked as a research fellow at Nanyang Technological University, Singapore, for three years. His research interests include various geometry processing algorithms, especially geodesic computation approaches and Voronoi/power tessellation methods.

Chang-He Tu is a professor in the School of Computer Science and Technology, Shandong University, Qingdao. He received his B.S., M.S., and Ph.D. degrees from Shandong University, Jinan, in 1990, 1993, and 2003, respectively. His research interests include computer graphics and robotics.

Wenping Wang received his Ph.D. degree in computer science from the University of Alberta, Edmonton, in 1992. He is a professor of computer science at Texas A&M University, Texas. His research interests include computer graphics, computer visualization, computer vision, robotics, medical image processing, and geometric computing.