IGReg: Image-Geometry-Assisted Point Cloud Registration via Selective Correlation Fusion

Zongyi Xu, Xinqi Jiang, Xinyu Gao, Rui Gao, Changjun Gu, Qianni Zhang, Weisheng Li, Xinbo Gao*, *Fellow, IEEE*

***Abstract*—Point cloud registration suffers from repeated patterns and low geometric structures in indoor scenes. The recent transformer utilises attention mechanism to capture the global correlations in feature space and improves the registration performance. However, for indoor scenarios, global correlation loses its advantages as it cannot distinguish real useful features and noise. To address this problem, we propose an image-geometry-assisted point cloud registration method by integrating image information into point features and selectively fusing the geometric consistency with respect to reliable salient areas. Firstly, an Intra-Image-Geometry fusion module is proposed to integrate the texture and structure information into the point feature space by the cross-attention mechanism. Initial corresponding superpoints are acquired as salient anchors in the source and target. Then, a selective correlation fusion module is designed to embed the correlations between the salient anchors and points. During training, the saliency location and selective correlation fusion modules exchange information iteratively to identify the most reliable salient anchors and achieve effective feature fusion. The obtained distinctive point cloud features allow for accurate correspondence matching, leading to the success of indoor point cloud registration. Extensive experiments are conducted on 3DMatch and 3DLoMatch datasets to demonstrate the outstanding performance of the proposed approach compared to the state-of-the-art, particularly in those geometrically challenging cases such as repetitive patterns and low-geometry regions.**

***Index Terms*—multimodal point cloud registration, low-geometry area, repetitive patterns**

I. Introduction

With the pervasion of 3D capturing sensors, the acquisition of point clouds has become more convenient than ever. Many industries benefit from the utilization of point clouds, such as autonomous driving [1]–[3], robotics [4], virtual reality [5], and shape modeling [6]. Since the view range of 3D sensors is usually limited, it is often required to register partial point clouds into a complete view in the applications. Thus, point cloud registration has become a fundamental problem for many tasks and draws a lot of attention in recent years [7]–[9].

This work is supported in part by National Natural Science Foundation of China (Grant No. 62206033, No. 62221005, No. U22A2096), the Natural Science Foundation of Chongqing (No. cstc2020jcyj-msxmX0855, No.cstc2021ycjh-bgzxm0339), the Chongqing Postdoctoral Research Special Funding Project (No. 2021XM2044), and the Chongqing Technology Innovation and Application Development Major Project (No. CSTB2023TIAD-STX0016). (Corresponding author: Xinbo Gao.)

Zongyi Xu (email: xuzy@cqupt.edu.cn), Xinqi Jiang, Xinyu Gao, Rui Gao, Changjun Gu, Weisheng Li, and Xinbo Gao (email: gaoxb@cqupt.edu.cn) are with Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China.

Qianni Zhang is with the Department of Electrical Engineering and Computer Science, Queen Mary Univeristy of London, London, UK, E1 4NS.

Traditional point cloud registration mainly relies on correspondence search and transformation estimation iteratively using classical optimization strategies [10]–[12]. Although the traditional registration methods generalize well to unknown scenes, they are prone to be affected by noise, outliers, partial overlap, and various densities. Such imperfections are prevalent in real-scanned point clouds.

With deep learning flourished, correspondence search can be obtained by the learned features [13], [14] and the transformation is achieved by one-step estimation (e.g. RANSAC [15]) without iteration. Based on learned features, a set of keypoints in the source and target point clouds are detected and matched to perform registration [13], [16]. However, as repetitive patterns and low-geometry areas are very common and sometimes occupy the majority of the areas in certain scenes, the extracted features are lack of discriminative power. In such cases, it is non-trivial to extract accurate point correspondences between the source and target point clouds. As shown in Figure 1, the source and target contain multiple sofas which are similar in appearance in the non-overlapping areas. The repetitive patterns are prone to produce incorrect correspondences. Also, accurate correspondences are hard to acquire for the low-geometry parts in the overlap region, such as floors that are composed of flat and smooth surface, as few geometric features can be extracted. These issues pose great challenges for locating accurate point correspondences for reliable registration and they are particularly prominent in indoor scenarios.

Recent methods in both 2D and 3D domains are proposed to enhance the discriminative power of feature representation. GeoTransformer adopts attention mechanism to merge global contexts into features for better superpoint matching [17]. Yu et al. aggregate the features of original points and superpoint for better region matching [18]. Zhang et al. propose a Gated NetVLAD by adopting a gating scheme to automatically estimate the weight for each residue vector [19]. However, weak geometry and repetitive patterns commonly occupy a major proportion of the point clouds, especially for indoor scenarios. Merely relying on geometry information is insufficient to extract accurate and distinct features for reliable registration in those challenging cases.

Thus, in this paper, we propose a multimodal point cloud registration framework that jointly adopts the image and geometry information to enrich the point cloud feature. To further enhance the discriminative power of point features, a selective correlation fusion module is proposed to merge the geometric correlations like the distance and angles concerning the reliable salient regions. Specifically, with the input

image and point cloud, the image and point cloud features are extracted by the image backbone and KPConv. During feature extraction, the point clouds are downsampled into superpoints and the associated features are jointly learned. In order to improve the richness of features, an Intra-Image-Geometry (IIG) fusion module is designed to fuse the image and geometry information. In the IIG fusion module, the 2D image and 3D point clouds are first aligned by projecting the point clouds on the 2D plane with the extrinsic camera parameters. Due to sparsity, each superpoint corresponds to an image patch on the 2D plane. Within each superpoint, the features of pixels in the corresponding patch are fused into the superpoint geometric features by the attention mechanism. In this way, the geometric features are enriched by the image information. Based on the image-enriched superpoint features, a saliency location module is then applied to select a set of reliable superpoint correspondences in the overlap region as salient anchors in the source and target point clouds. Non-maximum suppression is utilised to ensure the salient anchors are sparsely distributed and representative. Given the salient anchors, it is possible to enhance the superpoint feature for accurate matching by selectively fusing correlations between the superpoints and the salient anchors. The correlations including the superpoint-anchor distances and angles are incorporated by the structure cross-attention technique. To capture the most effective salient anchors in the overlap features which later become distinct through selective fusion, the positions of salient anchors and superpoint features are iteratively updated. This iterative process plays a crucial role in obtaining accurate superpoint correspondences. Finally, the point correspondences are established to generate the final transformation by the pose estimator.

We list our contributions as follows:

- We propose a multimodal point cloud registration framework, named *IGReg*, that enhances point cloud features by merging image textures and selectively fusing the correlations between superpoints in an iterative manner.
- An attention-based Intra-Image-Geometry (IIG) fusion module is designed to fully merge the image and point cloud information without introducing extra noise.
- A selective correlation fusion (SCF) method is proposed to incorporate the correlations between the anchors and superpoints for feature enhancement.

The remainder of this research is organized as follows. The related work is discussed in Section II. Section III presents the proposed method in detail. The experimental evaluation results are demonstrated in Section IV. And Section V concludes the paper.

II. Related Work

Feature-based Registration. The methods with feature-based registration mainly focus on four aspects: feature extraction, keypoint detection, outlier removal, and pose estimation. Qi et al. propose PointNet and PointNet++ successively [20], [21]. Although they provide a reference for the feature extraction of point clouds, these two methods do not consider the geometric structure features of point clouds. Deng et al. propose PPFNet [22]. The point pair feature (PPF) is combined with PointNet to improve the robustness of features against noise. The 3DFeatNet proposed by Yew and Lee et al. utilises a weakly supervised deep network to solve the difficult problem of accurate labeling of point cloud data and improve feature quality [23]. Gojcic et al. propose 3DSmoothNet, which utilises the Siamese network architecture to encode smoothed density value voxelization (SDV) [24]. Wang et al. design the edge convolution (EdgeConv) operation and construct DGCNN to capture topological information between points [25]. Thomas et al. propose KPConv to simulate operations in 2D convolutions to better capture local geometric information [26]. The 3DMatch network proposed by Zeng et al. takes voxels as input and utilises 3D convolutional neural networks to learn local geometric features [27]. Choy et al. propose FCGF which adopts sparse 3D convolution instead of traditional 3D convolution to alleviate the problem caused by point cloud sparsity [14]. SpinNet restricts the z-axis degrees of freedom through the estimated reference axis and uses

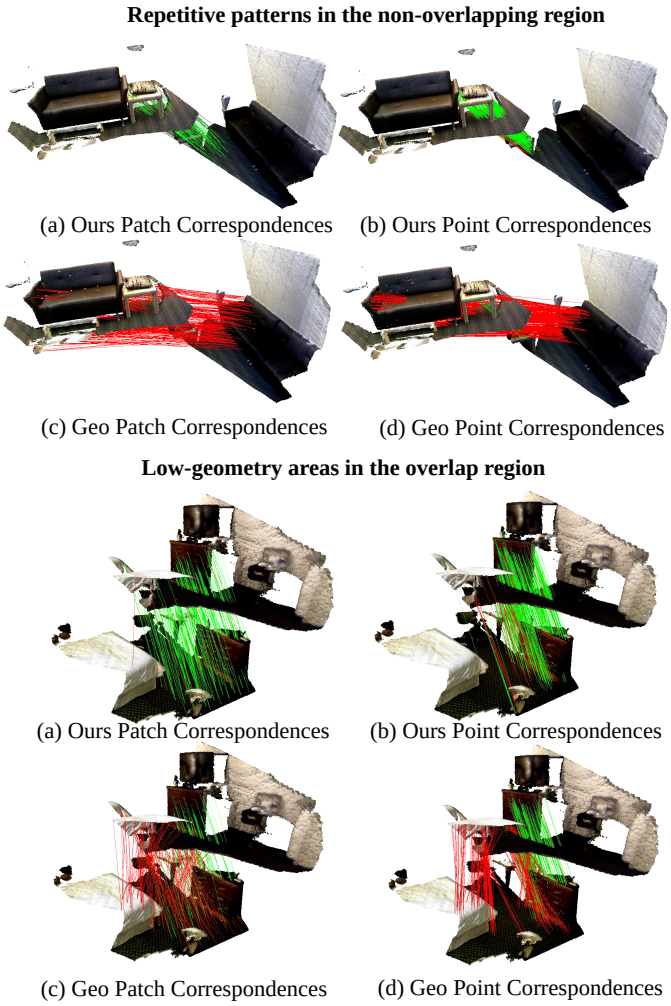


Fig. 1. The patch and point correspondences of the proposed method and GeoTransformer in the geometrically challenging cases: (1) Repetitive patterns (sofa parts) in the non-overlapping areas and (2) Low-geometry areas (floor parts) in the overlap region. Green lines are correct correspondences and red lines are incorrect ones.

spherical voxelization to eliminate the XY plane rotational degrees of freedom to extract features with high robustness [28]. D3feat proposed by Bai et al. uses a U-Net network composed of KPConv to detect keypoints while extracting point cloud features and uses a density-invariant saliency score to alleviate the effect of density on saliency [13]. Huang et al. improve the probability of correct detection by detecting the possibility of points in overlapping regions while extending the task to low-overlapping scenes [29]. Bai et al. propose PointDSC that adds the spatial geometric consistency constraints in the traditional method to the network and uses the neural network to extract the features of the correspondences, to select a set of spatially consistent point pairs [30]. DCP (deep closest point) and DeepVCP estimate relative pose by weighted confidence [31], [32]. While IDAM (iterative distance-aware similarity matrix) and DGR (Deep global registration) use weighted SVD to solve rigid transformations by selecting high-confidence point pairs [33], [34]. Although the aforementioned methods make efforts to extract robust and descriptive point cloud features. However, they often only consider the self-point cloud when embedding structures, which is insufficient for point clouds with low-geometry areas and repetitive patterns.

Direct Registration Methods. PointNetLK calculates the Jacobian matrix with the relative pose obtained by PointNet and the inverse compositional formula, and finally uses a differentiable Lucas & Kanade (LK) algorithm to calculate the rigid transformation [35]. Deng et al. input the PPF feature and point cloud into the PPF-FoldNet and PC-FoldNet networks respectively to obtain a new feature including the structure and pose, and used RelativeNet to predict the relative pose [36]. PCRNet uses a network similar to Siamese to predict the transformation matrix after splicing the global features of the two point clouds [37]. FMR draws on the idea of PointNetLK, utilises the reversible property of rigid transformation, and uses the encoder-decoder structure to supervise global features [38]. Xu et al. propose OMNet to predict overlapping masks of source and target point clouds in an iterative process, estimating rigid transformations from the global features of both through MLPs [39]. However, the direct registration approaches are usually limited in processing real scenes.

Multimodal Point Cloud Registration. Recent work starts to utilise images to assist point cloud registration. IMFNet merges the image and point cloud features using attention techniques [40]. ImLoveNet exploits an intermediate image to assist in obtaining accurate overlap region between the source and target point clouds [41]. However, without alignment between the 2D and 3D information, noise is fused to degrade the performance. PCR-CG firstly extracts the potential 2D correspondences predicted by image matching methods and then explicitly lifts the color signals into the 3D presentation with KPConv bottleneck [42]. Although PCR-CG uses an explicit 2D-3D projection method to reduce the introduction of noise, the approach is limited by the performance of image matching methods and is prone to being misguided by incorrect 2D prior information, which is fatal in low-overlap rate scenarios.

III. The Proposed Method

A. Problem Statement

Given two point clouds $\mathcal{P} = \{\mathbf{p}_i \in \mathbb{R}^3 \mid i = 1, \dots, N\}$ and $\mathcal{Q} = \{\mathbf{q}_i \in \mathbb{R}^3 \mid i = 1, \dots, M\}$ and their corresponding images $I_P \in \mathbb{R}^{W \times H}$ and $I_Q \in \mathbb{R}^{W \times H}$ from different viewpoints with a partial overlap in a scene, our goal is to solve a transformation $\mathbf{Trans} \in SE(3)$ which aligns the point clouds into a unified coordinate system. The transformation consists of two parts: rotation $\mathbf{R} \in SO(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$. They can be obtained by the formula:

$$\underset{\mathbf{R},\mathbf{t}}{\operatorname{argmin}} \sum_{(\mathbf{p}_{x_i}, \mathbf{q}_{y_i}) \in \mathcal{C}} \|\mathbf{q}_{y_i} - (\mathbf{R} \cdot \mathbf{p}_{x_i} + \mathbf{t})\|_2^2, \tag{1}$$

where $(\mathbf{p}_{x_i}, \mathbf{q}_{y_i})$ belongs to the correspondence set $\mathcal{C}$ and represents the point-wise correspondence between the source and the target point clouds. Thus the problem of solving the transformation is turned into the problem of finding the correct correspondence set $\mathcal{C}$.

B. Method Overview

As shown in Figure 2, the whole IGReg framework consists of the following parts: (1) feature extraction part where point clouds and images are input into backbones to extract point and image features and the point clouds are downsampled into superpoints; (2) the IIG fusion part where the superpoints are projected to corresponding image patches and the attention mechanism is used to fuse the image and superpoint features for enrichment; (3) the selective correlation fusion part where the anchor location and correlation fusion (i.e. the distance and angles) iteratively interact and finally the most reliable correlations are integrated into the superpoint features; (4) matching $\&$ registration with the enhanced features.

C. Image and Point Cloud Feature Extraction

We first feed the original source and target point cloud $\mathcal{P} \in \mathbb{R}^{|\mathcal{P}| \times 3}$ and $\mathcal{Q} \in \mathbb{R}^{|\mathcal{Q}| \times 3}$ into the shared KPConv backbone network, by which the raw point clouds are sampled into superpoints, denoted as $\hat{\mathcal{P}} = \{\hat{\mathbf{p}}_i\}_{i=1}^{|\hat{\mathcal{P}}|}$ and $\hat{\mathcal{Q}} = \{\hat{\mathbf{q}}_j\}_{j=1}^{|\hat{\mathcal{Q}}|}$ respectively, and superpoint features are extracted, represented as $\mathbf{F}^{\hat{\mathcal{P}}} \in \mathbb{R}^{|\hat{\mathcal{P}}| \times C}$ and $\mathbf{F}^{\hat{\mathcal{Q}}} \in \mathbb{R}^{|\hat{\mathcal{Q}}| \times C}$. The downsampling and feature extraction process can be formulated as $f : \left(\mathbb{R}^{|\mathcal{P}| \times 3}, \mathbb{R}^{|\mathcal{Q}| \times 3}\right) \rightarrow \left(\mathbb{R}^{|\hat{\mathcal{P}}| \times C}, \mathbb{R}^{|\hat{\mathcal{Q}}| \times C}\right)$, where C is the superpoint feature dimension.

We adopt the ResUNet-50 [43] backbone to extract image features. For the source and target images $I_P \in \mathbb{R}^{W \times H}$ and $I_Q \in \mathbb{R}^{W \times H}$, the acquired features are represented as $\mathbf{F}^{I_P} \in \mathbb{R}^{W \times H \times C}$ and $\mathbf{F}^{I_Q} \in \mathbb{R}^{W \times H \times C}$.

D. Intra-Image-Geometry Fusion Module

As point clouds only contain geometry information, it is nontrivial to distinguish different objects with similar structures. Thus, we propose an Intra-Image-Geometry (IIG) Fusion Module to integrate image and geometry information to enrich the point cloud features.

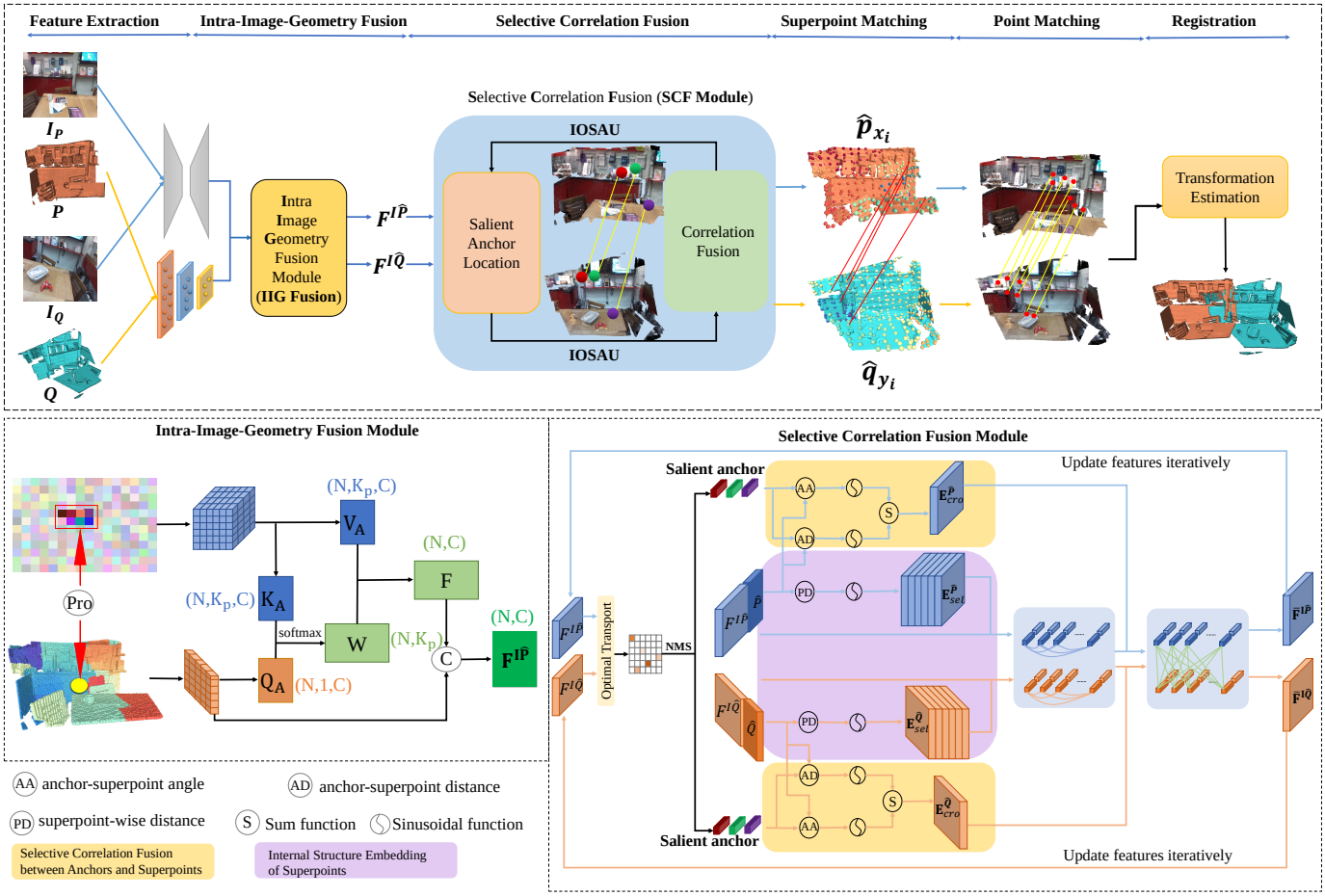


Fig. 2. Overview of the proposed registration method. Feature extraction is first performed to acquire the superpoint features and image features. Then the Intra-Image-Geometry Fusion (IIG Fusion) is proposed to integrate geometry information and corresponding image to enrich the point cloud features. Next the Selective Correlation Fusion Module (SCF Module) consists of Salient Anchor Location and Correlation Fusion Module. Specifically, the anchor location module is to capture the salient anchors with rich and discriminative geometric and image information. The Correlation Fusion Module integrates the structure information between salient anchors and superpoints for feature enhancement. The enhanced feature and the anchor positions are updated in an iterative manner to acquire the most effective anchors and distinct features. In the superpoint matching stage, superpoint correspondences are obtained under the guidance of enhanced superpoint features. In the point matching stage, the superpoint correspondences are extended to the point correspondences within the region and then the final transformation is obtained.

Projection. As the source and target point clouds are partially overlapped and sometimes with extremely low overlap. The integration of global image information like IMFNet [40] introduces noise. Thus, we first project the point cloud onto the 2D plane to locate a rough alignment. Given the i^{th} 3D point $\mathbf{p}_i$ in the point cloud, its pixel position $(\mathbf{w}_i, \mathbf{h}_i)$ in the image can be located with Eq. 2,

$$(\mathbf{w}_i, \mathbf{h}_i, 1) = h(\mathbf{K}_{int} \cdot (\mathbf{R}_{ext} \cdot \mathbf{p}_i + \mathbf{t}_{ext})) \tag{2}$$

where $(\mathbf{R}_{ext}, \mathbf{t}_{ext})$ is the extrinsic camera parameters including the rotation and translation matrices; $\mathbf{K}_{int} \in \mathbb{R}^{3\times3}$ is the intrinsic camera parameters; and $h(\cdot)$ represents the homogeneous function. As the superpoint is downsampled from the original points, each superpoint can be aligned with a set of image pixels which are denoted as an image patch.

Image-geometry fusion. Given the superpoints $\hat{\mathcal{P}}$ in the source point cloud and its corresponding image patches $I_{\hat{\mathcal{P}}}$ as an example, the cross-attention technique is adopted to extensively merge the pixel feature into the superpoints. After the projection model, the feature vector of superpoints and image patches is unsqueezed to $\mathbf{F}^{\hat{\mathcal{P}}} \in \mathbb{R}^{N\times1\times C}$ and $\mathbf{F}^{I_{\hat{\mathcal{P}}}} \in \mathbb{R}^{N\times K_p\times C}$. $\mathbf{F}^{\hat{\mathcal{P}}}$ is the feature of superpoints and $\mathbf{F}^{I_{\hat{\mathcal{P}}}}$ is the feature of corresponding pixels. Each superpoint is related to K_p pixels and N is the number of superpoints. $\mathbf{F}^{\hat{P}}$ is considered as the query array $\mathbf{Q_A} \in \mathbb{R}^{N\times1\times C}$. $\mathbf{F}^{I_P}$ is regarded as key array $\mathbf{K_A} \in \mathbb{R}^{N\times K_p\times C}$ and value array $\mathbf{V_A} \in \mathbb{R}^{N\times K_p\times C}$. C is the feature dimension. The weight matrix $\mathbf{W} \in \mathbb{R}^{N\times K_p} = softmax(\frac{\mathbf{Q_A}\mathbf{K_A}^T}{\sqrt{C}})$ represents the weight of each pixel's texture information that could contribute to describing the corresponding superpoint. Mathematically, the superpoint texture feature $\mathbf{F} \in \mathbb{R}^{N\times C}$ can be calculated with:

$$\mathbf{F} = MLP(\mathbf{W} * \mathbf{V_A}) \tag{3}$$

Finally, we concatenate the texture feature $\mathbf{F}$ and geometric feature $\mathbf{F}^{\hat{P}}$ to fuse multimodal information. Mathematically, the fused encoder feature $\mathbf{F}^{I\hat{\mathcal{P}}}$ is calculated as

$$\mathbf{F}^{I\hat{\mathcal{P}}} = (\mathbf{F}^{\hat{P}}_{ij})cat(\mathbf{F}_{ij}), \forall i \in [1, N], \forall j \in [1, C] \tag{4}$$

E. Selective Correlation Fusion Module (SCF Module)

1) Salient Anchor Location.: With the enriched features for superpoints, we can locate the sparse and shape-preserved anchor correspondences between the source and target point clouds. Once reliable anchor correspondences are obtained, they can be used as references to embed the correlations into each superpoint feature. In this way, the mismatching of those ambiguous superpoints can be eliminated. Thus, in this *Salient Anchor Location* module, our goal is to obtain those superpoints with discriminative features as salient anchors. Since the anchors should be consistent in the feature space of both source and target point clouds in order to embed consistent geometric information into point cloud features, the resultant salient anchors should be located in the overlap region.

Given the image-geometry fused superpoint features (i.e. $\mathbf{F}^{I\hat{\mathcal{P}}} \in \mathbb{R}^{|\hat{\mathcal{P}}|\times C}$ and $\mathbf{F}^{I\hat{\mathcal{Q}}} \in \mathbb{R}^{|\hat{\mathcal{Q}}|\times C}$) of source and target point clouds, the initial anchor correspondences can be chosen as those with high confidence in the similarity matrix $\mathbf{S}$ that is computed with L2 norm. As shown in Figure 2, in order to obtain the structure-preserved anchor correspondences that are distributed sparsely, we abandon the traditional top-K selection way that consecutively chooses the several highest-confidence superpoint matches to avoid the positions of picked anchor correspondences are concentrated. Instead, Non-Maximum Suppression (NMS) [44] is adopted to ensure the spatial uniformity and sparsity of selected anchor correspondences. The similarity scores between the superpoints of source and target point clouds are computed and ranked from the highest confidence to the lowest. Non-Maximum Suppression is applied with the fixed radius of r_{nms} around each superpoint. After choosing the highest-confidence superpoint, all the correspondences within the Euclidean distance of r_{nms} are removed. In the remaining superpoints, we pick the highest-confidence correspondence as the second anchor correspondence and remove those located within the radius of r_{nms}. The process is repeated until we acquire K initial anchor correspondences which are defined as follows.

$$\hat{\mathcal{C}} = \{(\hat{\mathbf{a}}_{x_k}, \hat{\mathbf{b}}_{y_k}) \in \mathbb{R}^{1\times 6} \mid k = 1, \dots, K\}, \tag{5}$$

where the anchor point $\hat{\mathbf{a}}_{x_k}$ is in correspondence with $\hat{\mathbf{b}}_{y_k}$, $x_k \in \{1, \dots \left|\hat{\mathcal{P}}\right|\}$ and $y_k \in \{1, \dots \left|\hat{\mathcal{Q}}\right|\}$ and K is the total number of anchor correspondences. The whole process of anchor location is illustrated in Algorithm 1.

2) Correlation Fusion Module: In this part, we first perform internal structure embedding within each point cloud. In order to enhance the discriminative ability of the superpoint features, we further selectively integrate the correlations between the anchors and superpoints at the stage of interaction between source and target point clouds.

Internal Structure Embedding of Superpoints. The internal structure within the point cloud contains abundant context information and can benefit the descriptive ability of superpoint features. Here, we adopt the self-attention mechanism to help perceive the context information and embed the internal structure into superpoint features.

Algorithm 1: Salient Anchor Location

Input: radius parameters r, similarity matrix $\mathbf{S}$, number of anchor correspondences K
Output: $\hat{\mathcal{C}} = \{(\hat{\mathbf{a}}_{x_k}, \hat{\mathbf{b}}_{y_k}) \mid k = 1, \dots, K\}$

$\hat{\mathcal{C}} = \phi$
while $k \in \{1, 2, ..., K\}$ **do**
  $S_{x,y} = max(\mathbf{S})$;
  $\hat{\mathcal{C}} = \hat{\mathcal{C}} \cup (\hat{\mathbf{a}}_x, \hat{\mathbf{b}}_y)$;
  for $i \in \{1, 2, ..., row(\mathbf{S})\}$ **do**
    if $||\hat{\mathbf{a}}_x - \hat{\mathbf{p}}_i||_2 < r$ **then**
      remove($\mathbf{S}_{i,-}$)
    end
  end
  for $j \in \{1, 2, ..., col(\mathbf{S})\}$ **do**
    if $||\hat{\mathbf{b}}_y - \hat{\mathbf{q}}_j||_2 < r$ **then**
      remove($\mathbf{S}_{-,j}$)
    end
  end
end

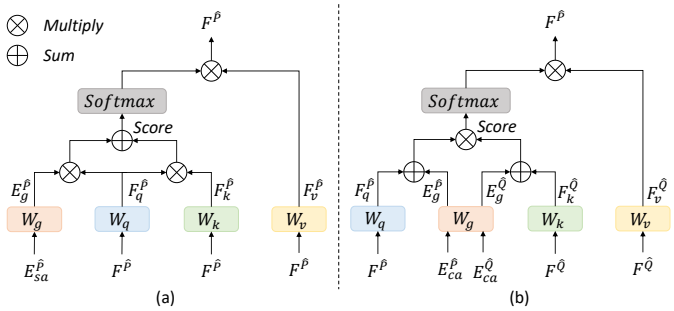


Fig. 3. The computation graph of self-attention (a) and cross-attention (b).

The source point cloud is taken as an example to explain the embedding process. Given a superpoint $\hat{\mathbf{p}}_i$ in the source, the superpoint-wise distance embedding from $\hat{\mathbf{p}}_i$ to another superpoint $\hat{\mathbf{p}}_j$ is computed using Eq. 6:

$$\mathbf{E}_{sa}^{\hat{\mathcal{P}}(i,j)} = \mathbf{W}_d f(d(\hat{\mathbf{p}}_i, \hat{\mathbf{p}}_j)/\sigma_d), \tag{6}$$

where $\hat{\mathbf{p}}_j \in \hat{\mathcal{P}} - \{\hat{\mathbf{p}}_i\}$ is the superpoint except $\hat{\mathbf{p}}_i$; $d(\hat{\mathbf{p}}_i, \hat{\mathbf{p}}_j) = \|\hat{\mathbf{p}}_i - \hat{\mathbf{p}}_j\|_2^2$; $f(\cdot)$ represents a sinusoidal function that maps a scalar to a high-dimensional feature; and σ_d is a coefficient to adjust distance sensitivity. It is usually set between 0.1m and 0.5m.

Given the superpoint feature $\mathbf{F}^{\hat{\mathcal{P}}}$ and the distance embedding $\mathbf{E}_{sa}^{\hat{\mathcal{P}}}$, the self-attention technique is used to merge the superpoint features and the superpoint-wise distance. Each superpoint and the superpoint-wise distance can be regarded as the token and relative positional embedding respectively in self-attention. Three learnable matrices $\mathbf{W}_q, \mathbf{W}_k$, and $\mathbf{W}_v$ are utilised to map each token $\mathbf{F}^{\hat{\mathcal{P}}}$ to its own query, keys, and values which are denoted by $\mathbf{F}_q^{\hat{\mathcal{P}}}, \mathbf{F}_k^{\hat{\mathcal{P}}}$, and $\mathbf{F}_v^{\hat{\mathcal{P}}}$, respectively. $\mathbf{W}_g$ is used to map $\mathbf{E}_{sa}^{\hat{\mathcal{P}}}$ to $\mathbf{E}_g^{\hat{\mathcal{P}}}$. Then $\mathbf{F}_q^{\hat{\mathcal{P}}}$, $\mathbf{F}_k^{\hat{\mathcal{P}}}$, and $\mathbf{E}_g^{\hat{\mathcal{P}}}$ are used to calculate the attention matrix $\mathbf{Score}$. Adding superpoint-wise distance can help the information fusion pay

attention to the position of each token. The calculation of each value in the attention matrix is as described in Eq. 7:

$$\mathbf{Score}_{(i,j)} = \frac{\mathbf{F}_q^{\hat{\mathcal{P}}_i} \cdot (\mathbf{F}_k^{\hat{\mathcal{P}}_j} + \mathbf{E}_g^{\hat{\mathcal{P}}_{(i,j)}})^{\mathrm{T}}}{\sqrt{C}} \tag{7}$$

The output of superpoint feature merged with the geometric information $\mathbf{F}^{\hat{\mathcal{P}}}$ is calculated by

$$\mathbf{F}^{\hat{\mathcal{P}}} = softmax(\mathbf{Score}) \cdot \mathbf{F}_v^{\hat{\mathcal{P}}} \tag{8}$$

In order to express the above calculation process more clearly, we use Fig.3 (a) to show the calculation graph of self-attention.

Correlation Fusion between Anchors and Superpoints. In the process of interacting the information across two point clouds, instead of aggregating all the correlations between superpoints, we selectively embed the correlations with respect to the selected anchors to further enhance the distinctiveness of superpoint features. There are several advantages of embedding the correlations with respect to salient anchors rather than all the superpoints: 1) In the case where some similar scenes of source and target are not in the overlapping area and the overlap ratio is rather low, interacting all the superpoint information across two point clouds unavoidably introduce noise and disturbance; 2) Selectively embedding the correlations between sparse and correct anchors and superpoints is useful to avoid the symmetric, upside-down and front-and-back flip problems in the low-geometry areas; 3) The computational cost is largely reduced. The process is shown in Fig. 3 (b). The features and correlations of the source and target superpoints are explicitly fused using a cross-attention mechanism.

For a superpoint $\hat{\mathbf{p}}_n$ in the source, the anchor point set is denoted as $\{\hat{\mathbf{a}}_i\}_{i=1}^{K}$. We calculate the correlations of the distance between the superpoint and the i_{th} anchor point using Eq. 9.

$$\rho_i = d(\hat{\mathbf{p}}_n, \hat{\mathbf{a}}_i), \tag{9}$$

The $f(\cdot)$ function is then used to map the anchor-superpoint distance to high-dimensional features like self-attention. Then, according to the anchor correspondence scores, they are weighted and added to form the anchor-superpoint distance embedding:

$$\mathbf{Ed}_{ca}^{\hat{\mathbf{p}}_n} = \sum_{i=1}^{K} \mathbf{W}_d f(\rho_i/\sigma_d), \tag{10}$$

where $\mathbf{Ed}_{ca}^{\hat{\mathcal{P}}_n}$ is the anchor-superpoint distance embedding; K is the number of anchor points in the source point clouds; $\mathbf{W}_d \in \mathbb{R}^{C\times C}$ is the projection matrix of the distance embedding; σ_d is used to adjust distance sensitivity as above.

Besides the correlation of distance, the angles between the superpoints and the anchors are also fused into the features. As shown in Figure 4, we take the superpoint $\hat{\mathbf{p}}_n$ as the vertex. The vertex and two anchor points compose the superpoint-anchor angles and are defined as:

$$\theta_k(\hat{\mathbf{p}}_n, \hat{\mathbf{a}}^l, \hat{\mathbf{a}}^s) = deg(\hat{\mathbf{a}}^l - \hat{\mathbf{p}}_n, \hat{\mathbf{a}}^s - \hat{\mathbf{p}}_n), \tag{11}$$

where θ_k represents the k_{th} anchor-superpoint angle; $\hat{\mathbf{a}}^l$ denotes the l_{th} anchor in the source and $l = \{1, 2, ..., K\}$; $\hat{\mathbf{a}}^s$ denotes the s_{th} anchor and $s = \{1, 2, ..., K\}$; $deg(\cdot)$ is the degree function that calculates the degree of angles.

After acquiring the anchor-superpoint angles, we use the sinusoidal function $f(\cdot)$ to map it to a high-dimensional feature. Then the confidence score weighted sum is used to obtain the anchor-superpoint angle embedding.

$$\mathbf{Ea}_{ca}^{\hat{\mathbf{p}}_n} = \sum_{k=1}^{C_K^2} \mathbf{W}_a f(\theta_k(\hat{\mathbf{p}}_n, \hat{\mathbf{a}}^l, \hat{\mathbf{a}}^s)/\sigma_\theta), \tag{12}$$

where $\mathbf{W}_a$ is the projection matrix for the angle embedding; the K is the number of anchor points and C_K^2 represents the number of combinations; σ_θ is a coefficient to adjust angle sensitivity and is usually set between 5° and 25°.

Then, we sum the anchor-superpoint distance $\mathbf{Ed}_{ca}^{\hat{\mathcal{P}}_n}$ and anchor-superpoint angle $\mathbf{Ea}_{ca}^{\hat{\mathcal{P}}_n}$ to get the final geometric embedding feature of the superpoint using Eq. 13.

$$\mathbf{E}_{ca}^{\hat{\mathbf{p}}_n} = \mathbf{Ed}_{ca}^{\hat{\mathbf{p}}_n} + \mathbf{Ea}_{ca}^{\hat{\mathbf{p}}_n}. \tag{13}$$

The above anchor-superpoint geometric structure embedding process is shown in Figure 4, the computed distances and angles between the salient anchors and superpoints are first obtained and summed up to merge the geometric information, forming the final geometric embedding. For the target superpoints, the geometric embedding, denoted as $\mathbf{E}_{ca}^{\hat{\mathbf{q}}_m}$, can be obtained in the same way.

With the extracted superpoint features $\mathbf{F}^{\hat{\mathbf{p}}_n}$ and $\mathbf{F}^{\hat{\mathbf{q}}_m}$ with KPConv and the geometric embedding $\mathbf{E}_{ca}^{\hat{\mathbf{p}}_n}$ and $\mathbf{E}_{ca}^{\hat{\mathbf{q}}_m}$ in the source and target point clouds, we fuse them using the cross attention technique. We map $\mathbf{F}^{\hat{\mathbf{p}}_n}$, $\mathbf{F}^{\hat{\mathbf{q}}_n}$ to $\mathbf{F}_q^{\hat{\mathbf{p}}_n}$, $\mathbf{F}_k^{\hat{\mathbf{q}}_n}$, and $\mathbf{F}_v^{\hat{\mathbf{q}}_n}$ with matrices $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ respectively. $\mathbf{W}_g$ is used to map $\mathbf{E}_{ca}^{\hat{\mathbf{p}}_n}$ and $\mathbf{E}_{ca}^{\hat{\mathbf{q}}_m}$ to $\mathbf{E}_g^{\hat{\mathbf{p}}_n}$ and $\mathbf{E}_g^{\hat{\mathbf{q}}_m}$ as follows.

$$\begin{aligned}
&\mathbf{F}_q^{\hat{\mathbf{p}}_n} = \mathbf{W}_q \cdot \mathbf{F}^{\hat{\mathbf{p}}_n}; \mathbf{F}_k^{\hat{\mathbf{q}}_m} = \mathbf{W}_k \cdot \mathbf{F}^{\hat{\mathbf{q}}_n};\\
&\mathbf{F}_v^{\hat{\mathbf{q}}_m} = \mathbf{W}_v \cdot \mathbf{F}^{\hat{\mathbf{q}}_n}; \mathbf{E}_g^{\hat{\mathbf{p}}_n} = \mathbf{W}_g \cdot \mathbf{E}_{ca}^{\hat{\mathbf{p}}_n};\\
&\mathbf{E}_g^{\hat{\mathbf{q}}_m} = \mathbf{W}_g \cdot \mathbf{E}_{ca}^{\hat{\mathbf{q}}_m},
\end{aligned} \tag{14}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$ and $\mathbf{W}_g \in \mathbb{R}^{C\times C}$ are the projection matrices for queries, keys, values, and geometric structure embeddings respectively. The attention score $\mathbf{Score}_{(n,m)}^{\hat{\mathcal{Q}}\to\hat{\mathcal{P}}}$ can be obtained with Eq. 15.

$$\mathbf{Score}_{(n,m)}^{\hat{\mathcal{Q}}\to\hat{\mathcal{P}}} = \frac{(\mathbf{F}_q^{\hat{\mathbf{p}}_n} + \mathbf{E}_g^{\hat{\mathbf{p}}_n}) \cdot (\mathbf{F}_q^{\hat{\mathbf{q}}_m} + \mathbf{E}_g^{\hat{\mathbf{q}}_m})^{\mathrm{T}}}{\sqrt{C}}. \tag{15}$$

After interacting with the information using cross-attention, the enhanced superpoint feature of the source point cloud $\bar{\mathbf{F}}^{\hat{\mathbf{p}}_n}$ can be calculated using Eq. 16.

$$\bar{\mathbf{F}}^{\hat{\mathbf{p}}_n} = softmax(\mathbf{Score}_{(n,m)}^{\hat{\mathcal{Q}}\to\hat{\mathcal{P}}}) \cdot \mathbf{F}_v^{\hat{\mathbf{q}}_m}. \tag{16}$$

Given a superpoint $\hat{\mathbf{q}}_m$ in the target point cloud, we can follow the above-mentioned cross-attention process and acquire the enhanced feature $\bar{\mathbf{F}}^{\hat{\mathbf{q}}_m}$.

With selective correlation fusion, we integrate the geometry consistency with respect to the salient anchors into the point cloud features. In this way, the features of those areas with weak geometry and repetitive patterns can be discriminative and can further improve the matching accuracy.

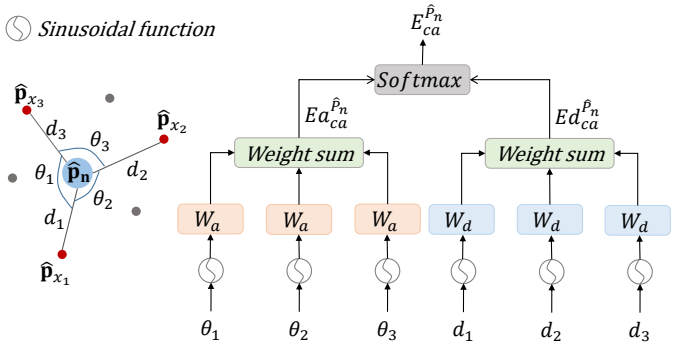


Fig. 4. An illustration of the correlation fusion and its computation.

3) Iterative-Optimization based Salient Anchor Updating (IOSAU) : As aforementioned, descriptive features are beneficial to locate the most distinctive anchors, and embedding the geometry of the distinctive anchors makes the superpoint features more discriminative. However, in the initial stage, the neural network is not well-trained and the initial anchor correspondences are not accurate and sparsely distributed enough, causing the anchor-superpoint geometry embedding cannot to capture structure information sufficiently.

In order to achieve accurate correlation fusion, we propose to update the anchor and the superpoint correspondences in an iterative manner. The ground-truth superpoint matches can promote obtaining salient anchors which are beneficial to obtain precise superpoint correspondences in turn. During the iteration, the initial wrongly-selected anchor correspondences have a less adverse effect on the features due to inconsistency. In contrast, the correct anchor correspondences with consistency are gradually enhanced during multiple iterations.

Given the initial superpoints $\hat{\mathcal{P}}$, $\hat{\mathcal{Q}}$ and the corresponding image fused superpoint features $\mathbf{F}^{I\hat{\mathcal{P}}}$, $\mathbf{F}^{I\hat{\mathcal{Q}}}$ of the input point clouds, the initial anchor correspondences $\hat{\mathcal{C}}$ are acquired by inputting them into the anchor location module. The geometric correlations of initial anchor correspondences and superpoints are incorporated using self and cross-attention mechanisms through selective correlation fusion module and the enhanced superpoint features $\bar{\mathbf{F}}^{I\hat{\mathcal{P}}}$ and $\bar{\mathbf{F}}^{I\hat{\mathcal{Q}}}$ are acquired. To update the anchor positions, the enhanced features concatenating superpoint coordinates are re-input into the anchor location and selective geometry embedding modules.

F. Matching and Registration

1) Superpoint Matching.: After the final superpoint features $\bar{\mathbf{F}}^{I\hat{\mathcal{P}}}$ and $\bar{\mathbf{F}}^{I\hat{\mathcal{Q}}}$ are obtained, we first normalize them and calculate the similarity matrix $\hat{\mathbf{S}} = \bar{\mathbf{F}}^{I\hat{\mathcal{P}}}(\bar{\mathbf{F}}^{I\hat{\mathcal{Q}}})^T/\sqrt{C}$. In this way, the problem of finding accurate superpoint correspondences is transformed into an optimal transportation problem. It is worth noting that some superpoints have no corresponding relationship, so we add a row and a column to the matrix $\hat{\mathbf{S}}$ as slack entries like CoFiNet [18]. Then we use the Sinkhorn algorithm to optimise the entire matrix $\hat{\mathbf{S}}$. We select K correspondences with the highest scores as the final superpoint matching result:

$$\hat{\mathcal{SC}} = \left\{ (\hat{\mathbf{p}}_{x_i}, \hat{\mathbf{q}}_{y_i}) \mid \hat{\mathbf{S}}_{(x_i,y_i)} \in \text{topk}(\hat{\mathbf{S}}) \right\}, \tag{17}$$

where the $\text{topk}(\hat{\mathbf{S}})$ is the function that returns the largest K entries in the matrix $\hat{\mathbf{S}}$; and x_i and y_i represent the i-th item corresponding to the index in the source point cloud and the target point cloud respectively.

Algorithm 2: IOSAU

Input: superpoint feature $\mathbf{F}^{I\hat{\mathcal{P}}}$ and $\mathbf{F}^{I\hat{\mathcal{Q}}}$, number of iterations k
Output: superpoint correspondences $\hat{\mathcal{SC}}$

while $i <= k$ **do**
  $\hat{\mathcal{C}}$ = Anchor Location($\mathbf{F}^{I\hat{\mathcal{P}}}$, $\mathbf{F}^{I\hat{\mathcal{Q}}}$)
  $\mathbf{F}^{I\hat{\mathcal{P}}}$, $\mathbf{F}^{I\hat{\mathcal{Q}}}$ = Correlation Fusion($\mathbf{F}^{I\hat{\mathcal{P}}}$, $\mathbf{F}^{I\hat{\mathcal{Q}}}$, $\hat{\mathcal{C}}$)
end
$\bar{\mathbf{F}}^{I\hat{\mathcal{P}}}$ = norm($\mathbf{F}^{I\hat{\mathcal{P}}}$); $\bar{\mathbf{F}}^{I\hat{\mathcal{Q}}}$ = norm($\mathbf{F}^{I\hat{\mathcal{Q}}}$)
superpoint similarity matrix $\hat{\mathbf{S}} = \bar{\mathbf{F}}^{I\hat{\mathcal{P}}}(\bar{\mathbf{F}}^{I\hat{\mathcal{Q}}})^T/\sqrt{C}$
$\hat{\mathbf{S}}$ = Sinkhorn($\hat{\mathbf{S}}$)
$\hat{\mathcal{SC}} = \left\{ (\hat{\mathbf{p}}_{x_i}, \hat{\mathbf{q}}_{y_i}) \mid \hat{\mathbf{S}}_{(x_i,y_i)} \in \text{topk}(\hat{\mathbf{S}}) \right\}$

2) Point Matching: After getting the superpoint correspondence, we use the decoder module of the backbone to decode the superpoint features, which are symmetric to the encoder process. In order to get point matching, we compare the point correspondences score matrix $\mathbf{S}$ within each superpoint correspondence based on the decoded point features. For the point correspondence score matrix $\mathbf{S}$ within the superpoints, we use the Sinkhorn algorithm to obtain the point correspondence set $\hat{\mathcal{PC}}_i$. A point match is achieved when its matching score is the k_p largest items in both row and column:

$$\hat{\mathcal{PC}}_i = \left\{ (\mathbf{p}_{x_i}, \mathbf{q}_{y_i}) \mid \mathbf{S}_{(x_i,y_i)} \in \text{topk}(\mathbf{S}) \right\}. \tag{18}$$

3) Transformation Estimation: Traditional point cloud descriptors usually cannot capture distinct enough features. Thus the acquired point correspondences often have a large number of outliers. In order to get accurate transformation, RANSAC is regarded as a robust estimator to compute the pose from the set of point correspondences with a high rate of outliers. However, RANSAC suffers from slow convergence. To achieve accurate and efficient transformation estimation, we first acquire the high-confidence region matches based on the proposed geometry-embedded features and then the point correspondences for the target and source point clouds are captured within the region matches. Thus, without removing the outliers like RANSAC, we estimate the transformation without iteration. For a region match, the rotation $\mathbf{R}$ and translation $\mathbf{t}$ using the i_{th} superpoint correspondence can be computed as Eq. 19.

$$\mathbf{R}_i, \mathbf{t}_i = \arg\min_{\mathbf{R},\mathbf{t}} \sum_{j}^{|\hat{\mathcal{SC}}_i|} w_j \left\| \mathbf{R} \cdot \mathbf{p}_{x_j} + \mathbf{t} - \mathbf{q}_{y_j} \right\|_2^2, \tag{19}$$

where $\hat{\mathcal{SC}}_i$ is the i_{th} superpoint correspondence; $|\hat{\mathcal{SC}}_i|$ represents the number of point correspondences in the superpoint; $(\mathbf{p}_{x_j}, \mathbf{q}_{y_j}) \in \hat{\mathcal{C}}_i$ denotes the j_{th} point correspondence, and w_j is the weight.

After calculating the transformation matrices using all the superpoints, the final transformation is obtained using Eq. 20:

$$\mathbf{R}, \mathbf{t} = \arg\max_{\mathbf{R}_i, \mathbf{t}_i} \sum_{k=1}^{|PC|} [\|\mathbf{R}_i \cdot \mathbf{p}_k + \mathbf{t}_i - \mathbf{q}_k\|_2^2 < \tau_a], \quad (20)$$

where $(\mathbf{p}_k, \mathbf{q}_k)$ are the k_{th} point correspondences in the point correspondence set PC and the τ_a is the threshold.

G. Loss Functions

The training loss function $\mathcal{L}$ consists of two parts: the loss for supervising the superpoint matching $\mathcal{L}_c$, and the loss for supervising point matching $\mathcal{L}_f$.

Superpoint Correspondence Loss. We follow [9] to employ a deformed circle loss to supervise the superpoint matching. For a superpoint in the point cloud, we first construct a local patch around the superpoint using the KNN search strategy. The overlap rate between the superpoints in the source and target point clouds is computed as follows. After the ground-truth transformation is applied, if the distance from one point in the source patch to its nearest neighbour in the target patch is less than a threshold, i.e. 0.05m in our method, these two points are considered as overlapped. If the proportion of overlapped points in the source and target patches is greater than 10%, we regard the pair of superpoints as overlapped. Otherwise, the superpoint pairs are treated as negative. Second, we select all superpoints with at least one positive relationship in $\mathcal{P}$ to form a basic set $\mathcal{P}_p$. For each superpoint $\hat{\mathbf{p}}_i \in \mathcal{P}_p$, positive superpoints in $\mathcal{Q}$ form the ε_p^i set, and negative superpoints form the ε_n^i set. Finally, the deformed circle loss on $\mathcal{P}$ is defined as:

$$\mathcal{L}_c^{\mathcal{P}} = \frac{1}{|\mathcal{P}_p|} \sum_{\hat{\mathbf{p}}_i \in \mathcal{P}_p} \log[1 + \sum_{\hat{\mathbf{q}}_j \in \varepsilon_p^i} e^{\lambda_i^j \beta_p^{i,j} \left(d_i^j - \Delta_p\right)} \cdot \sum_{\hat{\mathbf{q}}_k \in \varepsilon_n^i} e^{\beta_n^{i,k} \left(\Delta_n - d_i^k\right)}], \quad (21)$$

where d_i^j represents the Euclidean distance in the feature space and is calculated with $\left\|\mathbf{F}^{\hat{\mathcal{P}}_i} - \mathbf{F}^{\hat{\mathcal{Q}}_j}\right\|_2$; given the overlap ratio o_i^j between superpoints $\hat{\mathbf{p}}_i$ and $\hat{\mathbf{q}}_j$, λ_i^j is denoted by $\lambda_i^j = \sqrt{o_i^j}$. Like [9], the positive margins Δ_p and the negative margins Δ_n are set to 0.1 and 1.4 respectively. At the same time, for each sample, its positive weight $\beta_p^{i,j} = \gamma\left(d_i^j - \Delta_p\right)$ and negative weight $\beta_n^{i,k} = \gamma\left(\Delta_n - d_i^k\right)$ are calculated separately. $\mathcal{L}_r^{\mathcal{Q}}$ is defined in the same way with $\mathcal{L}_r^{\mathcal{P}}$.

The final superpoint matching loss consists of two parts $\mathcal{L}_r^{\mathcal{P}}$ and $\mathcal{L}_r^{\mathcal{Q}}$. It is defined as Eq. 22.

$$\mathcal{L}_c = \left(\mathcal{L}_c^{\mathcal{P}} + \mathcal{L}_c^{\mathcal{Q}}\right)/2. \quad (22)$$

Point Correspondence Loss For point matching, we use the negative log-likelihood function [45] to supervise the assignment matrix corresponding to each superpoint. In the training phase, we select N_g ground-truth superpoint matches and calculate the ground-truth point correspondences set $\mathcal{M}_i$ in the corresponding patches. Then we divide the unmatched points in the two patches into two sets of $\mathcal{I}_i$ and $\mathcal{J}_i$. The loss function for point correspondences within the region corresponding to the selected i_{th} superpoint is defined as:

$$\mathcal{L}_f^i = -\sum_{(x,y)\in\mathcal{M}_i} \log \mathbf{S}_i(x,y) - \sum_{x,m_i\in\mathcal{I}_i} \log \mathbf{S}_i(x, m_i+1) - \sum_{y\in\mathcal{J}_i} \log \mathbf{S}_i(n_i+1, y), \quad (23)$$

where $\mathbf{S}_i(x,y)$ represents the similarity between the x_{th} and the y_{th} points within the i_{th} superpoint correspondence.

The final fine correspondences loss function averages the point correspondences losses of the selected N_g superpoints and is defined in Eq. 24:

$$\mathcal{L}_f = \frac{1}{N_g} \sum_{i=1}^{N_g} \mathcal{L}_f^i. \quad (24)$$

IV. Experiments

In this section, extensive experiments are conducted to evaluate the effectiveness of our research. The experimental implementation details are first described in Sec. IV-A. Following current multimodal point cloud registration methods [42], our research is compared with state-of-the-art approaches on the datasets of 3DMatch [27] and 3DLoMatch [29]. Besides, to demonstrate the generalisability of IGReg, we directly apply the model trained on 3DMatch on the TUM RGB-D SLAM dataset and further evaluate the performance [46]. Finally, in Sec. IV-D, the ablation study is designed to verify the effectiveness of each module of the method.

A. Implementation Details

To extract the feature of the image, we first resize the RGB image to the resolution of 240×320 pixels. ResUNet-50 is used to extract the image features. The implementation is based on PyTorch [47] and trained on double NVIDIA RTX A5000 GPUs with an initial learning rate of 1e-4. The batch size is set to 1. The entire network structure is trained with Adam optimizer and its weight decay is set to 1e-6. For the dataset of 3DMatch and 3DLoMatch, each epoch decays at the rate of 0.05, and the number of epochs is set to 40. A 4-stage KPConv backbone is used in 3DMatch and 3DLoMatch. We select 3 anchor correspondences in the experiments. As for the distance sensitivity σ_d and the angle sensitivity σ_θ in the selective geometry embedding module, we set them to 0.2 and 15° respectively. Each attention module contains 4 attention heads. RANSAC, weighted SVD, and LGR are used for registration estimation.

B. Multimodal 3DMatch and 3DLoMatch Datesets

Both 3DMatch and 3DLoMatch are RGB-D datasets. Each point cloud is reconstructed from 50 consecutive frames with an RGB and a depth image. We select the first frame RGB image to construct the multimodal datasets that consist of paired images and point clouds that depict the same scenes. 3DMatch [27] consists of 62 scenes, in which 46 scenes are used for training, 8 scenes are used as a validation set, and

TABLE I
COMPARISONS OF OUR METHOD AND THE ADVANCED METHODS ON THE 3DMATCH AND 3DLOMATCH. THE EVALUATION RESULTS OF ALL METHODS ARE OBTAINED BY THE RANSAC ALGORITHM.

# Samples	3DMatch					3DLoMatch				
	5000	2500	1000	500	250	5000	2500	1000	500	250
Feature Matching Recall(%)↑										
FCGF [14]	97.4	97.3	97.0	96.7	96.6	76.6	75.4	74.2	71.7	67.3
D3Feat [13]	95.6	95.4	94.5	94.1	93.1	67.3	66.7	67.0	66.7	66.5
SpinNet [28]	97.6	97.2	96.8	95.5	94.3	75.3	74.9	72.5	70.0	63.6
Predator [29]	96.6	96.6	96.5	96.3	96.5	78.6	77.4	76.3	75.7	75.3
YOHO [48]	98.2	97.6	97.5	97.7	96.0	79.4	78.1	76.3	73.8	69.1
CoFiNet [18]	98.1	98.3	98.1	98.2	98.3	83.1	83.5	83.3	83.1	82.6
GeoTransformer [9]	97.9	97.9	97.9	97.9	97.6	88.3	88.6	**88.8**	88.6	88.3
IGReg	**98.7**	**98.7**	**98.4**	**98.4**	**98.4**	**88.8**	**88.8**	88.5	**88.6**	**88.5**
Inlier Ratio(%)↑										
FCGF [14]	56.8	54.1	48.7	42.5	34.1	21.4	20.0	17.2	14.8	11.6
D3Feat [13]	39.0	38.8	40.4	41.5	41.8	13.2	13.1	14.0	14.6	15.0
SpinNet [28]	47.5	44.7	39.4	33.9	27.6	20.5	19.0	16.3	13.8	11.1
Predator [29]	58.0	58.4	57.1	54.1	49.3	26.7	28.1	28.3	27.5	25.8
YOHO [48]	64.4	60.7	55.7	46.4	41.2	25.9	23.3	22.6	18.2	15.0
CoFiNet [18]	49.8	51.2	51.9	52.2	52.2	24.4	25.9	26.7	26.8	26.9
GeoTransformer [9]	**71.9**	75.2	76.0	82.2	85.1	**43.5**	45.3	46.2	52.9	57.7
IGReg	71.3	**78.2**	**82.9**	**84.8**	**86.1**	42.2	**48.1**	**54.2**	**57.0**	**58.9**
Registration Recall(%)↑										
FCGF [14]	85.1	84.7	83.3	81.6	71.4	40.1	41.7	38.2	35.4	26.8
D3Feat [13]	81.6	84.5	83.4	82.4	77.9	37.2	42.7	46.9	43.8	39.1
SpinNet [28]	88.6	86.6	85.5	83.5	70.2	59.8	54.9	48.3	39.8	26.8
Predator [29]	89.0	89.9	90.6	88.5	86.6	59.8	61.2	62.4	60.8	58.1
YOHO [48]	90.8	90.3	89.1	88.6	84.5	65.2	65.5	63.2	56.5	48.0
CoFiNet [18]	89.3	88.9	88.4	87.4	87.0	67.5	66.2	64.2	63.1	61.0
GeoTransformer [9]	92.0	91.8	91.8	91.4	91.2	75.0	74.8	74.2	74.1	**73.5**
IGReg	**93.6**	**93.2**	**92.4**	**92.1**	**91.3**	**76.4**	**75.8**	**75.0**	**74.5**	72.8

the remaining 8 scenes are used as the test set. For 3DMatch, the overlap rate of any to-be-registered point clouds is above 30%, and for 3DLoMatch [29], the overlap rate is between 10% and 30%.

Metrics. As with previous work, we mainly use 3 evaluation indicators to evaluate our methodology, Inlier Ratio (IR), Feature Matching Recall (FMR), and Registration Recall (RR). Specifically, IR represents the percentage of correspondences whose residual error is less than a certain threshold in the case of a real transformation, which is generally taken at $\tau_1 = 10cm$. FMR represents the proportion of point cloud pairs whose IR metrics are greater than a certain threshold, $\tau_2 = 5\%$. RR represents the ratio of registered point clouds whose RMSE is less than $0.2m$.

Comparisons to the State-of-the-Art. The experimental results of running RANSAC on 3DMatch and 3DLoMatch are shown in Table I. Following [29], we run RANSAC-50k to evaluate the transformation matrix under four sampling conditions of 250, 500, 1000, 2500, and 5000, respectively. For *Feature Matching Recall*, IGReg achieves the highest performance that reaches over 98.0% of all cases of sampled points on 3DMatch. *Inlier Ratio* improves by $1\% \sim 4.9\%$ on 3DMatch and $1.2\% \sim 8.0\%$ on 3DLoMatch, compared to the state-of-the-art GeoTransformer. For *Registration Recall*, with 5000 sampled points, our method achieves 93.6% on 3DMatch and 76.4% on 3DLoMatch, both surpassing the previous methods. This demonstrates that the proposed multimodal point cloud registration approach, i.e., IGReg, can achieve accurate point correspondences and effectively improve the registration results.

TABLE II
REGISTRATION RESULTS WITH DIFFERENT POSE ESTIMATORS ON 3DMATCH (3DM) AND 3DLOMATCH (3DLM).

Model	Estimator	# Samples	RR(%)		Times(s)		
			3DM	3DLM	Model	Pose	Total
FCGF [14]	RANSAC-50k	5000	85.1	40.1	0.119	12.394	12.513
D3Feat [13]	RANSAC-50k	5000	81.6	37.2	0.060	10.267	10.327
SpinNet [28]	RANSAC-50k	5000	88.6	59.8	97.808	0.788	98.596
Predator [29]	RANSAC-50k	5000	89.0	59.8	0.079	15.434	15.513
CoFiNet [18]	RANSAC-50k	5000	89.3	67.5	0.259	5.321	5.580
Geo [9]	RANSAC-50k	5000	92.0	75.0	0.184	4.805	4.989
IGReg	RANSAC-50k	5000	**93.6**	**76.4**	0.205	3.910	**4.115**
FCGF [14]	weighted SVD	250	42.1	3.9	0.119	0.009	0.128
D3Feat [13]	weighted SVD	250	37.4	2.8	0.060	0.009	**0.069**
SpinNet [28]	weighted SVD	250	34.0	2.5	97.808	0.008	97.816
Predator [29]	weighted SVD	250	50.0	6.4	0.079	0.010	0.089
CoFiNet [18]	weighted SVD	250	64.6	21.6	0.259	0.004	0.263
Geo [9]	weighted SVD	250	86.5	**59.9**	0.184	0.004	0.188
IGReg	weighted SVD	250	**87.1**	57.3	0.205	0.004	0.209
CoFiNet [18]	LGR	ALL	87.6	64.8	0.259	0.242	0.501
Geo [9]	LGR	ALL	91.5	74.0	0.184	0.115	**0.299**
IGReg	LGR	ALL	**93.5**	**75.1**	0.205	0.141	0.346

To further analyse the effectiveness of our method, we follow GeoTransformer to test the performance using different pose estimators, i.e., the RANSAC-50K, weighted SVD and local-to-global registration (LGR). It can be seen that IGReg achieves the best performance over all three different pose estimators on the 3DMatch dataset. For the time efficiency, we compare the average model time for feature extraction, the pose time for transformation estimation and the total time in Table II. We can see that our model time is slightly longer than GeoTranformer as additional image features are extracted using ResUNet50 and merged into geometric features with transformer, which consumes more time. However, the total time of our method based on the Ransac pose estimator is the shortest. This verifies that the correspondences located with our method are more accurate and the RANSAC can converge more quickly. For weighted SVD and LGR, our computation time is also comparable with GeoTransformer. Thus, by adding the image information for feature enhancement, we can achieve accurate and efficient registration.

Comparison with multimodal point cloud registration methods. As our method is based on geometry and image, we also compare our method with the current multimodal methods which are PCR-CG [49] and IMFNet [40]. As shown in Table III, our method outperforms PCR-CG, which proves that utilising the cross-attention mechanism to fuse the image and point cloud feature is more reasonable than initializing the point cloud feature as image features in an explicit way. In addition, our method also outperforms IMFNet mainly because there is no projection module in IMFNet so the points can not find the correct pixels to enhance the features.

Comparison of the repetitive patterns. In Figure 5, it can be seen that our approach can achieve accurate registration. The GeoTransformer finds incorrect superpoint and point correspondences in the non-overlap regions while our method

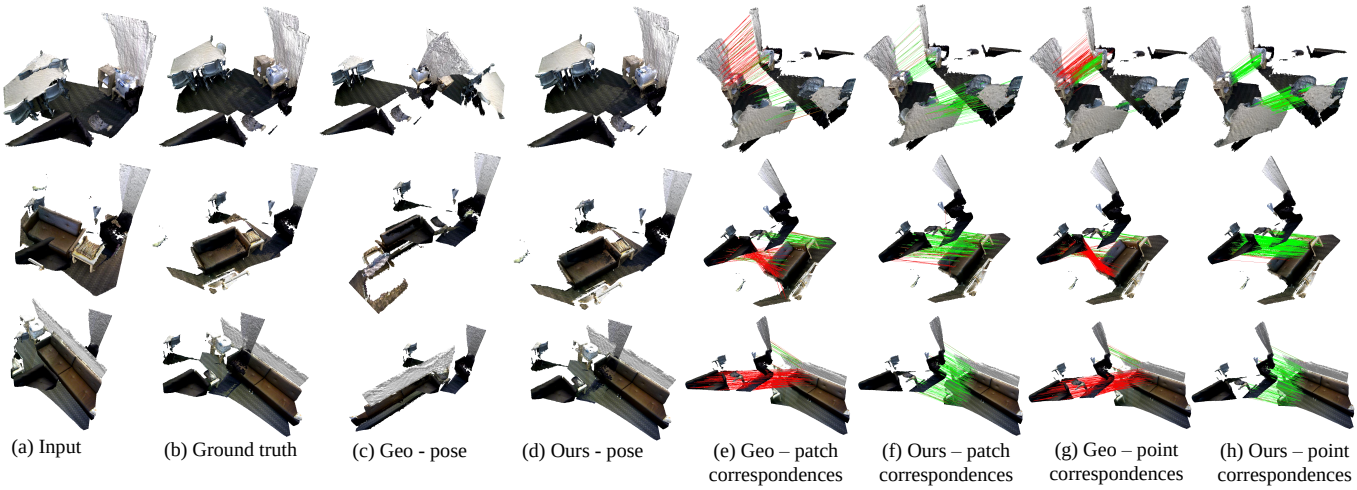


Fig. 5. Visual comparison of our method and GeoTransformer on the 3DLoMatch. The registration results, superpoint (patch) correspondences, and point correspondences are compared. Red lines are incorrect correspondences while green lines are correct ones.

TABLE III
COMPARISON WITH MULTIMODAL POINT CLOUD REGISTRATION.

Model	3DMatch		3DLoMatch	
	FMR	RR	FMR	RR
IMFNet [40]	98.5	91.0	80.5	48.4
PCR-CG [49]	97.4	89.4	80.4	66.3
IGReg	**98.7**	**93.6**	**88.8**	**76.4**

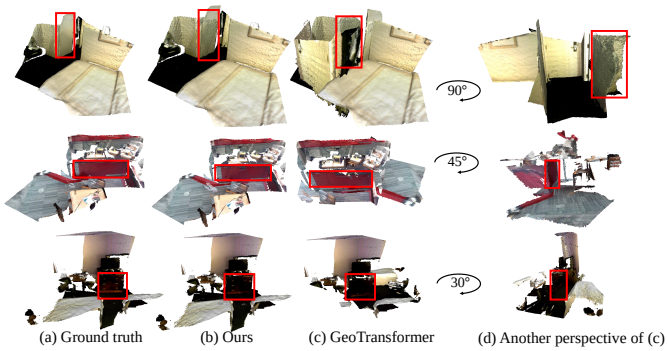


Fig. 6. The comparison of registration on low-geometry areas against GeoTransformer. The red frames represent the low-geometry overlap areas.

can remove all the incorrect correspondences. 3DLoMatch are indoor scenarios that contain a large number of repetitive walls, tables, and floors. These objects contain repetitive patterns like planes that have similar appearances to objects of different categories. Existing feature extraction methods cannot differentiate these areas, thus, as shown in the third row of column (e) in Figure 5, some superpoints of the table plane are in correspondence with those patches of the floor. In contrast, we enhance the geometry feature with image information and selective correlations with respect to the salient anchors, the features of appearance-similar patches and points can be differentiated. All the incorrect correspondences in the non-overlapping areas of our method are removed, which is also demonstrated by columns (f) and (h).

With the feature extracted by the proposed method, our results can also remove the incorrect correspondences outside the overlap region. When the non-overlapping but similar regions account for a large proportion, the correct matching is often suppressed for existing methods. For example, in the third row of Figure 5, in GeoTransformer, the superpoints and points that belong to the same category but not in the overlap region are incorrectly located as correspondences. In contrast, our method makes the features of similar regions in non-overlapping areas different after the geometric embedding based on anchor correspondences. Thus, we can find accurate correspondences even though non-overlapping of similar regions accounts for a large proportion.

Comparison of the low-geometry areas We also show the registration results on low-geometry overlap areas in Figure 6. Low-geometry overlap areas are widespread in real scenes such as the upright plane in the first two rows and horizontal planes in the third row. Finding correct correspondences within these low-geometry areas is non-trivial. It can be seen that the registration of the state-of-the-art GeoTransformer suffers from the front-back and upside-down flip problem shown in columns (c) and (d). This is because the geometric structures within the low-geometry areas are so weak that few features can be extracted to get accurate correspondences. With our proposed selective geometry embedding with respect to the salient anchors, our method can achieve superior performance.

Qualitative results. We also show more registration results with the single modal GeoTransformer and multimodal IMFNet on 3DMatch and 3DLoMatch in Figure 7. The input source and target point clouds, the ground-truth registration, the registration results of GeoTransformer, IMFNet and our method are presented in columns (a) to (e) respectively. We can see that our approach is robust to these challenging cases where a large amount of appearance-similar areas exist under different overlap ratios. These qualitative results further verify the effectiveness of our method.

Fig. 7. The visualization of registration results of IMFNet (multimodal), GeoTransformer (single modal) and the proposed IGReg. Examples in the first three rows are from 3DMatch whose overlap ratio is larger than 30% and examples in the remaining rows come from the 3DLoMatch where the overlap ratio is lower than 30%.

C. The Generalization of IGReg

In this section, we further demonstrate the generalisability of the proposed method by experimenting on the TUM RGB-D SLAM dataset [46] which is an indoor dataset proposed by the TUM Vision Group. Following [50], we choose 8 sequences for testing IGReg. For sequences "xyz", "360", "teddy", "desk" and "plant" captured with camera 1, we select one in every 5 frames for registration. For sequences "dishes", "coke" and "flower bouquet" captured with camera 2, we select one in every 20 frames for registration. Like the previous methods, we normalize the point clouds by aligning their centroids and applying the scaling operation to ensure the diagonal length of the bounding box equals 1.

Metrics. Following [50], the result of IGReg is evaluated using the Root Mean Square Error (RMSE) metric. The metric measures the accuracy of the estimated transformation $(\mathbf{R}, \mathbf{t})$ and is computed as the following:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{\mathbf{p}_i \in \mathcal{P}} \|\mathbf{R} \cdot \mathbf{p}_i + \mathbf{t} - \mathbf{R}^* \cdot \mathbf{p}_i - \mathbf{t}^*\|_2^2} \quad (25)$$

where $(\mathbf{R}^*, \mathbf{t}^*)$ is ground-truth transformation and $\mathbf{p}_i$ is the i_{th} point of point cloud $\mathcal{P}$.

Registration Results. We directly use the model trained on 3DMatch without fine-tuning and evaluate its performance on the TUM RGB-D SLAM dataset to show how well the proposed IGReg model can generalize to unseen data. The registration results are summarised in Table IV. We report the average and median RMSE of each sequence along with the mean value of them. IGReg achieves the minimal average and median RMSE of all sequences. For the mean value of average RMSE, IGReg is 0.82 lower than the second ranked method DGR [34]. Besides, on the challenging sequences captured from camera 2, the average RMSEs of our method are all below 1.2. The results indicate that IGReg still performs well when it is applied on unknown scenes, showing the proposed method has a good ability to generalise to data that comes from independent sources.

TABLE IV
REGISTRATION RESULTS ON EIGHT SEQUENCES FROM THE TUM RGB-D SLAM DATASET.

Method	xyz	360	teddy	desk	plant	dishes	coke	flower	mean
Average RMSE($\times 10^{-2}$)↓									
ICP [51]	2.1	5.1	2.1	2.3	1.6	3.7	3.1	2.7	2.8
AA-ICP [52]	2.1	5.1	2.1	2.3	1.6	3.7	3.1	2.7	2.8
Sparse ICP [53]	1.6	4.8	1.8	1.8	0.88	3.9	3.3	3.2	2.66
Robust ICP [50]	0.5	2.2	1	1.2	0.65	3.2	2.4	2.4	1.69
Symmetric ICP [54]	1.2	1.8	1.1	1.7	0.7	3.8	3.4	3.3	2.13
GMM-Reg [55]	4.4	6.3	3.2	5.7	2.7	3.7	3.1	2.4	3.94
CPD [56]	4.6	5.3	2.2	2.1	1.4	3.4	2.6	2.3	2.99
Teaser++ [57]	3.4	20	11	8	6.1	21	23	20	14.06
DCP [31]	6.5	10	6.6	7	5.6	7.4	7.5	6.3	7.11
DGR [34]	0.6	1.4	1	1.2	0.71	2.6	2.1	1.9	1.44
IGReg	**0.2**	**0.3**	**0.44**	**0.4**	**0.29**	**1.15**	**1.05**	**1.1**	**0.62**
Median RMSE($\times 10^{-2}$)↓									
ICP [51]	0.89	4	1.4	1.2	1.1	2.8	2.5	2.2	2.01
AA-ICP [52]	0.9	4	1.4	1.2	1.1	2.8	2.5	2.1	2.0
Sparse ICP [53]	0.86	3.7	1.1	1.1	0.67	3.6	3	3	2.13
Robust ICP [50]	0.43	0.75	0.76	0.77	0.56	2.6	2.1	1.8	1.22
Symmetric ICP [54]	0.44	0.73	0.82	0.7	0.59	3	2.9	3	1.52
GMM-Reg [55]	3.7	4.6	2.6	4.6	2.1	2.7	2	1.8	3.01
CPD [56]	3.6	3.3	1.5	1.8	1.1	2.7	1.8	1.8	2.2
Teaser++ [57]	2	14	3.2	2.5	2.4	15	21	11	8.89
DCP [31]	5.4	9.9	5.5	6	4.9	6	6	5.7	6.18
DGR [34]	0.53	0.86	0.84	0.96	0.65	1.7	1.2	1.1	0.86
IGReg	**0.17**	**0.23**	**0.35**	**0.32**	**0.27**	**0.66**	**0.5**	**0.77**	**0.42**

D. Ablation Experiment

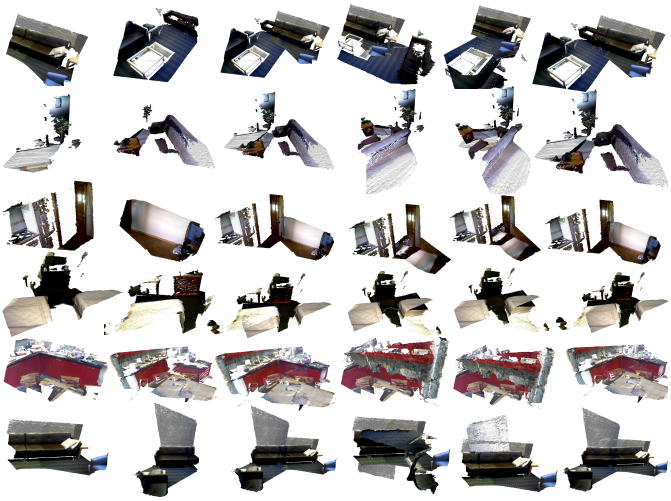
In this section, we design ablation experiments to analyse the effectiveness of hyperparameters and modules in our network. The datasets used for these ablation experiments are 3DMatch and 3DLoMatch, and the methods for evaluating the transformation matrix are both LGR.

Effect of each component in IGReg. As the proposed IGReg consists of the Intra-Image-Geometry (IIG) fusion module and the selective correlation fusion (SCF) module. We firstly evaluate the effect of each component in Table V.

To investigate the effectiveness of the SCF module, we remove it and fuse global correlations into the superpoint features. As shown in the first row of Table V, the Patch Inlier Ratio (PIR), IR, and RR decreases by 6.4%, 4.5%, 2.9% on 3DMatch and 6.7%, 4.7%, 0.5% on 3DLoMatch respectively.

TABLE V
ABLATION EXPERIMENTS ON EACH COMPONENT IN IGREG.

Model	3DMatch				3DLoMatch			
	PIR	FMR	IR	RR	PIR	FMR	IR	RR
w/o SCF Module	78.3	98.7	65.0	90.6	45.8	88.2	37.2	74.6
w/o IIG Fusion	**86.5**	98.3	**70.9**	91.7	**55.7**	86.5	**44.1**	75.1
IGReg	84.7	**98.7**	69.5	**93.5**	52.5	**88.8**	41.9	**75.1**



(a) Target (b) Source (c) Ground truth (d) IGReg w/o SCF (e) IGReg w/o IIG (f) IGReg

Fig. 8. More intuitive explanation of different modules of our method.

Without SCF module, the global correlation merges ambiguous geometry-texture information and results in registration failure. The results demonstrate that our SCF module can effectively enhance the feature distinctiveness and further improve the validity of point cloud registration.

Then we remove the IIG fusion module from the IGReg. In this way, our method is modified to be single modal. Following previous works like Predator [29], RR is the main metric which corresponds to the actual aim of point cloud registration. It is defined as the fraction of scan pairs from which the correct transformation parameters are found. As shown in Table V, with image information, IGReg gains 1.8% in RR on 3DMatch and keeps 75.1% on the 3DLoMatch. To explain the effectiveness of the IIG Module and SCF module more intuitively, we compare the registration results of three settings in Figure 8: (1) IGReg without the SCF module, (2) IGReg without IIG module and (3) IGReg. As shown in column (e) of Figure 8, if we remove the texture information, the points with similar geometry will be registered. For example, the floor parts in the first row are aligned incorrectly. When the SCF module is removed, only the image features are fused with the geometry information. The correlation with respect to the salient anchors in the overlap region is not considered, which results in some confusing parts being wrongly registered. In the 5th example of column (d) in Figure 8, the indistinctly red cabinet door parts are aligned, acquiring wrong registration results. After we adopt both the SCF and IIG modules, all these repetitive patterns and low geometry parts can be registered correctly, which verifies the effectiveness of our proposed modules.

TABLE VI
ABLATION EXPERIMENTS ON EACH PART IN IIG FUSION MODULE.

Model	3DMatch				3DLoMatch			
	PIR	FMR	IR	RR	PIR	FMR	IR	RR
w/o Projection	83.4	98	68.8	89	51.4	87.6	41.2	73.1
w/o CA Fusion	83.9	98.4	69.4	92.1	52.3	88.1	41.5	72.8
IGReg	**84.7**	**98.7**	**69.5**	**93.5**	**52.5**	**88.8**	**41.9**	**75.1**

Effect of projection and cross-attention in IIG fusion module. During the IIG fusion stage, the module is mainly composed of the projection part which performs 2D-3D alignment and the cross-attention part which fuses the multimodal features. We conduct ablation studies on these two parts to validate the effectiveness.

To evaluate the projection part, we compare it with a projection-free method, which fuses features of the point cloud and the whole image information directly with transformer just like IMFNet. However, the projection-free approach incorporates redundant feature information and involves ambiguity in the latter point cloud registration task, as it reports the lowest registration recall in the first row of Table VI.

For the sake of sufficiently utilising image features to enhance point cloud features, we explore different fusion methods and conduct ablation studies on these strategies. We replace the cross-attention based fusion with the average fusion as shown in the second row in Table VI. We use average pooling to process the features of pixels within the corresponding superpoint region, and then add the average feature to the geometry feature. It can be seen that average fusion can not deal with multimodal features effectively, especially in the low-overlap point cloud registration task. With cross attention technique, we can fuse multimodal features in a more reasonable way, which outperforms other fusion strategy.

Effect of anchor location in SCF module. As shown in Table VII, the anchor location module can effectively enhance the role of geometric embedding and further improve the results of point cloud registration. Compared with the anchor location module by selecting the $top-k$ correspondences, the anchor location module with NMS can improve the FMR and RR by 0.4% and 1.9% on 3DMatch respectively. It can also slightly improve the RR on 3DLoMatch by 0.3%. When using the top-k method to select anchor correspondences, there may be multiple anchor points close to each other so that these anchor points cannot preserve the structure of point clouds, which weakens the geometric embedding effect of the selective geometry embedding module.

TABLE VII
ABLATION EXPERIMENTS OF THE ANCHOR LOCATION.

Model	3DMatch				3DLoMatch			
	PIR	FMR	IR	RR	PIR	FMR	IR	RR
top-k	**86.7**	98.3	**70.9**	91.6	**55.3**	**89.3**	**43.8**	74.8
NMS	84.7	**98.7**	69.5	**93.5**	52.5	88.8	41.9	**75.1**

Effect of iterative-optimization based salient anchor updating (IOSAU). As shown in Table VIII, on 3DMatch, the RR of iteration=2 is 2.8% higher than that of iteration=1, and the performance of iteration=3 is lower than 0.7% of the RR of iteration=2. We analyse that as the iteration increases, the sampling of the anchor point is in some different areas. The anchor point is inevitably concentrated when the overlap rate is low. After the anchor point is concentrated, the embedded features lose their differences, leading to a decline in performance. This point can be verified from the 3DMatch experiment in Table VIII. When the overlap rate is high, the distribution of anchor points is relatively even and does not cause performance degradation. Here, we also evaluate the time consumption of each iteration in Table VIII. It can be seen that using two iterations can achieve the best trade-off between accuracy and time efficiency. Thus, we choose to iterate twice for the interaction between the salient location and the selective correlation fusion module.

TABLE VIII
ABLATION EXPERIMENTS OF THE NUMBER OF ITERATIONS.

Iteration	3DMatch				3DLoMatch				Model
	PIR	FMR	IR	RR	PIR	FMR	IR	RR	time(s)
1	84.3	98.6	69.0	90.7	52.0	88.8	41.5	74.8	0.183
2	84.7	98.7	69.5	**93.5**	52.5	**88.8**	41.9	**75.1**	0.205
3	**87.9**	**98.8**	**71.6**	92.8	**56.5**	88.5	**44.2**	74.8	0.229

Effect of the number of anchor point. We analyse that when the number of anchor points is 1, all superpoints on the sphere with anchor points as the radius have the same geometric constraints. When the number of anchor points is 2, the anchor points of two spherical circles have the same constraint condition, which makes the feature distinction lose the distinguishing ability in terms of symmetry. Therefore, we start the comparison experiment when the number of anchor points equals 3. It can be seen from table IX that the registration recall decreases as the number of anchor points increases. Thus, in the experiment, we select the number of anchors as 3.

TABLE IX
ABLATION EXPERIMENTS OF THE NUMBER OF ANCHOR POINTS.

Numbers	3DMatch				3DLoMatch			
	PIR	FMR	IR	RR	PIR	FMR	IR	RR
3	84.7	**98.7**	69.5	**93.5**	52.5	**88.8**	41.9	**75.1**
4	**87.1**	98.7	**71.0**	91.6	**55.8**	88.5	**43.9**	75.0
5	85.4	98.4	69.6	90.4	53.5	87.6	42.1	74.3

We also visualize the location of anchor points in the source and target in Figure 9. The blue spheres are the anchor location before we perform IOSAU and it can be seen that the initial anchors are located randomly in the whole point clouds. After completing iteration based optimization updating, we can see that the updated anchors (orange spheres) are located in overlapping areas shown as the yellow parts. In this way, the correlations of superpoints with respect to salient anchors are correct and can improve the distinction of features.

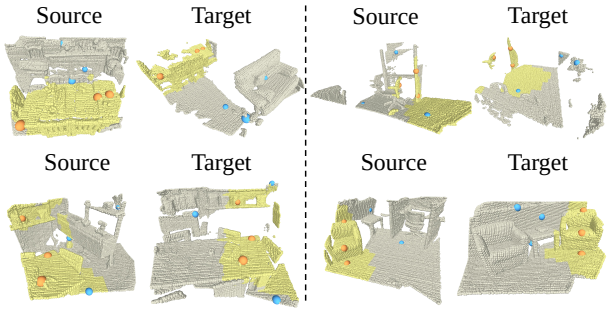


Fig. 9. Visualization of the location of anchor points in the source and target point clouds before and after iteration-based optimisation salient anchor updating (IOSAU). The blue spheres are the initial anchors before IOSAU. The orange ones are the salient anchors after IOSAU. We can see that through IOSAU, the salient anchors are located in the overlap areas (yellow regions).

We provide the ablation study of feature matching recall (FMR) and inlier ratio (IR) under different inlier distance thresholds in Figure 10. The blue lines and the yellow lines represent the performance on two datasets of 3DMatch and 3DLoMatch, respectively. We can see that FMR and IR continuously grow with the increase of the inlier distance threshold. For a fair comparison with current classical approaches, such as D3Feat [13], Predator [29] and GeoTransformer [9], we set the same inlier distance threshold to 0.1m.

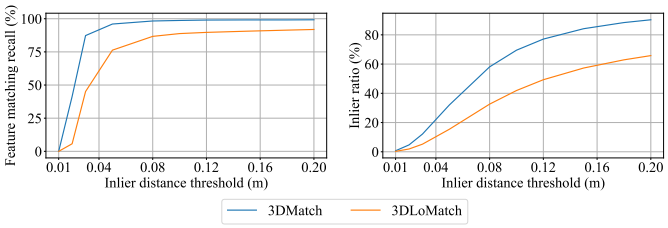


Fig. 10. Feature matching recall and inlier ratio under different inlier distance thresholds.

V. CONCLUSION

In this paper, we propose a multimodal point cloud registration method based on selectively fusing the correlations between salient anchors and superpoints. With the help of images, the point cloud features are enriched by the proposed IIG fusion module. Then we locate sparse and shape-preserved salient anchors and selectively embed the correlations between anchors and superpoints to enhance features for accurate superpoint matching. In order to fuse the most distinct information into the feature, we also propose an iterative-optimization-based salient anchor updating to achieve the most effective and accurate correlation fusion. The proposed approach utilises images to enrich point cloud features and uses salient anchors as the medium to exchange geometric information between source and target point clouds. The correlation information with respect to the salient anchor correspondences enables accurate and reliable superpoint correspondences, even for those appearance-similar areas in the non-overlap region, and those areas with low geometry in the overlap region.

REFERENCES

[1] Y. Yu, W. Zhang, F. Yang, and G. Li, "Rate-distortion optimized geometry compression for spinning lidar point cloud," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.

[2] J. Shan, S. Zhou, Y. Cui, and Z. Fang, "Real-time 3d single object tracking with transformer," *IEEE Transactions on Multimedia*, vol. 25, pp. 2339–2353, 2023.

[3] Y. Zheng, Y. Li, S. Yang, and H. Lu, "Global-pbnet: a novel point cloud registration for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 22 312–22 319, 2022.

[4] L. Zhao and W. Tao, "Jsnet++: Dynamic filters and pointwise correlation for 3d point cloud instance and semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–12, 2022, doi:10.1109/TCSVT.2022.3218076.

[5] A.-D. Nguyen, S. Choi, W. Kim, J. Kim, H. Oh, J. Kang, and S. Lee, "Single-image 3-d reconstruction: Rethinking point cloud deformation," *IEEE Transactions on Neural Networks and Learning Systems*, 2022, doi:10.1109/TNNLS.2022.3211929.

[6] C. Gu, Y. Cong, and G. Sun, "Three birds, one stone: Unified laser-based 3-d reconstruction across different media," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[7] Q. Wu, J. Wang, Y. Zhang, H. Dong, and C. Yi, "Accelerating point cloud registration with low overlap using graphs and sparse convolutions," *IEEE Transactions on Multimedia*, 2023.

[8] Y. Wang, C. Yan, Y. Feng, S. Du, Q. Dai, and Y. Gao, "Storm: Structure-based overlap matching for partial point cloud registration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 1135–1149, 2022.

[9] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, and K. Xu, "Geometric transformer for fast and robust point cloud registration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 143–11 152.

[10] Z. Xu, Q. Zhang, and S. Cheng, "Multilevel active registration for kinect human body scans: from low quality to high quality," *Multimedia Systems*, vol. 24, no. 3, pp. 257–270, 2018.

[11] P. Jauer, I. Kuhlemann, R. Bruder, A. Schweikard, and F. Ernst, "Efficient registration of high-resolution feature enhanced point clouds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 5, pp. 1102–1115, 2018.

[12] D. Campbell and L. Petersson, "An adaptive data representation for robust point-set registration and merging," in *2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4292–4300.

[13] 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.

[14] 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.

[15] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," in *Readings in Computer Vision*, M. A. Fischler and O. Firschein, Eds. San Francisco (CA): Morgan Kaufmann, 1987, pp. 726–740. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780080515816500702

[16] M. Saleh, S. Dehghani, B. Busam, N. Navab, and F. Tombari, "Graphite: Graph-induced feature extraction for point cloud registration," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 241–251.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, p. 5998–6008, 2017.

[18] 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.

[19] J. Zhang, Y. Cao, and Q. Wu, "Vector of locally and adaptively aggregated descriptors for image feature representation," *Pattern Recognition*, vol. 116, p. 107952, 2021.

[20] 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.

[21] R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in Neural Information Processing Systems*, vol. 30, p. 5099–5108, 2017.

[22] H. Deng, T. Birdal, and S. Ilic, "Ppfnet: Global context aware local features for robust 3d point matching," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 195–205.

[23] Z. J. Yew and G. H. Lee, "3dfeat-net: Weakly supervised local 3d features for point cloud registration," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 630–646.

[24] Z. Gojcic, C. Zhou, J. D. Wegner, and A. Wieser, "The perfect match: 3d point cloud matching with smoothed densities," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5540–5549.

[25] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, oct 2019. [Online]. Available: https://doi.org/10.1145/3326362

[26] 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.

[27] 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.

[28] S. Ao, Q. Hu, B. Yang, A. Markham, and Y. Guo, "Spinnet: Learning a general surface descriptor for 3d point cloud registration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 753–11 762.

[29] 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.

[30] X. Bai, Z. Luo, L. Zhou, H. Chen, L. Li, Z. Hu, H. Fu, and C.-L. Tai, "Pointdsc: Robust point cloud registration using deep spatial consistency," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 854–15 864.

[31] Y. Wang and J. Solomon, "Deep closest point: Learning representations for point cloud registration," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3522–3531.

[32] W. Lu, G. Wan, Y. Zhou, X. Fu, P. Yuan, and S. Song, "Deepvcp: An end-to-end deep neural network for point cloud registration," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 12–21.

[33] J. Li, C. Zhang, Z. Xu, H. Zhou, and C. Zhang, "Iterative distance-aware similarity matrix convolution with mutual-supervised point elimination for efficient point cloud registration," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 378–394.

[34] C. 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.

[35] Y. Aoki, H. Goforth, R. A. Srivatsan, and S. Lucey, "Pointnetlk: Robust & efficient point cloud registration using pointnet," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7156–7165.

[36] H. Deng, T. Birdal, and S. Ilic, "3d local features for direct pairwise registration," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3239–3248.

[37] V. Sarode, X. Li, H. Goforth, Y. Aoki, R. A. Srivatsan, S. Lucey, and H. Choset, "Pcrnet: Point cloud registration network using pointnet encoding," *ArXiv*, vol. abs/1908.07906, 2019.

[38] 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.

[39] H. Xu, S. Liu, G. Wang, G. Liu, and B. Zeng, "Omnet: Learning overlapping mask for partial-to-partial point cloud registration," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3112–3121.

[40] X. Huang, W. Qu, Y. Zuo, Y. Fang, and X. Zhao, "Imfnet: Interpretable multimodal fusion for point cloud registration," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 12 323–12 330, 2022.

[41] H. Chen, Z. Wei, Y. Xu, M. Wei, and J. Wang, "Imlovenet: Misaligned image-supported registration network for low-overlap point cloud pairs," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–9.
[42] Y. Zhang, J. Yu, X. Huang, W. Zhou, and J. Hou, "Pcr-cg: Point cloud registration via deep explicit color and geometry," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*. Springer, 2022, pp. 443–459.
[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[44] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[45] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4938–4947.
[46] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 573–580.
[47] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.
[48] H. Wang, Y. Liu, Z. Dong, and W. Wang, "You only hypothesize once: Point cloud registration with rotation-equivariant descriptors," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1630–1641.
[49] Y. Zhang, J. Yu, X. Huang, W. Zhou, and J. Hou, "Pcr-cg: Point cloud registration via deep color and geometry," *arXiv preprint arXiv:2302.14418*, 2023.
[50] J. Zhang, Y. Yao, and B. Deng, "Fast and robust iterative closest point," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3450–3466, 2021.
[51] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.
[52] A. L. Pavlov, G. W. Ovchinnikov, D. Y. Derbyshev, D. Tsetserukou, and I. V. Oseledets, "Aa-icp: Iterative closest point with anderson acceleration," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3407–3412.
[53] S. Bouaziz, A. Tagliasacchi, and M. Pauly, "Sparse iterative closest point," in *Computer graphics forum*, vol. 32, no. 5. Wiley Online Library, 2013, pp. 113–123.
[54] S. Rusinkiewicz, "A symmetric objective function for icp," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–7, 2019.
[55] B. Jian and B. C. Vemuri, "Robust point set registration using gaussian mixture models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 8, pp. 1633–1645, 2010.
[56] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 12, pp. 2262–2275, 2010.
[57] H. Yang, J. Shi, and L. Carlone, "Teaser: Fast and certifiable point cloud registration," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 314–333, 2020.

Zongyi Xu received the Ph.D. degree from Queen Mary University of London (QMUL) in 2019, the Master's degree in Computer Technology and the Bachelor's degree in Information Security from the University of Electronic Science and Technology of China (UESTC). She is currently an associate professor at Chongqing University of Posts and Telecommunications. Her research interests are 3D Vision, Point Cloud Processing and 3D Human Body Modelling.

Xinqi Jiang is currently an undergraduate student at the School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China. He is working toward the Bachelor's degree in computer science and technology. His research interests include computer graphics and point cloud registration.

Xinyu Gao received the Bachelor's degree in School of Automation Engineering from Northeast Electric Power University. She is currently working toward the Master's degree at Chongqing University of Posts and Telecommunications. Her research interests include point cloud registration and computer graphics.

Rui Gao received the Bachelor's degree from Changsha University of Science and Technology and the Master's degree from Chongqing University of Posts and Telecommunications. During the Master's degree, his research interest is point cloud registration.

Changjun Gu received the B.S. degree in automation from Shenyang Ligong University, Shenyang, China, in 2014, the M.E. degree in control theory and control engineering from Shenyang Ligong University, Shenyang, China, in 2017, and the Ph.D. degree in pattern recognition and intelligent system from the University of Chinese Academy of Sciences, Beijing, China, in 2022. He is currently a Lecturer at the Chongqing University of Posts and Telecommunications. His current research interests focus on robotics and 3D computer vision.

Qianni Zhang received the Ph.D.degree from Queen Mary University of London, U.K., in 2007. She is currently a senior lecturer at the School of Electronic Engineering and Computer Science, Queen Mary University of London. Her research interests include medical image analysis and understanding; deep learning networks for immunohistochemical data classification, 3D human modelling and animation, augmented reality for surgery planning and guidance.

Weisheng Li received the B.S. and M.S. degrees from the School of Electronics and Mechanical Engineering, Xidian University, Xi'an, China, in 1997 and 2000, respectively and the Ph.D. degree from the School of Computer Science and Technology, Xidian University, in 2004. He is currently a Professor with the Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include intelligent information processing and pattern recognition.

Xinbo Gao (M'02-SM'07-F'24) received the B.Eng., M.Sc. and Ph.D. degrees in electronic engineering, signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a research fellow at the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a post-doctoral research fellow at the Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong. Since 1999, he has been at the School of Electronic Engineering, Xidian University and now he is a Professor of Pattern Recognition and Intelligent System of Xidian University. Since 2020, he has been also a Professor of Computer Science and Technology of Chongqing University of Posts and Telecommunications. His current research interests include computer vision, machine learning and pattern recognition. He has published seven books and around 300 technical articles in refereed journals and proceedings. Prof. Gao is on the Editorial Boards of several journals, including Signal Processing (Elsevier) and Neurocomputing (Elsevier). He served as the General Chair/Co-Chair, Program Committee Chair/Co-Chair, or PC Member for around 30 major international conferences. He is Fellows of IEEE, IET, AAIA, CIE, CCF, and CAAI.