

GPI-Net: Gestalt-Guided Parallel Interaction Network via Orthogonal Geometric Consistency for Robust Point Cloud Registration

Weikang Gu¹, Mingyue Han¹, Li Xue¹ and Heng Dong² and Changcai Yang¹ and Riqing Chen¹ and Lifang Wei¹

¹College of Computer and Information Science, Fujian Agriculture and Forestry University

²School of Computing and Information Science, Fuzhou Institute of Technology

gwk154137@163.com, mingyue.han@163.sufe.edu.cn, 395391275@qq.com,
changcaiyang@gmail.com, {52311049004, riqing.chen, weilifang0028}@fafu.edu.cn

Abstract

The accurate identification of high-quality correspondences is a prerequisite task in feature-based point cloud registration. However, it is extremely challenging to handle the fusion of local and global features due to feature redundancy and complex spatial relationships. Given that Gestalt principles provide key advantages in analyzing local and global relationships, we propose a novel Gestalt-guided Parallel Interaction Network via orthogonal geometric consistency (GPI-Net) in this paper. It utilizes Gestalt principles to facilitate complementary communication between local and global information. Specifically, we introduce an orthogonal integration strategy to optimally reduce redundant information and generate a more compact global structure for high-quality correspondences. To capture geometric features in correspondences, we leverage a Gestalt Feature Attention (GFA) block through a hybrid utilization of self-attention and cross-attention mechanisms. Furthermore, to facilitate the integration of local detail information into the global structure, we design an innovative Dual-path Multi-Granularity parallel interaction aggregation (DMG) block to promote information exchange across different granularities. Extensive experiments on various challenging tasks demonstrate the superior performance of our proposed GPI-Net in comparison to existing methods. The code will be released at <https://github.com/gwk/GPI-Net>.

1 Introduction

The point cloud is now the predominant data format for the representation of the 3D world. In addition, a variety of point cloud processing algorithms have been designed for diverse applications, such as simultaneous localization and mapping [Xue and Lv, 2022], autonomous driving [Okuyama *et al.*, 2018], and robotics [Zhou *et al.*, 2022]. Sensors like LiDAR [Zhao *et al.*, 2024] and Microsoft Kinect [Yang *et al.*, 2016] are capable of generating point clouds directly. However, due to their limited line of sight, these sensors must be repositioned to obtain data from multiple viewpoints.

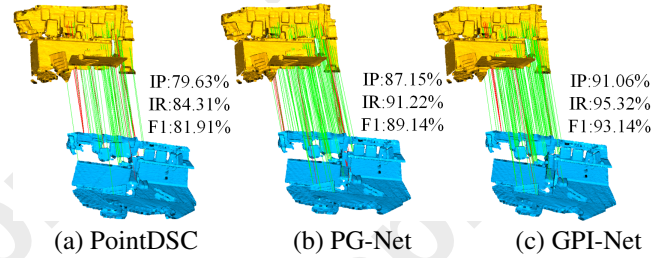


Figure 1: Visualization results of outlier removal. The green and red lines highlight inliers and outliers, respectively. On account of the power of our proposed GPI-Net to optimally aggregate local details and global information across multiple granularities, it demonstrates superior performance in identifying more inliers.

Point cloud registration has been developed to ensure the precise alignment of point clouds collected from different perspectives. Feature-based point cloud registration typically starts by generating initial correspondences utilizing feature descriptors, such as Fast Point Feature Histograms (FPFH) [Rusu *et al.*, 2009] and Fully Convolutional Geometric Feature (FCGF) [Choy *et al.*, 2019]. Nevertheless, owing to the lack of clarity in the feature descriptors and the ambiguity of local features, the initial correspondences frequently contain a substantial number of erroneous correspondences, giving rise to inaccurate alignment. Consequently, developing a robust and high-accuracy outlier removal network is crucial for increasing the inlier ratio and achieving accurate point cloud registration.

In recent decades, researchers have proposed various excellent outlier removal methods, including RANSAC [Fischler and Bolles, 1981], Spectral Matching (SM) [Leordeanu and Hebert, 2005], Fast Global Registration (FGR) [Zhou *et al.*, 2016], CG-SAC [Quan and Yang, 2020], and TEASER [Yang *et al.*, 2021]. Although these methods are effective in simple scenarios, they struggle to converge in complex scenes with high outlier ratios. It brings about inaccurate alignment under elevated outlier ratios. Recently, deep learning-based outlier removal methods have improved the accuracy and robustness of point cloud registration with stronger feature representation. Specifically, DGR [Choy *et al.*, 2020] adopts a feature embedding network to better extract global context information. 3DRegNet [Pais *et al.*, 2020] introduces classi-

fication and registration modules to recognize inliers. However, these methods rely solely on multilayer perceptrons and simple sparse convolution to capture the contextual information of correspondences, overlooking the importance of spatial consistency among inliers in 3D space. To overcome these challenges, PointDSC [Bai *et al.*, 2021] incorporates spatial consistency information to improve the accuracy of point cloud registration. PG-Net [Wang *et al.*, 2023] designs a grouped dense fusion attention module to collect rich contextual information and uses inlier probabilities to guide the classification of initial correspondences. Although the aforementioned methods perform well, they are insufficient to grasp the relationship between local details and global information. Figure 1(a) and (b) show that these methods left more outliers, which hinders effective complementarity and interaction among correspondences and may leave excessive redundant information. In summary, existing methods inevitably leave redundant information when exploring the relationship between local and global information, making it difficult to fully aggregate features and effectively capture more contextual information. To tackle these challenges, this paper proposes a novel Gestalt-guided Parallel Interaction Network via orthogonal geometric consistency (GPI-Net) designed for more robust and accurate outlier removal.

The Gestalt principles [Xu *et al.*, 2018] are visual perception principles that emphasize the superiority of whole perception over local perception, where the meaning of the whole is greater than the sum of the individual parts. In computer vision, object recognition is not solely based on the features of individual details but rather involves perceiving their overall structure and context. Leveraging Gestalt principles, GPI-Net not only extracts comprehensive and effective geometric features to support the acquisition of high-quality correspondences but also facilitates the complementarity and communication between local and global information, which reduces the generation of redundant information. Specifically, we introduce an Orthogonal Integration (OI) strategy in GPI-Net. The OI utilizes the spatial relationships of feature vectors to better aggregate local and global information while reducing redundancy during feature fusion. Then, we design a Gestalt Feature Attention (GFA) block based on the closure in Gestalt principles. The GFA accurately extracts geometric features of correspondences and effectively supplements missing contextual information by alternately applying self-attention [Gu *et al.*, 2021] and cross-attention [Kuang *et al.*, 2023], leveraging the relationships between surrounding features. Following the core ideas of the Gestalt principles, we present a novel Dual-path Multi-Granularity parallel interaction aggregation (DMG) block, which integrates contextual information from multiple granularities to better acquire local details and efficiently fuse them into global structural information. The GPI-Net boosts the differentiation between inliers and outliers to enhance understanding the internal dependencies of correspondences. Comprehensive experiments on outlier removal and pose estimation tasks reveal that our proposed framework exceeds existing methods.

We summarize the main contributions as follows.

- We propose a novel GPI-Net to address the feature re-

dundancy and complex spatial relationships, optimally guiding the classification of correspondences and alleviating the impact of outliers on point cloud registration.

- We incorporate Gestalt principles to guide GPI-Net in accurately identifying inliers on outlier removal task by emphasizing the superiority of whole perception over local perception.
- We design the OI and GFA to reduce redundant information by the spatial geometric features of correspondences, respectively. The DMG integrates spatial information across granularities via bidirectional hierarchical interaction, further enhancing the efficient fusion of local and global information.

2 Related Works

In this section, we briefly introduce some point cloud registration and outlier removal methods.

2.1 Point Cloud Registration

Iterative Closest Point (ICP) [Besl and McKay, 1992] is a classical traditional point cloud registration algorithm. The core idea of ICP is to iteratively optimize the objective function employing least squares to estimate the optimal rigid transformation. ICP is a more suitable choice when the point cloud already has good initial alignment or when there are fewer outliers in the scene. Nevertheless, traditional methods struggle to balance registration accuracy and computational cost. As a result, deep learning techniques have emerged to address these challenges. PointNet [Charles *et al.*, 2017] proposes a groundbreaking work that utilizes neural networks to extract features from point cloud. The core idea of PointNet is to employ symmetric functions to process unordered point cloud. However, PointNet has some limitations, particularly in extracting local geometric information. It neglects the local structure of point clouds and fails to fully utilize local information. On the basis of the limitations of PointNet in local feature learning, PointNet++ [Qi *et al.*, 2017] is presented to improve the network’s ability to perceive local structures and extract multi-scale features. By introducing a hierarchical feature extraction structure, PointNet++ effectively overcomes the shortcomings of PointNet in local feature learning and multi-scale feature representation.

2.2 Outlier Removal

RANSAC [Fischler and Bolles, 1981] achieves its goal by repeatedly selecting a random subset of the data. The selected subset is assumed to consist of inliers and the least-squares method is utilized for model fitting. Numerous RANSAC variants have introduced new sampling strategies and local optimization techniques to improve robustness. For instance, Locally Optimized RANSAC (LO-RANSAC) [Chum *et al.*, 2003] enhances robustness in the presence of local outliers by processing local data regions. However, RANSAC and its variants typically rely on random sampling and model evaluation, bringing about slow convergence and low accuracy. DHVR [Lee *et al.*, 2021] employs a decoupled multi-model fitting architecture based on deep Hough voting, separating and sequentially estimating rotation and translation. PG-Net

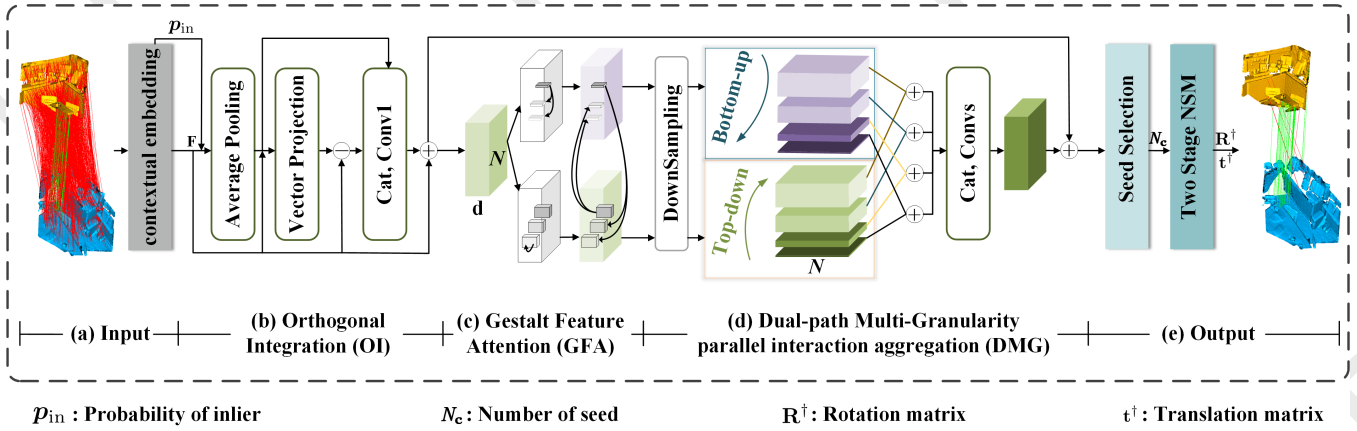


Figure 2: The framework of our proposed GPI-Net. GPI-Net takes the initial correspondences as input and generates N_c transformation matrices. It selects the optimal transformation matrix based on the number of inliers, achieving precise point cloud alignment. The OI, GFA, and DMG form the key components of GPI-Net.

[Wang *et al.*, 2023] designs a grouped dense fusion attention module to gather rich contextual information and an iterative structure to guide the classification of initial correspondences via inlier probabilities. However, these networks are insufficient to grasp the spatial relationship between local details and global information. Meanwhile, gathering spatial information at multiple granularities remains highly challenging. In this paper, our proposed GPI-Net uses a dual-path parallel interaction strategy with Gestalt principles, achieving exchange between local and global information across different granularities to improve pose estimation performance.

3 Method

In the following parts, we first introduce the problem formulation, then provide an overview of our proposed network, followed by a detailed explanation of each block.

3.1 Problem Formulation

Point cloud registration aims to align two partially overlapping point clouds, $P_s = \{p_k^s \in \mathbb{R}^3 \mid k = 1, 2, \dots, N\}$ and $P_t = \{p_j^t \in \mathbb{R}^3 \mid j = 1, 2, \dots, M\}$, into the same coordinate space by finding the transformation matrix between them. Firstly, we extract keypoint features from the point clouds via local descriptors (such as FPFH and FCGF) and generate the initial correspondences set C from these features through nearest-neighbor search.

$$C = \{c_1, c_2, \dots, c_N\} \in \mathbb{R}^{N \times 6}, \quad c_i = (p_i^s, p_i^t) \in C, \quad (1)$$

where c_i is the i th initial correspondence. $p_i^s \in P_s$ and $p_i^t \in P_t$ represent the spatial coordinates of the keypoints in the source and target point clouds. N is the number of initial correspondences.

Next, the initial correspondences are mapped to the probability that each correspondence is an inlier in the feature space, obtaining high-quality correspondences. At last, the high-quality correspondences are utilized to estimate the rigid transformation matrix, achieving point cloud registration.

Specifically, GPI-Net learns the probability of each correspondence being an inlier in high-dimensional feature space.

High-confidence correspondences are identified as seeds, and each seed obtains a consensus set in the metric space. Following the Two-Stage NSM [Wang *et al.*, 2023], each consensus set produces a transformation matrix $\mathbf{R}', \mathbf{t}'$. Through GPI-Net, we obtain N_c rigid transformation set $\mathbf{R}t_{set} = \{(\mathbf{R}'_1, \mathbf{t}'_1); (\mathbf{R}'_2, \mathbf{t}'_2); \dots; (\mathbf{R}'_{N_c}, \mathbf{t}'_{N_c})\}$, and the best transformation matrix $\mathbf{R}^\dagger, \mathbf{t}^\dagger$. The selection strategy is based on the number of correct correspondences to evaluate the quality of each transformation matrix. The transformation result of $\mathbf{R}^\dagger$ and $\mathbf{t}^\dagger$ is computed from the given set:

$$\mathbf{R}^\dagger, \mathbf{t}^\dagger = \arg \max_{(\mathbf{R}', \mathbf{t}') \in \mathbf{R}t_{set}} \sum_{i=1}^N [\|\mathbf{R}' p_i^s + \mathbf{t}' - p_i^t\| < \delta], \quad (2)$$

where $[\cdot]$ is the Iverson bracket, δ represents a predefined threshold for the inlier, and $\|\cdot\|$ denotes the Euclidean norm.

3.2 Overview

Our GPI-Net is shown in Figure 2. Initially, we input N pairs of initial correspondences into our proposed GPI-Net. The contextual embedding module [Wang *et al.*, 2023] maps the initial correspondences into a high-dimensional feature space and preliminarily extracts contextual information through grouped feature fusion. Next, the OI effectively filters out redundant information generated during feature aggregation by orthogonal integration. Then, the GFA leverages self-attention and cross-attention to further gather the geometric features of correspondences in high-dimensional space, enriching the contextual information. On this basis, the DMG collects features and facilitates information exchange across different granularities through a dual-path parallel interaction strategy, promoting the complementarity and interaction between local and global information. Subsequently, the seed selection module identifies high-quality correspondences. Finally, the Two-Stage NSM module is used to estimate the optimal point cloud transformation matrix.

3.3 Orthogonal Integration (OI)

To develop robust interactions between local details and global structures, we provide a feature map $F = \{f_p\}_{p=1}^N$,

where N represents the number of correspondences. It is necessary to gain global features, which are commonly obtained via global average pooling. However, this approach is unsuitable for the outlier removal task, on the ground that global average pooling assigns equal importance to each correspondence, while outliers have a detrimental effect on the network. To tackle this challenge, we adopt a weighted global average pooling to gather global features as shown in Figure 2(b). Specifically, we first map the feature map F to the inlier probability w of each correspondence by convolution. The probability value measures the importance of each correspondence within the network, guiding it to embed stronger global contextual information. The process is formalized as follows:

$$F_g = \xi(F, w) = \sum_{p=1}^N \frac{w_p}{\sum_{q=1}^N w_q} f_p, \quad (3)$$

here, ξ represents the weighted global average pooling operation. Then, we pass F_g through a bottleneck layer consisting of two 1D convolutional layers to further generate a more robust global feature F_g^T . Following this, we consider how to more optimally fuse local and global features while minimizing redundant information. Traditional methods simply employ operations like concatenation or element-wise addition, yet these methods frequently introduce considerable redundancy. Therefore, we apply orthogonal integration to achieve more compact feature aggregation. Specifically, we compute the projection of F in the direction of F_g^T , denoted as $F_{\text{projection}}$. Mathematically, the projection is as follows:

$$F_{\text{projection}} = \frac{F \cdot F_g^T}{\|F_g^T\|^2} \cdot F_g^T. \quad (4)$$

Owing to the difference in orthogonal components between the feature F and the projection vector $F_{\text{projection}}$, we further obtain the more crucial local detail feature map f within F , guided by the feature information in $F_{\text{projection}}$. The process is written as follows:

$$f = F - F_{\text{projection}}. \quad (5)$$

By this way, we gather a more crucial local information feature map from F , generating a more robust local feature representation f . The f not only contains more detailed feature information among correspondences, but also significantly reduces the interference of redundant information. Subsequently, we perform a concatenation operation between the f and F_g^T , followed by the convolutional layer to produce a more compact global feature representation F_o . The process is formalized as follows:

$$F_o = \text{Conv1}(\text{Cat}(f, F_g^T)) + F, \quad (6)$$

where $\text{Conv1}(\cdot)$ indicates a convolution kernel with a size of 1×1 . Through Equation (6), the network precisely captures the critical features of the task while maintaining optimal information representation capabilities through efficient integration of both local and global information.

3.4 Gestalt Feature Attention (GFA)

To comprehensively collect the geometric features among correspondences, we leverage the core idea of the Gestalt

principles: holistic perception is superior to local perception. As illustrated in Figure 2(c), we integrate self-attention and cross-attention mechanisms to identify similarities both within and between feature maps for capturing geometric information. The self-attention acquires long-range dependencies and global information by modeling relationships among correspondences. The cross-attention dynamically links connections between different feature maps, efficiently blending information from multiple feature maps. This ensures that information obtained from different feature maps is effectively intertwined and complementary. The GFA learns geometric information and complex relationships from local dependencies to global associations across different levels by stacking self-attention and cross-attention.

To tackle these complex variations, we combine the weights of different sizes, aiming to fully integrate the intrinsic local information into global structure. Specifically, the feature F_o is passed through a convolutional layer to generate the query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$). We compute self-attention across various correspondences and across multiple feature maps, using W_1 of size $\mathbb{R}^{N \times N}$ for correspondence-level attention and W_2 of size $\mathbb{R}^{d \times d}$ for feature-map-level attention. Here, N represents the number of correspondences and d denotes the number of feature channels.

$$f_{at1} = PWConv(\text{softmax}(\mathbf{Q} \cdot \mathbf{K}^T) \cdot \mathbf{V}) + F_o, \quad (7)$$

$$f_{at2} = PWConv(\text{softmax}(\mathbf{Q}^T \cdot \mathbf{K}) \cdot \mathbf{V}^T) + F_o, \quad (8)$$

where $PWConv(\cdot)$ denotes a point-wise convolution layer, f_{at1} represents the feature information obtained through the weight W_1 , which captures the interactions and dependencies among correspondences on a global granularity. The global information facilitates the network to understand the similarities and differences between correspondences from a holistic perspective. In contrast, f_{at2} focuses on collecting the internal structural information of features by calculating the similarity across each feature dimension, which emphasizes the dependencies between feature dimensions.

Self-attention struggles to fully explore the interactions and complementarities between different feature spaces. Based on the closure in Gestalt principles, missing information is supplemented by leveraging relationships between surrounding features. To boost dynamic connections and ensure complementarity between f_{at1} and f_{at2} , we adopt a cross-attention mechanism to alleviate the limitations of representations based on a single feature space. The cross-attention calculates the similarity between the $\mathbf{Q}$ and $\mathbf{K}$ from different feature maps, dynamically adjusting the weighted representation of the $\mathbf{V}$. Specifically, we utilize f_{at1} as the $\mathbf{Q}$, f_{at2} as the $\mathbf{K}$ and $\mathbf{V}$ to generate robust feature f'_{at1} , which incorporates contextual information from f_{at2} . Then, we reverse the roles to generate robust feature f'_{at2} that absorbs information from f_{at1} . The process enhances the information flow between different feature maps.

By combining self-attention and cross-attention, the network obtains geometric feature information and dependencies from a local to a global perspective.

3.5 Dual-Path Multi-Granularity Parallel Interaction Aggregation (DMG)

Extracting features at multiple granularities is highly challenging due to the complexity of the features and the difficulty in effectively balancing local and global features of the data. Grounded in the Gestalt principles that the whole is greater than the sum of its parts, we design a DMG block. The DMG adopts a dual-path parallel interaction strategy, achieving information exchange and integration between local and global features across different granularities.

Figure 2(d) shows that the features f'_{at1} and f'_{at2} from the GFA are downsampled into T granularities via average pooling. For f'_{at1} , the multi-granularity feature set is represented as $F_T = \{f_0, \dots, f_T\}$, where $f_t \in \mathbb{R}^{N \times d/2^t}$, and $t \in \{0, \dots, T\}$. N and d represent the number of correspondences and feature channels, respectively. Similarly, the f'_{at2} undergoes the same operation to generate a multi-granularity feature set $G_T = \{g_0, \dots, g_T\}$.

Gestalt principles suggest that relationships between local elements shape the perception of the whole. DMG integrates local and global information across granularities via bidirectional hierarchical interaction to produce a richer representation. Therefore, it is necessary to fuse features of different granularities to avoid the potential bias and to facilitate interaction between local and global information. Specifically, we apply dual-path mixing operations to different feature maps for promoting information exchange across multiple granularities. The process is defined by:

$$F_{\text{Ges}} = \text{Convs}(\text{Cat}(\{B_Mix(F_t) + T_Mix(G_t)\}_{t=0}^T)), \quad (9)$$

where the $\text{Convs}(\cdot)$ operation includes an InstanceNorm normalization layer, a BatchNorm layer, a ReLU activation function, a pointwise convolution layer, and a Shuffle channel mixing layer. The pointwise convolution layer takes input features of size $\mathbb{R}^{N \times 15d/8}$ and produces output of size $\mathbb{R}^{N \times d}$. $B_Mix(\cdot)$ and $T_Mix(\cdot)$ represent the coarse-to-fine mixing layer and fine-to-coarse mixing layer, respectively.

$B_Mix(\cdot)$ layer: Coarse-grained features provide a global and macro-level perspective, smoothing out local noise and facilitating the acquisition of global structural information by the network. Building on this, we adopt a bottom-up mixing strategy, starting with global features and gradually incorporating fine-grained local information, which guides the network to more accurately understand local details. Technically, for the multi-granularity feature set $F_T = \{f_0, \dots, f_T\}$, we utilize a residual structure to apply the $B_Mix(\cdot)$ layer to the t -th granularity, which achieves bottom-up mixing to facilitate effective information exchange between correspondences. The operation is formalized as follows:

$$F_t = F_t + \text{Mix_Bottom_up}(F_{t-1}), \quad (10)$$

where $t \in \{1, \dots, T\}$ and $\text{Mix_Bottom_up}(\cdot)$ consists of an InstanceNorm layer, a BatchNorm layer, a ReLU activation function, and a pointwise convolution layer. The pointwise convolution layer takes input and output features of size $\mathbb{R}^{N \times d/2^{t-1}}$ and $\mathbb{R}^{N \times d/2^t}$, respectively.

$T_Mix(\cdot)$ layer: The network incorporates broader context and global information to maintain sensitivity to details

after capturing fine-grained geometric features and structures. This fine-to-coarse strategy accurately captures local details first and gradually integrates global information, enhancing overall representation. Technically, for the multi-granularity features $G_T = \{g_0, \dots, g_T\}$, we also utilize a residual structure to employ the $T_Mix(\cdot)$ layer to the t -th granularity, forming a top-down mixing strategy. The operation is by:

$$G_t = G_t + \text{Mix_Top_down}(G_{t+1}), \quad (11)$$

where $t \in \{1, \dots, T\}$ and $\text{Mix_Top_down}(\cdot)$ consists of an InstanceNorm layer, a BatchNorm layer, a ReLU activation function, and a pointwise convolution layer. The pointwise convolution layer takes input and output features of size $\mathbb{R}^{N \times d/2^{t+1}}$ and $\mathbb{R}^{N \times d/2^t}$, respectively.

By introducing the DMG, the coarse-grained global perspective effectively smooths out noise in the fine-grained features. At the same time, the block extracts subtle geometric features at the fine-grained level and merges them with coarse-grained features to gather the global structure. This dual-path parallel interaction strategy, combining coarse-to-fine and fine-to-coarse approaches, has the power to achieve more precise global alignment.

4 Experiments

In this section, we first introduce the datasets and experimental parameters employed to verify the effectiveness of GPI-Net. Next, we evaluate GPI-Net in indoor and outdoor scenarios via FPFH and FCGF descriptors, respectively. Finally, we conduct an ablation study regarding our GPI-Net.

4.1 Datasets and Experimental Setup

Regarding indoor scenarios, we follow the same evaluation protocol as 3DMatch [Zeng *et al.*, 2017] to prepare the training and testing datasets. The testing set consists of 1623 pairs of point clouds from 8 indoor scenes. For each pair, we downsample the point clouds employing a 5cm voxel grid. For outdoor scenarios, we utilize the KITTI dataset [Geiger *et al.*, 2013] to test the model’s effectiveness. The testing set comprises 555 pairs of point clouds from 10 outdoor scenes, with each pair downsampled using a 30cm voxel grid.

We assess the effectiveness of GPI-Net through two tasks: outlier rejection and pose estimation. Regarding outlier rejection, we apply inlier precision (IP%), inlier recall (IR%), and F1 score (F1%) as evaluation metrics to measure the accuracy of inliers identification while minimizing false positives. For pose estimation, we adopt registration recall (RR%), rotation error (RE), and translation error (TE). Registration is considered successful when both RE and TE are below the specified thresholds: RE < 15° and TE < 30cm for indoor scenarios, and RE < 5° and TE < 60cm for outdoor scenarios. RE and TE are reported only on successfully registered point clouds.

4.2 Evaluation of Indoor Scenes

We first present the experimental results on the 3DMatch dataset to test the performance of our proposed GPI-Net in indoor scenarios. Table 1 presents the comparative results on the 3DMatch dataset.

Method		FPFH							FCGF						
		RR(%↑)	RE(° ↓)	TE(cm↓)	IP(%↑)	IR(%↑)	F1(%↑)	Time(s)	RR(%↑)	RE(° ↓)	TE(cm↓)	IP(%↑)	IR(%↑)	F1(%↑)	Time(s)
Traditional	SM	55.88	2.94	8.15	47.96	70.69	50.70	0.03	86.57	2.29	7.07	81.44	38.36	48.21	0.03
	FGR	40.91	4.96	10.25	6.84	38.90	11.23	0.14	78.93	2.90	8.41	25.63	53.90	33.58	0.17
	TEASER	75.48	2.48	7.31	73.01	62.63	66.93	0.04	85.77	2.73	8.66	82.43	68.08	73.96	0.11
	RANSAC-1M	64.20	4.05	11.35	63.96	57.90	60.13	0.39	88.42	3.05	9.42	77.96	79.86	78.55	0.45
	RANSAC-2M	65.25	4.07	11.56	64.41	58.37	60.51	3.87	90.88	2.71	8.31	78.52	83.52	80.68	4.13
Learning	3DRegNet	26.31	3.75	9.60	28.21	8.90	11.63	0.05	77.76	2.74	8.13	67.34	56.28	58.33	0.05
	DGR	32.84	2.45	7.53	29.51	16.78	21.35	2.49	88.85	2.28	7.02	68.51	79.92	73.15	1.36
	DHVR	67.10	2.78	7.84	60.19	64.90	62.11	0.46	90.93	2.25	7.08	78.35	78.15	78.94	0.46
	PointDSC	78.56	2.06	6.48	68.92	71.96	70.17	0.09	92.73	2.04	6.49	78.83	86.07	82.02	0.09
	PG-Net	81.96	2.07	6.62	72.45	77.72	74.82	0.18	92.98	2.04	6.47	79.12	86.03	82.16	0.18
	OURS	83.25	2.08	6.57	73.06	78.47	75.57	0.19	93.15	2.02	6.47	79.35	86.28	82.47	0.19

Table 1: Quantitative results on the 3DMatch dataset for outlier rejection and pose estimation.

Combined with FPFH. We first adopt FPFH descriptor to generate initial correspondences. From Table 1, we see that our method demonstrates satisfactory improvements in F1 and RR compared to other methods. Specifically, our method achieves an F1 approximately 6% higher than PointDSC and 1% higher than PG-Net, fully corroborating its effectiveness in the outlier rejection task. In terms of RR, our method surpasses PointDSC by nearly 5% and PG-Net by approximately 1.5%, further highlighting its superiority in the pose estimation task. This proves that GPI-Net, guided by Gestalt principles, optimally captures the complex relationships and internal dependencies between correspondences, leading to the acquisition of richer global structural information. As shown in Figure 3, we conduct a qualitative comparison with FPFH and our method reaches more accurate point cloud alignment relative to PointDSC and PG-Net. In contrast to traditional point cloud registration, GPI-Net outperforms RANSAC-2M in RR by nearly 20%, while maintaining low rotation and translation errors. Furthermore, we observe that 3DRegNet performs significantly worse than our approach in terms of RR with a gap of 56.94%. This indicates that using a simple MLP as the feature extractor is significantly limited in capturing global information. These limitations lead to poor representation of inliers, bringing about inaccurate point cloud alignment. It confirms that GPI-Net excels in inlier identification as well as in achieving lower registration errors, indicating an improvement in overall alignment accuracy.

Combined with FCGF. To comprehensively confirm the effectiveness of GPI-Net, we re-measure all methods exploiting FCGF descriptor. The results demonstrate that the FCGF descriptor consistently exhibits exceptional robustness and performance in different tasks. Our GPI-Net outperforms PG-Net in terms of F1 and RR, reaching 82.47% and 93.15%, respectively, which strongly affirms the superior performance of GPI-Net leveraging the FCGF descriptor. Moreover, we observe that while TEASER achieves a slightly higher IP than GPI-Net, GPI-Net significantly surpasses TEASER in IR and

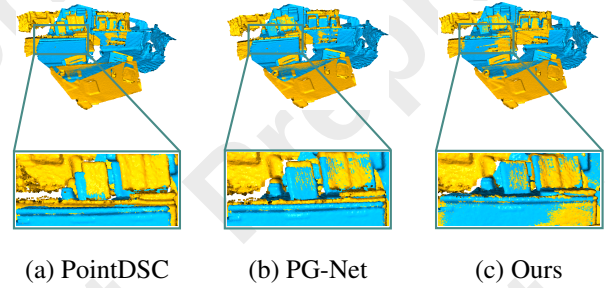


Figure 3: Qualitative comparison on 3DMatch dataset.

F1. It indicates that TEASER has limited capability in identifying inliers and struggles to collect structural information of inliers over a broader scope. Such a limitation affects its accuracy and robustness in point cloud registration. In contrast, GPI-Net demonstrates greater stability and effectiveness in the outlier removal task. Additionally, we find that employing the more robust FCGF descriptor consistently leads to better performance than the FPFH descriptor in both outlier rejection and pose estimation tasks. Table 1 reports that although our network experiences a slight decline in RR when applying FPFH in place of FCGF, the reduction is only 9.9%, whereas PG-Net shows a decrease of 11.02%. This underscores the robustness and adaptability of GPI-Net to different initial correspondences generated by varying descriptors.

Robustness with fewer correspondences. In the pose estimation task, the accuracy heavily depends on the number of correspondences. Therefore, to additionally verify the effectiveness and robustness of GPI-Net, we investigate the impact of the number of correspondences and compare it with two methods: PointDSC and PG-Net. Figure 4 shows that GPI-Net exhibits an RR that exceeds PointDSC by more than 10% at correspondence counts of 250, 500, and 1000. Similarly, GPI-Net surpasses PG-Net by over 2% when the number of correspondences is 500, 1000, and 2500. Overall, GPI-Net

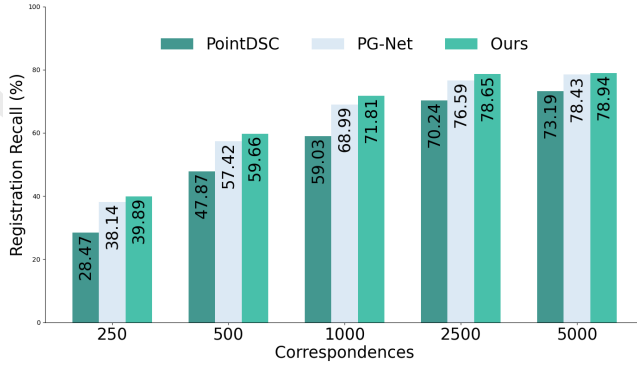


Figure 4: Comparison of RR utilizing different numbers of correspondences with FPFH.

Method	FPFH				FCGF			
	RR(% $\uparrow$)	RE(% $\downarrow$)	TE(cm $\downarrow$)	F1(% $\uparrow$)	RR(% $\uparrow$)	RE(% $\downarrow$)	TE(cm $\downarrow$)	F1(% $\uparrow$)
Traditional	FGR	5.23	0.86	43.84	—	89.54	0.46	25.72
	RANSAC	74.41	1.55	30.20	73.13	80.36	0.73	26.79
	CG-SAC	74.23	0.73	14.02	—	83.24	0.56	22.96
Learning	DGR	77.12	1.64	33.10	4.51	96.90	0.34	21.70
	PointDSC	97.95	0.30	7.23	88.03	97.96	0.33	21.29
	PG-Net	98.72	0.31	7.22	90.84	97.95	0.33	21.29
	OURS	98.96	0.30	7.17	91.52	98.21	0.33	21.26

Table 2: Quantitative Results on the KITTI dataset.

consistently outperforms PointDSC and PG-Net regardless of the number of correspondences. Notably, GPI-Net demonstrates increasingly pronounced advantages as the number of correspondences decreases, underscoring its effectiveness in scenarios with fewer correspondences.

4.3 Evaluation of Outdoor Scenes

To further confirm the effectiveness of GPI-Net in outdoor scenarios, we assess GPI-Net using the KITTI dataset. It is worth noting that our GPI-Net surpasses all these methods across multiple metrics in Table 2. Specifically, by utilizing the FPFH descriptor, our GPI-Net reaches a 24.55% improvement in RR over RANSAC, and with the FCGF descriptor, the gain is 17.85%. Meanwhile, our method preserves lower RE and TE while attaining the highest RR when paired with the FPFH descriptor. Furthermore, our method delivers competitive gains in RR and F1 relative to PointDSC, while achieving a lower TE. In comparison to PG-Net, our method not only attains higher RR and F1 but also minimizes errors with the FPFH descriptor. In summary, the results from both indoor and outdoor scenarios fully demonstrate the superiority and robustness of our network.

4.4 Ablation Study

In this section, we assess the effectiveness of the design decisions behind the block structures of GPI-Net. Table 3 gives the performance of different block combinations within the

datasets	baseline	OI	GFA	DMG	FPFH	
					RR(% $\uparrow$)	F1(% $\uparrow$)
3DMatch	✓	✓	✓	✓	81.96	74.82
					82.13	74.94
					82.18	74.89
					82.77	75.34
					82.99	75.46
					83.25	75.57
					98.72	90.84
KITTI	✓	✓	✓	✓	98.72	90.84
					98.76	90.88
					98.86	91.23
					98.92	91.42
					98.96	91.52

Table 3: Ablation Study. **Baseline**: PG-Net. **OI**: Orthogonal Integration. **GFA**: Gestalt Feature Attention. **DMG**: Dual-path Multi-Granularity parallel interaction aggregation.

GPI-Net with the FPFH descriptor. The first row represents PG-Net, serving as a baseline for comparison. It is evident that all block combinations outperform the baseline.

For instance, incorporating the DMG improves RR by 0.81% on the 3DMatch dataset, indicating that the DMG effectively aggregates local and global information through an innovative multi-granularity mixing strategy with dual-path parallel interactions. Next, adding the GFA on top of the DMG further boosts RR by 1%, highlighting that both GFA and DMG extract rich contextual information. This facilitates the separation of inliers and outliers. Ultimately, the inclusion of the OI contributes to an overall improvement of approximately 1.5%. The OI eliminates redundant information through orthogonal integration strategy, strengthening the power to recognize inliers, and significantly boosting the accuracy and robustness of point cloud registration.

5 Conclusion

In this paper, we propose an efficient and robust GPI-Net based on Gestalt principles. Specifically, the OI leverages the spatial information of feature vectors and adopts an orthogonal integration strategy to optimally reduce redundant information during feature fusion. The GFA employs a combination of self-attention and cross-attention to precisely obtain the geometric features of correspondences. The DMG performs an innovative multi-granularity feature extraction via dual-path parallel interaction strategy and facilitates information exchange across granularities, effectively grasping the relationship between local details and global information. GPI-Net promotes a more accurate differentiation between inliers and outliers. Extensive experiments demonstrate that our proposed GPI-Net delivers satisfactory improvements in both performance and robustness.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grant Nos. 62171130 and 62301160).

Contribution Statement

Weikang Gu and Mingyue Han contributed equally to this work. Lifang Wei is the corresponding author.

References

- [Bai *et al.*, 2021] Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. PointDSC: Robust point cloud registration using deep spatial consistency. In *Proc. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15854–15864. IEEE, jun 2021.
- [Besl and McKay, 1992] P.J. Besl and Neil D. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, February 1992. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
- [Charles *et al.*, 2017] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, July 2017. ISSN: 1063-6919.
- [Choy *et al.*, 2019] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully Convolutional Geometric Features. In *Proc. 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8957–8965, Seoul, Korea (South), October 2019. IEEE.
- [Choy *et al.*, 2020] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep Global Registration. In *Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2511–2520, June 2020.
- [Chum *et al.*, 2003] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally Optimized RANSAC. In *Proc. Pattern Recognition*, volume 2781, pages 236–243. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.
- [Fischler and Bolles, 1981] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, June 1981.
- [Geiger *et al.*, 2013] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, September 2013.
- [Gu *et al.*, 2021] Zongyun Gu, Qing Wang, Fangfang Li, and Junling Kan. Salient Object Detection Based on Self-attention Mechanism. In *Proc. 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 5, pages 969–975, October 2021.
- [Kuang *et al.*, 2023] Haopeng Kuang, Dingkan Yang, Shunli Wang, Xiaoying Wang, and Lihua Zhang. Towards Simultaneous Segmentation Of Liver Tumors And Intrahepatic Vessels Via Cross-Attention Mechanism. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, June 2023. ISSN: 2379-190X.
- [Lee *et al.*, 2021] Junha Lee, Seungwook Kim, Minsu Cho, and Jaesik Park. Deep Hough Voting for Robust Global Registration. In *Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15974–15983, October 2021.
- [Leordeanu and Hebert, 2005] M. Leordeanu and M. Hebert. A spectral technique for correspondence problems using pairwise constraints. In *Proc. Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1*, volume 2, pages 1482–1489 Vol. 2, October 2005.
- [Okuyama *et al.*, 2018] Takafumi Okuyama, Tad Gonsalves, and Jaychand Upadhyay. Autonomous Driving System based on Deep Q Learnig. In *Proc. 2018 International Conference on Intelligent Autonomous Systems (ICoIAS)*, pages 201–205, March 2018.
- [Pais *et al.*, 2020] G. Dias Pais, Srikumar Ramalingam, Venu Madhav Govindu, Jacinto C. Nascimento, Rama Chellappa, and Pedro Miraldo. 3DRegNet: A Deep Neural Network for 3D Point Registration. In *Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7191–7201, June 2020.
- [Qi *et al.*, 2017] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space, June 2017. arXiv:1706.02413 [cs].
- [Quan and Yang, 2020] Siwen Quan and Jiaqi Yang. Compatibility-Guided Sampling Consensus for 3-D Point Cloud Registration. *IEEE Transactions on Geoscience and Remote Sensing*, 58(10):7380–7392, October 2020.
- [Rusu *et al.*, 2009] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast Point Feature Histograms (FPFH) for 3D registration. In *Proc. 2009 IEEE International Conference on Robotics and Automation*, pages 3212–3217. IEEE, May 2009.
- [Wang *et al.*, 2023] Jingtao Wang, Xin Liu, Luanyuan Dai, Jiayi Ma, Lifang Wei, Changcai Yang, and Riqing Chen. PG-Net: Progressive Guidance Network via Robust Contextual Embedding for Efficient Point Cloud Registration. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023.
- [Xu *et al.*, 2018] Yusheng Xu, Wei Yao, Sebastian Tuttas, Ludwig Hoegner, and Uwe Stilla. Unsupervised Segmentation of Point Clouds From Buildings Using Hierarchical Clustering Based on Gestalt Principles. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(11):4270–4286, November 2018.
- [Xue and Lv, 2022] Xiujie Xue and Xiafu Lv. Stereo Visual Inertial SLAM Algorithm Fusing Point and Line Features. In *Proc. 2022 China Automation Congress (CAC)*, pages 1687–1692, November 2022.
- [Yang *et al.*, 2016] K. C. Yang, C. H. Huang, and C. F. Le. Applying microsoft kinect for windows to develop a Stroke Rehabilitation System. In *2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 1923–1927, December 2016. ISSN: 2157-362X.

- [Yang *et al.*, 2021] Heng Yang, Jingnan Shi, and Luca Carlone. TEASER: Fast and Certifiable Point Cloud Registration. *IEEE Transactions on Robotics*, 37(2):314–333, April 2021.
- [Zeng *et al.*, 2017] Andy Zeng, Shuran Song, Matthias Niessner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3DMatch: Learning Local Geometric Descriptors from RGB-D Reconstructions. In *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 199–208. IEEE, July 2017.
- [Zhao *et al.*, 2024] Lin Zhao, Hui Zhou, Xinge Zhu, Xiao Song, Hongsheng Li, and Wenbing Tao. LIF-Seg: LiDAR and Camera Image Fusion for 3D LiDAR Semantic Segmentation. *IEEE Transactions on Multimedia*, 26:1158–1168, 2024. Conference Name: IEEE Transactions on Multimedia.
- [Zhou *et al.*, 2016] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast Global Registration. In *proc. Computer Vision – ECCV 2016*, volume 9906, pages 766–782. Springer International Publishing, Cham, 2016.
- [Zhou *et al.*, 2022] Youjie Zhou, Yiming Wang, Fabio Poiesi, Qi Qin, and Yi Wan. Loop Closure Detection Using Local 3D Deep Descriptors. *IEEE Robotics and Automation Letters*, 7(3):6335–6342, July 2022.