SPEAL: Skeletal Prior Embedded Attention Learning for Cross-Source Point Cloud Registration

Kezheng Xiong[1], Maoji Zheng[1], Qingshan Xu[2], Chenglu Wen[1*], Siqi Shen[1*], Cheng Wang[1]

[1] Xiamen University [2] Nanyang Technological University

{xiongkezheng, zhengmaoji}@stu.xmu.edu.cn, qingshan.xu@ntu.edu.sg, {clwen, siqishen, cwang}@xmu.edu.cn,

Abstract

Point cloud registration, a fundamental task in 3D computer vision, has remained largely unexplored in cross-source point clouds and unstructured scenes. The primary challenges arise from noise, outliers, and variations in scale and density. However, neglected geometric natures of point clouds restrict the performance of current methods. In this paper, we propose a novel method, termed SPEAL, to leverage skeletal representations for effective learning of intrinsic topologies of point clouds, facilitating robust capture of geometric intricacy. Specifically, we design the Skeleton Extraction Module to extract skeleton points and skeletal features in an unsupervised manner, which is inherently robust to noise and density variances. Then, we propose the Skeleton-Aware GeoTransformer to encode high-level skeleton-aware features. It explicitly captures the topological natures and inter-point-cloud skeletal correlations with the noise-robust and density-invariant skeletal representations. Next, we introduce the Correspondence Dual-Sampler to facilitate correspondences by augmenting the correspondence set with skeletal correspondences. Furthermore, we construct a challenging novel cross-source point cloud dataset named KITTI CrossSource for benchmarking cross-source point cloud registration methods. Extensive quantitative and qualitative experiments are conducted to demonstrate our approach's superiority and robustness on both cross-source and same-source datasets. To the best of our knowledge, our approach is the first to facilitate point cloud registration with skeletal geometric priors.

Introduction

Point cloud registration is an essential task in graphics, vision, and robotics. It aims at estimating a rigid transformation to align two partially overlapping frames of point clouds. Recently, there has been a surge of interest in learning-based point cloud registration methods. These methods have made significant progress in addressing the sparsity, partial overlap, and complex distribution of point clouds in large outdoor scenes (Lu et al. 2021; Huang et al. 2021a; Yew and Lee 2022; Qin et al. 2022). However, the practical application and advances in point cloud acquisition present more challenges for point cloud registration, including unstructured scenes and cross-source data.

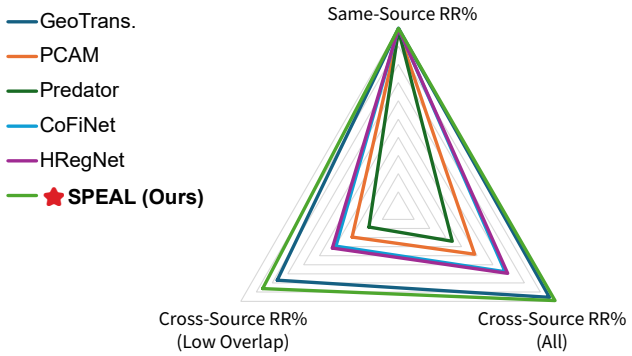


Figure 1: **Impossible Triangle of Current Methods.** Registration recalls under different settings are shown: KITTI Odometry (Same-Source), KITTI CrossSource and the low-overlap test split of KITTI CrossSource. Existing methods fail to perform as well as SPEAL on all three challenging circumstances.

In *unstructured scenes*, the complex natural scenes and objects often make it difficult to learn discriminative features for registration. This results in degraded performance of registration algorithms. In the case of *cross-source data*, challenges mainly arise from partial overlap, as well as considerable differences in scale and density, leading to difficulties in effective feature matching. The combination of noise and outliers from different sources further downgrades the quality of correspondences. Existing methods either focus solely on same-source point clouds, or overlook the intrinsic topological natures of the point clouds. This leads to suboptimal results for challenging scenarios such as cross-source point cloud registration and unstructured scenes.

We have observed that skeletons serve as an efficient and robust geometric representation for point clouds, exhibiting significant potential in various point cloud understanding tasks (Shi et al. 2021; Lin et al. 2021). They can effectively encode the geometric intricacy of point clouds. Inspired by this, we propose a novel transformer-based approach, termed **S**keletal **P**rior **E**mbedded **A**ttention **L**earning (SPEAL), to address the aforementioned challenges. Our method utilizes skeletal geometric priors to learn discriminative features for accurate and robust correspondences. To the

*Corresponding author.

best of our knowledge, our approach is the first to facilitate point cloud registration with skeletal geometric priors. Such skeletal geometric priors encourage robust feature learning by explicitly encoding the intrinsic topological characteristics, thereby facilitating the correspondences and registration results, as shown in Fig. 1

Specifically, to incorporate skeletal representations as a geometric prior, SPEAL comprises three key components: ***S**keleton **E**xtraction **M**odule (SEM)*, ***S**keleton-**A**ware **G**eo**TR**ansformer (SAGTR)*, and ***C**orrespondence **D**ual-**S**ampler (CDS)*. First, with the insights from the medial axis transform (MAT) (Blum 1967), SEM extracts a set of skeleton points with skeletal features from input point clouds in an unsupervised manner. It is robust to noise and density variances. Next, SAGTR is designed to learn skeleton-aware discriminative features, facilitating accurate and robust correspondences. It explicitly captures the topological natures and effectively learns inter-point-cloud geometric correlations with our skeletal representations. Finally, CDS samples reliable correspondences from both superpoints and skeleton points, which produces reliable coarse correspondences with awareness of the skeletal structure.

Extensive experiments are carried out on two datasets. One is *KITTI Odometry*, a large-scale outdoor registration benchmark (Geiger, Lenz, and Urtasun 2012), as well as its cross-source variant named *KITTI CrossSource* proposed by us. The other is a large-scale cross-source dataset mostly consisting of unstructured forest scenes (Weiser et al. 2022). The results demonstrate that SPEAL is effective and robust for both same-source and cross-source point cloud registration, as well as for point clouds of unstructured scenes.

Overall, our contributions are threefold:

- We propose a novel learning-based point cloud registration approach, SPEAL, which is the first to utilize skeletal representations as a geometric prior to achieve improved performance.
- The proposed SEM is an effective and portable skeleton extractor. Our SAGTR combined with CDS effectively produces accurate and robust correspondences for both same-source and cross-source point cloud registration.
- KITTI CrossSource, a novel cross-source point cloud dataset, meets the dire need (Huang et al. 2021b) of cross-source point cloud registration benchmarks. This opens up the possibility to bridge the gap between sensor technology and cross-source applications.

Related Work

Learning-based Registration Methods. Learning-based registration methods fall into two categories: correspondence-based methods and direct registration methods. Correspondence-based methods (Choy, Park, and Koltun 2019; Deng, Birdal, and Ilic 2018a,b; Gojcic et al. 2019; Yao et al. 2020) first extract correspondences between two point clouds, and then estimate the transformation with robust pose estimators. However, traditional robust estimators suffer from slow convergence and are sensitive to outliers. To address this, deep robust estimators (Choy, Dong, and Koltun 2020; Bai et al. 2021; Pais et al. 2020; Lee et al. 2021) utilize deep neural networks to reject outliers and compute the transformation. While these methods require a training procedure, they improve accuracy and speed. Direct registration methods directly estimate the transformation between two point clouds in an end-to-end way. Inspired by Iterative Closest Point (Besl and McKay 1992), some of them (Fu et al. 2021; Wang and Solomon 2019b,a; Yew and Lee 2020) iteratively build soft correspondences and then estimate the transformation with SVD. Others (Xu et al. 2021; Aoki et al. 2019; Huang, Mei, and Zhang 2020) extract a global feature vector and regress the transformation directly with a neural network. However, such methods could potentially fail in large-scale scenes.

Transformers in Point Cloud Registration. Originally designed for NLP tasks, Transformers (Vaswani et al. 2017) have shown remarkable efficacy in computer vision (Misra, Girdhar, and Joulin 2021; Carion et al. 2020; Dosovitskiy et al. 2020; Yu et al. 2021b). Recently, transformer-based methods for point cloud registration have also emerged. Geometric Transformer (Qin et al. 2022) leverages transformer layers for superpoint matching, while REGTR (Yew and Lee 2022) uses a transformer cross-encoder and a transformer decoder to directly predict overlap scores. PEAL(Yu et al. 2023) leverages additional overlap priors from 2D images.

Point Cloud Skeletal Representations. The curve skeleton is a widely-used skeletal representation due to its simplicity (Huang et al. 2013; Ma, Wu, and Ouhyoung 2003; Au et al. 2008; Cao et al. 2010). It has shown its potential in some learning-based methods (Xu et al. 2019; Shi et al. 2021) like keypoint extraction. However, it is only well-defined for tubular geometries, thus limiting its expressiveness for point clouds with complex shapes or in large-scale scenes. The **M**edial **A**xis **T**ransform (MAT) (Blum 1967) is another skeletal representation capable of encoding arbitrary shapes. Some methods (Sun et al. 2015; Yan, Letscher, and Ju 2018; Li et al. 2015) employ simplification techniques to alleviate the distortion caused by surface noise, but they are computationally ineffective and require watertight input surfaces. Recent learning-based efforts (Lin et al. 2021; Wen, Yu, and Tao 2023) use deep neural networks to predict MAT-based skeletons, thus greatly enhancing the robustness and computational efficiency. These methods have shown promising results in various 3D vision tasks, including shape reconstruction and point cloud sampling (Wen, Yu, and Tao 2023).

Method

Problem Statement. Given two point clouds $\mathcal{P} = \{\mathbf{p}_i \in \mathbb{R}^3 | i = 1, \dots, N\}$ and $\mathcal{Q} = \{\mathbf{q}_i \in \mathbb{R}^3 | i = 1, \dots, M\}$, our goal is to align the two point clouds by estimating a rigid transformation $\mathbf{T} = \{\mathbf{R}, \mathbf{t}\}$, where $\mathbf{R} \in SO(3)$ is a 3D rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ is a 3D translation vector. The transformation can be solved by:

$$\min_{\mathbf{R},\mathbf{t}} \sum\nolimits_{(\mathbf{p}_{x_i}, \mathbf{q}_{y_i}) \in \mathcal{C}^\star} \|\mathbf{R}\mathbf{p}_{x_i} + \mathbf{t} - \mathbf{q}_{y_i}\|^2, \tag{1}$$

where $\mathcal{C}^\star$ denotes the set of correspondences between two point clouds $\mathcal{P}$ and $\mathcal{Q}$. In reality, $\mathcal{C}^\star$ is usually unknown. Hence, we need to establish accurate correspondences $\mathcal{C}$ between two point clouds for a good transformation.

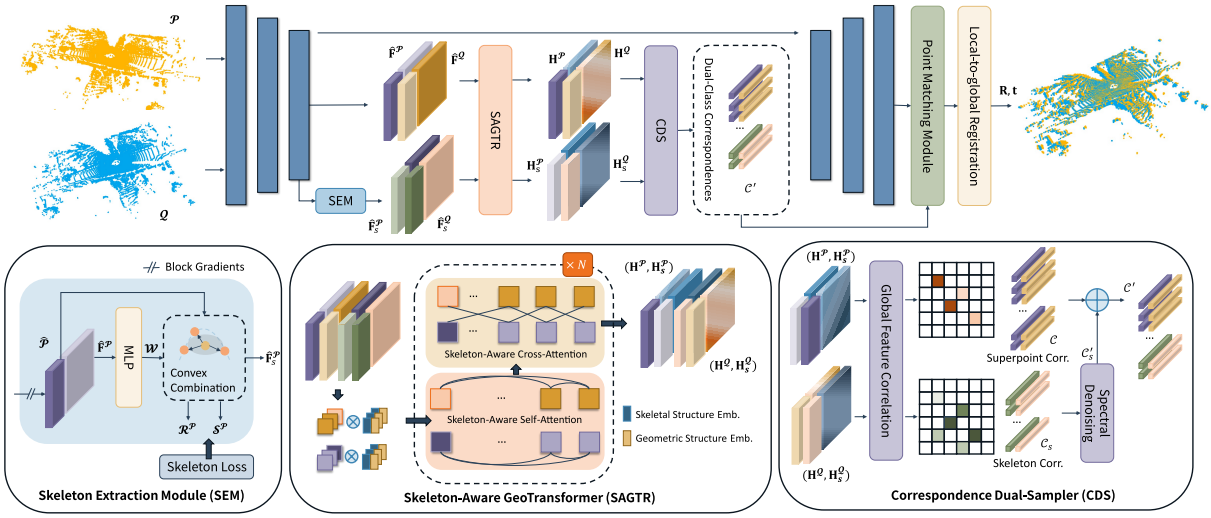


Figure 2: **The Overall Pipeline of SPEAL.** The backbone extracts superpoints and multi-level features from $\mathcal{P}$ and $\mathcal{Q}$. Then, SEM and SAGTR extract skeletal representations and learn discriminative skeleton-aware features, respectively. Finally, CDS extracts hybrid coarse correspondences with skeletal priors. The result transformation is computed with LGR.

Overview and Notations. Our work leverages skeletal priors in an end-to-end neural network to facilitate correspondences. The pipeline is shown in Fig. 2, following the hierarchical correspondence paradigm. To extract multi-level features for point clouds, we leverage the KPConv-FPN backbone (Lin et al. 2017; Thomas et al. 2019). The points at the coarsest level of the backbone are *superpoints*, denoted as $\hat{\mathcal{P}}$ and $\hat{\mathcal{Q}}$. Their associated features are $\hat{\mathbf{F}}^{\mathcal{P}} \in \mathbb{R}^{|\hat{\mathcal{P}}| \times d_t}$ and $\hat{\mathbf{F}}^{\mathcal{Q}} \in \mathbb{R}^{|\hat{\mathcal{Q}}| \times d_t}$. Then, our proposed SEM, SAGTR and CDS are used to extract reliable and accurate coarse correspondences with skeletal priors. Finally, we employ the Point Matching Module and Local-to-Global Registration (Qin et al. 2022) to obtain dense correspondences and estimate the final rigid transformation.

Skeleton Extraction Module

The Skeleton Extraction Module aims to approximate the Medial Axis Transform (MAT) by leveraging a convex combination of input points, which provides a well-defined skeletal representation for arbitrary shapes in an unsupervised manner. Inspired by existing methods (Lin et al. 2021), it overcomes the computational expense and sensitivity to surface noise of traditional MAT computation.

Specifically, for all points in $\hat{\mathcal{P}} \in \mathbb{R}^{|\hat{\mathcal{P}}| \times 3}$ and their features $\hat{\mathbf{F}}^{\mathcal{P}} \in \mathbb{R}^{|\hat{\mathcal{P}}| \times d_t}$, SEM aims to extract N_s skeleton points $\mathcal{S}^{\mathcal{P}} \in \mathbb{R}^{N_s \times 3}$, their skeletal features $\hat{\mathbf{F}}_s^{\mathcal{P}} \in \mathbb{R}^{N_s \times d_t}$, and their radii $\mathcal{R}^{\mathcal{P}} \in \mathbb{R}^{N_s \times 1}$. We extract skeletons for $\mathcal{Q}$ in the same way. To this end, we employ a multi-layer perceptron (MLP) to predict the weights $\mathcal{W} \in \mathbb{R}^{|\hat{\mathcal{P}}| \times N_s}$. The MLP is shared across $\hat{\mathcal{P}}$ and $\hat{\mathcal{Q}}$. Then, the skeleton points $\mathcal{S}^{\mathcal{P}}$ are obtained as the convex combination (Lin et al. 2021) of input points $\hat{\mathcal{P}}$:

$$\mathcal{S}^{\mathcal{P}} = \mathcal{W}^T \hat{\mathcal{P}} \; s.t. \; j = 1, \ldots, N_s, \; \sum_{i=1}^{|\hat{\mathcal{P}}|} \mathcal{W}(i,j) = 1 \quad (2)$$

The weighting scheme enhances the robustness of skeleton extraction by effectively filtering out noise and outliers. Similarly, we extract their skeletal features by $\hat{\mathbf{F}}_s^{\mathcal{P}} = \mathcal{W}^T \hat{\mathbf{F}}^{\mathcal{P}}$.

To predict the radius of each skeleton point, we first compute the closest distance for an input point $\hat{\mathbf{p}}$ to all skeleton points as follows:

$$d(\hat{\mathbf{p}}, \mathcal{S}^{\mathcal{P}}) = \min_{\mathbf{s} \in \mathcal{S}^{\mathcal{P}}} \|\hat{\mathbf{p}} - \mathbf{s}\|_2. \quad (3)$$

The distances for all input points are then summarized in a vector $\mathcal{D}^{\mathcal{P}} \in \mathbb{R}^{|\hat{\mathcal{P}}| \times 1}$. Next, the radii of all the skeleton points are computed through a linear combination of their closest distances from all the input points, *i.e.*, $\mathcal{R}^{\mathcal{P}} = \mathcal{W}^T \mathcal{D}^{\mathcal{P}}$. This approximation is based on the observation that the predicted weights for a skeleton point $\mathbf{s}$ are significant only for the input points that in a local neighborhood of $\mathbf{s}$, and diminish to 0 for the input points far away from $\mathbf{s}$.

The skeleton extraction is a fundamentally different task from the point cloud registration. Therefore, the module is separately supervised by the *skeleton loss* (Lin et al. 2021), and we block the gradient flow from this module to the backbone for a more stable training process and better performance (See supplementary materials).

Skeleton-Aware GeoTransformer

The registration of cross-source point clouds poses significant challenges, including noise, density differences, and scale variances. Skeleton points exhibit consistency and robustness against these challenges. Therefore, we propose the SAGTR module to encode the structure of point clouds. It comprises two key components: *Skeleton-aware Geometric Self-Attention* and *Skeleton-aware Cross-Attention*. They are interleaved for N_t times to further extract non-skeletal and skeletal hybrid features $(\mathbf{H}^{\mathcal{P}}, \mathbf{H}_s^{\mathcal{P}})$ and $(\mathbf{H}^{\mathcal{Q}}, \mathbf{H}_s^{\mathcal{Q}})$. These features encode inter-point-cloud and intra-point-cloud correlations and skeletal geometric priors. They contribute to accurate and robust coarse correspondences.

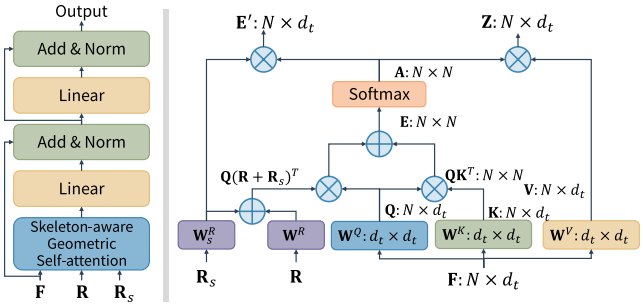


Figure 3: The structure (left) and computational graph (right) of skeleton-aware geometric self-attention.

Skeleton-Aware Geometric Self-Attention. In the following, we describe the computation for $\hat{\mathcal{P}}$, and the computation for $\hat{\mathcal{Q}}$ is exactly the same. Given an input feature matrix $\mathbf{X} \in \mathbb{R}^{L \times d_t}$ ($L = |\hat{\mathcal{P}}| + N_s$ is the length of the input sequence), the output feature matrix $\mathbf{Z} \in \mathbb{R}^{L \times d_t}$ is the weighted sum of all projected input features:

$$\mathbf{z}_i = \sum_{j=1}^{L} a_{i,j}(\mathbf{x}_j \mathbf{W}^V), \tag{4}$$

where $a_{i,j}$ is the weight coefficient computed by a row-wise softmax on the attention score $e_{i,j}$, and $e_{i,j}$ is computed as:

$$e_{i,j} = (\mathbf{x_i}\mathbf{W}^Q)(\mathbf{x_j}\mathbf{W}^K + \mathbf{r}^p_{i,j}\mathbf{W}^P + \mathbf{r}^s_{i,j}\mathbf{W}^S)^T / \sqrt{d_t}. \tag{5}$$

Fig. 3 shows the computation of Skeleton-aware Geometric Self-attention. Here, $\mathbf{r}^s_{i,j} \in \mathbf{R}_s$ and $\mathbf{r}^p_{i,j} \in \mathbf{R}$ are *Skeleton-Aware Structure Embedding* and *Point-Wise Structure Embedding*, respectively. We follow (Qin et al. 2022) to compute $\mathbf{r}^p_{i,j}$, which encodes non-skeletal geometric structures between superpoints. $\mathbf{r}^s_{i,j}$ encodes skeletal latent geometric information of point clouds, which will be described next. $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V, \mathbf{W}^P, \mathbf{W}^S \in \mathbb{R}^{d_t \times d_t}$ are the respective projections for queries, keys, values, point-wise structure embedding and skeleton-aware structure embedding.

We design a novel approach, termed *Skeleton-Aware Structure Embedding*, to encode skeletal latent structural information in the geometric space. The insight is to leverage the transformation invariance and robustness in the local geometric structure formed by the skeleton points. This embedding includes *skeleton-wise distance embedding* and *skeleton-wise angular embedding*. They respectively capture distance and angle information of the local geometric structure formed by skeleton points around superpoints.

Specifically, given two superpoints $\hat{\mathbf{p}}_i, \hat{\mathbf{p}}_j \in \hat{\mathcal{P}}$, their k-NN skeleton points are $\mathbf{s}^i_1, \ldots, \mathbf{s}^i_k \in \mathcal{K}^s_i$ and $\mathbf{s}^j_1, \ldots, \mathbf{s}^j_k \in \mathcal{K}^s_j$, respectively. Based on them, as shown in Fig. 4, the computation of skeleton-wise structure embedding $\mathbf{r}^s_{i,j}$ is twofold: 1) *Skeleton-Aware Distance Embedding.* For each superpoint $\mathbf{p}_j$, we first compute $\rho^s_j = \sum_{\mathbf{s}^j_x \in \mathcal{K}^s_j} d(\mathbf{s}^j_x, \mathbf{p}_j)$, where $d(\mathbf{s}^j_x, \mathbf{p}_j) = \|\mathbf{s}^j_x - \mathbf{p}_j\|_2$ denotes the distance between s_i and p_j in the Euclidean space. Then, the skeleton-aware distance embedding $\mathbf{d}^s_j$ is computed by applying a sinusoidal function on $(\rho^s_i - \rho^s_j)/\sigma^s_d$. 2) *Skeleton-Aware Angular Embedding.* For each skeleton point $s^j_x \in \mathcal{K}^s_j$, we first compute

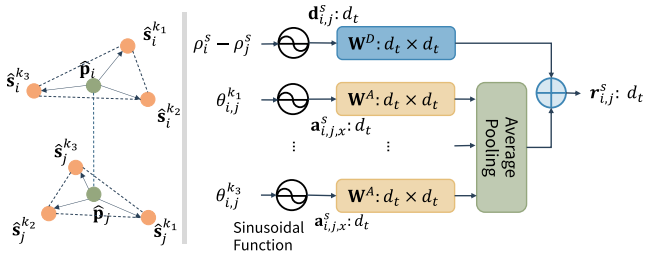


Figure 4: The computation of skeleton-aware structure embedding.

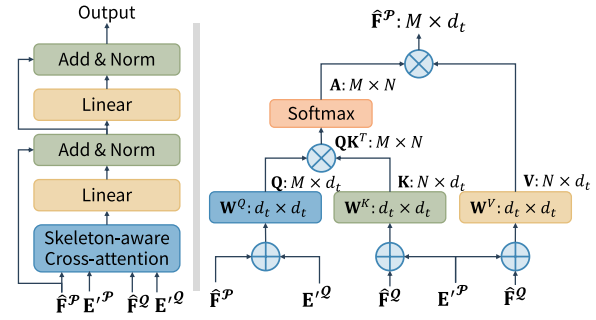


Figure 5: The structure (left) and computational graph (right) of skeleton-aware cross-attention.

the angle $\theta^x_{i,j} = \angle(\mathbf{s}^j_x - \mathbf{p}_j, \mathbf{p}_i - \mathbf{p}_j)$. Based on the angles, the skeleton-wise angular embedding $\mathbf{a}^s_{i,j,x}$ is computed by applying a sinusoidal function on $\theta^x_{i,j}/\sigma^s_a$. Herein, σ^s_d and σ^s_a control the sensitivity on skeleton-wise distances and angles respectively. The final skeleton-aware structure embedding $\mathbf{r}^s_{i,j}$ is the aggregation of the skeleton-aware angular embedding $\mathbf{a}^s$ and the skeleton-aware distance embedding $\mathbf{d}^s$:

$$\mathbf{r}^s_{i,j} = \mathbf{d}^s_{i,j}\mathbf{W}^D + \text{mean}_x\{\mathbf{a}^s_{i,j,x}\mathbf{W}^A\}, \tag{6}$$

where $\mathbf{W}^D$ and $\mathbf{W}^A$ are trainable weights.

To help the successive cross-attention layers to capture the geometric structure with skeletal priors, our skeleton-aware self-attention layers also produce the skeleton-aware positional encoding $\mathbf{E}'$ by applying the attention scores on the skeleton-aware structure embedding $\mathbf{r}^s_{i,j}$:

$$\mathbf{E}'_{i,k} = \sum_{j=1}^{L} a_{i,j} \cdot r^s_{i,j,k}. \tag{7}$$

Skeleton-Aware Cross-Attention. Several existing works (Qin et al. 2022; Yew and Lee 2022) have utilized the cross-attention mechanism for inter-point-cloud feature exchange. However, they either lack positional encoding or fail to explicitly consider the geometric structure of point clouds, leading to suboptimal performance. To address this, we propose the skeleton-aware cross-attention to explicitly learn the correlation of point clouds with skeletal priors, as is shown in Fig. 5.

Given feature maps with their skeleton-aware positional encoding $(\mathbf{X}^{\mathcal{P}}, \mathbf{E}'_{\mathcal{P}})$ and $(\mathbf{X}^{\mathcal{Q}}, \mathbf{E}'_{\mathcal{Q}})$ for $\hat{\mathcal{P}}$ and $\hat{\mathcal{Q}}$ respectively. A skeleton-aware cross-attention layer first adds the

positional encoding to features to produce skeleton-aware features $\mathbf{X}^{'\mathcal{P}}$ and $\mathbf{X}^{'\mathcal{Q}}$. Then, the output for $\hat{\mathcal{P}}$ are computed with the features of $\hat{\mathcal{Q}}$:

$$\mathbf{z}_i^{\mathcal{P}} = \sum_{j=1}^{|\hat{\mathcal{Q}}|} a_{i,j}(\mathbf{x}_j^{'\mathcal{Q}}\mathbf{W}^V). \tag{8}$$

Similarly, the weights $a_{i,j}$ are computed by a row-wise softmax on the attention score $e_{i,j}$:

$$e_{i,j} = (\mathbf{x}_i^{'\mathcal{P}}\mathbf{W}^Q)(\mathbf{x}_j^{'\mathcal{Q}}\mathbf{W}^K)^T/\sqrt{d_t}. \tag{9}$$

The same cross-attention implementation goes for $\hat{\mathcal{Q}}$. In contrast to skeleton-aware geometric self-attention that captures the intra-point-cloud transformation-invariant geometric structure, the cross-attention here captures the inter-point-cloud geometric corrections and consistency. The hybrid features obtained from SAGTR are therefore discriminative enough for matching.

Correspondence Dual-Sampler

With discriminative features, it is vital to extract accurate coarse correspondences. Geometric Transformer (Qin et al. 2022) only matches the superpoints. Despite its efficiency, superpoints are sparse and may be unrepeatable, leading to outlier correspondences. Existing efforts strive to tackle this issue with sophisticated sampling strategies (Li et al. 2023) or overlap priors from 2D images (Yu et al. 2023). Unfortunately, their required complicated sampling and extra 2D images result in suboptimal computational efficiency, which hinders the application of such methods. To this end, we propose CDS to effectively augment the correspondence set with our effective skeletal representation, leading to a more accurate and robust hybrid coarse correspondence set.

We separately construct the non-skeletal correspondence set $\mathcal{C}$ and skeletal correspondence set $\mathcal{C}_s$ by feature matching: We first compute the Gaussian correlation matrix $\mathbf{S} \in \mathbb{R}^{|\hat{\mathcal{P}}|\times|\hat{\mathcal{Q}}|}$ for the normalized features $\mathbf{F}^{\mathcal{P}}$ and $\mathbf{F}^{\mathcal{Q}}$, and then use a dual-normalization operation (Sun et al. 2021; Rocco et al. 2018) to suppress ambiguous matches. Finally, we select at most N_c largest entries for each correspondence set.

Since skeleton points lying on non-overlap regions may introduce outlier correspondences, we introduce the *Spectral Denoising* procedure to filter $\mathcal{C}_s$ with a spectral matching algorithm (Leordeanu and Hebert 2005): We firstly compute a compatibility matrix based on the 3D spatial consistency of $\mathcal{C}_s$. Then, we iteratively remove components conflicting with the item of the maximum principal eigenvector until either the principal eigenvector becomes zero or $|\mathcal{C}_s|$ equals the minimum number of the main cluster. The main cluster, denoted as $\mathcal{C}_s^{'}$, is the final skeletal correspondence set.

Finally, we resample the least confident N_s entries of $\mathcal{C}$ with top N_k correspondences in $\mathcal{C}_s^{'}$ to obtain the hybrid correspondence set $\mathcal{C}^{'}$, thereby improving the accuracy and robustness of the hybrid coarse correspondence set by replacing potential outliers with more reliable correspondences.

Losses

We use a registration loss and a skeleton loss to supervise SPEAL. The registration loss consists of *Overlap-aware Circle Loss* ($\mathcal{L}_{oc}$) and *Point Matching Loss* ($\mathcal{L}_p$) from Geometric Transformer (Qin et al. 2022):

$$\mathcal{L} = \mathcal{L}_{oc} + \mathcal{L}_p. \tag{10}$$

The skeleton loss in Lin et al. (2021) is used to supervise the SEM. It is the weighted sum of *Sampling Loss* $\mathcal{L}_{\text{s}}$, *Point-to-sphere Loss* $\mathcal{L}_{\text{r}}$ and *Radius Regularizing Loss* $\mathcal{L}_{\text{p2s}}$:

$$\mathcal{L}_{\text{skeleton}} = \mathcal{L}_{\text{s}} + \lambda_1\mathcal{L}_{\text{p2s}} + \lambda_2\mathcal{L}_{\text{r}}, \tag{11}$$

where λ_1 and λ_2 are hyperparameters to balance the losses. Please refer to the supplementary material for more details.

Experiments

Datasets and Experimental Setup

Same-Source Dataset. The KITTI Odometry dataset (Geiger, Lenz, and Urtasun 2012) serves as a widely-used dataset for odometry and SLAM evaluation. It can also be employed to test same-source point cloud registration. This dataset comprises 11 sequences of LiDAR point clouds. We follow the existing practices (Qin et al. 2022; Huang et al. 2021a) to use sequences 00-06 for training, sequences 07-08 for evaluation and sequences 09-10 for testing.

Cross-Source Datasets. Currently, there are few cross-source datasets of large-scale outdoor scenes available for registration tasks. This hinders the development of cross-source registration methods. Therefore, we have developed a novel dataset[1] termed *KITTI CrossSource* derived from KITTI Odometry. Our proposed dataset includes 11 sequences of LiDAR point clouds and reconstructed point clouds generated from stereo images using MonoRec (Wimbauer et al. 2021). We improve the reconstruction quality with a filter-and-combine strategy. Please refer to the supplementary material for details.

The GermanyForest3D dataset is derived from an existing large-scale forest scene dataset (Weiser et al. 2022). It contains cross-source point cloud data acquired in 12 forest plots in south-west Germany under leaf-on and leaf-off conditions. Each plot provides Airborne Laser Scanning (ALS), Terrestrial Laser Scanning (TLS) and UAV-borne Laser Scanning (ULS) point clouds. In this paper, we use ALS and ULS scans to evaluate the cross-source registration performance. In experiments, we use 10 plots for training, 1 for validation and 1 for testing.

Data Preprocessing. For the GermanyForest3D dataset, the point clouds of each plot are subdivided into $30\text{m} \times 30\text{m} \times 30\text{m}$ blocks to make them suitable for the registration task. For all datasets, the Iterative Closest Point (ICP) algorithm from the Open3D library (Zhou, Park, and Koltun 2018) is used to refine the noisy ground truth transformation, following previous works (Qin et al. 2022; Lu et al. 2021). The point clouds are downsampled with a voxel size of 0.3m.

Metrics. Following previous practices (Qin et al. 2022; Lu et al. 2021; Huang et al. 2021a), we evaluate the registration performance using following metrics: *Relative Rotation Error* (RRE), *Relative Translation Error* (RRE), and *Registration Recall* (RR). We use a RRE threshold and a RTE

[1]The dataset will be made publicly available. Please refer to github.com/kezheng1204/KITTI-CrossSource for updates.

Method	KITTI CrossSource			KITTI Odometry		
	RRE(°)	RTE(m)	RR(%)	RRE(°)	RTE(m)	RR(%)
RANSAC	6.14	9.46	0.8	0.54	0.13	91.9
FGR	–	–	–	0.96	0.93	39.4
FCGF	–	–	–	0.30	0.095	96.6
DGR	–	–	–	0.37	0.320	98.7
HRegNet	2.19	0.84	69.3	0.29	0.120	99.7
CoFiNet	1.99	0.81	67.6	0.41	0.085	99.8
Predator	5.06	2.59	34.2	0.27	**0.068**	98.8
PCAM	4.07	2.40	45.9	0.79	0.12	98.0
GeoTrans.	1.87	0.63	96.8	0.24	**0.068**	**99.8**
SPEAL	**1.41**	**0.58**	**97.3**	**0.23**	0.069	**99.8**

Table 1: Cross-source (the proposed KITTI CrossSource) and same-source (KITTI Odometry) registration results on the KITTI datasets. "–" indicates the method is not applicable to the dataset.

threshold to compute RR for all datasets (RRE $< 0.5°$ and RTE $<$ 0.3m for GermanyForest3D and RRE $< 5°$ and RTE $<$ 2m for KITTI datasets). Additionally, we measure the quality of correspondences with Inlier Ratio (IR), which is the fraction of extracted correspondences whose residuals are below a certain threshold under the ground-truth transformation.

Implementation Details. To train SPEAL, we use Adam (Kingma and Ba 2014) optimizer with an initial learning rate of $1e-4$ and a weight decay of $1e-6$. We train SPEAL for 200 epochs with a batch size of 1 on a NVIDIA RTX 3090 GPU.

Baselines. We compare our method with state-of-the-art methods of three classes: (a) Traditional methods, including RANSAC (Fischler and Bolles 1981) and FGR (Zhou, Park, and Koltun 2016). (b) Transformer-based methods, including CoFiNet (Yu et al. 2021a), Predator(Huang et al. 2021a), PCAM (Cao et al. 2021), REGTR (Yew and Lee 2022) and Geometric Transformer (abbreviated as GeoTrans.) (Qin et al. 2022). (c) Other learning-based methods, including FCGF (Choy, Park, and Koltun 2019), DGR (Choy, Dong, and Koltun 2020) and HRegNet (Lu et al. 2021).

Cross-Source Results

KITTI CrossSource. The quantitative results are reported in Table 1. Our method achieves state-of-the-art performance on this dataset. For traditional methods, FGR is not applicable to this dataset and our approach outperforms RANSAC by a large margin. HRegNet is a recent SOTA for outdoor large-scale scenes. However, it presents suboptimal performance on this dataset, showing considerable performance decay for cross-source data. In contrast, our SPEAL is more accurate in terms of all metrics, and has a 28% higher RR than HRegNet. Among transformer-based methods, our method surpasses GeoTrans. by a large margin, showing the effectiveness of the integrated skeletal priors.

GermanyForest3D. This cross-source dataset is with large scale and unstructured scenes, which are challenging for registration. The evaluation results under different overlap ratios are shown in Table 2. Traditional methods show suboptimal performance, and RANSAC even fails to register low overlap point clouds. Learning-based methods overall perform better. However, current methods still suffer from considerable performance decay especially under low overlap condition. SPEAL outperforms all the others by a large margin, showing outstanding robustness introduced by skeletal priors in low overlap and unstructured condition.

Overlap	RRE (°) ↓		RTE (m) ↓		RR(%) ↑	
	$\leq 30\%$	$> 30\%$	$\leq 30\%$	$> 30\%$	$\leq 30\%$	$> 30\%$
RANSAC	112.0	91.1	24.66	17.4	–	–
FGR	41.86	28.3	14.32	8.49	–	–
FCGF	1.54	0.53	0.49	0.19	8.7	56.6
DGR	1.06	0.38	0.36	0.10	32.8	76.2
HRegNet	1.16	1.40	0.141	0.238	24.7	41.7
REGTR	3.11	2.48	0.89	0.73	11.3	23.9
GeoTrans.	0.328	0.176	0.097	0.053	88.1	96.5
SPEAL$_{(ours)}$	**0.296**	**0.165**	**0.088**	**0.048**	**91.7**	**99.3**

Table 2: Cross-source registration results on the GermanyForest3D dataset. "–" indicates that the method is not applicable to the dataset.

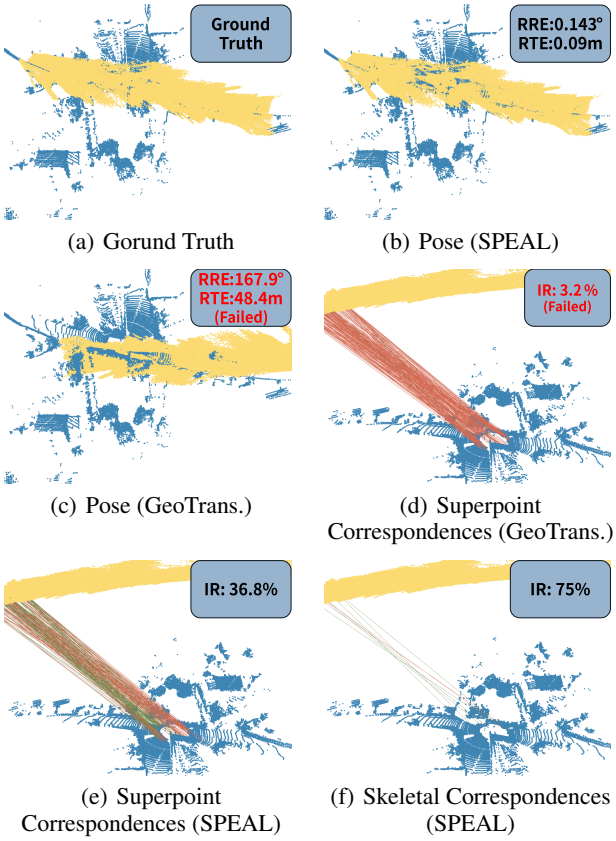
Same-Source Results

Table 1 also lists the quantitative results on the same-source dataset KITTI Odometry. Compared with recent state-of-the-arts, our method achieves comparable performance in terms of RTE and RR, and outperforms all the other methods in terms of RRE. This result indicates that our method is also effective for same-source registration, while achieving state-of-the-art performance for cross-source registration.

Analysis

Effectiveness of the Skeletal Priors. To qualitatively verify the effectiveness of the skeletal representation, we visualize the dual-class correspondences from the CDS module, including superpoints and skeleton points. The qualitative results are shown in Fig. 6. In addition to the challenges of partial overlap and density differences, this scan also presents the challenge of unstructured objects. However, SPEAL is still able to extract right correspondences with the help of skeletons, while the current state-of-the-art, GeoTrans., completely fails. SPEAL achieves an IR of 36.8%, which is nearly $10\times$ higher than GeoTrans. It is worth noting that with our spectral denoising step, the skeletal correspondences achieve an IR of 75%, demonstrating the effectiveness of the spectral denoising step.

Robustness. Fig. 7(a) displays registration recalls with different RRE and RTE thresholds in KITTI CrossSource. SPEAL consistently outperforms the other methods. In particular, it achieves a registration recall of 86.04% in the challenging low overlap of $30\% \sim 50\%$, which is 9.3% higher than the second-best method. In addition, Fig. 7(b) compares registration recalls and inlier ratios under different overlap with GeoTrans. Our method consistently achieves higher inlier ratios under all overlap ratios. This demonstrates the superior quality of correspondences generated by our method with skeletal priors. The results prove that our SPEAL is robust to various threshold settings, and demonstrates outstanding performance in low overlap condition.

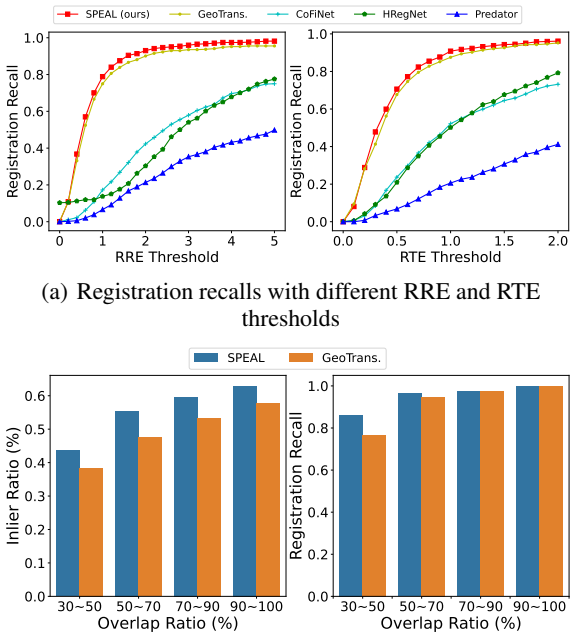


(a) Gorund Truth (b) Pose (SPEAL)

(c) Pose (GeoTrans.) (d) Superpoint Correspondences (GeoTrans.)

(e) Superpoint Correspondences (SPEAL) (f) Skeletal Correspondences (SPEAL)

Figure 6: Qualitative results on KITTI CrossSource. Red and green denotes outlier and inlier correspondences, respectively.



(a) Registration recalls with different RRE and RTE thresholds

(b) Correspondence and registration results under different overlap ratios

Figure 7: Quantitative results on robustness of SPEAL on the KITTI CrossSource dataset.

Ablation Studies

Overall Effectiveness. We conduct ablation studies to assess the effectiveness of SPEAL on GermanyForest3D. We compare different configurations of SPEAL, including: (a) vanilla geometric self-attention and vanilla cross attention, (b) skeleton-aware self-attention and vanilla cross attention, (c) skeleton-aware self-attention and skeleton-aware cross attention. In addition, we also compare with (d) the method without the CDS module, which only samples the coarse correspondences from superpoints. The results in Table 3 demonstrate the effectiveness of our design.

Model	RRE(°)↓	RTE(m)↓	RR(%)↑
(a) vanilla	0.19	0.054	96.2
(b) cross attn. w/o SPE	0.18	0.051	97.4
(c) w/ SSE & SPE	0.18	0.049	97.1
(d) corr. w/o CDS	0.19	0.051	96.3
(e) SPEAL (Ours)	**0.16**	**0.048**	**99.3**

Table 3: Ablation studies on SPEAL.

Spectral Denoising in CDS. To validate the effectiveness of spectral denoising, we also ablate the CDS module. We compare different schemes for applying the spectral denoising.

Spectral Denoising		RRE(°)↓	RTE(m)↓	RR(%)↑
Sup. Corr.	Skel. Corr.			
		0.17	0.051	98.1
✓		0.20	0.063	97.8
✓	✓	0.19	0.062	98.1
	✓	**0.16**	**0.048**	**99.3**

Table 4: Ablation studies on the CDS module. Sup. Corr. and Skel. Corr. denote the superpoint correspondences and skeletal correspondences, respectively.

The results in Table 4 demonstrate that the spectral denoising step is only necessary for the skeletal correspondences and leads to inferior performance in other configurations.

Conclusion

In this paper, we have proposed SPEAL, a novel point cloud registration method that leverages a MAT-based skeletal representation to capture the geometric intricacies, thereby facilitating registration. Our method introduces SEM to extract the skeleton points and their skeletal features. Furthermore, we design SAGTR and CDS which explicitly integrate skeletal priors to ensure robust and accurate correspondences. Extensive experiments demonstrate that SPEAL is effective for both same-source and cross-source point cloud registration.

Acknowledgement. This work was supported in part by the National Key R&D Program of China under Grant 2021YFF0704600, the Fundamental Research Funds for the Central Universities (No. 20720220064).

References

Aoki, Y.; Goforth, H.; Srivatsan, R. A.; and Lucey, S. 2019. Pointnetlk: Robust & efficient point cloud registration using pointnet. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7163–7172.

Au, O. K.-C.; Tai, C.-L.; Chu, H.-K.; Cohen-Or, D.; and Lee, T.-Y. 2008. Skeleton extraction by mesh contraction. *ACM transactions on graphics (TOG)*, 27(3): 1–10.

Bai, X.; Luo, Z.; Zhou, L.; Chen, H.; Li, L.; Hu, Z.; Fu, H.; and Tai, C.-L. 2021. Pointdsc: Robust point cloud registration using deep spatial consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15859–15869.

Besl, P. J.; and McKay, N. D. 1992. Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, 586–606. Spie.

Blum, H. 1967. A transformation for extracting new descriptions of shape. *Models for the perception of speech and visual form*, 362–380.

Cao, A.-Q.; Puy, G.; Boulch, A.; and Marlet, R. 2021. PCAM: Product of cross-attention matrices for rigid registration of point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13229–13238.

Cao, J.; Tagliasacchi, A.; Olson, M.; Zhang, H.; and Su, Z. 2010. Point cloud skeletons via laplacian based contraction. In *2010 Shape Modeling International Conference*, 187–197. IEEE.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 213–229. Springer.

Choy, C.; Dong, W.; and Koltun, V. 2020. Deep global registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2514–2523.

Choy, C.; Park, J.; and Koltun, V. 2019. Fully convolutional geometric features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8958–8966.

Deng, H.; Birdal, T.; and Ilic, S. 2018a. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In *Proceedings of the European conference on computer vision (ECCV)*, 602–618.

Deng, H.; Birdal, T.; and Ilic, S. 2018b. Ppfnet: Global context aware local features for robust 3d point matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 195–205.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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

Fu, K.; Liu, S.; Luo, X.; and Wang, M. 2021. Robust point cloud registration framework based on deep graph matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8893–8902.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

Gojcic, Z.; Zhou, C.; Wegner, J. D.; and Wieser, A. 2019. The perfect match: 3d point cloud matching with smoothed densities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5545–5554.

Huang, H.; Wu, S.; Cohen-Or, D.; Gong, M.; Zhang, H.; Li, G.; and Chen, B. 2013. L1-medial skeleton of point cloud. *ACM Trans. Graph.*, 32(4): 65–1.

Huang, S.; Gojcic, Z.; Usvyatsov, M.; Wieser, A.; and Schindler, K. 2021a. Predator: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 4267–4276.

Huang, X.; Mei, G.; and Zhang, J. 2020. Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11366–11374.

Huang, X.; Mei, G.; Zhang, J.; and Abbas, R. 2021b. A comprehensive survey on point cloud registration. *arXiv preprint arXiv:2103.02690*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, J.; Kim, S.; Cho, M.; and Park, J. 2021. Deep hough voting for robust global registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15994–16003.

Leordeanu, M.; and Hebert, M. 2005. A spectral technique for correspondence problems using pairwise constraints. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, 1482–1489. IEEE.

Li, P.; Wang, B.; Sun, F.; Guo, X.; Zhang, C.; and Wang, W. 2015. Q-mat: Computing medial axis transform by quadratic error minimization. *ACM Transactions on Graphics (TOG)*, 35(1): 1–16.

Li, Y.; Tang, C.; Yao, R.; Ye, A.; Wen, F.; and Du, S. 2023. HybridPoint: Point Cloud Registration Based on Hybrid Point Sampling and Matching. *arXiv preprint arXiv:2303.16526*.

Lin, C.; Li, C.; Liu, Y.; Chen, N.; Choi, Y.-K.; and Wang, W. 2021. Point2skeleton: Learning skeletal representations from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4277–4286.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lu, F.; Chen, G.; Liu, Y.; Zhang, L.; Qu, S.; Liu, S.; and Gu, R. 2021. Hregnet: A hierarchical network for large-scale outdoor lidar point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16014–16023.

Ma, W.-C.; Wu, F.-C.; and Ouhyoung, M. 2003. Skeleton extraction of 3D objects with radial basis functions. In *2003 Shape Modeling International.*, 207–215. IEEE.

Misra, I.; Girdhar, R.; and Joulin, A. 2021. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2906–2917.

Pais, G. D.; Ramalingam, S.; Govindu, V. M.; Nascimento, J. C.; Chellappa, R.; and Miraldo, P. 2020. 3dregnet: A deep neural network for 3d point registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7193–7203.

Qin, Z.; Yu, H.; Wang, C.; Guo, Y.; Peng, Y.; and Xu, K. 2022. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11143–11152.

Rocco, I.; Cimpoi, M.; Arandjelović, R.; Torii, A.; Pajdla, T.; and Sivic, J. 2018. Neighbourhood consensus networks. *Advances in neural information processing systems*, 31.

Shi, R.; Xue, Z.; You, Y.; and Lu, C. 2021. Skeleton merger: an unsupervised aligned keypoint detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 43–52.

Sun, F.; Choi, Y.-K.; Yu, Y.; and Wang, W. 2015. Medial meshes–a compact and accurate representation of medial axis transform. *IEEE transactions on visualization and computer graphics*, 22(3): 1278–1290.

Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; and Zhou, X. 2021. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8922–8931.

Thomas, H.; Qi, C. R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; and Guibas, L. J. 2019. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6411–6420.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y.; and Solomon, J. M. 2019a. Deep closest point: Learning representations for point cloud registration. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3523–3532.

Wang, Y.; and Solomon, J. M. 2019b. Prnet: Self-supervised learning for partial-to-partial registration. *Advances in neural information processing systems*, 32.

Weiser, H.; Schäfer, J.; Winiwarter, L.; Krašovec, N.; Fassnacht, F. E.; and Höfle, B. 2022. Individual tree point clouds and tree measurements from multi-platform laser scanning in German forests. *Earth System Science Data*, 14(7): 2989–3012.

Wen, C.; Yu, B.; and Tao, D. 2023. Learnable Skeleton-Aware 3D Point Cloud Sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17671–17681.

Wimbauer, F.; Yang, N.; Von Stumberg, L.; Zeller, N.; and Cremers, D. 2021. MonoRec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6112–6122.

Xu, H.; Liu, S.; Wang, G.; Liu, G.; and Zeng, B. 2021. Omnet: Learning overlapping mask for partial-to-partial point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3132–3141.

Xu, Z.; Zhou, Y.; Kalogerakis, E.; and Singh, K. 2019. Predicting animation skeletons for 3d articulated models via volumetric nets. In *2019 international conference on 3D vision (3DV)*, 298–307. IEEE.

Yan, Y.; Letscher, D.; and Ju, T. 2018. Voxel cores: Efficient, robust, and provably good approximation of 3d medial axes. *ACM Transactions on Graphics (TOG)*, 37(4): 1–13.

Yao, Y.; Deng, B.; Xu, W.; and Zhang, J. 2020. Quasi-newton solver for robust non-rigid registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7600–7609.

Yew, Z. J.; and Lee, G. H. 2020. Rpm-net: Robust point matching using learned features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11824–11833.

Yew, Z. J.; and Lee, G. H. 2022. Regtr: End-to-end point cloud correspondences with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6677–6686.

Yu, H.; Li, F.; Saleh, M.; Busam, B.; and Ilic, S. 2021a. Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *Advances in Neural Information Processing Systems*, 34: 23872–23884.

Yu, J.; Ren, L.; Zhang, Y.; Zhou, W.; Lin, L.; and Dai, G. 2023. PEAL: Prior-Embedded Explicit Attention Learning for Low-Overlap Point Cloud Registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17702–17711.

Yu, X.; Rao, Y.; Wang, Z.; Liu, Z.; Lu, J.; and Zhou, J. 2021b. Pointr: Diverse point cloud completion with geometry-aware transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12498–12507.

Zhou, Q.-Y.; Park, J.; and Koltun, V. 2016. Fast global registration. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 766–782. Springer.

Zhou, Q.-Y.; Park, J.; and Koltun, V. 2018. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847*.