Heritage Science

COMMENT

Open Access

Diffusion Transformer for point cloud registration: digital modeling of cultural heritage

Li An[1†], Pengbo Zhou[2†], Mingquan Zhou[1*], Yong Wang[1*] and Guohua Geng[1]

Abstract

Digital modeling is an essential means for preserving and passing down historical culture within cultural heritage. Point cloud registration technology, by aligning point cloud data captured from multiple perspectives, enhances the accuracy of reconstructing the complex structures of artifacts and buildings and provides a reliable digital foundation for their protection, exhibition, and research. Due to the challenges posed by complex morphology, noise, and missing data when processing cultural heritage data, this paper proposes a point cloud registration method based on the Diffusion Transformer (PointDT). Compared to traditional methods, the Diffusion Transformer can better capture both the global features and local structures of point cloud data, more accurately capturing the geometric and semantic information of the target point cloud, thereby achieving precise digital reconstruction. In this study, we trained our method using indoor datasets such as 3DMatch and large-scale outdoor datasets like KITTI, and validated it on various cultural heritage datasets, including those of the Terracotta Warriors and heritage buildings. The results demonstrate that this method not only significantly improves accuracy but also shows advantages in computational efficiency.

Keywords Point cloud registration, Diffusion transformer, Cultural heritage, Digital modeling

Introduction

Modeling of digital cultural heritage is an interdisciplinary field at the intersection of contemporary technology and cultural heritage preservation, with profound and underestimated significance. Over time and under the influence of natural forces, many precious cultural heritages have gradually lost their original appearance and are even on the brink of destruction. With the rapid development of digital technology, point cloud technology, as one of the important means of digital modeling, has received increasing attention and application. Point cloud technology converts the vast coordinate data collected from object surfaces through methods such as laser scanning into digital three-dimensional models, providing strong support for the protection, research, and exhibition of cultural heritage [1–3].

In the process of digital modeling, point cloud technology plays a crucial role. Point cloud data consists of a large number of discrete points in three-dimensional space, accurately describing the surface morphology and structural characteristics of objects. By capturing and processing point cloud data, we can generate high-precision three-dimensional models, providing important technical support for the protection and research of cultural heritage. However, due to the complexity

†Li An and Pengbo Zhou contribute equally to this work.

*Correspondence:
Mingquan Zhou
mqzhou@nwu.edu.cn
Yong Wang
202110326@stumail.nwu.edu.cn
[1] School of Information Science and Technology, Northwest University, Xian 710127, China
[2] School of Arts and Communication, Beijing Normal University, Beijing 100875, China

and incompleteness of point cloud data, point cloud registration has become a crucial issue [4–6].

With the continuous advancement of technology, point cloud registration technology is also constantly developing and evolving. Traditional registration methods are mainly based on feature matching or optimization techniques, which have achieved good results in some cases but still face many challenges when dealing with complex cultural heritage data [7, 8]. For example, cultural heritage itself has complex forms and structures, often accompanied by a large amount of noise and missing data, which place higher demands on the accuracy and robustness of registration. In addition, point cloud data from different sources may have differences, such as resolution, sampling density, etc., which also pose difficulties for the registration process.

In recent years, researchers have proposed many innovative methods to improve the effectiveness of point cloud registration. For example, learning-based local feature descriptors can enhance the robustness and accuracy of registration [9, 10]; registration methods based on graph neural networks [11, 12] can fully utilize the topological structure information of point clouds; and using deep learning techniques [4, 13, 14] for noise and missing data repair has also become a research hotspot.

In response to the various challenges faced by traditional registration techniques in dealing with complex cultural heritage data, including issues such as complex morphology, noise, and missing data, this paper proposes an end-to-end diffusion Transformer approach. Innovatively introducing the diffusion transformer, it can more effectively capture the global features and local structures of point cloud data. The diffusion process allows nodes to gradually adjust their own feature representations during iteration, effectively handling noise. Meanwhile, shape descriptors and local features are introduced into the diffusion model to enhance the model's understanding of point cloud data, further improving the accuracy and robustness of registration. The remaining sections of this paper are organized as follows: Section II will review related work, presenting the latest developments and relevant technologies in the field of point cloud registration. Section III will provide a detailed description of our proposed Diffusion Transformer (PointDT) framework, including Shared and Encoded Point Clouds, Co-Context Information Extraction, and Overlapping Region Matches. Section IV will introduce our experimental design and results, followed by analysis and discussion of the experimental outcomes. Finally, we will conclude the paper and propose future research directions.

Related work

Deep feature learning: In recent years, the advancement of deep learning and neural networks has led to significant breakthroughs in point cloud registration, which is of great importance for the digital modeling of cultural heritage. These methods achieve registration by learning feature representations and registration models for point clouds. Zeng et al. [15] introduced the data-driven 3DMatch algorithm, which employs a 3D ConvNet to learn local geometric descriptors. It utilizes an optimized method to calculate the transformation matrix between two overlapping point clouds. Different from the 3DMatch algorithm, the Ppfnet algorithm [16] and unsupervised Ppf-foldnet algorithm [10] are based on PointNet. These algorithms directly extract local features from point cloud data and calculate global features by aggregating the local features of other point sets. Xu et al. [17] proposed a local-local point cloud registration algorithm based on global features. It addresses the negative impact of non-overlapping regions on the entire network by learning overlapping templates. To address the issues related to the poor robustness of rotation-invariant feature structures and low repeatability of key point detection in existing point cloud registration algorithms, Wang et al. [18] effectively improved the accuracy of point cloud registration. Yan et al. [19] introduced a new hybrid optimization method that effectively solves the local and global point matching problem by optimizing local and projective losses. Piotr et al. [8] proposed a Fast Adaptive Multimodal Feature Registration (FAMFR) method capable of accurately registering two point clouds representing various cultural heritage interiors. FAMFR is based on two different handcrafted features, utilizing the color and shape of objects to precisely register point clouds with either rich surface geometric details or those with geometric deficiencies but rich color decorations. Markiewicz et al. [1] evaluated the quality and completeness of the TLS registration process using 2D raster data in the form of spherical images and affine handcrafted and learned-based detectors in multi-stage terrestrial laser scanning (TLS) point cloud registration as test data. This approach effectively addresses the registration for both less textured buildings and test sites with rich textures and numerous decorations.

End-to-end registration: With significant progress in deep learning, point cloud registration research has seen further exploration of learning-based methods. Among them, end-to-end point cloud registration offers the advantage of faster calculation speed and the ability to learn advanced features, leading to higher robustness. Lu et al. [20] proposed the DeepVCP algorithm, which introduced an end-to-end learning framework for point cloud registration. Based on the PointNet++

framework, the algorithm learned semantic features and effectively addressed the problem of local sparsity in point clouds through a point cloud generation method during the feature descriptor extraction process. Bai et al. [21] focused on convolution problems on irregular point clouds and the extraction of local geometric information. Using KPConv [22], they extracted features from dense and irregular point clouds and combined normalization operations to address the sparsity issue in point clouds. The Predator algorithm [23] aimed to enrich the information of two sparsely overlapping point clouds. It utilized cross-attention blocks to exchange information between the two sets of point clouds and predicted point saliency and the overlapping probability of the target point cloud during decoding. To enhance the robustness of the algorithm model and handle outliers, Zhang et al. [24] proposed a neural network-based method for point cloud registration by learning a partial permutation matrix. This method enabled an end-to-end point cloud registration process by directly learning the local correspondence between point clouds. For the issue of partial overlaps relying too heavily on labels, Mei et al. [4] proposed a framework for unsupervised depth probabilistic registration of point clouds with partial overlaps. The algorithm employed a network to learn the posterior probability distribution of a Gaussian mixture model (GMM) from a point cloud and used the Sinkhorn algorithm to predict distribution-level correspondences constrained by the GMM mixture weights. Unsupervised learning was achieved through the loss of distribution consistency.

Transformer: In recent years, researchers have started to apply the Transformer model to point cloud registration. To address the issue of local optima in the ICP class of point cloud registration algorithms, Wang et al. [25] proposed a learning-based method. The algorithm first learns initial features using the DGCNN algorithm, embeds the position information, and passes it through a Transformer structure. The relative pose is then estimated using SVD. However, this method requires a reference point cloud for alignment, which must be of good quality and accuracy. Fu et al. [26] tackled the problem of outlier sensitivity in learning-based point cloud registration algorithms by proposing a method based on depth map matching. The algorithm establishes a graph structure and corresponding relationships for initial features, which are then inputted into a Transformer structure to create an edge generator, improving correspondence quality. Liu et al. [27] introduced an end-to-end Transformer network for large-scale point cloud alignment. This method addresses challenges such as a large number of points, complex distribution, and outlier sensitivity in registering large-scale scenes. It captures long-range correlations and filters outliers by globally extracting point features. However, the algorithm divides the point cloud data into local blocks for feature extraction, which can reduce computational complexity but may result in the loss or inaccuracy of local relationships. The REGTR algorithm [28] directly learns a feature matrix obtained through dimensionality reduction using a Transformer network. This feature matrix is then used to predict the probability of overlapping point cloud regions and their corresponding positions in the source point cloud. Compared to traditional methods, this approach can directly learn more accurate correspondences from downsampling feature matrix without additional feature learning and optimization processes. However, representing point cloud data as a sequence in this algorithm may lead to the loss of local structure and geometric information. While the Transformer model can capture the global relationship of point cloud sequences, it may not be sufficient for modeling local features.

Diffusion models: Recently, diffusion models have played a significant role in tasks such as image denoising, image enhancement, and image segmentation. For instance, in image denoising, diffusion models can smooth images and reduce noise by averaging the values of each pixel with its neighboring pixels. In image segmentation, diffusion models facilitate information propagation across the image, grouping similar pixels together and enabling pixel clustering for object segmentation. In image tasks, Transformers establish connections between different regions of an image, capturing global contextual information. For example, in image generation tasks, Transformers can learn to transform different parts of an input image into corresponding parts of an output image, achieving high-quality image generation. In image classification tasks, Transformers can learn to weight features at different scales and orientations, thus improving classification accuracy.

The integration of diffusion models and Transformers effectively extracts local features and gains comprehensive information when processing image data. Initially, diffusion models propagate information within local regions by interacting between pixels, smoothing and consolidating data. This propagation process aids in fusing features of adjacent pixels, thereby reducing noise, smoothing details, and capturing local structures. Through diffusion, local features propagate and are shared within local areas. Subsequently, Transformers introduce attention mechanisms that establish connections between different regions of an image. This empowers Transformers to capture global contextual information and identify crucial relationships among diverse regions. When diffusion models are combined with Transformers, the latter can leverage the smooth information generated by the diffusion propagation

process, leading to a better understanding of local features within the global context.

In summary, the amalgamation of diffusion models and Transformers efficiently extracts local features and obtains comprehensive information while processing image data. This combined approach holds potential advantages in tasks such as image recognition, segmentation, and generation.

Inspired by the diffusion processes used in image processing and computer vision tasks to simulate the spread of information between image pixels or feature points, thereby enhancing feature representation and information fusion, we propose a novel point cloud registration algorithm utilizing a diffusion Transformer to address the challenges of complexity and diversity in digital modeling of cultural heritage artifacts. Compared to traditional methods, the diffusion Transformer captures the subtle structures and global features of artifacts more effectively by simulating the diffusion process of features between point cloud nodes, significantly improving the accuracy and robustness of registration.

Method

Similar to the structures of D3Feat [21] and Predator [23], our method adopts a hierarchical structure as a whole, and the specific process is illustrated in Fig. 1. The overall workflow can be summarized as follows:

Shared and Encoded Point Clouds: Initially, we apply the KPConv method to share and encode the two input point clouds of the cultural heritage artifacts. This process generates fewer hyperpoint sets and their corresponding features, enhancing the efficiency and accuracy of data processing.

Co-Context Information Extraction: Next, we use the diffusion Transformer module to extract co-context information from the two overlapping artifact point clouds. This module captures the relevant relationships and interactions between points in the overlapping regions, thereby reflecting the detailed features of the artifacts more accurately.

Overlapping Region Matches: Finally, the decoding network takes the co-context information obtained from the diffusion Transformer module as input and outputs the number of matches in the overlapping region. This step provides the precise registration results required for the digital modeling of artifacts.

Problem setting

For two point clouds defined as the source point cloud $P = \{p_i \in \mathbb{R}^3 \mid i = 1, 2, \ldots, N\}$ and the target point cloud $Q = \{q_i \in \mathbb{R}^3 \mid i = 1, 2, \ldots, M\}$, where N and M are the number of points in point clouds P and Q, respectively, the goal of point cloud registration is to align the two point clouds using an unknown 3D rigid transformation $RT = \{R, T\}$, which consists of a rotation $R \in SO(3)$ and a translation $T \in \mathbb{R}^3$. The transformation matrix can be defined as:

$$\min_{R,T} \sum_{(p_i,q_i)\in\vartheta} \|R \cdot p_i + T - q_i\|_2^2 \tag{1}$$

where ϑ represents the ground truth correspondences between the points in P and Q. The notation $\|\bullet\|$ denotes the Euclidean norm.

To address this problem, we propose the PointDT model algorithm. The algorithm takes a pair of point clouds and their correspondences as input, optimizing the 3D rigid transformation to align the source point

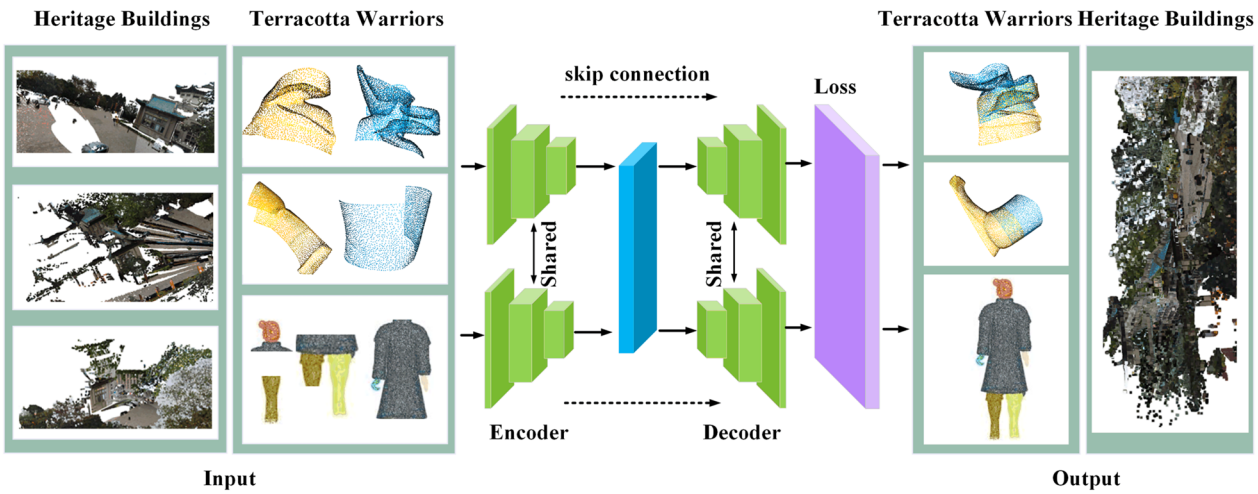

Fig. 1 Network architecture of PointDT

cloud with the target point cloud, thereby achieving precise digital modeling of cultural heritage. This process effectively captures the details and shape characteristics of cultural heritage, providing robust technical support for the preservation, restoration, and study of artifacts.

Encoder-decoder

Encoder: For the denser original point clouds $P \in \mathbb{R}^{N\times 3}$ and $Q \in \mathbb{R}^{M\times 3}$, we utilize the KPConv module as the backbone, which consists of a series of residual modules and strided convolutions, to perform downsampling and reduce the number of keypoints to $P' \in \mathbb{R}^{N'\times 3}$ and $Q' \in \mathbb{R}^{M'\times 3}$, respectively (where $N > N'$ and $M > M'$). Furthermore, we employ a shared encoding method to extract relevant features, resulting in $F'_{P'} \in \mathbb{R}^{N'\times D}$ and $F'_{Q'} \in \mathbb{R}^{M'\times D}$, where D represents the feature dimension.

Decoder: The decoder module follows a standard approach and consists of a 3-layer network structure, including upsampling, linear transformations, and skip connections.

Transformer

To further determine the overlapping area between the two downsampled and feature-extracted point clouds and improve the overall robustness of the network, we introduce an Adaptive Graph Convolution MLP (AGCM) network and integrate it with the PointDT model to form a local–global co-context information-local network architecture. For the digitization modeling of cultural heritage artifacts, this step is crucial. By determining the overlapping regions, we can align the source and target point clouds more accurately, capturing the details and shape features of artifacts more precisely. This provides a solid foundation for subsequent operations.

The AGCM network aims to enhance the contextual understanding of the two point clouds and increase the flexibility of the receptive field. Here, we describe the process using the source point cloud P' as an example.

Firstly, we define the graph structure $G(V, E)$, where $V = \{1, 2, \ldots, N'\}$ represents the set of vertices and $E \subseteq |V| \times |V|$ represents the set of edges, with each vertex corresponding to a point p'_i in P'. The purpose of this step is to establish the connectivity of the point clouds, enabling the network to better understand the relationships between them.

Next, we design a convolutional filter that can be applied at any relative position within the graph. This filter allows us to focus our attention on the most relevant parts of the neighborhood for learning, enabling the convolutional kernel to dynamically adapt to the local structure of the object and better capture the features of artifacts. By computing the feature mapping function h_θ for the filter weights and applying pointwise MLP β_m, we can obtain the feature representation of the overlapping region, further enhancing the understanding and modeling accuracy of artifacts.

$$\Psi_{ijm} = h_\theta\left(\Delta p_{ij}\right) \tag{2}$$

$$^{(k+1)}P' = \max \beta_m\left[\left[\Psi_{ijm}, F_j\right]\right] \tag{3}$$

$$F'_{AGCM} = \beta_m\left[^{(0)}P', {}^{(1)}P', {}^{(2)}P'\right] \tag{4}$$

where, the weight of the M filters is represented as $\Psi_{ijm} = (\psi_1, \psi_2, \ldots, \psi_M)$. The feature mapping function h_θ is implemented using a multi-layer perceptron (MLP). The relative positions of the graph vertices are denoted as $\Delta p_{ij} = [p_i, p_j - p_i]$, where $[\bullet, \bullet]$ represents the concatenation operation.

To compute the output of the AGCM network, we apply the pointwise MLP β_m to each filter weight ψ_m. The inner product of two vectors $[\![\bullet, \bullet]\!]$ is used to define the correspondence F_j, which is calculated as $[F_i, F_j - F_i]$.

Diffusion Transformer: To address the limitations of existing algorithms in capturing local information and improving the effectiveness in cases of low overlap, we propose a diffusion Transformer model algorithm. We will continue using the source point cloud as an example.

We define the position relational embedding $a^{P'}_{j-i}$, which represents the relative position between point p_i and point p_j in the source point cloud. This embedding is used in the cross-attention module to capture the spatial relationships between points. By capturing these relationships, the digital model can more accurately reflect the true geometric shape of the artifact.

Next, we introduce a 4-layer multi-head self-attention module. This module builds upon the local feature information from the AGCM module and allows it to focus on global information by considering different subspaces of representations. The multi-head mechanism enables the model to attend to multiple aspects of the point cloud simultaneously, enhancing its ability to capture both local and global features. In the digitization of artifacts, this mechanism can comprehensively capture the overall shape and subtle features of the artifact, providing a foundation for high-fidelity digitization.

By incorporating the diffusion Transformer module, our algorithm improves the global co-context information and effectively captures the overlapping information between two point clouds, even in cases of low overlap. Thus, even in cases where the artifact is incomplete or damaged, the digitization process can achieve high-precision restoration.

Firstly, we define $F' = (x_1^{P'}, x_2^{P'} \cdots x_{N'}^{P'})$ as the input of the self-attention module in the i-th layer, where N' is the number of points, $Z' = (z_1^{P'}, z_2^{P'} \cdots z_{N'}^{P'})$ is the output matrix, which is the weighted sum of all input matrix transformations, the formula is as follows:

$$z_i^{P'} = \sum_{j=1}^{N'} soft\max\left(\alpha_{i,j}^{Self-}\right)\left(x_j^{P'} W^{V,P'}\right) \tag{5}$$

where, $\alpha_{i,j}^{Self-}$ is the weight coefficient that has not been normalized, and its definition is as follows:

$$\alpha_{i,j}^{Self-} = \frac{1}{\sqrt{d_{head}}}\left(x_i^{P'} W^{Q,P'}\right)\left(x_j^{P'} W^{K,P'}\right)^T \tag{6}$$

Next, we define $F'' = \left(x_1^{P'} + z_1^{P'}, x_2^{P'} + z_2^{P'} \cdots x_{N'}^{P'} + z_{N'}^{P'}\right)$ and $F'_{Q'} = (x_1^{Q'}, x_2^{Q'} \cdots x_{N'}^{Q'})$ as the input $MHAttn(F'_{P'}, F'_{Q'}, F'_{Q'})$ of the cross-attention module in the i-th layer, $Z'' = (z_1^{P',Q'}, z_2^{P',Q'} \cdots z_{N'}^{P',Q'})$ is the output matrix, and its formula is as follows:

$$z_i^{P',Q'} = \sum_{j=1}^{N'} soft\max\left(\alpha_{i,j}^{Cross-}\right) x_j^{Q'} W^{V,Q'} \tag{7}$$

where, $\alpha_{i,j}^{Cross-}$ is the weight coefficient that has not been normalized, and its definition is as follows:

$$\alpha_{i,j}^{Cross-} = \frac{1}{\sqrt{d_{head}}}\left(x_i^{Q'} W^{Q,Q'}\right)\left(x_j^{P'} W^{K,P'} + a_{j-i}^{P'}\right)^T \tag{8}$$

This cross-attention mechanism is particularly important in the digitization of artifacts, as it can combine point cloud data from different perspectives to generate a more complete and consistent artifact model.

Finally, the co-global context information between the two point clouds is output, which is defined as follows:

$$F_i^{DT} = F''_i + MLP(F''_i + z_i^{P',Q'}) \tag{9}$$

It is important to note that the processing of the source point cloud P' and the target point cloud Q' is consistent, except for the exchange of P' and Q' when crossing the attention module. This consistency ensures the unified processing of different data sources in the digitization of artifacts, making the final digital model more accurate and reliable.

Through the Diffusion Transformer module, our algorithm significantly enhances the ability to capture global contextual information in the digitization of cultural heritage artifacts. Especially in low-overlap scenarios, it can effectively reconstruct and preserve detailed structures and features of the artifacts.

Loss function

The proposed PointDT network is trained in an end-to-end manner and supervised with ground truth. The specific loss function is as follows:

Feature Loss: Similar to the D3Feat and Predator approaches, we introduce a circle loss function during the training of 3D point clouds to evaluate feature loss and constrain point feature descriptors. The circle loss function is defined as follows:

$$\mathcal{L}_{FL}^P = \frac{1}{N_P}\sum_{i=1}^{N_P} \log\left[1 + \sum_{j\in\varepsilon_p} e^{c\beta_p^j\left(d_i^j - \Delta p\right)} \bullet \sum_{k\in\varepsilon_n} e^{\lambda\beta_p^k\left(\Delta n - d_i^k\right)}\right] \tag{10}$$

where, d_i^j represents the Euclidean distance between features, $d_i^j = \left\|f_{p_i} - f_{q_j}\right\|_2$. ε_p and ε_n respectively represent the matching and unmatching points of the point set P_{RS} (random sampling points of the source point cloud), i.e., the positive and negative areas. Δp and Δn represent positive and negative regions, respectively. λ stands for predefined parameters. Similarly, the feature loss $\mathcal{L}_{FL}^P$ of the target point cloud is also calculated in the same way, so the total feature loss $\mathcal{L}_{FL} = \frac{1}{2}(\mathcal{L}_{FL}^P + \mathcal{L}_{FL}^Q)$.

Overlap loss: For supervised training, we employ a binary cross-entropy loss function, defined as follows:

$$\begin{aligned}\mathcal{L}_{OL}^P = &\frac{1}{N}\sum_{i=1}^{N} O_{p_i}^{label}\log\left(O_{p_i}\right) \\ &+ \left(1 - O_{p_i}^{label}\right)\log\left(1 - O_{p_i}\right)\end{aligned} \tag{11}$$

$$\begin{aligned}\mathcal{L}_{OL} = &\frac{1}{N}\sum_{i=1}^{N} O_{p_i}^{label}\log\left(O_{p_i}\right) \\ &+ \left(1 - O_{p_i}^{label}\right)\log\left(1 - O_{p_i}\right)\end{aligned} \tag{12}$$

where, $O_{p_i}^{label}$ represents the overlapping mark of ground truth at point p_i, which is defined as follows:

$$O_{p_i}^{label} = \begin{cases} 1, & \left\|T_{P,Q}^{GT}(p_i) - NN\left(T_{P,Q}^{GT}(p_i), Q\right)\right\| < \tau_1 \\ 0, & \text{otherwise}\end{cases} \tag{13}$$

where $T_{P,Q}^{GT}$ represents the ground truth rigid transformation between the overlapping point clouds, and NN represents the nearest neighbor. τ_1 represents the overlap

threshold. Similarly, the overlap loss $\mathcal{L}_{OL}^{Q}$ of the target point cloud is calculated in the same way. Therefore, the total overlap loss is defined as $\mathcal{L}_{OL} = \frac{1}{2}(\mathcal{L}_{OL}^{P} + \mathcal{L}_{OL}^{Q})$.

Matching loss: For each point p_i in the source point cloud, there is a corresponding feature at point q_i in the target point cloud, forming a pair of matching points. To address the sparsity of ground-truth points after downsampling, we employ a matching loss function to handle this challenge.

$$L_{ML}^{P} = \frac{1}{N}\sum_{i=1}^{N} M_{p_i}^{label} \log\left(M_{p_i}\right) + \left(1 - M_{p_i}^{label}\right) \log\left(1 - M_{p_i}\right) \tag{14}$$

where, $M_{p_i}^{label}$ represents the ground truth mark of the registration point p_i, which is defined as follows:

$$M_{p_i}^{label} = \begin{cases} 1, & \left\| T_{P,Q}^{GT}(p_i) - NN\left(T_{P,Q}^{GT}(p_i), Q\right) \right\| < \tau_2 \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

where, τ_2 represents the overlap threshold. Similarly, the overlap loss $\mathcal{L}_{ML}^{Q}$ of the target point cloud is also calculated in the same way, so the total feature loss $\mathcal{L}_{ML} = \frac{1}{2}(\mathcal{L}_{ML}^{P} + \mathcal{L}_{ML}^{Q})$.

To sum up, the overall loss function is $\mathcal{L} = \mathcal{L}_{FL} + \mathcal{L}_{OL} + \mathcal{L}_{ML}$.

Experiments

Implementation details

Our method is implemented in PyTorch [29] and trained on a single Nvidia RTX 4090 GPU with an Intel(R) Core(TM) i7-13700KF CPU @ 3.40GHz and 128GB RAM. For the 3DMatch [15] and 3DLoMatch [23] datasets, we set the learning rate to 0.001, batch size to 4, and train for 40 epochs. For the odometry KITTI [30] dataset, we set the learning rate to 0.01, batch size to 1, and train for 150 epochs.

Indoor datasets: 3DMatch and 3DLoMatch

Dataset: The 3DMatch dataset contains real indoor data from 62 scenarios, with 46 scenes for training, 8 for validation, and 8 for testing. We first pretrain our model using the Predator method and then evaluate it on the 3DMatch and 3DLoMatch datasets. The overlapping areas of the 3DMatch and 3DLoMatch datasets are greater than 30% and between 10% and 30%, respectively.

Metrics: Following the metrics used in D3Feat, Predator, we measure the recall rate of successfully registered pairs using the Registration Recall (RR) metric. A successful registration is considered when the conversion error is less than 0.2m (i.e., RMSE is less than 0.2m). Additionally, we define the relative rotation error (RRE) and relative translation error (RTE) to measure the accuracy of successful registrations. As baselines, we compare our results with state-of-the-art methods such as FCGF [31], D3Feat [21], DGR [32], Predator [23], OMNet [17], CoFiNet [33], REGTR [28], UDPReg [4], MAC [34], RIGA [13].

$$RMSE = \sqrt{\frac{1}{\left|C_{ij}^{GT}\right|}\sum_{(p,q)\in C_{ij}^{GT}} \left\| T_{P,Q}^{GT}(p) - q \right\|_2^2} \tag{16}$$

Table 1 Performance on 3DMatch and 3DLoMatch datasets

Method	3DMatch			3DLoMatch			Param. (M)	Time (s)
	RR (%)	RRE (°)	RTE (m)	RR (%)	RRE(°)	RTE (m)		
FCGF	85.1	1.949	0.066	40.1	3.147	0.100	8.76	0.16
D3Feat	81.6	2.161	0.067	37.2	3.361	0.103	24.3	0.40
DGR	62.7	2.103	0.067	48.7	3.954	0.113	–	–
OMNet	90.5	4.166	0.105	8.4	7.299	0.151	–	–
CoFiNet	89.7	2.147	0.067	67.2	3.271	0.090	–	–
Predator	89.0	2.029	0.064	59.8	3.048	0.093	7.43	0.54
REGTR	92.0	1.567	0.049	64.8	2.827	0.077	–	0.11
UDPReg	91.4	1.642	0.064	64.3	2.951	0.086	12.7	2.67
MAC	93.7	1.890	0.062	59.8	3.500	0.098	–	–
RIGA	89.3	1.798	0.056	65.1	3.016	0.089	–	–
PointDT	**93.8**	**1.536**	**0.045**	**69.6**	**2.75**	**0.061**	9.26	0.19

Bold represents the best result

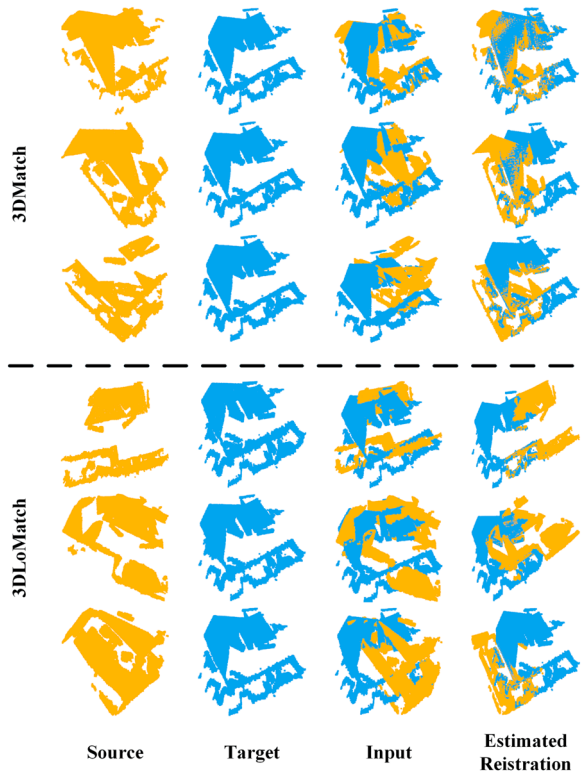

Fig. 2 Registration visualization on 3DMatch, 3DLoMatch

$$RTE = \left\| t - t^{GT} \right\|_2 \tag{17}$$

$$RRE = \arccos\left(\frac{trace\left(R^T R^{GT}\right) - 1}{2}\right) \tag{18}$$

Registration Results: Table 1 presents the comparison of different algorithms, with the best performing ones highlighted in bold. Additionally, Fig. 2 showcases the registration results obtained on the 3DMatch and 3DLoMatch datasets. When tested on the 3DMatch and 3DLoMatch datasets, our algorithm achieves the highest average registration recall across scenarios. Notably, we also achieve the lowest RTE and RRE values. In terms of average registration recall: Compared to the Predator, MAC, and RIGA algorithms, our algorithm achieves an increase of 4.8%, 0.1%, and 4.5% on the 3DMatch dataset, and 9.8%, 9.8%, and 4.5% on the 3DLoMatch dataset.

Outdoor dataset: odometry KITTI

Dataset: The odometry KITTI dataset consists of large-scale LiDAR scanning data from 11 scenarios, with scenarios 0-5 used for training, scenarios 6-7 for validation, and scenarios 8-10 for testing. We refined the provided ground truth data using ICP based on D3Feat, Predator, and GeoTrans algorithms. The evaluation is performed on point clouds spaced up to 10 m apart from each other.

Table 2 Evaluation results on Odometry KITTI dataset

Method	RTE (cm)	RRE (°)	RR (%)
3DFeat-Net	25.9	0.25	96.0
FCGF	9.5	0.30	96.6
D3Feat	7.2	0.30	**99.8**
DGR	32	0.37	98.7
CoFiNet	8.5	0.41	**99.8**
Predator	6.8	0.27	**99.8**
MAC	8.4	0.40	99.5
RoCNet++	7.3	0.23	**99.8**
RIGA	13.5	0.45	99.1
PointDT	**6.3**	**0.22**	**99.8**

Bold represents the best result

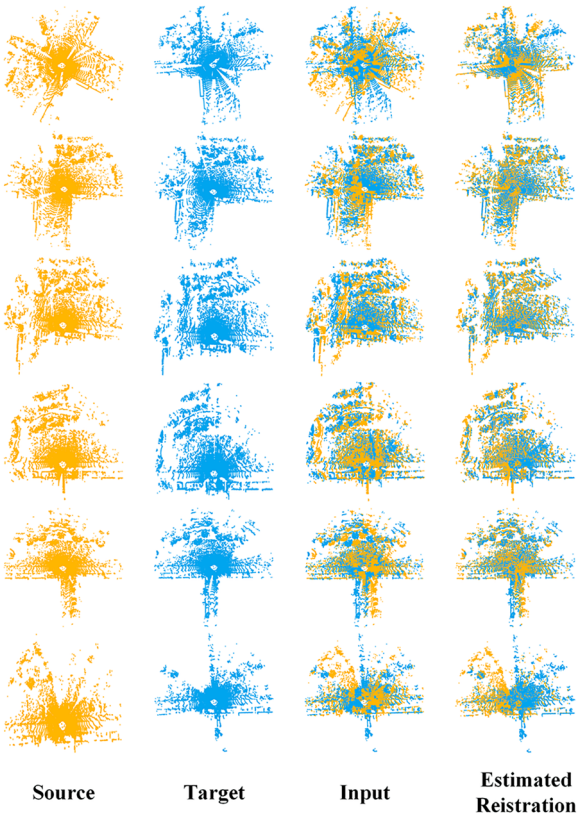

Fig. 3 Registration visualization on Odometry KITTI

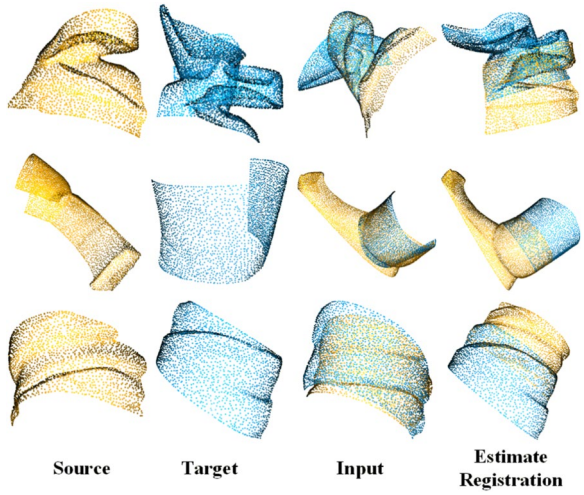

Fig. 4 Registration visualization of 3DMatch → Terracotta Warriors data

Metrics: For the odometry KITTI dataset registration evaluation, we use RR, RRE, and RTE methods following D3Feat and Predator. We select FCGF, DGR, Predator, CoFiNet, MAC, RoCNet++ [35] and RIGA as our baseline algorithms.

Registration Results: Table 2 presents the comparison of different algorithms, with the best performing ones highlighted in bold. Additionally, Fig. 3 illustrates the registration results obtained on the odometry KITTI dataset. When tested on the odometry KITTI dataset, our algorithm achieves the highest mean registration recall. Notably, we also achieve the lowest RTE and RRE values.

Cultural heritage dataset

To evaluate the registration performance of the proposed algorithm in cultural heritage datasets, we first validate it using the dataset of the Terracotta Warriors and Horses in the Mausoleum of the First Qin Emperor collected by Northwestern University, as shown in the Fig. 4.

From the Fig. 4, it can be seen that there are good registration results in the head, feet, and arms of the Terracotta Warriors.

To further verify the registration fusion performance of the proposed algorithm in large-scale cultural heritage datasets, we utilize cultural heritage buildings from the Wuhan University Terrestrial Laser Scanning(WHU-TLS) [36–38] dataset by Wuhan University for validation, as shown in the Fig. 5.

From the Fig. 5, it can be observed that the proposed algorithm still achieves good registration fusion results in large-scale cultural heritage buildings experimentation.

To verify the registration fusion efficiency of the proposed algorithm in large-scale scenes, we used cultural heritage buildings from the WHU-TLS dataset as the basis and tested different fusion times, as shown in Table 3. Here, since the PointDT algorithm adopts

Table 3 Comparison of registration time for WHU-TLS heritage buildings

Number	ICP	FPFH+RANSAC	PointDT
2	5.39	>>100	0.93
7	1292.01	>>100	13.69

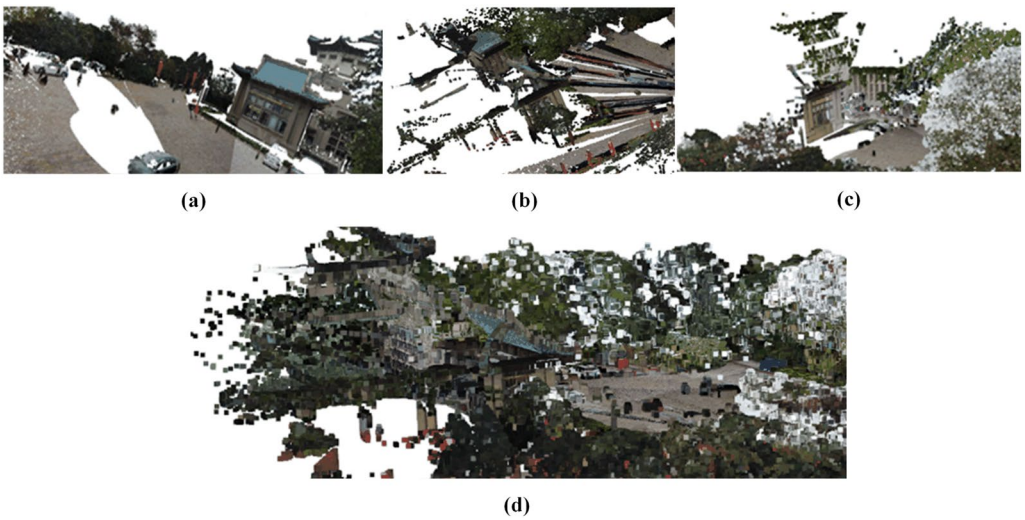

Fig. 5 Registration visualization of 3DMatch → WHU-TLS Heritage Building data. **a** Shows a real outdoor scene. **b** and **c** Show point cloud data collected from different perspectives in real scenes. **d** Shows the registration results of (**b**) and (**c**)

Table 4 Ablation of different modules on the 3DMatch, 3DloMatch dataset

Method	3DMatch			3DLoMatch		
	RR (%)	RRE (°)	RTE (m)	RR (%)	RRE (°)	RTE (m)
Base	50.6	3.668	0.93	12.5	5.632	0.117
AGCM	92.3	1.706	0.05	67.8	3.105	0.084
Diffusion Transformer	91.9	1.847	0.060	69.1	3.130	0.083
Position Encoding	92.8	1.665	0.062	68.3	2.77	0.077
PointDT	**93.8**	**1.536**	**0.045**	**69.6**	**2.75**	**0.061**

Bold represents the best result

transfer learning, only the time used for testing verification after model training completion is calculated.

From Table 3, it can be seen that compared to traditional algorithms, the proposed algorithm demonstrates outstanding performance. Especially with an increase in fusion quantity, the proposed algorithm exhibits a significant advantage in terms of time required.

Ablation study

Importance of Individual Modules: To evaluate the effectiveness of individual module selections in our model, we conducted ablation experiments on the 3DMatch and 3DLoMatch datasets.

It can be seen from Table 4 that the PointDT algorithm has better performance than the other four individual module effects.

To further ascertain the robustness of the proposed algorithm, we introduced a set of Gaussian noise with standard deviations ranging from 0.02 to 0.1 in our experiments for validation purposes. The outcomes are depicted in Fig. 6.

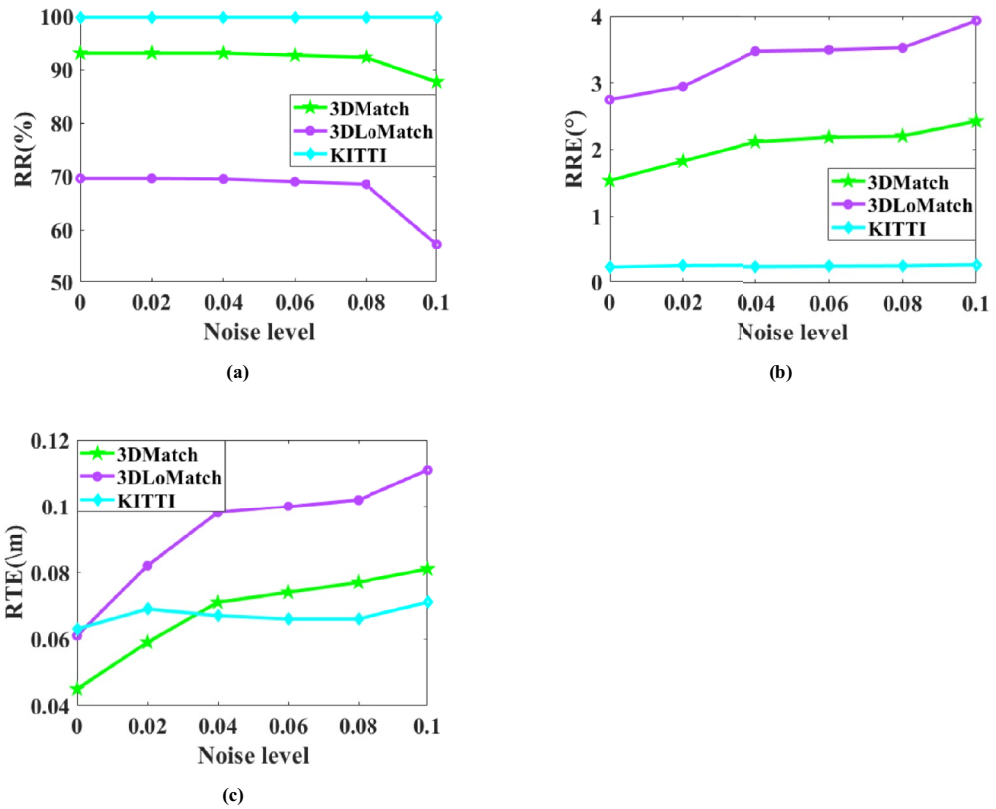

Fig. 6 Robustness tests of 3DMatch, 3DLoMatch and KITTI datasets on RR, RRE and RTE respectively

From Fig. 6, it can be clearly seen that the proposed algorithm demonstrates excellent robustness in both indoor and outdoor testing scenarios.

Conclusion

In this study, we explored the application of point cloud registration technology in digital cultural heritage modeling and introduced the Diffusion Transformer model to enhance registration performance. This method effectively handles large-scale point cloud data, improving the accuracy and efficiency of registration. Firstly, our research demonstrates the outstanding performance of the Diffusion Transformer model in point cloud registration tasks. Through experiments in different types of cultural heritage scenarios, we found that the Diffusion Transformer model can more accurately capture the geometry and semantic information of target point clouds, thereby achieving faster and more precise registration results. This is crucial for the accuracy and reliability of digital cultural heritage modeling. Secondly, our research also revealed the advantages of the Diffusion Transformer model in handling data noise.

Furthermore, our study provides some insights into future research directions in the field of digital cultural heritage modeling. We can explore how to integrate point cloud registration technology with other digital modeling methods, such as image processing and virtual reality technology, to create more realistic and comprehensive digital cultural heritage models.

Acknowledgements
We thank the National and Local Joint Engineering Research Center for Cultural Heritage Digitization for providing the Terracotta Warriors data.

Author contributions
Li An: Conceptualization, Methodology, Resources, Writing, original draft preparation, Writing, review and editing. Pengbo Zhou: Writing, review and editing, Writing, review and editing, Visualization. Mingquan Zhou: Conceptualization, Methodology, Writing, review and editing. Yong Wang: Methodology, Writing, review and editing, Visualization. Guohua Geng: Conceptualization, Methodology, original draft preparation.

Funding
This research was funded by the National Natural Science Foundation of China: 62271393. Key Laboratory Project of the Ministry of Culture and Tourism: 1222000812, crrt2021K01. Xi'an Science and Technology Plan Project: 2024JH-CXSF-0014. National key research and development plan: 2020YFC1523301, 2020YFC1523303.

Data availability
The data will be available upon reasonable request.

Declarations

Ethics approval and consent to participate
Written informed consent has been obtained from the School of Information Science and Technology of Northwest University and all authors for this article, and consent has been obtained for the data used.

Competing interests
The authors declare that they have no conflict of interest.

Received: 14 March 2024 Accepted: 2 June 2024
Published online: 14 June 2024

References

1. Markiewicz J, Kot P, Markiewicz Ł, Muradov M. The evaluation of hand-crafted and learned-based features in Terrestrial Laser Scanning-Structure-from-Motion (TLS-SfM) indoor point cloud registration: the case study of cultural heritage objects and public interiors. Heritage Sci. 2023;11(1):254.
2. Cotella VA. From 3D point clouds to HBIM: application of artificial intelligence in cultural heritage. Autom Constr. 2023;152:104936.
3. Tabib RA, Hegde D, Anvekar T, Mudenagudi U. DeFi: detection and filling of holes in point clouds towards restoration of digitized cultural heritage models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023. p. 1603–1612.
4. Mei G, Tang H, Huang X, Wang W, Liu J, Zhang J, et al. Unsupervised deep probabilistic approach for partial point cloud registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 13611–13620.
5. Lyu M, Yang J, Qi Z, Xu R, Liu J. Rigid pairwise 3D point cloud registration: a survey. Pattern Recognition. 2024;110408.
6. Ao S, Hu Q, Wang H, Xu K, Guo Y. Buffer: balancing accuracy, efficiency, and generalizability in point cloud registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 1255–1264.
7. Galanakis D, Maravelakis E, Pocobelli DP, Vidakis N, Petousis M, Konstantaras A, et al. SVD-based point cloud 3D stone by stone segmentation for cultural heritage structural analysis-the case of the Apollo Temple at Delphi. J Cult Herit. 2023;61:177–87.
8. Foryś P, Sitnik R, Markiewicz J, Bunsch E. Fast adaptive multimodal feature registration (FAMFR): an effective high-resolution point clouds registration workflow for cultural heritage interiors. Herit Sci. 2023;11(1):190.
9. Besl PJ, McKay ND. Method for registration of 3-D shapes. In: Sensor fusion IV: control paradigms and data structures. vol. 1611. Spie; 1992. p. 586–606.
10. Deng H, Birdal T, Ilic S. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In: Proceedings of the European conference on computer vision (ECCV); 2018. p. 602–618.
11. Cao F, Wang L, Ye H. SharpGConv: a novel graph method with plug-and-play sharpening convolution for point cloud registration. IEEE Transactions on Circuits and Systems for Video Technology. 2024;1–1. https://doi.org/10.1109/TCSVT.2024.3369468.
12. Liu S, Wang T, Zhang Y, Zhou R, Li L, Dai C, et al. Deep semantic graph matching for large-scale outdoor point cloud registration. IEEE Trans Geosci Remote Sens. 2024;62:1–4. https://doi.org/10.1109/TGRS.2024.3355707.
13. Yu H, Hou J, Qin Z, Saleh M, Shugurov I, Wang K, et al. RIGA: rotation-invariant and globally-aware descriptors for point cloud registration. IEEE Trans Pattern Anal Mach Intell. 2024. https://doi.org/10.1109/TPAMI.2023.3349199.
14. Wang Y, Zhou P, Geng G, An L, Liu Y. CCAG: end-to-end point cloud registration. IEEE Robot Autom Lett. 2023;9(1):435–42.
15. Zeng A, Song S, Nießner M, Fisher M, Xiao J, Funkhouser T. 3dmatch: learning local geometric descriptors from rgb-d reconstructions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1802–1811.
16. Deng H, Birdal T, Ilic S. Ppfnet: global context aware local features for robust 3d point matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 195–205.

17. Xu H, Liu S, Wang G, Liu G, Zeng B. Omnet: learning overlapping mask for partial-to-partial point cloud registration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 3132–3141.
18. Wang H, Liu Y, Hu Q, Wang B, Chen J, Dong Z, et al. RoReg: pairwise point cloud registration with oriented descriptors and local rotations. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023;1–18.
19. Yan Y, An J, Zhao J, Shen F. Hybrid optimization with unconstrained variables on partial point cloud registration. Pattern Recogn. 2023;136:109267.
20. Lu W, Wan G, Zhou Y, Fu X, Yuan P, Song S. Deepvcp: an end-to-end deep neural network for point cloud registration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019. p. 12–21.
21. Bai X, Luo Z, Zhou L, Fu H, Quan L, Tai CL. D3feat: joint learning of dense detection and description of 3d local features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. p. 6359–6367.
22. Thomas H, Qi CR, Deschaud JE, Marcotegui B, Goulette F, Guibas LJ. Kpconv: flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 6411–6420.
23. Huang S, Gojcic Z, Usvyatsov M, Wieser A, Schindler K. Predator: registration of 3d point clouds with low overlap. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition; 2021. p. 4267–4276.
24. Zhang Z, Sun J, Dai Y, Zhou D, Song X, He M. End-to-end learning the partial permutation matrix for robust 3D point cloud registration. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36; 2022. p. 3399–3407.
25. Wang Y, Solomon JM. Deep closest point: learning representations for point cloud registration. In: Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 3523–3532.
26. Fu K, Liu S, Luo X, Wang M. Robust point cloud registration framework based on deep graph matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 8893–8902.
27. Liu J, Wang G, Liu Z, Jiang C, Pollefeys M, Wang H. RegFormer: An Efficient Projection-Aware Transformer Network for Large-Scale Point Cloud Registration. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Los Alamitos, CA, USA: IEEE Computer Society; 2023. p. 8417–8426. Available from: https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00776.
28. Yew ZJ, Lee GH. Regtr: end-to-end point cloud correspondences with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 6677–6686.
29. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst. 2019;32:1–12.
30. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The Kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE; 2012. p. 3354–3361.
31. Choy C, Park J, Koltun V. Fully convolutional geometric features. In: Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 8958–8966.
32. Choy C, Dong W, Koltun V. Deep global registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. p. 2514–2523.
33. Yu H, Li F, Saleh M, Busam B, Ilic S. Cofinet: reliable coarse-to-fine correspondences for robust pointcloud registration. Adv Neural Inf Process Syst. 2021;34:23872–84.
34. Zhang X, Yang J, Zhang S, Zhang Y. 3D registration with maximal cliques. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 17745–17754.
35. Slimani K, Achard C, Tamadazte B. RoCNet++: triangle-based descriptor for accurate and robust point cloud registration. Pattern Recogn. 2024;147:110108.
36. Dong Z, Liang F, Yang B, Xu Y, Zang Y, Li J, et al. Registration of large-scale terrestrial laser scanner point clouds: a review and benchmark. ISPRS J Photogramm Remote Sens. 2020;163:327–42.
37. Dong Z, Yang B, Liang F, Huang R, Scherer S. Hierarchical registration of unordered TLS point clouds based on binary shape context descriptor. ISPRS J Photogramm Remote Sens. 2018;144:61–79.
38. Dong Z, Yang B, Liu Y, Liang F, Li B, Zang Y. A novel binary shape context for 3D local surface description. ISPRS J Photogramm Remote Sens. 2017;130:431–52.

Publisher's Note