Neural Networks 192 (2025) 107930

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Full length Article

CMFF: Cross-modal feature fusion network for robust point cloud completion

Jian Gao [a], Yuhe Zhang [a,*], Pengbo Zhou [b], Xinda Liu [a], Longquan Yan [a], Mingquan Zhou [b], Guohua Geng [a,*]

[a] College of Information Science and Technology, Northwest University, Xi'an, 710127, China
[b] School of Arts and Communication, Beijing Normal University, Beijing, China

ARTICLE INFO

Keywords:
Point cloud completion
Feature information fusion
Differential point transformer
Differential cross transformer

ABSTRACT

In the field of 3D vision, point cloud data often suffers from partial missing regions due to occlusion, reflection, or viewpoint limitations, which severely affects downstream tasks. Existing cross modal point cloud completion methods directly employ cross attention mechanisms for feature fusion, without considering the feature distributions and noise between different modalities, leading to suboptimal completion results. In this paper, we propose a novel cross modal point cloud completion framework, CMFF, which takes partial point clouds and single view images as inputs. Specifically, firstly, it uses a point cloud encoder and an image encoder to extract features from the input. In the point cloud encoder, we propose a differential point transformer module for extracting local geometric details and global structural features of the point cloud, which enhances the representation and robustness of complex geometries. Second, we propose a differential cross transformer module for feature fusion. The redundant and conflicting cross modal features are filtered by differential operations to enhance the correlation of cross modal features and improve the completion accuracy. Third, coarse point cloud is generated using a point cloud patch generator. Finally, we propose the fine point cloud module, which optimizes multi modal features using simple attention mechanism and generates an offset vector to optimize the point cloud. Extensive experiments on the view-guided point cloud completion benchmark ShapeNet-ViPC and the Terracotta Warriors dataset show that CMFF outperforms 15 current methods to state-of-the-art on several point cloud completion metrics, exhibiting excellent performance and generalization ability.

1. Introduction

As an important form of spatial information expression, point cloud data are widely used in the fields of 3D reconstruction (Gao et al., 2023; He et al., 2025), automatic driving (Liu et al., 2022; Shao et al., 2024) and cultural heritage protection (Gao et al., 2024; Yang et al., 2024). However, due to factors such as occlusion, reflex, or view angle limitation during the acquisition process, point cloud data are usually missing and incomplete, which largely affects the subsequent analysis and processing results. Therefore, it is necessary to complete the scanned point cloud data before applying it to downstream tasks.

Traditional point cloud completion methods (Chang et al., 2021; Gao et al., 2024; Tchapmi et al., 2019; Xia et al., 2021; Yu et al., 2021; Yuan et al., 2018; Zhou et al., 2022) mainly rely on a single partial point cloud data for completion, and usually use neural networks (e.g., PointNet Charles et al., 2017, PointNet++ Qi et al., 2017, etc.) to learn the local and global structural information, so they usually face the problem of insufficient feature representation capability. Especially when dealing with complex geometries or large-scale data, the performance of existing methods often fails to meet the demands of high-precision.

In recent years, several cross modal completion methods (Aiello et al., 2022; Xu et al., 2025; Zhang et al., 2021a; Zhu et al., 2023) have been proposed, which use semantic information provided by single view images to assist in point cloud completion. These cross modal methods can compensate for the shortcomings of uni-modal methods and improve the accuracy of point cloud completion. However, existing cross modal methods also have some issues.

The key step in cross modal point cloud completion is feature fusion, and the main challenge in feature fusion is to overcome inter-modal differences and noise. Data in different modalities have different feature distributions and representations, direct fusion may result in key features being overwhelmed or irrelevant information in certain modalities being overemphasized. Existing methods rely on self attention and cross attention to fuse features without sufficiently considering the feature

* Corresponding authors.
E-mail addresses: gaojian1@stumail.nwu.edu.cn (J. Gao), zhangyuhe0601@nwu.edu.cn (Y. Zhang), ghgeng@nwu.edu.cn (G. Geng).

https://doi.org/10.1016/j.neunet.2025.107930
Received 4 March 2025; Received in revised form 7 May 2025; Accepted 29 July 2025
Available online 5 August 2025

distributions and noise of different modalities, resulting in lower performance information fusion with unsatisfactory completion. Second, most of the existing methods use DGCNN (Phan et al., 2018) and PointNet++ (Qi et al., 2017) for feature extraction, they usually rely on k-nearest neighbor maps to capture local features, with relatively low sensitivity to local geometric details, and are unable to perceive small changes in local point clouds. This can lead to the omission of key geometric features, making it difficult to accurately model local structures, so they are usually invalid to complete complex models.

In order to solve the above problems, we propose a novel cross modal point cloud completion framework called CMFF, which takes partial point clouds and single view images as inputs. Specifically, firstly, it uses a point cloud encoder and an image encoder to extract features from the input. In the point cloud encoder, we propose the Differential Point Transformer (DPT) module, the DPT introduces relative positional encoding for point clouds and employs a differential mechanism to learn the differences between two sets of attention weights. This enables effective capture of variations between different regions or contexts within the point cloud, thereby enhancing the model's robustness to complex point cloud structures. Second, we propose the Differential Cross Transformer module (DCT) for feature fusion. The DCT eliminates noise in cross modal features by separately mapping queries, keys, and values to different spaces, and calculating the difference between two softmax attention matrices. This approach allows effective weighted fusion of features from different modalities while adaptively enhancing focus on relevant contextual information, leading to more precise fusion results. Third, coarse point cloud is generated using a point cloud patch generator. Finally, we propose a Fine Point Cloud module. This component employs a simple attention mechanism (SAM) to optimize multimodal features from both coarse point clouds and images. The enhanced multimodal features are then used to generate a set of offset vectors aimed at refining the coarse point cloud. We conducted experiments on the view-guided point cloud completion dataset ShapeNet-ViPC and the Terracotta Warriors dataset. The results demonstrate that CMFF outperforms 15 mainstream methods to state-of-the-art on several point cloud completion metrics.

Our main contributions are as follows:

- we propose a novel cross modal point cloud completion framework called CMFF, which achieves state-of-the-art performance on cross modal point cloud completion benchmark datasets.
- We propose a differential point transformer (DPT) module, which can effectively extract both local details and global structural features of the point cloud, enhancing the ability to express and the robustness of complex geometric shapes.
- The Differential Cross Transformer module (DCT) is proposed to filter out redundant and conflicting features between modalities through differential operations, strengthening the feature correlation across different modalities and improving the accuracy and robustness of point cloud completion.
- A Point Cloud Refinement module is proposed to further optimize the details of point cloud completion by fusing features across different modalities.

2. Related work

2.1. Traditional point cloud completion

Geometry-based methods: These methods rely on the shape information of the known portion of the point cloud and infer the missing parts using assumed geometric rules or models. It is usually assumed that the point cloud structure has some regularity or symmetry. Some methods (Berger et al., 2014; Davis et al., 2002; Nguyen et al., 2016) use local interpolation techniques to fill small holes or missing areas on the surface, which is especially suitable for recovering smooth surfaces. Other methods (Mitra et al., 2006, 2013; Sung et al., 2015) infer the complete shape of an object by utilizing symmetry axes or the symmetry of the object, which is particularly effective when the object has clear symmetry. In addition, the advantage of geometric methods lies in their high computational efficiency, as they can quickly complete the point cloud, especially when the missing parts are small or regular. However, these methods also have clear limitations: when large portions are missing or the object shape is complex, the inferred results may be inaccurate, and they are difficult to adapt to highly irregular or complex geometric point cloud data.

Alignment-based methods: These methods complete the point cloud by matching the missing parts with models from a pre-constructed template database. Compared to geometry-based methods, alignment-based methods can handle more complex point cloud missing situations, as they use the shape information from complete or partial models to recover the missing data. For example, some methods (Li et al., 2015; Pauly et al., 2005; Shao et al., 2012) retrieve entire 3D models and match them with the target point cloud to infer the structure of the missing area; other methods (Kim et al., 2013; Mitra et al., 2006; Shen et al., 2012) perform local matching, repairing the missing part by comparing its local shape with similar parts in the database. Alignment-based methods usually generate more accurate completion results, particularly for point cloud data with substantial missing areas or complex shapes. However, alignment-based methods also have some limitations. First, constructing a large template database and performing efficient matching and inference optimization require significant computational resources, which presents a major challenge for large-scale applications. Furthermore, the quality of the template database directly affects the accuracy of the completion results, so how to construct an efficient and precise template library is one of the key bottlenecks for this approach.

2.2. Uni-modal point cloud completion

Point-based methods: With the introduction of PointNet (Charles et al., 2017) and PointNet++ (Qi et al., 2017), researchers shifted from voxelization-based approaches to directly handling point clouds. PCN (Yuan et al., 2018) is a pioneer in this category, generating dense and complete point clouds in a coarse-to-fine manner. It uses two PointNet layers to extract features and combines the advantages of fully connected layers and folding-based decoders, surpassing the performance of single decoders. FinerPCN (Chang et al., 2021) introduces a per-point convolution to optimize local structure and achieve high-fidelity point cloud completion. PUNet (Yu et al., 2018) learns multi-scale features of point clouds through subpixel convolution layers, which are mainly used for dense reconstruction but not for completing missing parts. ASFM-Net (Xia et al., 2021) uses a Siamese autoencoder neural network to map partial and complete input point clouds into a shared latent space and utilizes iterative refinement units to generate complete shapes with fine-grained details. CS-Net (Ma et al., 2022) combines segmentation and completion models to handle noisy or anomalous point clouds. TopNet (Tchapmi et al., 2019) introduces a hierarchical rooted tree structure decoder to generate points at different detail levels. AML (Stutz & Geiger, 2018) introduces a weakly-supervised method using the maximum likelihood approach. Pcl2Pcl (Chen et al., 2019) proposes a GAN framework to bridge the semantic gap between incomplete and complete shapes. AtlasNet (Groueix et al., 2018) represents 3D shapes as a collection of parameterized surface elements, where these learnable parameters transform a set of 2D squares into surfaces, covering the surface in a manner similar to forming a paper model by placing paper strips on a shape. MSN (Liu et al., 2020) takes multi-scale neighborhood information as input and captures geometric details at different scales through multi-level feature learning and fusion, achieving precise point cloud completion. LAKe-Net (Tang et al., 2022) performs point cloud completion by aligning keypoint localization, generating surface skeletons, and refining shapes.

Graph-based methods: Graph neural networks (GNNs) are well-suited for handling point clouds since point cloud data can be viewed

as a graph with nodes representing points, where information is propagated across nodes, ignoring the order of input nodes, and learning the dependency relationship between nodes as edges. Inspired by DGCNN (Phan et al., 2018), LDGCNN (Zhang et al., 2019) continues its feature extraction method and further improves performance by connecting multi-level features from different layers, while optimizing the model scale. PMP-Net (Wen et al., 2021) formulates the shape completion problem as a point cloud deformation problem and generates a complete point cloud by iteratively moving input points to appropriate positions with minimal displacement. RL-GAN-Net (Sarmad et al., 2019) introduces the use of a reinforcement learning (RL) agent to drive a generative adversarial network (GAN) to predict the complete point cloud. As GANs are unstable and difficult to train, RL-GAN-Net trains the GAN on latent space representations to address this issue. ECG (Pan, 2020) first generates a rough skeleton to facilitate the capture of useful edge features for noisy measurements, and then refines point cloud details through a refinement stage. Each stage is a generation subnetwork conditioned on the incomplete point cloud input, using deep hierarchical encoders with graph convolution functions to propagate multi-scale edge features and refine local geometric structures. PC-RGNN (Zhang et al., 2021b) designs a graph neural network module that captures relationships between points comprehensively through local-global attention mechanisms and multi-scale graph-based context aggregation, significantly enhancing encoded features. Shi et al. (2021) takes the input data and intermediate generation as control points and support points, respectively, and modeling the point cloud completion task with graph convolution network (GCN)-guided optimization.

Attention-based methods: Point cloud processing methods based on attention mechanisms, such as Point Transformers (Zhao et al., 2021) and Point Cloud Transformers (Guo et al., 2021), have successfully introduced attention mechanisms into computer vision, achieving significant results. Initially, these methods were primarily applied to point cloud classification and later expanded to point cloud completion tasks. PUI-Net (Zhao et al., 2020) uses a channel attention mechanism to extract discriminative features from point clouds and generates high-resolution dense feature maps to fill in missing point cloud parts using non-local feature expansion units. N-DPC (Li et al., 2020) introduces a self-attention mechanism to extract global features of point clouds, strengthening the dependencies between points, and adopts a coarse-to-fine strategy to generate refined point cloud data. PointGrow (Sun et al., 2020) designs a dedicated self-attention module recursive neural network, which samples points by analyzing the conditional distribution of the previous network output, enhancing correlations between points. PointTr (Yu et al., 2021) represents the point cloud as a set of unordered points and uses a Transformer encoder-decoder structure to rebuild the missing point cloud, proposing a geometry-sensitive point cloud completion Transformer. ProxyFormer (Li et al., 2023) discards the Transformer (Vaswani et al., 2017) decoder in traditional Transformer-based point cloud completion methods (such as PointTr Yu et al., 2021) and proposes a different design strategy. SnowflakeNet (Xiang et al., 2021) simulates the snowflake-like growth process of points in 3D space to generate a complete point cloud, where each sub-point is progressively generated by splitting its parent point. PointAttN (Wang et al., 2024) directly establishes structural relationships between points using attention mechanisms, efficiently completing point clouds through geometric detail perception.

2.3. Cross modal point cloud completion

Due to the sparsity and incompleteness of point cloud data, unimodality is often insufficient to provide enough geometric information for accurately completing the missing parts, especially when dealing with complex object shapes and large-scale missing regions. As a result, multi-modal data (such as images, depth maps, etc.) have been introduced. By fusing surface textures, color information from images with the spatial structure of point clouds, it is possible to more effectively complete missing regions, improve completion accuracy and detail restoration, and enhance the adaptability to complex object shapes and variable scenes. ViPC (Zhang et al., 2021a) is a pioneering work on image-guided point cloud completion based on single-view images. It constructs a benchmark dataset, ShapeNet-ViPC, for cross modal point cloud completion. ViPC first maps the image to a rough representation of the point cloud, then aligns the reconstructed point cloud with the incomplete point cloud, and finally refines it using the connected image features and point cloud features. CSDN (Zhu et al., 2023) proposes IPadaIN, which embeds global features from images and partial point clouds to guide the geometric generation and completion of the missing point cloud regions, refining the rough output by adjusting the generated point locations. XMFnet (Aiello et al., 2022) uses stacked cross attention and self-attention layers to fuse image and point cloud features, and employs multi-branch generation to complete the point cloud using the fused features. CDPNet (Du et al., 2024) introduces a two-stage strategy that utilizes global information from images to predict a rough shape, and then completes the point cloud using a multi-branch generation method. EGIInet (Xu et al., 2025) unifies the processes of image and encoding to facilitate modality alignment and proposes an explicit guided information interaction strategy based on modality alignment for point cloud completion. FTIGNet (Song et al., 2023) overcomes the lack of prior information caused by small-scale datasets by using a pre-trained visual language model, while effectively fusing visual and textual information to predict both the semantic and geometric features of incomplete shapes. DMF-Net (Mao et al., 2024) introduces a novel dual-channel modality fusion network for image-guided point cloud completion.

3. Methodology

Given single-view image $I \in \mathbb{R}^{H\times W\times C}$ and partial point cloud $P \in \mathbb{R}^{N\times 3}$, our goal is to predict the missing regions and generate a more complete point cloud $P_1 \in \mathbb{R}^{N_1\times 3}$. Here, H and W represent the height and width of the image, respectively, and C represents the number of channels. N and N_1 represent the number of input and output points, respectively. In this paper, we propose a cross modal feature fusion network (CMFF) for robust point cloud completion. Fig. 1 shows an architecture of the proposed method, the following sections describe this architecture in detail.

3.1. Feature extraction module

The primary distinction between images and point clouds is in their data organization, and effective feature extraction relies on unifying the representation of both modalities. Inspired by EGIINet (Xu et al., 2025), we represent both images and point clouds as a series of tokens. This approach simplifies the structural differences between images and point clouds and improves the ability of the model to reason about cross modal data.

For the image $I \in \mathbb{R}^{H\times W\times C}$, we use 16×16 convolution kernel and 16×16 steps to convert it into a series of tokens. Each token represents a local region in the image. In this way, the model learns the mapping from image pixels to token representations, providing a unified input format for subsequent cross modal alignment. We then use a ViT (Alexey, 2020) block based on self-attention as the backbone network to extract features from the image token sequence, ultimately obtaining the image feature $F_I \in \mathbb{R}^{N_i\times L}$, where N_i represents the number of image tokens and L represents the length of the token features.

For the point cloud $P \in \mathbb{R}^{N\times 3}$, we use the set abstraction (Qi et al., 2017) to serialize it as shown in Fig. 2. Specifically, we first apply Farthest point sampling (FPS) to extract the key tokens P^t of the point cloud P, while using Ball-query clustering to aggregate the local neighborhood of each key token. Then, we use an MLP to extract the initial point features $F'_{P^t} \in \mathbb{R}^{N_p\times L}$ of each token. Here, N_p represents the number of point cloud tokens, and L represents the length of the token features.

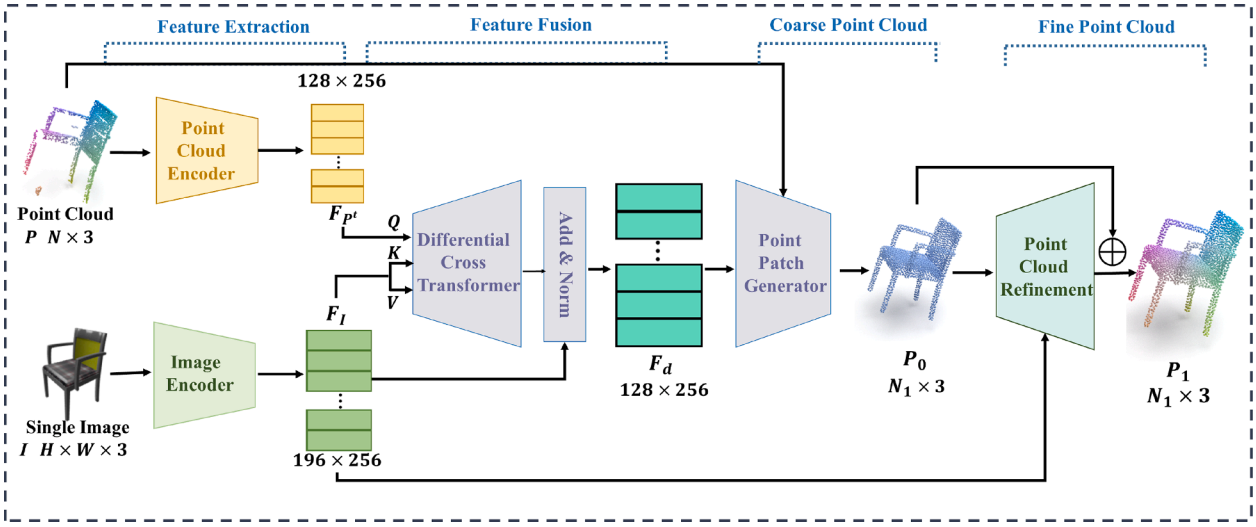

Fig. 1. CMFF Architecture. CMFF consists of four main components; (i) feature extraction module, (ii) feature fusion module, (iii) coarse point cloud module and (iv) fine point cloud module. It takes partial point cloud and single view image as input. First, it extracts features from the input using point cloud encoder and image encoder. Then, It then fuses the two features and removes inter-modal noise using a differential cross transformer. Next, coarse point cloud is generated using a point cloud patch generator. Finally, the coarse point cloud and image features are used to generate a set of offsets to obtain a fine point cloud.

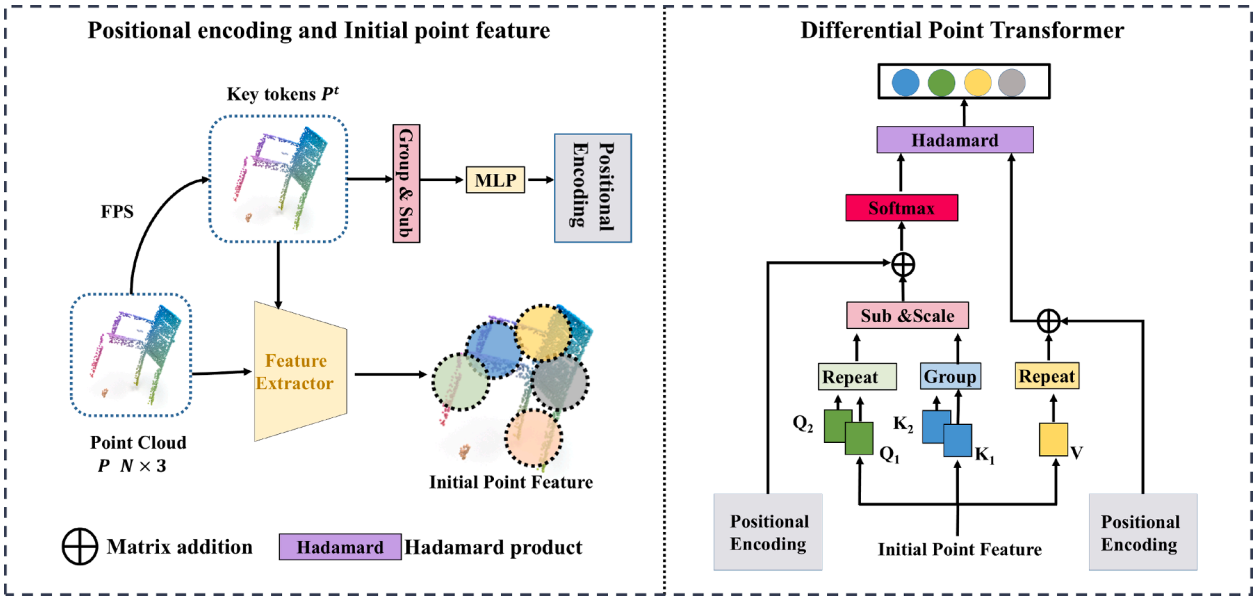

Fig. 2. The architecture of point cloud encoder. We first perform position coding and initial feature extraction for the point cloud, and subsequently propose differential point transformer (DPT) to further enhance the performance of the point cloud features.

However, set abstraction is easy to over-smooth local details and difficult to model long distance depends, resulting in limited point cloud completion accuracy.

To overcome this deficiency, we propose a differential point transformer (DPT) model. Specifically, given the key tokens $P^t \in \mathbb{R}^{N_p \times 3}$ of a point cloud, first we perform positional encoding:

$$\phi_{ij} = \theta\left(P_i^t - P_j^t\right), P_j^t \in \gamma(P_i^t) \tag{1}$$

where P_i^t is the ith point of P^t, the γ denotes the list of the k nearest neighbors, P_j^t is the jth point in the list of P_i^t neighbors. The θ is an MLP with two linear layers and a ReLU nonlinearity.

Given the initial features $F'_{P^t} \in \mathbb{R}^{N_p \times L}$ of key tokens, we first project them into queries, keys, and values $Q_1, Q_2, K_1, K_2, V \in \mathbb{R}^{N_p \times L}$:

$$[Q_1; Q_2] = F'_{p^t} W^Q, [K_1; K_2] = F'_{p^t} W^K, V = F'_{p^t} W^V \tag{2}$$

where $W^Q \in \mathbb{R}^{L \times 2L}, W^K \in \mathbb{R}^{L \times 2L}, W^V \in \mathbb{R}^{L \times L}$ are the parameters of the convolutional layer in the neural network. The dimensional shape of $F'_{p^t} W^Q$ is $N_p \times 2L$, the slice of $F'_{p^t} W^Q[:, :L]$ is Q_1, the slice of $F'_{p^t} W^Q[:, L:2L]$ is Q_2. K_1 and K_2 are obtained using the same method. Then, the Differential Point Transformer computes the output as follows:

$$A_1^{ij} = softmax\left(\frac{Q_1^i - K_1^j + \phi_{ij}}{\sqrt{L}}\right), K_1^j \in \gamma(K_1^i) \tag{3}$$

$$A_2^{ij} = softmax\left(\frac{Q_2^i - K_2^j + \phi_{ij}}{\sqrt{L}}\right), K_2^j \in \gamma(K_2^i) \tag{4}$$

where γ denotes the list of the k nearest neighbors. K^j is the jth vector in the list of K^i neighbors. ϕ denotes positional encoding. Ultimately, the feature $F_{p^t} \in \mathbb{R}^{N_p \times L}$ of key tokens is obtained from the following

equation:

$$F_{p^t}^i = \sum_{j=0}^{k}(A_2^{ij} - \lambda A_1^{ij}) \odot \left(\varphi(V) + \phi_{ij}\right) \tag{5}$$

where φ denotes the vector of replicas k times, λ is the learnable scalar, and $\odot$ is the Hadamard product. The introduction of positional encoding during attention computation can further strengthen local geometric relationships. Difference learning enhances the ability to capture long-distance dependencies by calculating variations between different queries and keys.

3.2. Feature fusion module

Once we have obtained the tokens features of the image and the point cloud, we need to effectively fuse the features of both so that the tokens of the image can effectively complement the tokens of the point cloud. Cross Transformer is a common approach to cross modal fusion, but it often tends to focus too much on irrelevant contexts and fails to adequately account for possible feature noise between different modalities.

For this reason, we propose a feature fusion method based on Differential Cross Transformer (DCT), which consists of three stacked multi-head differential cross attention (MDCA). As shown in Fig. 3, the point cloud features $F_{p^t} \in \mathbb{R}^{N_p \times L}$ are projected into two query tensors and the image features $F_I \in \mathbb{R}^{N_i \times L}$ are projected into two key tensors and a value tensor:

$$[Q_1; Q_2] = F_{p^t} W^Q, [K_1; K_2] = F_I W^K, V = F_I W^V \tag{6}$$

where $W^Q \in \mathbb{R}^{L \times 2L}, W^K \in \mathbb{R}^{L \times 2L}, W^V \in \mathbb{R}^{L \times L}$ are the parameters of the convolutional layer in the neural network. The dimensional shape of $F_{p^t} W^Q$ is $N_p \times 2L$, the slice of $F_{p^t} W^Q[:, :L]$ is Q_1, the slice of $F_{p^t} W^Q[:, L:2L]$ is Q_2. K_1 and K_2 are obtained using the same method. The differential cross attention (DCA) mechanism maps queries, keys and value vectors to outputs:

$$M_1 = \frac{Q_1 K_1^T}{\sqrt{L}} \tag{7}$$

$$M_2 = \frac{Q_2 K_2^T}{\sqrt{L}} \tag{8}$$

$$DCA = (softmax(M_1) - \lambda softmax(M_2))V \tag{9}$$

where λ is a learnable scalar, and to synchronize the learning dynamics, we reparameterize the scalar λ as:

$$\lambda = exp(\lambda_{q1} \bullet \lambda_{k1}) - exp(\lambda_{q2} \bullet \lambda_{k2}) + \lambda_{init} \tag{10}$$

where $\lambda_{q1}, \lambda_{k1}, \lambda_{q2}, \lambda_{k2} \in \mathbb{R}^{N_p}$ are learnable vectors and $\lambda_{init} \in (0, 1)$ is a constant used to initialize λ. In the experiments, the λ_{init} is set to 0.8.

In multi-head differential cross attention, Let h denote the number of attention heads. For each head $i \in [1, h]$, we use a different projection matrix to compute the attention weights. It is worth noting that the scalar λ is shared among all attention heads within the same layer. The outputs of each head are normalized and finally stitched together to generate the fused feature representation, as shown in the following computational procedure:

$$head_i = (1 - \lambda_{init}) GN(DCA(F_{p^t}; F_I; W^Q; W^K; W^V; \lambda)) \tag{11}$$

$$F_d = concat(head_1, \ldots, head_h) W^D \tag{12}$$

Where $GN(\cdot)$ denotes using RMS Norm (Zhang & Sennrich, 2019) to normalize each head. $Concat(\cdot)$ denotes connects the heads along the channel dimensions. The DCT uses the difference between the two softmax attention functions to eliminate the cross modal attention noise. By introducing disparity weights for modal features, the contributions of different modal features are dynamically adjusted to avoid overemphasis of irrelevant information and effectively suppress noise interference while enhancing attention to the relevant context.

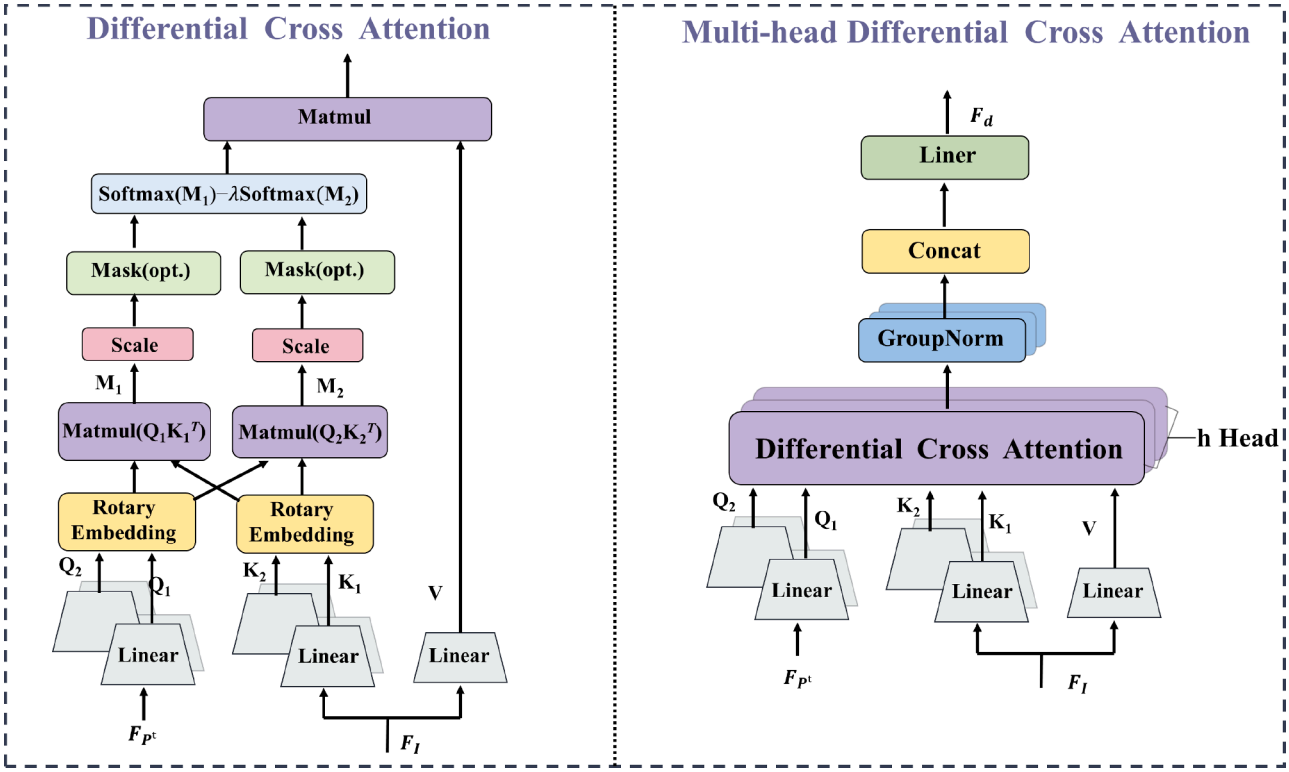

Fig. 3. The architecture of Differential Cross Attention and Multi-head Differential Cross Attention.The core idea of DCA is to compute the difference between two softmax attention maps to suppress the noise interference between cross modal features to improve the performance of feature fusion.

3.3. Coarse point cloud module

The coarse point cloud module is used to convert $F_d \in \mathbb{R}^{N_p \times L}$ into a point cloud. In this process, we employ a multi-branch structure of Axform (Zhang et al., 2022a) to generate point cloud patches for the missing regions. Each branch focuses on generating different parts of the object. To make the feature transformation more flexible and adaptive, we add a 4-layer fully connected network (with the structure of 512, 512, 512, L) before each branch to project the features into multiple subspaces. Next, we transpose the output matrix of each branch and apply the softmax activation function over its N_p dimension to generate an attention map. This attention map is used to weight the features before transposition, and through weighted aggregation and a fully connected network, N_p new points are formed. Please refer to Axform (Zhang et al., 2022a) for specific network details. The input point cloud after FPS, is connected with the output of each branch to form the coarse point cloud $P_0 \in \mathbb{R}^{N_1 \times 3}$.

3.4. Fine point cloud module

The fine point cloud module aims to utilize the information of the two modalities to generate a set of coordinate offsets $\triangle C$ of dimension $N_1 \times 3$, which are added to the coarse point cloud P_0 to generate point-by-point displacements for point cloud refinement.

As shown in Fig. 4, we first use min-pointNet to encode point cloud P_0, extracting the per-point features $F_{P_0} \in \mathbb{R}^{N_1 \times 256}$. To extract global information from the image, we apply a max-pooling operation to obtain the global features of the image, and replicate these features N_o times so that they can be concatenated with the per-point features of the P_0. Then, the point cloud features and global image features are fused through convolution operation to generate multimodal features $F_G \in \mathbb{R}^{N_1 \times 256}$. At this time, the generated feature representation contains not only the local geometric information of the point cloud, but also the global semantic information provided by the image.

To further optimize the multi-modal features F_G, we introduce a simple attention mechanism (SAM):

$$SAM(F_G) = \lambda \cdot softmax(\rho(F_G)) \cdot F_G \tag{13}$$

Where λ is a learnable parameter and ρ denotes the MLP. The softmax function is used to normalize the attention weights . This enables the model to compute the relative importance of each feature and dynamically adjust the contribution of different parts of the multimodal features, ultimately leading to a more effective fusion of point cloud and image information.

Finally, through the above feature processing, the multi-modal features we obtained are inputted into the MLP layer to predict the coordinate offset $\triangle C$. The point-by-point refinement of the coarse point cloud P_0 by the coordinate offset $\triangle C$ results in a more detailed point cloud P_1 that conforms to the global structure, which improves the accuracy and visual effect of the recovered point cloud.

3.5. Loss functions

In order to realize the coarse-to-fine generation process, we use the Chamfer distance ($L1_{cd}$) between the coarse points clouds P_0 and the ground truth P_{GT}, as well as between the fine point clouds P_1 and the ground truth P_{GT}, as the network loss:

$$Loss = L1_{cd}(P_0, P_{GT}) + L1_{cd}(P_1, P_{GT}) \tag{14}$$

where $L1_{cd}$ is defined as:

$$L1_{cd}(X, Y) = \frac{1}{|X|} \sum_{x \in X} \min_{y \in Y} \|x - y\| + \frac{1}{|Y|} \sum_{y \in Y} \min_{x \in X} \|y - x\| \tag{15}$$

4. Experiments

4.1. Dataset

In our experiments, we evaluate the performance of CMFF using ShapeNet-ViPC (Zhang et al., 2021a), a benchmark dataset based on view-guided point cloud completion. ShapeNet-ViPC is constructed based on ShapeNetRendering, which contains 38,328 objects from 13 categories, i.e., Airplanes, Monitors, Automobiles, Chairs, Lamps, Sofas, Phones Boats, Benches, Cabinets, Tables, Speakers, Guns. For each object, we generate 24 incomplete point clouds under 24 viewpoints. The 24 viewpoint settings are the same as ShapeNetRendering. Specifically, each 3D shape is normalized into a bounding sphere of radius 1 and finally rotated to a pose corresponding to a specific viewpoint. For each 3D shape, we uniformly sampled 2048 points from the mesh surface of the target shape under the corresponding viewpoint setting as a ground truth complete point cloud. For the image data, we use the same 24 rendered views as ShapeNetRendering, each with a resolution of 224 × 224 For experiments performed on known categories, we use the same setup as VIPC (Zhang et al., 2021a), i.e., all experiments are performed using 31,650 objects from eight categories, of which 80 % are used for training and 20 % for testing. For generalization ability experiment, the training set was kept the same and we used objects from the four categories not used in the training set for testing.

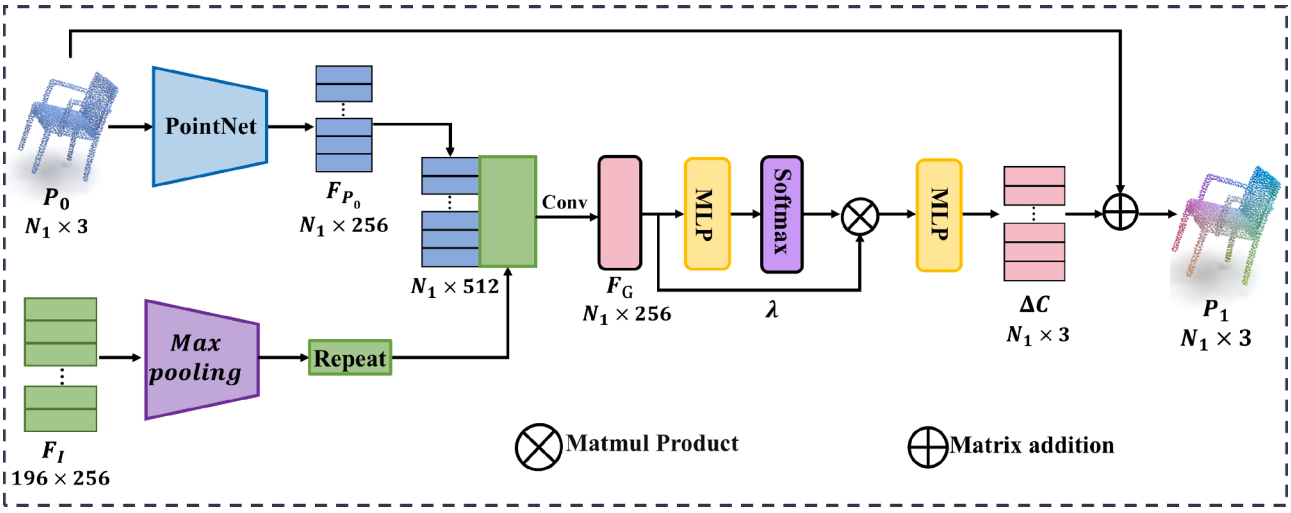

Fig. 4. The architecture of fine point cloud module. P_0 is the coarse point cloud and F_I is the image feature. A set of offset vectors is generated by combining the two modal features to optimize the coarse point cloud.

4.2. Implementation details and evaluation metrics

In our implementation, the input point cloud contains 2048 points. N_p is set to 128, and the neighborhood size k is set to 16. The output feature size of the point cloud encoder is 128 × 256. The output feature map size of the 2D encoder is 196 × 256, and the size of the features after fusion by features is 128 × 256. The coarse point cloud module uses 8 parallel Axform (Zhang et al., 2022a) branching structures, each of which generates a 128 × 3 dimensional point cloud. We use farthest point sampling to sample 1024 points from the input point cloud, which is spliced with the points generated by the Axform (Zhang et al., 2022a) branching structures to produce a point cloud with 2048 points. We performed our experiments on an ubuntu 22.04 system accompanied by a GeForce RTX 4090 graphics card, and during training we used the Adam optimizer with an initial learning rate of 0.001, we trained for 200 epochs, and at the 70th and 160th epochs the learning rate was scaled down by a factor of 10. The complete point cloud contains 2048 points. We evaluate our model using $L2_{cd}$ and F-score, as in previous work. The $L2_{cd}$ can be expressed as:

$$L2_{cd}(x,y) = \frac{1}{2|X|}\sum_{x\in X}\min_{y\in Y}\|x-y\|_2^2 + \frac{1}{2|Y|}\sum_{y\in Y}\min_{x\in X}\|y-x\|_2^2 \tag{16}$$

The F-score is defined as follows, where the threshold d is equal to 0.001, as in previous work:

$$F(X,Y) = \frac{2X(d)Y(d)}{X(d)+Y(d)} \tag{17}$$

4.3. Results on the ShapeNet-ViPC dataset

Baseline methods: In this section, for a comprehensive comparison, we select 11 typical uni-modal point cloud completion methods: AtlasNet (Groueix et al., 2018), FoldingNet (Yang et al., 2018), PCN (Yuan et al., 2018), TopNet (Tchapmi et al., 2019), PF-Net (Huang et al., 2020), MSN (Liu et al., 2020), GRNet (Xie et al., 2020), PoinTr (Yu et al., 2021), PointAttn (Wang et al., 2024), SDT (Zhang et al., 2022b), Seedformer (Zhou et al., 2022) and as well as 4 recent cross modal point cloud completion methods VIPC (Zhang et al., 2021a), CSDN (Zhu et al., 2023), XMFNet (Aiello et al., 2022), and EGIINet (Xu et al., 2025). These methods are representative of recent deep learning-based methods, and their implementations are based on author code available on GitHub. AtlasNet (Groueix et al., 2018) reconstructs the point cloud by estimating the parametric surface elements. FoldingNet (Yang et al., 2018) generates the complete point cloud by using a 2D mesh to perform a folding operation. PCN (Yuan et al., 2018) and MSN (Liu et al., 2020) use a coarse-to-fine point cloud completion method. TopNet (Tchapmi et al., 2019) generates structured point clouds using a tree decoder. PF-Net (Huang et al., 2020) estimates the missing point cloud hierarchically by utilizing a feature-points-based multi-scale generating network. GRNet (Xie et al., 2020) uses voxelization for point cloud completion. PoinTr (Yu et al., 2021), PointAttn (Wang et al., 2024), SDT (Zhang et al., 2022b), and Seedformer (Zhou et al., 2022) are the most recent point cloud completion methods based on the tansformer for point cloud completion. ViPC (Zhang et al., 2021a) is the pioneering work on point cloud completion for single view images.XMFnet (Aiello et al., 2022) uses cross attentive layer and self attentive layer modal fusion techniques for fusing multivariate features. CSDN (Zhu et al., 2023) transforms the modal fusion problem into a style transformation problem. EGIINet (Xu et al., 2025) uses modal alignment with an explicitly guided informative interaction strategy for aligning modalities. For the above method, the output predicted complete point cloud contains 2048 points.

Quantitative Evaluation: we evaluated the performance of different methods on point cloud completion tasks using two metrics: $L2_{cd} \times 10^3$ (CD) and F-Score @ 0.001. Table 1 presents the comparison results for CD, where lower values indicate a smaller discrepancy between the completed and ground-truth point clouds. Our model achieved the best results across all categories, with an average CD of 1.055, significantly outperforming other methods, such as EGIINet (1.211) and XMFNet (1.443). This represents reductions of 12.9 % and 26.9 %, respectively. For individual categories, our method achieved the lowest CD values in complex geometries, such as Chair (1.073) and Lamp (0.574), demonstrating its superiority in capturing detailed features and preserving global geometric structures. Similarly, for simpler geometries like Airplane and Watercraft, our method also achieved the best results, with CD values of 0.469 and 0.706, showcasing its strong generalization capabilities.

Table 2 shows the comparison results for F-Score, where higher values indicate better accuracy and completeness of the model. Our method achieved the best performance across all categories, with an average F-Score of 0.867, significantly surpassing EGIINet (0.836) and XMFNet (0.796) by 3.7 % and 8.9 %, respectively. Specifically, in categories with complex geometries, such as Chair and Lamp, our method achieved F-Scores of 0.883 and 0.849, exceeding other methods, which highlights the model's effectiveness in handling scenes with intricate details and rich local features. Moreover, for categories with regular structures, such as Airplane and Watercraft, our method achieved outstanding F-Scores of 0.976 and 0.964, demonstrating exceptional adaptability to simple geometric shapes. Notably, our method also showed significant improvements in categories such as Table and Sofa, further proving its superiority in completing global structures. Overall, our method

Table 1
Detailed results of CMFF and baseline methods using $L2_{cd} \times 10^3 \downarrow$ evaluation metrics on the ShapeNet-ViPC dataset. Smaller values indicate better complementary results, and bold indicates best.

Methods	Avg	Airplane	Cabinet	Car	Chair	Lamp	Sofa	Table	Watercraft
AtlasNet (Groueix et al., 2018)	6.062	5.032	6.414	4.868	8.161	7.182	6.023	6.561	4.261
FoldingNet (Yang et al., 2018)	6.271	5.242	6.958	5.307	8.823	6.504	6.368	7.080	3.882
PCN (Yuan et al., 2018)	5.619	4.246	6.409	4.840	7.441	6.331	5.668	6.508	3.510
TopNet (Tchapmi et al., 2019)	4.976	3.710	5.629	4.530	6.391	5.547	5.281	5.381	3.350
PF-Net (Huang et al., 2020)	3.873	2.515	4.453	3.602	4.478	5.185	4.113	3.838	2.871
MSN (Liu et al., 2020)	3.793	2.038	5.060	4.322	4.135	4.247	4.183	3.976	2.379
GRNet (Xie et al., 2020)	3.171	1.916	4.468	3.915	3.402	3.034	3.812	3.071	2.160
PoinTr (Yu et al., 2021)	2.851	1.686	4.001	3.203	3.111	2.928	3.507	2.845	1.737
PointAttN (Wang et al., 2024)	2.853	1.613	3.969	3.257	3.157	3.058	3.406	2.787	1.872
SDT (Zhang et al., 2022b)	4.246	3.166	4.807	3.607	5.056	6.101	4.525	3.995	2.856
Seedformer (Zhou et al., 2022)	2.902	1.716	4.049	3.392	3.151	3.226	3.603	2.803	1.679
ViPC (Zhang et al., 2021a)	3.308	1.760	4.558	3.183	2.476	2.867	4.481	4.990	2.197
CSDN (Zhu et al., 2023)	2.570	1.251	3.670	2.977	2.835	2.554	3.240	2.575	1.742
XMFNet (Aiello et al., 2022)	1.443	0.572	1.980	1.754	1.403	1.810	1.702	1.386	0.945
EGIINet (Xu et al., 2025)	1.211	0.534	1.921	1.655	1.204	0.776	1.552	1.227	0.802
Ours	**1.055**	**0.469**	**1.723**	**1.411**	**1.073**	**0.574**	**1.386**	**1.095**	**0.706**

Table 2
Detailed results of CMFF and baseline methods using F-Score @ 0.001↑ evaluation metrics on the ShapeNet-ViPC dataset. Larger values indicate better completion results, and bold indicates best.

Methods	Avg	Airplane	Cabinet	Car	Chair	Lamp	Sofa	Table	Watercraft
AtlasNet (Groueix et al., 2018)	0.410	0.509	0.304	0.379	0.326	0.426	0.318	0.469	0.551
FoldingNet (Yang et al., 2018)	0.331	0.432	0.237	0.300	0.204	0.360	0.249	0.351	0.518
PCN (Yuan et al., 2018)	0.407	0.578	0.270	0.331	0.323	0.456	0.293	0.431	0.577
TopNet (Tchapmi et al., 2019)	0.467	0.593	0.358	0.405	0.388	0.491	0.361	0.528	0.615
PF-Net (Huang et al., 2020)	0.551	0.718	0.399	0.453	0.489	0.559	0.409	0.614	0.656
MSN (Liu et al., 2020)	0.578	0.798	0.378	0.380	0.562	0.652	0.410	0.615	0.708
GRNet (Xie et al., 2020)	0.601	0.767	0.426	0.446	0.575	0.694	0.450	0.639	0.704
PointTr (Yu et al., 2021)	0.683	0.842	0.516	0.545	0.662	0.742	0.547	0.723	0.780
PointAttN (Wang et al., 2024)	0.662	0.841	0.483	0.515	0.638	0.729	0.515	0.699	0.774
SDT (Zhang et al., 2022b)	0.473	0.636	0.291	0.363	0.398	0.442	0.307	0.574	0.602
Seedformer (Zhou et al., 2022)	0.688	0.835	0.551	0.544	0.668	0.777	0.555	0.716	0.786
ViPC (Zhang et al., 2021a)	0.591	0.803	0.451	0.512	0.529	0.706	0.434	0.594	0.730
CSDN (Zhu et al., 2023)	0.695	0.862	0.548	0.560	0.669	0.761	0.557	0.729	0.782
XMFnet (Aiello et al., 2022)	0.796	0.961	0.662	0.691	0.809	0.792	0.723	0.830	0.901
EGIInet (Xu et al., 2025)	0.836	0.969	0.693	0.723	0.847	0.919	0.756	0.857	0.927
Ours	**0.867**	**0.976**	**0.755**	**0.801**	**0.883**	**0.849**	**0.802**	**0.904**	**0.964**

consistently achieved the lowest CD and highest F-Score across all categories, demonstrating its stability and outstanding performance.

Qualitative Evaluation: As can be seen from Fig. 5, our approach exhibits significant advantages over baseline methods (VIPC, CSDN, XMFNet, and EGIINet) for several object classes, including airplanes, cabinets, cars, chairs, lamps, sofas, tables, and watercraft. Specifically, on complex geometries (e.g., chairs and lamps), our method is able to accurately reconstruct details such as the backrest and legs of chairs, and the shades and bases of lamps, which are more difficult to realize with baseline methods. In simple geometric structures (e.g., airplanes, cars, and watercraft), our method is able to generate smooth surfaces and well-defined edges, whereas baseline methods tend to introduce noise or lead to structural incompleteness. In addition, in challenging categories such as cabinets, chairs, and sofas, our method is able to reproduce localized details in greater detail while preserving the overall structure. For example, in the cabinet category, our method completely reconstructs the cabinet doors and frames; in the sofa category, it is able to restore the details of the cushions and armrests, while the results of other methods are blurred or missing. In the table category, our method performs particularly well, successfully reconstructing the complete table legs and avoiding the interference of noise points, which is often difficult to achieve with baseline methods. This result shows the accuracy of the algorithm, which can efficiently finish the point cloud completion task and provide high-quality 3D data to support the subsequent analysis. More visualization results of our method are displayed in Fig. 6.

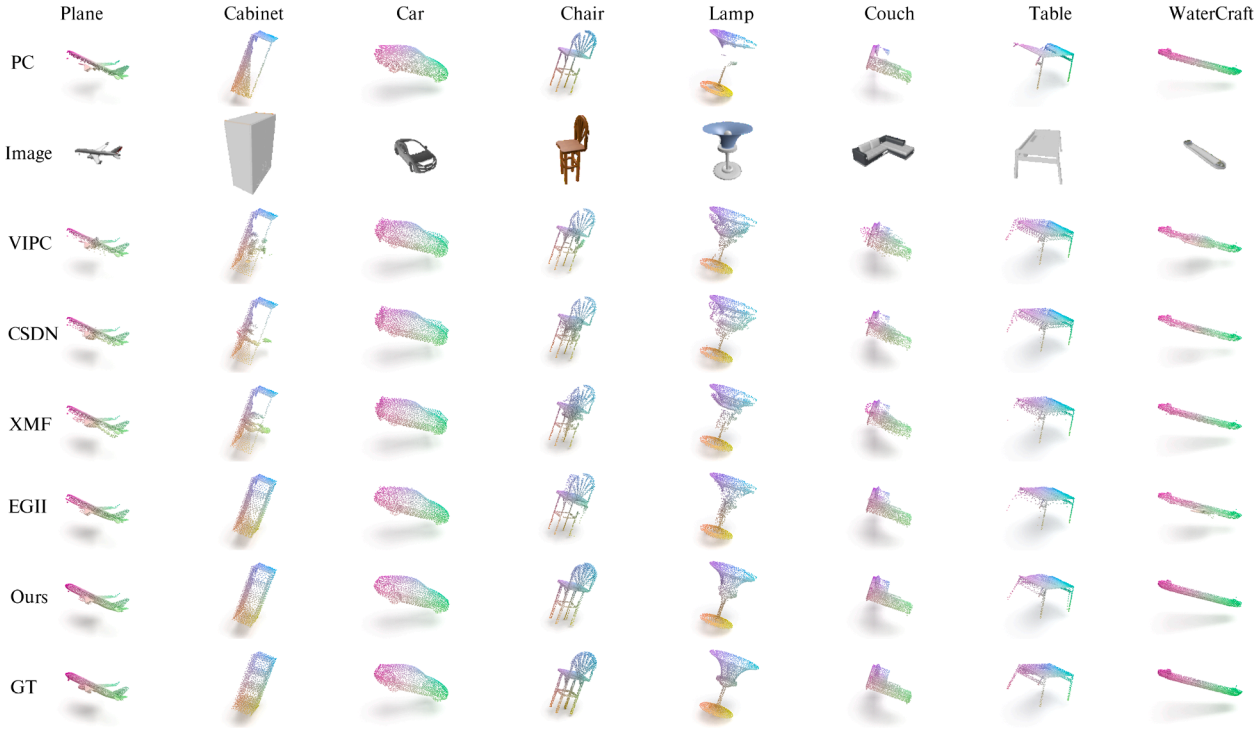

Fig. 5. Qualitative comparison of Our method with four other open source cross modal point cloud completion methods VIPC (Zhang et al., 2021a), CSDN (Zhu et al., 2023), XMFNet (Aiello et al., 2022), EGIINet (Xu et al., 2025) on the ShapeNet-ViPC (Zhang et al., 2021a) dataset. We show results for only some of the objects. Our method achieves less noise and smoother completion.

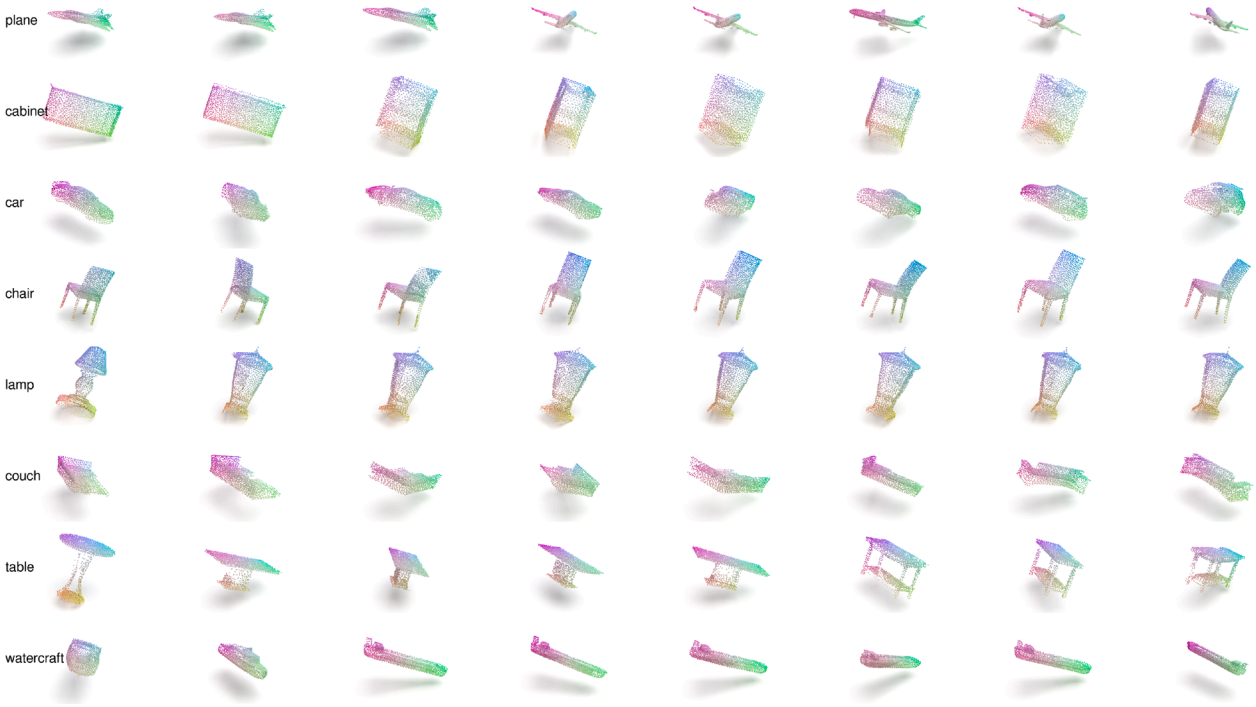

Fig. 6. More visual displays of our method on the ShapeNet-ViPC dataset.

Table 3
Detailed results of CMFF and baseline methods using the $L2_{cd} \times 10^3 \downarrow$ and F-score @ 0.001↑ evaluation metrics on the unknown category of the ShapeNet-ViPC dataset. Bold indicates best.

Methods	Avg		Bench		Monitor		Speaker		Cellphone	
	CD	F-score	CD	F-score	CD	F-score	CD	F-score	CD	F-score
PF-Net (Huang et al., 2020)	5.011	0.468	3.683	0.584	5.304	0.433	7.663	0.319	3.392	0.534
MSN (Liu et al., 2020)	4.684	0.533	2.613	0.706	4.818	0.527	8.259	0.291	3.047	0.607
GRNet (Xie et al., 2020)	4.096	0.548	2.367	0.711	4.102	0.537	6.493	0.376	3.422	0.569
PointTr (Yu et al., 2021)	3.755	0.619	1.976	0.797	4.084	0.599	5.913	0.454	3.049	0.627
PointAttN (Wang et al., 2024)	3.674	0.605	2.135	0.764	3.741	0.591	5.973	0.428	2.848	0.637
SDT (Zhang et al., 2022b)	6.001	0.327	4.096	0.479	6.222	0.268	9.499	0.197	4.189	0.362
ViPC (Zhang et al., 2021a)	4.601	0.498	3.091	0.654	4.419	0.491	7.674	0.313	3.219	0.535
CSDN (Zhu et al., 2023)	3.656	0.631	1.834	0.798	4.115	0.598	5.690	0.485	2.985	0.644
XMFnet (Aiello et al., 2022)	2.671	0.710	1.278	0.862	2.806	0.677	4.823	0.556	1.779	0.748
EGIInet (Xu et al., 2025)	2.354	0.750	1.047	0.902	2.513	0.716	4.282	0.591	1.575	0.792
Ours	**2.147**	**0.767**	**0.916**	**0.913**	**2.134**	**0.731**	**4.105**	**0.596**	**1.435**	**0.823**

4.4. Generalization ability experiment

In the generalization ability experiment, we compare the performance of our method with other baseline methods on the ShapeNet-ViPC dataset for the unknown categories (benches, monitors, stereos, cell phones). From the results in Table 3, our method achieves the best performance with the lowest CD error (2.147) and the highest F-score (0.767). Specifically, for the Bench category (CD = 0.916, F-score = 0.913) and the Cellphone category (CD = 1.435, F-score = 0.823), our method surpasses the best baseline, EGIInet (Bench: CD = 1.047, F-score = 0.902; Cellphone: CD = 1.575, F-score = 0.792), reducing CD error by 12.5 % and 8.9 %, and improving F-score by 1.2 % and 3.9 %, respectively. Moreover, in categories such as Monitor and Speaker, our method also achieves the best results, with CD error reductions of 15.1 % (2.134 vs. 2.513) and 4.1 % (4.105 vs. 4.282), and F-score improvements of 2.1 % (0.731 vs. 0.716) and 0.8 % (0.596 vs. 0.591) compared to EGIInet. These results show that our method is not only able to provide high-quality complementary results in known categories, but also still able to effectively handle point cloud data with different shapes and complexities when facing unknown categories, demonstrating strong generalization capabilities. The qualitative evaluation is displayed in Fig. 7.

4.5. Results on terracotta warriors dataset

To further validate the application of our method in real data, we tested our method on the Terracotta Warriors Dataset (Gao et al., 2024). The experimental results are shown in Table 4 and Fig. 8. The experimental results demonstrate that the proposed method achieves the best performance across all evaluation metrics. Compared to the state-of-the-art baseline GaNet, the average $L2_{cd}$ error is reduced by 34.2 %, while the F-score improves by 8.7 %, consistently outperforming existing methods across all categories. Notably, for complex structures such as War House 1 and Kneeling Archer, the $L2_{cd}$ error decreases by 39.2 % and 37.2 %, respectively, indicating that the proposed method not only effectively reconstructs the overall shape of point clouds but also preserves fine structural details with higher accuracy. This indicates that the method has great potential for application in the field of virtual restoration of cultural heritage.

Table 4
Detailed results of CMFF and baseline methods using the $L2_{cd} \times 10^3 \downarrow$ and F-score @ 0.001↑ evaluation metrics on the Terracotta Warriors dataset. Bold indicates best.

Methods	Avg		Infantryman		War House 1		Kneeling Archer		Warrior		War House 2	
	CD	F-score	CD	F-score	CD	F-score	CD	F-score	CD	F-score	CD	F-score
PCN (Yuan et al., 2018)	4.242	0.320	2.439	0.477	6.497	0.175	5.511	0.207	3.369	0.534	3.395	0.344
PMPNet (Wen et al., 2021)	3.785	0.369	2.183	0.545	5.804	0.201	4.917	0.241	3.304	0.458	2.977	0.398
SaNet (Wen et al., 2020)	2.638	0.565	1.480	0.787	4.101	0.327	3.445	0.400	2.188	0.677	1.975	0.633
SPD (Xiang et al., 2021)	1.862	0.659	1.167	0.864	2.726	0.429	2.400	0.502	1.373	0.811	1.641	0.690
PointAttN (Wang et al., 2024)	2.267	0.555	1.439	0.759	3.312	0.342	2.921	0.399	1.643	0.711	2.019	0.566
GaNet (Gao et al., 2024)	1.717	0.787	0.942	0.919	2.731	0.591	2.280	0.684	1.439	0.848	1.192	0.892
Ours	**1.129**	**0.856**	**0.726**	**0.981**	**1.661**	**0.674**	**1.433**	**0.759**	**0.789**	**0.969**	**1.034**	**0.900**

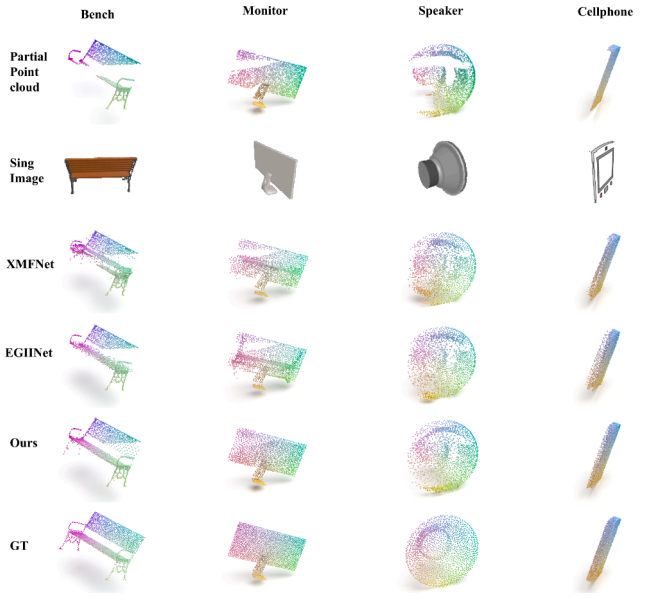

Fig. 7. Qualitative comparison of CMFF with four other open source methods VIPC (Zhang et al., 2021a), CSDN (Zhu et al., 2023), XMFNet (Aiello et al., 2022), and EGIINet (Xu et al., 2025) on unknown categories of the ShapeNet-ViPC dataset.

4.6. Complexity analysis

We conduct a detailed complexity analysis of cross-modal point cloud completion methods. Table 5 reports the theoretical computational cost (FLOPs) and the number of parameters for each model. The scores are measured on the ShapeNet-ViPC dataset with 2048 points and a batch size of 1. For reference, we also provide the overall Chamfer Distance.

As shown in Table 5, CMFF achieves a good balance between completion accuracy and complexity. Among all compared methods, CMFF obtains the lowest average Chamfer Distance (CD-Avg) of 1.055, demonstrating superior completion quality. Although CMFF does not have the absolute lowest parameter count (10.32M) or FLOPs (3.35G), it maintains a relatively low computational cost while delivering the best performance. Compared to VIPC (18.13M) and CSDN (14.20G), which have the highest parameter count and FLOPs respectively, CMFF offers a clear advantage in complexity. Furthermore, although XMFNet (8.72M) and EGII-Net (8.96M) have fewer parameters, CMFF achieves significantly better accuracy with only a slight increase in FLOPs. This indicates that CMFF effectively enhances completion performance without substantially increasing computational resources, highlighting its practicality and deployment potential.

Table 5
Comparison of model parameters, FLOPs, and average Chamfer Distance ($L2_{cd} \times 10^3 \downarrow$).

Methods	Params	FLOPs	CD-Avg
VIPC	18.131M	3.20G	3.308
CSDN	17.67M	14.20G	2.570
XMFNet	**8.72M**	**2.97G**	1.443
EGIINet	8.96M	3.05G	1.211
CMFF	10.32M	3.35G	**1.055**

4.7. Ablation study

In this section, we analyze the effect of different modules on the performance of the proposed method through ablation experiments on the ShapeNet-ViPC dataset. Table 6 lists the results of each experimental setup with $L2_{cd} \times 10^3$ (CD), where smaller values indicate better completion result. By comparing the results of the different setups, we can gain insight into the contribution of each module.

First, to validate the effectiveness of the differential point transformer (DPT) module, we replace the DPT with DGCNN (Phan et al., 2018). The results are shown in the first row of Table 6. w/o DPT's performance is generally lower than that of our method, especially in several categories such as cabinets and watercraft, where the CD are respectively increased from 1.723 to 1.916 and 0.706 to 0.897. In contrast, our method is able to better extract the global and local features of the point cloud with a customized 3D encoder, thus improving the reconstruction accuracy.

In order to demonstrate the effectiveness of differential cross transformer (DCT), we used a common cross attention mechanism to fuse cross modal features. The experimental results are demonstrated in the second row of Table 6. w/o DCT performs poorly, especially in the chair and watercraft categories, where the CD increases from 1.073 to 1.298 and 0.706 to 0.882, respectively. this indicates that the ordinary cross attention mechanism cannot effectively fuse the point cloud and the image information, which leads to poor reconstruction results in some categories. In contrast, our proposed DCT can better combine the information of different modalities, thus improving the reconstruction quality in each category.

Table 6
Quantitative comparison of ablation experiments on the ShapeNet-ViPC dataset. We investigate the effect of our proposed differential point transformer (DPT), differential cross transformer (DCT), fine point cloud (FPC) module and image features on network performance. ($L2_{cd} \times 10^3 \downarrow$).

Methods	Avg	Airplane	Cabinet	Car	Chair	Lamp	Sofa	Table	Watercraft
w/o DPT	1.167	0.482	1.916	1.505	1.087	0.686	1.497	1.269	0.897
w/o DCT	1.250	0.739	2.011	1.528	1.298	0.757	1.493	1.296	0.882
w/o FPC	1.202	0.698	1.831	1.588	1.284	0.631	1.438	1.193	0.951
w/o Image	1.625	0.949	2.131	1.905	1.379	0.978	2.181	1.829	1.645
Ours	**1.055**	**0.469**	**1.723**	**1.411**	**1.073**	**0.574**	**1.386**	**1.095**	**0.706**

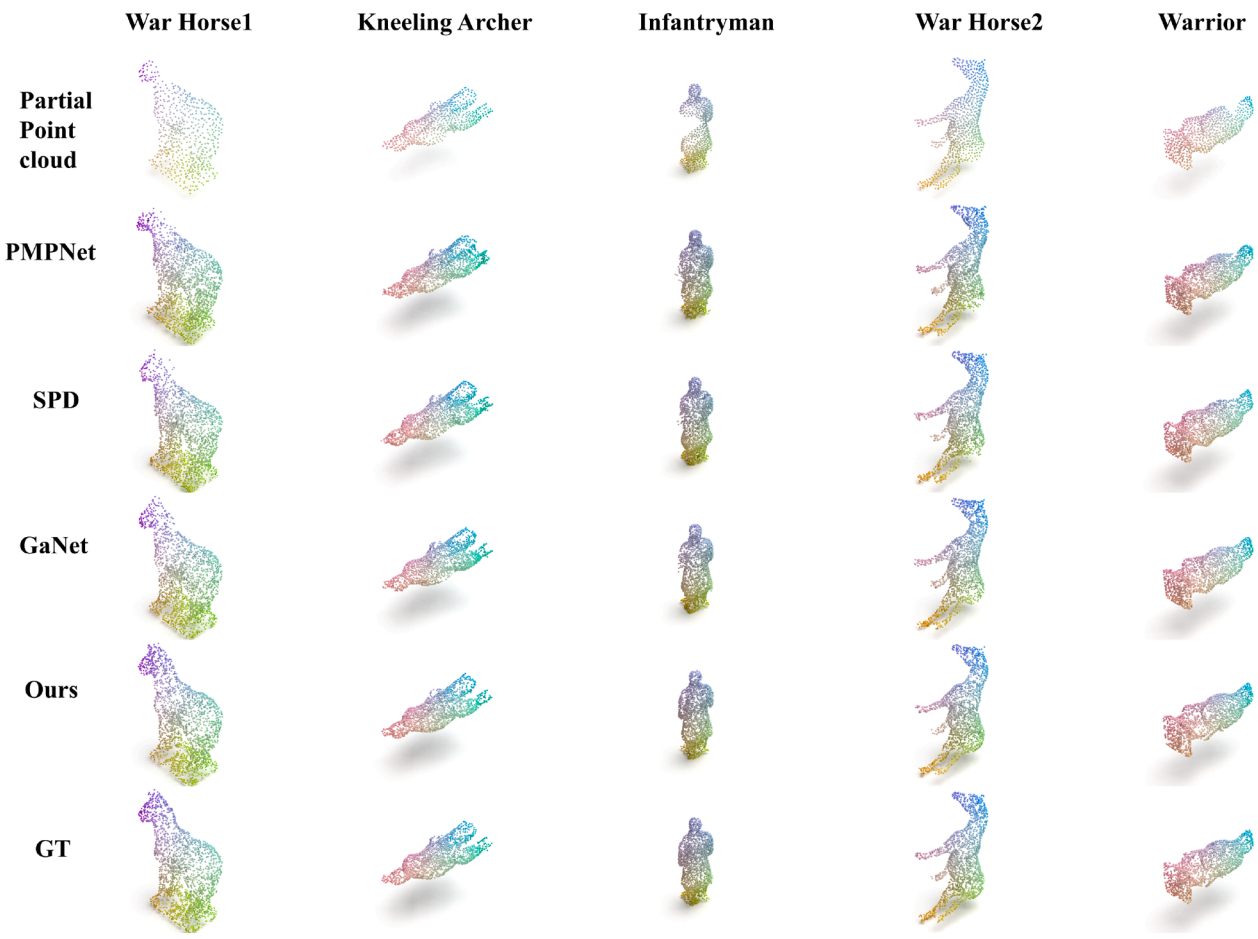

Fig. 8. Qualitative comparison of CMFF with three other open source methods PMPNet (Wen et al., 2021), SPD (Xiang et al., 2021),GaNet (Gao et al., 2024) on Terracotta Warriors dataset. We show the results for only some of the objects.

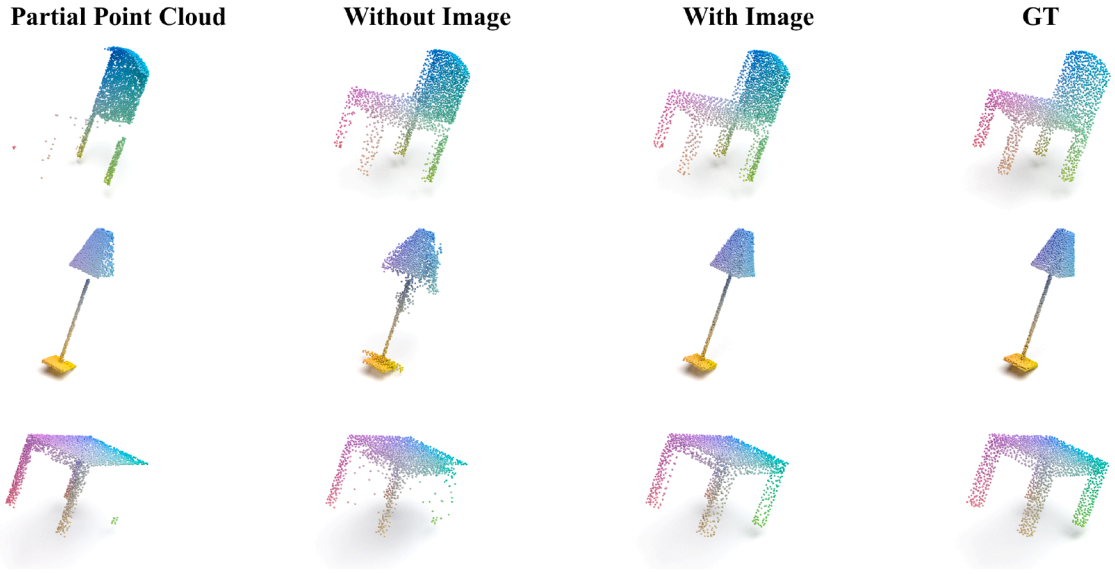

Fig. 9. Comparison of ablation experiments with and without image information.

In order to demonstrate the effectiveness of the fine point cloud module (FPC), we removed the module. As can be seen in the third row of Table 6, the lack of FPC leads to a decrease in the overall performance, especially in the cabinet and sofa categories, where the CD increases from 1.723 to 1.831 and 1.386 to 1.497, respectively, which suggests that FPC plays a crucial role in optimizing the point cloud details and improving the accuracy of the point cloud. Removing the FPC decreases the fineness and accuracy of the reconstruction results, especially when dealing with objects with complex shapes.

To demonstrate the effectiveness of the image features, we used only the point cloud as input and ignored the image information. The

experimental results are displayed in the fourth row of Table 6. This setup led to a significant degradation in performance, especially in the Chair and Table categories, where the CD increased from 1.073 to 1.379 and 1.095 to 1.829, respectively. Fig. 9 illustrates the results of the visualization with and without the image information as an aid. The absence of the image aid during the training process caused the complementary results to converge to the average of the other categories, thus resulting in a significant degradation of the reconstruction quality.

5. Conclusion

In this paper, we propose a novel cross modal point cloud completion network, CMFF, which aims to repair missing parts of 3D models by effectively combining image information. We design a differential point transformer that effectively extracts local details and global structural features in the point cloud, which significantly improves the expressiveness and robustness of the algorithm in dealing with complex geometries, and in particular shows strong advantages in detail restoration and structural integrity. In addition, our proposed differential cross transformer effectively filters out redundant and conflicting features between different modes and enhances the inter-modal feature correlation, and this innovative design significantly improves the accuracy and stability of point cloud completion. In addition, the fine point cloud module facilitates the fusion of different modal features, which further improves the detailed performance of point cloud completion. Experimental results show that CMFF not only outperforms existing techniques in terms of completion accuracy and detail performance, but also has high robustness.

Although CMFF has strong performance, it may perform poorly in certain situations, such as the extreme occlusion of most point clouds or extremely poor selection of single view images. In these cases, when occlusion or single view perspective significantly affects the correlation between point clouds and image features, the model may have difficulty accurately recovering missing point cloud data. Future work will focus on addressing these limitations, and further explore the application of this algorithm in more complex scenarios such as cultural heritage preservation, intelligent manufacturing, and automated driving, thus promoting the practical application of point cloud completion techniques.

CRediT authorship contribution statement

Jian Gao: Writing – original draft, Data curation, Visualization, Methodology; **Yuhe Zhang:** Writing – review & editing; **Pengbo Zhou:** Methodology, Conceptualization; **Xinda Liu:** Visualization, Project administration; **Longquan Yan:** Writing – review & editing, Visualization; **Mingquan Zhou:** Funding acquisition; **Guohua Geng:** Writing – review & editing, Funding acquisition.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was funded by the National Natural Science Foundation of China (62271393), Technology Innovation Leading Project of Shaanxi (2024QY-SZX-11), Xi'an Science and Technology Programme Demonstration Project on Science and Technology Innovation for Social Development (24SFSF0002).

References

Aiello, E., Valsesia, D., & Magli, E. (2022). Cross-modal learning for image-guided point cloud shape completion. *Advances in Neural Information Processing Systems*, *35*, 37349–37362.

Alexey, D. (2020). An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

Berger, M., Tagliasacchi, A., Seversky, L. M., Alliez, P., Levine, J. A., Sharf, A., & Silva, C. T. (2014). State of the art in surface reconstruction from point clouds. In *35th annual conference of the European association for computer graphics, eurographics 2014-state of the art reports*. The Eurographics Association.

Chang, Y., Jung, C., & Xu, Y. (2021). FinerPCN: High fidelity point cloud completion network using pointwise convolution. *Neurocomputing*, *460*, 266–276.

Charles, R. Q., Su, H., Kaichun, M., & Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 77–85).

Chen, X., Chen, B., & Mitra, N. J. (2019). Unpaired point cloud completion on real scans using adversarial training. arXiv preprint arXiv:1904.00069

Davis, J., Marschner, S. R., Garr, M., & Levoy, M. (2002). Filling holes in complex surfaces using volumetric diffusion. In *Proceedings. first international symposium on 3d data processing visualization and transmission* (pp. 428–441). IEEE.

Du, Z., Dou, J., Liu, Z., Wei, J., Wang, G., Xie, N., & Yang, Y. (2024). Cdpnet: Cross-modal dual phases network for point cloud completion. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 1635–1643). (vol. *38*).

Gao, J., Zhang, Y., Liu, Z., & Li, S. (2023). Hdrnet: High-dimensional regression network for point cloud registration. In *Computer graphics forum* (pp. 33–46). Wiley Online Library (vol. *42*).

Gao, J., Zhang, Y., Shiqin, G., Zhou, P., Wen, Y., & Geng, G. (2024). Ganet: Graph attention based terracotta warriors point cloud completion network. *Heritage Science*, *12*(1), 406–419.

Groueix, T., Fisher, M., Kim, V. G., Russell, B. C., & Aubry, M. (2018). A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 216–224).

Guo, M.-H., Cai, J.-X., Liu, Z.-N., Mu, T.-J., Martin, R. R., & Hu, S.-M. (2021). Pct: Point cloud transformer. *Computational Visual Media*, *7*, 187–199.

He, C., Zhao, Z., Zhang, X., Yu, H., & Wang, R. (2025). Rotinv-PCT: Rotation-invariant point cloud transformer via feature separation and aggregation. *Neural Networks*, *185*, 107223.

Huang, Z., Yu, Y., Xu, J., Ni, F., & Le, X. (2020). Pf-net: Point fractal network for 3d point cloud completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7662–7670).

Kim, V. G., Li, W., Mitra, N. J., Chaudhuri, S., DiVerdi, S., & Funkhouser, T. (2013). Learning part-based templates from large collections of 3d shapes. *ACM Transactions on Graphics (TOG)*, *32*(4), 1–12.

Li, G., Chen, Y., Cheng, M., Wang, C., & Li, J. (2020). N-DPC: Dense 3d point cloud completion based on improved multi-stage network. In *Proceedings of the 2020 9th international conference on computing and pattern recognition* (pp. 274–279).

Li, S., Gao, P., Tan, X., & Wei, M. (2023). Proxyformer: Proxy alignment assisted point cloud completion with missing part sensitive transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9466–9475).

Li, Y., Dai, A., Guibas, L., & Nießner, M. (2015). Database-assisted object retrieval for real-time 3d reconstruction. In *Computer graphics forum* (pp. 435–446). Wiley Online Library (vol. *34*).

Liu, M., Sheng, L., Yang, S., Shao, J., & Hu, S.-M. (2020). Morphing and sampling network for dense point cloud completion. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11596–11603). (vol. *34*).

Liu, Z., Zhang, Y., Gao, J., & Wang, S. (2022). Vfmvac: View-filtering-based multi-view aggregating convolution for 3d shape recognition and retrieval. *Pattern Recognition*, *129*, 108774.

Ma, C., Yang, Y., Guo, J., Wang, C., & Guo, Y. (2022). Completing partial point clouds with outliers by collaborative completion and segmentation. arXiv preprint arXiv:2203.09772

Mao, A., Tang, Y., Huang, J., & He, Y. (2024). Dmf-net: Image-guided point cloud completion with dual-channel modality fusion and shape-aware upsampling transformer. arXiv preprint arXiv:2406.17319

Mitra, N. J., Guibas, L. J., & Pauly, M. (2006). Partial and approximate symmetry detection for 3d geometry. *ACM Transactions on Graphics (ToG)*, *25*(3), 560–568.

Mitra, N. J., Pauly, M., Wand, M., & Ceylan, D. (2013). Symmetry in 3d geometry: Extraction and applications. In *Computer graphics forum* (pp. 1–23). Wiley Online Library (vol. *32*).

Nguyen, D. T., Hua, B.-S., Tran, K., Pham, Q.-H., & Yeung, S.-K. (2016). A field model for repairing 3d shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5676–5684).

Pan, L. (2020). Ecg: Edge-aware point cloud completion with graph convolution. *IEEE Robotics and Automation Letters*, *5*(3), 4392–4398.

Pauly, M., Mitra, N. J., Giesen, J., Gross, M. H., & Guibas, L. J. (2005). Example-based 3d scan completion. In *Symposium on geometry processing* (pp. 23–32).

Phan, A. V., Le Nguyen, M., Nguyen, Y. L. H., & Bui, L. T. (2018). Dgcnn: A convolutional neural network over large-scale labeled graphs. *Neural Networks*, *108*, 533–543.

Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017). Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, *30*, 5105–5114.

Sarmad, M., Lee, H. J., & Kim, Y. M. (2019). Rl-gan-net: A reinforcement learning agent controlled gan network for real-time point cloud shape completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5898–5907).

Shao, T., Xu, W., Zhou, K., Wang, J., Li, D., & Guo, B. (2012). An interactive approach to semantic modeling of indoor scenes with an rgbd camera. *ACM Transactions on Graphics (TOG)*, *31*(6), 1–11.

Shao, Y., Tan, A., Wang, B., Yan, T., Sun, Z., Zhang, Y., & Liu, J. (2024). Ms23d: A 3d object detection method using multi-scale semantic feature points to construct 3d feature layer. *Neural Networks*, *179*, 106623.

Shen, C.-H., Fu, H., Chen, K., & Hu, S.-M. (2012). Structure recovery by part assembly. *ACM Transactions on Graphics (TOG)*, *31*(6), 1–11.

Shi, J., Xu, L., Heng, L., & Shen, S. (2021). Graph-guided deformation for point cloud completion. *IEEE Robotics and Automation Letters*, *6*(4), 7081–7088.

Song, W., Zhou, J., Wang, M., Tan, H., Li, N., & Liu, X. (2023). Fine-grained text and image guided point cloud completion with CLIP model. arXiv preprint arXiv:2308.08754

Stutz, D., & Geiger, A. (2018). Learning 3d shape completion from laser scan data with weak supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1955–1964).

Sun, Y., Wang, Y., Liu, Z., Siegel, J., & Sarma, S. (2020). Pointgrow: Autoregressively learned point cloud generation with self-attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 61–70).

Sung, M., Kim, V. G., Angst, R., & Guibas, L. (2015). Data-driven structural priors for shape completion. *ACM Transactions on Graphics (TOG)*, *34*(6), 1–11.

Tang, J., Gong, Z., Yi, R., Xie, Y., & Ma, L. (2022). Lake-net: Topology-aware point cloud completion by localizing aligned keypoints. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1726–1735).

Tchapmi, L. P., Kosaraju, V., Rezatofighi, H., Reid, I., & Savarese, S. (2019). Topnet: Structural point cloud decoder. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 6000–6010.

Wang, J., Cui, Y., Guo, D., Li, J., Liu, Q., & Shen, C. (2024). Pointattn: You only need attention for point cloud completion. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 5472–5480). (vol. *38*).

Wen, X., Li, T., Han, Z., & Liu, Y.-S. (2020). Point cloud completion by skip-attention network with hierarchical folding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1939–1948).

Wen, X., Xiang, P., Han, Z., Cao, Y.-P., Wan, P., Zheng, W., & Liu, Y.-S. (2021). Pmp-net: Point cloud completion by learning multi-step point moving paths. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7443–7452).

Xia, Y., Xia, Y., Li, W., Song, R., Cao, K., & Stilla, U. (2021). Asfm-net: Asymmetrical siamese feature matching network for point completion. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 1938–1947).

Xiang, P., Wen, X., Liu, Y.-S., Cao, Y.-P., Wan, P., Zheng, W., & Han, Z. (2021). Snowflakenet: Point cloud completion by snowflake point deconvolution with skip-transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5499–5509).

Xie, H., Yao, H., Zhou, S., Mao, J., Zhang, S., & Sun, W. (2020). Grnet: Gridding residual network for dense point cloud completion. In *European conference on computer vision* (pp. 365–381). Springer.

Xu, H., Long, C., Zhang, W., Liu, Y., Cao, Z., Dong, Z., & Yang, B. (2025). Explicitly guided information interaction network for cross-modal point cloud completion. In *European conference on computer vision* (pp. 414–432). Springer.

Yang, W., Huang, J., He, X., & Wen, S. (2024). Fixed-time synchronization of complex-valued neural networks for image protection and 3d point cloud information protection. *Neural Networks*, *172*, 106089. https://doi.org/10.1016/j.neunet.2023.12.043

Yang, Y., Feng, C., Shen, Y., & Tian, D. (2018). Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 206–215).

Yu, L., Li, X., Fu, C.-W., Cohen-Or, D., & Heng, P.-A. (2018). Pu-net: Point cloud upsampling network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2790–2799).

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

Yuan, W., Khot, T., Held, D., Mertz, C., & Hebert, M. (2018). Pcn: Point completion network. In *2018 international conference on 3d vision (3DV)* (pp. 728–737). IEEE.

Zhang, B., & Sennrich, R. (2019). Root mean square layer normalization. *Advances in Neural Information Processing Systems*, *32*, 12381–12392.

Zhang, K., Hao, M., Wang, J., de Silva, C. W., & Fu, C. (2019). Linked dynamic graph CNN: Learning on point cloud via linking hierarchical features. arXiv preprint arXiv:1904.10014

Zhang, K., Yang, X., Wu, Y., & Jin, C. (2022a). Attention-based transformation from latent features to point clouds. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 3291–3299). (vol. *36*).

Zhang, W., Zhou, H., Dong, Z., Liu, J., Yan, Q., & Xiao, C. (2022b). Point cloud completion via skeleton-detail transformer. *IEEE Transactions on Visualization and Computer Graphics*, *29*(10), 4229–4242.

Zhang, X., Feng, Y., Li, S., Zou, C., Wan, H., Zhao, X., Guo, Y., & Gao, Y. (2021a). View-guided point cloud completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15890–15899).

Zhang, Y., Huang, D., & Wang, Y. (2021b). Pc-rgnn: Point cloud completion and graph neural network for 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 3430–3437). (vol. *35*).

Zhao, H., Jiang, L., Jia, J., Torr, P. H. S., & Koltun, V. (2021). Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 16259–16268).

Zhao, Y., Xie, J., Qian, J., & Yang, J. (2020). Pui-net: A point cloud upsampling and inpainting network. In *Pattern recognition and computer vision: Third chinese conference, PRCV 2020, nanjing, china, october 16–18, 2020, proceedings, part i 3* (pp. 328–340). Springer.

Zhou, H., Cao, Y., Chu, W., Zhu, J., Lu, T., Tai, Y., & Wang, C. (2022). Seedformer: Patch seeds based point cloud completion with upsample transformer. In *European conference on computer vision* (pp. 416–432). Springer.

Zhu, Z., Nan, L., Xie, H., Chen, H., Wang, J., Wei, M., & Qin, J. (2023). Csdn: Cross-modal shape-transfer dual-refinement network for point cloud completion. *IEEE Transactions on Visualization and Computer Graphics*, *30*(7), 3545–3563.