# Multi-Modal Point Cloud Completion with Interleaved Attention Enhanced Transformer

**Chenghao Fang**[1], **Jianqing Liang**[1,*], **Jiye Liang**[1], **Hangkun Wang**[4], **Kaixuan Yao**[1] and **Feilong Cao**[2,3]

[1] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi Taihang Laboratory, School of Computer and Information Technology, Shanxi University

[2] School of Mathematical Sciences, Zhejiang Normal University

[3] Institute of Mathematics and Cross-disciplinary Science, Zhejiang Normal University

[4] School of Mathematics, Northwest University

chenghaofang1014@163.com, {liangjq, ljy, ykx}@sxu.edu.cn, {whk123zxc, icteam}@163.com

## Abstract

Multi-modal point cloud completion, which utilizes a complete image and a partial point cloud as input, is a crucial task in 3D computer vision. Previous methods commonly employ a cross-attention mechanism to fuse point clouds and images. However, these approaches often fail to fully leverage image information and overlook the intrinsic geometric details of point clouds that could complement the image modality. To address these challenges, we propose an interleaved attention enhanced Transformer (IAET) with three main components, i.e., token embedding, bidirectional token supplement, and coarse-to-fine decoding. IAET incorporates a novel interleaved attention mechanism to enable bidirectional information supplementation between the point cloud and image modalities. Additionally, to maximize the use of the supplemented image information, we introduce a view-guided upsampling module that leverages image tokens as queries to guide the generation of detailed point cloud structures. Extensive experiments demonstrate the effectiveness of IAET, highlighting its state-of-the-art performance on multi-modal point cloud completion benchmarks in various scenarios. The source code is freely accessible at https://github.com/doldolOuO/IAET.

## 1 Introduction

Point cloud, as a common 3D data format, is widely used in real-world scenarios such as autonomous driving [Chen *et al.*, 2024], medical treatment [Liu *et al.*, 2023], and robotics [Cheng *et al.*, 2022]. Given its importance and versatility, many researchers are currently interested in 3D computer vision. Owing to inherent limitations of 3D sensor devices and environmental factors, point cloud data collected in real-world settings is often sparse and incomplete, which adversely impacts downstream tasks such as point cloud segmentation [Du *et al.*, 2024c; Liang *et al.*, 2023; Du *et al.*, 2024b], registration [Zhang *et al.*, 2023b; Mu *et al.*, 2024], and detection [Pei *et al.*, 2023; Zhang *et al.*, 2021b].

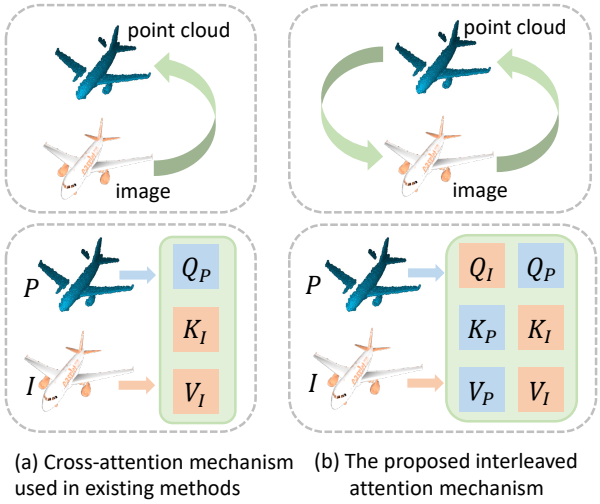


Figure 1: An illustration of our core idea. Subfigure (a) illustrates that existing methods typically employ a cross-attention mechanism to fuse point cloud and image data, where information transfer is unidirectional (from the image to the point cloud). In contrast, Subfigure (b) demonstrates that the proposed method utilizes an interleaved attention mechanism to enable bidirectional information transfer between the point cloud and image modalities.

Single-modal point cloud completion [Fei *et al.*, 2022; Tesema *et al.*, 2024], restoring the incomplete input to 3D objects with complete shapes, is proposed to address this challenge. From the perspective of the generated completion results, existing methods can be divided into two categories. Some approaches [Huang *et al.*, 2020; Alliegro *et al.*, 2021; Yu *et al.*, 2021] focus on generating missing shapes, while others [Yuan *et al.*, 2018; Zhou *et al.*, 2022; Wang *et al.*, 2024] directly produce complete geometric structures. Both

*Corresponding author

usually encode the incomplete input into a global representation. The difference lies in the decoding part; the former usually restores the global representation to the shape of the missing area, while the latter restores the global representation to a complete 3D object. While the above methods have achieved remarkable results, they are still difficult to recover enough details of the missing areas due to the incompleteness of partial input [Zhang *et al.*, 2021a; Zhang *et al.*, 2022].

To address this challenge, a straightforward approach introduces additional modality information to assist point cloud completion. Zhang et al. [Zhang *et al.*, 2021a] propose a view-guided point cloud completion (ViPC) framework that, for the first time, combines image information with point cloud information for point cloud completion. Notably, the major challenge in view-guided point cloud completion is how to efficiently utilize image modality. To address this challenge, CSDN [Zhu *et al.*, 2024] proposes a novel IPAdaIN module to transfer global shape information of image modality to that of point cloud modality. Different from CSDN, most methods such as XMFNet [Aiello *et al.*, 2022], CDPNet [Du *et al.*, 2024a], and EGIINet [Xu *et al.*, 2024] adopt a cross-attention mechanism to fuse point cloud and image modalities. As shown in Figure 1, although the above methods have achieved certain performance, they do not fully utilize the image information and ignore that the structural information in the point cloud representation can enhance the geometry of the image representation.

To address above challenges, we propose the interleaved attention enhanced Transformer (IAET) consisting of three parts, i.e., token embedding, bidirectional token supplement, and coarse-fine decoding. Different from previous methods that use a cross-attention mechanism to fuse point cloud and image modalities in a unidirectional manner, we adopt a bidirectional information transfer strategy to enhance the interaction between modalities. Specifically, we employ existing 2D and 3D feature extractors to generate image tokens and point cloud tokens. Subsequently, we propose the interleaved attention mechanism which enables mutual supplementation between the point cloud and image tokens. Finally, we utilize the supplemented image tokens to guide the transformation of point cloud tokens into the completion results. IAET demonstrates state-of-the-art performance on the multi-modal point cloud completion benchmark datasets. The main contributions of this paper are as follows.

- For multi-modal point cloud completion, we suggest the interleaved attention enhanced Transformer (IAET) which aims to fully use image information to assist the point cloud completion process.
- We propose an interleaved attention mechanism to enable bidirectional information supplement across modalities, that is, enhancing point cloud representations with richer semantic detail and infusing image representations with supplementary geometric information.
- We propose a view-guided upsampling module introducing image information to enhance the generation of more detailed and accurate completion results. Specifically, this module utilizes image information as a query to retrieve relevant point cloud representations, enabling it to capture point clouds' local geometric structures.

## 2 Related Work

### 2.1 Single-modal Point Cloud Completion

PCN [Yuan *et al.*, 2018] is introduced as the first end-to-end framework for point cloud completion. Building on PCN, TopNet [Tchapmi *et al.*, 2019] proposes a tree-structured decoder that allows generation of completion results with arbitrary resolutions. Subsequently, Liu et al. [Liu *et al.*, 2020] develop a morphing and sampling network that achieves completion results with smooth surfaces. PF-Net [Huang *et al.*, 2020] marks a notable advance by focusing on generating the shape of missing regions using a multi-resolution encoder and a point pyramid decoder. To further refine missing shapes, PoinTr [Yu *et al.*, 2021] pioneers the use of a full Transformer [Vaswani *et al.*, 2017] architecture for point cloud completion, signaling the potential of Transformers for this task. Recently, SDT [Zhang *et al.*, 2023a] introduces a skeleton-detail Transformer that emphasizes skeleton points in coarse completion results, while Seedformer [Zhou *et al.*, 2022] applies an upsample Transformer to increase point density. More recently, PointAttN [Wang *et al.*, 2024] presents a fully attention-based network architecture to enhance the completion of partial objects.

### 2.2 Multi-modal Point Cloud Completion

The development of devices such as cross-modal sensors has enabled multi-modal point cloud completion. As pioneers in this field, Zhang et al. [Zhang *et al.*, 2021a] are the first to construct a multi-modal point cloud completion dataset incorporating both point cloud and image data and propose ViPC for multi-modal point cloud completion. A key challenge in multi-modal point cloud completion lies in effectively leveraging image information to aid point cloud completion. To address this challenge, CSDN [Zhu *et al.*, 2024] introduces a novel IPAdaIN module, which facilitates the transfer of global shape information from images to point clouds. XMFNet [Aiello *et al.*, 2022] first uses the cross-attention mechanism to fuse modality and point cloud representation, achieving excellent performance. CDPNet [Du *et al.*, 2024a] and EGIINet [Xu *et al.*, 2024] also use the cross-attention mechanism to fuse the two modalities. Specifically, the former uses a two-stage network to complete partial input. The latter focuses on the extraction and aligning of two modalities' features. Compared with ViPC and CSDN, XMFNet, CDPNet, and EGIINet, which use the cross-attention mechanism to fuse image and point cloud features, have better completion performance. However, the information transmission approach that utilizes the cross-attention mechanism to fuse point clouds and images is unidirectional, specifically transferring information from the image to the point cloud. This approach does not fully exploit the potential of the image data and overlooks the fact that the geometric information inherent in the partial point cloud could enhance the geometric representation of the image modality. IAET aims to explore the impact of bidirectional information transmission between point cloud and image on multi-modal point cloud completion.

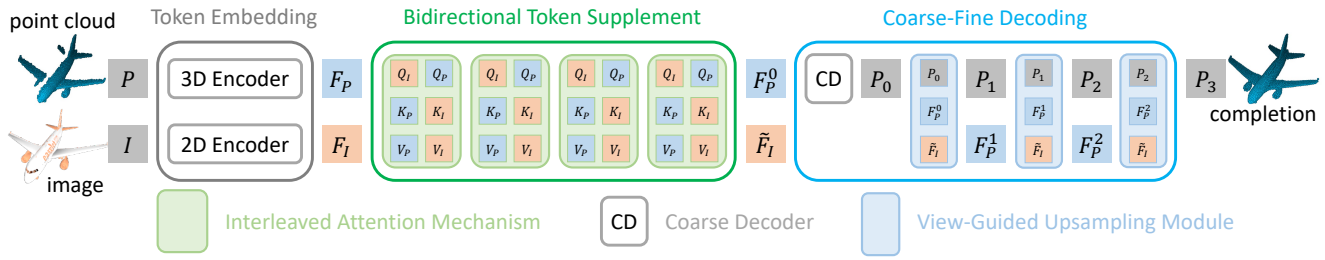


Figure 2: Overview of the proposed IAET consists of token embedding, bidirectional token supplement, and coarse-fine decoding.

# 3 Method

## 3.1 Overview

As shown in Figure 2, IAET mainly consists of three parts, i.e., token embedding, bidirectional token supplement, and coarse-fine decoding. During the token embedding stage, a partial point cloud $P$ and a full image $I$ are transformed into point cloud tokens $F_P$ and image tokens $F_I$ using PointNet++ [Qi *et al.*, 2017b] and ResNet [He *et al.*, 2016], respectively. In the bidirectional token supplement stage, interleaved attention mechanisms are applied to enhance the point cloud tokens and image tokens, resulting in supplemented point cloud tokens $F_P^0$ and supplemented image tokens $\tilde{F}_I$. Finally, in the coarse-fine decoding stage, the supplemented point cloud tokens $F_P^0$ are processed through a coarse decoder, producing a coarse point cloud $P_0$. The coarse point cloud $P_0$, along with the supplemented point cloud tokens $F_P^0$ and supplemented image tokens $\tilde{F}_I$, is then refined using three view-guided upsampling modules to generate the completion outcome $P_3$. Next, we will provide a detailed introduction to the two parts, i.e., bidirectional token supplement and coarse-fine decoding, along with their key components.

## 3.2 Bidirectional Token Supplement

Existing methods usually use a cross-attention mechanism to fuse the representation of point clouds and images. This unidirectional method of information transmission fails to fully utilize the rich information embedded in images and overlooks the geometric features of point clouds, which have the potential to enhance image representation. Unlike previous approaches, IAET employs a bidirectional token supplementation strategy during the fusion of the two modalities. This strategy facilitates the generation of a more informative point cloud representation while incorporating geometric information into the image representation. As illustrated in Figure 2, this process is implemented through multiple interleaved attention mechanisms.

**Interleaved Attention Mechanism.** To concisely illustrate the principle underlying the interleaved attention mechanism, we designate the inputs to this mechanism as point cloud tokens $F_P$ and image tokens $F_I$, as depicted in Figure 3. Initially, the point cloud enhances the information content of the image. In this process, image tokens $F_I$ serve as the query, while point cloud tokens $F_P$ serve as the key and value. The

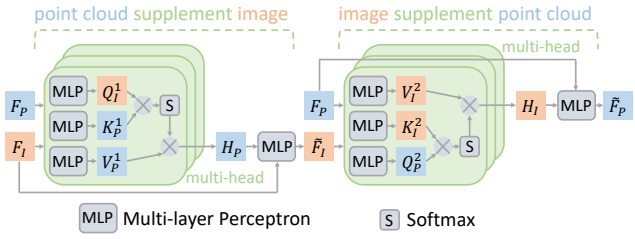


Figure 3: Overview of the interleaved attention mechanism.

corresponding mathematical formulation is as

$$\begin{cases} Q_I^1 = \mathrm{MLP}(F_I), \\ K_P^1 = \mathrm{MLP}(F_P), \\ V_P^1 = \mathrm{MLP}(F_P), \end{cases} \tag{1}$$

where $\mathrm{MLP}(\cdot)$ is the multi-layer perceptron. Then, the geometric information $H_P$ is obtained from the query $Q_I^1$, key $K_P^1$, and value $V_P^1$, and the formula is as

$$H_P = \mathrm{Softmax}(\frac{Q_I^1 {K_P^1}^T}{\sqrt{d_1}})V_P^1, \tag{2}$$

where $\mathrm{Softmax}(\cdot)$ is the Softmax function and $\sqrt{d_1}$ is the dimension of query $Q_I^1$, key $K_P^1$, and value $V_P^1$. Finally, the geometric information extracted from point cloud tokens is supplemented to the image tokens, and its formula is as

$$\tilde{F}_I = \mathrm{MLP}(H_P + F_I), \tag{3}$$

where $\tilde{F}_I$ is the supplemented image tokens. Subsequently, a similar operation is performed to facilitate the information transfer from supplemented image tokens to point cloud tokens. In this step, point cloud tokens $F_P$ are utilized as query, while supplemented image tokens $\tilde{F}_I$ serve as key and value. The corresponding mathematical formulation is as

$$\begin{cases} Q_P^2 = \mathrm{MLP}(F_P), \\ K_I^2 = \mathrm{MLP}(\tilde{F}_I), \\ V_I^2 = \mathrm{MLP}(\tilde{F}_I). \end{cases} \tag{4}$$

After that, we use complete image tokens with geometric information to supplement incomplete point cloud tokens and its formula is as

$$H_I = \mathrm{Softmax}(\frac{Q_P^2 {K_I^2}^T}{\sqrt{d_2}})V_I^2, \tag{5}$$

where $\sqrt{d_2}$ is the dimension of query $Q_P^2$, key $K_I^2$, and value $V_I^2$. Finally, point cloud tokens are supplemented with extracted image information $H_I$, and its formula is as

$$\tilde{F}_P = \mathrm{MLP}(H_I + F_P). \tag{6}$$

The output of an interleaved attention mechanism is supplemented point cloud tokens $\tilde{F}_P$ and supplemented image tokens $\tilde{F}_I$.

In summary, the proposed bidirectional token supplement integrates several interleaved attention mechanisms to enable effective mutual supplementation of information between point clouds and images. Moreover, the enriched image representation is leveraged to guide the subsequent point generation process, ensuring the comprehensive utilization of image information.

### 3.3 Coarse-Fine Decoding

This section provides a detailed explanation of the final stage of IAET. In the coarse-fine decoding stage, IAET utilizes the supplemented point cloud tokens $F_P^0$ to reconstruct a coarse point cloud $P_0$ through deconvolution and MLPs. Subsequently, as shown in Figure 2, the coarse point cloud $P_0$, supplemented point cloud tokens $F_P^0$, and supplemented image tokens $\tilde{F}_I$ are processed through three view-guided upsampling modules to produce the final completion result $P_3$.

**View-Guided Upsampling Module.** To fully leverage image information, IAET embeds the supplemented image tokens into the proposed upsampling module. This enables the use of image data to guide the capture of fine-grained details within the point cloud tokens. As illustrated in Figure 4, to succinctly explain the principle of the view-guided upsampling module, we consider the inputs of the $i$-th module to be image tokens $\tilde{F}_I \in \mathbb{R}^{M\times C}$, point cloud tokens $F_P^i \in \mathbb{R}^{N\times C}$, and a low-resolution point cloud $P_i \in \mathbb{R}^{N\times 3}$. First, we use cosine similarity to calculate the similarity matrix $A \in \mathbb{R}^{N\times M}$ between each point cloud token and the image token. The formula is as

$$\begin{cases} \tilde{A}_{ij} = \dfrac{f_i \cdot \tilde{f}_j^T}{\|{f_i}^T\|_2 \cdot \|\tilde{f}_j\|_2}, \\ A = \mathrm{Softmax}(\tilde{A}), \end{cases} \tag{7}$$

where $f_i \in \mathbb{R}^{1\times C}$ and $\tilde{f}_j \in \mathbb{R}^{1\times C}$ denote $i$-th point cloud token and $j$-th image token, respectively. $\mathrm{Softmax}(\cdot)$ denotes the Softmax function. After that, the image tokens and point cloud tokens are processed through an MLP to generate the queries and keys, while the point cloud is processed via PointNet [Qi *et al.*, 2017a] to derive the values. The formula is as

$$\begin{cases} Q = \mathrm{MLP}(\tilde{F}_I), \\ K = \mathrm{MLP}(F_P^i), \\ V = \mathrm{PointNet}(P_i), \end{cases} \tag{8}$$

where $Q \in \mathbb{R}^{M\times C}$, $K \in \mathbb{R}^{N\times C}$, and $V \in \mathbb{R}^{N\times C}$ denote queries, keys, and values. Each query $q_i (i = 1, 2, \ldots, M)$ represents the shape features of a local region within the image, while each key $k_i (i = 1, 2, \ldots, N)$ represents the geometric features of a local region in the point cloud. Next, we

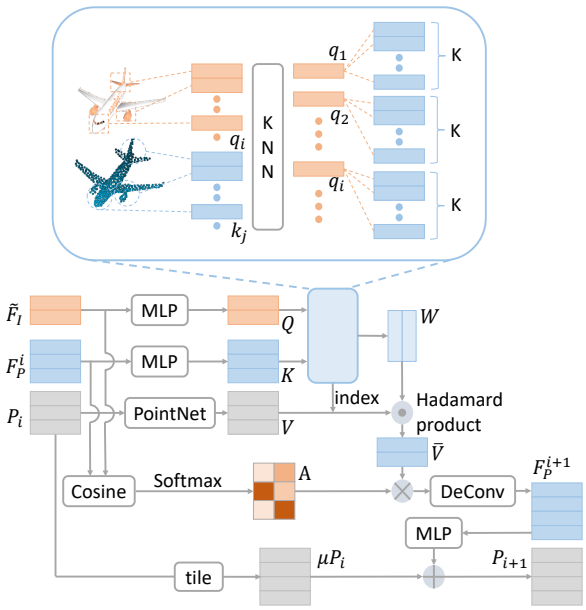


Figure 4: Overview of the view-guided upsampling module.

employ the K-nearest neighbor (KNN) algorithm to identify the K nearest keys for each query, with the goal of retrieving relevant point cloud shapes corresponding to each image shape. Specifically, we use Euclidean distance to quantify the similarity between a query and a key, and the resulting values are passed through a Softmax function to compute the corresponding weights $W \in \mathbb{R}^{M\times K\times C}$. Then, point cloud features are aggregated based on the shape of each image and its formulation is as

$$\tilde{v}_i = \sum_{j=1}^{\mathrm{K}} (w_j^i \odot v_j), i = 1, 2, \ldots, M, \tag{9}$$

where $\tilde{V} \in \mathbb{R}^{M\times C}$ contains $M$ values and $w_j^i \in \mathbb{R}^{1\times C}$ denote $j$-th weight produced by $i$-th image token. Finally, the point cloud tokens $F_P^{i+1} \in \mathbb{R}^{\mu N\times C}$ output by the $i$-th module are produced, and its formulation as

$$F_P^{i+1} = \mathrm{DeConv}(A\tilde{V}). \tag{10}$$

And a high-resolution point cloud $P_{i+1} \in \mathbb{R}^{\mu N\times 3}$ output by the $i$-th module is produced, and its formulation as

$$P_{i+1} = \mathrm{tile}(P_i) + \mathrm{MLP}(F_P^{i+1}), \tag{11}$$

where $\mathrm{tile}(\cdot)$ is the copy operation aiming to copy the coordinates of the point cloud $P_i$ $\mu$ times.

In conclusion, the coarse-fine decoding framework of IAET consists of a coarse decoder and several view-guided upsampling modules to generate the completed point clouds. Specifically, image information is integrated into the proposed view-guided upsampling module, where image tokens serve as queries to guide the generation of offset values for high-resolution point clouds. This approach is designed to produce fine-grained completion results with sufficient local details.

| Methods | CD ↓ / F-Score ↑ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Avg | Airplane | Cabinet | Car | Chair | Lamp | Sofa | Table | Watercraft |
| Single-modal Methods | | | | | | | | | |
| AtlasNet [Groueix *et al.*, 2018] | 6.062 / 0.410 | 5.032 / 0.509 | 6.414 / 0.304 | 4.868 / 0.379 | 8.161 / 0.326 | 7.182 / 0.426 | 6.023 / 0.318 | 6.561 / 0.469 | 4.261 / 0.551 |
| FoldingNet [Yang *et al.*, 2018] | 6.271 / 0.331 | 5.242 / 0.432 | 6.958 / 0.237 | 5.307 / 0.300 | 8.823 / 0.204 | 6.504 / 0.360 | 6.368 / 0.249 | 7.080 / 0.351 | 3.882 / 0.518 |
| PCN [Yuan *et al.*, 2018] | 5.619 / 0.407 | 4.246 / 0.578 | 6.409 / 0.270 | 4.840 / 0.331 | 7.441 / 0.323 | 6.331 / 0.456 | 5.668 / 0.293 | 6.508 / 0.431 | 3.510 / 0.577 |
| TopNet [Tchapmi *et al.*, 2019] | 4.976 / 0.467 | 3.710 / 0.593 | 5.629 / 0.358 | 4.530 / 0.405 | 6.391 / 0.388 | 5.547 / 0.491 | 5.281 / 0.361 | 5.381 / 0.528 | 3.350 / 0.615 |
| PF-Net [Huang *et al.*, 2020] | 3.873 / 0.551 | 2.515 / 0.551 | 4.453 / 0.399 | 3.602 / 0.453 | 4.478 / 0.489 | 5.185 / 0.559 | 4.113 / 0.409 | 3.838 / 0.614 | 2.871 / 0.656 |
| MSN [Liu *et al.*, 2020] | 3.793 / 0.578 | 2.038 / 0.798 | 5.060 / 0.378 | 4.322 / 0.380 | 4.135 / 0.562 | 4.247 / 0.652 | 4.183 / 0.410 | 3.976 / 0.615 | 2.379 / 0.708 |
| GRNet [Xie *et al.*, 2020] | 3.171 / 0.601 | 1.916 / 0.767 | 4.468 / 0.426 | 3.915 / 0.446 | 3.402 / 0.575 | 3.034 / 0.694 | 3.872 / 0.450 | 3.071 / 0.639 | 2.160 / 0.704 |
| PoinTr [Yu *et al.*, 2021] | 2.851 / 0.683 | 1.686 / 0.842 | 4.001 / 0.516 | 3.203 / 0.545 | 3.111 / 0.662 | 2.928 / 0.742 | 3.507 / 0.547 | 2.845 / 0.723 | 1.737 / 0.780 |
| SDT [Zhang *et al.*, 2023a] | 4.246 / 0.473 | 3.166 / 0.636 | 4.807 / 0.291 | 3.607 / 0.363 | 5.056 / 0.398 | 6.101 / 0.442 | 4.525 / 0.307 | 3.995 / 0.574 | 2.856 / 0.602 |
| SeedFormer [Zhou *et al.*, 2022] | 2.902 / 0.688 | 1.716 / 0.835 | 4.049 / 0.551 | 3.392 / 0.544 | 3.151 / 0.668 | 3.226 / 0.777 | 3.603 / 0.555 | 2.803 / 0.716 | 1.679 / 0.786 |
| PointAttN [Wang *et al.*, 2024] | 2.853 / 0.662 | 1.613 / 0.841 | 3.969 / 0.483 | 3.257 / 0.515 | 3.157 / 0.638 | 3.058 / 0.729 | 3.406 / 0.512 | 2.787 / 0.699 | 1.872 / 0.774 |
| Multi-modal Methods | | | | | | | | | |
| ViPC [Zhang *et al.*, 2021a] | 3.308 / 0.591 | 1.760 / 0.803 | 4.558 / 0.451 | 3.183 / 0.512 | 2.476 / 0.529 | 2.867 / 0.706 | 4.481 / 0.434 | 4.990 / 0.594 | 2.197 / 0.730 |
| CSDN [Zhu *et al.*, 2024] | 2.570 / 0.695 | 1.251 / 0.862 | 3.670 / 0.548 | 2.977 / 0.560 | 2.835 / 0.669 | 2.554 / 0.761 | 3.240 / 0.557 | 2.575 / 0.729 | 1.742 / 0.782 |
| CDPNet [Du *et al.*, 2024a] | 1.706 / 0.758 | 0.764 / 0.934 | 2.755 / 0.587 | 2.141 / 0.638 | 1.769 / 0.752 | 1.213 / 0.850 | 2.231 / 0.641 | 1.675 / 0.789 | 1.102 / 0.869 |
| XMFNet [Aiello *et al.*, 2022] | 1.443 / 0.796 | 0.572 / 0.961 | 1.980 / 0.662 | 1.754 / 0.691 | 1.403 / 0.809 | 1.810 / 0.792 | 1.702 / 0.723 | 1.386 / 0.830 | 0.945 / 0.901 |
| EGIINet [Xu *et al.*, 2024] | 1.211 / 0.836 | 0.534 / 0.969 | 1.921 / 0.693 | 1.655 / 0.723 | 1.204 / 0.847 | 0.776 / 0.919 | 1.552 / 0.756 | 1.227 / 0.857 | 0.802 / 0.927 |
| IAET (Ours) | **1.090 / 0.860** | **0.503 / 0.980** | **1.397 / 0.782** | **1.648 / 0.725** | **1.196 / 0.850** | **0.668 / 0.934** | **1.401 / 0.795** | **1.200 / 0.867** | **0.704 / 0.950** |

Table 1: Comparison of CD and F-Score under known categories on ShapeNet-ViPC.

### 3.4 Loss

Consistent with previous methods, IAET uses chamfer distance (CD) [Fan *et al.*, 2017] as the loss function, and its formula is as

$$\begin{aligned}\mathcal{L}_{\mathrm{CD}}(P,\hat{P}) = &\frac{1}{|P|}\sum_{p\in P}\min_{\hat{p}\in\hat{P}}\|p-\hat{p}\|_2^2 \\ &+\frac{1}{|\hat{P}|}\sum_{\hat{p}\in\hat{P}}\min_{p\in P}\|\hat{p}-p\|_2^2,\end{aligned} \tag{12}$$

where $P$ and $\hat{P}$ are the completion result and ground truth, respectively. CD is commonly used to quantify the similarity between two point clouds. Since IAET employs a coarse-to-fine decoding strategy, it generates completion results at multiple resolutions, namely $P_0$, $P_1$, $P_2$, and $P_3$. The ground truth corresponding to each resolution is utilized as the final loss function to train the model and specific formulation is as

$$\mathcal{L} = \sum_{i=0}^{3}\mathcal{L}_{\mathrm{CD}}(P_i,\hat{P}_i), \tag{13}$$

where $\hat{P}_i(i=0,1,2,3)$ denotes the ground truth corresponding to the resolution of completion results $P_i(i=0,1,2,3)$, which are generated using the farthest point sampling [Qi *et al.*, 2017b] strategy.

## 4 Experiments and Analyses

### 4.1 Datasets

**ShapeNet-ViPC.** ShapeNet-ViPC [Zhang *et al.*, 2021a] is a multi-modal dataset containing images and point clouds. It contains 13 categories and a total of 38,328 objects. Each object contains 24 views, so ShapeNet-ViPC contains 919,872 samples in total. Each sample consists of a partial point cloud, a complete point cloud, and an image sizing $138 \times 138$. Same as with previous methods [Zhang *et al.*, 2021a; Zhu *et al.*, 2024; Aiello *et al.*, 2022; Xu *et al.*, 2024], we conduct known category experiments on 8 categories and generalization experiments on the remaining categories.

**KITTI.** KITTI [Geiger *et al.*, 2012] is a real-world dataset. KITTI used in this experiment was created by Wu et al. [Wu *et al.*, 2025] Specifically, it contains 156 different car samples, each of which contains an incomplete point cloud and a complete image sizing $224 \times 224$. Since KITTI does not contain real labels, we train each method on the car category of a multi-modal point cloud benchmark [Wu *et al.*, 2025] and test them on KITTI.

### 4.2 Experimental Settings

**Metrics.** Consistent with all previous multi-modal methods [Zhang *et al.*, 2021a; Zhu *et al.*, 2024; Aiello *et al.*, 2022; Du *et al.*, 2024a; Xu *et al.*, 2024], we use CD and F-Score as evaluation metrics. CD is the same as formula (12). F-Score is a harmonic mean of precision and recall. Precision measures the number of points in the predicted point cloud that are close to the ground-truth point cloud, while recall quantifies the number of points in the ground-truth point cloud that are covered by the predicted point cloud. The formula of F-Score is as

$$\begin{cases}\mathrm{F\text{-}Score} = 2\cdot\dfrac{\mathrm{Precision}\cdot\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}, \\ \mathrm{Precision} = \dfrac{1}{|P|}\displaystyle\sum_{p\in P} I\left(\min_{\hat{p}\in\hat{P}}\|p-\hat{p}\|_2^2 \le \epsilon\right), \\ \mathrm{Recall} = \dfrac{1}{|\hat{P}|}\displaystyle\sum_{\hat{p}\in\hat{P}} I\left(\min_{p\in P}\|\hat{p}-p\|_2^2 \le \epsilon\right),\end{cases} \tag{14}$$

where $\epsilon > 0$ is a threshold, $P$ and $\hat{P}$ are predicted point cloud and ground-truth point cloud, and $I(\cdot)$ is the indicator function. Consistent with previous methods [Zhang *et al.*, 2021a; Zhu *et al.*, 2024; Aiello *et al.*, 2022; Du *et al.*, 2024a; Xu *et al.*, 2024], we set $\epsilon = 0.01$.

**Implementation Details.** All experiments are performed on a NVIDIA RTX A6000. We use the Adam [Kingma and Ba, 2015] optimizer with an initial learning rate of 0.001. The learning rate decayed every 20 epochs with a decay rate of 0.7. Our method converges after 200 epochs with a batch size of 64.

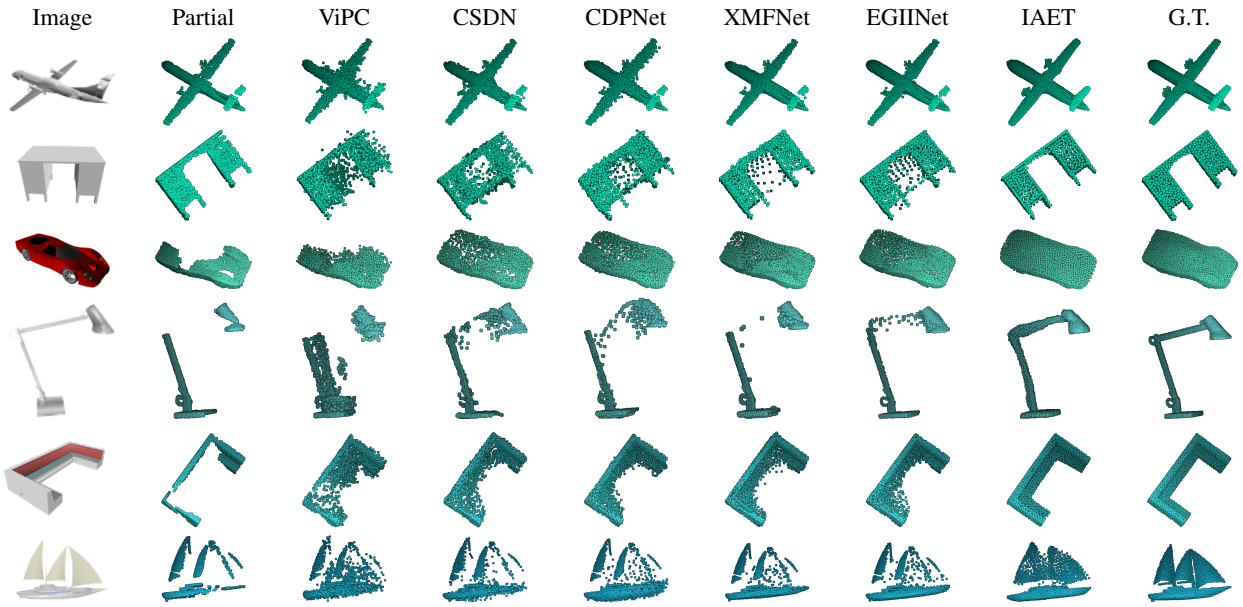


Figure 5: Comparison of visualization under known categories on ShapeNet-ViPC.

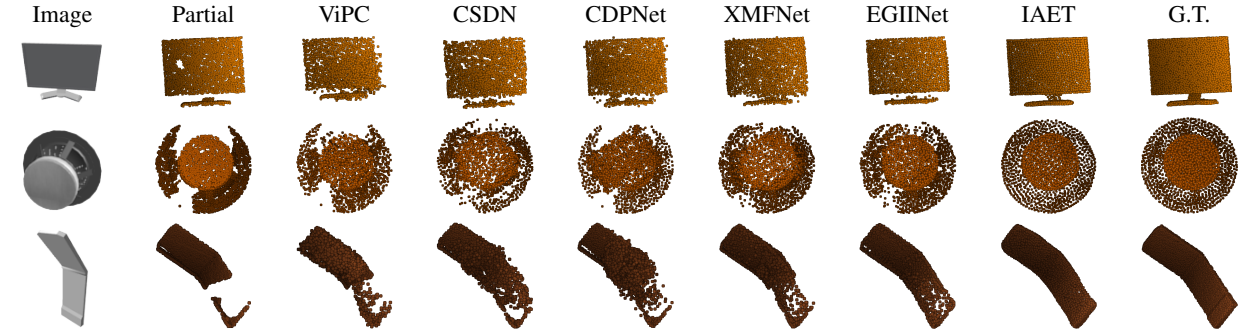


Figure 6: Comparison of visualization under unknown categories on ShapeNet-ViPC.

### 4.3 Results on ShapeNet-ViPC

**Results on Known Categories.** In eight known categories, we compare the proposed IAET with all multi-modal point cloud completion methods, such as ViPC [Zhang *et al.*, 2021a], CSDN [Zhu *et al.*, 2024], XMFNet [Aiello *et al.*, 2022], CDPNet [Du *et al.*, 2024a], and EGIINet [Xu *et al.*, 2024], and some representative single-modal point cloud completion methods, such as AtlasNet [Groueix *et al.*, 2018], FoldingNet [Yang *et al.*, 2018], PCN [Yuan *et al.*, 2018], TopNet [Tchapmi *et al.*, 2019], PF-Net [Tchapmi *et al.*, 2019], MSN [Liu *et al.*, 2020], GRNet [Xie *et al.*, 2020], PoinTr [Yu *et al.*, 2021], SDT [Zhang *et al.*, 2023a], SeedFormer [Zhou *et al.*, 2022], and PointAttN [Wang *et al.*, 2024]. Table 1 presents the quantitative experimental results. Our method has the best result in all categories. Figure 5 presents the qualitative comparison of the experimental results. The first, second, third, fourth, fifth, and sixth rows correspond to the airplane, cabinet, car, lamp, sofa, and watercraft categories, respectively. In the airplane, cabinet, and sofa categories, only our method generates objects with smooth surfaces, whereas the completion results from other methods exhibit outlier points. In the car category, our method is the only one capable of generating smooth surfaces. In the lamp example, only our method can generate the missing lamp arm, while other methods seem unable to recover this shape. A similar observation can be made in the watercraft category. Despite the challenges in generating irregular shapes such as sails, only our method successfully reconstructs portions of the sail structures.

### 4.4 Generalization Ability Evaluation

**Results on Unknown Categories of ShapeNet-ViPC.** To assess the generalization capability of the proposed method on unseen categories, we follow the experimental setup of previous studies [Zhu *et al.*, 2024; Xu *et al.*, 2024], training on eight known categories and testing on four unknown categories. Specifically, we compare the performance of IAET against all existing multi-modal point cloud completion meth-

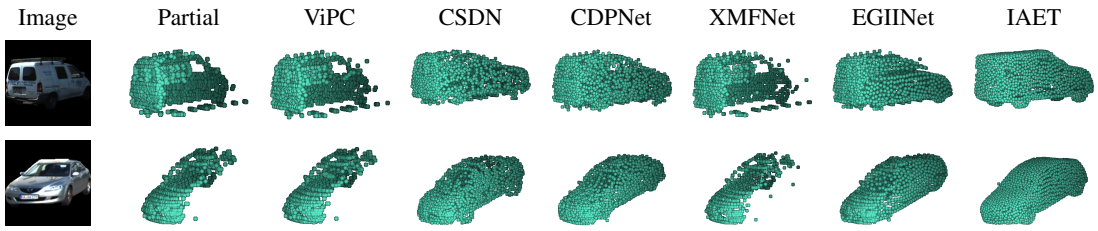


Figure 7: Comparison of visualization on KITTI.

| Methods | CD ↓ | | | | |
|---|---|---|---|---|---|
| | Avg | Bench | Monitor | Speaker | Cellphone |
| Single-modal Methods | | | | | |
| PF-Net [Huang *et al.*, 2020] | 5.011 | 3.684 | 5.304 | 7.663 | 3.392 |
| MSN [Liu *et al.*, 2020] | 4.684 | 2.613 | 4.818 | 8.259 | 3.047 |
| GRNet [Xie *et al.*, 2020] | 4.096 | 2.367 | 4.102 | 6.493 | 3.422 |
| PoinTr [Yu *et al.*, 2021] | 3.755 | 1.976 | 4.084 | 5.913 | 3.049 |
| SDT [Zhang *et al.*, 2023a] | 6.001 | 4.096 | 6.222 | 9.499 | 4.189 |
| SeedFormer [Zhou *et al.*, 2022] | 5.215 | 3.228 | 4.464 | 8.520 | 4.646 |
| PointAttN [Wang *et al.*, 2024] | 3.674 | 2.135 | 3.741 | 5.973 | 2.848 |
| Multi-modal Methods | | | | | |
| ViPC [Zhang *et al.*, 2021a] | 4.601 | 3.091 | 4.419 | 7.674 | 3.219 |
| CSDN [Zhu *et al.*, 2024] | 3.656 | 1.834 | 4.115 | 5.690 | 2.985 |
| CDPNet [Du *et al.*, 2024a] | 4.462 | 3.122 | 4.100 | 7.611 | 3.013 |
| XMFNet [Aiello *et al.*, 2022] | 2.671 | 1.278 | 2.806 | 4.823 | 1.779 |
| EGIINet [Xu *et al.*, 2024] | 2.354 | 1.047 | 2.513 | 4.282 | 1.575 |
| IAET (Ours) | **1.573** | **1.030** | **1.497** | **2.682** | **1.084** |

Table 2: Comparison of CD under unknown categories on ShapeNet-ViPC.

| Methods | F-Score ↑ | | | | |
|---|---|---|---|---|---|
| | Avg | Bench | Monitor | Speaker | Cellphone |
| Single-modal Methods | | | | | |
| PF-Net [Huang *et al.*, 2020] | 0.468 | 0.584 | 0.433 | 0.319 | 0.534 |
| MSN [Liu *et al.*, 2020] | 0.533 | 0.706 | 0.527 | 0.291 | 0.607 |
| GRNet [Xie *et al.*, 2020] | 0.548 | 0.711 | 0.537 | 0.376 | 0.569 |
| PoinTr [Yu *et al.*, 2021] | 0.619 | 0.797 | 0.599 | 0.454 | 0.627 |
| SDT [Zhang *et al.*, 2023a] | 0.327 | 0.479 | 0.268 | 0.197 | 0.362 |
| SeedFormer [Zhou *et al.*, 2022] | 0.590 | 0.736 | 0.598 | 0.410 | 0.615 |
| PointAttN [Wang *et al.*, 2024] | 0.605 | 0.764 | 0.591 | 0.428 | 0.637 |
| Multi-modal Methods | | | | | |
| ViPC [Zhang *et al.*, 2021a] | 0.498 | 0.654 | 0.491 | 0.313 | 0.535 |
| CSDN [Zhu *et al.*, 2024] | 0.631 | 0.798 | 0.598 | 0.485 | 0.644 |
| CDPNet [Du *et al.*, 2024a] | 0.589 | 0.714 | 0.593 | 0.418 | 0.629 |
| XMFNet [Aiello *et al.*, 2022] | 0.710 | 0.862 | 0.677 | 0.556 | 0.748 |
| EGIINet [Xu *et al.*, 2024] | 0.750 | 0.902 | 0.716 | 0.591 | 0.792 |
| IAET (Ours) | **0.807** | **0.903** | **0.810** | **0.615** | **0.899** |

Table 3: Comparison of F-Score under unknown categories on ShapeNet-ViPC.

ods, such as ViPC [Zhang *et al.*, 2021a], CSDN [Zhu *et al.*, 2024], XMFNet [Aiello *et al.*, 2022], CDPNet [Du *et al.*, 2024a], and EGIINet [Xu *et al.*, 2024], and some representative single-modal point cloud completion methods, such as PF-Net [Tchapmi *et al.*, 2019], MSN [Liu *et al.*, 2020], GRNet [Xie *et al.*, 2020], PoinTr [Yu *et al.*, 2021], SDT [Zhang *et al.*, 2023a], SeedFormer [Zhou *et al.*, 2022], and PointAttN [Wang *et al.*, 2024]. Tables 2 and 3 present the quantitative results of the experiments. Our method achieves the best performance across four unknown categories, demonstrating its superior generalization ability. Figure 6 presents a qualitative comparison of the experimental results, with the first, second, and third rows corresponding to the monitor, speaker, and cellphone categories, respectively. In the monitor example, IAET generates a more compact screen along with detailed features of the bracket. For the speaker sample, only IAET is able to produce a circular global shape. Additionally, IAET successfully generates a complete cell phone shape in the cellphone sample.

**Results on Real-Scene KITTI.** To assess the effectiveness of IAET in real-world scenarios, we compare the qualitative results of all multi-modal point cloud completion methods, such as ViPC [Zhang *et al.*, 2021a], CSDN [Zhu *et al.*, 2024], XMFNet [Aiello *et al.*, 2022], CDPNet [Du *et al.*, 2024a], and EGIINet [Xu *et al.*, 2024], on the KITTI dataset. As shown in Figure 7, each row corresponds to a real car sample. Only IAET generates completion results that closely resemble the vehicle shape depicted in the image.

| Model | CD ↓ | F-Score ↑ |
|---|---|---|
| w/o interleaved attention mechanism | 1.213 | 0.837 |
| w/o view-guided upsampling module | 1.251 | 0.832 |
| IAET (Ours) | **1.090** | **0.860** |

Table 4: Ablation study on IAET.

### 4.5 Ablation Study

This section evaluates the effectiveness of IAET's core components from a quantitative perspective. As shown in Table 4, the first row corresponds to IAET without interleaved attention mechanisms, while the second row represents IAET without view-guided upsampling modules. The experimental results indicate that both components are critical to the overall performance of IAET.

## 5 Conclusion

We propose IAET, a novel framework for multi-modal point cloud completion. To fully leverage image information and enhance its contribution to point cloud completion, we propose an interleaved attention mechanism to facilitate the mutual supplement of point cloud and image features. Furthermore, we propose a view-guided upsampling module to utilize image information effectively, guiding the generation of high-quality completion results. Extensive qualitative and quantitative experiments demonstrate that IAET achieves state-of-the-art performance on multi-modal point cloud completion benchmarks.

In future work, we aim to enhance the performance of multi-modal point cloud completion with more modalities and large-scale model technologies.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.U21A20473, 62032022, 62376141, 62176244) and the Key Technologies Program of Taihang Laboratory in Shanxi Province, China (THYF-JSZX-24010600).

## References

[Aiello *et al.*, 2022] Emanuele Aiello, Diego Valsesia, and Enrico Magli. Cross-modal learning for image-guided point cloud shape completion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 37349–37362, New Orleans, USA, November 2022.

[Alliegro *et al.*, 2021] Antonio Alliegro, Diego Valsesia, Giulia Fracastoro, Enrico Magli, and Tatiana Tommasi. Denoise and contrast for category agnostic shape completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4629–4638, Virtual, June 2021.

[Chen *et al.*, 2024] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(12):10164–10183, December 2024.

[Cheng *et al.*, 2022] Yuwei Cheng, Jingran Su, Mengxin Jiang, and Yimin Liu. A novel radar point cloud generation method for robot environment perception. *IEEE Transactions on Robotics (TRO)*, 38(6):3754–3773, December 2022.

[Du *et al.*, 2024a] Zhenjiang Du, Jiale Dou, Zhitao Liu, Jiwei Wei, Guan Wang, Ning Xie, and Yang Yang. CDPNet: Cross-modal dual phases network for point cloud completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 1635–1643, Vancouver, Canada, February 2024.

[Du *et al.*, 2024b] Zijin Du, Jianqing Liang, Jiye Liang, Kaixuan Yao, and Feilong Cao. Graph regulation network for point cloud segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(12):7940–7955, December 2024.

[Du *et al.*, 2024c] Zijin Du, Hailiang Ye, and Feilong Cao. A novel local-global graph convolutional method for point cloud semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 35(4):4798–4812, April 2024.

[Fan *et al.*, 2017] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3D object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 605–613, Honolulu, USA, July 2017.

[Fei *et al.*, 2022] Ben Fei, Weidong Yang, Wen-Ming Chen, Zhijun Li, Yikang Li, Tao Ma, Xing Hu, and Lipeng Ma. Comprehensive review of deep learning-based 3D point cloud completion processing and analysis. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 23(12):22862–22883, December 2022.

[Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, Providence, USA, June 2012.

[Groueix *et al.*, 2018] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3D surface generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 216–224, Salt Lake City, USA, June 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, USA, June 2016.

[Huang *et al.*, 2020] Zitian Huang, Yikuan Yu, Jiawen Xu, Feng Ni, and Xinyi Le. PF-Net: Point fractal network for 3D point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7662–7670, Virtual, June 2020.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diego, USA, May 2015.

[Liang *et al.*, 2023] Jiye Liang, Zijin Du, Jianqing Liang, Kaixuan Yao, and Feilong Cao. Long and short-range dependency graph structure learning framework on point cloud. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(12):14975–14989, December 2023.

[Liu *et al.*, 2020] Minghua Liu, Lu Sheng, Sheng Yang, Jing Shao, and Shi-Min Hu. Morphing and sampling network for dense point cloud completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 11596–11603, New York, USA, February 2020.

[Liu *et al.*, 2023] Yifan Liu, Wuyang Li, Jie Liu, Hui Chen, and Yixuan Yuan. GRAB-Net: Graph-based boundary-aware network for medical point cloud segmentation. *IEEE Transactions on Medical Imaging (TMI)*, 42(9):2776–2786, September 2023.

[Mu *et al.*, 2024] Juncheng Mu, Lin Bie, Shaoyi Du, and Yue Gao. ColorPCR: Color point cloud registration with multi-stage geometric-color fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21061–21070, Seattle, USA, June 2024.

[Pei *et al.*, 2023] Yu Pei, Xian Zhao, Hao Li, Jingyuan Ma, Jingwei Zhang, and Shiliang Pu. Clusterformer: Cluster-based Transformer for 3D object detection in point clouds.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6664–6673, Paris, France, October 2023.

[Qi *et al.*, 2017a] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, Honolulu, USA, July 2017.

[Qi *et al.*, 2017b] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5105–5114, Long Beach, USA, December 2017.

[Tchapmi *et al.*, 2019] Lyne P Tchapmi, Vineet Kosaraju, Hamid Rezatofighi, Ian Reid, and Silvio Savarese. TopNet: Structural point cloud decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 383–392, Long Beach, USA, June 2019.

[Tesema *et al.*, 2024] Keneni W. Tesema, Lyndon Hill, Mark W. Jones, Muneeb I. Ahmad, and Gary K.L. Tam. Point cloud completion: A survey. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 30(10):6880–6899, October 2024.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 6000–6010, Long Beach, USA, December 2017.

[Wang *et al.*, 2024] Jun Wang, Ying Cui, Dongyan Guo, Junxia Li, Qingshan Liu, and Chunhua Shen. PointAttN: You only need attention for point cloud completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 5472–5480, Vancouver, Canada, February 2024.

[Wu *et al.*, 2025] Lintai Wu, Qijian Zhang, Junhui Hou, and Yong Xu. Leveraging single-view images for unsupervised 3D point cloud completion. *IEEE Transactions on Multimedia (TMM)*, 27:940–953, February 2025.

[Xie *et al.*, 2020] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Jiageng Mao, Shengping Zhang, and Wenxiu Sun. GRNet: Gridding residual network for dense point cloud completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 365–381, Glasgow, Scotland, August 2020.

[Xu *et al.*, 2024] Hang Xu, Chen Long, Wenxiao Zhang, Yuan Liu, Zhen Cao, Zhen Dong, and Bisheng Yang. Explicitly guided information interaction network for cross-modal point cloud completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, page 414–432, Milan, Italy, September 2024.

[Yang *et al.*, 2018] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. FoldingNet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 206–215, Salt Lake City, USA, June 2018.

[Yu *et al.*, 2021] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. PoinTr: Diverse point cloud completion with geometry-aware Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12478–12487, Montreal, Canada, October 2021.

[Yuan *et al.*, 2018] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. PCN: Point completion network. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 728–737, Verona, Italy, September 2018.

[Zhang *et al.*, 2021a] Xuancheng Zhang, Yutong Feng, Siqi Li, Changqing Zou, Hai Wan, Xibin Zhao, Yandong Guo, and Yue Gao. View-guided point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15885–15894, Virtual, June 2021.

[Zhang *et al.*, 2021b] Yanan Zhang, Di Huang, and Yunhong Wang. PC-RGNN: Point cloud completion and graph neural network for 3D object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 3430–3437, Virtual, February 2021.

[Zhang *et al.*, 2022] Bowen Zhang, Xi Zhao, He Wang, and Ruizhen Hu. Shape completion with points in the shadow. In *SIGGRAPH Asia 2022 Conference Papers (SIGGRAPH)*, page 28, Daegu, Korea, December 2022.

[Zhang *et al.*, 2023a] Wenxiao Zhang, Huajian Zhou, Zhen Dong, Jun Liu, Qingan Yan, and Chunxia Xiao. Point cloud completion via skeleton-detail Transformer. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 29(10):4229–4242, October 2023.

[Zhang *et al.*, 2023b] Xiyu Zhang, Jiaqi Yang, Shikun Zhang, and Yanning Zhang. 3D registration with maximal cliques. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17745–17754, Vancouver, Canada, June 2023.

[Zhou *et al.*, 2022] Haoran Zhou, Yun Cao, Wenqing Chu, Junwei Zhu, Tong Lu, Ying Tai, and Chengjie Wang. SeedFormer: Patch seeds based point cloud completion with upsample Transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 416–432, Tel Aviv, Israel, October 2022.

[Zhu *et al.*, 2024] Zhe Zhu, Liangliang Nan, Haoran Xie, Honghua Chen, Jun Wang, Mingqiang Wei, and Jing Qin. CSDN: Cross-modal shape-transfer dual-refinement network for point cloud completion. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 30(7):3545–3563, July 2024.