DMF-Net: Image-Guided Point Cloud Completion with Dual-Channel Modality Fusion and Shape-Aware Upsampling Transformer

Aihua Mao[1*†], **Yuxuan Tang**[1†], **Jiangtao Huang**[1], **Ying He**[2]

[1]School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
[2]School of Computer Science and Engineering, Nanyang Technological University
ahmao@scut.edu.cn, 202120144052, cs_jetthwang@mail.scut.edu.cn, yhe@ntu.edu.sg

Abstract

In this paper we study the task of a single-view image-guided point cloud completion. Existing methods have got promising results by fusing the information of image into point cloud explicitly or implicitly. However, given that the image has global shape information and the partial point cloud has rich local details, We believe that both modalities need to be given equal attention when performing modality fusion. To this end, we propose a novel dual-channel modality fusion network for image-guided point cloud completion(named DMF-Net), in a coarse-to-fine manner. In the first stage, DMF-Net takes a partial point cloud and corresponding image as input to recover a coarse point cloud. In the second stage, the coarse point cloud will be upsampled twice with shape-aware upsampling transformer to get the dense and complete point cloud. Extensive quantitative and qualitative experimental results show that DMF-Net outperforms the state-of-the-art unimodal and multimodal point cloud completion works on ShapeNet-ViPC dataset.

Introduction

With the development of 3D sensing equipment such as LiDARs, laser scanners and RGB-D cameras, it is easier to acquire point clouds data, which has abundant applications like autonomous driving (Cui et al. 2021), scene understanding (Hou, Dai, and Nießner 2019), robotic vision (Varley et al. 2017) and augmented reality (Park, Lepetit, and Woo 2008). However, point clouds captured by these devices are often incomplete and sparse due to self-occlusions, occlusions between objects, uneven illumination and low scanning resolution. The poor quality point clouds can make it difficult to understand 3D shape, which will leads to ambiguity in many downstream tasks like point cloud detection (Zhou and Tuzel 2018; Qi et al. 2019), point cloud reconstruction (Yang et al. 2018; Mandikal and Radhakrishnan 2019) and point cloud upsampling (Yu et al. 2018; Mao et al. 2022). Therefore, point cloud completion, which focus on recovering complete high quality point clouds from sparse partial input, has become a hot research topic.

Inspired by the seminal work (Qi et al. 2017a) employing shared-MLP for point-wise feature extraction, an increasing number of learning-based approaches for point cloud completion have emerged. Most of these learning-based works utilize an encoder-decoder architecture, wherein the encoder extracts a global latent vector from the partial input, and the decoder subsequently generates the complete point cloud based on this global latent representation. However, the global information from an incomplete point cloud can be ambiguous and misleading.

Compared to 3D point cloud data, it is obvious that the corresponding 2D image data is easier to acquire (Nan, Sharf, and Chen 2014). Moreover, we know that humans can easily infer the 3D shape of an object from its 2D image, which intuitively proves that 2D image can play an important role in 3D shape completion task. Specifically, an image has the global shape information of an object, while the partial point cloud has rich local details. The two complementary information mentioned before can make the extracted feature more representative so that the decoder can generate a more plausible shape. Previous multimodal completion networks (Zhang et al. 2021; Zhu et al. 2023; Aiello, Valsesia, and Magli 2022) have successfully improved the performance by introducing a single-view image into the point cloud completion process. However, the modality fusion process of these methods is dominated by point cloud modality, lacking reasonable utilization of image modality.

To address the above issue, we design an image-guided point cloud completion network with dual-channel modality fusion and shape-aware upsampling transformer (named DMF-Net). DMF-Net recovers the partial point cloud in two stages. At the first stage, DMF-Net employs an encoder-decoder architecture to generate a sparse but complete point cloud. We observe that image modality should not just serve as a guidance for point cloud modality, but should complement point cloud modality, which indicates the two modalities should be considerd equally important when performing modality fusion. Thus, in the encoding phase, we propose a dual-channel modality fusion strategy, which is capable of capturing the complementary information in both image and point cloud by fusing the two modalities in a symmetric way. At the second stage, inspired by these works (Xiang et al. 2021; Zhou et al. 2022; Wang et al. 2024), to further recover the local details and improve the uniformity of the complete point cloud, based on the coarse point cloud generated by the first stage, DMF-Net utilizes a shape-aware

*Corresponding author.
†These authors contributed equally.

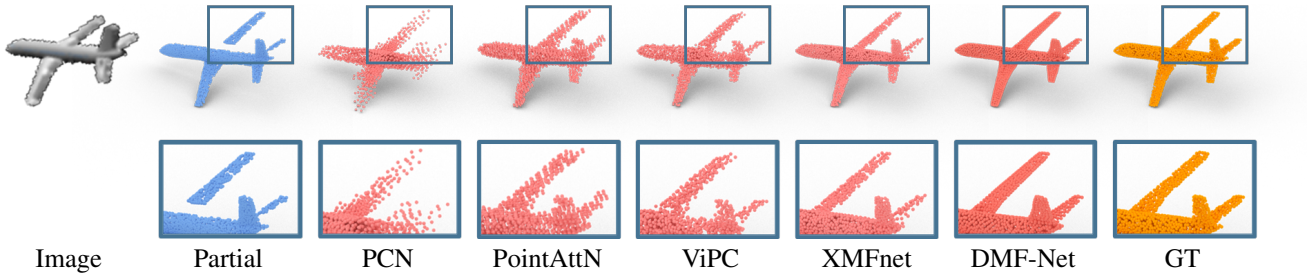

Figure 1: The visualization of point cloud completion results produced by PCN (Yuan et al. 2018), PointAttN (Wang et al. 2024), ViPC (Zhang et al. 2021), XMFnet (Aiello, Valsesia, and Magli 2022) and DMF-Net. Given a partial point cloud and a corresponding single-view image, our DMF-Net reconstructs a high-quality complete point cloud with more local details and better uniformity.

upsampler to recover a dense and complete point cloud. The term shape-aware means during the upsampling process, the local feature context is utilized for better recovering the local geometric details.

In summary, the main contributions of this work are threefold:

- We propose a novel point cloud completion network with dual-channel modality fusion and shape-aware upsampling transformer, which utilizes the complementary information in image and partial point cloud to recover a high quality complete point cloud in a coarse-to-fine manner.
- We propose a dual-channel modality fusion module, which fuses image modality and point cloud modality in a symmetric way, allowing the two modalities contribute equally to the fused global feature.
- We propose a novel shape-aware upsampling transformer for point cloud completion, which is capable of capturing the local geometric details through encouraging the communication of local neighborhood points with self-attention mechanism.

Related Work

Traditional Point Cloud Completion

Traditional methods for point cloud completion can be roughly categorize into two types: geometry-based and alignment-based. Geometry-based methods usually utilize the visible part to predict the missing part by some prior geometric assumptions. Methods like (Mitra, Guibas, and Pauly 2006; Mitra et al. 2013; Sung et al. 2015) employ symmetric axes to repeat the regular part of objects from the observed regions. Other methods such as (Davis et al. 2002; Berger et al. 2014; Nguyen et al. 2016) fill holes of a smooth surface by local interpolations. These works can only handle partial scans with a small degree of incompleteness like small holes on the surface. Alignment-based methods involve matching partial shapes with template models from a large database. Some methods (Pauly et al. 2005; Shao et al. 2012; Li et al. 2015) retrieve the whole 3D shapes directly. Other methods (Kalogerakis et al. 2012; Shen et al. 2012; Kim et al. 2013) retrieve part of the 3D shapes to do the completion task. These methods can handle variant incompleteness of point cloud, but they are not suitable for practical application scenarios due to the expensive cost of inference optimization and building large datasets, as well as their high sensitivity to noise.

Unimodal Point Cloud Completion

In recent years, deep learning-based methods have become the mainstream direction of point cloud completion. Early learning-based methods (Dai, Ruizhongtai Qi, and Nießner 2017; Han et al. 2017) apply 3D CNN to voxel-based representations of 3D objects, which are computational expensive. Thanks to PointNet (Qi et al. 2017a) and PointNet++ (Qi et al. 2017b), most methods utilize an encoder-decoder architecture, where the encoder part extracts a global latent vector of the partial input and the decoder part reconstructs the complete point cloud through the global vector. PCN (Yuan et al. 2018) is the pioneering work which first avoids any geometric priors and annotations about the underlying shape. In the decoding phase, PCN employs the folding operation proposed by FoldingNet (Yang et al. 2018) to achieve point cloud refinement and upsampling. AtlasNet (Groueix et al. 2018) represents a 3D shape as a collection of parametric surface elements and tackles the point cloud completion by learning the transformation of 2D square to 3D surface. TopNet (Tchapmi et al. 2019) proposes a hierarchical rooted tree structure decoder which can generate structured point clouds iteratively. MSN (Liu et al. 2020) introduces a coarse-to-fine method which generates different parts of a shape and refines the coarse point cloud with a residual network. GRNet (Xie et al. 2020) proposes a technique to convert the point cloud to 3D grid so that 3D CNN can be applied. PF-Net (Huang et al. 2020) proposes a multi-scale encoder to extract both global and local features and a point pyramid decoder to generate points in missing regions hierarchically. Inspired by GAN (Goodfellow et al. 2014), adversarial loss is utilized in the training process to get better completion performance. CRN (Wang, Ang Jr, and Lee 2020) proposes a cascaded refinement network to iteratively upsample and refine the coarse point cloud. ECG (Pan 2020) utilizes an encoder-decoder architecture to first restore the skeleton of a complete point cloud, then constructs a graph and applies convolution for edge perception, while expand-

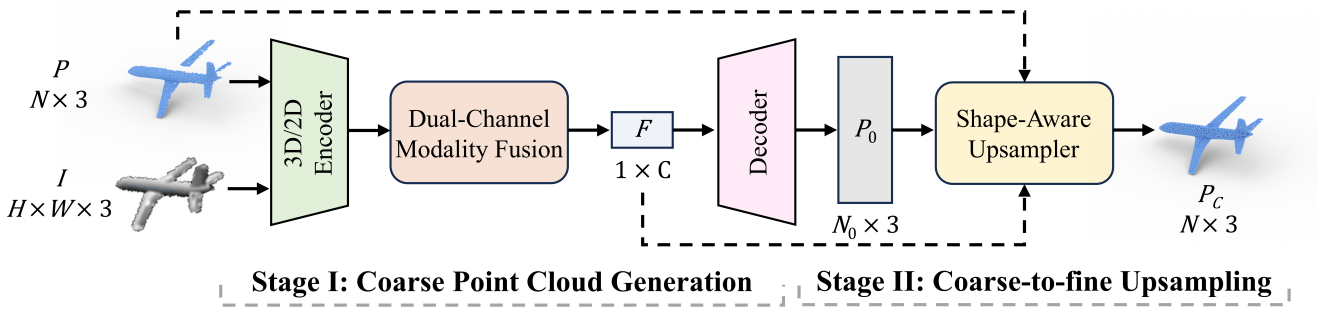

Figure 2: The architecture of our proposed DMF-Net. It takes two stages to recover the complete point cloud. In the first stage, a coarse point cloud is generated according to the partial point cloud and corresponding single-view image. In the second stage, the coarse point cloud will be upsampled to get the dense and complete output.

ing the point cloud features to achieve upsampling and refinement. VRC-Net (Pan et al. 2021) reconstructs a coarse point cloud by employing a dual-path probabilistic modeling network and do the refinement by employing a relational enhancement network which can learn multi-scale structural relations between partial input and the coarse point cloud. Works like (Yu et al. 2021; Xiang et al. 2021; Zhou et al. 2022; Wang et al. 2024) have successfully integrated the Transformer (Vaswani et al. 2017) architecture into point cloud completion, further improve the performance of the completion network. PoinTr (Yu et al. 2021) reformulates the point cloud completion task into a set-to-set translation task, and introduces a geometric perception module to better capture the spatial neighborhood information of the point cloud. SnowflakeNet (Xiang et al. 2021) regards the generation of a complete point cloud as a snowflake-like growth process of points in 3D space, and utilizes skip-Transformer to learn the point splitting mode that best adapts to the local area to generate a locally compact and structured complete point cloud. Seedformer (Zhou et al. 2022) introduces local seed shape representation and integrates the spatial and semantic relationships between neighborhood points through Transformer during upsampling. PointAttN (Wang et al. 2024) utilizes attention mechanism to achieve geometric detail perception and feature enhancement, and establish the structural relationship between points to generate a complete point cloud with detailed geometric structure.

However, the above methods only takes partial point cloud as input. When performing feature extraction on the partial input, the outcoming global latent vector actually does not contain the information of the missing part, which means the encoding process suffers from information loss that lowers the performance of the completion network.

Multimodal Point Cloud Completion

To solve the lack of global shape information in the partial point cloud, researchers have introduced image modality into the point cloud completion task, which is pioneered by ViPC (Zhang et al. 2021). Extra modalities can provide complementary information which can be used to generated more plausible 3D shapes, and inspired by this idea, ViPC (Zhang et al. 2021) utilizes a pre-trained neural network to generate a point cloud according to the corresponding image, and then explicitly fuse the two modalities through concatenation between the partial point cloud and the generated point cloud. In the refinement stage, a point-wise offset vector is predict by utilizing the information of the two modalities. XMFnet (Aiello, Valsesia, and Magli 2022), Unlike ViPC, performs feature-level fusion to implicitly fuse the two modalities through cross-attention and self-attention, and employs multiple independent branches to predict different parts of the complete point cloud. Inspried by StyleGAN (Karras, Laine, and Aila 2019), CSDN (Zhu et al. 2023) treats the modality fusion task as a style transfer task, and transfer the style of the image to the partial point cloud in the folding-based decoding process. CSDN proposes a dual-refinement module to predict a offset vector by leveraging both the global information in image and local information in partial input. These methods got promising results by utilizing image modality to provide the information of the missing part, however, the fusion process is dominated by the point cloud modality, and the uniformity of the final output is less satisfied.

Method

Overview

The overall framework of our proposed DMF-Net is shown in Fig. 2, in which the cross-modal point cloud completion task consists of two stages: coarse point cloud generation and coarse-to-fine upsampling. Specifically, given the partial point cloud $P \in \mathbb{R}^{N\times 3}$, and the corresponding single-view image $I \in \mathbb{R}^{H\times W\times 3}$, our goal is to generate the complete point cloud $P_C \in \mathbb{R}^{N\times 3}$. N denotes the number of points in P and P_C. H and W denote the height and width of I. In the first stage, we adopt a 3D encoder and a 2D encoder to extract point cloud features and image features respectively. Then, we fuse the feature of the two modalities through the proposed dual-channel modality fusion module, which yields a global vector. Next, the decoder is utilized to generate a coarse point cloud according to the global vector. In the second stage, the coarse point cloud is upsampled by the shape-aware upsampler to get the complete point cloud P_C.

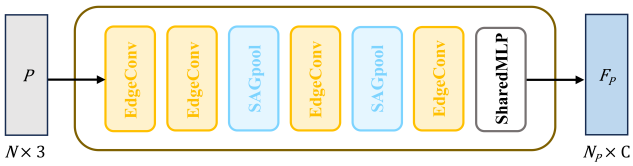

Figure 3: The architecture of 3D encoder, which is employed to extract the feature of partial point cloud.

Point Cloud and Image Encoders

3D Encoder. The partial point cloud P contains rich local details. In order to capture the local geometric information of P, we feed the partial point cloud into the 3D encoder, as shown in Fig. 3. Inspired by (Aiello, Valsesia, and Magli 2022), the architecture of 3D encoder is a sequence of graph-convolutional layers (EdgeConv (Phan et al. 2018)) interleaved by graph-pooling layers (SAGPool (Lee, Lee, and Kang 2019)). The EdgeConv layer employs KNN algorithm to build graph of the partial point cloud and MLPs to extract point-wise feature of the graph. Directly use maxpooling operation on the point-wise feature will inevitably suffer from information loss, so the SAGPool layer is utilized for the purpose of reducing the number of points while preserving more local information. The last layer of the 3D encoder is a shared MLP, which is utilized to yield the point-wise feature $F_P \in \mathbb{R}^{N_P \times C}$. N_P is the number of points and C is the dimensions of the point-wise feature F_P.

2D Encoder. The single-view image I contains the global shape information which serves as a guiding information during the point cloud completion process. The simple but effective convolutional network ResNet18 (He et al. 2016) is chosen as our 2D encoder. The extracted $7 \times 7 \times C$ feature map is reshaped to get the pixel-wise feature $F_I \in \mathbb{R}^{N_I \times C}$, where N_I is the number of pixels and C is the dimensions of F_I.

Dual-channel Modality Fusion

To better leverage the complementary information of the two modalities, we design a Dual-channel Modality Fusion (DMF) module, as shown in Fig. 4. Specifically, after getting the point-wise feature F_P and the pixel-wise feature F_I, we utilize maxpooling operation on F_P to get the point cloud global feature $G_P \in \mathbb{R}^{1 \times C}$ and F_I to get the image global feature $G_I \in \mathbb{R}^{1 \times C}$. The upper path aims to calculate point cloud feature enhanced by image feature $F_{IP} \in \mathbb{R}^{N_I \times C}$, where G_P is replicated for N_I times and concatenate with F_I. Then, the attention matrix $W_{IP} \in \mathbb{R}^{N_I \times N_P}$ is calculated by

$$W_{IP} = softmax\left(\mu\left(F_1\right)\right), \tag{1}$$

where $F_1 = [F_I, replicate\left(G_P, N_I\right)]$ is the enhanced pixel-wise feature and $replicate(\cdot, \cdot)$ denotes the replication operation, $[\cdot, \cdot]$ denotes the concatenation operation and μ is a non-linear function implemented by shared-MLPs. Next, F_{IP} is calulated by

$$F_{IP} = W_{IP} \otimes F_P, \tag{2}$$

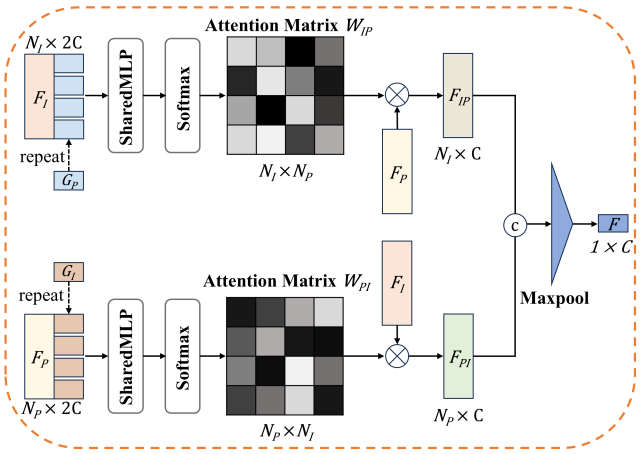

Figure 4: The architecture of Dual-channel Modality Fusion (DMF) module, which fuses the point-wise feature and the pixel-wise feature in a symmetric way.

where $\otimes$ denotes the matrix multiplication operation. Similarly, the lower path aims to calculate image feature enhanced by point cloud feature $F_{PI} \in \mathbb{R}^{N_P \times C}$. Attention matrix $W_{PI} \in \mathbb{R}^{N_P \times N_I}$ is calculated by

$$W_{PI} = softmax\left(\theta\left(F_2\right)\right), \tag{3}$$

where $F_2 = [F_P, replicate\left(G_I, N_P\right)]$ is the enhanced point-wise feature and θ is a non-linear function implemented by shared-MLPs. Then, F_{PI} is calulated by

$$F_{PI} = W_{PI} \otimes F_I, \tag{4}$$

The enhanced features F_{IP} and F_{PI} are concatenated and maxpooling is utilized to yield the fused global feature $F \in \mathbb{R}^{1 \times C}$. DMF module fuses the feature of the two modalities in a symmetric way, and by doing so, the image modality can serve as a better guidance, which can enhance the representational ability of the fused global feature F.

Coarse Point Cloud Generation

To reconstruct a coarse point cloud $P_0 \in \mathbb{R}^{N_0 \times 3}$ from the fused feature F (where N_0 is the number of points in P_0), we utilize the seed generator of (Xiang et al. 2021) as our decoder, as shown in Fig. 5. The decoder first employs 1D transpose convolution to expand the global feature

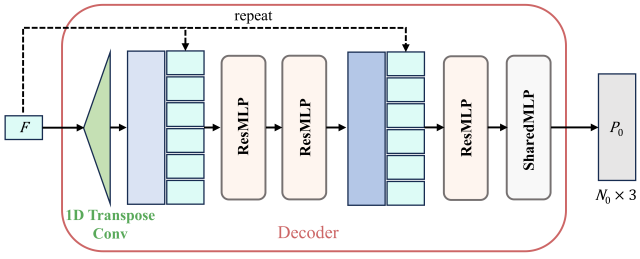

Figure 5: The architecture of decoder, which takes in the fused global feature and reconstructs a coarse point cloud.

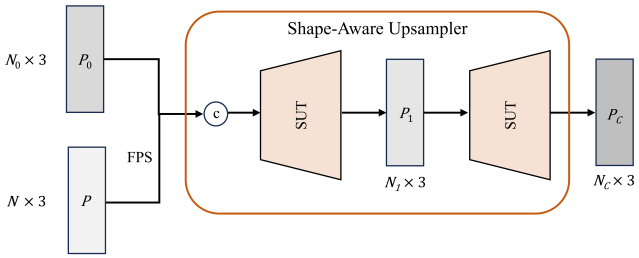

Figure 6: The architecture of upsampler, which consists of two consecutive shape-aware Upsample Transformers (SUTs) to generate the complete point cloud in a coarse-to-fine manner.

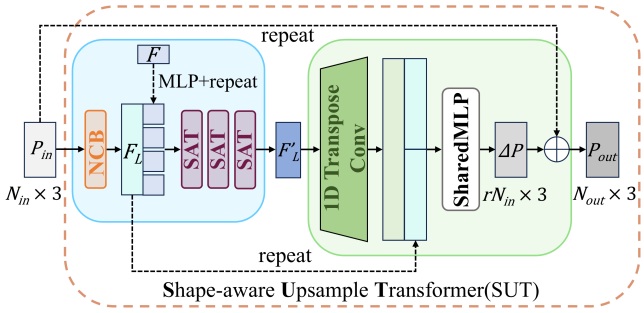

Figure 7: The architecture of SUT, which upsamples the coarse point cloud in two steps: 1) local feature embedding and enhancement, 2) feature expansion and displacement generation.

to a point-wise feature. Then concatenates it with duplicated global feature to preserve more information. Next, two residual shared MLPs are utilized to enhance the point-wise feature. The last layer of the decoder is a shared MLP to adjust the feature dimension to 3, thus outputs the coordinates of the coarse point P_0.

Coarse-to-Fine Upsampling

Existing multimodal completion networks like ViPC (Zhang et al. 2021) and CSDN (Zhu et al. 2023) both have adopted a coarse-to-fine completion strategy. Moreover, in the refinement stage, both of them leverage information of the two modalities to predict a displacement vector and add to the coarse point cloud to generate a point-wise displacement to achieve the refinement goal. However, the generated complete point cloud lacks local details due to the ignorance of the rich neighborhood information. Besides, the uniformity of the complete point cloud is less satisfied since the coarse point cloud and the complete point cloud have same points which means the refinement can only handle some noisy points. The coarse point cloud generated by the first stage is sparse, so we design a shape-aware upsampler to increase the number of points to get the dense and complete point cloud $P_C \in \mathbb{R}^{N_C \times 3}$, as shown in Fig. 6. In order to preserve the local information in partial point cloud P, before upsampling, P is downsampled by the FPS algorithm and concatenated with the coarse point cloud P_0. Then, intermediate point cloud $P_1 \in \mathbb{R}^{N_1 \times 3}$ is obtained through one shape-aware upsample Transformer (SUT). Another SUT is utilized to obatain the complete point cloud P_C. The architecture of SUT is illustrated in Fig. 7.

Local Feature Embedding and Enhancement. To make use of the rich neighborhood information during the upsampling process, given a sparse point cloud $P_{in} = \{p_i\}_{i=1}^{N_{in}}$, where N_{in} is the number of points in P_{in} and p_i is the 3D coordinate, we design a Neighborhood Communication Block (NCB) to extract the local feature F_L of P_{in}. Firstly, the KNN algorithm is utilized to obtain a neighborhood set $\mathcal{N}(p_i) = \{p_{i1}, p_{i2}, \ldots, p_{ik}\}$ of every point p_i, where k is the neighborhood size. We compute the geometric context of p_i as $[p_i, p_i - p_{ik}]$, where $[\cdot, \cdot]$ denotes the concatenate operation, and the geometric context vector $C_{geo} \in \mathbb{R}^{N_{in} \times k \times 6}$ aggregates the geometric context of all points in P_{in}. Then, a simple MLP is utilized to extract the point-wise feature $F_{in} = \{f_i\}_{i=1}^{N_{in}}$, where $f_i \in \mathbb{R}^c$ is the corresponding feature of p_i, c is the dimension of f_i. Similarly, KNN algorithm is utilized to obtain a neighborhood set $\mathcal{N}(f_i) = \{f_{i1}, f_{i2}, \ldots, f_{ik}\}$ of every feature f_i. The feature context of f_i is calculated by $[f_i, f_i - f_{ik}]$ and the feature context vector $C_f \in \mathbb{R}^{N_{in} \times k \times 2c}$ aggregates the feature context of all features in F_{in}. The local feature $F_L \in \mathbb{R}^{N_{in} \times C_L}$ of the sparse point cloud can be calculated by

$$F_L = \max_k \left[\alpha\left(C_{geo}\right), \beta\left(C_f\right)\right], \tag{5}$$

where α and β are two independent non-linear function implemented by MLPs. Before doing feature enhancement, the fused global feature F is replicated and concatenate with F_L. The attention mechanism is suitable for the self semantic communication of the point-wise feature. Thus, we employs Self-Attention Transformer (SAT) block for feature enhancement and the architecture of SAT is the same as the self-attention layer in (Vaswani et al. 2017). SAT integrates the information from different points of the point-wise feature by applying the multi-head self-attention with residual connection, which can be formulated as

$$\begin{aligned} Z &= LayerNorm(Q + MultiHead(Q, K, V)), \\ Q &= XW^Q, K = XW^K, V = XW^V, \end{aligned} \tag{6}$$

where X is the input feature, $LayerNorm(\cdot)$ denotes the layer normalization operation, $MultiHead(\cdot)$ denotes the multi-head self-attetion layer, $W^Q \in \mathbb{R}^{2C_L \times C_Q}$, $W^K \in \mathbb{R}^{2C_L \times C_K}$ and $W^V \in \mathbb{R}^{2C_L \times C_V}$ are linear transformation matrices. Subsequently, a feed forward network is utilized to add non-linearity, thus the output of SAT can be formulated as

$$SAT(X) = Z + FFN(Z), \tag{7}$$

After 3 consecutive SAT blocks, we get the enhanced local feature $F'_L \in \mathbb{R}^{N_{in} \times C'_L}$.

Feature Expansion and Displacement Generation. Note that during the local feature embedding and enhancement step, the number of points in point-wise feature F'_L remain unchanged. In order to achieve the goal of upsampling, we

use 1D transpose convolution for feature expansion. This operation is called point splitting (Xiang et al. 2021), which can be intuitively interpreted as splitting each point of the point-wise feature into multiple points, thus increase the number of points. Then, the local feature F_L is replicated and concatenated with the expanded feature. In this way, local feature serves as a guidance throughout the entire upsampling process, which is crucial to recovering local geometric details. Inspired by ViPC (Zhang et al. 2021) and CSDN (Zhu et al. 2023), we also adopt the idea of predicting point-wise displacement for refinement. The displacement vector $\Delta P \in \mathbb{R}^{rN_{in} \times 3}$ is obtained by applying a shared MLP to the concatenated feature, where r is the upsample ratio. The output of SUT module can be formulated as

$$P_{out} = replicate\left(P_{in}, r\right) + \Delta P. \tag{8}$$

Loss Function

Like the current state-of-the-art work (Aiello, Valsesia, and Magli 2022), we utilize L1 Chamfer Distance (L1-CD) as our loss function, which can be written as:

$$\begin{aligned}\mathcal{L}_{CD}(Y, Y_{gt}) =& \frac{1}{2|Y|} \sum_{y \in Y} \min_{\hat{y} \in Y_{gt}} ||y - \hat{y}|| + \\ & \frac{1}{2|Y_{gt}|} \sum_{\hat{y} \in Y_{gt}} \min_{y \in Y} ||\hat{y} - y||,\end{aligned} \tag{9}$$

where Y is the predicted point cloud and Y_{gt} is the ground truth. Chamfer Distance (CD) calculates the average closest point distance between the predicted point cloud and the ground truth point cloud, where the first term forces the predicted points to lie close to the ground truth points and the second term ensures the ground truth point cloud is covered by the predicted point cloud.

In order to explicitly constrain point clouds generated at each stage of our network, we calculate L1-CD for coarse point cloud P_0, the intermediate point cloud P_1 and the complete point cloud P_C. The corresponding ground-truth point clouds Y_0 and Y_1 are downsampled by the FPS algorithm from the ground truth Y_{gt} to the same resolution as P_0 and P_1 respectively. The total training loss of our network is defined as

$$\mathcal{L} = \mathcal{L}_{CD}\left(P_0, Y_0\right) + \mathcal{L}_{CD}\left(P_1, Y_1\right) + \mathcal{L}_{CD}\left(C, Y_{gt}\right). \tag{10}$$

Experiments

Datasets

The dataset used in our experiment is the benchmark dataset ShapeNet-ViPC (Zhang et al. 2021), which is derived from ShapeNet (Chang et al. 2015). It contains 38,328 objects from 13 categories, including airplane, bench, cabinet, car, chair, monitor, lamp, speaker, firearm, sofa, table, cellphone and watercraft. For each object, there are 24 partial point clouds with occlusions generated under 24 viewpoints, which follow the same settings as ShapeNetRendering (Chang et al. 2015). The complete ground-truth point cloud is generated by uniformly sampling 2,048 points from the mesh surface of a target in ShapeNet (Chang et al. 2015). Moreover, each 3D shape is rotated to the pose corresponding to a specific view point and normalized into the bounding sphere with radius of 1. Images are generated from the same 24 viewpoints as ShapeNetRendering with a resolution of 224×224. Each object has 24 partial point clouds and 24 corresponding images rendered from 24 viewpoints and a ground-truth point cloud. During training, we randomly chose a viewpoint for the image and the partial point cloud. (The viewpoint of the two modalities can be different.) For the experiments conducted on the known categories, we employ the same experimental setting as ViPC (Zhang et al. 2021), i.e., we use 31,560 objects from eight categories with 80% for training and 20% for testing. And for the experiments conducted on novel categories, the training set remains unchanged and we use objects from 4 categories for testing which are not used in the training set.

Implementation Details and Evaluation Metrics

In our implementation, the input partial point cloud P contains $N = 2048$ points. Follow XMFnet (Aiello, Valsesia, and Magli 2022), the EdgeConv layers chose $k = 20$ as neighborhood size and the SAGPool layers chose $k = 16$ and $k = 6$ as neighborhood size. Each SAGPool layer downsamples the partial point cloud by a factor of 4. The output feature size of the 3D encoder is $2,048/16 \times 512 = 128 \times 512$. The output feature map of the 2D encoder has a size of $7 \times 7 \times 512$, and by reshaping, the size of the pixel-wise feature is 49×512. The output feature dimensions of the DMF module is 512. The coarse point cloud P_0 contains $N_0 = 256$ points and the partial point cloud P is downsampled by the FPS algorithm to 256 points and concatenated with P_0 leading to a point cloud with 512 points. The NCB chose $k = 16$ as neighborhood size. The dimension of the local feature F_L is $C_L = 128$. The dimension of the enhanced local feature F'_L is $C'_L = 512$. Each SUT module has the upsample ratio of 2, which means the output point cloud of the first SUT module contains $N_1 = 512 \times 2 = 1024$ points and the complete point cloud contains $N_C = 1024 \times 2 = 2048$ points.

For quantitative evaluation, like previous works, we chose L2 Chamfer Distance (L2-CD) and F-Score as evaluation metrics. L2-CD can be formulated as

$$\begin{aligned}\mathcal{L}_{CD}(Y, Y_{gt}) =& \frac{1}{2|Y|} \sum_{y \in Y} \min_{\hat{y} \in Y_{gt}} ||y - \hat{y}||_2^2 + \\ & \frac{1}{2|Y_{gt}|} \sum_{\hat{y} \in Y_{gt}} \min_{y \in Y} ||\hat{y} - y||_2^2,\end{aligned} \tag{11}$$

The whole network is trained end-to-end with an initial learning rate of 1×10^{-4} for 120 epochs with a batch size of 16. Adam optimizer (Kingma and Ba 2014) is utilized during the training process with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is decayed by 0.7 for every 20 epochs. The proposed network is implemented by Pytorch and trained on NVIDIA RTX 3090 GPU.

Methods	Avg	Airplane	Cabinet	Car	Chair	Lamp	Sofa	Table	Watercraft
Unimodal Methods									
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
ECG (Pan 2020)	4.957	2.952	6.721	5.243	5.867	4.602	6.813	4.332	3.127
VRC-Net (Pan et al. 2021)	4.598	2.813	6.108	4.932	5.342	4.103	6.614	3.953	2.925
PointAttn (Wang et al. 2024)	2.853	1.613	3.969	3.257	3.157	3.058	3.406	2.787	1.872
Multimodal Methods									
ViPC (Zhang et al. 2021)	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, Valsesia, and Magli 2022)	1.443	0.572	1.980	1.754	1.403	1.810	1.702	1.386	0.945
Ours	**1.038**	**0.453**	**1.678**	**1.462**	**1.019**	**0.583**	**1.340**	**1.072**	**0.696**

Table 1: Quantitative comparison for the point cloud completion on ShapeNet-ViPC dataset using per-point L2 Chamfer distance $\times 10^{-3}$ (lower is better). The best results are highlighted in bold.

Experiment Results on Known Categories

In this section, we will compare our DMF-Net with several existing point cloud completion networks. We compare the results of several unimodal methods which only take the partial point cloud as input, including AtlasNet (Groueix et al. 2018), FoldingNet (Yang et al. 2018), PCN(Yuan et al. 2018), TopNet (Tchapmi et al. 2019), ECG(Pan 2020), VRC-Net (Pan et al. 2021) and PointAttN (Wang et al. 2024). AtlasNet and FoldingNet generate the complete point cloud by using 2D grids for folding operation. PCN employs an encoder-decoder architeture to recover the 3D shape in a coarse-to-fine manner. TopNet generates structured point clouds with a tree-like decoder. ECG emloys an auto-encoder to recover the skeleton of a complete point cloud and applies convolution on edges for refinement. VRC-Net proposes a probabilistic modeling and relational enhancement network for coarse-to-fine point cloud completion. PointAttN proposes a coarse-to-fine completion network with attention mechanism throughout the whole completion process. We also compare the results of existing multi-modal point cloud completion methods, including ViPC (Zhang et al. 2021), XMFnet (Aiello, Valsesia, and Magli 2022) and CSDN (Zhu et al. 2023). ViPC is the pioneering work for leveraging a single-view image for point cloud completion. XMFnet adopts cross- and self-attention layers for modality fusion and proposes a decoder with several independent braches for generating different regions of the complete point cloud. CSDN reformulate the modality fusion problem as a style transfer problem and predicts offset vectors for refinement. For the above methods, the output predicted complete point cloud contains 2,048 points.

Quantitative Results. Follow XMFnet (Aiello, Valsesia, and Magli 2022), we calculated CD on the 2,048 points of each object. A lower CD value indicates better performance. Category-specific training is performed and the results for each category is reported in Table 1. It can be seen that the performance of multimodal completion networks is generally better than that of unimodal completion networks, proving that the introduction of image modality is effective for point cloud completion. The proposed DMF-Net achieves the best results across all 8 categories in terms of CD. It is worth noting that compared with the current top-ranked XMFnet, DMF-Net reduces the average CD value by 0.405, which is 28.1% lower than the results of XMFnet. Moreover, each category has different degrees of improvement in completion results, with a significant improvement in the lamp category, proving that DMF-Net can handle objects with slender structures well.

Qualitative Results. Due to space limitations, the presentation of qualitative results can be found in the supplementary materials, which shows the point clouds completed by the proposed DMF-Net and its competitors. Our DMF-Net not only restores fine local geometric structures, but also ensures the uniformity of points. Besides, the overall shapes of complete point clouds generated by the proposed DMF-Net are cloest to the ground-truth.

Conclusion

In this paper, we propose a novel image-guided point cloud completion network with dual-channel modality fusion and shape-aware upsampling transformer. Unlike existing multimodal completion network, DMF-Net employs an upsampling transformer to tackle the task in a coarse-to-fine manner. Moreover, Our recognition of the equal importance of both modalities in feature fusion enables DMF-Net to effectively capture complementary information from both image and point cloud. Extensive experiments demonstrate the significant improvement of our DMF-Net over other state-of-the-art methods. However, our work lacks exploration of real-world scenarios which have lower resolution image and point cloud with higher degree of incompleteness. In the future work, we will focus on improving the generalization ability of the model in real-world scans and how to introduce more modalities like text to help enhance the performance of the completion network.

Acknowledgments

This work was supported by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515012791 and the Ministry of Education, Singapore, under its Academic Research Fund Grants (MOE-T2EP20220-0005 & RT19/22).

References

Aiello, E.; Valsesia, D.; and Magli, E. 2022. Cross-modal Learning for Image-Guided Point Cloud Shape Completion. In *Advances in Neural Information Processing Systems*.

Berger, M.; Tagliasacchi, A.; Seversky, L. M.; Alliez, P.; Levine, J. A.; Sharf, A.; and 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, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Cui, Y.; Chen, R.; Chu, W.; Chen, L.; Tian, D.; Li, Y.; and Cao, D. 2021. Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23(2): 722–739.

Dai, A.; Ruizhongtai Qi, C.; and Nießner, M. 2017. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5868–5877.

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

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

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

Han, X.; Li, Z.; Huang, H.; Kalogerakis, E.; and Yu, Y. 2017. High-resolution shape completion using deep neural networks for global structure and local geometry inference. In *Proceedings of the IEEE international conference on computer vision*, 85–93.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hou, J.; Dai, A.; and Nießner, M. 2019. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4421–4430.

Huang, Z.; Yu, Y.; Xu, J.; Ni, F.; and 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*, 7662–7670.

Kalogerakis, E.; Chaudhuri, S.; Koller, D.; and Koltun, V. 2012. A probabilistic model for component-based shape synthesis. *Acm Transactions on Graphics (TOG)*, 31(4): 1–11.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

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

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

Lee, J.; Lee, I.; and Kang, J. 2019. Self-attention graph pooling. In *International conference on machine learning*, 3734–3743. PMLR.

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

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

Mandikal, P.; and Radhakrishnan, V. B. 2019. Dense 3d point cloud reconstruction using a deep pyramid network. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1052–1060. IEEE.

Mao, A.; Du, Z.; Hou, J.; Duan, Y.; Liu, Y.-j.; and He, Y. 2022. Pu-flow: A point cloud upsampling network with normalizing flows. *IEEE Transactions on Visualization and Computer Graphics*.

Mitra, N. J.; Guibas, L. J.; and 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.; and Ceylan, D. 2013. Symmetry in 3d geometry: Extraction and applications. In *Computer graphics forum*, volume 32, 1–23. Wiley Online Library.

Nan, L.; Sharf, A.; and Chen, B. 2014. 2D-D Lifting for Shape Reconstruction. In *Computer Graphics Forum*, volume 33, 249–258. Wiley Online Library.

Nguyen, D. T.; Hua, B.-S.; Tran, K.; Pham, Q.-H.; and Yeung, S.-K. 2016. A field model for repairing 3d shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5676–5684.

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

Pan, L.; Chen, X.; Cai, Z.; Zhang, J.; Zhao, H.; Yi, S.; and Liu, Z. 2021. Variational relational point completion network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8524–8533.

Park, Y.; Lepetit, V.; and Woo, W. 2008. Multiple 3d object tracking for augmented reality. In *2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, 117–120. IEEE.

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

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

Qi, C. R.; Litany, O.; He, K.; and Guibas, L. J. 2019. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, 9277–9286.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Shao, T.; Xu, W.; Zhou, K.; Wang, J.; Li, D.; and 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.

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

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

Tchapmi, L. P.; Kosaraju, V.; Rezatofighi, H.; Reid, I.; and Savarese, S. 2019. Topnet: Structural point cloud decoder. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 383–392.

Varley, J.; DeChant, C.; Richardson, A.; Ruales, J.; and Allen, P. 2017. Shape completion enabled robotic grasping. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 2442–2447. IEEE.

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

Wang, J.; Cui, Y.; Guo, D.; Li, J.; Liu, Q.; and Shen, C. 2024. Pointattn: You only need attention for point cloud completion. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 38, 5472–5480.

Wang, X.; Ang Jr, M. H.; and Lee, G. H. 2020. Cascaded refinement network for point cloud completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 790–799.

Xiang, P.; Wen, X.; Liu, Y.-S.; Cao, Y.-P.; Wan, P.; Zheng, W.; and 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*, 5499–5509.

Xie, H.; Yao, H.; Zhou, S.; Mao, J.; Zhang, S.; and Sun, W. 2020. Grnet: Gridding residual network for dense point cloud completion. In *European Conference on Computer Vision*, 365–381. Springer.

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

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

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

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

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

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

Zhou, Y.; and Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4490–4499.

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