Zero-Shot Multi-Object Scene Completion

Shun Iwase[1,2], Katherine Liu[2], Vitor Guizilini[2], Adrien Gaidon[2], Kris Kitani[1,⋆], Rareș Ambruș[2,⋆], and Sergey Zakharov[2,⋆]

[1] Carnegie Mellon University
[2] Toyota Research Institute

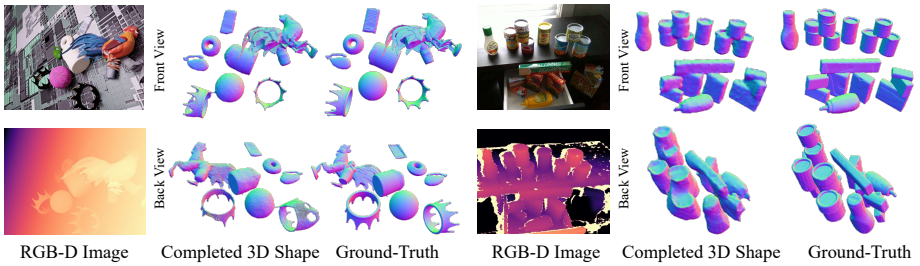

Fig. 1: Given an RGB-D image and the foreground mask of multiple objects not seen during training, our method predicts their complete 3D shapes quickly and accurately, including occluded areas. (**Left**) Synthetic image results. (**Right**) Zero-shot generalization to a real-world image of household objects with noisy depth data. Our 3D results are rotated with respect to the input to highlight completions in occluded regions.

Abstract. We present a 3D scene completion method that recovers the complete geometry of multiple unseen objects in complex scenes from a single RGB-D image. Despite notable advancements in single-object 3D shape completion, high-quality reconstructions in highly cluttered real-world multi-object scenes remains a challenge. To address this issue, we propose OctMAE, an architecture that leverages an Octree U-Net and a latent 3D MAE to achieve high-quality and near real-time multi-object scene completion through both local and global geometric reasoning. Because a naive 3D MAE can be computationally intractable and memory intensive even in the latent space, we introduce a novel occlusion masking strategy and adopt 3D rotary embeddings, which significantly improve the runtime and scene completion quality. To generalize to a wide range of objects in diverse scenes, we create a large-scale photorealistic dataset, featuring a diverse set of 12K 3D object models from the Objaverse dataset that are rendered in multi-object scenes with physics-based positioning. Our method outperforms the current state-of-the-art on both synthetic and real-world datasets and demonstrates a strong zero-shot capability. https://sh8.io/#/oct_mae

⋆ Equal advising.

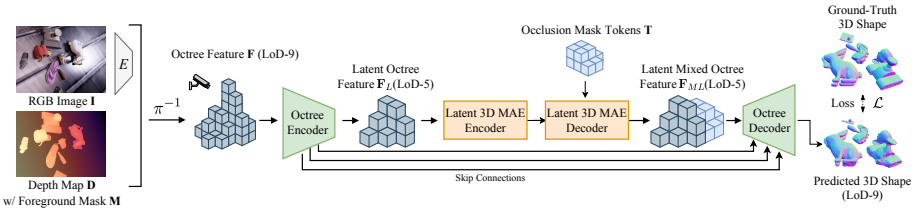

Fig. 2: Overview of our proposed method (OctMAE). Given an input RGB Image **I**, depth map **D**, and a foreground mask **M**, the octree feature **F** is obtained by unprojecting an image feature encoded by a pre-trained image encoder **E**. The octree feature is then encoded by the Octree encoder and downsampled to the Level of Detail (LoD) of 5. The notation LoD-h indicates that each axis of the voxel grid has resolution of 2^h. The latent 3D MAE takes the encoded Octree feature **F** as input and its output feature is concatenated with the occlusion mask tokens **T**. Next, the masked decoded feature $\mathbf{F}_{ML}$ is computed by sparse 3D MAE decoder. Finally, the Octree decoder predicts a completed surface at LoD-9.

1 Introduction

Humans can instantly imagine complete shapes of multiple novel objects in a cluttered scene via advanced geometric and semantic reasoning. This ability is also essential for robots if they are to effectively perform useful tasks in the real world [26, 27, 46, 60]. In this work, we propose a method that can quickly and accurately complete a wide number of objects in diverse real-world scenes.

Prior works [31, 34, 36, 43, 47, 71] have achieved phenomenal progress in scene and object shape completion from a single RGB-D image. Object-centric methods [17, 25] in particular can achieve very high reconstruction accuracy by relying on category-specific shape priors. However, when deployed on entire scenes such methods require bespoke instance detection/segmentation models, and often perform test-time optimization which is time consuming and would hinder real-time deployment on a robot. Moreover, existing methods are typically limited to a small set of categories. Thus, zero-shot multi-object scene completion remains a challenging and open problem that has seen little success to date. This is in stark contrast to the sudden increase in powerful algorithms for 2D computer vision tasks such as object detection [33, 75] and image segmentation [35, 70]. We attribute this progress to a great extent to the availability of large-scale datasets [8, 54] coupled with neural architectures and learning objectives [22, 50, 53, 57] that can effectively exploit the highly structured data occurring in the natural world [20].

Taking inspiration from the latest developments in the 2D domain, we propose a scene completion algorithm at the scene level that generalizes across a large number of shapes and that only supposes an RGB-D image and foreground mask as input. Our method consists of Octree masked autoencoders (OctMAE) — a hybrid architecture of Octree U-Net and a latent 3D MAE (Figure 2). Although a recent work, VoxFormer [34], also extends MAE architecture to 3D

using deformable 3D attention and shows great improvement in semantic scene completion tasks, its memory utilization is still prohibitive to handle a higher resolution voxel grid. We address this issue by integrating 3D MAE into the latent space of Octree U-Net. Our experiments show that the latent 3D MAE is the key to global structure understanding and leads to strong performance and generalization across all datasets. Moreover, we find that the choice of a masking strategy and 3D positional embeddings is crucial to achieve better performance. We provide extensive ablations to verify that our 3D latent MAE design is effective.

Our second contribution consists of the creation of a novel synthetic dataset to counteract the lack of large-scale and diverse 3D datasets. The dataset contains 12K 3D models of hand-held objects from Objaverse [12] and GSO [16] datasets (Figure 3). We utilize the dataset to conduct a comprehensive evaluation of our method as well as other baselines and show that our method scales and achieves better results. Finally, we perform zero-shot evaluations on synthetic as well as real datasets and show that a combination of 3D diversity coupled with an appropriate architecture is key to generalizable scene completion in the wild.

Our contributions can be summarized as follows:

- We present a novel network architecture, Octree Masked Autoencoders (OctMAE), a hybrid architecture of Octree U-Net and latent 3D MAE, which achieves state-of-the-art results on all the benchmarks. Further, we introduce a simple occlusion masking strategy with full attention, which boosts the performance of a latent 3D MAE.
- We create the first large-scale and diverse synthetic dataset using Objaverse [12] dataset for zero-shot multi-object scene completion, and provide a wide range of benchmark and analysis.

2 Related Work

3D reconstruction and completion. Reconstructing indoor scenes and objects from a noisy point cloud has been widely explored [1, 2, 4, 6, 9, 10, 23, 24, 34, 40, 42, 47, 48, 56, 65, 66]. Several works [4, 5, 43, 44, 47, 58, 60, 63, 71, 72, 74, 76] tackle more challenging shape completion tasks where large parts of a target is missing. While these methods achieve impressive results, they do not explicitly consider semantic information, which may limit their capability for accurate shape completion. Recent methods [31, 32, 34, 76] in Semantic Scene Completion (SSC) leverage semantic information via an RGB image. Nevertheless, the number of target categories is quite limited, restricting its utility for a broad range of applications in the real world. In addition, many methods adopt occupancy or SDF as an output representation, which necessitates post-processing such as the marching cubes [41] and sphere tracing to extract an explicit surface. As another direction, GeNVS [3], Zero-1-to-3 [39], and 3DiM [64] explore single-view 3D reconstruction via novel view synthesis. However, expensive test-time optimization is required. Recently, One-2-3-45 [38] and MCC [66] attempt to improve the generation speed, however, their runtime for multi-object scenes is still far from near

real-time. Further, since these methods are object-centric, multiple objects in a single scene are not handled well due to the complicated geometric reasoning especially caused by occlusions by other objects. In this paper, we propose a general and near real-time framework for multi-object 3D scene completion in the wild using only an RGB-D image and foreground mask without expensive test-time optimization.

Implicit 3D representations. Recently, various types of implicit 3D representation have become popular in 3D reconstruction and completion tasks. Early works [18, 42, 47] use a one-dimensional latent feature to represent a 3D shape as occupancy and SDF fields. Several works [31, 48, 58] employ voxels, groundplanes, and triplanes, demonstrating that the retention of geometric information using 3D CNNs enhances performance. Although the voxel representation typically performs well among these three, its cubic memory and computational costs make increasing resolution challenging. To mitigate this issue, sparse voxels [6, 21, 37, 55, 62] treat a 3D representation as a sparse set of structured points using the octree and hash table and perform convolutions only on non-empty voxels and its neighbors. Further, the high-resolution sparse voxel enables a direct prediction of a target surface. As another direction, [1, 67, 77] leverage point cloud. Nonetheless, an unstructured set of points can be non-uniformly distributed in the 3D space and requires running the k-NN algorithm at every operation. This aspect often renders point-based methods less appealing compared to the sparse voxel representation. Therefore, our method adopts an octree-based representation used in [62] for efficient training and direct surface prediction.

Masked Autoencoders (MAE). Inspired by the success of ViTs [15, 73] and masked language modeling [14, 51], [22] demonstrates that masked autoencoders (MAE) with ViTs can learn powerful image representation by reconstructing masked images. To improve the efficiency and performance of MAE, ConvMAE [19] proposes a hybrid approach that performs masked autoencoding at the latent space of 2D CNN-based autoencoder network. Recently, VoxFormer [34] extends the MAE design to 3D for semantic scene completion using 3D deformable attention, and shows great improvement over previous works. However, it is not trivial to scale up the MAE architecture to a higher resolution voxel due to memory constraints. Motivated by ConvMAE [19] and OCNN [62], we propose an efficient OctMAE architecture using sparse 3D operations.

3 Proposed Method

Given an RGB image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, depth map $\mathbf{D} \in \mathbb{R}^{H \times W}$, and foreground mask $\mathbf{M} \in \mathbb{R}^{H \times W}$ containing all objects of interest, we aim to predict their complete 3D shapes quickly and accurately. Our framework first encodes an RGB image $\mathbf{I}$ with a pre-trained image encoder E such as ResNeXt [69] and then lifts the resulting features up to 3D space using a depth map $\mathbf{D}$ and foreground mask

$\mathbf{M}$ to acquire 3D point cloud features $\mathbf{F} \in \mathbb{R}^{N \times D}$ and its locations $\mathbf{P} \in \mathbb{R}^{N \times 3}$ (Section 3.1). Second, we convert the 3D features into an octree using the same algorithm used in [63] and pass it to OctMAE to predict a surface at each LoD (Section 3.2). The diagram of our method is visualized in Figure 2.

3.1 Octree Feature Aggregation

We adopt ResNeXt-50 [69] as an image encoder to obtain dense and robust image features $\mathbf{W} = E(\mathbf{I}) \in \mathbb{R}^{H \times W \times D}$ from an RGB image. The image features are unprojected into the 3D space using a depth image with $(\mathbf{F}, \mathbf{P}) = \pi^{-1}(\mathbf{W}, \mathbf{D}, \mathbf{M}, \mathbf{K})$ where a point cloud feature and its corresponding coordinates are represented as $\mathbf{F}$ and $\mathbf{P}$. π^{-1} unprojects the image features $\mathbf{W}$ to the camera coordinate system using a depth map $\mathbf{D}$, foreground mask $\mathbf{M}$, and an intrinsic matrix $\mathbf{K}$. Next, we define an octree at the level of detail (LoD) of 9 (512^3) with the grid and cell size being 1.28 m and 2.5 mm respectively, and use the point features to populate the voxel grid, averaging features when multiple points fall into the same voxel. Here, LoD-h simply represents resolution of an octree. For instance, the voxel grid of LoD-9 has the maximum dimension of $2^9 = 512$ for each axis. An octree is represented as a set of 8 octants with features at non-empty regions; therefore, it is more memory-efficient than a dense voxel grid. The octree is centered around the z-axis in the camera coordinate system, and its front plane is aligned with the nearest point to the camera along with the z-axis.

3.2 OctMAE: Octree Masked Autoencoders

We design OctMAE which leverages Octree U-Net [62] and latent 3D MAE to achieve accurate and efficient zero-shot multi-object scene completion. Octree U-Net consists of multiple sparse 3D convolutional layers. While the Octree U-Net architecture can efficiently encode octree features to low resolution, only local regions are considered at each operation. On the contrary, 3D MAE can capture global object information which helps predict globally consistent 3D shapes. However, unlike an image, a dense voxel grid contains a prohibitive number of tokens even in the latent space, which makes it challenging to adopt an MAE architecture directly for 3D tasks. Recently, ConvMAE [19] proposed to leverage the advantages of both CNNs and MAE in 2D for efficient training. Nevertheless, a naïve extension of ConvMAE [19] to 3D also leads to prohibitive computational and memory costs. To address this issue, we propose a novel occlusion masking strategy and adopt 3D rotary embeddings, enabling efficient masked autoencoding in the latent space.

Encoder architecture. The encoder of Octree U-Net [63] takes the octree feature at LoD-9 and computes a latent octree feature $\mathbf{F}_L \in \mathbb{R}^{N' \times D'}$ at LoD-5 where N' is the number of non-empty voxels and D' is the latent feature dimension. To incorporate global symmetric and object scale information which gives more cues about completed shapes, we use S layers of the full self-attention

Transformer blocks in the latent 3D MAE encoder. Since N' is typically the order of the hundreds to thousands, we resort to memory-efficient attention algorithms [11, 49]. Ideally, learnable relative positional encodings [77] are used to deal with the different alignments of point cloud features inside an octree. However, it requires computing the one-to-one relative positional encoding $N' \times N'$ times, which largely slows down the training and makes it computationally impractical. Therefore, we use RoPE [59] to encode 3D axial information between voxels. Concretely, we embed position information with RoPE at every multi-head attention layer as

$$\mathbf{R}_i = \mathrm{diag}\left(R(p_i^x), R(p_i^y), R(p_i^z), \mathbf{I}\right) \in \mathbb{R}^{D' \times D'}, \quad \mathbf{f}'_i = \mathbf{R}_i \mathbf{f}_i, \tag{1}$$

where $\mathbf{f}_i \in \mathbb{R}^{D'}$, and $\mathbf{p}_i \in \mathbb{R}^3$ is i-th octree feature and coordinates. $R : \mathbb{R} \to \mathbb{R}^{\lfloor D'/3 \rfloor \times \lfloor D'/3 \rfloor}$ is a function to generate a rotation matrix given normalized 1D axial coordinate. The detailed derivation of $\mathbf{R}$ can be found in the supplemental.

Occlusion masking. Next, we concatenate mask tokens $\mathbf{T} \in \mathbb{R}^{M \times D'}$ to the encoded latent octree feature where M is the number of the mask tokens. Note that each of the mask tokens has identical learnable parameters. The key question is how to place them in 3D space. Although previous methods [34] put mask tokens inside all the empty cells of a dense voxel grid, it is unlikely that visible regions extending from the camera to the input depth are occupied unless the error of a depth map is enormous. Further, this dense masking strategy forces to use a local attention mechanism such as deformable 3D attention used in VoxFormer [34], due to the highly expensive memory and computational cost. To address this issue, we introduce an occlusion masking strategy in which the mask tokens $\mathbf{T}$ are placed only into occluded voxels. Concretely, we perform depth testing on every voxel within a voxel grid to determine if they are positioned behind objects. Mask tokens are assigned to their respective locations only after passing this test. The proposed occlusion masking strategy and efficient positional encoding enable our latent 3D MAE (Figure 4) to leverage full attention instead of local attention.

Decoder architecture. The masked octree feature is given to the latent 3D MAE decoder which consists of S layers of the full cross-attention Transformer blocks with RoPE [59] to learn global reasoning including occluded regions. Finally, the decoder of Octree U-Net takes the mixed latent octree feature of the Transformer decoder $\mathbf{F}_{ML} \in \mathbb{R}^{(N'+M) \times D'}$ as input and upsamples features with skip connections. The decoded feature is passed to a two-layer MLP which estimates an occupancy at LoD-h. In addition, normals and SDF values are predicted only at the final LoD. To avoid unnecessary computation, we prune grid cells predicted as empty with a threshold of 0.5 at every LoD, following [63].

3.3 Training Details and Loss Functions

We use all surface points extracted through OpenVDB [45] during training. The loss function is defined as

Fig. 3: Example images of our synthetic dataset. We use BlenderProc [13] to acquire high-quality images under various and realistic illumination conditions.

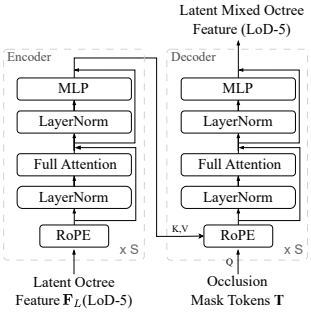

Fig. 4: Overall architecture of Latent 3D MAE.

Table 1: Dataset comparisons. We create the first large-scale and diverse 3D scene completion dataset for novel multiple objects using a subset of 3D models from Objaverse dataset [12]. The number of categories is reported by using the LVIS categories, and $R^{\text{LVIS}}(\%)$ represents a ratio of the number of the categories covered by the dataset. † denotes the number of objects with actual size.

Dataset	Type	3D Models	# Frames	# Objs	# Cats	$R^{\text{LVIS}}(\%)$
YCB-V [68]	Real	✓	133K	21	5	0.4
HB [28]	Real	✓	17K	33	13	1.0
HOPE [36]	Real	✓	2K	28	3	0.3
CO3D V2 [52]	Real		6M	**40K**	50	4.2
MegaPose [30]	Synthetic	✓	1M	1K†	17	0.9
Ours	Synthetic	✓	1M	12K	**601**	**50.0**

$$\mathcal{L} = \mathcal{L}_{nrm} + \mathcal{L}_{SDF} + \sum_{h \in \{5,6,7,8,9\}} \mathcal{L}^h_{occ}, \tag{2}$$

where $\mathcal{L}_{nrm}$ and $\mathcal{L}_{SDF}$ measure the averaged L2 norm of normals and SDF values. $\mathcal{L}^h_{occ}$ computes a mean of binary cross entropy function of each LoD-h.

4 Dataset

As shown in Table 1, existing datasets are limited in the diversity of object categories. Although the CO3D V2 dataset [52] contains data for 40k objects, because the provided ground-truth 3D shapes are reconstructed from unposed multi-view images, they tend to be highly noisy and parts of the object missing due to lack of visibility. To tackle this problem, we leverage Objaverse [12], a large-scale 1M 3D object dataset containing 46k objects with LVIS category annotations. To focus on completion of hand-held objects, we select 601 categories and ensure that the largest dimension of the objects in each category

falls approximately within the range of 4 cm to 40 cm. In addition, for high-quality rendering, we omit objects that lack textures, contain more than $10,000$ vertices, or are articulated. To increase the number of objects, we add objects from Google Scanned Objects (GSO) [16], which results in $12,655$ objects in total. We render 1M images of $25,000$ scenes using physics-based rendering and positioning via BlenderProc [13] to simulate realistic scenes (Figure 3). For each image, we randomly choose a camera view such that at least one object is within the camera frame. We also generate $1,000$ images using 250 withheld objects for evaluation.

5 Experimental Results

Implementation details. We train all the models for 2 epochs using the Adam [29] optimizer with a learning rate of 0.002 and batch size of 16 on NVIDIA A100. Note that the models are only trained on the synthetic dataset introduced in Section 4. In addition, the number of Transformer blocks K, the feature dimension D, and D' are set to 3, 32, and 192 respectively. We use a pre-trained model of ResNeXt-50 [69] as an image encoder for all the experiments. The ground-truth occupancy, SDF and normals are computed from meshes with OpenVDB [45]. During training, we dilate ground-truth masks using the radius randomly selected from 1, 3 and 5 pixels to deal with the segmentation error around the object edges. During evaluation, we use ground-truth masks provided by the datasets.

Evaluation metrics. We report Chamfer distance (CD), F1-Score@10mm (F1), and normal consistency (NC) to evaluate the quality of a completed surface. For surface-based methods, we use a predicted surface directly for evaluation. For the methods that predict occupancy, the marching cubes algorithm [41] is used to extract a surface and uniformly sample $100,000$ points from its surface such that the number of points are roughly equal to the surface prediction methods. We use mm as a unit for all the reported metrics.

Evaluation datasets. We evaluate the baselines and our model on one synthetic and three real-world datasets. For the synthetic dataset, we render $1,000$ images using textured 3D scans from Objaverse [12], following the same procedure described in Section 4. We randomly choose 3 to 5 objects per image from the withheld objects for Objavese dataset. Since these 3D scans are relatively more complex than the objects seen in the real-world datasets we use, they can provide a good scene completion quality estimate for complex objects. For the real-world dataset, we use the YCB-Video [68], HOPE [36] and HomebrewedDB (HB) [28] datasets. YCB-Video consists of 21 everyday objects with diverse shapes. HOPE contains 28 simple household objects with mostly rectangular and cylindrical everyday shapes, and the images are captured in various lighting conditions in indoor scenes using a RealSense D415 RGBD camera. HB includes 33 objects (*e.g.*, toy, household, and industrial objects). Their images are taken by PrimeSense Carmine in lab-like environments.

Table 2: Quantitative evaluation of multi-object scene completion on Ours, YCB-Video [68], HOPE [36], and HomebrewedDB [28] datasets. Chamfer distance (CD), F1-Score@10mm (F1), and normal consistency (NC) are reported. Chamfer distance is reported in the unit of mm.

Method	3D Rep.	Synthetic			Real								
		Ours			YCB-Video [68]			HB [28]			HOPE [36]		
		CD↓	F1↑	NC↑	CD↓	F1↑	NC↑	CD↓	F1↑	NC↑	CD↓	F1↑	NC↑
VoxFormer [34]	Dense	44.54	0.382	0.653	30.32	0.438	0.641	34.84	0.366	0.608	47.75	0.323	0.594
ShapeFormer [71]	Dense	39.50	0.401	0.593	38.21	0.385	0.588	40.93	0.328	0.594	39.54	0.306	0.591
MCC [66]	Implicit	43.37	0.459	0.700	35.85	0.289	0.608	19.59	0.371	0.655	17.53	0.357	0.658
ConvONet [48]	Dense	23.68	0.541	0.710	32.87	0.458	0.649	26.71	0.504	0.643	20.95	0.581	0.678
POCO [1]	Implicit	21.11	0.634	0.753	15.45	0.587	0.699	13.17	0.624	0.709	13.20	0.602	0.706
AICNet [31]	Dense	15.64	0.573	0.741	12.26	0.545	0.702	11.87	0.557	0.674	11.40	0.564	0.670
Minkowski [6]	Sparse	11.47	0.746	0.802	8.04	0.761	0.717	8.81	0.728	0.719	8.56	0.734	0.709
OCNN [63]	Sparse	9.05	0.782	0.828	7.10	0.778	0.771	7.02	0.792	0.736	8.05	0.742	0.736
Ours	Sparse	**6.48**	**0.839**	**0.848**	**6.40**	**0.800**	**0.785**	**6.14**	**0.819**	**0.770**	**6.97**	**0.803**	**0.750**

Baselines. As discussed in Secs. 1 and 2, multi-object scene completion from a single RGB-D image is relatively not explored due to the lack of large-scale and diverse multi-object scene completion datasets. We carefully choose baseline architectures that can support this task with simple or no adaptation. We focus on three primary method types from related fields. Firstly, we select Semantic Scene Completion (SSC) methods [6, 31, 34, 63] that do not heavily rely on domain or categorical knowledge of indoor or outdoor scenes. Secondly, we opt for object shape completion methods [6, 63, 66, 71] that can be extended to multi-object scene completion without an architectural modification and prohibitive memory utilization. Thirdly, we consider voxel or octree-based 3D reconstruction methods [1, 6, 48, 63] that predict a complete and plausible shape using noisy and sparse point cloud data. For dense voxel-based (*e.g.*, AICNet [31], ConvONet [48] and VoxFormer [34]) and sparse voxel-based methods (*e.g.*, MinkowskiNet [6], OCNN [63], and our method), we use LoD-6 and LoD-9 as an input resolution respectively. All the experiments are conducted using the original implementation provided by the authors, with few simple modifications to adapt for multi-object scene completion and a fair comparison. For instance, we extend the baselines that take the point cloud as input by concatenating the image features to the point cloud features. For occupancy-based methods, though their output voxel grid resolution is LoD-6, we use trilinear interpolation to predict occupancy at LoD-7 [48]. For MinkowskiNet [6] and OCNN [62, 63], we use the U-Net architecture with the depth of 5 (LoD-9 to LoD-4). We discuss further details about the baseline architectures, their modifications, and hyperparameters in the supplemental.

5.1 Quantitative Results

Table 2 shows that our method outperforms the baselines on all the metrics and datasets. Although our model is only trained on synthetic data, it demonstrates strong generalizability to real-world datasets. We also remark that our

Table 3: Ablation Study of positional encoding on our synthetic dataset. We compare w/o positional encoding, conditional positional encoding (CPE) [7], absolute postional encoding (APE) used in [34], and RoPE [59].

Type	CD↓	F1↑	NC↑
w/o	11.32	0.778	0.808
CPE [7]	9.91	0.785	0.811
APE [34]	8.61	0.782	0.825
RPE [61]	<u>7.81</u>	<u>0.804</u>	<u>0.830</u>
RoPE [59]	**6.48**	**0.839**	**0.848**

Table 4: Ablation study on 3D attention algorithms. The scores are reported on the HOPE dataset [36].

Method	Occ. Masking	CD↓	F1↑	Runtime↓
3D DSA [34]		12.14	0.703	93.3
Neighbor. Attn. [77]		9.26	0.727	130.8
Octree Attn. [61]		7.99	0.752	116.4
Neighbor. Attn. [77]	✓	8.81	0.759	111.9
Octree Attn. [61]	✓	7.54	0.772	105.3
Full + Self Attn.	✓	<u>7.21</u>	<u>0.785</u>	<u>86.2</u>
Full + Cross Attn.	✓	**6.97**	**0.803**	**85.1**

method exhibits robustness to the noise characteristics present in depth data captured by typical RGB-D cameras despite being trained on noise-free depth data in simulation. The comparisons show that hierarchical structures and the latent 3D MAE are key to predicting 3D shapes of unseen objects more accurately than the baselines. Unlike our method, VoxFormer [34] uses an MAE with 3D deformable attention where only 8 neighbors of the reference points at the finest resolution are considered. Figure 8 also demonstrates that methods using a dense voxel grid or implicit representation fail to generalize to novel shapes. This implies that capturing a right choice of a network architecture is crucial to learn generalizable shape priors for zero-shot multi-object scene completion. Our method has the similar U-Net architecture used in MinkowskiNet [6] and OCNN [62] except we use the latent 3D MAE at LoD-5 instead of making the network deeper. This indicates that the latent 3D MAE can better approximate the shape distribution of the training dataset by leveraging an attention mechanism to capture global 3D contexts. Table 7 also confirms that our method achieves the best scene completion quality by measuring Chamfer distance in visible and occluded regions separately.

Positional encoding. As shown in Table 3, we explore the effect of RoPE [59] on the validation set of our synthetic dataset. The first row shows that all the metrics significantly drop if positional encoding is not used. In addition, we test CPE [7], APE [34], and RPE [61] and obtain slightly better scores. CPE [7] is typically more effective than APE in tasks such as 3D instance/semantic segmentation and object detection where a complete 3D point cloud is given. However, this result highlights the challenge of capturing position information from mask tokens which initially have the identical parameters. Our method employs RoPE [59] for relative positional embedding. One of the important aspect of RoPE [59] is that it does not have any learnable parameters. Despite this, it demonstrates superior performance compared to other approaches. Although RoPE was originally proposed in the domain of natural language processing, our experiment reveals its effectiveness in multi-object 3D scene completion.

Table 5: Ablation study of the number of MAE layers on our synthetic dataset.

#Layers	CD↓	F1↑	NC↑	Runtime↓
1	9.01	0.784	0.828	**76.4**
3	6.48	0.839	0.848	85.1
5	**5.75**	**0.850**	**0.855**	96.2

Table 6: Ablation study of U-Net architectures on HomebrewedDB dataset [28].

Architecture	CD↓	F1↑	NC↑	Runtime↓
Mink. U-Net [6]	7.26	0.788	0.743	**83.8**
OctFormer [61]	7.45	0.756	0.728	114.4
Octree U-Net [62]	**6.14**	**0.819**	**0.770**	85.1

Table 7: Comparisons of the runtime (ms). For reference, we also show Chamfer distance of visible CD_{vis} and occluded CD_{occ} regions on our synthetic dataset.

Method	3D Rep.	Resolution	CD_{vis} ↓	CD_{occ}↓	CD↓	Runtime↓
VoxFormer [34]	Dense	128^3	18.25	66.32	44.54	79.5
ShapeFormer [71]	Dense	128^3	14.61	63.33	39.50	1.8×10^4
MCC [66]	Implicit	128^3	15.39	63.41	44.37	9.1×10^3
ConvONet [48]	Dense	128^3	17.09	34.09	23.68	48.4
POCO [1]	Implicit	128^3	10.37	31.55	21.11	758.8
AICNet [31]	Dense	128^3	9.98	21.43	15.64	**24.2**
Minkowski [6]	Sparse	512^3	7.12	15.44	11.47	78.5
OCNN [63]	Sparse	512^3	3.87	12.16	9.05	80.1
Ours	Sparse	512^3	**3.29**	**9.40**	**6.48**	85.1

3D Attention algorithms. Table 4 reveals that occlusion masking yields better runtime and metrics than dense masking. Furthermore, our experiments suggest that full attention and Octree attention, both characterized by their wider receptive fields, are more effective compared to local attention algorithms such as 3D deformable self-attention (3D DSA) [34] and neighborhood attention [77].

Number of layers in 3D latent MAE. We further explore the design of 3D latent MAE in Table 5. Increasing the number of layers in 3D latent MAE improves the scene completion quality while making the runtime slower. Consequently, we select 3 layers for a good trade-off between the accuracy and runtime.

U-Net architectures. In Table 6, we investigate U-Net architectures. The key difference of Minkowski U-Net [6] is the use of a sparse tensor as an underlying data structure instead of an octree, which gives a slightly better performance than Octree U-Net [62]. OctFormer [61] proposes an octree-based window attention mechanism using the 3D Z-order curve to support a much larger kernel size than Octree U-Net. In general, a wider range of an effective receptive field helps achieve better performance. Nonetheless, OctFormer achieves a chamfer distance and F-1 score of 7.45 and 0.756, which is worse than Octree U-Net by 1.31 and 0.063 respectively. This indicates that the OctFormer's attention mechanism is less effective compared to an Octree U-Net architecture especially in the presence of latent 3D MAE, playing the similar role in the latent space.

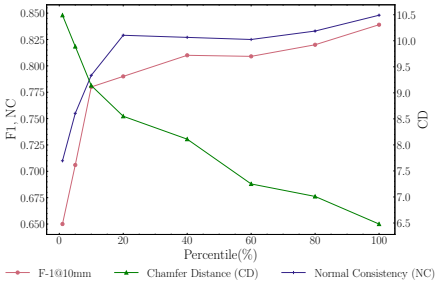

Fig. 5: Scaling of the metrics with the number of objects in a training dataset. We conduct the experiments by changing the ratio of the number of objects to 1%, 5%, 10%, 20%, 40%, 60%, 80%, and 100%.

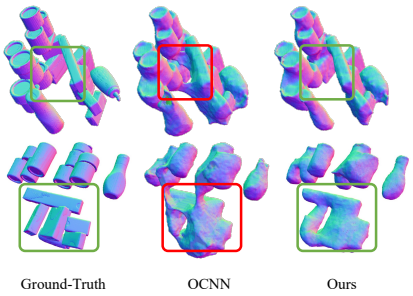

Fig. 6: Qualitative comparison of OCNN [62] and our method. Our proposed latent 3D MAE helps predict globally consistent scene completion.

Runtime analysis. Table 7 shows the runtime performance of the baselines and our method. For a fair comparison, we run inference over the 50 samples of the HOPE dataset and report the average time. For occupancy-based methods, we predict occupancy on object surfaces and occluded regions. Due to the memory-intensive nature of MCC [1]'s Transformer architecture, we run inference multiple times with the maximum chunk size of 10,000 points. Our experiments demonstrate that implicit 3D representations used in POCO [1] and MCC [66] become slower when the voxel grid resolution is higher. Further, an autoregressive Transformer adopted in ShapeFormer [71] greatly increases the runtime. Conversely, the methods which leverage sparse voxel grids (*e.g.*, MinkowskiNet [6], OCNN [63], and Ours) achieve much faster runtime thanks to efficient sparse 3D convolutions, and hierarchical pruning on predicted surfaces. Our method offers runtimes comparable to the fastest method, while implementing attention operations over the scene via latent 3D MAE, and achieving superior reconstruction.

Dataset scale analysis. To assess the importance of the large-scale 3D scene completion datasets, we train our model on splits of increasing sizes which contain 1%, 5%, 10%, 20%, 40%, 60%, 80%, and 100% of the total number of the objects in our dataset. We report metrics on the test split of our dataset. Section 5.1 shows that all the metrics have a strong correlation with respect to the number of objects. This could imply that the model benefits significantly from increased data diversity and volume, enhancing its ability to understand and complete 3D shapes. We believe that this analysis is crucial for understanding the relationship between data quantity and model performance.

5.2 Qualitative Results

Figure 7 shows the qualitative results of our method on both of the synthetic and real-world datasets from three different views. Unlike the synthetic dataset,

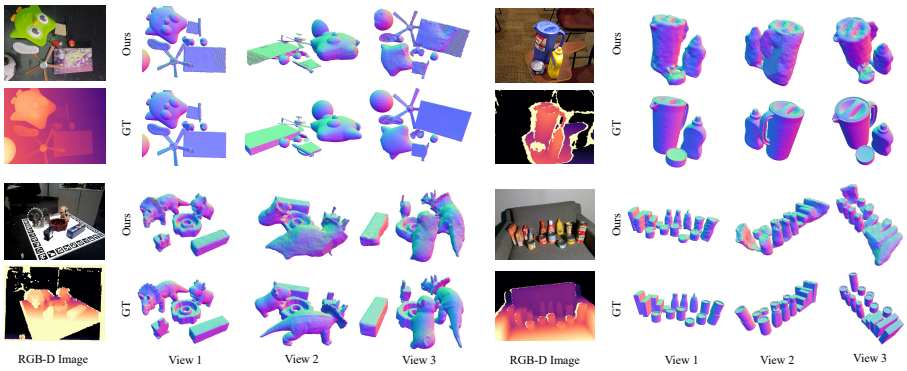

Fig. 7: Qualitative results on our synthetic dataset (**Top Left**), YCB-Video (**Top Right**), HomebrewedDB (**Bottom Left**), and HOPE (**Bottom Right**) datasets. These results demonstrate the strong generalization to the real-world images on multi-object scene completion. We choose 3 different views for better visibility.

the real-world depth measurements are more noisy and erroneous, however, we observe that our method can generate faithful and consistent 3D shapes on different types of objects. These results indicate that our model successfully learns geometric and semantic priors of real-world objects only from the synthetic data. Moreover, Figure 6 provides a comparison between our method and the second-best baseline, OCNN [62]. OCNN struggles with multi-object reasoning, resulting in unnatural artifacts improperly merging multiple objects. We believe this finding further supports that using 3D latent MAE helps capture a global context for better scene completion.

6 Conclusion and Future Work

In this paper, we present OctMAE, a hybrid architecture combining an Octree U-Net and a latent 3D MAE, for efficient and generalizable scene completion. Further, we create the first large-scale and diverse 3D scene completion dataset, which consists of 1M images rendered with 12K objects with realistic scale. Our experimental results on a wide range of the datasets demonstrate accurate zero-shot multi-object scene completion is possible with a proper choice of the network architecture and dataset, which potentially facilitates several challenging robotics tasks such as robotic manipulation and motion planning. Although our method achieves superior performance, it comes with some limitations. First, truncated objects are not reconstructed properly since depth measurements are not available. We believe we can overcome this problem by incorporating techniques for query proposal [72] and amodal segmentation [78]. The second limitation is that the semantic information of completed shapes is not predicted. Although our focus in this work is geometric scene completion, we believe it is an interesting direction to integrate a technique from an open-vocabulary segmen-

tation methods to obtain instance-level completed shapes. Third, our method does not handle uncertainty of surface prediction explicitly. In future work, we plan to extend our method to model uncertainty to improve the scene completion quality and diversity.

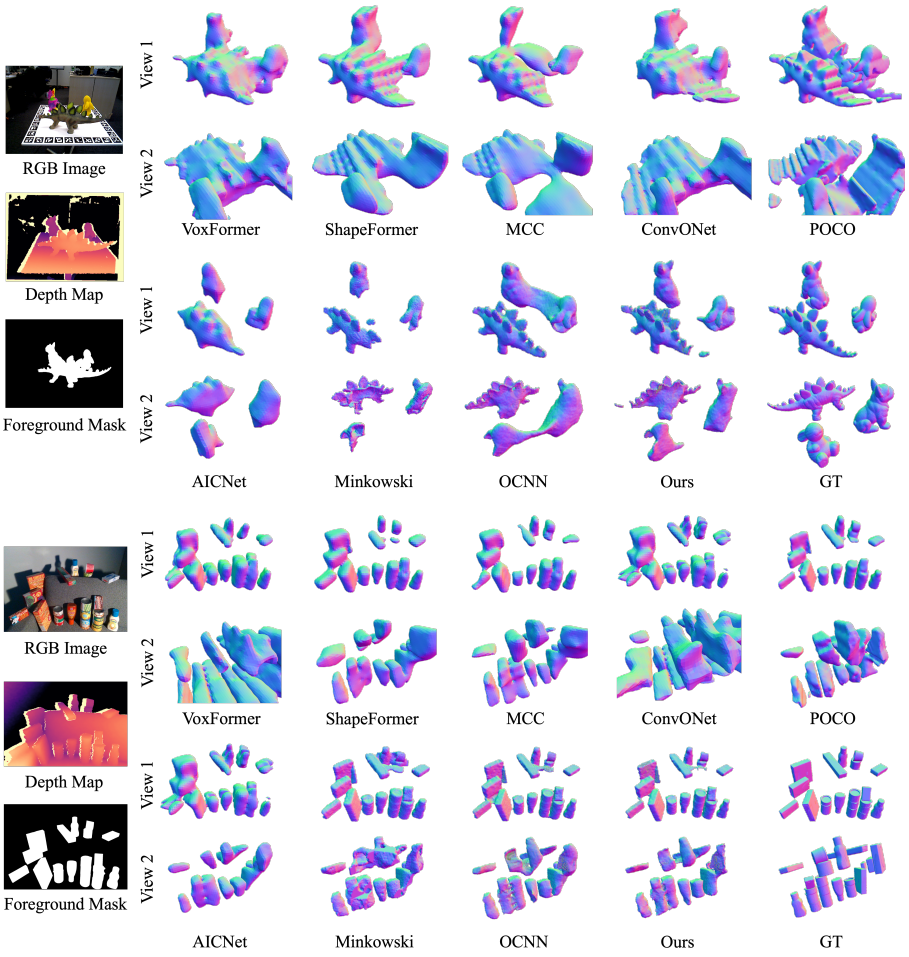

Fig. 8: Comparisons on HomebrewedDB dataset (**Top**), and HOPE (**Bottom**) datasets. For better visibility, we show the generated and ground truth shapes. The top and bottom rows show an image from near camera and back views respectively. Compared to the other methods, our method predicts accurate and consistent shapes on a challenging scene completion task for novel objects.

Acknowledgment

We thank Zubair Irshad and Jenny Nan for valuable feedback and comments. This research is supported by Toyota Research Institute.

References

1. Boulch, A., Marlet, R.: POCO: Point Convolution for Surface Reconstruction. In: CVPR (2022)
2. Bozic, A., Palafox, P., Thies, J., Dai, A., Nießner, M.: TransformerFusion: Monocular rgb scene reconstruction using transformers. In: NeurIPS (2021)
3. Chan, E.R., Nagano, K., Chan, M.A., Bergman, A.W., Park, J.J., Levy, A., Aittala, M., Mello, S.D., Karras, T., Wetzstein, G.: GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In: CoRR (2023)
4. Chen, H.X., Huang, J., Mu, T.J., Hu, S.M.: CIRCLE: Convolutional Implicit Reconstruction And Completion For Large-Scale Indoor Scene. In: ECCV (2022)
5. Cheng, Y.C., Lee, H.Y., Tulyakov, S., Schwing, A.G., Gui, L.Y.: SDFusion: Multimodal 3d shape completion, reconstruction, and generation. In: CVPR (2023)
6. Choy, C., Gwak, J., Savarese, S.: 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In: CVPR (2019)
7. Chu, X., Tian, Z., Zhang, B., Wang, X., Shen, C.: Conditional Positional Encodings for Vision Transformers. In: ICLR (2023)
8. Computer, T.: RedPajama: an Open Dataset for Training Large Language Models (2023)
9. Dai, A., Diller, C., Nießner, M.: SG-NN: Sparse generative neural networks for self-supervised scene completion of rgb-d scans. In: CVPR (2020)
10. Dai, A., Ritchie, D., Bokeloh, M., Reed, S., Sturm, J., Nießner, M.: ScanComplete: Large-Scale Scene Completion and Semantic Segmentation for 3D Scans. In: CVPR (2018)
11. Dao, T.: FlashAttention-2: Faster attention with better parallelism and work partitioning (2023)
12. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A Universe of Annotated 3D Objects. CVPR (2022)
13. Denninger, M., Winkelbauer, D., Sundermeyer, M., Boerdijk, W., Knauer, M., Strobl, K.H., Humt, M., Triebel, R.: BlenderProc2: A Procedural Pipeline for Photorealistic Rendering. Journal of Open Source Software (2023)
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL (2019)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR (2021)
16. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google Scanned Objects: A High-Quality Dataset of 3D Scanned Household Items. In: ICRA (2022)
17. Duan, Y., Zhu, H., Wang, H., Yi, L., Nevatia, R., Guibas, L.J.: Curriculum deepsdf. In: ECCV (2020)

18. Dupont, E., Kim, H., Eslami, S.M.A., Rezende, D.J., Rosenbaum, D.: From data to functa: Your data point is a function and you can treat it like one. In: ICML (2022)
19. Gao, P., Ma, T., Li, H., Dai, J., Qiao, Y.: ConvMAE: Masked Convolution Meets Masked Autoencoders. NeurIPS (2022)
20. Goldblum, M., Finzi, M., Rowan, K., Wilson, A.G.: The No Free Lunch Theorem, Kolmogorov Complexity, and the Role of Inductive Biases in Machine Learning. CoRR (2023)
21. Graham, B., Engelcke, M., van der Maaten, L.: 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks. CVPR (2018)
22. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
23. Hou, J., Dai, A., Nießner, M.: RevealNet: Seeing Behind Objects in RGB-D Scans. In: CVPR (2020)
24. Huang, J., Gojcic, Z., Atzmon, M., Litany, O., Fidler, S., Williams, F.: Neural Kernel Surface Reconstruction. In: CVPR (2023)
25. Irshad, M.Z., Zakharov, S., Ambrus, R., Kollar, T., Kira, Z., Gaidon, A.: Shapo: Implicit representations for multi-object shape, appearance, and pose optimization. In: ECCV (2022)
26. Kappler, D., Meier, F., Issac, J., Mainprice, J., Garcia Cifuentes, C., Wüthrich, M., Berenz, V., Schaal, S., Ratliff, N., Bohg, J.: Real-time Perception meets Reactive Motion Generation. RA-L (2018)
27. Karaman, S., Frazzoli, E.: Sampling-Based Algorithms for Optimal Motion Planning. Int. J. Rob. Res. (2011)
28. Kaskman, R., Zakharov, S., Shugurov, I., Ilic, S.: HomebrewedDB: RGB-D Dataset for 6D Pose Estimation of 3D Objects. ICCVW (2019)
29. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
30. Labbé, Y., Manuelli, L., Mousavian, A., Tyree, S., Birchfield, S., Tremblay, J., Carpentier, J., Aubry, M., Fox, D., Sivic, J.: MegaPose: 6d pose estimation of novel objects via render & compare. In: CoRL (2022)
31. Li, J., Han, K., Wang, P., Liu, Y., Yuan, X.: Anisotropic Convolutional Networks for 3D Semantic Scene Completion. In: CVPR (2020)
32. Li, J., Liu, Y., Gong, D., Shi, Q., Yuan, X., Zhao, C., Reid, I.: RGBD Based Dimensional Decomposition Residual Network for 3D Semantic Scene Completion. In: CVPR. pp. 7693–7702 (June 2019)
33. Li*, L.H., Zhang*, P., Zhang*, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: CVPR (2022)
34. Li, Y., Yu, Z., Choy, C., Xiao, C., Alvarez, J.M., Fidler, S., Feng, C., Anandkumar, A.: VoxFormer: Sparse Voxel Transformer for Camera-based 3D Semantic Scene Completion. In: CVPR (2023)
35. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: CVPR (2023)
36. Lin, Y., Tremblay, J., Tyree, S., Vela, P.A., Birchfield, S.: Multi-view Fusion for Multi-level Robotic Scene Understanding. In: IROS (2021)
37. Liu, L., Gu, J., Lin, K.Z., Chua, T.S., Theobalt, C.: Neural Sparse Voxel Fields. NeurIPS (2020)
38. Liu, M., Xu, C., Jin, H., Chen, L., Xu, Z., Su, H., et al.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. NeurIPS (2023)

39. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot One Image to 3D Object. In: CVPR (2023)
40. Liu, Z., Feng, Y., Black, M.J., Nowrouzezahrai, D., Paull, L., Liu, W.: MeshDiffusion: Score-based Generative 3D Mesh Modeling. In: ICLR (2023)
41. Lorensen, W.E., Cline, H.E.: Marching Cubes: A High Resolution 3D Surface Construction Algorithm. SIGGRAPH (1987)
42. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy Networks: Learning 3D Reconstruction in Function Space. In: CVPR (2019)
43. Mittal, P., Cheng, Y.C., Singh, M., Tulsiani, S.: AutoSDF: Shape Priors for 3D Completion, Reconstruction and Generation. In: CVPR (2022)
44. Mohammadi, S.S., Duarte, N.F., Dimou, D., Wang, Y., Taiana, M., Morerio, P., Dehban, A., Moreno, P., Bernardino, A., Del Bue, A., Santos-Victor, J.: 3DSGrasp: 3D Shape-Completion for Robotic Grasp. In: ICRA (2023)
45. Museth, K.: VDB: High-resolution sparse volumes with dynamic topology (2013)
46. Okumura, K., Défago, X.: Quick Multi-Robot Motion Planning by Combining Sampling and Search. In: IJCAI (2023)
47. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In: CVPR (2019)
48. Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., Geiger, A.: Convolutional Occupancy Networks. In: ECCV (2020)
49. Rabe, M.N., Staats, C.: Self-attention Does Not Need $O(n^2)$ Memory (2021)
50. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
51. Radford, A., Narasimhan, K.: Improving Language Understanding by Generative Pre-Training (2018)
52. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common Objects in 3D: Large-Scale Learning and Evaluation of Real-life 3D Category Reconstruction. In: ICCV (2021)
53. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models (2021)
54. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. NeurIPS (2022)
55. Shao, T., Yang, Y., Weng, Y., Hou, Q., Zhou, K.: H-CNN: Spatial Hashing Based CNN for 3D Shape Analysis. TVCG (2020)
56. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. In: NeurIPS (2021)
57. Shi, Z., Zhou, X., Qiu, X., Zhu, X.: Improving image captioning with better use of captions. CoRR (2020)
58. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic Scene Completion from a Single Depth Image. CVPR (2017)
59. Su, J., Lu, Y., Pan, S., Wen, B., Liu, Y.: RoFormer: Enhanced Transformer with Rotary Position Embedding. In: ICLR (2020)
60. Varley, J., DeChant, C., Richardson, A., Ruales, J., Allen, P.: Shape completion enabled robotic grasping. In: IROS (2017)
61. Wang, P.S.: OctFormer: Octree-based Transformers for 3D Point Clouds. SIGGRAPH (2023)
62. Wang, P.S., Liu, Y., Guo, Y.X., Sun, C.Y., Tong, X.: O-CNN: Octree-Based Convolutional Neural Networks for 3D Shape Analysis. SIGGRAPH (2017)

63. Wang, P.S., Liu, Y., Tong, X.: Deep Octree-based CNNs with Output-Guided Skip Connections for 3D Shape and Scene Completion. In: CVPRW (2020)
64. Watson, D., Chan, W., Martin-Brualla, R., Ho, J., Tagliasacchi, A., Norouzi, M.: Novel View Synthesis with Diffusion Models. CoRR (2022)
65. Williams, F., Gojcic, Z., Khamis, S., Zorin, D., Bruna, J., Fidler, S., Litany, O.: Neural Fields as Learnable Kernels for 3D Reconstruction. In: CVPR (2022)
66. Wu, C.Y., Johnson, J., Malik, J., Feichtenhofer, C., Gkioxari, G.: Multiview Compressive Coding for 3D Reconstruction. In: CVPR (2023)
67. Wu, X., Lao, Y., Jiang, L., Liu, X., Zhao, H.: Point transformer V2: Grouped Vector Attention and Partition-based Pooling. In: NeurIPS (2022)
68. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes (2018)
69. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated Residual Transformations for Deep Neural Networks. CVPR (2017)
70. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: ODISE: Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. CVPR (2023)
71. Yan, X., Lin, L., Mitra, N.J., Lischinski, D., Cohen-Or, D., Huang, H.: ShapeFormer: Transformer-based Shape Completion via Sparse Representation. In: CVPR (2022)
72. Yu, X., Rao, Y., Wang, Z., Liu, Z., Lu, J., Zhou, J.: PoinTr: Diverse Point Cloud Completion with Geometry-Aware Transformers. In: ICCV (2021)
73. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. CVPR (2022)
74. Zhang, D., Choi, C., Park, I., Kim, Y.M.: Probabilistic Implicit Scene Completion. In: ICLR (2022)
75. Zhang, H., Zhang, P., Hu, X., Chen, Y.C., Li, L.H., Dai, X., Wang, L., Yuan, L., Hwang, J.N., Gao, J.: GLIPv2: Unifying Localization and Vision-Language Understanding. CoRR (2022)
76. Zhang, P., Liu, W., Lei, Y., Lu, H., Yang, X.: Cascaded Context Pyramid for Full-Resolution 3D Semantic Scene Completion. In: ICCV (2019)
77. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: ICCV (2021)
78. Zhu, Y., Tian, Y., Mexatas, D., Dollár, P.: Semantic Amodal Segmentation. In: CVPR (2017)