PatchComplete: Learning Multi-Resolution Patch Priors for 3D Shape Completion on Unseen Categories

Anonymous Author(s)
Affiliation
Address
email

Abstract

While 3D shape representations enable powerful reasoning in many visual and perception applications, learning 3D shape priors tends to be constrained to the specific categories trained on, leading to an inefficient learning process, particularly for general applications with unseen categories. Thus, we propose *PatchComplete*, which learns effective shape priors based on multi-resolution local patches, which are often more general than full shapes (e.g., chairs and tables often both share legs) and thus enable geometric reasoning about unseen class categories. To learn these shared substructures, we learn multi-resolution patch priors across all train categories, which are then associated to input partial shape observations by attention across the patch priors, and finally decoded into a complete shape reconstruction. Such patch-based priors avoid overfitting to specific train categories and enable reconstruction on entirely unseen categories at test time. We demonstrate the effectiveness of our approach on synthetic ShapeNet data as well as challenging real-scanned objects from ScanNet, which include noise and clutter, improving over state of the art in novel-category shape completion by 19.3% in chamfer distance on ShapeNet, and 9.3% for ScanNet.

1 Introduction

The prevalence of commodity RGB-D sensors (e.g., Intel RealSense, Microsoft Kinect, iPhone, etc.) has enabled significant progress in 3D reconstruction, achieving impressive tracking quality [21, 14, 23, 7, 34, 11] and even large-scale reconstructed datasets [9, 3]. Unfortunately, 3D scanned reconstructions remain limited in geometric quality due to clutter, noise, and incompleteness (e.g., as seen in the objects in Figure 1). Understanding complete object structures is fundamental towards constructing effective 3D representations, that can then be used to fuel many applications in robotic perception, mixed reality, content creation, and more.

Recently, significant progress has been made in shape reconstruction, across a variety of 3D representations, including voxels [8, 12, 28], points [13, 37], meshes [33, 10], and implicit field representations [24, 17]. However, these methods tend to rely heavily on strong synthetic supervision, producing impressive reconstructions on similar train class categories but struggling to generalize to unseen categories. This leads to an expensive compute and data requirement in adapting to new objects in different scenarios, whose class categories may not necessarily have been seen during training and so must be re-trained or fine-tuned for.

In order to encourage more generalizable 3D feature learning to represent shape characteristics, we observe that while different class categories may have very different global structures, local geometric structures are often shared (e.g., a long, thin structure could represent a chair leg, a table leg, a lamp rod, etc.). We thus propose to learn a set of multi-resolution patch-based priors that captures such shared local substructures across the training set of shapes, which can be applied to shapes outside

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

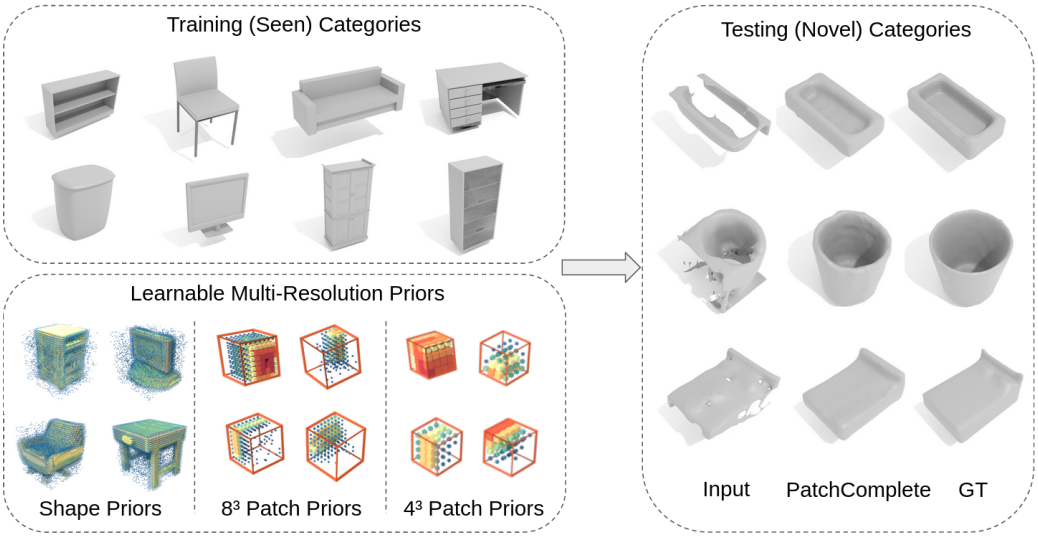

Figure 1: PatchComplete learns strong local priors for 3D shape completion by constructing multi-resolution patch priors from train shapes, which can then be applied for effective shape completion on unseen categories.

of the train set of categories. Our local patch-based priors can thus capture shared local structures, across different resolutions, that enable effective shape completion on novel class categories of not only synthetic data but also challenging real-world observations with noise and clutter.

We propose PatchComplete, which first learns patch priors for shape completion by correlating regions of observed partial inputs to the learned patch priors through an attention-based association, and decoding to reconstruct a complete shape. These patch priors are learned at different resolutions to encompass potentially different sizes of local substructures; we then learn to fuse the multi-resolution priors together to reconstruct the output complete shape. This enables learning generalizable local 3D priors that facilitate effective shape completion even for unseen categories, outperforming state of the art in synthetic and real-world observations by 19.3% and 9.3% on Chamfer Distance.

In summary, our contributions are:

- We propose generalizable 3D shape priors by learning patch-based priors that characterize shared local substructures that can be associated with input observations by cross-attention. This intermediate representation preserves structure explicitly, and can be effectively leveraged to compose complete shapes for entirely unseen categories.
- We design a multi-resolution fusion of different patch priors at various resolutions in order to effectively reconstruct a complete shape, enabling multi-resolution reasoning about the most informative learned patch priors to recover both global and local shape structures.

2 Related Work

2.1 3D Shape Reconstruction and Completion

Understanding how to reconstruct 3D shapes is an essential task for 3D machine perception. In particular, the task of shape completion to predict a complete shape from partial input observations has been studied by various works toward understanding 3D shape structures. Recently, many works have leveraged deep learning techniques to learn strong data-driven priors for shape completion, focusing on various different representations, e.g. volumetric grids [35, 12], continuous implicit functions [25, 17, 24, 6, 30], point clouds [26, 27], and meshes [10, 15]. These works tend to focus on strong synthetic supervision on a small set of train categories, achieving impressive performance on unseen shapes from train categories, but often struggling to generalize to unseen classes. We focus on learning more generalizable, local shape priors in order to effectively reconstruct complete shapes on unseen class categories.

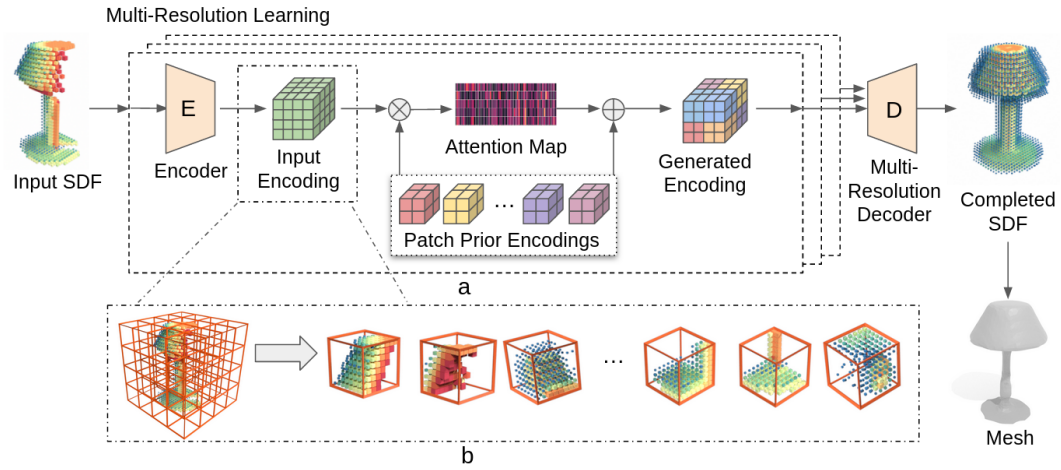

Figure 2: Overview of our approach. (a) shows our multi-resolution prior learning for shape completion. Each dotted block indicates patch prior learning for a single resolution, and the three different resolution encodings are then fused in a multi-resolution decoder that outputs a complete shape as an SDF grid. (b) illustrates our input partial observations and local input patches, from which we learn mappings to learned patch priors as to how to best compose the complete shape.

2.2 Few-Shot and Zero-Shot 3D Shape Reconstruction

Several works have been developed to tackle the challenging task of few-shot or zero-shot shape reconstruction, as observations in-the-wild often contain a wide diversity of objects. In the few-shot scenario where several examples of novel categories are available, Wallace and Hariharan [32] learn to refine a given category prior. Michalkiewicz et al. [18] further propose to learn compositional shape priors for single-view reconstruction. In the zero-shot scenario without any examples seen for novel categories, Naeem et al. [20] learn priors from seen categories to generate segmentation masks for unseen categories. Zhang et al. [38] additionally proposed to use spherical map representations to learn priors for the reconstruction of novel categories. Thai et al. [29] recently developed an approach to transfer knowledge from an RGB image for shape reconstruction. We also tackle a zero-shot shape reconstruction task, by learning a multi-resolution set of strong local shape priors to compose a reconstructed shape.

Several recent works have explored learning shape priors by leveraging a VQ-VAE backbone with autoregressive prediction to perform shape reconstruction [19, 36]. In contrast, we propose to learn multi-resolution shape priors without requiring any sequence interpretation, which enables direct applicability to real-world scan data that often contains noise and clutter. Additionally, hierarchical reconstruction has shown promising results for shape reconstruction [2, 5]. Our approach also takes a multi-resolution perspective, but learns explicit shape and local patch priors and their correlation to input partial observations for robust shape completion.

3 Method

3.1 Overview

Our method aims to learn effective 3D shape priors that enable general shape completion from partial input scan data, on novel class categories not seen during training. Key to our approach is the observation that 3D shapes often share repeated local patterns – for instance, chairs, tables, and nightstands all share a support surface, and chair or table legs can share a similar structure with lamp rods. Inspired by this, we regard a complete object as a set of substructures, where each substructure represents a local geometric region. We thus propose PatchComplete to learn such local priors and assemble them into a complete shape from a partial scan. An overview of our approach is shown in Figure. 2.

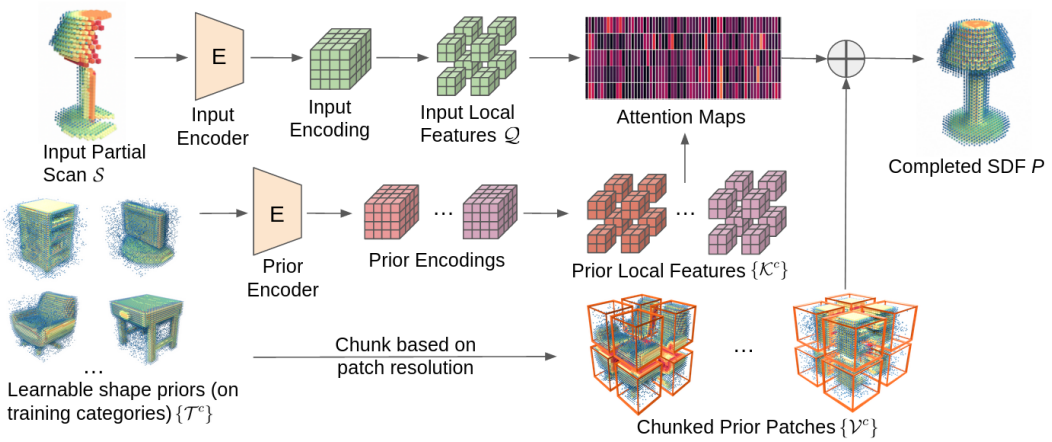

Figure 3: Network architecture for local patch-based shape prior learning under a single resolution. We learn to build mappings from local regions in a partial object scan $\mathcal{S}$ to local priors based on complete learnable shape priors $\{\mathcal{T}^c\}$. An input encoder takes the input of $\mathcal{S}$ to encode its local features (represented by patches $\mathcal{Q}$). Analogously, we adopt a prior encoder to process each prior in $\{\mathcal{T}^c\}$ to construct a local prior feature pool $\{\mathcal{K}^c\}$. In parallel, we chunk priors in $\{\mathcal{T}^c\}$ into patch volumes $\{\mathcal{V}^c\}$, which are correspondingly fused by the attention to input patches to compose the most informative patch priors: we use each incomplete local patch in $\mathcal{Q}$ to query the keys $\{\mathcal{K}^c\}$ from complete prior patches, and assemble their corresponding patch volumes $\{\mathcal{V}^c\}$ for shape completion.

3.2 Learning Local Patch-Based Shape Priors

We first learn local shape priors from ground-truth train objects. We represent both the input partial scan $\mathcal{S}$ and the ground-truth shape $\mathcal{G}$ as 3D truncated signed distance field grids of size D^3. Figure 3 illustrates the process of learning local shape priors. We learn to build mappings from local regions in incomplete input observations to local priors based on complete shapes, in order to robustly reconstruct the complete shape output.

We first aim to learn patch-based priors, which we extract from learnable global shape priors $\{\mathcal{T}^c\}$. We denote $\mathcal{G}^c$ as the set of ground-truth train objects in the c-th category. These $\{\mathcal{T}^c\}$ priors are initialized per train category c based on mean-shift clustering within shapes in $\mathcal{G}^c$. Thus $\mathcal{T}^c$ is a set of representative samples in each category, which are encoded parallel to the input scan $\mathcal{S}$.

Both encoders are analogously structured as 3D convolutional encoders which spatially downsample by a factor of R, resulting in a 3D encoding of size N^3; $N = D/R$. The input encoding (from S) are uniformly chunked into patches $\{\mathcal{Q}_i\}$, i=1,...,N^3; similarly, the encoded priors (from $\{\mathcal{T}^c\}$) are chunked into patches $\{\mathcal{K}_i^c\}$.

We then use each local part $\mathcal{Q}_i$ to query for the most informative encodings of complete local regions $\{\mathcal{K}_i^c\}$, building this mapping based on cross-attention [31]. Then for each input patch, we calculate its similarity with all local patches in representative shapes on training categories:

$$\text{Attention}(\mathcal{Q}_i, \mathcal{K}) = \text{softmax}(\frac{\mathcal{Q}_i \cdot \mathcal{K}^T}{d/2}), \mathcal{K} = \{\mathcal{K}_i^c | i = 1, ..., N^3, c = 1, ..., C\}, \tag{1}$$

where d is the dimension of the encoded vectors of $\mathcal{Q}_i, \mathcal{K}_i^c$; C is the category number. We then reconstruct complete shape patches $\{\mathcal{P}_i\}$ by:

$$\mathcal{P}_i = \text{Attention}(\mathcal{Q}_i, \mathcal{K}) \cdot \mathcal{V} \tag{2}$$

where $\mathcal{V} = \{\mathcal{V}_i^c\}$ is the set of N^3 chunks (of resolution R) from shapes in all categories $\{\mathcal{T}^c\}$, where each $\mathcal{V}_i^c$ is paired with $\mathcal{K}_i^c \in \mathcal{K}$. We can then recompose $\{\mathcal{P}_i\}$ to the predicted full shape P.

Loss. We use an ℓ_1 reconstruction loss to train the learned local patch priors. Note that $\{\mathcal{T}^c\}$ are learned along with the network weights, enabling learning the most effective global shape priors for the shape completion task.

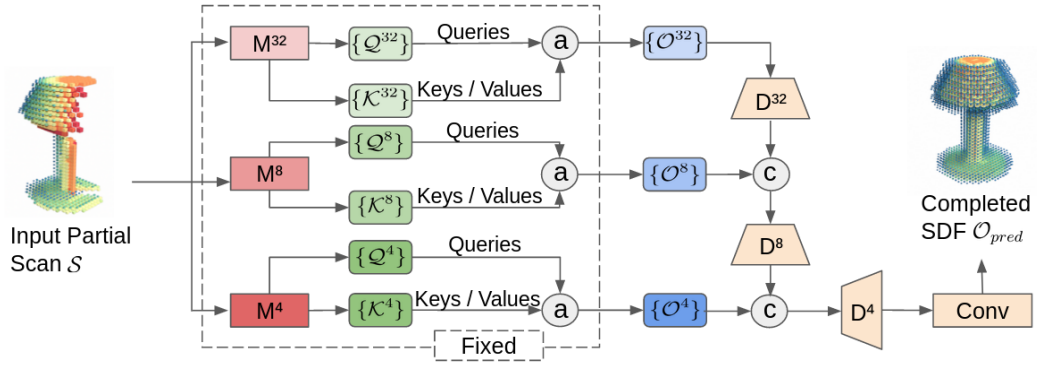

Figure 4: Network architecture for multi-resolution learning pipeline. We generate a complete shape from a partial input scan $\mathcal{S}$ by fusing input local features and learned local priors in a multi-scale fashion. We first extract the input local features $\{\mathcal{Q}^R\}$ and the learned local priors $\{\mathcal{K}^R\}$ using the prior learning model (M^R) in Section 3.2 under different resolutions (R=32, 8, 4). Then, we use attention (Equation 3) to generate intermediate features $\{\mathcal{O}^R\}$, which we then recursively fuse to decode a complete shape.

3.3 Multi-Resolution Patch Learning

In Section 3.2, we learn to complete shapes in a single patch resolution R. Since local substructures may exist at various scales, we learn patch priors under varying resolutions, which we then use to decode a complete shape. We use three patch resolutions (R=4, 8, 32), for which we learn patch priors. This results in three pairs of trained {input encoder, prior encoder} (see Figure 3). Given a partial input scan $\mathcal{S}$, each pair outputs a set of 1) input patch features $\{\mathcal{Q}_i^R\}$, and 2) prior patch features $\mathcal{K}^R$, under the resolution of R=4, 8, 32. In this section, we decode a complete shape from these multi-resolution patch priors.

Since $\mathcal{K}^R$ stores the patch priors, we use it as the (key, value) term into an attention module, where each input patch feature $\mathcal{Q}_i^R$ is used to query for the most informative prior features in $\mathcal{K}^R$ under different resolutions, from which we complete partial patches in feature spaces with a multi-scale fashion. We formulate this process by

$$\mathcal{O}_i^R = \text{Concat}\left[\mathcal{Q}_i^R, \text{Attention}(\mathcal{Q}_i^R, \mathcal{K}^R) \cdot \mathcal{K}^R\right], R = 4, 8, 32, \tag{3}$$

where we concatenate input patch feature with the attention result to compensate information loss on observable regions. This outputs $\{\mathcal{O}_i^R\}$ as the intermediate feature of the i-th patch; $i = 1, ..., N_R^3$. N_R denotes the number of patches under the resolution R; $N_R = D/R$. We recompose all generated patch features into a volume $\mathcal{O}^R = \{\mathcal{O}_i^R | i = 1, ..., N_R^3\}$. Note that the dimension of each grid feature in $\{\mathcal{Q}_i^R\}$ and $\mathcal{K}^R$ equals to d (see Eq. 1). Then $\mathcal{O}^R$ is with the dimension of $N_R^3 \times 2d$.

We then sequentially use 3D convolution decoders to upsample and concatenate $\mathcal{O}^R$ from low to high resolution to fuse all the shape features as in

$$\begin{aligned} \mathcal{O}^{R_{r+1}} &= \text{Concat}\left[\text{Upsample}(\mathcal{O}^{R_r}), \mathcal{O}^{R_{r+1}}\right], R_r|_{r=1,2,3} = 32, 8, 4, \\ \mathcal{O}_{pred} &= \text{Conv}(\text{Upsample}(\mathcal{O}^{R_3})). \end{aligned} \tag{4}$$

In Eq. 4, $\mathcal{O}^{R_r}$ has a lower resolution than $\mathcal{O}^{R_{r+1}}$. We then use deconvolution layers to upsample $\mathcal{O}^{R_r}$ to match the resolution of $\mathcal{O}^{R_{r+1}}$, and then concatenate them together. We recursively fuse $\mathcal{O}^{R_r}$ into $\mathcal{O}^{R_{r+1}}$, which produces $\mathcal{O}^{R_3}$ as the final fusion result with the resolution of $(D/4)^3$. An extra upsampling followed by a convolution layer is adopted to upsample $\mathcal{O}^{R_3}$ into our shape prediction $\mathcal{O}_{pred}$ with the dimension of D^3.

In training the feature fusion, we fix all parameters from the {input encoder, prior encoder} under the three resolutions, since they are pre-trained on Section 3.2 and provide better-learned priors and attention maps under these strong constraints. The whole pipeline for this section can be found in Figure 4.

We use an ℓ_1 loss to supervise the TSDF value regression. We weight the ℓ_1 loss to penalize false predictions based on the predicted signs as Eq. 5 shows, which represents whether this voxel grid is occupied or not. It is also used in the loss function in Section 3.2.

$$\mathcal{L} = w_{occ}\ell_1(\mathcal{O}^{occ}_{pred}, \mathcal{O}_{gt}) + w_{empty}\ell_1(\mathcal{O}^{empty}_{pred}, \mathcal{O}_{gt}) + \ell_1(\mathcal{O}^{correct}_{pred}, \mathcal{O}_{gt}). \quad (5)$$

In Eq. 5, $\mathcal{O}^{empty}_{pred}$ represents the false positive TSDF values, where the ground truth has negative signs and the prediction has positive signs, which in general indicates the missing predictions. $\mathcal{O}^{occ}_{pred}$ represents false negative TSDF values, where the ground truth has positive signs and the prediction has negative signs, which indicates extra predictions. $\mathcal{O}^{correct}_{pred}$ means those predicted TSDF values with the same signs as the ground-truth. During training, we choose the weight for false positive (w_{empty}) as 5, and the weight for false negative (w_{occ}) as 3.

3.4 Implementation Details

We train our approach on a single NVIDIA A6000 GPU, using an Adam optimizer with batch size 32 for a synthetic dataset and batch size 16 for a real-world dataset, and an initial learning rate of 0.001. We train for 80 epochs until convergence, and then decrease the learning rate by half after epoch 50. We use the same settings for learning priors and the multi-resolution decoding, which train for 4 and 15.5 hours respectively. For additional network architecture details, we refer to the supplemental material.

Note that for training for real scan data, we first pre-train on synthetic data and then fine-tune only the input encoder.

4 Experiments and Analysis

4.1 Experimental Setup

Datasets. We train and evaluate our approach on synthetic shape data from ShapeNet [4] as well as on challenging real-world scan data from ScanNet [9]. For ShapeNet data, we virtually scan the objects to create partial input scans, following [22, 12]. We use 18 categories during training, and test on 8 novel categories, resulting in 3,202/1,325 train/test models with 4 partial scans for each model.

For real data, we use real-scanned objects from ScanNet extracted by their bounding boxes, with corresponding complete target data given by Scan2CAD [1]. We use 8 categories for training, comprising 7,537 train samples, and test on 6 novel categories of 1,191 test samples.

For all experiments, objects are represented as 32^3 signed distance grids with truncation of 2.5 voxel units for ShapeNet and 3 voxel units for ScanNet. The objects in ShapeNet are normalized into the unit cube, while we keep the scaling for ScanNet objects, and save their voxel sizes separately to keep the real size information. Additionally, to train and evaluate on real data, all methods are first pre-trained on ShapeNet and then fine-tuned on ScanNet.

Baselines. We evaluate our approach against various state-of-the-art shape completion methods. We compare with state-of-the-art shape completion methods 3D-EPN [12] and IF-Nets [6], which learn effective shape completion on dense voxel grids and with implicit neural field representations, respectively, without any focus on unseen class categories. We further compare to the state-of-the-art few-shot shape reconstruction approach of Wallace and Hariharan [32] (referred to as Few-Shot) leveraging global shape priors, which we apply in our zero-shot unseen category scenario. Finally, AutoSDF [19] uses a VQ-VAE module with a transformer-based autoregressive model over latent patch priors to produce TSDF shape reconstruction.

Evaluation Metrics. To evaluate the quality of reconstructed shape geometry, we use $\mathcal{L}_1$ Chamfer Distance (CD) and Intersection over Union (IoU) between predicted and ground truth shapes. To evaluate methods that output occupancy grids, we use occupancy thresholds used by the respective methods to obtain voxel predictions, i.e. 0.4 for [32] and 0.5 for [6]. To evaluate methods that output signed distance fields, we extract the iso-surface at level zero with marching cubes [16]. 10K points are sampled on surfaces for CD calculation. Both Chamfer distance and IoU are evaluated on objects in the canonical system, and we report Chamfer Distance scaled by $\times 10^2$.

Table 1: Quantitative comparison for shape completion on synthetic ShapeNet [4] data.

	Chamfer Distance ($\times 10^2$) ↓					IoU↑				
	3D-EPN [12]	Few-Shot [32]	IF-Nets [6]	Auto-SDF [19]	Ours	3D-EPN [12]	Few-Shot [32]	IF-Nets [6]	Auto-SDF [19]	Ours
Bag	5.01	8.00	4.77	5.81	**3.94**	0.738	0.561	0.698	0.563	**0.776**
Lamp	8.07	15.10	5.70	6.57	**4.68**	0.472	0.254	0.508	0.391	**0.564**
Bathtub	4.21	7.05	4.72	5.17	**3.78**	0.579	0.457	0.550	0.410	**0.663**
Bed	5.84	10.03	5.34	6.01	**4.49**	0.584	0.396	0.607	0.446	**0.668**
Basket	7.90	8.72	**4.44**	6.70	5.15	0.540	0.406	0.502	0.398	**0.610**
Printer	5.15	9.26	5.83	7.52	**4.63**	0.736	0.567	0.705	0.499	**0.776**
Laptop	3.90	10.35	6.47	4.81	**3.77**	0.620	0.313	0.583	0.511	**0.638**
Bench	4.54	8.11	5.03	4.31	**3.70**	0.483	0.272	0.497	0.395	**0.539**
Inst-Avg	5.48 $\pm 2e^{-1}$	9.75 $\pm 9e^{-2}$	5.37 $\pm 1e^{-1}$	5.76 $\pm 3e^{-2}$	**4.23** $\pm 4e^{-2}$	0.582 $\pm 9e^{-3}$	0.386 $\pm 1e^{-3}$	0.574 $\pm 4e^{-5}$	0.446 $\pm 6e^{-3}$	**0.644** $\pm 1e^{-3}$
Cat-Avg	5.58 $\pm 2e^{-1}$	9.58 $\pm 1e^{-1}$	5.29 $\pm 1e^{-1}$	5.86 $\pm 5e^{-3}$	**4.27** $\pm 5e^{-2}$	0.594 $\pm 8e^{-3}$	0.403 $\pm 1e^{-3}$	0.581 $\pm 3e^{-4}$	0.452 $\pm 7e^{-3}$	**0.654** $\pm 1e^{-3}$

4.2 Evaluation on Synthetic Data

In Table 1, we evaluate our approach in comparison with prior arts on unseen class categories of synthetic ShapeNet [4] data. Our approach on learning attention-based correlation to learned local shape priors results in notably improved reconstruction performance, with coherent global and local structures, as shown in Figure 5. In Table 1, our work outperforms other baselines both instance-wise and category-wise. One of the key factors is that our method learns multi-scale patch information from seen categories to complete unseen categories with enough flexibility, while most of the other baselines are designed for 3D shape completion on known categories, which hardly leverage shape priors across categories.

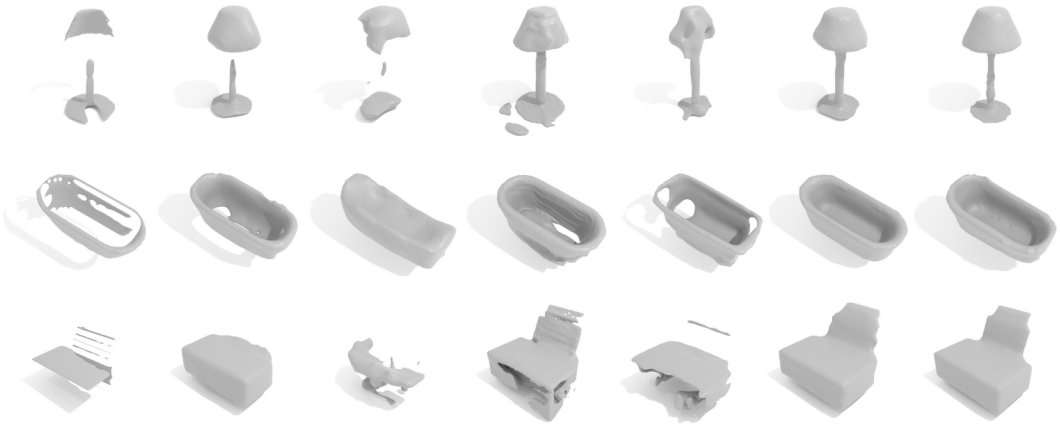

Figure 5: Qualitative comparison for shape completion on synthetic ShapeNet [4] dataset.

4.3 Evaluation on Real Scan Data

Table 2 evaluates our approach in comparison with prior arts on real scanned objects from unseen categories in ScanNet [9]. Here, input scans are not only partial but often contain noise and clutter; our multi-resolution learned priors enable more robust shape completion in this challenging scenario. Results in Figure 6 further demonstrate that our approach presents more coherent shape completion than prior methods by using cross-attention with learnable priors, which better preserves the global structures in coarse and cluttered environments.

Table 2: Quantitative comparison with state of the art on real-world ScanNet shape completion.

	Chamfer Distance ($\times 10^2$) ↓					IoU↑				
	3D-EPN [12]	Few-Shot [32]	IF-Nets [6]	Auto-SDF [19]	Ours	3D-EPN [12]	Few-Shot [32]	IF-Nets [6]	Auto-SDF [19]	Ours
Bag	8.83	9.10	8.96	9.30	**8.23**	0.537	0.449	0.442	0.487	**0.583**
Lamp	14.27	11.88	10.16	11.17	**9.42**	0.207	0.196	0.249	0.244	**0.284**
Bathtub	7.56	7.77	7.19	7.84	**6.77**	0.410	0.382	0.395	0.366	**0.480**
Bed	7.76	9.07	8.24	7.91	**7.24**	0.478	0.349	0.449	0.380	**0.484**
Basket	7.74	8.02	6.74	7.54	**6.60**	0.365	0.343	0.427	0.361	**0.455**
Printer	8.36	8.30	8.28	9.66	**6.84**	0.630	0.622	0.607	0.499	**0.705**
Inst-Avg	8.60 $\pm 2e^{-1}$	8.83 $\pm 2e^{-2}$	8.12 $\pm 7e^{-2}$	8.56 $\pm 2e^{-2}$	**7.37** $\pm 7e^{-2}$	0.441 $\pm 2e^{-3}$	0.387 $\pm 1e^{-3}$	0.426 $\pm 3e^{-3}$	0.386 $\pm 1e^{-4}$	**0.498** $\pm 9e^{-3}$
Cat-Avg	9.09 $\pm 3e^{-1}$	9.02 $\pm 8e^{-2}$	8.26 $\pm 8e^{-2}$	8.90 $\pm 2e^{-2}$	**7.49** $\pm 4e^{-2}$	0.440 $\pm 3e^{-3}$	0.386 $\pm 6e^{-3}$	0.426 $\pm 7e^{-3}$	0.389 $\pm 3e^{-4}$	**0.495** $\pm 5e^{-3}$

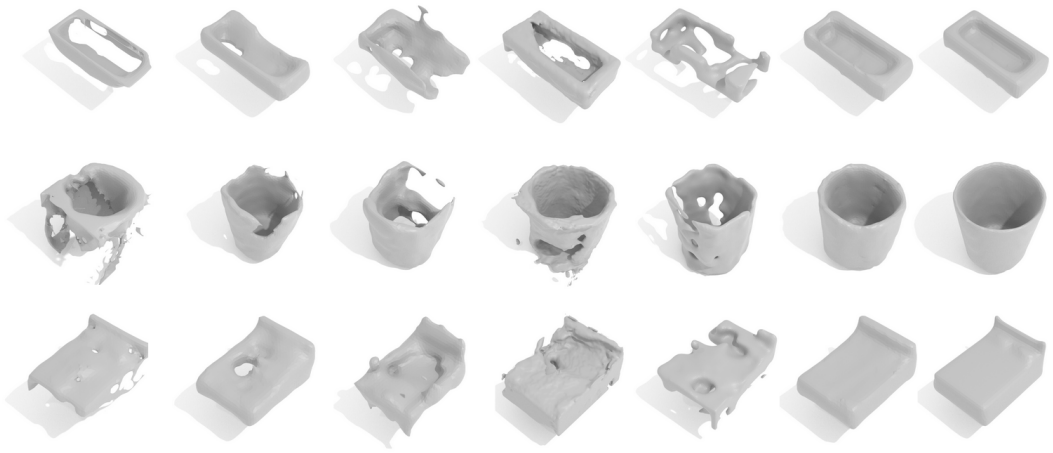

Figure 6: Shape completion on real-world ScanNet [9] object scans. Our method to learn mappings to multi-resolution learnable patch priors enables more coherent shape completion on novel categories.

4.4 Ablation Analysis

Does multi-resolution patch learning help shape completion for novel categories? In Table 3, we evaluate shape completion with each resolution in comparison with our multi-resolution approach. Learning only global shape priors (i.e., 32^3) tends to overfit to seen train categories, while the local patch resolutions can provide more generalizable priors. Combining all results in complementary feature learning for the most effective shape completion results.

Table 3: Ablation study on different patch resolutions. A multi-resolution approach gains benefits from both global and local reasoning.

	ShapeNet [4]				ScanNet [9]			
	Inst-CD↓	Cat-CD↓	Inst-IoU↑	Cat-IoU↑	Inst-CD↓	Cat-CD↓	Inst-IoU↑	Cat-IoU↑
Ours (32^3 priors only)	11.94	11.61	0.35	0.37	10.43	11.23	0.41	0.40
Ours (8^3 priors only)	4.86	4.89	0.61	0.62	7.64	7.81	0.48	0.49
Ours (4^3 priors only)	4.44	4.50	**0.64**	0.64	**7.34**	7.58	0.49	0.49
Ours	**4.23**	**4.27**	**0.64**	**0.65**	7.37	**7.49**	**0.50**	**0.50**

Table 4: Ablation study on attention used to learn input and patch prior correlations.

	ShapeNet [4]				ScanNet [9]			
	Inst-CD↓	Cat-CD↓	Inst-IoU↑	Cat-IoU↑	Inst-CD↓	Cat-CD↓	Inst-IoU↑	Cat-IoU↑
Ours (no attention)	4.90	4.98	**0.61**	0.62	7.80	8.09	**0.49**	0.48
Ours	**4.69**	**4.74**	**0.61**	**0.63**	**7.58**	**7.84**	0.48	**0.49**

Does cross-attention to learn local priors help? We evaluate our approach to learn both local priors and their correlation to input observations with cross-attention in Table 4, which shows that this enables more effective shape completion on unseen categories.

Table 5: Effect of synthetic pre-training on real-world ScanNet object completion vs. training from scratch.

	Inst-CD↓	Cat-CD↓	Inst-IoU↑	Cat-IoU↑
Scratch	7.61	7.73	**0.50**	**0.50**
Ours	**7.37**	**7.49**	**0.50**	**0.50**

What is the effect of synthetic pre-training for real scan completion? Table 5 shows the effect of synthetic pre-training for shape completion on real scanned objects. This encourages learning more robust priors to output cleaner local structures as given in the synthetic data, resulting in improved performance on real scanned objects.

Does learning the priors help completion? In Table 6, we evaluate our learnable priors in comparison with using fixed priors (by mean-shift clustering of train objects) for shape completion on ShapeNet [4]. Learned priors receive gradient information to adapt to best reconstruct the shapes, enabling improved performance over a fixed set of priors.

Table 6: Ablation on learnable priors in comparison with fixed priors on ShapeNet.

	Inst-CD↓	Cat-CD↓	Inst-IoU↑	Cat-IoU↑
Ours (fixed priors)	4.31	4.34	**0.64**	**0.65**
Ours	**4.23**	**4.27**	**0.64**	**0.65**

4.5 Limitations

While PatchComplete has presented a promising step towards learning more generalizable shape priors, various limitations remain. For instance, output shape completion is limited by the dense voxel grid resolution in representing fine-scale geometric details. Additionally, detected object bounding boxes are required for real scan data as input to independent shape completion predictions, while formulation considering other objects in the scene or an end-to-end framework could learn more effectively from global scene contextual information.

4.6 Societal Impact

This work proposes a method for 3D completion for unseen categories with less dependency on data capturing, with potential application to robotics or content creation scenarios. For instance, a robot could potentially more efficiently grasp unseen objects when estimating complete shape structure; or self-driving cars could perform obstacle avoidance by robust 3D shape estimation for a variety of objects. While potential real-world applications are promising, this requires careful consideration of personal data privacy, as well as any biases embedded in the training dataset.

5 Conclusion

We have proposed PatchComplete to learn effective local shape priors for shape completion that enables robust reconstruction for novel class categories at test time. This enables learning shared local substructures across a variety of shapes that can be mapped to local incomplete observations by cross attention, with multi-resolution fusion producing coherent geometry at global and local scales. This enables robust shape completion even for unseen categories with different global structures, across synthetic as well as challenging real-world scanned objects with noise and clutter. We believe that such robust reconstruction of real scanned objects takes an important step towards understanding 3D shapes, and hope this inspires future work in understanding real-world shapes and scene structures.

References

[1] Avetisyan, A., Dahnert, M., Dai, A., Savva, M., Chang, A.X., Nießner, M.: Scan2cad: Learning cad model alignment in rgb-d scans (2018) 6

[2] Bechtold, J., Tatarchenko, M., Fischer, V., Brox, T.: Fostering generalization in single-view 3d reconstruction by learning a hierarchy of local and global shape priors (2021) 3

[3] Chang, A.X., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D data in indoor environments. International Conference on 3D Vision (3DV) (2017) 1

[4] Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: Shapenet: An information-rich 3d model repository (2015) 6, 7, 8, 9, 12

[5] Chen, Z., Zhang, Y., Genova, K., Fanello, S., Bouaziz, S., Haene, C., Du, R., Keskin, C., Funkhouser, T., Tang, D.: Multiresolution deep implicit functions for 3d shape representation (2021) 3

[6] Chibane, J., Alldieck, T., Pons-Moll, G.: Implicit functions in feature space for 3d shape reconstruction and completion (2020) 2, 6, 7, 8

[7] Choi, S., Zhou, Q.Y., Koltun, V.: Robust reconstruction of indoor scenes. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5556–5565. IEEE (2015) 1

[8] Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: European conference on computer vision. pp. 628–644. Springer (2016) 1

[9] Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes (2017) 1, 6, 7, 8, 9, 12

[10] Dai, A., Nießner, M.: Scan2mesh: From unstructured range scans to 3d meshes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5574–5583 (2019) 1, 2

[11] Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., Theobalt, C.: Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. ACM Transactions on Graphics (TOG) **36**(3), 24 (2017) 1

[12] Dai, A., Qi, C.R., Nießner, M.: Shape completion using 3d-encoder-predictor cnns and shape synthesis (2017) 1, 2, 6, 7, 8

[13] Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 605–613 (2017) 1

[14] Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., et al.: Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In: Proceedings of the 24th annual ACM symposium on User interface software and technology. pp. 559–568. ACM (2011) 1

[15] Li, X., Liu, S., Kim, K., Mello, S.D., Jampani, V., Yang, M.H., Kautz, J.: Self-supervised single-view 3d reconstruction via semantic consistency. In: European Conference on Computer Vision. pp. 677–693. Springer (2020) 2

[16] Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. ACM siggraph computer graphics **21**(4), 163–169 (1987) 6

[17] Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space (2019) 1, 2

[18] Michalkiewicz, M., Tsogkas, S., Parisot, S., Baktashmotlagh, M., Eriksson, A., Belilovsky, E.: Learning compositional shape priors for few-shot 3d reconstruction (2021) 3

[19] Mittal, P., Cheng, Y.C., Singh, M., Tulsiani, S.: AutoSDF: Shape priors for 3d completion, reconstruction and generation. In: CVPR (2022) 3, 6, 7, 8

[20] Naeem, M.F., Örnek, E.P., Xian, Y., Gool, L.V., Tombari, F.: 3d compositional zero-shot learning with decompositional consensus (2021) 3

[21] Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: Real-time dense surface mapping and tracking. In: Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on. pp. 127–136. IEEE (2011) 1

[22] Nie, Y., Lin, Y., Han, X., Guo, S., Chang, J., Cui, S., Zhang, J., et al.: Skeleton-bridged point completion: From global inference to local adjustment. Advances in Neural Information Processing Systems **33**, 16119–16130 (2020) 6

[23] Nießner, M., Zollhöfer, M., Izadi, S., Stamminger, M.: Real-time 3d reconstruction at scale using voxel hashing. ACM Transactions on Graphics (TOG) (2013) 1

[24] Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation (2019) 1, 2

[25] Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., Geiger, A.: Convolutional occupancy networks (2020) 2

[26] Stutz, D., Geiger, A.: Learning 3d shape completion under weak supervision. International Journal of Computer Vision **128**(5), 1162–1181 (2020) 2

[27] Tang, J., Lei, J., Xu, D., Ma, F., Jia, K., Zhang, L.: Sa-convonet: Sign-agnostic optimization of convolutional occupancy networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6504–6513 (2021) 2

[28] Tatarchenko, M., Dosovitskiy, A., Brox, T.: Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In: Proceedings of the IEEE international conference on computer vision. pp. 2088–2096 (2017) 1

[29] Thai, A., Stojanov, S., Upadhya, V., Rehg, J.M.: 3d reconstruction of novel object shapes from single images (2021) 3

[30] Tretschk, E., Tewari, A., Golyanik, V., Zollhöfer, M., Stoll, C., Theobalt, C.: Patchnets: Patch-based generalizable deep implicit 3d shape representations (2021) 2

[31] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 4

[32] Wallace, B., Hariharan, B.: Few-shot generalization for single-image 3d reconstruction via priors (2019) 3, 6, 7, 8

[33] Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In: Proceedings of the European conference on computer vision (ECCV). pp. 52–67 (2018) 1

[34] Whelan, T., Leutenegger, S., Salas-Moreno, R.F., Glocker, B., Davison, A.J.: Elasticfusion: Dense slam without a pose graph. Proc. Robotics: Science and Systems, Rome, Italy (2015) 1

[35] Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes (2015) 2

[36] Yan, X., Lin, L., Mitra, N.J., Lischinski, D., Cohen-Or, D., Huang, H.: Shapeformer: Transformer-based shape completion via sparse representation. arXiv preprint arXiv:2201.10326 (2022) 3

[37] Yang, G., Huang, X., Hao, Z., Liu, M.Y., Belongie, S., Hariharan, B.: Pointflow: 3d point cloud generation with continuous normalizing flows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4541–4550 (2019) 1

[38] Zhang, X., Zhang, Z., Zhang, C., Tenenbaum, J.B., Freeman, W.T., Wu, J.: Learning to reconstruct shapes from unseen classes (2018) 3

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See abstract, Section 1 and Section 4.
   (b) Did you describe the limitations of your work? [Yes] Section 4.5
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Section 4.6
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.1 and Section 3.4, more details can be found in supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Table 1 and Table 2, more details can be found in supplemental material.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 3.4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We use and cite ShapeNet [4] and ScanNet [9].
   (b) Did you mention the license of the assets? [Yes] See the data section of the supplemental material.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See the data section of the supplemental material.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See the data section of the supplemental material.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]