Learning Local Pattern Modularization for Point Cloud Reconstruction from Unseen Classes

Chao Chen[1], Yu-Shen Liu[1]*, and Zhizhong Han[2]

[1] School of Software, Tsinghua University, Beijing, China
[2] Department of Computer Science, Wayne State University, Detroit, USA
chenchao19@mails.tsinghua.edu.cn, liuyushen@tsinghua.edu.cn, h312h@wayne.edu

Abstract. It is challenging to reconstruct 3D point clouds in unseen classes from single 2D images. Instead of object-centered coordinate system, current methods generalized global priors learned in seen classes to reconstruct 3D shapes from unseen classes in viewer-centered coordinate system. However, the reconstruction accuracy and interpretability are still eager to get improved. To resolve this issue, we introduce to learn local pattern modularization for reconstructing 3D shapes in unseen classes, which achieves both good generalization ability and high reconstruction accuracy. Our insight is to learn a local prior which is class-agnostic and easy to generalize in object-centered coordinate system. Specifically, the local prior is learned via a process of learning and customizing local pattern modularization in seen classes. During this process, we first learn a set of patterns in local regions, which is the basis in the object-centered coordinate system to represent an arbitrary region on shapes across different classes. Then, we modularize each region on an initially reconstructed shape using the learned local patterns. Based on that, we customize the local pattern modularization using the input image by refining the reconstruction with more details. Our method enables to reconstruct high fidelity point clouds from unseen classes in object-centered coordinate system without requiring a large number of patterns or any additional information, such as segmentation supervision or camera poses. Our experimental results under widely used benchmarks show that our method achieves the state-of-the-art reconstruction accuracy for shapes from unseen classes. The code is available at https://github.com/chenchao15/Unseen.

1 Introduction

It is challenging and vital to reconstruct point clouds from unseen classes. A widely used strategy [8, 15, 18, 44, 49] for point cloud reconstruction is to learn

* The corresponding author is Yu-Shen Liu. This work was supported by National Key R&D Program of China (2022YFC3800600), the National Natural Science Foundation of China (62272263, 62072268), and in part by Tsinghua-Kuaishou Institute of Future Media Data.

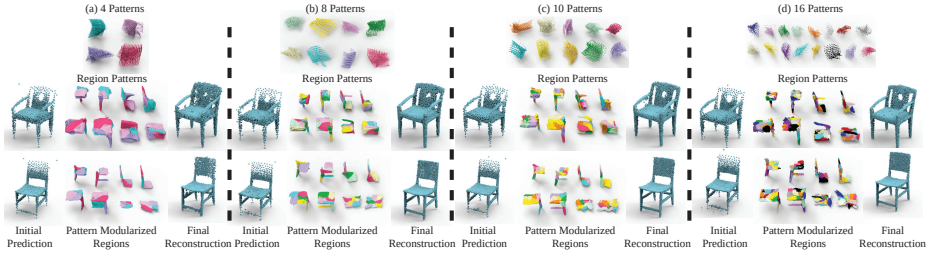

Fig. 1: We learn to reconstruct shapes from unseen classes by learning a local class-agnostic prior with region patterns, such as (a) 4 patterns, (b) 8 patterns, (c) 10 patterns, or (d) 16 patterns. Using region patterns, we modularize each region of the initial reconstruction. Based on that, we obtain our final reconstruction by customizing these pattern modularized regions.

a global prior from 2D images to 3D point clouds in object-centered coordinate system. The global prior is the key for good reconstruction accuracy, since shapes from the same classes are well aligned in the object-centered coordinate system during training. However, the global prior can not generalize well to infer shapes from unseen classes.

Some methods learn global priors in viewer-centered coordinate system which achieve better generalization to unseen classes, where the ground truth 3D shape is rotated to match the pose of the object in the input image. These methods can reasonably generalize to unseen classes by requiring additional information to support the generalization, such as camera parameters [1, 55, 62] or category shape prior [54]. However, they still struggle to improve reconstruction accuracy for unseen classes due to the large geometry variation across classes and various camera poses. More recent methods [1, 2, 37–40, 52, 60] represent 3D shapes as implicit functions to increase reconstruction accuracy. However, these methods require a large number of queries as training samples for each shape, and lack of interpretability due to the limited capability of implicit functions for representing open surfaces of parts.

To improve accuracy and interpretability, we propose to learn local pattern modularization for reconstructing 3D shapes from unseen classes. Our insight is to learn a local prior in object-centered coordinate system which can not only generalize well to unseen classes but also reconstruct high fidelity point clouds. Our idea comes from the observation that shapes from different classes may share some similar local structures, which makes it feasible to learn a local prior that is class-agnostic. Therefore, rather than a global prior, we learn a local prior from 3D local regions which can be generalized better to unseen classes. Specifically, we aim to learn a set of patterns in local regions, as demonstrated in Fig. 1, which can be used as the basis in the object-centered coordinate system to represent each region on shapes across different classes. From an input image, we first predict an initial shape reconstruction, and then modularize each region using the learned region patterns. Based on that, we learn to customize these pattern

modularized regions according to the input image by refining them with more details using a learned modularization shift. Our method enables to reconstruct high fidelity point clouds from unseen classes in object-centered coordinate system without requiring a large number of patterns or any additional information, such as segmentation supervision or camera poses. Our experimental results under widely used benchmarks show that our method achieves the state-of-the-art reconstruction accuracy for shapes from unseen classes. Our contributions are listed below.

- We introduce to learn local pattern modularization for 3D shape reconstruction from unseen classes. By further customizing the modularization, we obtain a local prior which gets better generalized to unseen classes than current global priors.
- We justify the feasibility of class-agnostic local prior in the object-centered coordinate system, which significantly improves the reconstruction accuracy in point cloud reconstruction from unseen classes.
- Our method achieves the state-of-the-art reconstruction accuracy of point clouds in both seen and unseen classes under the widely used benchmarks.

2 Related Work

Deep learning based 3D shape reconstruction has made a big progress with different 3D representations including voxel grids [13,53,61], triangle meshes [11, 21,22,34], point clouds [3,18–20,24,27,35,36,47,58], and implicit functions [4,7,9, 10,12,16,23,30,41,42,46,63,64]. The widely used strategy aims to leverage deep learning models to learn a global prior for shape reconstruction from 2D images. In the following, we focus on reviewing studies for point clouds reconstruction.

Supervised Point Clouds Reconstruction from Seen Classes. PointNet [47] is a pioneer work in point clouds understanding. Supervised methods learn to reconstruct point clouds from pairs of 2D image and its corresponding point cloud. With an encoder for 2D image understanding, Fan et al. [15] built up an encoder-decoder architecture with various shortcuts to reconstruct the point clouds. By fusing multiple depth and silhouette images generated from different view angles, Soltani et al. [49] reconstruct dense point clouds using a 2D neural network. Nguyen et al [44] tried to deform a random point cloud to the object shape with image feature blending to increase the point cloud reconstruction accuracy. Toward dense point cloud reconstruction with texture, Hu et al. [26] reformulated the reconstruction as an object coordinate map prediction and shape completion problem. AtlasNet [18] generates a point cloud as multiple 3D patches which are transformed from a set of 2D sampled points.

Unsupervised Point Clouds Reconstruction from Seen Classes. Without ground truth point clouds as supervision, unsupervised methods learn to reconstruct point clouds using various differentiable renders to compare the reconstructed point clouds and ground truth 2D images. Lin et al. [33] introduced

a pseudo-renderer to model the visibility using pooling in the dense points projection. Other rendering based methods [27, 32, 43, 57] leveraged surface splatting [57], Gaussian functions in 3D space [27] or on 2D images [32,43] to rasterize point clouds. CapNet [32] introduced a loss to match rendered pixels and pixels on ground truth silhouette images. Without pixel-wise interpolation, visibility handling, or shading in rendering, DRWR [19] directly inferred losses to adjust each 3D point from pixel values and its 2D projection.

3D Shape Reconstruction from Unseen Classes. The studies mentioned above only learn a global prior for the reconstruction of point clouds from classes that have been seen in the training. However, these learned prior is hard to be generalized to reconstruct point clouds from unseen classes. To learn more generalized global prior, GenRe [62] disentangled geometric projections from shape reconstruction, where depth prediction and spherical map inpainting are used for class-agnostic reconstruction. With a provided category shape prior, Wallace et al. [54] introduced few-shot 3D shape generation by category agnostic refinement of the provided category-specific prior. Similarly, GSIR [55] jointly learned interpretation and reconstruction to capture class-agnostic prior to recover 3D structures as cuboids. Recent work [1, 2, 52, 60] employed implicit functions for 3D reconstruction from unseen classes. These methods extend the potentials of generalization for unseen classes shown in some local implicit function based methods [5, 17, 28, 50]. However, these methods require a large number of queries as training samples for each shape, and lack of interpretability due to the incapability of implicit functions for representing open surfaces of parts. Instead, we use point clouds to interpret the reconstruction with much fewer points for each shape. Recent large visual model [31, 45] aims to learn reconstruction on a large scale of classes.

Different from these methods, our method learns a local prior for point clouds reconstruction without requiring camera parameters or category shape prior, which is much more generalizable to unseen classes. Moreover, the customization of pattern modularized regions also enables us to reconstruct point clouds in object-centered coordinate system, which achieves much higher accuracy.

3 Method

Overview. Our framework is demonstrated in Fig. 2. We aim to reconstruct a point cloud F from an input image I, where F is from a class that is not seen during training. We represent point clouds involved in our network in the object-centered coordinate system.

We first reconstruct an initial shape prediction S from image I using an encoder and decoder network. The 2D encoder extracts the information of I as a latent code f_I, which is further used to generate the initial shape prediction S by a shape decoder. Here, we leverage a shape constraint L_{Shape} to make the predicted S plausible.

Then, we split the initial shape prediction S into regions $\{R_m, m \in [1, M]\}$ to reduce the bias on seen classes during training, since regions across different

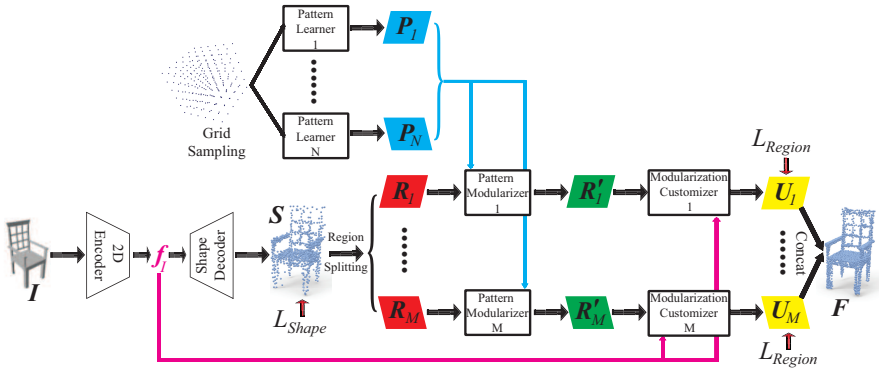

Fig. 2: The demonstration of our method. We aim to reconstruct a point clouds F from input image I, where F may come from classes that are not seen during training.

classes may share similar local structures. $\{R_m\}$ is further used to learn the region patterns $\{P_n, n \in [1, N]\}$. We use $\{P_n\}$ as the basis to represent various local regions across different classes in the object-centered coordinate system, which is one key to improve the generalization ability. We learn each region pattern P_n by transforming a grid sampling using a pattern learner. We use all region patterns $\{P_n\}$ to modularize each region R_m in a pattern modularizer, so that each region can be represented based on the same set of patterns $\{P_n\}$, which results in a pattern modularized region R'_m. Learning local pattern modularization is our first step to learn a local prior for unseen classes.

We further learn to customize each pattern modularized region R'_m in a modularization customizer. Since R'_m only represents the structure of regions but without geometry details, we introduce to leverage the input image to provide geometry details, which is another key to improve the generalization ability. Our insight here is that getting images involved in part generation would further achieve class-agnostic reconstruction. The modularization customizer customizes R'_m into a pattern customized region U_m according to the latent code f_I of input image I. This aims to push the modularization customizer to generate regions U_m that fits f_I better without a bias on classes. We push the modularization customizer to produce a set of pattern customized regions $\{U_m, m \in [1, M]\}$ which form the final shape reconstruction F by concatenation. We further add a region constraint L_{Region} to $\{U_m\}$ to supervise the customization procedure.

Finally, we train our network to capture a local prior by minimizing a loss function combining L_{Shape} and L_{Region},

$$L = L_{Region} + \alpha L_{Shape}, \tag{1}$$

where α is a balance weight and we will elaborate on L_{Shape} and L_{Region} in the following.

Initial Shape Prediction. We start from learning a mapping from input image I to a shape $S \in \mathbb{R}^{S\times 3}$. The mapping produces an intermediate representation

as a latent code $f_I \in \mathbb{R}^{1\times H}$ to bridge the image and shape space. We aim to capture a weak global prior to make the initial shape prediction S plausible, which helps our network to have a good start without relying on specific classes. We leverage a Chamfer Distance (CD) to generate a plausible S below,

$$L_{Shape} = \sum_{g\in G} \min_{s\in S} ||s-g||_2 + \sum_{s\in S} \min_{g\in G} ||s-g||_2, \tag{2}$$

where $G \in \mathbb{R}^{G\times 3}$ is the ground truth point clouds and we leverage a small weight $\alpha = 0.1$ in front of L_{Shape} in Eq. 1 to keep the global prior weak and not biased on seen classes during training.

Region Splitting. We split initial shape prediction S into regions $\{R_m \in \mathbb{R}^{R\times 3}, m \in [1, M]\}$ to learn the region patterns in the object-centered coordinate system. During training, we determine the range of each region R_m by voxelizing the bounding box of ground truth G, such that $G = \{G_m, m \in [1, M]\}$. We split each edge of the bounding box into $M^{1/3}$ segments, and regard the points on S which are located in the same voxel of G_m as R_m. While we get R_m during test by directly voxelizing the bounding box of the initial shape prediction S. Note that we keep the number of points in each region R_m the same by padding zero points for more convenient manipulation in network.

Pattern Learner. We learn region patterns $\{P_n \in \mathbb{R}^{P\times 3}, n \in [1, N]\}$ in local region coordinate system by translating each region to the origin, as defined in Eq. 3. This centering makes structures in local regions comparable to each other, which also helps region patterns easily learn more reasonable common structures.

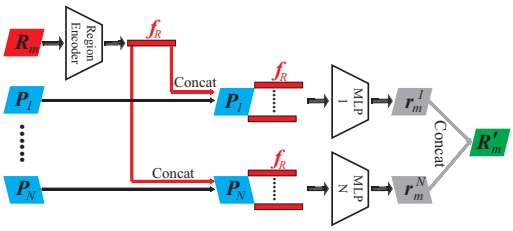

Fig. 3: The architecture of pattern modularizer.

$$R_m \leftarrow R_m - c_m, \text{where } c_m = mean(R_m). \tag{3}$$

We use N pattern learners to learn $\{P_n\}$. Each learner transforms a set of points sampled on the voxel grid into a region pattern. We share the similar idea of AtlasNet [14, 18] to generate a pattern with strong neighboring relationship. However, we translate the same set of sampled points to make each pattern learner have a different start, which results in more discriminative region patterns. All region patterns $\{P_n\}$ are involved in the following pattern modularization process.

Pattern Modularizer. We modularize each region R_m using all the region patterns $\{P_n\}$ in pattern modularizer. We push the network to represent regions from different classes using the same set of region patterns $\{P_n\}$, which reduces the bias on seen classes during training. As shown in Fig. 3, we first leverage a region encoder to map R_m as a feature f_R. Then, we concatenate $f_R \in \mathbb{R}^{1\times E}$ to each point of P_n to form an intermediate representation with a dimensionality of $P \times (3 + E)$ which is further transformed into a modularization $r_m^n \in \mathbb{R}^{P\times 3}$.

Finally, we concatenate all modularization from different patterns into one pattern modularized region $R'_m \in \mathbb{R}^{NP\times 3}$. Note that we also conduct this pattern modularization procedure in local region coordinate system, so we translate the pattern modularized regions R'_m back to object-centered coordinate system by reversing the centering procedure defined in Eq. 3 below,

$$R'_m \leftarrow R'_m + c_m. \tag{4}$$

Modularization Customizer. Based on the pattern modularized region R'_m, we further customize it using the input image I in a modularization customizer. Our purpose is to get more detailed geometry from I since R'_m merely covers a coarse structure of local regions. Our solution is to push the network to learn how to adjust a region using the content in the image accordingly, which further increases the generalization ability of the local prior. We demonstrate modularization customizer in Fig. 4. We leverage the idea of ResNet [25] to predict the modularization shift $t_m \in \mathbb{R}^{NP\times 3}$ for each R'_m. We concatenate the latent code f_I of the image I to each point of R'_m, which forms an intermediate representation with a dimensionality of $NP \times (3+H)$. This intermediate representation is further transformed into a modularization shift t_m by an MLP. Finally, we got the pattern customized region U_m below,

$$U_m = R'_m + t_m. \tag{5}$$

We reconstruct the final point cloud $F \in \mathbb{R}^{F\times 3}$ by concatenating all pattern customized regions $\{U_m, m \in [1, M]\}$ together. Note that we remove the points on each U_m that have the same indexes of zero points padded to R_m to reduce the redundance. To regulate the pattern customized region U_m in a specific region, we add a local shape constraint to minimize the CD distance between each U_m and the corresponding GT region G_m,

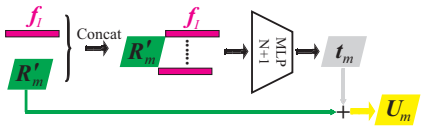

Fig. 4: The architecture of modularization customizer.

$$L_{Region} = \frac{1}{M}\sum_{m=1}^{m=M}\sum_{g\in G_m}\min_{u\in U_m}||u-g||_2 + \sum_{u\in U_m}\min_{g\in G_m}||u-g||_2. \tag{6}$$

4 Experiments and Anaylysis

We evaluate our performance by comparing our method with the state-of-the-art ones in point cloud reconstruction from seen classes and unseen classes.

4.1 Setup

Details. To highlight the effectiveness of our idea, we leverage a simple neural network in our experiments. We use a network introduced in Differentiable Point Clouds [27] as 2D encoder and shape decoder in Fig. 2. Each one of the N pattern learners is an MLP with 3 fully connected layers. In pattern modularizer in Fig. 3, the region encoder is a fully connected layer, and each one of the N MLP has 4 fully connected layers, while the MLP is formed by 3 fully connected layers in modularization customizer in Fig. 4.

We reconstruct initial shape prediction S and final reconstruction F as $S = F = 2048$ points. We split S into $M = 8$ regions. Each region is represented by $N = 8$ patterns, and each pattern is formed by $P = 256$ points. The feature f_I of the input image I is $H = 1024$ dimensional, while the feature f_R of region R_m is $E = 64$ dimensional.

Dataset and Metric. For fair comparison with the state-of-the-art methods, we conduct experiments using ShapeNet [6] and Pixel3D [51] under different experiment conditions. In numerical comparison, we will elaborate on the experiment conditions including the classes used during training and test and the number of points used in evaluation. Moreover, we employ L1-CD defined in Eq. 2 and IoU to evaluate the results. The results of L1-CD are produced by comparing our reconstruction and the ground truth with the same number of points. Our results of IoU are produced using voxel grids obtained by the method introduced in DPC [27] at a resolution of 32^3 which keeps the same as others.

	Methods	Airplane	Bench	Cabinet	Car	Chair	Display	Lamp	Speaker	Rifle	Sofa	Table	Phone	Vessel	Mean
CD	R2N2	0.227	0.194	0.217	0.213	0.270	0.605	0.778	0.318	0.183	0.229	0.239	0.195	0.238	0.278
	PSGN	0.137	0.181	0.215	0.169	0.247	0.284	0.314	0.316	0.134	0.224	0.222	0.161	0.188	0.215
	Pix2Mesh	0.187	0.201	0.196	0.180	0.265	0.239	0.308	0.285	0.164	0.212	0.218	0.149	0.212	0.216
	AtlasNet	0.104	0.138	0.175	0.141	0.209	0.198	0.305	0.245	0.115	0.177	0.190	0.128	0.151	0.175
	OccNet	0.134	0.150	0.153	0.149	0.206	0.258	0.368	0.266	0.143	0.181	0.182	0.127	0.201	0.194
	3D43D	0.096	0.112	0.119	0.122	0.193	0.166	0.561	0.229	0.248	0.125	0.146	0.107	0.175	0.184
	GraphX	0.024	0.037	0.039	0.033	0.047	0.050	0.048	0.054	0.026	0.057	0.051	0.024	0.037	0.041
	SDT	0.042	0.034	0.049	0.029	**0.036**	0.047	0.062	0.064	0.054	**0.041**	**0.033**	0.032	0.038	0.039
	Ours	**0.019**	**0.032**	**0.037**	**0.027**	0.040	**0.046**	**0.043**	**0.046**	**0.018**	0.049	0.044	**0.020**	**0.033**	**0.035**
IoU	R2N2	0.561	0.527	0.772	0.836	0.550	0.565	0.421	0.717	0.600	0.706	0.580	0.754	0.610	0.631
	PSGN	0.601	0.550	0.771	0.831	0.544	0.552	0.462	0.737	0.604	0.708	0.606	0.749	0.611	0.640
	GAL	0.685	0.709	0.772	0.737	0.700	0.804	0.670	0.698	0.715	0.739	0.714	0.773	0.675	0.712
	GraphX	0.791	0.746	0.770	0.821	0.704	0.765	0.573	0.715	0.765	0.786	0.688	**0.848**	0.772	0.750
	Ours	**0.802**	**0.765**	**0.808**	**0.841**	**0.715**	**0.812**	**0.679**	**0.746**	**0.780**	**0.790**	**0.732**	0.844	**0.783**	**0.776**

Table 1: Accuracy of reconstruction with 2048 points under ShapeNet for seen classes in terms of L1-CD and IoU.

4.2 Reconstruction from Seen Classes

Numerical Evaluation. We first evaluate our method under all 13 seen classes in ShapeNet dataset. We train our model using the training set of all 13 classes, while testing the trained model using the test set from the 13 seen classes. We

compare our method with the latest methods designed for different 3D representations, including voxel based method R2N2 [13], mesh based methods Pix2Mesh [56], point cloud based methods PSGN [15], AtlasNet [18], GraphX [44], and SDT [26], and implicit function based method OccNet [41] and 3D43D [1]. We report our numerical comparison under each one of 13 classes in Table 1. The comparison demonstrates that our method outperforms other methods in shape reconstruction. Similarly, our IoU comparison with R2N2 [13], PSGN [15], GAL [29], and GraphX [44] in Table 1 also shows that our method can reveal more accurate structures in reconstructions.

Additionally, we compare with some methods that are just trained under a subset of ShapeNet dataset which includes Chair, Car, Plane, Table and Motorcycle. To highlight our advantage, we train our model only using 3 classes including Chair, Car, and Plane, but test our model under all the 5 classes, which keeps the same as others. For fair comparison, we down sample our reconstruction results to 1024 points to compare it with the ground truth, which keeps the same as our compared methods.

Fig. 5: The visual comparison under seen classes in ShapeNet.

We report our average accuracy over the 5 classes by comparing with viewer-centered methods and object-centered methods for different 3D representations in the "Seen" column in Table 2. The viewer-centered methods, including DRC [53], MarrNet [59], GenRe [62], GSIR [55], reconstruct shapes in camera coordinate system, which require camera poses to align ground truth shapes to the images. This makes it hard to train neural network to converge to high accurate reconstructions, but the network will be more generalized to unseen classes [62]. While object-centered methods, including IMNet [12], OccNet [41], DeepSDF [46], AtlasNet [18], DRWR [19], can reconstruct more accurate shapes in canonical coordinate system. Our method not only achieves the best per-

formance among all object-centered methods even we are using much less seen classes during training, but also generalized better to unseen classes. Note that we reproduce the results of DRWR by training it under the same 3 classes as ours using its code.

Visual Comparison. We visually compare our method with the state-of-the-art in Fig. 5. We can see that our method can reveal more accurate geometry than others, where we also show our baseline reconstruction as "CD" which is obtained by training 2D encoder and shape decoder merely using the L_{Shape} loss.

4.3 Reconstruction from Unseen Classes

Evaluation in ShapeNet. We first evaluate our trained model which produces the seen results under ShapeNet in Table 2 under 4 unseen classes including Bench, Sofa, Bed, and Vessel.

We report our average reconstruction accuracy with 1024 points compared to the ground truth point clouds in the "Unseen" column in Table 2, where we down sample our reconstruction with 2048 points to 1024 points. The comparison shows that our method can significantly outperform the other methods which learn a global prior for shape reconstruction from unseen classes. Moreover, our method also shows much better generalization ability than GSIR [55] which aims to generalize a learned global prior. The superior over GSIR [55] justifies that our idea of generalizing local prior of reconstruction is more promising.

Method		Seen	Unseen
Viewer-Centered	DRC	0.0970	0.1270
	MarrNet	0.0810	0.1160
	GenRe	0.0680	0.1080
	GSIR	0.0680	0.0990
Object-Centered	IMNet	0.0550	0.1190
	OccNet	0.0600	0.1280
	DeepSDF	0.0530	0.1150
	AtlasNet	0.0630	0.1260
	DRWR	0.0536	0.0715
	Ours	**0.0527**	**0.0540**

Table 2: L1-CD accuracy of reconstruction with 1024 points.

Method		Bench	Vessel	Rifle	Sofa	Table	Phone	Cabinet	Speaker	Lamp	Display	Mean
Viewer-Centered	DRC	0.120	0.109	0.121	0.107	0.129	0.132	0.142	0.141	0.131	0.156	0.129
	MarrNet	0.107	0.094	0.125	0.090	0.122	0.117	0.125	0.123	0.144	0.149	0.120
	Multi-View	0.092	0.092	0.102	0.085	0.105	0.110	0.119	0.117	0.142	0.142	0.111
	GenRe	0.089	0.092	0.112	0.082	0.096	0.107	0.116	0.115	0.124	0.130	0.106
Object-Centered	DRC	0.112	0.100	0.104	0.108	0.133	0.199	0.168	0.164	0.145	0.188	0.142
	AtlasNet	0.102	0.092	0.088	0.098	0.130	0.146	0.149	0.158	0.131	0.173	0.127
	GraphX	0.111	0.065	0.119	0.098	0.138	0.120	0.113	0.111	0.134	0.114	0.112
	DRWR	0.075	0.059	0.104	0.070	0.100	0.094	0.088	0.086	0.102	0.097	0.088
	CD	0.110	0.084	0.121	0.122	0.114	0.136	0.126	0.122	0.143	0.160	0.124
	Ours	**0.054**	**0.046**	**0.046**	**0.058**	**0.070**	**0.061**	**0.071**	**0.072**	**0.089**	**0.077**	**0.064**

Table 3: L1-CD accuracy of reconstruction with 1024 points for unseen classes.

Then, we report our numerical comparison under more unseen classes in ShapeNet. In this experiment, we also use our model trained under Chair, Plane, Car, in Table 2, while testing under 10 unseen classes shown in Table 3. We

Method	Bench	Vessel	Rifle	Sofa	Table	Phone	Cabinet	Speaker	Lamp	Display	Mean
3D43D	0.357	0.521	0.707	0.421	0.583	0.996	0.529	0.744	1.997	1.389	0.824
SDFNet	0.133	0.209	0.199	0.306	0.288	0.434	0.241	0.374	0.554	0.487	0.323
HPN	0.079	0.071	0.070	0.144	0.148	0.064	0.114	0.110	0.147	0.163	0.111
Point-e	0.084	0.155	0.103	0.100	0.135	0.207	0.102	0.104	0.195	0.112	0.130
Ours	**0.049**	**0.042**	**0.042**	**0.051**	**0.064**	**0.054**	**0.062**	**0.063**	**0.082**	**0.070**	**0.058**

Table 4: L1-CD accuracy of reconstruction with 2048 points for unseen classes.

conduct the comparison with viewer-centered methods including MarrNet [59], Multi-View [48], GenRe [62], and object-centered methods including DRC [53], AtlasNet [18], GraphX [44], DRWR [19].

The comparison shows that our learned local prior for shape reconstruction can generalize to unseen classes better than the methods learning global prior including AtlasNet, GraphX, DRWR. Although our model is trained using ground truth point clouds in object-centered coordinate system, it still generalizes to unseen classes with higher reconstruction accuracy than viewer-centered methods including MarrNet, Multi-View, and GenRe. In addition, we also highlight our learned local prior by comparing our final reconstruction F with the initial shape prediction S obtained by merely using L_{Shape} in Eq. (2). Our significant improvement over the results of "CD" demonstrates that learning local prior is very helpful for the reconstruction generalization to unseen classes.

Fig. 6: The comparison under unseen classes in ShapeNet.

Under the same condition, we report the comparison with methods learning implicit functions, including the viewer-centered method 3D43D [1], SDFNet [52], and HPN [2], and Point-E [45] for 2048 point reconstruction. The comparison in Table 4 shows that our method significantly outperforms these methods, although they require lots of queries sampled in 3D space to learn implicit functions, which is more detailed supervision than our surface points.

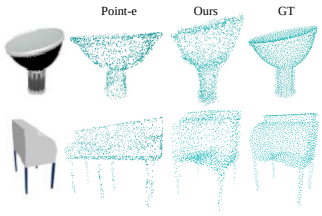

Fig. 7: The comparison under unseen classes with Point-e.

Evaluation in Pixel3D. Finally, we evaluate our generalization performance in Pixel3D dataset.

We leverage our model trained under Chair, Plane, and Car in ShapeNet in Table 2 to reconstruct shapes from 5 unseen classes in Pixel3D dataset, including Bed, Bookcase, Desk, Sofa, and Wardrobe. We compare object-centered methods for point clouds reconstruction including Atlasnet [18], GraphX [44], viewer-centered methods from unseen classes including GenRe [62], GSIR [55], and also our initial shape prediction ("CD") with merely L_{Shape} as loss. The comparison in Table 5 also demonstrates our superior over the state-of-the-art.

Visual Comparison. Our visual comparison with the state-of-the-art under unseen classes in ShapeNet is shown in Fig. 6 and Fig. 8. It demonstrates that the compared methods do not generalize well to unseen classes to reconstruct plausible shapes, such as the baseline CD, AtlasNet, GraphX, and DRWR, while GenRe that learns a global prior generalizes to unseen classes with low accuracy in viewer-centered coordinate system. Fig. 8 shows our superiority over Point-E [45] which does not perform well on images with occlusion. Our method can leverage the learned local prior to reconstruct more plausible shapes in higher accuracy in object-centered coordinate system. We also conduct a visual comparison under Pixel3D in Fig. 8, which also demonstrates our significant improvements over others. Moreover, we also show more shape reconstructions from unseen classes in Fig. 9 under ShapeNet and Pixel3D.

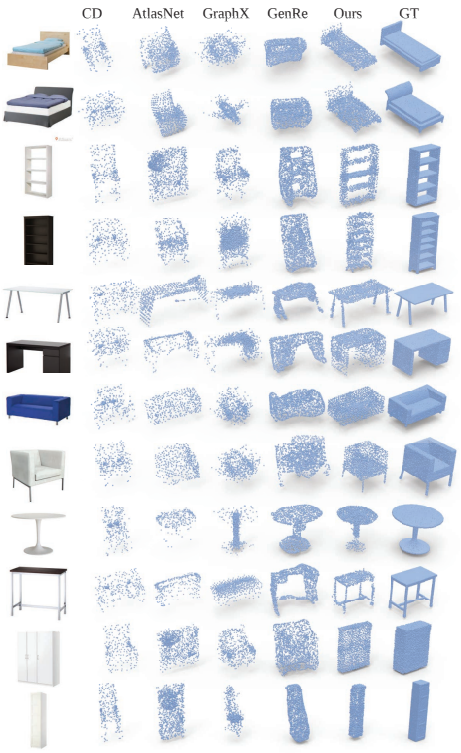

Fig. 8: The visual comparison under unseen classes in Pixel3D.

4.4 Analysis

We conduct experiments under Bench class in ShapeNet, we reconstruct point clouds from seen classes with 2048 points.

Ablation Studies. We conduct ablation studies to justify the effectiveness of the elements in our model. We first highlight our local prior. We only use the 2D encoder and point decoder to minimize L_{Shape} in training, and report the result as "No local" in Table 6. The degenerated results indicate that the local prior is important to improve the reconstruction accuracy. Similarly, we explore

Method	Bed	Bookcase	Desk	Sofa	Wardrobe
GraphX	0.141	0.122	0.132	0.094	0.116
AtlasNet	0.115	0.137	0.124	0.096	0.119
GenRe	0.111	0.101	0.107	0.085	0.111
GSIR	0.107	0.095	0.100	0.083	0.103
CD	0.165	0.102	0.163	0.104	0.132
Ours	**0.085**	**0.094**	**0.089**	**0.074**	**0.067**

Table 5: L1-CD accuracy of reconstruction with 2048 points for unseen classes under Pixel3D.

Fig. 9: More reconstruction from unseen classes under (a) ShapeNet and (b) Pixel3D.

the effectiveness of pattern modularization by removing the pattern learner and pattern modularization, and the effectiveness of modularization shift by removing the customization procedure, respectively.

The results of "No patterns" and "No shift" degenerates, which indicates that the network requires region patterns and their adjustment to reconstruct various regions on different shapes. We also evaluate the effect of L_{Region} by replacing it into another L_{Shape}. The result of "No L_{Region}" justifies that L_{Region} is important for the detailed local structures. Although we aim to learn a local prior, the weak global prior captured by initial shape reconstruction S is also helpful to provide the network a good start, as shown by the degenerated results of "No L_{Shape}".

No local	No patterns	No shift	No L_{Region}	No L_{Shape}	Ours
0.054	0.049	0.041	0.038	0.048	**0.032**

Table 6: Ablation studies in terms of L1-CD.

Sampled Points. We learn region patterns by transforming points sampled from voxel grids, since the sampled points occupy all the space where we hold local regions in local coordinate system. We compare these sampled points with the ones sampled on a 2D plane which is introduced to reconstruct 3D patches [14, 18]. Comparison in Table 7 shows that sampling on 2D plane is harder to be transformed to represent 3D local structures in voxel grids.

	Sampling on 2D plane	Sampling in 3D voxel
L1CD	0.034	**0.032**

Table 7: Sampling effect.

Region Number M. We explore the effect of region number M by trying different region number candidates including $\{1, 8, 27\}$. The comparison in Table 8 demonstrates that it is hard to capture structures in local regions if the regions are too large ("1") or too small ("27"), both of which results in reconstructions with low accuracy.

M	1	8	27
L1CD	0.035	**0.032**	0.049

Table 8: Region number effect.

Pattern Number N. We also report the effect of pattern number N by reconstructing point clouds using $\{2, 4, 8, 16\}$ patterns. The comparison in Table 10 shows that it is adequate to use $N = 8$ patterns to represent local regions with $M = 8$. Since we have modularization customizer to further

	1 pattern,1 region	8 patterns,8 regions	GraphX
Bench	0.038	**0.032**	0.037
Plane	0.121	**0.114**	0.121
Car	0.100	**0.090**	0.095

Table 9: Comparison with one region in terms of L1-CD.

adjust the pattern modularized regions, our model also does not require a large number of region patterns.

Moreover, we also conduct an experiment to evaluate our generalization ability with only one region (whole shape) and one pattern under seen Bench (training) and unseen Plane and Car (testing). Our method shows much better performance for unseen class reconstruction even with one pattern in Table 9.

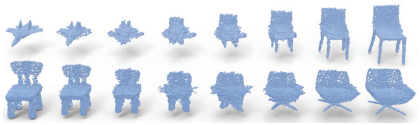

Fig. 10: The interpolated shapes.

Learned Latent Space. We visualize the learned latent space by reconstructing interpolated shapes from uniformly interpolated latent codes between two point clouds. We use the feature f_I of input image to represent each point cloud. The plausible interpolated shapes in Fig. 10 demonstrate the semantic meaning of the learned latent space.

N	2	4	8	16
L1CD	0.036	0.034	**0.032**	0.035

Table 10: Pattern number effect.

Pattern Modularization. We visualize the interpolation with or without pattern modularization. As shown in Fig. 11, It will be hard to learn semantic and meaningful space Without pattern modularization. . We can see that the poorly interpolated shapes without pattern modularization do not show a smooth transition.

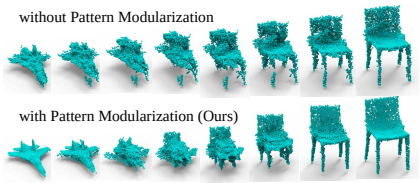

Fig. 11: Pattern modularization effect.

Visualization. We visualize initial shape prediction S, final reconstruction F, region patterns P_n, and pattern modularized regions R'_m in Fig. 1. We can see that S produces a coarse shape of the reconstruction, based on which we reconstruct a more accurate F using the learned local prior. All region patterns P_n are involved in modularizing each region R'_m, as color shown, where each pattern represents some structures in the local region and further gets customized to better fit the geometry of a region on F.

5 Conclusion

We introduce to reconstruct point clouds from unseen classes by learning local pattern modularization. Our local prior captured by learning and customizing local pattern modularization in seen classes can be effectively generalized to unseen classes in object-centered coordinate system, which leads to much higher reconstruction accuracy. Moreover, our method significantly improves the interpretability of reconstruction from unseen classes using our learned region patterns. We justify the idea of reconstructing regions using only few patterns without requiring any additional information. Our experimental results achieve the state-of-the-art under the widely used benchmarks.

References

1. Bautista, M.A., Talbott, W., Zhai, S., Srivastava, N., Susskind, J.M.: On the generalization of learning-based 3D reconstruction. In: IEEE Winter Conference on Applications of Computer Vision. pp. 2179–2188 (2021) 2, 4, 9, 11
2. Bechtold, J., Maxim, T., Volker, F., Thomas, B.: Fostering generalization in single-view 3D reconstruction by learning a hierarchy of local and global shape priors. In: IEEE Conference on Computer Vision and Pattern Recognition (2021) 2, 4, 11
3. Bednarik, J., Parashar, S., Gundogdu, E., Salzmann, Mathieu andFua, P.: Shape reconstruction by learning differentiable surface representations. In: IEEE Conference on Computer Vision and Pattern Recognition (2020) 3
4. Ben-Shabat, Y., Hewa Koneputugodage, C., Gould, S.: Digs: Divergence guided shape implicit neural representation for unoriented point clouds. In: IEEE Conference on Computer Vision and Pattern Recognition (2022) 3
5. Chabra, R., Lenssen, J.E., Ilg, E., Schmidt, T., Straub, J., Lovegrove, S., Newcombe, R.A.: Deep local shapes: Learning local SDF priors for detailed 3D reconstruction. In: European Conference on Computer Vision. vol. 12374, pp. 608–625 (2020) 4
6. Chang, A.X., Funkhouser, T.A., Guibas, L.J., 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. CoRR **abs/1512.03012** (2015) 8
7. Chen, C., Han, Z., Liu, Y.S.: Unsupervised inference of signed distance functions from single sparse point clouds without learning priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 3
8. Chen, C., Han, Z., Liu, Y.S., Zwicker, M.: Unsupervised learning of fine structure generation for 3D point clouds by 2D projection matching. In: IEEE International Conference on Computer Vision (2021) 1
9. Chen, C., Liu, Y.S., Han, Z.: Latent partition implicit with surface codes for 3d representation. In: European Conference on Computer Vision (2022) 3
10. Chen, C., Liu, Y.S., Han, Z.: Gridpull: Towards scalability in learning implicit representations from 3d point clouds. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2023) 3
11. Chen, W., Gao, J., Ling, H., Smith, E., Lehtinen, J., Jacobson, A., Fidler, S.: Learning to predict 3D objects with an interpolation-based differentiable renderer. In: Advances In Neural Information Processing Systems (2019) 3
12. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. IEEE Conference on Computer Vision and Pattern Recognition (2019) 3, 9
13. 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 (2016) 3, 9
14. Deprelle, T., Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M.: Learning elementary structures for 3D shape generation and matching. In: Advances in Neural Information Processing Systems. pp. 7433–7443 (2019) 6, 13
15. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3D object reconstruction from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2463–2471 (2017) 1, 3, 9
16. Feng, W., Li, J., Cai, H., Luo, X., Zhang, J.: Neural points: Point cloud representation with neural fields for arbitrary upsampling. In: IEEE Conference on Computer Vision and Pattern Recognition (2022) 3

17. Genova, K., Cole, F., Sud, A., Sarna, A., Funkhouser, T.: Local deep implicit functions for 3D shape. In: IEEE Conference on Computer Vision and Pattern Recognition (June 2020) 4
18. Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M.: A papier-mâché approach to learning 3D surface generation. In: IEEE Conference on Computer Vision and Pattern Recognition (2018) 1, 3, 6, 9, 11, 12, 13
19. Han, Z., Chen, C., Liu, Y.S., Zwicker, M.: DRWR: A differentiable renderer without rendering for unsupervised 3D structure learning from silhouette images. In: International Conference on Machine Learning (2020) 3, 4, 9, 11
20. Han, Z., Chen, C., Liu, Y.S., Zwicker, M.: ShapeCaptioner: Generative caption network for 3D shapes by learning a mapping from parts detected in multiple views to sentences. In: ACM International Conference on Multimedia (2020) 3
21. Han, Z., Liu, Z., Han, J., Vong, C.M., Bu, S., Chen, C.: Mesh convolutional restricted boltzmann machines for unsupervised learning of features with structure preservation on 3D meshes. IEEE Transactions on Neural Network and Learning Systems **28**(10), 2268 – 2281 (2017) 3
22. Han, Z., Liu, Z., Han, J., Vong, C.M., Bu, S., Li, X.: Unsupervised 3D local feature learning by circle convolutional restricted boltzmann machine. IEEE Transactions on Image Processing **25**(11), 5331–5344 (2016) 3
23. Han, Z., Qiao, G., Liu, Y.S., Zwicker, M.: SeqXY2SeqZ: Structure learning for 3D shapes by sequentially predicting 1D occupancy segments from 2D coordinates. In: European Conference on Computer Vision (2020) 3
24. Han, Z., Wang, X., Liu, Y.S., Zwicker, M.: Multi-angle point cloud-vae:unsupervised feature learning for 3D point clouds from multiple angles by joint self-reconstruction and half-to-half prediction. In: IEEE International Conference on Computer Vision (2019) 3
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016) 7
26. Hu, T., Lin, G., Han, Z., Zwicker, M.: Learning to generate dense point clouds with textures on multiple categories. In: IEEE Winter Conference on Applications of Computer Vision. pp. 2169–2178 (2021) 3, 9
27. Insafutdinov, E., Dosovitskiy, A.: Unsupervised learning of shape and pose with differentiable point clouds. In: Advances in Neural Information Processing Systems. pp. 2807–2817 (2018) 3, 4, 8
28. Jiang, C., Sud, A., Makadia, A., Huang, J., Nießner, M., Funkhouser, T.: Local implicit grid representations for 3D scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (2020) 4
29. Jiang, L., Shi, S., Qi, X., Jia, J.: GAL: geometric adversarial loss for single-view 3D-object reconstruction. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) European Conference on Computer Vision. vol. 11212, pp. 820–834 (2018) 9
30. Jiang, Y., Ji, D., Han, Z., Zwicker, M.: SDFDiff: Differentiable rendering of signed distance fields for 3D shape optimization. In: IEEE Conference on Computer Vision and Pattern Recognition (2020) 3
31. Jun, H., Nichol, A.: Shap-E: Generating conditional 3D implicit functions. arXiv preprint arXiv:2305.02463 (2023) 4
32. L., N.K., Mandikal, P., Agarwal, M., Babu, R.V.: CAPNet: Continuous approximation projection for 3D point cloud reconstruction using 2D supervision. AAAI (2019) 4

33. Lin, C.H., Kong, C., Lucey, S.: Learning efficient point cloud generation for dense 3D object reconstruction. In: AAAI (2018) 3
34. Liu, S., Li, T., Chen, W., Li, H.: Soft rasterizer: A differentiable renderer for image-based 3D reasoning. The IEEE International Conference on Computer Vision (Oct 2019) 3
35. Liu, X., Han, Z., Liu, Y.S., Zwicker, M.: Point2Sequence: Learning the shape representation of 3D point clouds with an attention-based sequence to sequence network. In: AAAI. pp. 8778–8785 (2019) 3
36. Liu, X., Han, Z., Xin, W., Liu, Y.S., Zwicker, M.: L2G auto-encoder: Understanding point clouds by local-to-global reconstruction with hierarchical self-attention. In: ACM International Conference on Multimedia (2019) 3
37. Ma, B., Han, Z., Liu, Y.S., Zwicker, M.: Neural-Pull: Learning signed distance functions from point clouds by learning to pull space onto surfaces. In: International Conference on Machine Learning (2021) 2
38. Ma, B., Liu, Y.S., Han, Z.: Learning signed distance functions from noisy 3D point clouds via noise to noise mapping. In: International Conference on Machine Learning (ICML) (2023) 2
39. Ma, B., Liu, Y.S., Zwicker, M., Han, Z.: Reconstructing surfaces for sparse point clouds with on-surface priors. In: IEEE Conference on Computer Vision and Pattern Recognition (2022) 2
40. Ma, B., Liu, Y.S., Zwicker, M., Han, Z.: Surface reconstruction from point clouds by learning predictive context priors. In: IEEE Conference on Computer Vision and Pattern Recognition (2022) 2
41. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3D reconstruction in function space. In: IEEE Conference on Computer Vision and Pattern Recognition (2019) 3, 9
42. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision (2020) 3
43. Navaneet, K.L., Mandikal, P., Jampani, V., Babu, R.V.: DIFFER: Moving beyond 3D reconstruction with differentiable feature rendering. In: CVPR Workshops (2019) 4
44. Nguyen, D., Choi, S., Kim, W., Lee, S.: Graphx-convolution for point cloud deformation in 2D-to-3D conversion. In: IEEE International Conference on Computer Vision. pp. 8627–8636 (2019) 1, 3, 9, 11, 12
45. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022) 4, 11, 12
46. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: DeepSDF: Learning continuous signed distance functions for shape representation. In: IEEE Conference on Computer Vision and Pattern Recognition (2019) 3, 9
47. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep learning on point sets for 3D classification and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (2017) 3
48. Shin, D., Fowlkes, C.C., Hoiem, D.: Pixels, voxels, and views: A study of shape representations for single view 3D object shape prediction. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3061–3069 (2018) 11
49. Soltani, A.A., Huang, H., Wu, J., Kulkarni, T.D., Tenenbaum, J.B.: Synthesizing 3D shapes via modeling multi-view depth maps and silhouettes with deep generative networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2511–2519 (2017) 1, 3

50. Songyou Peng, Michael Niemeyer, L.M.M.P.A.G.: Convolutional occupancy networks. In: European Conference on Computer Vision (2020) 4
51. Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J.B., Freeman, W.T.: Pix3D: Dataset and methods for single-image 3D shape modeling. In: IEEE Conference on Computer Vision and Pattern Recognition (2018) 8
52. Thai, A., Stojanov, S., Upadhya, V., Rehg, J.M.: 3D reconstruction of novel object shapes from single images. In: International Conference on 3D Vision (2021) 2, 4, 11
53. Tulsiani, S., Zhou, T., Efros, A.A., Malik, J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 209–217 (2017) 3, 9, 11
54. Wallace, B., Hariharan, B.: Few-shot generalization for single-image 3D reconstruction via priors. In: IEEE International Conference on Computer Vision. pp. 3817–3826 2, 4
55. Wang, J., Fang, Z.: GSIR: generalizable 3D shape interpretation and reconstruction. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) European Conference on Computer Vision. vol. 12358, pp. 498–514 (2020) 2, 4, 9, 10, 12
56. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.: Pixel2Mesh: Generating 3D mesh models from single RGB images. In: European Conference on Computer Vision. pp. 55–71 (2018) 9
57. Wang, Y., Felice, S., Shihao, W., Cengiz, Ö., Olga, S.: Differentiable surface splatting for point-based geometry processing. ACM Transactions on Graphics **38**(6) (2019) 4
58. Wen, X., Xiang, P., Han, Z., Cao, Y.P., Wan, P., Zheng, W., Liu, Y.S.: PMP-Net++: Point cloud completion by transformer-enhanced multi-step point moving paths. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2022) 3
59. Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, B., Tenenbaum, J.: MarrNet: 3D shape reconstruction via 2.5D sketches. In: Advances in Neural Information Processing Systems. pp. 540–550 (2017) 9, 11
60. Xiang, Y., Chibane, J., Bhatnagar, B.L., Schiele, B., Akata, Z., Pons-Moll, G.: Any-shot gin: Generalizing implicit networks for reconstructing novel classes. In: International Conference on 3D Vision (2022) 2, 4
61. Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: Learning single-view 3D object reconstruction without 3D supervision. In: Advances in Neural Information Processing Systems, pp. 1696–1704 (2016) 3
62. Zhang, X., Zhang, Z., Zhang, C., Tenenbaum, J., Freeman, B., Wu, J.: Learning to reconstruct shapes from unseen classes. In: Advances in Neural Information Processing Systems, pp. 2257–2268 (2018) 2, 4, 9, 11, 12
63. Zhou, J., Ma, B., Li, S., Liu, Y.S., Fang, Y., Han, Z.: CAP-UDF: Learning unsigned distance functions progressively from raw point clouds with consistency-aware field optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–18 (2024) 3
64. Zhou, J., Ma, B., Liu, Y.S., Fang, Y., Han, Z.: Learning consistency-aware unsigned distance functions progressively from raw point clouds. In: Advances in Neural Information Processing Systems (2022) 3