Any-Shot GIN: Generalizing Implicit Networks for Reconstructing Novel Classes

Yongqin Xian[1,2] Julian Chibane[2,3] Bharat Lal Bhatnagar[2,3] Bernt Schiele[2]
Zeynep Akata[2,3,4] Gerard Pons-Moll[2,3]

[1] ETH Zurich [2] Max Planck Institute for Informatics
[3] University of Tübingen [4] Max Planck Institute for Intelligent System

Abstract

We address the task of estimating the 3D shapes of novel shape classes from a single RGB image. Prior works are either limited to reconstructing known training classes or are unable to reconstruct high-quality shapes. To solve those issues, we propose Generalizing Implicit Networks (GIN) which decomposes 3D reconstruction into 1.) front-back depth estimation followed by differentiable depth voxelization, and 2.) implicit shape completion with 3D features. The key insight is that the depth estimation network learns local class-agnostic shape priors, allowing us to generalize to novel classes, while our implicit shape completion network is able to predict accurate shapes with rich details by learning implicit surfaces in 3D voxel space. We conduct extensive experiments on a large-scale benchmark using 55 classes of ShapeNet and real images of Pix3D. We qualitatively and quantitatively show that the proposed GIN significantly outperforms the state of the art on both seen and novel shape classes for single-image 3D reconstruction. We also illustrate that our GIN can be further improved by using only few-shot depth supervision from novel classes.

1. Introduction

Humans have the remarkable ability to recognize and imagine the 3D shapes of an arbitrary object presented in a single RGB image, even when the object has not been seen before. Such generalization ability is important because new object categories could appear in the real world. The goal of this work is to develop a single RGB-image 3D reconstruction method that generalizes well to unseen shapes within and beyond the training classes.

Single-image 3D reconstruction has been revolutionized by neural implicit models [24, 35, 27, 50] because they are not limited by resolution like earlier voxel-based methods. The idea is to represent the 3D shape with implicit networks which classify 3D points as being inside or outside the surfaces, based on the image. While results on seen classes are remarkable, specially in terms of achieved details, a natural question is whether they generalize to novel classes.

Our experiments show that generalization of those methods is poor. Most neural implicit models encode the full image into a single global latent vector [24, 27], thereby loosing geometric structure crucial for sharing information across classes (e.g., man-made shapes look locally similar and have similar attributes). The single global vector makes those models behave more like a classification into a set of known templates [39], which completely fails for novel classes. Another class of models make 3D predictions based on pixel-aligned features [51], but they do not reason about shapes in 3D, which makes learning about symmetries and common object shapes hard. There exist works that complete 3D shapes by reasoning in 3D [8, 28], but they require 3D pointclouds as input. Motivated by David Marr's [23] studies of perception, earlier models [45, 54] have proposed to split the reconstruction process into 2.5D sketch and 3D completion to improve generalization, but these models are based on voxels, and hence limited by resolution.

In this paper we propose GIN, a novel neural implicit model which generalizes well on novel classes. Inspired by early work [23, 45, 54], we also predict depth as an intermediate representation, but we couple it with a powerful 3D reasoning network based on neural fields. Specifically, GIN consists of two differentiable steps: depth prediction, and 3D implicit reconstruction based on partial depth pointclouds. Since depth is a local prediction task, the information across classes can be effectively leveraged – intuitively, each pixel counts as one training data point. In the second step, we propose to extract local voxel features by unprojecting depth into a 3D voxel grid and applying 3D convolutions, obtaining a multi-scale feature volume. The occupancy for each 3D point is then predicted based on deep features extracted at continuous locations of the volume. This allows GIN to reason about object shape globally and locally, while being invariant to image texture and appearance. Such built in invariance allows us to learn from fewer examples, and thereby generalizes better than SOTA, as evidenced by our exhaustive experiments.

This paper makes the following contributions. First, we

[1] Yongqin Xian is currently with ETHZ. Majority of this work was done at MPI Informatics. Project page: https://virtualhumans.mpi-inf.mpg.de/gin/

propose a hybrid method, named Generalizing Implicit Networks (GIN), consisting of two cascaded modules which sequentially predict front-back depth maps and estimate shapes with implicit surfaces which reason in 3D. Combining those modules is new because applying those components in isolation (prior works) does not generalize well to novel shape classes. Secondly, we quantitatively and qualitatively demonstrate that the proposed GIN significantly outperforms the SOTA on seen classes as well as novel classes on ShapeNet. Despite trained only on synthetic renderings, our GIN shows strong generalization ability on real images of Pix3D. We also illustrate that our GIN is able to further improve novel classes using only few-shot depth supervision, which motivates our title "Any-Shot GIN". Finally, we shed new light into the single image 3D reconstruction problem by conducting extensive ablations and analysis such as comparing model components and the choice of coordinate system, and evaluating the importance of the depth versus 3D shape reasoning.

2. Related work

Here we describe related works in three groups: 3D reconstruction, reconstructing novel shape classes and zero/few-shot learning for 2D recognition.

3D reconstruction and completion. Voxel is probably the most popular shape representation due to its simplicity [29, 36, 11, 46, 47, 44]. However, they were limited by the output resolution due to the memory requirement. Another important set of works use meshes as the output representation [4, 15, 14, 17, 43, 30]. Unfortunately, most of them are limited to generate meshes with simple topology [43], therefore do not generalize well to novel classes.

Recent advances in 3D reconstruction have been achieved by learning neural fields to represent object surfaces [27, 7, 24, 51, 8, 16, 33, 34, 9, 53]. The pioneer works [24, 27, 7] encode the input image as a global feature vector and learn a decoder to predict the occupancy or signed distance of query points. This, however, only represents surfaces of closed shapes, and was generalized to include color via the NeRF [26] representation and to represent also open-surfaces via the NDF [10] representation. Moreover, local shape features have been incorporated into the input encoding to better capture surface details [51, 8, 33, 34, 16, 55, 21, 9, 52]. Specifically, we build on the feature encoding of [8]: incomplete 3D shapes are encoded into rich, multi-scale feature fields and decoded via an MLP into occupancy. Our work also utilizes implicit representation but differs from existing works in three aspects: (1) we address a different problem to reconstruct novel shape classes from single RGB image, (2) we explicitly use depth maps as an intermediate representation, (3) we extract local features in 3D voxel space, capturing rich surfaces details in 3D.

Reconstructing novel shape classes. Despite its importance, there are only a few prior works in single-image 3D reconstruction of novel shape classes. GenRe [54] was the first approach to reconstruct novel shapes classes, consisting of three cascaded networks, i.e., depth estimation network, spherical map inpainting network and voxel refinement network. Built upon GenRe, GSIR [42] improves its shape representation with cuboids [18]. Both GenRe [54] and GSIR [42] are voxel-based methods which are limited by the output resolution. In contrast, our method uses implicit surfaces which allow to generate high-resolution shapes. Recently, SDFNet [40] proposes to incorporate depth maps into an implicit shape networks. While SDFNet [40] encodes the input depth map as a global vector, we extract local shape features to capture surface details. Although Ray-ONet [3] learns local features in image feature space, our method lifts depth maps into 3D and extract local features in voxel feature space.

Zero/few-shot learning for 2D recognition. Most of zero/few-shot works address 2D recognition tasks where the goal is to predict the class labels of novel classes given an input image [1, 48, 6, 13]. In contrast, our main focus is to reconstruct 3D shapes of novel classes from a RGB image. In general, zero-shot recognition relies on semantic embeddings, e.g., annotated attributes [20] or word embeddings [25] to transfer knowledge from seen to novel classes. On the other hand, few-shot recognition explores meta-learning [13, 31] and metric learning [41, 6] to learn novel classes efficiently.

3. Generalizing Implicit Networks (GIN)

We identify three key problems in existing works on single image based 3D reconstruction. First, works using voxel shape representation [54, 45], are restricted by output resolution and typically produce lower quality details. Second, implicit surfaces based methods alleviate this problem but directly regressing a 3D shape from a single RGB image gives the network the option to reason about shape structure via recognition [24, 51, 39], harming the generalization on novel shape classes. Lastly, these works often compress the 3D information into a global vector [24, 40] thus losing 3D structure and surface details. To address these issues, we propose to factorize the problem into front-back depth estimation and implicit shape completion for better generalization to novel shape classes. To capture surface details, we further propose to lift depth maps into 3D and extract point-aligned multi-scale features directly in voxel feature space. Fig. 1 shows an overview of our method.

Problem formulation. Let $\mathcal{T} = \{m_k\}_{k=1}^{K}$ be our training set, consisting of K meshes. Each mesh m_k is associated with 2D-renderings, i.e., color and depth, from J different view-points. Furthermore, each training mesh belongs to

seen shape classes $\mathcal{Y}^s = \{\text{car, ship, ...}\}$. Our goal is to learn a parametric function f that is able to predict the 3D shapes from a single RGB image belonging to novel classes $\mathcal{Y}^n = \{\text{bookshelf, suitcase, ...}\}$. Note that the seen classes and novel classes are completely disjoint, i.e., $\mathcal{Y}^s \cap \mathcal{Y}^n = \emptyset$.

3.1. Front-Back Depth Estimation

The first component of our model predicts front and back depth maps from an RGB color image. Using depth as an intermediate representation has several advantages. First, depth contains essential geometric information shared across different classes which facilitates knowledge transfer from seen to novel classes. Furthermore, depth estimation encourages the networks to focus on using geometric cues rather than recognition [39, 45].

Given an input RGB image I, we learn convolutional neural networks Ψ_f and Ψ_b to predict its front and back depth maps respectively. Here the front depth map refers to the per-pixel ray-distance to the visible object surfaces from the viewpoint of the input image, while the back depth map is defined via the last surface intersection from the same viewpoint. We adopt a U-Net [32] architecture for depth estimation. The networks Ψ_f and Ψ_b share the same image encoder while they have two separate decoders.

Loss function. Instead of optimizing the $\mathcal{L}_2$ loss like [54], we choose to minimize the following continuous version of the reverse Huber (berHu) loss [19],

$$\mathcal{L}_{berHu}(\hat{D}_f, D_f) = \begin{cases} |\hat{D}_f - D_f|, & |\hat{D}_f - D_f| \leq t \\ \frac{(\hat{D}_f - D_f)^2 + t^2}{2t} & \text{otherwise.} \end{cases}$$

where $\hat{D}_f = \Psi_f(I)$ denotes the predicted front depth and D_f is its ground truth. t is a constant that controls where the switch from $\mathcal{L}_2$ and $\mathcal{L}_1$ occurs. Following [19], we set $t = 0.2 * \max_i(|\hat{D}_f^i - D_f^i|)$, where i indexes all pixels in the current batch. Compared to the $\mathcal{L}_2$ loss, the berHu loss yields larger gradients at small errors by switching to $\mathcal{L}_1$, which is beneficial for small-error regions. Similarly, the back depth estimator Ψ_b optimizes the same berHu loss.

3.2. Implicit Shape Completion

The depth maps contain only partial object surfaces. The second part of our method is a shape completion network with implicit shape representation to recover the missing information. Given a latent representation c of the input observation, e.g., voxel grids, we aim to learn an implicit network $f(c, p) : \mathcal{C} \times \mathbb{R}^3 \rightarrow [0, 1]$ that predicts the occupancy at a continuous query point $p \in \mathbb{R}^3$. Thereby, the 3D surface is implicitly represented as the decision boundary of the classifier. In this section, we describe three important elements of our shape completion, i.e., depth voxelization, point-aligned multi-scale features i.e., the latent encoding c, and the implicit shape decoder.

Depth voxelization. One of our key contributions is to extract local and global shape features in 3D which requires the input to be a 3D voxel grid. To this end, we unproject front and back depth maps into a combined 3D point cloud in the viewer-centered coordinate, which requires only the intrinsic camera parameters. We assume that camera intrinsics are fixed, same as prior work [33, 34, 54, 51]. Afterwards, we convert the unordered and irregular point cloud into a regular 3D grid $V \in [-0.5, 0.5]^{N^3}$ where N denotes the voxel resolution. To make the voxelization differentiable, we compute the value of each voxel cell by averaging the distance between all points inside this voxel to its center followed by negating the value and adding one. If a voxel does not contain any point, we set it to be zero.

Point-aligned multi-scale features. In order to learn detailed shape structures, we follow [8] to extract a rich encoding of the input voxel V with 3D convolutions. Our shape encoder network Φ consists of L layers of standard 3D convolution, non-linearity, batch norm, and max-pooling. Applying the encoder on the input voxel yields a multi-scale feature pyramid $F_1, \ldots, F_L$ with growing receptive fields. While early layer features, e.g., F_1 capture local information (shape details), the deeper layer features, e.g. F_L, encode global structures. For an intermediate layer l with feature maps F_l of C_l channels, we extract a C_l-dim feature vector by trilinearly interpolating F_l at the query point $p \in \mathbb{R}^3$. In addition, we also take into account the context information by extracting features for the 6 surrounding points of the query point p. Afterwards, we concatenate features from those 7 points (including the query point itself) to produce the point-aligned feature $F_l(p) \in \mathbb{R}^{7*C_l}$. We repeat the above process for each intermediate layer and concatenate them to obtain the point-aligned multi-scale feature denoted as $\Phi(V, p) = cat(\{F_i(p)\}_{i=0}^{L})$.

Implicit shape decoder. Given the point-aligned multi-scale features, our implicit shape decoder $f(\Phi(V, p))$ predicts the occupancy of the query point p, i.e., the occupancy is 1 if p is inside the surface otherwise 0. While ONet and DISN feed the query point p i.e., the x, y, z coordinates to the decoder, our implicit decoder only relies on the local shape features of query points $\Phi(V, p)$, preventing the network from memorizing the points of training shapes.

Loss function. Since it is a binary classification problem, we minimize the following binary cross-entropy loss defined for a single query point,

$$\mathcal{L}_{CE}(o, \hat{o}) = -o \log \hat{o} - (1 - o) \log(1 - \hat{o}) \qquad (1)$$

where $\hat{o}$ denotes the predicted occupancy and o is the ground truth occupancy of the query point p.

Discussion. Our novelty lies in the combination of using depth as an intermediate representation, encoding local and global shape features in 3D voxel space, and decoding with

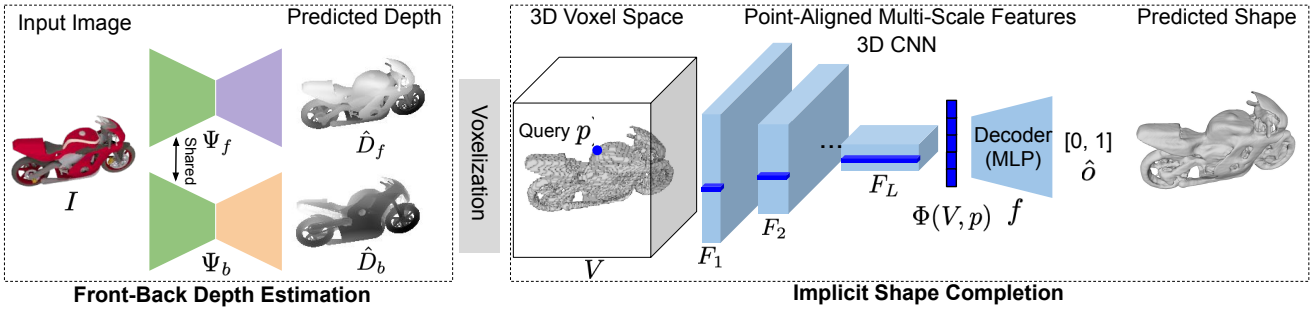

Figure 1. Overview of our proposed method GIN for single-image reconstruction of novel classes. Given an input image I, we first predict its front and back depth maps, i.e., $\hat{D}_f$ and $\hat{D}_b$ with depth estimation networks, i.e., Ψ_f and Ψ_b with a shared encoder. Then we unproject the depth maps into 3D points followed by voxelizing them into a 3D voxel grid V in viewer-centered coordinate system. Afterwards, we apply 3D convolutions and extract point-aligned multi-scale features by trilinearly interpolating intermediate feature maps $\{F_1, F_2, \ldots, F_L\}$ at query point p. Finally, our implicit decoder predicts the occupancy $\hat{o}$ of the query point p.

an implicit network. We empirically show that prior works that apply those techniques in isolation do not generalize well to novel shape classes (Tab. 1). To our knowledge, such a combination has not been proposed before, and in particular not for reconstructing novel shape classes.

3.3. Training and Inference

Here we describe training and inference pipelines. More implementation details are provided in supplement.

Training pipeline. We adopt a two-stage training algorithm. At the first stage, we train the depth estimation network and shape completion network separately using the ground truth depth maps and shapes. Afterwards, we finetune the depth estimation and shape completion networks in an end-to-end manner.

Viewer-centered supervision. While many prior works [24, 51, 8] are trained with 3D shapes in the canonical view, we train our network with viewer-centered supervision where the ground truth shape is present in the view point of the input image. This has two advantages, i) we don't need to estimate camera pose at inference time and ii) our desired output is pixel aligned. We found that networks trained with the viewer-centered supervision generalizes better to novel categories (see Tab. 2).

Query point sampling. To train the implicit network, we combine two different point sampling strategies with equal probabilities (0.5). The first strategy samples points near surfaces by adding a Gaussian noise to the surface points, i.e., $p = p^s + n$ with p^s being a surface point and $n \sim \mathcal{N}(0, 0.01)$. In addition, we sample continuous points uniformly from the unit 3D cube, i.e., $p \sim U(-0.5, 0.5)$.

Inference. At the test time, given a single RGB image, we want to reconstruct the 3D surface of the object shown in the image. First, we predict its front and back depth maps using the learned depth estimation networks. Then we convert the depth maps into 3D voxel grids as described in Sec. 3.2. Afterwards, we construct a 3D grid at desired resolution and compute occupancies of all grid points with the learned implicit shape completion network. Finally, we transform the resulting occupancy grid into a mesh by running the classical marching cubes algorithm [22]. It is worth noting that our method does not need to estimate the camera pose because it predicts 3D shapes in viewer-centered coordinate.

4. Experiments

Here we first describe baselines and evaluation settings. In Section 4.1, we compare with SOTA quantitatively and qualitatively. In Section 4.2, we ablate our model components, back depth, choice of coordinates, and show per-class performance analysis. Finally, we present results in few-shot single-image 3D reconstruction setting in Section 4.3.

Dataset. We follow the SDFNet [40] benchmark because it is the largest scale evaluation benchmark for single-image 3D reconstruction. The benchmark utilizes the ShapeNetCore.v2 [5] dataset which consists of over 50K meshes from 55 object categories. It treats the 13 largest categories as seen classes and the remaining 42 categories as novel classes. Within each class, the shapes are randomly split into training, validation, and test sets. In addition, we also test our model, trained on synthetic renderings of ShapeNet, on real images from Pix3D [37], mainly consisting of furniture images from the IKEA website.

Data generation. To generate synthetic renderings for ShapeNet, we adopt the rendering code provided by SDFNet [40] to generate images in Blender [12]. Specifically, each shape is rendered from $J = 25$ different view points uniformly sampled from $\theta_{azimuth} \in [0°, 360°)$, $\theta_{elevation} \in [-50°, 50°)$. Our viewpoints are the same as for SDFNet (2DOF setting). We normalize each mesh to fit inside a unit cube after the viewpoint rotation such that the resulting depth maps are in a normalized range ($[-0.5, 0.5]$). The distance to the camera is fixed to be 2.2, which is standard for single-image 3D reconstruction [11, 24, 54]. The rendering pipeline in total results in

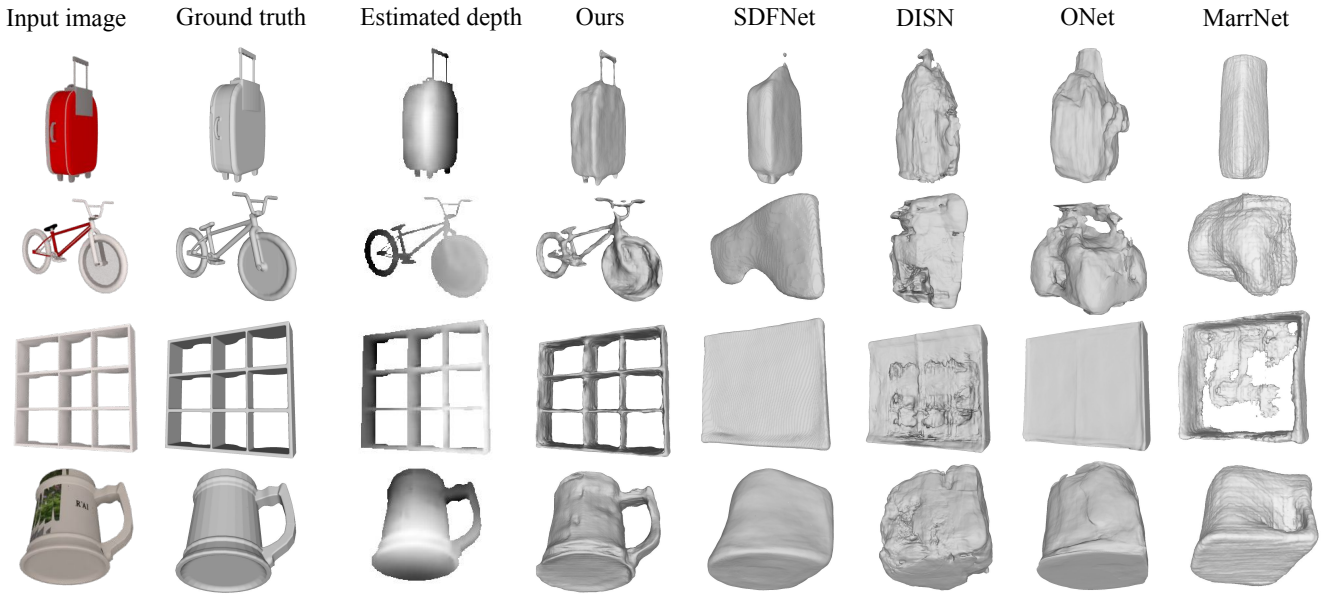

Figure 2. Qualitative comparison on ShapeNet. We visualize the reconstruction results from a single RGB image of novel classes. All the methods are trained on the 13 seen classes on ShapeNet. Our results are obviously more consistent with the input image than competitors.

over 1.3M images of size 256×256. In addition to the RGB images, we also render their corresponding depth images. We compute watertight meshes using the code provided by DISN [51]. For a fair comparison, these watertight meshes are used as ground-truth for all methods.

Metric. As our goal is to evaluate the quality of surface reconstruction, we use the widely-used F-Score (FS) [39], Chamfer distance (CD) [2], and normal consistency (NC) which are reliable metrics for measuring the distance between object surfaces [39]. In all of our experiments, the distance is computed with 100K points sampled from predicted and ground-truth mesh pairs. We follow SDFNet [40] to report the FS at distance threshold 0.01. We compute the FS, CD, and NC within each class and report the mean.

Evaluation protocol. Unless otherwise specified, we train a single model on all the training examples of seen classes. The network is then evaluated on the unseen shapes from both seen and novel classes. All the 25 random views per training shape are used for training, while only a random view per test shape is selected for evaluation. For the methods that predict 3D shapes in the camera view, i.e., ours, GenRe [54] and SDFNet [40], we rotate the ground truth mesh to the input viewpoint for evaluation.

4.1. Comparison with SOTA

In the following, we compare with SOTA baselines on ShapeNet and Pix3D quantitatively and qualitatively .

Baselines. We compare our method against the following state-of-the-art baselines. GenRe [54] and SDFNet [40] define the state-of-the-art in single-image 3D reconstruction for novel classes. MarrNet [45] and GenRe [54] also predict depth but are based on voxels. ONet [24] and DISN [51] are representative methods with implicit shape representation. We directly take results of GenRe and SDFNet from [40] because we use the same renderings and data splits. We obtain the results of other methods by running the publicly released codes on our renderings.

	Seen Classes			Novel Classes		
Method	CD ↓	NC ↑	FS ↑	CD ↓	NC ↑	FS ↑
GenRe [54]	0.153	0.60	0.12	0.172	0.61	0.11
MarrNet [45]	0.116	0.68	0.15	0.127	0.69	0.13
ONet [24]	0.081	0.78	0.25	0.145	0.72	0.15
DISN [51]	0.070	0.77	0.33	0.124	0.72	0.20
SDFNet [40]	0.050	**0.79**	0.42	0.080	0.76	0.31
GIN (ours)	**0.042**	**0.79**	**0.47**	**0.056**	**0.79**	**0.40**

Table 1. Comparison with SOTA on ShapeNet. We report Chamfer distance (CD), normal consistency (NC) and F-score (FS) at distance threshold 0.01. All methods are trained on 13 seen classes and evaluated on both seen and novel classes.

Reconstructing unseen shapes of seen classes. We first conduct evaluation on the seen classes, which is the standard evaluation setting used in single-image 3D reconstruction [24, 51]. As shown in Tab. 1 (Seen Classes columns), our method achieves the best CD and FS. In particular, our results exceed the second best method, i.e., SDFNet by 16.0% and 5% in terms of CD and FS respectively. This empirically indicates that extracting local shape features is critical for the performance.

Reconstructing unseen shapes of novel classes. Furthermore, we compare with different methods on 42 novel classes that are excluded from the training set. This is a highly challenging task because it is not possible to learn class-specific shape priors of novel classes from training

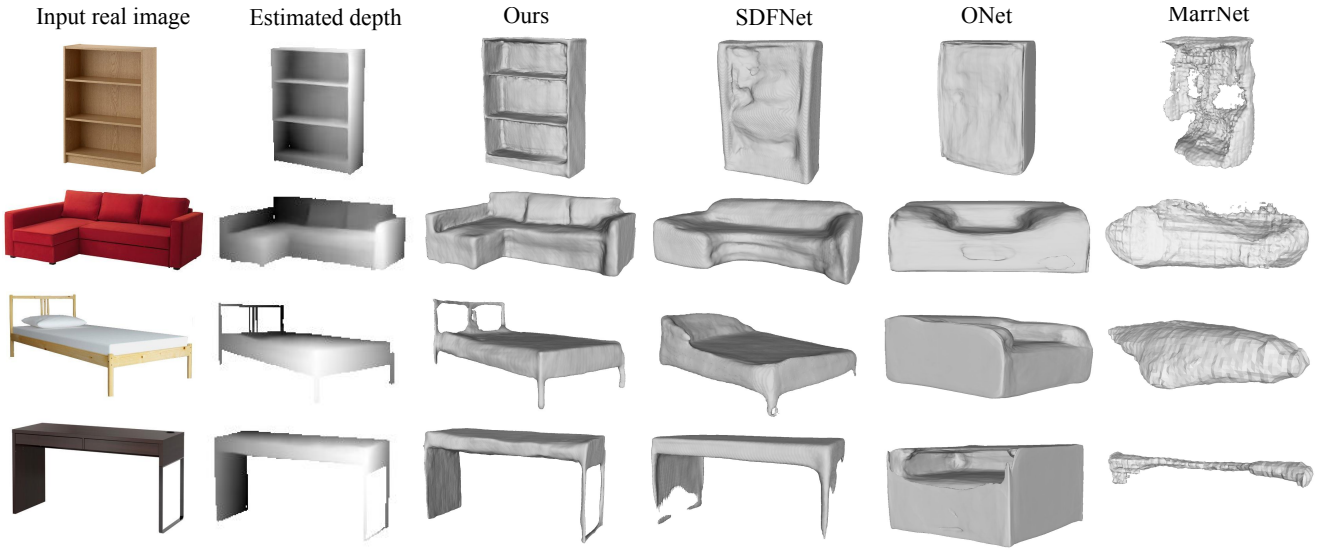

Figure 3. Qualitative comparison on real images from Pix3D [37]. All models are trained on the synthetic renderings of ShapeNet. Our results are significantly better aligned with the inputs than the baselines.

data and accomplishing this task requires strong generalization abilities. The results are presented in Tab. 1 (Novel Classes columns). We have the following observations. First, our method significantly outperforms all baselines by a wide margin, surpassing the second best method SDFNet [40] by 30.0% and 9% in terms of CD and FS respectively. Secondly, while all methods suffer from a performance drop from seen to novel classes, our method has the smallest drop among implicit methods, which indicates the strong generalization ability of our approach.

Qualitative results on novel classes of ShapeNet. We present qualitative results in Fig. 2. Our results are obviously more consistent with input images while competitors either fail to recover input details or output shapes with wrong topology. For example, our method is able to reconstruct the handle of the suitcase (1-st row), the structural details of the bike (2-nd row), the empty shelves of the bookcase (3-th row), and the handle of the mug (the fourth row). However, competitors like SDFNet [40] or ONet [24], always produce overly smooth surfaces. This is because our method extracts point-aligned multi-scale shape features to capture details, while SDFNet encodes the input with a global feature vector where detailed information is lost. DISN [51] attempts to learn local features from 2D image features, but resulting in global topology errors. These results confirm the importance of using depth as an intermediate representation, learning local shape features in 3D voxel space and decoding with implicit networks.

Qualitative results on real images. We further evaluate how our model trained on synthetic renderings, generalizes to real images from Pix3D [37]. Note that this is a highly challenging setting due to the substantial domain gap between synthetic renderings and real images. To reduce the gap and increase the training data diversity, we apply strong data augmentations on synthetic renderings i.e., color jittering, adding lighting noise and overlaying synthetic renderings with real background images from SUN [49]. For a fair comparison, all the methods are trained on the same synthetic renderings of the 13 ShapeNet classes. Some qualitative results are shown in Fig. 3 where the first two rows (bed and sofa) are seen classes and last two rows (bed and desk) are novel classes. Despite only trained on synthetic ShapeNet, we find that our method generates significantly better reconstructions with rich details than the baselines, validating the domain transferability of our approach. More qualitative results can be found in the supplement.

4.2. Model Analysis

In this section, we ablate different components of our method and analyze per-class performance.

Impact of depth estimation and point-aligned multi-scale features. We ablate the two important modules of our method by: (1) replacing the depth estimation network with a 3D CNN that directly predicts the 3D voxel shapes; (2) replacing point-aligned multi-scale features module with a 2D CNN that extracts features from estimated depth images. In Tab. 2, we observe that both changes (Ours w/o PAMSF and Ours w/o Depth) lead to a significant performance drop compared to our full method, implying the importance of using depth as an intermediate representation and learning local features in 3D.

Impact of back depth. We show the impact of using back-view depth (BD) in Tab. 2. If the input to our implicit shape completion network is the oracle depth, removing BD leads to a significant performance drop. This is expected because the oracle front and back depth almost provide a complete

	Method	Coordinate	Seen Classes CD ↓	Seen Classes FS ↑	Novel Classes CD ↓	Novel Classes FS ↑
EST Depth	Ours	VC	**0.042**	0.47	**0.056**	**0.40**
	Ours w/o BD	VC	0.043	**0.48**	0.059	**0.40**
	Ours	OC	0.042	**0.48**	0.059	0.38
	Ours w/o Depth	VC	0.060	0.35	0.086	0.25
	Ours w/o PAMSF	VC	0.056	0.36	0.073	0.28
Oracle Depth	Ours	VC	**0.012**	**0.91**	**0.015**	**0.87**
	Ours w/o BD	VC	0.022	0.77	0.030	0.71
	Ours	OC	**0.012**	**0.91**	0.019	0.83

Table 2. Ablation studies on using estimated depth (EST Depth), ground truth depth (oracle depth), back-view depth (BD) and choices of coordinate system: viewer-centered (VC) and object-centered (OC) coordinates. We also ablate the depth estimation (Depth) module and point-aligned multi-scale features (PAMSF). We report the Chamfer distance (CD) and F-score (FS) on both seen and novel classes.

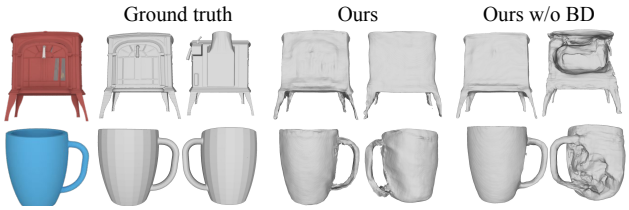

Figure 4. Qualitative results of ablating the back-view depth (BD) estimation. Without BD, the invisible side of the reconstruction does not have reasonable surfaces qualitatively.

object surface. If the estimated depth is used, using BD leads to a slight reduction in CD. Moreover, comparing the qualitative results in Fig. 4, we observe that BD yields a more globally coherent surface on the invisible side which may not be reflected quantitatively.

Impact of coordinate system. Furthermore, we study how the choices of coordinate systems affect the generalization ability. While many prior methods [24, 51, 27, 38] predict shapes in an object-centered (OC) coordinate system, i.e., a canonical view, our method uses a viewer-centered (VC) coordinate system. As shown in Tab. 2, while OC achieves comparable results with VC on seen classes, VC generalizes significantly better to novel classes than OC consistently. Intuitively, OC aligns all the ground truth meshes to a canonical view which makes it easier to fit the training data. However, as a canonical view for novel classes is not defined, OC does not generalize well to the object categories that are not seen during training.

Per-class F-score vs depth error. In Fig. 5, we show per-class F-score and depth estimation RMSE of both seen and novel classes. We have the following observations. First, our depth estimation network generalizes well to novel classes e.g., most of novel classes have a depth RMSE lower than 0.04, laying a good foundation for our shape completion network. Moreover, the per-class F-scores are strongly correlated with the depth RMS, i.e., a lower depth RMSE tends to have a higher F-score.

In addition, compared to the seen classes, we observe

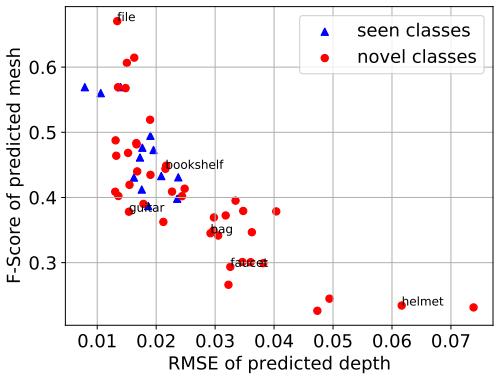

Figure 5. Per-class depth error vs F-score of our method. We report root-mean-squared error (RMSE) for depth estimation. It shows that F-score strongly correlates with the depth RMSE.

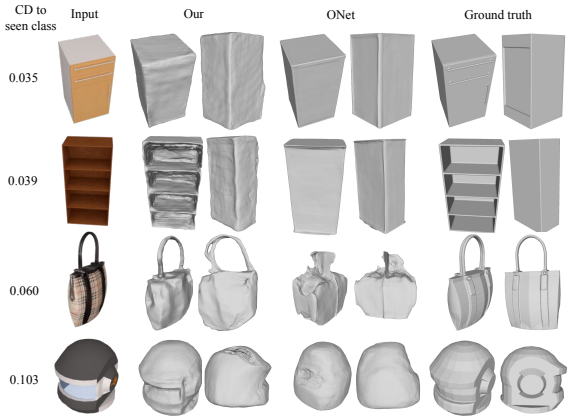

Figure 6. Qualitative results of 4 novel classes from easy to difficult. CD to seen classes indicates the distance of a given class to seen classes (smaller distance means higher similarity).

that results of novel classes have a higher deviation, ranging from 0.23 to 0.67 in terms of F-score. We speculate that this is related to their similarities to seen classes. Formally, for each novel class with shapes, we estimate its dissimilarity to seen classes by computing its Chamfer distance to the training set. Indeed, we found that the novel class F-score is strongly correlated with its dissimilarity to training classes (Pearson correlation -0.66). In Fig. 6, we show that the qualitative results gradually become worse with an increasing dissimilarity to seen classes. Nevertheless, our results are able to capture input details while ONet tends to output a over-smooth shape. These results again confirm that learning local features in 3D voxel space is crucial.

4.3. Analysis in Few-Shot Reconstruction

Previous experiments show that our method generalizes well to the novel classes that are not observed during training. While it is expensive to collect a large number of training shapes, it is often realistic to assume that a few training shapes are available for novel classes, namely few-shot

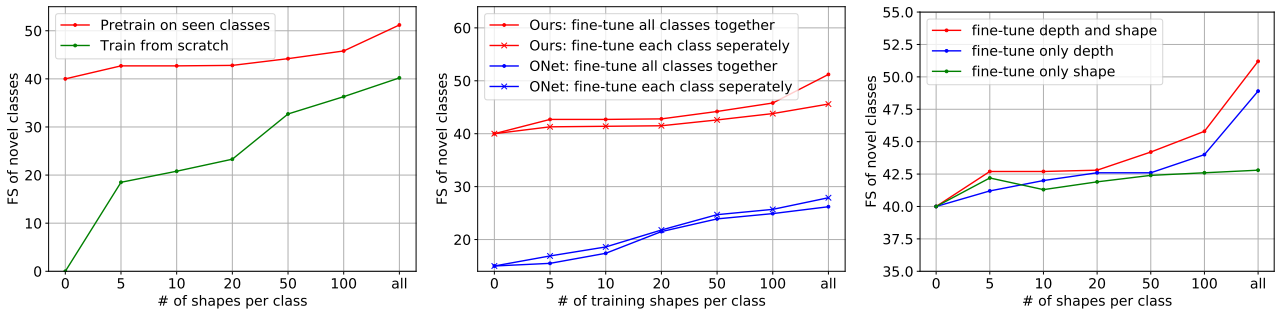

Figure 7. Results of novel classes under the few-shot single-image 3D reconstruction setting. Left: the effect of pretraining on seen classes, middle: fine-tune each class separately vs fine-tune all classes together, showing that the depth allows our method to benefit more from the geometric knowledge shared across shape categories, right: the effect of using different supervision signals, indicating that our method can be further improved by using only few-shot depth supervision.

single-image reconstruction. In this section, we conduct extensive analyses under this setting to show interesting insights regarding the effect of gradually enlarging training set, learning strategies, and different level of supervisions.

Experimental setup. We conduct experiments on ShapeNetv2 [5] and use the same class splits (13/42 seen/novel classes) introduced in Sec. 4. The training set of seen classes and test set remain unchanged too. In addition, for each novel class, we randomly draw $k = \{5, 10, 20, 50, 100\}$ shapes incrementally from its training set. Note that we simply take all examples when the number of available training shapes is smaller than k. In total, there are over 10K training meshes from 42 novel classes.

Effect of pretraining. We compare two different initializations i.e., random initialization and the model pretrained on seen classes. We present the results with different numbers of training shapes in Fig. 7 (left). The performance gap tends to decrease with the increasing number of training shapes. The key message is that it is beneficial to first pretrain the reconstruction model on some relevant classes when training data is scarce. In the subsequent experiments, we always initialize our model with the pretrained one.

Fine-tune each class separately vs fine-tune all classes. Here we investigate two different learning strategies i.e., fine-tune the pretrained model on each novel class separately (42 models) and fine-tune all novel classes together (one model). We present the results in Fig. 7 (middle). For ONet, we observe that fine-tuning each class separately performs slightly better than fine-tuning all classes, which is expected because learning to reconstruct a single class is easier than to reconstruct all classes. In contrast, for our method, fine-tuning all classes consistently outperforms fine-tuning each class separately. This can be attributed to the intermediate depth maps used in our method, encoding class-agnostic priors which facilitate knowledge sharing across different classes. In the subsequent experiments, we always fine-tune all classes together.

Depth is almost all you need. Furthermore, we compare three different supervision signals used to improve novel classes: (1) using 3D shapes as supervision to fine-tune our method (2) using only ground truth depth maps as supervision to fine-tune depth (front only) estimation network (3) using 3D shapes as supervision to only fine-tune shape completion network. As shown in Fig. 7 (right), there is a small performance gap between fine-tuning only depth and fine-tuning both shape and depth. This is encouraging because collecting ground truth depth images is far cheaper than 3D shapes. These experiments demonstrate that just by improving depth prediction using fine-tuning, we can significantly improve the reconstruction quality.

5. Conclusion

In this work, we have introduced Generalizing Implicit Networks (GIN) for single-image 3D reconstruction of novel shape classes. First, we argue that using depth maps as an intermediate representation for learning implicit surfaces enhances the generalizability, facilitating knowledge sharing across shape categories. Secondly, we propose to lift the depth maps into 3D and complete them with an implicit network, preserving the local details and reasoning globally. We empirically show that our GIN significantly outperforms the state of the art (e.g., SDFNet), improving the Chamfer distance by 16.0% and 30.0% on seen and novel shape classes respectively. Qualitatively, our reconstruction preserves better surface details contained in the input image than the competitors (Fig. 2). Although trained on only synthetic renderings, our GIN generalizes well on real images (Fig. 3).

Acknowledgements This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans) and the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. G. Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1, Project number 390727645. J. Chibane is a fellow of the Meta Research PhD Fellowship Program - area: AR/VR Human Understanding.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *TPAMI*, 2015.

[2] Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. Technical report, SRI INTERNATIONAL MENLO PARK CA ARTIFICIAL INTELLIGENCE CENTER, 1977.

[3] Wenjing Bian, Zirui Wang, Kejie Li, and Victor Adrian Prisacariu. Ray-onet: Efficient 3d reconstruction from a single rgb image. In *BMVC*, 2021.

[4] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[6] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR*, 2019.

[7] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, pages 5939–5948, 2019.

[8] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *CVPR*, 2020.

[9] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis from sparse views of novel scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021.

[10] Julian Chibane, Aymen Mir, and Gerard Pons-Moll. Neural unsigned distance fields for implicit function learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, December 2020.

[11] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016.

[12] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.

[13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[14] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *CVPR*, 2018.

[15] Kan Guo, Dongqing Zou, and Xiaowu Chen. 3d mesh labeling via deep convolutional neural networks. *ACM Transactions on Graphics (TOG)*, 35(1):1–12, 2015.

[16] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *NeurIPS*, 2020.

[17] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018.

[18] Florian Kluger, Hanno Ackermann, Eric Brachmann, Michael Ying Yang, and Bodo Rosenhahn. Cuboids revisited: Learning robust 3d shape fitting to single rgb images. In *CVPR*, 2021.

[19] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016.

[20] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[21] Manyi Li and Hao Zhang. D2im-net: Learning detail disentangled implicit fields from single images. In *CVPR*, 2021.

[22] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987.

[23] David Marr. Vision. *W. H. Freeman and Company*, 1982.

[24] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019.

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013.

[26] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[27] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019.

[28] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *ECCV*. Springer, 2020.

[29] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *CVPR*, pages 5648–5656, 2016.

[30] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *ECCV*, pages 704–720, 2018.

[31] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2016.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 2015.

[33] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *CVPR*, pages 2304–2314, 2019.

[34] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, pages 84–93, 2020.

[35] Daeyun Shin, Charless C Fowlkes, and Derek Hoiem. Pixels, voxels, and views: A study of shape representations for single view 3d object shape prediction. In *CVPR*, 2018.

[36] Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *CVPR*, pages 808–816, 2016.

[37] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *CVPR*, 2018.

[38] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *CVPR*, 2017.

[39] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, 2019.

[40] Anh Thai, Stefan Stojanov, Vijay Upadhya, and James M Rehg. 3d reconstruction of novel object shapes from single images. In *2021 International Conference on 3D Vision (3DV)*, 2021.

[41] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, 2016.

[42] Jianren Wang and Zhaoyuan Fang. Gsir: Generalizable 3d shape interpretation and reconstruction. In *ECCV*. Springer, 2020.

[43] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, pages 52–67, 2018.

[44] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. Forknet: Multi-branch volumetric semantic completion from a single depth image. In *ICCV*, pages 8608–8617, 2019.

[45] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, William T Freeman, and Joshua B Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. In *NeurIPS*, 2017.

[46] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *CVPR*, pages 82–90, 2016.

[47] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. Learning shape priors for single-view 3d completion and reconstruction. In *ECCV*, pages 646–662, 2018.

[48] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 2018.

[49] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[50] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *Computer Graphics Forum*, 2022.

[51] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. 2019.

[52] Yuan Yao, Nico Schertler, Enrique Rosales, Helge Rhodin, Leonid Sigal, and Alla Sheffer. Front2back: Single view 3d shape reconstruction via front to back prediction. In *CVPR*, 2020.

[53] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021.

[54] Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Joshua B Tenenbaum, William T Freeman, and Jiajun Wu. Learning to Reconstruct Shapes From Unseen Classes. In *NeurIPS*, 2018.

[55] Fang Zhao, Wenhao Wang, Shengcai Liao, and Ling Shao. Learning anchored unsigned distance functions with gradient direction alignment for single-view garment reconstruction. In *ICCV*, 2021.