Planes vs. Chairs: Category-guided 3D shape learning without any 3D cues

Zixuan Huang[1], Stefan Stojanov[1], Anh Thai[1],
Varun Jampani[2], and James M. Rehg[1]

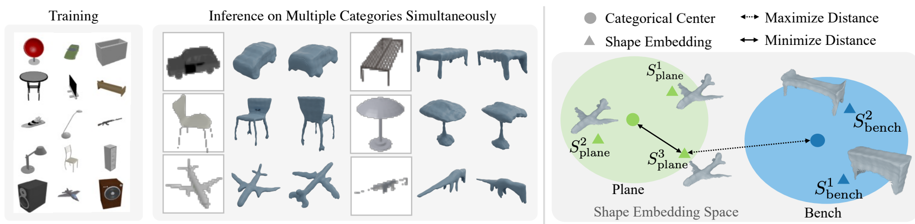
[1] Georgia Institute of Technology
[2] Google Research

Abstract. We present a novel 3D shape reconstruction method which learns to predict an implicit 3D shape representation from a single RGB image. Our approach uses a set of single-view images of multiple object categories without viewpoint annotation, forcing the model to learn across multiple object categories without 3D supervision. To facilitate learning with such minimal supervision, we use category labels to guide shape learning with a novel categorical metric learning approach. We also utilize adversarial and viewpoint regularization techniques to further disentangle the effects of viewpoint and shape. We obtain the first results for large-scale (more than 50 categories) single-viewpoint shape prediction using a single model. We are also the first to examine and quantify the benefit of class information in single-view supervised 3D shape reconstruction. Our method achieves superior performance over state-of-the-art methods on ShapeNet-13, ShapeNet-55 and Pascal3D+.

1 Introduction

Reconstructing the 3D shape of objects from 2D images is a fundamental computer vision problem. Recent works have demonstrated that high-quality 3D shape reconstructions can be obtained from a single RGB input image [4, 7, 10, 30]. In the most straightforward approach, explicit 3D supervision from ground-truth (GT) object shapes is used to train a shape reconstruction model [43, 30, 39, 51]. While this approach yields strong performance, GT 3D shape is difficult to obtain on a large scale. This limitation can be addressed via multi-view supervision [52, 45, 53, 24, 17, 21, 26, 37], in which the learner optimizes a reprojection loss over a set of viewpoints. This approach becomes more difficult when fewer viewpoints are available during training. The limiting case is single-view supervision, which is challenging due to the well-known entanglement between shape and pose.[1] Prior works (see Table 1) have addressed this challenge in two ways. The first approach assumes that the camera viewpoint is known at training time [25, 20]. While this strong assumption effectively reduces entanglement, it is not scalable as it requires pose annotations.

[1] When only a single view of each object instance is available for training, there are an infinite number of possible 3D shapes that could explain the image under appropriate camera viewpoints.



(a) Multi-category Single-view Shape Learning without Viewpoint (b) Shape Metric Learning via Category label

Fig. 1: (a) We present the first method to learn 3D shape reconstruction from single-viewpoint images over *multiple object categories simultaneously*, without 3D supervision or viewpoint annotations. (b) To facilitate shape learning under such a challenging scenario, We leverage category labels to perform metric learning in the shape embedding space. For each category, we learn a shape center. Then with a given sample, we minimize the distance between its shape embedding and the corresponding category center, while contrasting this distance with other inter-category distances.

A second line of attack trains separate models for each category of 3D objects [18, 44, 8, 23, 15, 14, 55, 36, 49]. This single category approach provides the learner with a strong constraint – each model only needs to learn the "pattern" of shape and pose entanglement associated with a single object category. This approach has the disadvantage that data cannot be pooled across categories: pooling data can be beneficial to tasks such as generalization to unseen categories of objects (zero-shot [19, 57] or few-shot [53]) and viewpoint learning. For example, tables and chairs have many features (e.g. legs) in common, and a multi-class reconstruction model could leverage such similarities in constructing shared feature representations.

We introduce a novel *multi-category, single-view (MCSV)* 3D shape learning method, shown in Fig. 1 (a), that does not use shape or viewpoint supervision, and can learn to represent multiple object categories within a single reconstruction model. Our method exploits the observation that shapes within a category (e.g. planes) are more similar to each other than they are to the shapes in other categories (e.g. planes vs. chairs). We use a shape metric learning approach, illustrated in Fig. 1 (b), which learns category shape embeddings (category centroids denoted as green/blue dots) together with instance shape embeddings (green/blue triangles) and minimizes the distance between shape instances and their corresponding category centers (solid arrow) while maximizing the distance to the other categories (dotted arrow). This method has two benefits. First, the learned shape manifold captures a strong prior knowledge about shape distance, which helps to eliminate erroneous shapes that might otherwise explain the input image. Second, similar shapes tend to be clustered together, enabling the supervision received by a particular shape to affect its neighbors, and helping to overcome the limitation of having only a single viewpoint. As a consequence of this approach, we are able to train a single model on all 55 object categories in ShapeNet-55, which has not been accomplished in prior work.

Another important aspect of our approach is that we represent object shape as an implicit signed distance field (SDF) instead of the more commonly used mesh representation, in order to accommodate objects with varying topologies. This is important for scaling MCSV models to handle large numbers of object categories. Specifically, our model (Fig. 1 (a)) takes a single input image and produces an implicit SDF function for the 3D object shape. In contrast to mesh and voxel representations, SDF representations create challenges in training because they are less constrained and require a more complex differentiable rendering pipeline. A beneficial aspect of our approach is the use of adversarial learning [15, 55] and cycle-consistency [35, 55] for regularization.

In summary, our 3D reconstruction learning approach has the following favorable properties:

- **Extremely weak supervision.** We use only single-view masked images with category labels during training (considered as 'self-supervised' in prior works).
- **Categorical shape metric learning.** We make effective use of class labels with our novel shape metric learning.
- **Single template-free multi-category model.** We learn a single model to reconstruct multiple categories without using any category-specific templates.
- **Implicit SDF representation.** We use implicit SDFs, instead of meshes, to represent diverse object topologies.
- **Order of magnitude increase in object categories learned in a single model.** Due to the scalability of our approach, we are the first to train a MCSV model on all 55 categories of ShapetNetCoreV2 [2].
- **State-of-the-art reconstructions.** Experiments on ShapeNet [2] and Pascal3D+ [50] show consistently better performance than previous methods.

To the best of our knowledge, our work is the first successful exploration of the challenging MCSV problem without using known viewpoints or supervised pretraining.

2 Related Work

Single-View Supervision. The body of work that is most closely-related to this paper are methods that use single-view supervision to learn 3D shape reconstruction with a back-projection loss [18, 44, 8, 23, 15, 14, 6, 55, 36, 20, 49, 25]. These works can be organized as in Table 1 and largely differ in their choice of 1) learning a single multi-category (multi-class) vs multiple single-category models for reconstruction; 2) shape representation (e.g. implicit SDF vs explicit mesh); 3) known vs. unknown viewpoint assumption. We are the first to demonstrate the feasibility of single-view, multi-category learning of an implicit shape representation (SDF) under the unknown viewpoint condition via the use of category labels. We are also the first to provide results on all of the ShapeNet-55 classes in a single model, an order of magnitude more than prior works.

Table 1: Single-view supervised methods for shape reconstruction. For methods without supervision, they are mainly verified under a limited setting [14] or for specific objects [49]. K: keypoints, T: category templates, M: masks, m: mesh, vox: voxel, pc: pointcloud, im: implicit representation.

Model	[18]	[44]	[8]	[23]	[6]	[15]	[14]	[55]	[36]	[25]	[49]	[20]	[40]	Ours
Multi-Class Rec.	-	-	-	-	-	-	-	-	-	-	-	✓	✓	✓
Supervision	K,M	M,T	M,T	M	M	M	-	M	M	M,V	-	M,V	M,V	M
3D Rep.	m	m	m	m	vox	vox	m	m	pc	im	d	m	m	im

Within this body of work [20, 25, 55] and the concurrent work of [40] are the most closely related. Kato et al. [20] (VPL) and Simoni et al. [40] are the only prior works to address multi-category shape learning with a single model, but both assume access to ground truth (GT) viewpoint and use mesh representations. Similar to our work, Ye et al. [55] do not use viewpoint GT, but they train one model for each object category. Further, their mesh prediction is refined from a low-resolution voxel prediction by optimizing on each individual sample. The latter is less efficient than feedforward models and can potentially harm the prediction of concave shapes or details. In contrast with [21, 40, 55] we use an implicit shape representation. Further, as shown in Table 1, some earlier approaches use fixed meshes as deformable shape templates. In contrast with implicit functions, meshes with fixed topology restrict the possible 3D shapes that can be accurately represented.

Lin et. al. [25] is the only prior work to use an implicit SDF representation for single view supervision, and we adopt their SDF-SRN network architecture and associated reprojection losses in our formulation. This work trains single-category models and assumes that the camera viewpoint is known, whereas one of the main goals of our work is to remove these assumptions and to demonstrate that the unconstrained SDF representation can be successful in the multi-category, unknown viewpoint case. We consistently outperform a version of SDF-SRN modified for our setting in Table 2, which highlights our innovations.

Shape Supervision. Prior works on single image 3D reconstruction with explicit 3D geometric supervision have achieved impressive results [4, 7, 10, 30, 51]. However, the requirement of 3D supervision limits the applicability of these methods. To overcome this, subsequent works employ multi-view images for supervision. The use of image-only supervision is enabled by differentiable rendering, which enables the generation of 2D images/masks from the predicted 3D shape and the comparison to the ground truth reference images as supervision. These methods can be grouped by their representation of shape, including voxels [52, 45, 53], pointclouds [24, 17], meshes [21, 26] and implicit representation [37]. In contrast to these works, we allow only *a single image per object instance* in the training dataset.

Category Information. Few prior works have explored using category-specific priors for few-shot shape reconstruction [46, 31], where they assume voxel templates are available for each category. Our method uses category priors for shape

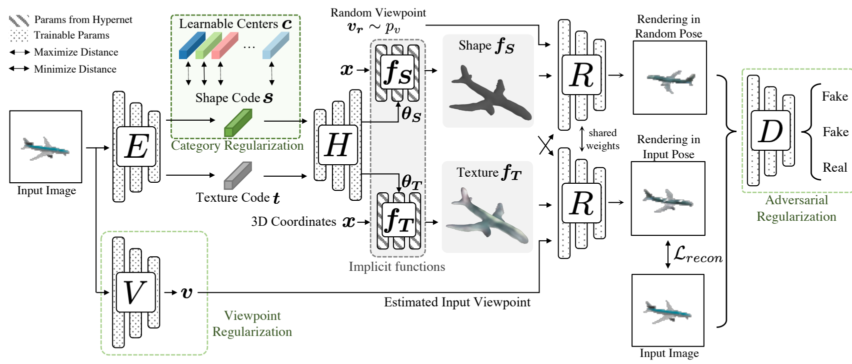


Fig. 2: **Approach Overview.** Given the input image, the encoders $\boldsymbol{E}$ and $\boldsymbol{V}$ predict the shape $\boldsymbol{s}$ and texture $\boldsymbol{t}$ codes and viewpoint $\boldsymbol{v}$. The hypernetwork $\boldsymbol{H}$ takes both latent codes and outputs the parameters $\boldsymbol{\theta_S}$, $\boldsymbol{\theta_T}$ of the implicit functions $\boldsymbol{f_S}$ and $\boldsymbol{f_T}$. These functions are sampled with a learnable ray tracer $\boldsymbol{R}$ to render the image with the predicted viewpoint $\boldsymbol{v}$ and thus to compute the reconstruction loss. We also render an extra image from a random viewpoint for the adversarial and viewpoint regularizations. Our shape metric learning technique is performed over the shape latent code.

reconstruction, but by only leveraging the significantly weaker supervision of category labels.

Deep Metric Learning. To harness shape learning with category labels, our method also shares ideas with metric learning. The key idea of metric learning is to learn an embedding space where similar instances are together and dissimilar instances are far according to some distance metric. This paradigm has shown success in self-supervised [3, 13] and supervised [22] learning, as well as many downstream tasks e.g. person re-identification [16]. We use the idea of metric learning for single-view shape reconstruction and demonstrate its effectiveness for the first time.

3 Approach

Given a collection of n images concatenated with their masks $\{I_i \in \mathbb{R}^{h\times w\times 4}\}_{i=1}^{n}$ and class labels $\{\mathbf{y}_i \in \{0, 1, \cdots, c\}\}$, our goal is to learn a single-view 3D reconstruction model without any 3D, viewpoint, or multi-view supervision. In contrast to most existing work, we learn a single network that works across all object categories. We represent shape using an implicit shape representation function $\boldsymbol{f_S} : \mathbb{R}^3 \to \mathbb{R}$, a multi-layer perceptron (MLP) that maps 3D coordinates to signed distance function (SDF) values $s \in \mathbb{R}$. Following a standard framework for learning 3D shape via differentiable rendering, our model first infers shape, texture and viewpoints from the input images, which are then rendered into images. The rendered image is then compared to the input image, providing a training signal for model learning. However, in our challenging multi-category setting, without viewpoint or 3D supervision, providing supervision only by comparing the rendered image and input images of the object results in poor performance. To mitigate this, we propose a set of regularization techniques which improve the reconstruction quality and stabilize model training. We first present an overview of the different modules of our approach in Section 3.1, and then introduce our category-based shape metric learning in Section 3.2. Finally, we present other regularization methods in Section 3.3. Implementation details and licenses are described in the supplement.

3.1 Network Overview

Our model is trained end-to-end and consists of four main modules (Fig. 2).

Image encoder. The image encoder $\boldsymbol{E}$ maps an image with mask channel $I \in \mathbb{R}^{h\times w\times 4}$ to a latent shape vector $\boldsymbol{s} \in \mathbb{R}^l$ for the downstream shape predictor, and a latent texture vector $\boldsymbol{t} \in \mathbb{R}^l$ for texture predictor.

Shape and texture prediction module. Following the design of hypernetworks [11, 25], the shape and texture prediction module uses latent codes $\boldsymbol{s}$ and $\boldsymbol{t}$ to predict the parameters of shape and texture implicit functions $\boldsymbol{f_S}$ and $\boldsymbol{f_T}$ that map 3D query points $\boldsymbol{x}$ to SDF and texture predictions:

$$\boldsymbol{\theta_S}, \boldsymbol{\theta_T} \leftarrow \boldsymbol{H}(\boldsymbol{s}, \boldsymbol{t}; \boldsymbol{\theta_H}) \tag{1}$$

$$\boldsymbol{f_S} : (\boldsymbol{x}; \boldsymbol{\theta_S}) - \text{SDF prediction} \tag{2}$$

$$\boldsymbol{f_T} : (\boldsymbol{x}; \boldsymbol{\theta_T}) - \text{Texture prediction} \tag{3}$$

$\boldsymbol{f_S}$ and $\boldsymbol{f_T}$ are MLPs with predefined structure and parameters estimated by the hypernet $\boldsymbol{H}$ with parameters $\boldsymbol{\theta_H}$.

Viewpoint prediction module. In our model, the viewpoint is represented as the trigonometric functions of Euler angles for continuity [1], i.e. $\boldsymbol{v} = [\cos\boldsymbol{\gamma}, \sin\boldsymbol{\gamma}]$ with $\boldsymbol{\gamma}$ denoting 3 Euler angles. The viewpoint prediction network $\boldsymbol{V}(I; \boldsymbol{\theta_V})$ predicts the viewpoint $\boldsymbol{v}$ of the object in the input image I with regard to a canonical pose (can be different from the human-defined canonical pose), where $\boldsymbol{\theta_V}$ are the learnable parameters.

Differentiable renderer. We use an SDF-based implicit differentiable renderer from [25] which takes the shape, texture and the viewpoint as inputs and renders the corresponding image. Formally, we denote it as a functional, $\boldsymbol{R}(\boldsymbol{f_S}, \boldsymbol{f_T}, \boldsymbol{v}; \boldsymbol{\theta_R})$, which maps shape SDF, texture implicit function and the viewpoint into 2D RGB image with an alpha channel, in $\mathbb{R}^{h\times w\times 4}$. The renderer itself is also learnable with $\boldsymbol{\theta_R}$ parameters. Note that as the renderer in [25] cannot render masks, we modify it to render an extra alpha channel from the SDF field following [54].

Before we describe different regularizations, we briefly discuss the major challenge in learning our model. During training, the only supervision signals are the input images and the masks of the objects. The common approach is to minimize a reconstruction loss via differentiable rendering[2]

$$\mathcal{L}_{recon} = \|I - \boldsymbol{R}(\boldsymbol{f_S}, \boldsymbol{f_T}, \boldsymbol{v}; \boldsymbol{\theta_R})\|_2^2. \tag{4}$$

This learning scheme works well with multi-view supervision [37], or even single-view images with viewpoint ground truth [25]. However, when there is only single-view supervision without viewpoint ground truth, this is significantly more difficult. There are infinite combinations of shapes and viewpoints that perfectly render the given image, and the model does not have any guidance to identify the correct shape-viewpoint combination. To tackle this problem, we utilize category labels to guide the learning of shape via metric learning. We also make use of other regularizations including adversarial learning and cycle consistency to facilitate learning.

3.2 Shape metric learning with category guidance

Our novel shape metric learning approach is the key ingredient that enables us to train a single model containing 55 shape categories. The key idea is to learn a metric which maps shapes within the same category (e.g. chairs) to be close to each other, while mapping shapes in different categories (e.g. planes vs. chairs) to be farther apart. This approach leverages qualitative label information to obtain a continuous metric for shape similarity, resulting in two benefits. First, the shape manifold defined by the learned metric constrains the space of possible 3D shapes, helping to eliminate spurious shape solutions that would otherwise be consistent with the input image (see Fig. 3 (a) for an example). Second, the shape manifold tends to group similar shapes together, creating a neighborhood structure over the training samples. For example, without category guidance it can be difficult for the network to learn the relationship between two images of different chairs captured from different, unknown, camera viewpoints. However, with the learned shape manifold these instances are grouped together, facilitating the sharing of supervisory signals from the rendering-based losses. A qualitative reconstruction comparison in Fig. 3 (a) and t-SNE visualizations of shape embedding space in Fig. 3 (b) provide further insight into these benefits. Note that as the image discriminator used for adversarial regularization uses images and labels as input, category information is still available during training even without metric learning. Therefore, the current comparison demonstrates the effectiveness of metric learning for further utilizing category label information.

Formally, for a batch of n images $\{I_i\}_{i=1}^n$, we extract both shape latent codes and texture latent codes via an image encoder $\boldsymbol{E}$. For the i^{th} sample, we denote the latent shape code as $\boldsymbol{s}_i$ and the category label as $y_i \in \mathcal{C}$ where $\mathcal{C}$ is the set of integer labels. There are two types of losses commonly used for metric learning: embedding losses [12, 48, 38] that compare different instances (e.g., triplet

[2] Masks are leveraged with additional losses that enforce consistency between the shape and the input mask. For conciseness, we omit them here and refer readers to [25] for more details.

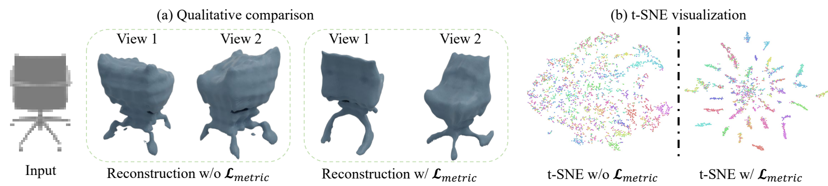


Fig. 3: (a) Category-guided metric learning is beneficial for eliminating erroneous shapes that can explain input image. In the first view (close to input viewpoint), both shapes looks plausible and can explain input image. But from another view, it is clear that our metric-learned shape is more reasonable as a chair, while the baseline can be closer to a table/nightstand. (b) t-SNE visualization of the shape embedding space on ShapeNet-55. Category-guided metric learning results in a more structured representation and facilitates joint learning.

loss), and proxy-based losses [47, 27, 56] which learn parameters representing categories (e.g. weight vectors in normalized softmax loss or ProxyNCA [34]). As the embedding losses usually require specific sampling strategies to work well, we follow the latter for metric learning in this work.

Specifically, for each category we define a learnable category center vector (or category embedding) $\{\boldsymbol{c}_k\}_{k=1}^{c}$ and optimize a distance metric so that instances of the same category are close and instances of different categories are far apart, as illustrated in the green shaded region of Fig. 2 and Fig. 1(b). We optimize the following loss between learned shape instance embeddings $\boldsymbol{s}$ and category center embeddings $\boldsymbol{c}$:

$$\mathcal{L}_{metric} = -\sum_{i=0}^{N} log \frac{exp(d(\boldsymbol{s}_i, \boldsymbol{c}_{y_i})/\tau)}{\sum_{k \in \mathcal{C}} exp(d(\boldsymbol{s}_i, \boldsymbol{c}_k)/\tau)}, \tag{5}$$

where d is a similarity measure and τ is the temperature. Following normalized softmax loss [47] we use cosine similarity as a similarity measure. The centers are randomly initialized from a uniform distribution and updated directly through gradient descent. We set the temperature τ to be 0.3 across all our experiments.

3.3 Shape and viewpoint regularization

Adversarial regularization. To further facilitate learning, we use adversarial regularization to ensure that predicted shapes and texture fields will render realistic images from any viewpoints in addition to the input view. Given the training image collection, we estimate the data manifold via an image discriminator. If the rendered images also lie on the training data manifold, most erroneous shapes and texture predictions will be eliminated. To achieve this, similar to [15, 55], we use a discriminator with the adversarial training [9]. Furthermore, because our multi-category setting results in a complex image distribution, we condition our discriminator on category labels for easier training following [33]. While our work is not the first work to use GANs for shape learning in general, we believe it still provides useful new insights by combining adversarial training with unconstrained implicit representation and class conditioning under our challenging MCSV setting.

Formally, suppose $I_{recon} = \boldsymbol{R}(\boldsymbol{f_S}, \boldsymbol{f_T}, \boldsymbol{v})$ is the output of the renderer for estimated shape, texture and viewpoint given an input image. We sample another random viewpoint $\boldsymbol{v}'$ from a prior distribution $P_v(\boldsymbol{v})$ and render another image, $I_{rnd} = R(\boldsymbol{f_S}, \boldsymbol{f_T}, \boldsymbol{v}')$ from this viewpoint. We match the distribution of I and I_{recon}, I_{rnd} with adversarial learning as shown in Fig. 2. Specifically, we optimize the objective $\mathcal{L}_{gan}$ below by alternatively updating our model (the $\boldsymbol{E}, \boldsymbol{H}, \boldsymbol{R}, \boldsymbol{V}$ networks) and the discriminator $\boldsymbol{D}(I; \boldsymbol{\theta_D})$

$$\min_{\boldsymbol{\theta_E}, \boldsymbol{\theta_H}, \boldsymbol{\theta_V}, \boldsymbol{\theta_R}} \max_{\boldsymbol{\theta_D}} \mathcal{L}_{gan} = \mathbb{E}[\log \boldsymbol{D}(I)] + \mathbb{E}[\log(1 - \boldsymbol{D}([I_{recon}, I_{rnd}]))]. \quad (6)$$

Here, $[I_{recon}, I_{rnd}]$ is the stack of reconstructed and randomly rendered images (on batch dimension). For the update step of the reconstruction model, we follow a non-saturating scheme [9] where we maximize $\mathbb{E}[\log(\boldsymbol{D}([I_{recon}, I_{rnd}]))]$ instead of minimizing $\mathbb{E}[\log(1-\boldsymbol{D}([I_{recon}, I_{rnd}]))]$. We also use R_1 regularizer [29] and spectral normalization [32] to stabilize the training process.

Viewpoint regularization via cycle-consistency. We regularize the viewpoint prediction module via cycle-consistency [58], with a similar approach as the state-of-the-art at self-supervised viewpoint estimation [35] and Ye et al. [55]. However, they both rely on a strong shape symmetry assumptions to facilitate the view learning. In our approach, we do not use any such restrictive assumptions. The key idea of viewpoint regularization, $\mathcal{L}_{cam}$, is that we can render an arbitrary number of images by sampling viewpoints with a given shape and texture. This provides us many rendered image-viewpoint pairs to supervise the viewpoint predictor. Please see the supplement for a formal description of $\mathcal{L}_{cam}$.

4 Experiments

This section presents the experiments of applying our method to multiple synthetic and real datasets, as well as ablations of its individual components. Following an overview of the datasets, metrics and baselines, we give results on synthetic and real data.

An additional goal is to quantify the value of category labels in the MCSV setting. By comparing category label-based supervision with a standard two-view reconstruction approach, we find that the value of the category label is approximately equal to 20% of an additional view. We also show that models trained under MCSV setting can outperform and generalize better than category-specific models. See the supplement for details.

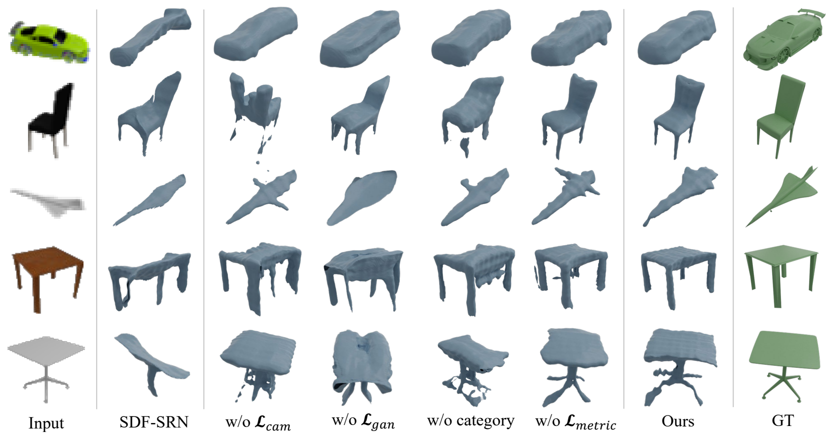


Fig. 4: Qualitative comparison on ShapeNet-13. Our method learns both better global structures and details on various categories.

4.1 Datasets

For synthetic data we use two splits of the ShapeNet [2] dataset, and for real data we use Pascal3D+[50]. We provide additional results on Pix3D [41] in the supplement.

ShapeNet-13. This dataset consists of images from the 13 biggest categories of ShapeNet, originally used by Kato et al. [21] and following works learning 3D shape without explicit 3D supervision [37, 25]. 24 views per object are generated by placing cameras at a fixed 30° elevation and with azimuthal increments around the object of 15°. Following SDF-SRN [25], we use a 70/10/20 split for training, validation and testing, resulting in 30643, 4378 and 8762 objects respectively. We use only one view per object instance, which is randomly pre-selected out of the 24 views for each object.

ShapeNet-55. To create a more challenging setting in comparison to prior work, we use all of the categories of ShapeNet.v2. This is challenging due to (1) approximately four times more shape categories; and (2) uniformly sampling at random the azimuth in the $[0°, 360°]$ range and elevation in $[20°, 40°]$. As a result, the image and shape distributions that the model needs to learn are significantly more complex. To reduce the data imbalance, we randomly sample at most 500 objects per category. Like ShapeNet13 we use only one image per object instance and a 70/10/20 split, with 15031, 1705, 3427 images respectively.

Pascal 3D+. This dataset contains real-world images with 3D ground truth obtained by CAD model alignment. Its challenges compared to ShapeNet come from (1) inaccurate object masks that include noisy backgrounds—an especially difficult setting for adversarial learning; and (2) the viewpoints vary in azimuth, elevation and tilt, creating challenges because different categories have different and unknown viewpoint distributions. We combine the commonly used motorcycle and chair categories for a total of 2307 images, and as in [25] we use the ImageNet [5] subset of Pascal3D+ with 1132/1175 images for training/testing.

4.2 Evaluation Metrics

We use Chamfer Distance (CD) and F-score under different thresholds to evaluate shape reconstruction performance following [25, 10, 42, 43]. We use the Marching Cubes algorithm [28] to convert the implicit representation to meshes prior to computing the metrics. Shapes are transformed into camera coordinates for evaluation. See the supplement for evaluation details.

Chamfer Distance. Following [37, 25], CD is defined as an average of accuracy and completeness. Given two surface point sets S_1 and S_2, CD is given as:

$$d_{CD}(S_1, S_2) = \frac{1}{2|S_1|} \sum_{x \in S_1} \min_{y \in S_2} \|x - y\|_2 + \frac{1}{2|S_2|} \sum_{y \in S_2} \min_{x \in S_1} \|x - y\|_2 \qquad (7)$$

F-score. Compared to CD which accumulates distances, F-score measures accuracy and completeness by thresholding the distances. For a threshold d (F-Score@d), precision is the percentage of predicted points that have neighboring ground truth points within distance d, recall is the percentage of ground truth points that have neighboring predicted points within d. F-score, the harmonic mean of precision and recall, can be intuitively interpreted as the percentage of surface that is correctly reconstructed.

4.3 Baselines

There are no directly comparable state-of-the-art methods under our MCSV setting. Instead, we slightly alter the setting of two recent methods SDF-SRN [25] and Ye et al. [55] to compare with our model.

SDF-SRN [25] learns to predict implicit shapes using single images and known camera poses in training. For comparison purposes, we attach our viewpoint prediction module to SDF-SRN and train it across multiple categories simultaneously. We compare to SDF-SRN on all datasets.

Ye et al. [55] learns category-specific voxel prediction without viewpoint GT.[3] We compare our method with Ye et al. on Pascal3D+ (see supplement for additional ShapeNet-13 comparison). For Ye et al., we train different models for different categories following their original setting and compute the average metric over categories. We use GT masks as supervision.

4.4 ShapeNet-13

We perform experiments on ShapeNet-13 and show quantitative and qualitative results in Table 2 and Figure 4. See additional results in the supplement.

[3] Their method has an optional per-sample test-time optimization to further refine the voxels and convert into meshes. Using the authors' implementation of this optimization did not lead to improved results in our experiments, so we use the voxel prediction as the output for evaluation.

Table 2: Quantitative result measured by CD and F-score on ShapeNet-13. Our method performs favorably to baselines and other SOTA methods.

Methods	F-Score@1.0↑	F-Score@5.0↑	F-Score@10.0↑	CD↓
w/o category	0.1589	0.6261	0.8527	0.520
w/o $\mathcal{L}_{metric}$	0.1875	0.6864	0.8805	0.458
w/o $\mathcal{L}_{cam}$	0.1837	0.6741	0.8758	0.463
w/o $\mathcal{L}_{gan}$	0.1846	0.6437	0.8422	0.532
Ours	**0.2005**	**0.7168**	**0.8949**	**0.430**
SDF-SRN	0.1606	0.5441	0.7584	0.682

Ablation Study. We first analyze the results of ablating the multiple regularization techniques used in our approach. In Table 2, 'w/o category' shows the results of the model without any category information and 'w/o $\mathcal{L}_{metric}$' is the model with only GAN conditioning to leverage category labels, by removing the metric learning component (both use all losses except $\mathcal{L}_{metric}$). Comparing them to our full model, we clearly see our metric learning helps utilize category information in a more efficient way (w/o $\mathcal{L}_{metric}$ vs Ours). We also see the benefit of having category information for shape reconstruction tasks (w/o category vs Ours). We further ablate on our camera (w/o $\mathcal{L}_{cam}$ vs Ours) and GAN regularization (w/o $\mathcal{L}_{gan}$ vs Ours) and quantify their value. Our qualitative result verifies these findings as well. We include more quantitative findings about the value of category labels for shape learning in the supplement.
SOTA Comparison. Comparing with SDF-SRN in Table 2 we see that our approach improves on the adapted SDF-SRN by a 23.8% (w/o category) and 37.0% (with category, Ours) CD decrease. Qualitatively, our approach captures both the overall shape topology and details (legs of table in the bottom row) whereas SDF-SRN fails. These results demonstrate the effectiveness of our proposed model.

4.5 ShapeNet-55

Compared to ShapeNet-13, ShapeNet-55 is significantly more challenging due to both the number of categories and viewpoint variability. For results see Table 3 and Fig. 5 (a).
Ablation Study. On this dataset, we again ablate on the value of category labels and our proposed shape metric learning. Similar to ShapeNet-13, we clearly see our metric learning method utilizes category information more effectively than GAN conditioning only (w/o $\mathcal{L}_{metric}$ vs Ours in Table 3). Meanwhile, we demonstrate the huge benefit of having category information for shape reconstruction on this dataset (w/o category vs Ours, 32.4% CD decrease). This is much more significant than ShapeNet-13 because ShapeNet-55 has much higher data complexity, where a better structured shape latent space can be quite beneficial to shape learning. Qualitatively, without category regularization the model cannot even capture the global shape successfully, whereas without shape metric

learning it only partially succeeds at highly symmetric objects (two vases in Figure 5 (a)). For others (e.g. bench at row 2) its reconstruction can only explain the input view. This further strengthen the discussion in Fig. 3 (a), where shape metric learning can help eliminate erroneous shape predictions.

Table 3: Quantitative result measured by CD and F-score on ShapeNet-55. Our method performs favorably to baselines and other SOTA methods.

Methods	F-Score@1.0↑	F-Score@5.0↑	F-Score@10.0↑	CD↓
w/o category	0.0977	0.4365	0.6815	0.801
w/o $\mathcal{L}_{metric}$	0.1431	0.5758	0.8047	0.620
Ours	**0.1619**	**0.6164**	**0.8386**	**0.541**
SDF-SRN [25]	0.0707	0.2750	0.4806	1.172

SOTA Comparison. Comparing with row 4 of Table 3 we see that our approach improves on the adapted SDF-SRN by 31.6% (without category) and 53.8% (with category) CD decrease. The significant performance improvement justifies the effectiveness of our overall framework, as well as the effectiveness of the categorical shape metric learning. Qualitatively, SDF-SRN collapses and fails to capture the topology of most inputs, while our method with still demonstrate a great qualitative reconstruction performance. More qualitative results on various categories and reconstructed textures are included in the supplement.

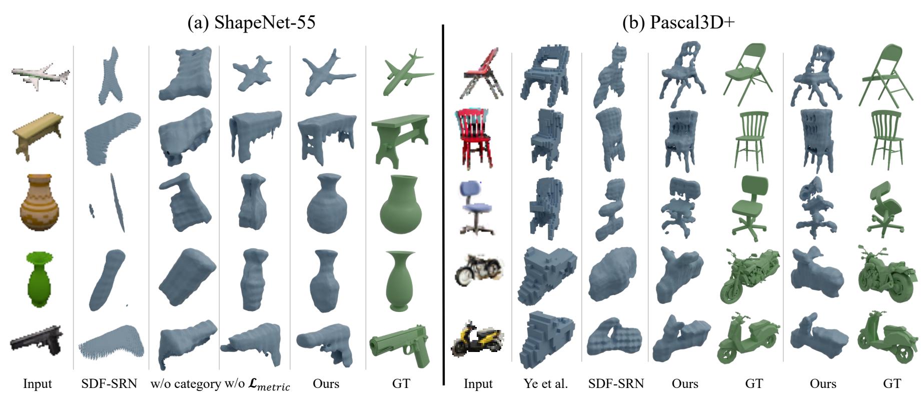


Fig. 5: Qualitative comparison on ShapeNet55 and Pascal3D+. Our method learns both better global 3D structure and shape details on various categories.

4.6 Pascal3D+

We compare our method to SDF-SRN [25] and Ye et al. [55] on Pascal3D+, as shown in Table 4 and Fig. 5 (b). It is clear that our method outperforms the

SOTA on both metrics, and our metric learning method is critical to the good performance. We also compare reconstructions qualitatively. As shown in Fig. 5 (b), SDF-SRN collapses to a thin flake that can only explain input images, while Ye et al. lack some important details. This further verifies the effectiveness of our proposed method.

Table 4: Quantitative result measured by CD and F-score on Pascal3D+. Our method performs favorably to other SOTA methods.

Methods	F-Score@1.0↑	F-Score@5.0↑	F-Score@10.0↑	CD↓
SDF-SRN [25]	0.0954	0.4656	0.7474	0.777
Ye et al. [55]	**0.1195**	0.5429	0.8252	0.625
Ours w/o $\mathcal{L}_{metric}$	0.1038	0.5108	0.7799	0.673
Ours	0.1139	**0.5460**	**0.8455**	**0.580**

5 Limitation and Discussion

One limitation is that reconstruction accuracy decreases for some samples with large concavity. For example, for bowls or bathtubs on ShapeNet-55, our method cannot capture the concavity of the inner surface. Concavity is hard to model without explicit guidance, and this can potentially be improved by explicitly modeling lighting and shading. On the other hand, ShapeNet-55 represents a class imbalance challenge, where images in different categories range from 40 to 500 images. This makes learning on rare classes difficult.

We also observe a training instability of our model caused by the usage of adversarial regularization. Meanwhile, we think our shape metric learning can be further improved by explicitly modeling the multi-modal nature of some categories. This could be achieved by using several proxies for each category. We leave this to future work. Please see the supplement for more discussion and qualitative examples on limitations.

6 Conclusion

We have presented the first 3D shape reconstruction method with an SDF shape representation under the challenging *multi-category, single-view (MCSV)* setting without viewpoint supervision. Our method leverages category labels to guide implicit shape learning via a novel metric learning approach and additional regularization. Our results on ShapeNet-13, ShapeNet-55 and Pascal3D+, demonstrate the superior quantitative and qualitative performance of our method over prior works. Our findings are the first to quantify the benefit of category information in single-image 3D reconstruction.

Acknowledgments. We thank Miao Liu for helpful comments and discussions.

References

1. Beyer, L., Hermans, A., Leibe, B.: Biternion nets: Continuous head pose regression from discrete training labels. In: German Conference on Pattern Recognition. pp. 157–168. Springer (2015)
2. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
4. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: European conference on computer vision. pp. 628–644. Springer (2016)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Gadelha, M., Maji, S., Wang, R.: 3d shape induction from 2d views of multiple objects. In: 2017 International Conference on 3D Vision (3DV). pp. 402–411. IEEE (2017)
7. Gkioxari, G., Malik, J., Johnson, J.: Mesh r-cnn. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
8. Goel, S., Kanazawa, A., Malik, J.: Shape and viewpoint without keypoints. In: European Conference on Computer Vision. pp. 88–104. Springer (2020)
9. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. arXiv preprint arXiv:1406.2661 (2014)
10. Groueix, T., Fisher, M., Kim, V.G., Russell, B., Aubry, M.: AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2018)
11. Ha, D., Dai, A., Le, Q.V.: Hypernetworks. arXiv preprint arXiv:1609.09106 (2016)
12. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1735–1742. IEEE (2006)
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
14. Henderson, P., Ferrari, V.: Learning single-image 3d reconstruction by generative modelling of shape, pose and shading. International Journal of Computer Vision pp. 1–20 (2019)
15. Henzler, P., Mitra, N.J., Ritschel, T.: Escaping plato's cave: 3d shape from adversarial rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9984–9993 (2019)
16. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
17. Insafutdinov, E., Dosovitskiy, A.: Unsupervised learning of shape and pose with differentiable point clouds. arXiv preprint arXiv:1810.09381 (2018)
18. Kanazawa, A., Tulsiani, S., Efros, A.A., Malik, J.: Learning category-specific mesh reconstruction from image collections. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 371–386 (2018)

19. Kar, A., Häne, C., Malik, J.: Learning a multi-view stereo machine. arXiv preprint arXiv:1708.05375 (2017)
20. Kato, H., Harada, T.: Learning view priors for single-view 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9778–9787 (2019)
21. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3907–3916 (2018)
22. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. arXiv preprint arXiv:2004.11362 (2020)
23. Li, X., Liu, S., Kim, K., De Mello, S., Jampani, V., Yang, M.H., Kautz, J.: Self-supervised single-view 3d reconstruction via semantic consistency. In: European Conference on Computer Vision. pp. 677–693. Springer (2020)
24. Lin, C.H., Kong, C., Lucey, S.: Learning efficient point cloud generation for dense 3d object reconstruction. In: AAAI Conference on Artificial Intelligence (AAAI) (2018)
25. Lin, C.H., Wang, C., Lucey, S.: Sdf-srn: Learning signed distance 3d object reconstruction from static images. arXiv preprint arXiv:2010.10505 (2020)
26. Liu, S., Li, T., Chen, W., Li, H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7708–7717 (2019)
27. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017)
28. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. ACM siggraph computer graphics **21**(4), 163–169 (1987)
29. Mescheder, L., Geiger, A., Nowozin, S.: Which training methods for gans do actually converge? In: International conference on machine learning. pp. 3481–3490. PMLR (2018)
30. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4460–4470 (2019)
31. Michalkiewicz, M., Parisot, S., Tsogkas, S., Baktashmotlagh, M., Eriksson, A., Belilovsky, E.: Few-shot single-view 3-d object reconstruction with compositional priors. In: European Conference on Computer Vision. pp. 614–630. Springer (2020)
32. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018)
33. Miyato, T., Koyama, M.: cgans with projection discriminator. arXiv preprint arXiv:1802.05637 (2018)
34. Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 360–368 (2017)
35. Mustikovela, S.K., Jampani, V., Mello, S.D., Liu, S., Iqbal, U., Rother, C., Kautz, J.: Self-supervised viewpoint learning from image collections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3971–3981 (2020)
36. Navaneet, K., Mathew, A., Kashyap, S., Hung, W.C., Jampani, V., Babu, R.V.: From image collections to point clouds with self-supervised shape and pose net-

works. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1132–1140 (2020)
37. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3504–3515 (2020)
38. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
39. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2304–2314 (2019)
40. Simoni, A., Pini, S., Vezzani, R., Cucchiara, R.: Multi-category mesh reconstruction from image collections. In: 2021 International Conference on 3D Vision (3DV). pp. 1321–1330. IEEE (2021)
41. 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 (CVPR) (2018)
42. Tatarchenko, M., Richter, S.R., Ranftl, R., Li, Z., Koltun, V., Brox, T.: What do single-view 3d reconstruction networks learn? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3405–3414 (2019)
43. Thai, A., Stojanov, S., Upadhya, V., Rehg, J.M.: 3d reconstruction of novel object shapes from single images. arXiv preprint arXiv:2006.07752 (2020)
44. Tulsiani, S., Kulkarni, N., Gupta, A.: Implicit mesh reconstruction from unannotated image collections. arXiv preprint arXiv:2007.08504 (2020)
45. Tulsiani, S., Zhou, T., Efros, A.A., Malik, J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2626–2634 (2017)
46. Wallace, B., Hariharan, B.: Few-shot generalization for single-image 3d reconstruction via priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3818–3827 (2019)
47. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: Normface: L2 hypersphere embedding for face verification. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1041–1049 (2017)
48. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. Journal of machine learning research **10**(2) (2009)
49. Wu, S., Rupprecht, C., Vedaldi, A.: Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1–10 (2020)
50. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond pascal: A benchmark for 3d object detection in the wild. In: IEEE winter conference on applications of computer vision. pp. 75–82. IEEE (2014)
51. Xu, Q., Wang, W., Ceylan, D., Mech, R., Neumann, U.: Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. arXiv preprint arXiv:1905.10711 (2019)
52. Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. arXiv preprint arXiv:1612.00814 (2016)
53. Yang, G., Cui, Y., Belongie, S., Hariharan, B.: Learning single-view 3d reconstruction with limited pose supervision. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 86–101 (2018)

54. Yariv, L., Kasten, Y., Moran, D., Galun, M., Atzmon, M., Basri, R., Lipman, Y.: Multiview neural surface reconstruction by disentangling geometry and appearance. arXiv preprint arXiv:2003.09852 (2020)
55. Ye, Y., Tulsiani, S., Gupta, A.: Shelf-supervised mesh prediction in the wild. arXiv preprint arXiv:2102.06195 (2021)
56. Zhai, A., Wu, H.Y.: Classification is a strong baseline for deep metric learning. arXiv preprint arXiv:1811.12649 (2018)
57. Zhang, X., Zhang, Z., Zhang, C., Tenenbaum, J.B., Freeman, W.T., Wu, J.: Learning to reconstruct shapes from unseen classes. arXiv preprint arXiv:1812.11166 (2018)
58. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)