Pre-train, Self-train, Distill: A simple recipe for Supersizing 3D Reconstruction

Kalyan Vasudev Alwala[1]   Abhinav Gupta[12]   Shubham Tulsiani[2]

[1]Meta AI Research   [2]Carnegie Mellon University
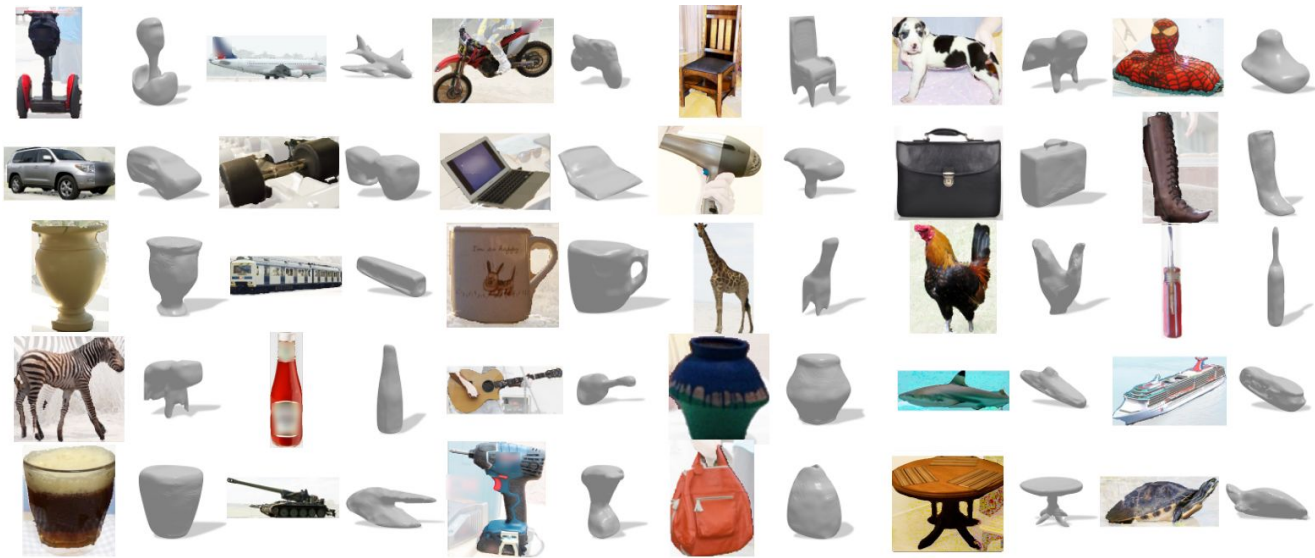
https://shubhtuls.github.io/ss3d/

Figure 1. We present an approach to learn a *single* reconstruction model across diverse object categories. Given an input image depicting any (segmented) object spanning over 150 categories, this unified reconstruction model can infer its 3D shape.

Abstract

Our work learns a unified model for single-view 3D reconstruction of objects from hundreds of semantic categories. As a scalable alternative to direct 3D supervision, our work relies on segmented image collections for learning 3D of generic categories. Unlike prior works that use similar supervision but learn independent category-specific models from scratch, our approach of learning a unified model simplifies the training process while also allowing the model to benefit from the common structure across categories. Using image collections from standard recognition datasets, we show that our approach allows learning 3D inference for over 150 object categories. We evaluate using two datasets and qualitatively and quantitatively show that our unified reconstruction approach improves over prior category-specific reconstruction baselines. Our final 3D reconstruction model is also capable of zero-shot inference on images from unseen object categories and we empirically show that increasing the number of training categories improves the reconstruction quality.

1. Introduction

We live in a rich, diverse world comprising of a variety of objects, from naturally occurring birds and bears, to ingeniously engineered cars and airplanes, or the pragmatically simple cups and vases. While we have made significant advances in semantic understanding of the visual world, developing approaches which can recognize hundreds and thousands of these objects, the state-of-the-art single-view reconstruction systems are only capable of inferring 3D for a handful among them. Why is that?

The primary reason for success in case of semantic recognition is the use of supervision. Availability of large amounts of training data for thousands of categories for classification [4], detection [26] and even segmentation [9] has been the key to progress. On the other hand, obtaining 3D supervision for images of generic objects is extremely hard. Even the biggest 3D datasets contain only tens of categories and even in that case, the images do not reflect the complexity of the real-world.

To tackle this problem of supervision, a number of recent works have instead investigated learning 3D inference
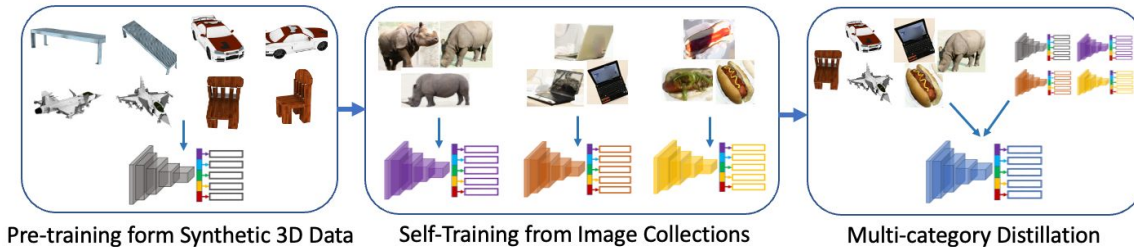
Figure 2. **Approach Overview.** We first pre-train a reconstruction model using multi-view renderings of synthetic data. We then self-train category-specific models on diverse image collections in-the-wild with only foreground mask annotations. We finally distill the learned models from prior training stages into a unified reconstruction model.

using only image collections and 2D annotations (such as foreground masks) as supervisory signal. However, as these approaches learn category-level models from scratch, they have no 3D priors and need to leverage additional constraints or regularization to make learning tractable. Unfortunately, this either restricts their applicability to generic classes *e.g.* commonly used morphable models cannot capture categories with large variation, or makes the training process unstable *e.g.* when using adversarial constraints.

Most approaches in the area have adopted one of the two extremes: supervised or self-supervised and therefore compromised either on scalability or generalization. In this paper, instead of adopting two extreme viewpoints – we adopt the middle ground. More specifically, we adopt three key strategies to achieve both scalability and tractability of learning 3D reconstruction. First, we observe that while semantic recognition approaches train one model to detect hundreds of categories; single-view reconstruction approaches train one model for each category. This bottlenecks the sharing of primitives and features across multiple categories. We argue this sharing is critical for generalization. In this work, we propose to train a *single* model to reconstruct objects across hundreds of diverse object categories. Second, instead of learning this joint model via regularization or extra constraints, we use synthetic 3D training data for *some* categories to help us learn priors. More specifically, we pre-train a base model using the synthetic data in a supervised manner. Third and finally, we observe that a pre-trained model cannot be used to just fine-tune on all the new categories in one shot. This is because mask-supervision is weak supervision at best and joint learning leads to averaging errors. Instead, we propose a self-training label propagation procedure: first fit a category specific model followed by a distillation process to learn a joint model that provides better generalization.

Our final approach is combination of pre-training and self-training that provides robustness and scalablilty. We first pretrain a 3D reconstruction model using the 3D synthetic data. This pre-training stage helps the model learn right 3D priors to learn from weak mask supervision. One

we learn.a pre-trained base model, we then adapt this base model trained on synthetic data to reconstruct novel categories from image collections, and show that this obviates the need for additional regularization such as limited intra-class variation or adversarial priors. In the final step, we then distill the per-category adapted models into a unified high-capacity 3D reconstruction network which is finally able to reconstruct objects from across hundreds of object categories (see Fig.1). Interestingly, not only can this unified model better reconstruct the categories seen during training, it can even be used for *zero-shot* 3D prediction for objects from *unseen* categories – something that prior per-category learning methods are fundamentally incapable of.

2. Related Work

Learning 3D from Supervision. With the resurgence of neural networks, several works have leveraged deep learning techniques for single-view 3D reconstruction. Using ground-truth 3D as supervision, prior approaches have pursued inference of representations such as voxels [2, 7] , meshes [42], Point Clouds [6], octrees [10] or implicit functions [30, 33]. However, due their reliance on 3D ground truth data, these methods mostly rely on synthetic training data and are limited in their scalability to generic object categories. To overcome this reliance on 3D supervision, subsequent approaches [29, 40, 43] instead used multi-view image collections to provide supervisory signal. However, as multi-view data is also difficult to acquire in-the-wild, other methods proposed to learn 3D from single-view images, although relying on additional annotations such as camera poses [12, 13, 20, 28] or semantic keypoints [18, 23]. While these works have all demonstrated impressive results, they critically require some form of annotations (ground truth 3D, multiple views, 2D keypoints, or camera poses) that are hard to acquire for generic classes. While our approach similarly relies on synthetic 3D data for a base set of classes to bootstrap learning, we show that 3D inference for *novel* categories can be learned without any such annotations, and thus allows our method to scale to hundreds of object classes using in-the-wild image collections.

	SRN-SDF [28]	CMR [18]	U-CMR [8]	IMR [39]	UMR [27]	SSMP [46]	Ours
camera	✓	✓					
template		✓	✓	✓			
keypoints		✓					
mask	✓	✓	✓	✓	✓	✓	✓
Per-Category Model		✓	✓	✓	✓	✓	
Number of Categories	<15	<5	<15	<25	<10	<75	>150
Shape-Regularizers		Def	Def	Def	Adv	Adv	

Table 1. A summary of the requirements and limitations of existing single-view 3D reconstruction approaches. Def and Adv denote morphable model and adversarial regularizers respectively.
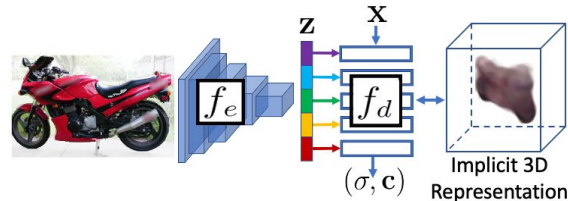

Figure 3. **Network Architecture.** Our image-conditioned implicit reconstruction network uses a ResNet-based encoder. The predicted encoding is used to conditionally modulate the outputs of the intermediate layers of a coordinate-based implicit network.

Learning 3D from Unannotated Image Collections. Closer to our goal of scalably learning 3D prediction, recent works have shown that single-view 3D inference can be learned from widely available category-level image collections, requiring only foreground masks as additional supervision. All these methods, whose training and inference setups we summarize in Table.1, learn 3D prediction by enforcing reprojection consistency with the available observations. While this removes the dependency on expensive annotations, these approaches require additional constraints to avoid degenerate solutions. For example, one common solution is to enforce that the variation within a category can be captured by a linear morphable model [8, 27, 39] and optionally use category-level templates [8, 24, 25, 39] or self-supervised semantics [27] to resolve pose ambiguities. Unfortunately, this prevents these approaches from reconstructing categories with non-spherical topology *e.g.* handbags or significant variation *e.g.* chairs. An alternative approach is to allow more expressive shape models and rely on adversarial training to encourage novel-view renderings to be realistic [14, 32, 46]. However, these methods can be more difficult to tune and also require a prior over training viewpoints. Our key insight is that these adhoc regularizers are required because all these methods learn per-category models from scratch. In contrast, we propose a simple and scalable approach for learning 3D that leverages synthetic data for pretraining and learns a unified model across object categories, thus allowing us to learn 3D for novel categories without requiring any such explicit regularizers.

Implicit Shape Representations. Approaches tackling unsupervised 3D reconstruction typically infer meshes [8, 27, 39] or discrete volumes [32, 46], but these are limited in their expressiveness or require learned neural renderers to model appearance [32]. Inspired by their suitability for volume rendering and success in modeling complex scenes [31], we instead opt to use neural implicit functions to represent the 3D shape (and appearance). Although the more striking recent applications of these representations have been to model single instances [31, 36, 45], our approach requires modeling *different* objects via a single network, and we therefore use image-conditioned implicit networks. Prior works have proposed latent variable-based [30,

33], hyper-network based [37] or pixel-aligned [47] mechanisms for such conditioning, but we instead adapt a modulation mechanism commonly used in generative networks [1, 19]. While the choice of our representation is shared with these recent approaches, our work shows that these representations can be learned from unannotated image collections, and that a single conditional implicit network can capture 3D across several categories.

3. Approach

Our goal in this work is to learn a unified model that can infer 3D from a single image of any object from over hundreds of categories. As 3D supervision is naturally difficult to obtain for such diverse set of categories, our approach relies on single-view image collections with approximate foreground masks as a more scalable alternative. However, instead of solely relying on these image collections, our approach is driven by the insight that (synthetic) 3D data for *some* categories does exist. While image collections are perhaps the only scalable source for learning about categories like balloons, or bananas, or starfishes, we do have 3D data available for other categories like cars and chairs. To incorporate this observation, we propose a multi-stage training approach that leverages both, synthetic shape collections and in-the-wild image collections for learning 3D inference across a broad set of categories.

In the first stage of training (Section 3.1), we pretrain our image-conditioned implicit reconstruction network using synthetic textured 3D models for about 50 object categories. The availability of 3D models allows us to render multiple views of the same instance while also knowing the precise camera poses for each image, thus allowing us to train using a novel-view rendering objective. However, as this initial model is only trained with synthetic renderings on a limited set of categories, it unsurprisingly does not perform well in-the-wild. In the second stage of learning (Section 3.2), we finetune this initial model using in-the-wild category-level image collections. Although neither camera viewpoints nor multiple views are available for learning, we find that the generic 3D priors learned in the
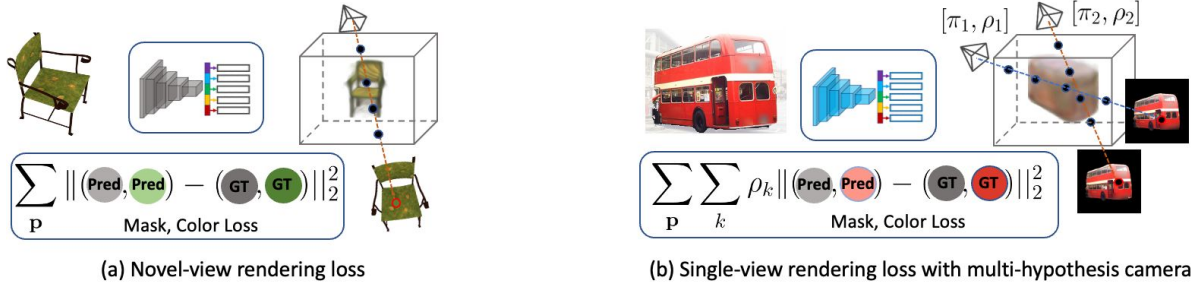
(a) Novel-view rendering loss

(b) Single-view rendering loss with multi-hypothesis camera

Figure 4. **Training Objectives. left:** When pre-training with synthetic data, we supervise the reconstruction network via a novel-view rendering loss. For each pixel/ray in a novel view, we volume-render the predicted mask and color using the implicit representation and penalize deviations from the ground-truth. **right:** To self-train a category-level expert from image collections, we minimize the expected rendering loss under a multi-hypothesis camera parametrization.
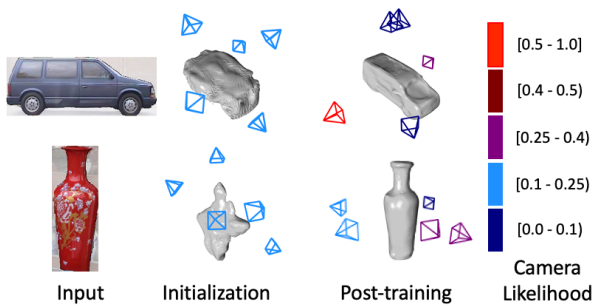

Figure 5. **Shape and Camera optimization in Self-training.** We visualize the initial and final configurations for two images. Although the symmetric vase maintains a multi-modal camera distribution, the shape for both instances improves with training.

first stage help prevent degenerate solutions. This yields multiple category-specific models, and we propose a simple technique to then distill these into a unified reconstruction model which can be used to perform 3D inference across generic classes (Section 3.3).

Image-conditioned Implicit Reconstruction Network. Our reconstruction model(s) can be viewed as having an encoder-decoder structure. Given an input image I, the encoder $f_e : I \rightarrow \mathbb{R}^d$ maps it to a latent code $\mathbf{z} = f_e(I)$. Our implicit decoder uses this latent code as a conditioning to predict the density $\sigma \in \mathbb{R}^1$ and color $\mathbf{c} \in \mathbb{R}^3$ for any query point $\mathbf{x}$ *i.e.* $f_d(\mathbf{x}, \mathbf{z}) = (\sigma, \mathbf{c})$. The image-conditioned decoder $f_d(\cdot, \mathbf{z})$ can thus be viewed as a implicitly representing the predicted geometry and appearance for the object depicted in the input image, and can be volume-rendered to generate images from any query viewpoint.

Our network architecture is highlighted in Figure 3. The encoder is based on a ResNet-34 [11] architecture and outputs a high-dimensional latent code given the input image. Inspired by Chan *et al.* [1] who used a similar architecture for unconditionally generating radiance fields, we use a SiREN [36] based decoder network with FiLM condition-

ing [34], where the parts of the latent code modulate the output of each decoder layer. Note that we do not condition the reconstruction on any explicit category labels, and the network is tasked to infer 3D for generic objects given only the (segmented) input image.

Volume Rendering. We denote by $\mathcal{V}(f, \pi, \mathbf{p})$ the process of volume rendering an implicit function f over the ray corresponding to a pixel $\mathbf{p}$ in an image taken from viewpoint π. Given a query ray, we follow the approach in NeRF [31] for differentiable volume rendering [5] with a neural representation by uniformly sampling 3D points along the ray, and aggregating the outputs using the predicted densities (and colors) to render. We use $\mathcal{V}_m$ and $\mathcal{V}_c$ to denote mask and color rendering respectively, and note the mask rendering process is equivalent to using a constant unit color at each 3D point. We refer the reader to [31] for more details.

3.1. Pre-Training from Synthetic 3D Data

We pretrain our implicit reconstruction network using a dataset of synthetic 3D shapes. As we have access to textured 3D meshes, we are able to render multiple views for each object with known camera viewpoints. We then train our reconstruction network by simply enforcing that the single-view (implicit) 3D predictions, when volumetrically rendered from the available viewpoints, match the known color and masks at each pixel.

Given an input image I depicting a *segmented* object, our network predicts an implicit 3D representation $f_I \equiv f_d(\cdot, f_e(I))$. We then use an image $\bar{I}$ of the same instance, captured from a novel camera viewpoint $\bar{\pi}$ to supervised our prediction. Denoting by $\bar{I}_m[\mathbf{p}]$ and $\bar{I}_c[\mathbf{p}]$ the observed color and foreground mask label for a pixel $\mathbf{p}$, and using $\{m, c\}$ as shorthand for indexing over mask and color, we train our network to minimize the reconstruction error between the predictions and the renderings across the dataset:

$$L_{\text{synth}} = \sum_{\mathbf{p}} \|\mathcal{V}_{\{m,c\}}(f_I, \bar{\pi}, \mathbf{p}) - \bar{I}_{\{m,c\}}[\mathbf{p}]\|_2^2$$

3.2. Self-Training from Image Collections

Our approach aims to adapt the pretrained implicit reconstruction network to generic novel categories. While the methodology described in Section 3.1 allows learning single-view 3D prediction, it crucially relies on available multi-view renderings and associated camera poses. However, when learning 3D reconstruction for generic categories in-the-wild, this form of supervision is not available. Instead, we must rely on single-view image collections with approximate foreground masks for training. Our approach is to derive supervisory signal by enforcing consistency between renderings of the predicted 3D and the input image, but we need to overcome the challenges posed by unknown camera viewpoints in order to do so.

Volume Rendering with Multi-hypothesis Cameras. When jointly learning (unknown) shape and camera viewpoints with differentiable rendering, the optimization is susceptible to local minima. To make the optimization landscape easier, several prior works [8, 17, 25, 38] have advocated using a *multi-hypothesis* viewpoint parametrization $\Pi \equiv \{(\pi_1, \rho_1), \ldots, (\pi_K, \rho_K)\}$, with a probability ρ_k associated with hypothesis π_k. Under this probabilistic parametrization for camera viewpoint, a rendering loss can be formalized as an *expected* loss – with the loss under each hypothesis weighted by its likelihood. Given an image I, and camera Π, the volume rendering loss w.r.t. an implicit representation f can be computed as:

$$L(f, I, \Pi) = \sum_{\mathbf{p}} \sum_{k} \rho_k \|\mathcal{V}_{\{m,c\}}(f, \pi_k, \mathbf{p}) - I_{\{m,c\}}[\mathbf{p}]\|_2^2$$

In practice, as volume rendering is computationally expensive, we approximate this loss by only sampling a small number of pixels/rays for each camera hypothesis.

Finetuning 3D Inference in-the-wild. Given a category-level collection of images with associated foreground masks $\mathcal{D} \equiv \{(I^n, I_m^n)\}$, we adapt the implicit reconstruction network learned in Section 3.1 to reconstruct instances in this collection. Inspired by work from Goel *et al*. [8], we first associate each image I^n with a randomly initialized multi-hypothesis camera Π^n. We then formulate our learning problem as that of jointly optimizing the 3D reconstruction network and the per-image camera hypotheses Π^n, and minimize the following objective:

$$L_{\text{img}} = \sum_{n} L(f_d(\cdot, f_e(I^n)), I^n, \Pi^n)$$

Similar to prior approaches using multi-hypothesis cameras [8, 25], we follow a two-step optimization approach where only the camera parameters are initially learned while keeping the 3D model fixed. Unlike U-CMR [8], we also optimize the probabilities ρ_k^n instead of treating these as
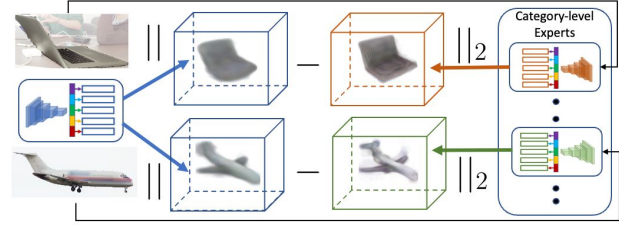

Figure 6. **Distillation.** We use a point-wise regression loss to distill multiple experts into a unified reconstruction model.

deterministic functions of the rendering loss. A more fundamental difference with the prior approaches [8,25] though is that we do not rely on a template 3D shape to make the camera optimization well-behaved. Instead, we find the shapes inferred by our pretrained reconstruction network are sufficient to guide initial camera optimization. After this initial optimization, we then jointly finetune the reconstruction network while continuing to optimize the cameras $\{\Pi^n\}$ over the image collection.

3.3. Multi-category Distillation

The approach described in Section 3.2 allows learning category-specific reconstruction networks from in-the-wild image collections. Denoting by $f^c \equiv (f_e^c, f_d^c)$ a category-specific reconstruction network learned for a category c, we now aim to learn a unified reconstruction network $\bar{f} \equiv (\bar{f}_e, \bar{f}_d)$ that can mimic the learned category-specific 'experts' across all categories $c \in \mathcal{C}$. Not only would such a unified network make inference easier, it could also improve performance by leveraging the common structure across classes, and perform better than individual networks, while also allowing *zero-shot* reconstruction for unseen classes.

We propose a simple approach of distilling [15] the per-category networks to learn a unified network (see Figure 6). Concretely, given the category-level image collections $\mathcal{D}^c$, we train our unified network to match the point-wise density and color predictions from the corresponding category-level experts across all images.

$$L_{\text{dist}} = \sum_{c \in C} \sum_{I \in \mathcal{D}^c} \sum_{\mathbf{x}} \|f_d^c(\mathbf{x}, f_e^c(I)) - \bar{f}_d(\mathbf{x}, \bar{f}_e(I))\|_2^2$$

While we find it beneficial to sample the points $\mathbf{x}$ along rays, the pointwise distillation objective allows efficient training without any computationally expensive volume rendering steps to aggregate the predictions.

4. Experiments

4.1. Datasets

Synthetic-data. We use a curated collection of CAD models from [16], spanning 51 object categories and a total of 40k instances used for training. We render color and

Figure 7. **Comparisons.** Visualization of 3D reconstructions predicted by the baselines and our final unified model on OpenImages, CUB, quadrapeds, Co3D and Pascal3D+. The first, second, third and fourth row of visualizations corresponds to predictions of SSMP [46], synthetic data-only model, category-level expert and our final unified model respectively.

mask images for each CAD model from 20 views using Blender [3], with azimuth and elevation angles randomly sampled from $[0°, 360°]$ and $[-75°, 75°]$ respectively.

In-the-wild Image Collections. As our approach in Section 3.2 can learn from image collections with approximate foreground masks, we are able to leverage various image recognition datasets for training. In particular, we use images from the following five datasets:

CUB-200-2011 [41]: This consists of 6,000 images with annotated foreground masks from over 200 bird subspecies. We treat the entire image collection as representing a single 'bird' category.

Quadrupeds from ImageNet: We use images spanning 25+ quadruped categories from the ImageNet dataset [4], along with approximate foreground masks extracted from an instance segmentation system [22]. In particular, we use the curated splits from Kulkarni *et al.* [24] where they discard images with severe occlusion and truncation, resulting in a total of 25k+ images.

PASCAL3D+ [44]: We use the unoccluded instances spanning 11 rigid-categories, with about 3k images per category. We use an off-the-shelf image segmentation system [22] to extract foreground masks for the ImageNet subset of images where ground-truth masks are not available. While this dataset provides annotations for approximate 3D shape in the form of template CAD models, we do not use these during training, and only use these for quantitative evaluation on held-out test images.

Open Images [26]: We use images from across 77 diverse categories, with 500 to 20k images with annotated foreground masks for each category. We adapt the train and test splits from Ye *et al.* [46], where a simple image classifier was used to filter out truncated and occluded instances.

4.2. Evaluation Setup

Training Details. All our networks share the same architecture. The encoder is a ResNet-34 [11] with an additional residual linear layer that produces $\mathbf{z} \in \mathbb{R}^{2560}$. Conditioned on $\mathbf{z}$, the decoder is a 5-layer FiLM-ed SIREN network [1] with an embedding size of 256 in each layer. We use Adam optimizer [21] with a Reduce-on-plateau learning rate scheduler. To ensure better convergence, while learning category-specific models in Section 3.1, we freeze the network for the first 10 epochs wherein we only optimize the multi-hypothesis camera parameters. For additional training details please refer to the supplementary material.

Baselines. We compare our final model with SSMP [46], the current state of the art method under similar data assumption as ours *i.e.* learning 3D for generic classes using image collections with approximate foreground masks. We also report the performance of intermediate stages of our approach – a) the synthetic data-only model (Synth.), and ii) category-specific experts (Cat-spec.).

Evaluation. To convert our neural implicit representation $f_d(\cdot, \mathbf{z})$, into an explicit mesh, we follow the Marching Cubes approach suggested by Mescheder *et al.* [30]. All our quantitative and qualitative evaluations are reported on this explicit representation. As we have access to (approximate) 3D ground truth meshes in Pascal3D+, we report intersection over union (IoU) on a held out test split. Following CMR [18], while computing IoU, we search over scale, rotate and position to align the predicted meshes with the ground truth. We also evaluate all approaches on Co3D [35], in which we have access to 3D ground truth point clouds. We evaluate on manually selected 100 occlu-

	SSMP [46]	Synth.	Cat-spec.	Ours
Aeroplane	0.38	0.44	0.32	**0.49**
Bicycle	**0.22**	0.11	0.18	0.18
Bottle	0.43	0.54	**0.66**	0.65
Bus	0.21	0.42	0.53	**0.57**
Car	0.46	**0.72**	0.54	0.70
Chair	0.44	0.52	0.57	**0.59**
D.Table	0.26	0.28	0.31	**0.34**
Motorbike	0.47	0.47	0.44	**0.53**
Sofa	0.24	0.43	0.46	**0.52**
Train	0.09	0.25	0.31	**0.37**
TV	0.17	0.19	0.27	**0.33**
Mean	0.31	0.40	0.42	**0.48**

Table 2. IoU metric comparison on Pascal3D+ [44]. Higher IoU implies superior 3D reconstruction.

	SSMP [46]	Synth.	Cat-spec.	Ours
Banana	0.256	0.409	0.265	**0.224**
Bottle	0.239	0.182	**0.163**	0.186
C.Phone	0.208	0.251	0.184	**0.172**
Donut	0.183	0.440	**0.176**	0.181
Hydrant	0.184	0.198	0.182	**0.166**
Orange	0.103	0.307	0.089	**0.082**
Suitcase	0.162	0.310	**0.146**	0.149
T.Bear	**0.197**	0.427	0.239	0.245
Toaster	0.194	0.316	0.161	**0.153**
Vase	0.210	0.286	0.211	**0.197**
Mean	0.194	0.313	0.182	**0.176**

Table 3. Chamfer distance comparison on Co3D [35]. Lower Chamfer distance implies superior 3D reconstruction.

sion and truncation free instances spanning 10 categories. We report Chamfer distance [6] between ground truth point cloud and points uniformly sampled from the predicted surfaces. We follow a similar approach to Pascal3D+ evaluation to align the predictions with ground truth point clouds.

4.3. Results

Quantitative Results. We report empirical results for Pascal3D+ and Co3D dataset in Tables 2 and 3 respectively, and note that our final model consistently improves over the prior state-of-the-art as well as the synthetic baseline. We also observe that the joint model yields some overall improvements over the per-category experts.

Qualitative Evaluation. We demonstrate the 3D reconstruction quality of our final unified model in a wide range of settings in Fig. 1, 9. We also visualize qualitative comparisons with baselines in Fig. 7, and find that these corroborate the quantitative analysis – that our unified model improves over the baselines and is capable of producing high-quality reconstructions for a wide variety of classes. We also depict some representative failure modes in Fig. 8.

Figure 8. **Failure modes.** Our model is not robust to occlusion and does not perform well on some difficult categories.

	Synth.	Ours-75+	Ours-125+
Banana	0.409	0.378	**0.304**
Donut	0.440	0.412	**0.244**
Orange	0.307	0.218	**0.213**
Suitcase	0.310	0.286	**0.189**
T.Bear	0.427	0.364	**0.320**
Toaster	0.316	0.194	**0.167**
Mean	0.368	0.308	**0.239**

Table 4. **Zero-shot reconstruction on unseen classes**. We report the Chamfer loss on unseen classes in Co3D dataset. As the model is trained over more classes, its zero-shot reconstruction performance on held-out unseen object classes improves.

Zero-shot Reconstruction. A benefit unique to learning a unified 3D reconstruction model is that, because of the common structure across generic objects, it can perform meaningful reconstruction even on *unseen* categories. We evaluate this ability of the final unified model to perform zero-shot reconstruction by varying the number of category-specific experts available to distill. We consider two unified models (Ours-75+) and (Ours-125+), which are distilled from over 75 and 125 experts respectively, but crucially, have not train on the evaluation categories. We summarize our observations in Table. 4, and find a clear and encouraging trend that observing more classes in training improves performance on *unseen* categories.

5. Conclusion

We presented a simple and scalable approach to learn a unified 3D reconstruction model across a diverse set of object categories. We achieved this by pre-training on synthetic data, which then allowed self-training on a variety of image collections. This simple strategy enabled us to learn a state-of-the art reconstruction model that can infer 3D for over hundred categories, and our results also highlight the empirical benefits of this sharing. While these results are encouraging, we note that our approach has some common failure modes. In particular, our reconstructions are unable to capture fine shape details. Our approach also crucially relies on images of unoccluded isolated objects for learning, and thus can neither learn from, nor infer 3D in more challenging images.

Figure 9. **In-the-wild Results.** Single-view 3D inference of our final unified model in the wild. We visualize from two viewpoints the meshes extracted from the predicted neural representation in a canonical frame.

References

[1] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Conference on Computer Vision and Pattern Recognition*, 2021. 3, 4, 6

[2] 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 *European*

Conference on Computer Vision, 2016. 2

[3] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 6

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009. 1, 6

[5] Robert A Drebin, Loren Carpenter, and Pat Hanrahan. Volume rendering. *ACM Siggraph Computer Graphics*, 1988. 4

[6] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Conference on Computer Vision and Pattern Recognition*, 2017. 2, 7

[7] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision*, 2016. 2

[8] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *European Conference on Computer Vision*, 2020. 3, 5

[9] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1

[10] Christian Häne, Shubham Tulsiani, and Jitendra Malik. Hierarchical surface prediction for 3d object reconstruction. In *International Conference on 3D Vision*, 2017. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016. 4, 6

[12] Paul Henderson and Vittorio Ferrari. Learning single-image 3d reconstruction by generative modelling of shape, pose and shading. *International Journal of Computer Vision*, 2020. 2

[13] Paul Henderson, Vagia Tsiminaki, and Christoph H Lampert. Leveraging 2d data to learn textured 3d mesh generation. In *Conference on Computer Vision and Pattern Recognition*, 2020. 2

[14] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *International Conference on Computer Vision*, 2019. 3

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 5

[16] Trimble Inc. 3D Warehouse. https://3dwarehouse.sketchup.com/. Accessed: 2021-06-25. 5

[17] Eldar Insafutdinov and Alexey Dosovitskiy. Unsupervised learning of shape and pose with differentiable point clouds. *Advances in Neural Information Processing Systems*, 2018. 5

[18] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *European Conference on Computer Vision*, 2018. 2, 3, 6

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks.

In *Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 3

[20] Hiroharu Kato and Tatsuya Harada. Learning view priors for single-view 3d reconstruction. In *Conference on Computer Vision and Pattern Recognition*, 2019. 2

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015. 6

[22] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Conference on Computer Vision and Pattern Recognition*, 2020. 6

[23] Chen Kong and Simon Lucey. Deep non-rigid structure from motion. In *International Conference on Computer Vision*, 2019. 2

[24] Nilesh Kulkarni, Abhinav Gupta, David F Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *Conference on Computer Vision and Pattern Recognition*, 2020. 3, 6

[25] Nilesh Kulkarni, Abhinav Gupta, and Shubham Tulsiani. Canonical surface mapping via geometric cycle consistency. In *International Conference on Computer Vision*, 2019. 3, 5

[26] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 2020. 1, 6

[27] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *European Conference on Computer Vision*, 2020. 3

[28] Chen-Hsuan Lin, Chaoyang Wang, and Simon Lucey. Sdfsrn: Learning signed distance 3d object reconstruction from static images. *Advances in Neural Information Processing Systems*, 2020. 2, 3

[29] Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3d supervision. *Advances in Neural Information Processing Systems*, 2019. 2

[30] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Conference on Computer Vision and Pattern Recognition*, 2019. 2, 3, 6

[31] 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 *European Conference on Computer Vision*, 2020. 3, 4

[32] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *International Conference on Computer Vision*, 2019. 3

[33] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Conference on Computer Vision and Pattern Recognition*, 2019. 2, 3

[34] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018. 4

[35] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 6, 7

[36] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 2020. 3, 4

[37] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 2019. 3

[38] Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Multi-view consistency as supervisory signal for learning shape and pose prediction. In *Conference on Computer Vision and Pattern Recognition*, 2018. 5

[39] Shubham Tulsiani, Nilesh Kulkarni, and Abhinav Gupta. Implicit mesh reconstruction from unannotated image collections. *arXiv preprint arXiv:2007.08504*, 2020. 3

[40] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Conference on Computer Vision and Pattern Recognition*, 2017. 2

[41] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 6

[42] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *European Conference on Computer Vision*, 2018. 2

[43] Olivia Wiles and Andrew Zisserman. Silnet: Single- and multi-view reconstruction by learning from silhouettes. *British Machine Vision Conference*, 2017. 2

[44] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *Winter Conference on Applications of Computer Vision*. IEEE, 2014. 6, 7

[45] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 2020. 3

[46] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *Conference on Computer Vision and Pattern Recognition*, 2021. 3, 6, 7

[47] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Conference on Computer Vision and Pattern Recognition*, 2021. 3