Few-Shot Generalization for Single-Image 3D Reconstruction via Priors

Bram Wallace
Cornell University
bw462@cornell.edu

Bharath Hariharan
Cornell University
bharathh@cs.cornell.edu

Abstract

Recent work on single-view 3D reconstruction shows impressive results, but has been restricted to a few fixed categories where extensive training data is available. The problem of generalizing these models to new classes with limited training data is largely open. To address this problem, we present a new model architecture that reframes single-view 3D reconstruction as learnt, category agnostic refinement of a provided, category-specific prior. The provided prior shape for a novel class can be obtained from as few as one 3D shape from this class. Our model can start reconstructing objects from the novel class using this prior without seeing any training image for this class and without any retraining. Our model outperforms category-agnostic baselines and remains competitive with more sophisticated baselines that finetune on the novel categories. Additionally, our network is capable of improving the reconstruction given multiple views despite not being trained on task of multi-view reconstruction.

1. Introduction

A key aspect of visual understanding is recovering the 3D structure of a scene. While classically such recovery of 3D structure has used multiple views of a scene, there has been recent research on 3D reconstruction *from a single image* using machine learning techniques. However, recovering 3D structure from a single image is a challenging learning problem. First, the output space is not just very large (e.g., represented as voxels, a $100 \times 100 \times 100$ grid is already a million-dimensional space) but also very *structured*: of all possible 3D shapes that are consistent with an image of a chair, a vanishingly small number are valid chair shapes. To perform well, the machine learning algorithm needs to capture a prior over possible chair shapes. Large, deep networks can indeed capture such priors when provided enough chairs for training, and this has been the dominant approach taken by prior work. However this leads to the second challenge: the cost of acquiring training data.

Training data for single view 3D reconstruction requires either 3D shapes [5] or at the very least multiple views of the same physical object [33]. Such training data can be acquired for a small number of categories, but is too expensive to obtain for every single object class we might want to reconstruct. Prior work attempts to circumvent this issue by training a *category-agnostic* model [32, 33], but such models might underperform due to ignoring category-specific structure in the output space. Therefore, we ask: is it possible to capture category-specific shape priors for single-image 3D reconstruction from very limited training data?

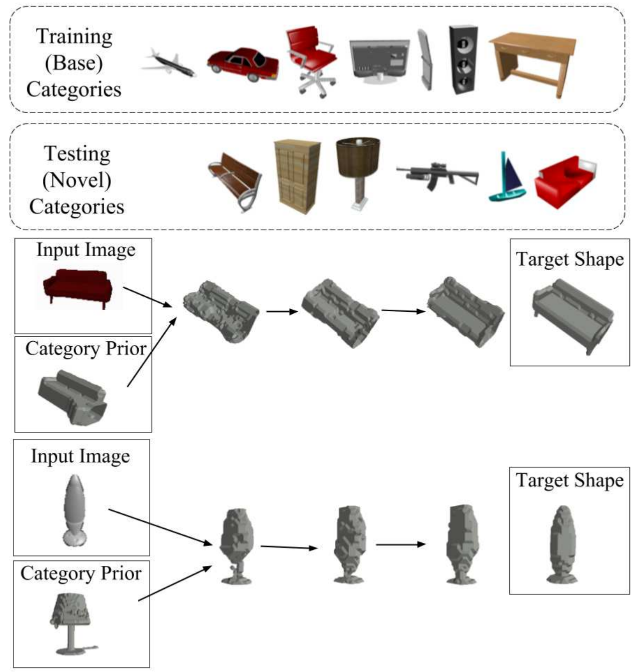

Figure 1. We train on 7 base categories and test the model's few-shot transfer ability on 6 novel categories. Our model takes in an image of the object to reconstruct along with a category-specific prior shape which can be as simple as a single novel class example. It then optionally iteratively refines this prior to produce a reconstruction.

In this paper, we show that the answer is yes. We present a simple *few-shot transfer learning* approach that can very quickly learn to reconstruct new classes of objects using very little training data. Instead of training a direct mapping from RGB images to 3D shapes, we train a model that uses image input to *refine* an input prior shape. This simple reparametrization allows us to swap in new priors for new classes *at runtime*, enabling single view reconstruction of novel object classes *with no additional training*. We show that this boosts reconstruction accuracy significantly over category-agnostic models.

We find an additional benefit to implementing the prior in this way: the output of our model can used as a new prior and fed back into the model to iteratively refine the prediction. While the notion of iterative prediction for better accuracy has been proposed before [19, 15, 2], the connection to few-shot learning in this context is new. We demonstrate that this iterative strategy can also be used out-of-the-box for competitive multi-view reconstruction *without any multi-view training*. Our approach is shown in Figure 1.

Summarizing the contributions of this paper, we:

1. Propose an augmented network architecture and training protocol that can incorporate prior categorical information at runtime
2. Demonstrate this network's ability on few-shot learning
3. Demonstrate this network's ability to perform competitive multiview reconstruction without being trained on the task

2. Related Work

Classically, the problem of 3D reconstruction has been solved using multiple views and geometric or photometric constraints [6, 13]. However, recently the success of convolutional networks on recognition tasks has prompted research into using machine learning for *single-view* 3D reconstruction. Early success in this paradigm was shown by R2N2 [5]. R2N2 iteratively refines a 3D reconstruction based on multiple views; this is similar in spirit to our approach of refining a prior shape, but the focus is on multi-view reconstruction and not generalization. Later work has since improved the underlying representation of 3D shapes [7, 9, 34, 23, 17, 24], replaced 3D training data with multiple calibrated views of each training object [20, 32, 33], incorporated insights from geometry to improve performance [10, 36], or made other improvements to the learning procedure [26, 22]. However, the question of generalizing to novel classes with limited training data is under-explored.

Work on generalization in the context of 3D reconstruction is limited. Recently Tatarchenko et al. demonstrated that single view 3D reconstruction models tend to memorize and retrieve similar shapes from the training set; an indication of overfitting [18]. This suggests that more generalizable models are necessary. Yang et al. are one of the first to attempt transfer learning for 3D reconstruction and find the best solutions to be using class-agnostic models and finetuning. [33]. We show that our approach outperforms both of these solutions when training data for novel classes is limited. Class agnostic models might be more generalizable if they incorporate geometrical constraints [36] or leverage pose information[11]. This idea of using geometry is orthogonal to, and indeed complementary to, our insight of separating out the category-specific prior.

The notion of using or learning priors has also been explored before. One approach to using priors is to use an adversary to enforce realistic reconstruction [28, 12]. Cherabier et al. use shape priors to learn from relatively little data, but focus on scene reconstruction with semantically labeled depth maps as inputs [3]. 3D-VAE-GAN is similar to our work in leveraging categorical knowledge[27]. Closer in spirit to our work are single-view reconstruction methods that use meshes as their underlying representation, which often function by deforming a prior mesh [23, 9] However, in all these approaches, the focus is on improving the in-category performance rather than on generalization or transfer; which are often not even evaluated. In contemporary work, Wang et al. propose to deform a source mesh to match a taarget shape, but their focus is on point cloud registration rather thaan single view reconstruction [25].

The approach we propose also has connections to models which use iterated inference in structured prediction problems. This idea was originally proposed for more classical approaches based on graphical models [19, 15] but has recently been applied to deep networks [2]. An iterative approach to single-view reconstruction is that of Zou et al., who build reconstructions via sequential concatenation of shape primitives [37]. Although shape primitives can sometimes lack expressivity for complex shapes, they also capture some priors about shape.

Our work is also related to few-shot transfer learning. Most prior work on the few-shot learning problem focuses on classification tasks. A large class of recent work in this domain uses the idea of meta-learning, where the model is trained using simulated few-shot learning scenarios it will encounter in deployment [21, 16, 30]. Our training procedure is similar in this regard, but focuses on structured prediction instead of classification. Some early work on few-shot learning also has the notion of separating out a class-specific mean shape from class-agnostic aspects of the classification problem [14], but again the key difference in this paper is the structured prediction problem.

3. Problem Setup

We are interested in learning single view 3D reconstruction for novel classes from very limited training data. We approach this broad problem through the lens of *few-shot* *learning* and *transfer learning*. We assume that we have a large dataset of 3D shapes with corresponding images for some set of classes, which we call *base classes* [8]. We will train our model on these base classes.

After training, the model will encounter *novel classes* for which we have very limited ground truth 3D shape information. In general, we will assume access to between 1 and 25 examples of 3D models for each class. Note that we do not assume that these 3D models come equipped with corresponding images; the model we propose below only uses the 3D models themselves to construct a category-specific prior. The model must use these example 3D models to reconstruct 3D shape for *test examples* of each class. The final performance metric will be its reconstruction accuracy on these test examples. In particular, we follow prior work and look at intersection-over-union between the predicted shape and the ground truth.

4. Approach

4.1. Model Architecture

We first create a *category-specific shape prior* in the form of a voxel grid by averaging the voxel representations for a small number of 3D shapes available for the novel class. Note that the individual voxels can take floating point values in this grid. We then design a *category-agnostic* neural network that *refines* this category-specific prior based on image input. This neural network uses two encoders to encode the image and the category prior into a common embedding space. The embeddings of the image and the category prior are added together and fed into a decoder that produces the refined shape.

This scheme for few-shot prediction offers several major advantages:

1. Very little runtime is required to incorporate the few-shot information. The shapes must simply be loaded and averaged, a negligible operation compared to the network's forward pass.
2. No retraining of the network is performed.
3. There is no difference in the predictive method for new or old categories.
4. Multiple types of priors can be incorporated in this fashion.
5. No corresponding images are required for transfer learning, only shapes. These might be obtained from CAD models created by designers.

Iterative prediction: Because our model refines an input shape, its output can be fed back in again to refine the shape further. Such iterative refinement has been shown to be useful for structured prediction problems [19, 15, 2]. We evaluate both iterative and non-iterative versions in our experiments.

Implementation details: The precise architecture is shown in Figure 2. The image encoder takes in a 127×127 RGB image as input and feeds it through a series of convolutions (3×3 except for an initial 7×7) alternating with max pooling layers and finishing with a fully connected layer. The shape encoder takes in the category prior as input. The shape encoder is a series of 3-dimensional convolutions followed by two dense layers. The image encoder is the same as that used by R2N2 and the shape encoder and decoder are similar to architectures employed by Yang et al [33]. The output of the two encoders are feature vectors of length 128 which are summed before being fed into the generator.

LeakyRelu is used in both encoders, with $\alpha = 0.01$ in the image encoder and $\alpha = 0.3$ for the shape half [31]. Traditional Relu was used in the shape generator. A sigmoid activation is applied at the final step of the shape generator.

4.2. Training

For every training datapoint, we sample an image from one of the base classes and the corresponding ground truth 3D shape as the target. Our secondary input, the prior shape, consists of an average of some number of other same-category shapes from the training set. For some models, this prior shape is the "Full Prior": all the shapes in the train dataset are averaged. When a "k-shot" prior is used, it consists of k averaged shapes, always from the training dataset. The "Full Prior" models always have the same initial input shape within a category while the "k-shot" prior networks use a different randomly generated prior for each image-target pair. We display the "Full Prior" shapes for each category in Figure 3. The loss is the binary cross-entropy loss.

Training iterative models: To train the model in an iterative setting, we repeat each training batch multiple times, with the model outputs of one iteration being fed as inputs in the next. For each batch from the generator, the same input images and target shapes are used each time and the input shapes change after the first step, being the output of the previous forward pass (Algorithm 1).

Implementation details: All experiments are done using Keras with a Tensorflow backend [4, 1]. Training is done in batches of 32 using an Adadelta optimizer [35]. Early stopping is used, with our metric of accuracy being the per-category average Intersection-over-Union (IoU) on the base classes with a threshold on the output of 0.4. This threshold is standard in literature, and in our case as well offered good performance. Relative performances of the models were maintained at different thresholds.

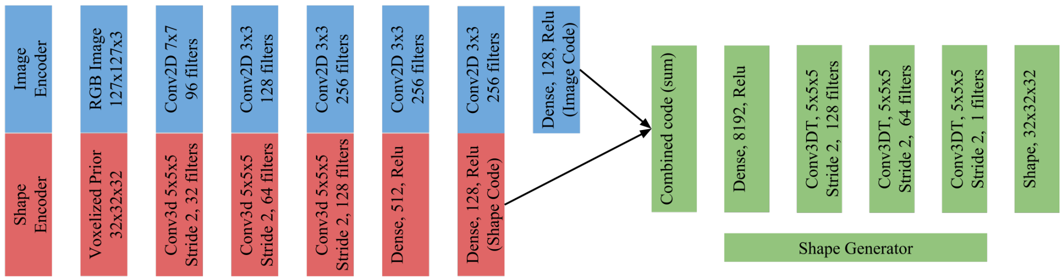

Figure 2. Our model is dual-input. The first input is an image encoded using the exact same architecture as 3D-R2N2 [5]. The second is a voxelized prior shape encoded via 3D convolutions, similar to Yang et al. [33]. The generator is similar to that of Yang et al. The 128-dimensional output of the encoders are summed. Each Conv2D layer is followed by 2x2 MaxPooling and LeakyRelu with $\alpha = 0.01$ and each Conv3D layer is followed by LeakyRelu with $\alpha = 0.3$. ReLu activations are used in the generator.

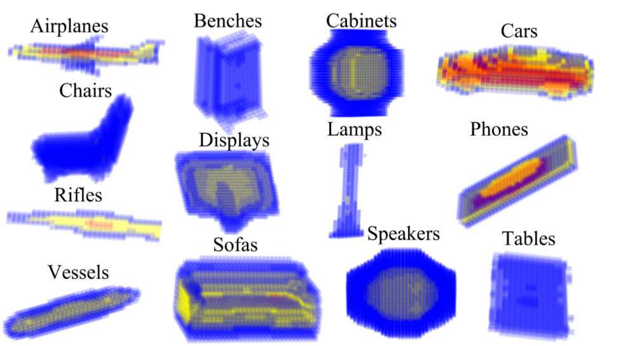

Figure 3. The averaged shapes of the entire training dataset for each category. The color represents the frequency of models in which a given gridpoint was occupied. Red indicates 90-100%, yellow 60-90% and blue 30-60%. We see that **airplanes**, **cars**, and **rifles** have an extremely consistent shape, while other categories such as **lamps** and **tables** have relatively weak priors, with no visible non-blue gridpoints.

Algorithm 1 Training for iterative refinement.

for epoch in epochs **do**
  for batch in batches **do**
    Load input images, input shapes, target shapes from generator
    for $iter\_i$ in $1..\#iters$ **do**
      Train on input images, input shapes, target shapes with backprop
      Set the input shapes equal to the output of the model
    end for
  end for
end for

5. Results

5.1. Experimental setup

We experiment with the ShapeNet dataset. Seven of the 13 categories are designated *base classes* and are used during training: **airplanes, cars, chairs, displays, phones, speakers,** and **tables** (matching the work of Yang et al [33]). We use 127×127 RGB rendered images of models and $32 \times 32 \times 32$ voxelized representations. Examples of the data as input-target pairs can be seen in Figures 1 and 5. Each model has 24 associated images from random viewpoints. We use the same training-testing split as R2N2 which was an 80-20 split. We further divide this into a 75-5-20 split to obtain a validation set.

When testing on base classes, we use the full prior unless otherwise noted. For novel-category testing, we always report the number of shapes being averaged into the prior. We consider both iterative and non-iterative models.

Baselines: We compare against multiple baselines. The first baseline is a category-agnostic mapping from images to 3D shapes. This model uses the same image encoder and shape decoder architecture, but does not use any category-specific prior as input or employ novel-category data at all. Such a category-agnostic model has been shown to perform very well in prior work [5, 33] and is thus a strong baseline. The second baseline *finetunes* the image-only model on the novel classes. k shapes are rendered from up to 24 viewpoints, resulting in between k and $24k$ image-pairs (depending on the model) which are then finetuned on. Note that this baseline uses paired images, which are not available to our approach. We finetune the models for 200 iterations using SGD with a learning rate of 0.005.

5.2. Main results

We first present results for our best model variant under the few-shot learning setting along with multiple baselines in Table 1. We vary the number of novel-class example

# Novel class Examples (k)	Image-Only Baseline	Finetune 1 Render	Finetune 5 Renders	Finetune 24 Renders	1-Iteration 1-Shot
1-shot	0.36	0.38	0.38	0.39	0.38
2-shot	0.36	0.38	0.39	0.40	0.39
3-shot	0.36	0.38	0.39	0.41	0.39
4-shot	0.36	0.39	0.40	0.42	0.39
5-shot	0.36	0.39	0.40	0.42	0.40
10-shot	0.36	0.39	0.42	0.44	0.40
Full Prior	0.36			0.40	

Table 1. Few-shot learning results on novel classes. The Image-Only Baseline does not incorporate new-category information at all. The "1-Iteration 1-shot" model is a non-iterative model trained with 1-shot priors and tested with priors consisting of k averaged shapes from the training category. We see that our model offers competitive performance, especially in very low-shot regimes, despite *no image supervision or retraining*. Scores reported are category-wise average IoU. The same Image-Only Baseline architecture achieved 0.55 IoU when trained on *all* of the classes at once. We perform 3-5 runs of each experiment with $\sigma_{IoU} < 0.01$.

shapes available and evaluate models on the average IoU across all novel classes. Figure 4 plots the performances of the models for priors containing varying amounts of information.

We observe that our best model variant (1 iteration trained on 1-shot priors) significantly outperforms the category-agnostic baseline across the board on the novel classes indicating the usefulness of the category prior. Compared to the finetuning-based approaches, our method outperforms the variant that sees one rendering per model, and is competitive with the variant that sees five images per model. Note that the finetuning approaches see significantly more information than our approach, which gets *no novel-class images at all*. Furthermore, unlike the finetuning approaches, our model *requires no retraining on the target class at all*. Any new class can thus be added to our models' repertoire simply by adding a corresponding prior. Further-

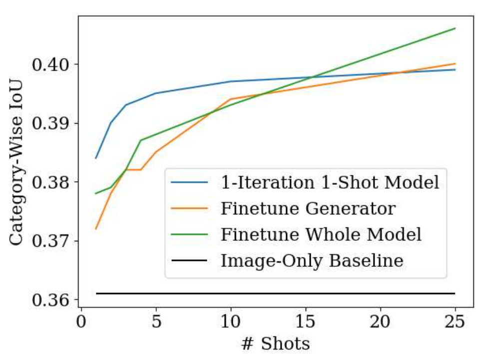

Figure 4. Performance of the 1-iteration 1-shot-trained model against various baselines tuned with 1 view per model. We see that the majority of improvement (60%) comes within the first 1 to 3 shots.

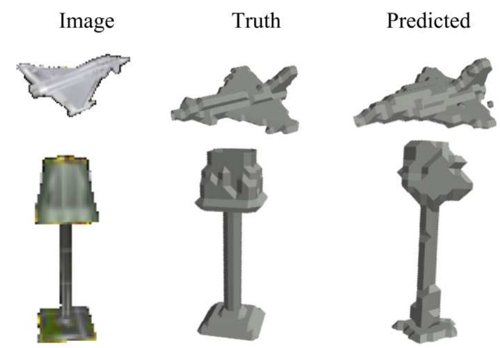

Figure 5. Examples of ground-truth and predicted shapes from an image. Note that the category **lamps** is not in our training set, we use a prior to enable generalization to this previously unseen category.

# Novel class Examples	1-Iteration Full Prior	2-Iteration Full Prior	3-Iteration Full Prior	2-Iteration 1-Shot Prior
1-shot	0.34	0.36/0.37	0.34/0.37/0.38	0.38/0.38
2-shot	0.36	0.38/ 0.38	0.37/0.39/0.38	0.39/0.38
3-shot	0.36	0.38/0.38	0.38/0.39/0.39	0.39/0.38
4-shot	0.36	0.39/0.38	0.39/0.39/0.39	0.39/0.38
5-shot	0.37	0.39/0.38	0.39/0.39/0.39	0.39/0.38
10-shot	0.37	0.39/0.39	0.40/0.40/0.39	0.39/0.38
25-shot	0.37	0.40/0.39	0.40/0.40/0.39	0.40/0.38
Full Prior	0.37	0.40/0.39	0.41/0.40/0.39	0.40/0.38

Table 2. Few-shot learning results on novel classes for additional model variants. Models are trained and tested for the same number of iterations. Setup as in Table 1. The best-performing iteration for each model is underlined.

more, as shown in Figure 4, we find that *very few novel-class shapes* are needed for this prior: with only 5 shapes, our model sees a gain of 4 points over the category-agnostic baseline. Example predictions are shown in Figure 5.

We include results from additional variants of our model in Table 2. We note that among the different model variants, models that perform iterated inference do not outperform the 1-iteration 1-shot model. Furthermore, for more informative priors, iteration buys no gain and sometimes even hurts the novel classes. Despite these shortcomings, we do find that they prove useful in the multi-view setting (Sec. 5.3).

In Table 3 we see that the improvements on novel classes do not come at the expense of performance within base classes. We also include the averaged performance of the R2N2 network as presented in the original work, to show that our baseline when trained on all 13 categories is slightly better, and thus a very strong control architecture to use.

5.3. Multi-view Reconstruction

In practice, it is often the case that we have more than one view of the object we want to reconstruct. Neural network

Training Procedure	Training Classes	Base-class performance
R2N2	All	0.58
1 Iteration No Prior	All	0.59
1 Iteration No Prior	Base	0.62
1 Iteration Full Prior	Base	0.63
2 Iteration Full Prior	Base	0.62/0.62
3 Iteration Full Prior	Base	0.61/0.61/0.61
1 Iteration 1-Shot Prior	Base	0.62
2 Iteration 1-Shot Prior	Base	0.61/0.61

Table 3. Training Category Results Summary. Models are tested on the test dataset of the training categories. The prior used is the same as during training. Our models perform comparably to an image-only baseline fitted on the training categories. This baseline outperforms R2N2 substantially, which we see is primarily due to the reduced categorical load.

approaches to multi-view reconstruction from uncalibrated views typically use recurrent neural networks as in R2N2. However, since our model is framed as *refining* a prior, we can use it iteratively, feeding in new images at each step.

Table 4 shows the performance of two of our best variants in the multi-view settings for both the base and the novel classes. We show both the non-iterative model trained on 1-shot priors (best performer in Table 1) as well as the 3-iteration model trained on the full prior. For the base classes, we use the full category prior and compare to R2N2 (with the caveat that R2N2 is trained on more classes). For the novel classes, we use a 1-shot prior.

We find that on the base classes, our 3-iteration model significantly improves on its single-view accuracy and achieves competitive performance to R2N2 *without any multi-view training*. Access to multiple views is even more beneficial for novel classes, where performance increases by close to 7 points. This again is despite not being trained on the multiview task, and only being given 1 example shape to learn from.

Interestingly, the non-iterative model is unable to benefit from the additional images. This suggests that when the target task requires iterative refinement, *training* for iterative refinement might be necessary, even if it is only single-view training.

5.4. Analysis

As shown above, our approach demonstrates strong performance on novel classes with very limited training data, for both single view and multiview reconstruction. We now perform a thorough analysis of our results, including the following questions:

1. How do the performance improvements break down over categories and over examples?

Method	# Views=1	2	3	4	5
Base classes					
R2N2 [5]	0.58	0.62	0.64	0.64	0.65
LSM [11]	0.60	0.71	0.75	0.77	-
3-Iteration	0.61	0.63	0.63	0.63	0.64
1-iteration 1-shot	0.62	0.62	0.62	0.62	0.62
Novel classes					
3-Iteration	0.34	0.38	0.40	0.40	0.41
1-iteration 1-shot	0.39	0.39	0.39	0.39	0.39

Table 4. Multi-view performance (IoU) on base categories (top) and novel categories (bottom). For base classes we compare to R2N2 (of which our architecture is an augmented version) and Learned Stereo Machines (an approach which uses provided pose information to backproject the pixels into a canonical, shared, reference frame. A full prior is used for the base classes and a 1-shot prior is used for the novel classes. The models iterative scheme can be adapted to multi-view reconstruction and shows substantial benefit despite not being trained on the task.

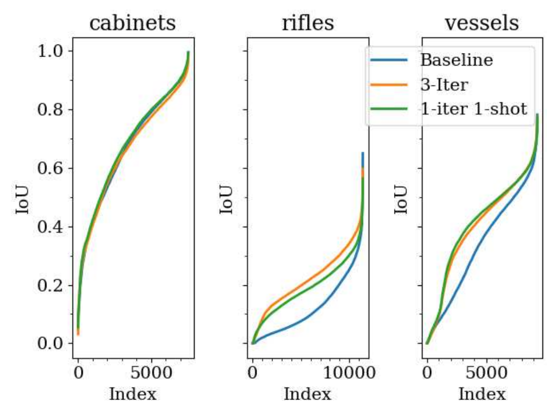

Figure 6. Here we plot the IoUs in increasing order for each model-category pair. We see that both of our new models substantially outperform the baseline on **rifles** and **vessels**. A 10-shot prior was used. Note that this is the same data as shown in Figure 7.

2. How important is the prior?
3. Can this approach be used on real-world images?

5.4.1 Analyzing Performance Distribution

While we have presented the average IoU on transfer learning tasks, this doesn't address the question of how these statistical results are achieved (e.g. translation of the distribution or a few exceptionally strong reconstructions). To determine the cause, we first look at the error distributions by plotting the IoUs for three categories and models in Figure 6. Here a 10-shot prior is used.

We see that increases in accuracy primarily come not from substantially increasing the number of highly accurately reconstructed shapes, but reducing the number of poorly reconstructed shapes. In **rifles** for example, the baseline has an IoU of less than 0.1 for over half the instances, whereas for our models this number is less than 17%.

Having analyzed the distribution of performances, we now graph the relations between model performances on the same input in Figure 7. We see that our models improve upon baseline performance for the vast majority of datapoints. We confirm that our new models mitigate many bad predictions, evidenced by clusters where the Baseline IoU is approximately 0.2 while our models achieve double that or more.

Figure 7 also shows an example instance for which the reconstruction changes significantly, and demonstrates the cause of this performance difference. **Vessels** are very elongated, and the only elongated category in the training set is **airplanes**. However, **airplanes** have wings and **vessels** do not. The baseline, relying on the prior it has learnt on **airplanes**, erroneously includes the wings in the reconstruction. In contrast, our model uses the provided prior to avoid this mistake. It is important to note that our model does this *without any retraining*, simply by virtue of disentangling the prior from other aspects of the reconstruction problem.

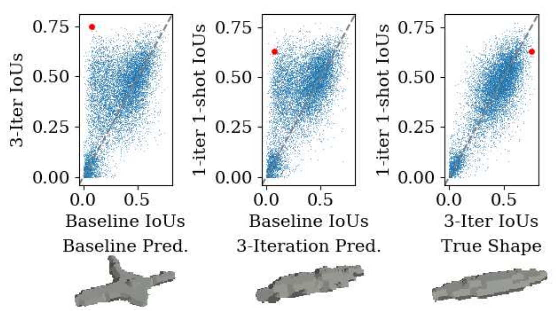

Figure 7. Scatter plots of model performances vs each other for **vessels**. Note that a point on the identity line has equal reconstruction IoU across two models. Predictions from the Baseline and 3-Iteration models for the red datapoint are shown in the bottom row. A 10-shot prior was used.

The previous discussion also suggests that improvements on different categories should vary depending on how far the class is from the set of base categories in terms of its distribution of shapes. To see if this is true, we present the per-category accuracies of the baseline, 3-iteration full-prior, and 1-iteration 1-shot models in Table 5. We see that both of the new models perform impressively on **rifles** and **vessels** and neutrally to poorly on **cabinets**. Referring back to the average shapes presented in Figure 3, we note that **vessels** and **rifles**, the two categories that our models perform best on, are both very elongated. The only elongated category in the training set is **airplanes**. Meanwhile, **cabinets** have a simple blocky prior. We hypothesize that this makes the prior less useful for learning, as such a basic shape is very simple to extrapolate from an image.

5.4.2 Importance of the prior

A neccessary question to ask when implementing a prior for a model is whether the observed performance stems from model or the prior itself. One could hypothesize that the improved results we see could be due to the model simply regurgitating the input prior. To test this, we performed experiments with a naive baseline that simply outputs the prior without taking into account the image at all. In the far-right column of Table 5, we show the average IoU for such a baseline using a 1-shot prior. We see that, while the performance of this naive guess is *correlated* with our model in terms of its difference from baseline performance, it performs significantly worse than both of our models. We also tested the performance of the naive prior guess with up to 25 shot priors, never observing category-wise IoU greater than 0.30. This shows that our model does provide valuable inference, and it is the combination of the prior with this inference that yields the performance.

At the other extreme of prior quality, we experimented with using the target shape as the input prior, where the 1-iteration 1-shot model achieved an IoU of 0.64 on the training categories and 0.41 on the novel categories This might be because the network is combining the provided prior with both the image information as well as general shape priors it has learnt from other classes, this is indeed intended behavior. Finally, we note that using different 1-shot shapes on the same image-target pair results in a score distribution with $\sigma \approx 0.05$.

Performance With Inaccurate Priors An assumption of our framework is categorical knowledge at runtime, allowing the selection of a prior shape. As we have shown, this assumption enables boosted performance on novel categories. In Figure 8 we perform experiments to observe what happens when that assumption breaks down. We run our model on the novel categories with priors drawn from

Category	Baseline	3-Iteration	1-Iteration 1-Shot	1-Shot Guess
Benches	0.37	0.39 (5.4%)	0.37 (0.0%)	0.13 (-64%)
Cabinets	0.66	0.62 (-6.5%)	0.66 (0.0%)	0.29 (-56%)
Lamps	0.18	0.19 (5.6%)	0.19 (5.6%)	0.11 (-40%)
Sofas	0.50	0.51 (2.0%)	0.52 (4.0%)	0.33 (-35%)
Vessels	0.33	0.37 (12%)	0.38 (15%)	0.22 (-34%)
Rifles	0.12	0.16 (33%)	0.19 (58%)	0.27 (120%)
Mean	0.36	0.37	0.39	0.23

Table 5. Per-category transfer performances. A 1-shot prior was used for both models. The far-right column is the result of naively guessing a random shape from the training set. The accuracy of our models are correlated with the accuracy of the 1-shot guess, yet avoid large errors when 1-shot guesses are very poor.

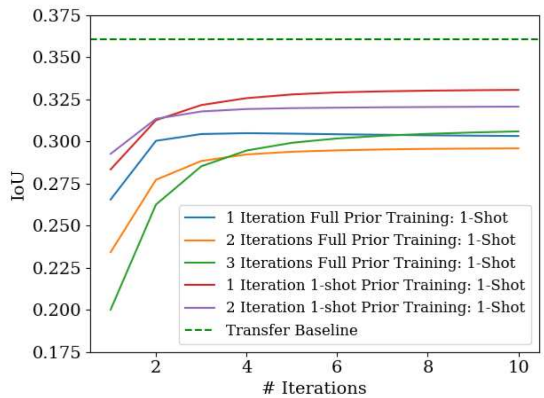

Figure 8. Performance across iterations with the model being fed a 1-shot prior from a random training/testing category. The green line is the transfer baseline of 0.36. We see that the models never achieve baseline performance, confirming the neccessity of categorical knowledge when implementing the presented framework.

other randomly selected categories.

We see that the models never achieve baseline performance, implying that categorical information is necessary to obtain the improvement that we have seen. This might be construed as a disadvantage of the presented framework, but it is also evidence that the model has disentangled the category-specific and the category-agnostic aspects of the problem and is relying more on the input prior to provide the category-specific information. In practice, given the advanced state of image classification, knowledge of the category at test time is a valid assumption. This assumption is in fact common in prior single-view reconstruction work [33, 7].

It is interesting to note that the performance of the 1-iteration model trained with 1-shot priors suffers substantially less than other models on the transfer task when given an incorrect prior. We hypothesize that, given the high variability of 1-shot input priors, this model has come to rely less on the prior than others.

5.4.3 Application to In-the-Wild Images

We finetune the 1-shot 1-iteration model on PASCAL 3D+ [29]. We train on all 13 ShapeNet categories and 7 of the 10 non-deformable PASCAL 3D+ categories. We hold out **bicycles, motorcycles** and **trains** as these categories are not present in the ShapeNet dataset. As seen in Table6 our model far outperforms the image-only architecture on both the training and testing categories. These results should be considered cautiously due to extremely low variation of PASCAL models, as noted in the original PASCAL 3D+ paper [29]. As observed by Tatarchenko et al., retrieval techniques work extremely well on PASCAL, explaining why a

Model	Training Categories (Validation)	Novel Categories
Image-Only	0.40	0.26
1-Shot 1-Iteration	0.50	0.37

Table 6. Results of finetuning a ShapeNet-trained model on the common categories of PASCAL3D+.

shape prior is so useful[18].

6. Future Work

The proposed idea of separating out the category-specific prior as an additional input can apply to other single-view reconstruction approaches using other representations of shape too. The prior can be derived from other sources, such as CAD models or geometry-based reasoning. The results of Tatarchenko et al. also suggest that a simple category-based approach can yield state-of-the-art results on reconstruction, implying a possible crossover between our technique and theirs. This viewpoint of separating out category-specific priors and learnt category-agnostic refinement can also be applied to many computer vision regression problems (e.g. segmentation or shape completion) that have had relatively little few-shot transfer work done on them.

7. Conclusion

In conclusion, we presented a new framework for 3D reconstruction that significantly improves generalization to new classes with limited training data, and offers multi-view reconstruction for free. Our models take two inputs: the typical image of the object to reconstruct along with a shape prior. Few-shot knowledge consisting of shape models can be used by inputting the average in as the prior. Such a model can then make iterative predictions by using its own output as a prior. Our model requires no novel class images and no retraining. We identified that our model offers far less extremely poor reconstructions than the baseline. We found that this framework peformed well on the multi-view reconstruction task. This finding in particular is surprising given that this model is never trained on multiview. The results here show that explicit categorical information and priors can be a powerful tool in 3D reconstruction.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fer-

nanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 3

[2] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 3

[3] Ian Cherabier, Johannes L Schonberger, Martin R Oswald, Marc Pollefeys, and Andreas Geiger. Learning priors for semantic 3d reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 314–330, 2018. 2

[4] François Chollet et al. Keras. https://keras.io, 2015. 3

[5] 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*, pages 628–644. Springer, 2016. 1, 2, 4, 6

[6] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *TPAMI*, 32(8):1362–1376, 2010. 2

[7] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *CVPR*, 2018. 2, 8

[8] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3

[9] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 2

[10] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *Advances in neural information processing systems*, pages 365–376, 2017. 2

[11] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. 2017. 2, 6

[12] Hiroharu Kato and Tatsuya Harada. Learning view priors for single-view 3d reconstruction. *CoRR*, abs/1811.10719, 2018. 2

[13] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International journal of computer vision*, 38(3):199–218, 2000. 2

[14] Erik G Miller, Nicholas E Matsakis, and Paul A Viola. Learning from one example through shared densities on transforms. In *CVPR*, 2000. 2

[15] Stephane Ross, Daniel Munoz, Martial Hebert, and J. Andrew Bagnell. Learning message-passing inference machines for structured prediction. In *CVPR*, 2011. 2, 3

[16] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. In *NIPS*, 2017. 2

[17] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2088–2096, 2017. 2

[18] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2019. 2, 8

[19] Zhuowen Tu. Auto-context and its application to high-level vision tasks. In *CVPR*, pages 1–8, 2008. 2, 3

[20] Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Multi-view consistency as supervisory signal for learning shape and pose prediction. In *CVPR*, 2018. 2

[21] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016. 2

[22] Meng Wang, Lingjing Wang, and Yi Fang. 3densinet: A robust neural network architecture towards 3d volumetric object prediction from 2d image. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 961–969. ACM, 2017. 2

[23] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67, 2018. 2

[24] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions on Graphics (TOG)*, 36(4):72, 2017. 2

[25] Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. 3dn: 3d deformation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1038–1046, 2019. 2

[26] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. In *Advances in neural information processing systems*, pages 540–550, 2017. 2

[27] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in neural information processing systems*, pages 82–90, 2016. 2

[28] 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 *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[29] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 75–82, 2014. 8

[30] Yu xiong Wang, Ross Girshick, Martial Herbert, and Bharath Hariharan. Low-shot learning from imaginary data. In *CVPR*, 2018. 2

[31] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015. 3

[32] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *NeurIPS*, pages 1696–1704, 2016. 1, 2

[33] Guandao Yang, Yin Cui, Serge Belongie, and Bharath Hariharan. Learning single-view 3d reconstruction with limited pose supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 86–101, 2018. 1, 2, 3, 4, 8

[34] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *CVPR*, 2018. 2

[35] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012. 3

[36] Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Josh Tenenbaum, Bill Freeman, and Jiajun Wu. Learning to reconstruct shapes from unseen classes. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2257–2268. Curran Associates, Inc., 2018. 2

[37] Chuhang Zou, Ersin Yumer, Jimei Yang, Duygu Ceylan, and Derek Hoiem. 3d-prnn: Generating shape primitives with recurrent neural networks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 2