Few-Shot Single-View 3-D Object Reconstruction with Compositional Priors

Mateusz Michalkiewicz[1], Sarah Parisot[2,5*], Stavros Tsogkas[3,4⋆], Mahsa Baktashmotlagh[1], Anders Eriksson[1], and Eugene Belilovsky[2]

[1] University of Queensland m.michalkiewicz@uq.net.au, {m.baktashmotlagh, a.eriksson}@uq.edu.au
[2] Mila, University of Montreal eugene.belilovsky@umontreal.ca
[3] University of Toronto tsogkas@cs.toronto.edu
[4] Samsung AI Research Center, Toronto
[5] Huawei Noah's Ark Lab. London sarah.parisot@huawei.com

Abstract. The impressive performance of deep convolutional neural networks in single-view 3D reconstruction suggests that these models perform non-trivial reasoning about the 3D structure of the output space. Recent work has challenged this belief, showing that complex encoder-decoder architectures perform similarly to nearest-neighbor baselines or simple linear decoder models that exploit large amounts of per-category data, in standard benchmarks. A more realistic setting, however, involves inferring 3D shapes for categories with few available training examples; this requires a model that can successfully *generalize* to novel object classes. In this work we experimentally demonstrate that naive baselines fail in this *few-shot* learning setting, where the network must learn informative shape priors for inference of new categories. We propose three ways to learn a class-specific global shape prior, directly from data. Using these techniques, our learned prior is able to capture multi-scale information about the 3D shape, and account for intra-class variability by virtue of an implicit compositional structure. Experiments on the popular ShapeNet dataset show that our method outperforms a zero-shot baseline by over 50% and the current state-of-the-art by over 10% in terms of relative performance, in the few-shot setting.

Keywords: 3D reconstruction, few-shot learning, compositionality

1 Introduction

Inferring the 3D geometry of an object, or a scene, from its 2D projection on the image plane is a classical computer vision problem with a plethora of applications, including object recognition, scene understanding, medical diagnosis,

* Stavros Tsogkas and Sarah Parisot contributed to this article in their personal capacity as an Adjunct Professor at the University of Toronto and Visiting Scholar at Mila, respectively. The views expressed (or the conclusions reached) are their own and do not necessarily represent the views of Samsung Research America, Inc and Huawei Technologies Co., Ltd.

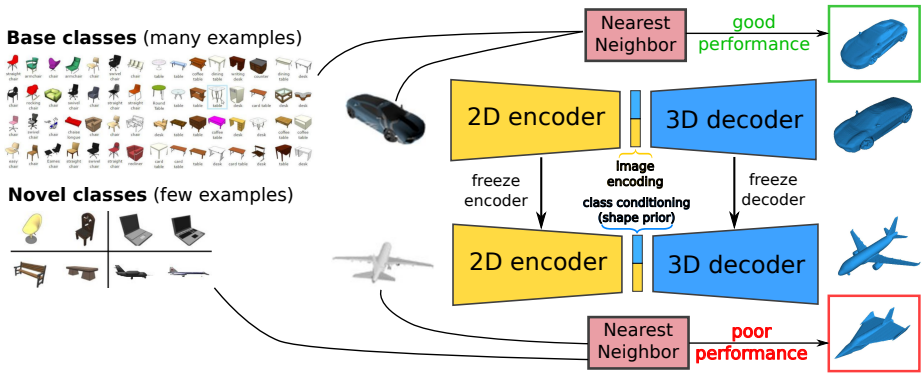


Fig. 1: We tackle the problem of single-view 3D reconstruction in the few-shot learning setup. [31] showed that naive baselines such as nearest neighbor, can outperform complicated models when data is abundant. However, such baselines cannot generalize to new classes for which only few training examples are available. We propose to use a deep encoder-decoder architecture whose output is conditioned on *learned* category-specific shape embeddings; our shape priors capture intra-class variability more effectively than previous works, significantly improving generalization.

animation, and more. After decades of research this problem remains challenging as it is inherently ill-posed: there are many valid 3D objects shapes (or scenes) that correspond to the same 2D projection.

Traditional multi-view geometry and shape-from-X methods try to resolve this ambiguity by using multiple images of the same object/scene from different viewpoints to find a mathematical solution to the inverse 2D-to-3D reconstruction mapping. Notable examples of such methods include [27,14,36,11,10].

In contrast to the challenges faced by all these methods, humans can solve this ill-posed problem relatively easily, even using *just a single image*. Through experience and interaction with objects, people accumulate prior knowledge about their 3D structure, and develop mental models of the world that allow them to accurately predict how a 2D scene could be "lifted" in 3D, or how an object would look from a different viewpoint.

The question then becomes: "how can we incorporate similar priors into our models?". Some early works rely on CAD models[13,24,33,37], while Xu et al. [40] use low-level priors and mid-level Gestalt principles such as curvature, symmetry, and parallelism, to regularize the 3D reconstruction of a 2D sketch. The downside of such methods is that they require an extremely specific specification of the model priors, which often limits their applicability.

Motivated by the success of deep convolutional networks (CNN) in multiple domains, the community has recently switched to an alternative paradigm, where more sophisticated priors are directly *learned from data*. The idea is straightforward: given a an appropriate set of paired 2D-3D data, one can train a model that takes as input a 2D image and outputs a 3D shape. Most of these works rely on an encoder-decoder architecture, where the encoder extracts a latent

representation of the object depicted in the image, and the decoder maps that representation into a 3D shape [26,3,16]. Many works have studied ways to make the 3D decoder more efficient and improve shape representation. The high quality outputs obtained suggest that, indeed, these models learn to perform non-trivial reasoning about 3D object structure.

Surprisingly, recent works [17,31] have shown that this is not the case. Tatarchenko et al. [31] argue that, because of the way current benchmarks are constructed, even the most sophisticated learning methods end up finding shortcuts, and rely primarily on recognition to solve single-view 3D reconstruction. Their experiments show that modern CNNs for 3D reconstruction are outperformed by simple nearest neighbor (NN) or classification baselines, both quantitatively, and qualitatively. Similarly, [17] showed that simple linear decoder models, learned by PCA, are sufficient to achieve competitive performance. There is one caveat though: to achieve good performance with these baselines, having a large dataset is crucial. More importantly, true 3D shape understanding implies good *generalization* to new object classes. This is trivial to humans –we reason about the 3D structure of unknown objects, drawing on our inductive bias from *similar* objects we have seen– but still remains an open computer vision problem.

Based on this observation, we argue that single-view 3D reconstruction is of particular interest in the few-shot learning setting. Our hypothesis is that learning to recover 3D shapes using few examples, while promoting generalization to novel classes, provides a good setup for the development and evaluation of models that go beyond simple categorization and actually learn about shape.

To the best of our knowledge, the first work of that kind is by Wallace and Hariharan [32]. Instead of directly learning a mapping from 2D images to 3D shapes, they train a model that uses features extracted from 2D images to refine an input *shape prior* into a final 3D output. Their framework allows one to easily adapt the shape prior and use it when inferring new classes. However, their approach has several restrictions. First, the shape prior for an object class is computed either by i) averaging the available examples for that class or ii) randomly selecting one of them. Both of these operations collapse intra-class variability, failing to fully exploit the already limited available training data. Second, the method does not explicitly force inter-class concepts to be learned.

In this work, we first demonstrate empirically that naive baselines that are quite effective for general single-view object reconstruction [31] come up short when generalizing to novel classes in a few-shot learning setup [32], highlighting the importance of this setup for the design and evaluation of methods with generalization capability. Furthermore, we address the shortcomings of [32] by introducing three strategies for constructing the shape prior, focusing on modelling intra-class variability, compositionality and multi-scale conditioning. More specifically, we first learn a shape prior that captures intra-class variability by solving an optimization problem involving all shapes available for the new class. We then introduce a compositional bias in the shape prior that allows learning concepts that can be shared across different classes or transferred to new ones.

Finally, we make use of conditional batch normalisation [22] to impose class conditioning explicitly at multiple scales of the decoding process.

In summary, we make the following contributions:

- We investigate the few-shot learning setting for 3D shape reconstruction and demonstrate that this setup constitutes an ideal testbed for the development of methods that reason about shapes.
- We introduce three strategies for shape prior modelling, including a compositional approach that successfully exploits similarities across classes.
- We conduct experiments demonstrating that we outperform the state of the art by a significant margin, while generalizing to new classes more accurately.

2 Related Work

2.1 Single-view 3D Reconstruction

Single- and multi-view 3D reconstruction have recently focused on improving learning efficiency and generation quality by finding better alternatives to the typically used 3D CNN decoder and voxelized shape representation [3,8,38,41,42]. Such alternatives include point clouds [5], meshes [34], and representations based on the signed distance transform [21,18,2]. Although each one of these representations has its pros and cons, [31,17] showed that they do not beat naive baselines such as nearest neighbor (NN) or linear decoder.

2.2 Few-shot Learning

Few-shot learning has become a highly popular research topic in computer vision and machine learning [25,7]. Most works focus on the classification task, with few investigating more complex problems such as segmentation [29] or object detection [39,35]. Our work considers the few-shot setting in the practical 3-d shape reconstruction which has only been considered in [32].Existing methods can be divided into two categories: meta-learning/meta-gradient based approaches [6], and metric-learning/prototype based approaches [23,30]. The former aims to teach models to adapt quickly, in a few gradient updates, to new unseen classes, while the latter learns a distance metric such that the distance of a query image to the few annotated examples of the same class is minimal.

3 Methods

Let $\mathcal{D}_b = \{(I_i^b, S_i^b)\}$ be a set of image-shape pairs, belonging to one of N_b *base* object classes. We assume that $|\mathcal{D}_b|$ is *large*, i.e., $\mathcal{D}_b$ contains enough training examples for our purposes. We also consider a *much smaller* set of *novel* classes, $\mathcal{D}_n^K$. Each class in $\mathcal{D}_n^K$ comprises only a small set of K image-shape pairs $\{(I_1^n, S_1^n), \ldots, (I_K^n, S_K^n)\}$, and a large set of test or query images.

Our objective is to use the abundant data in $\mathcal{D}_b$ to train a model that takes a 2D input image I, containing a single object, and outputs its 3D reconstruction,

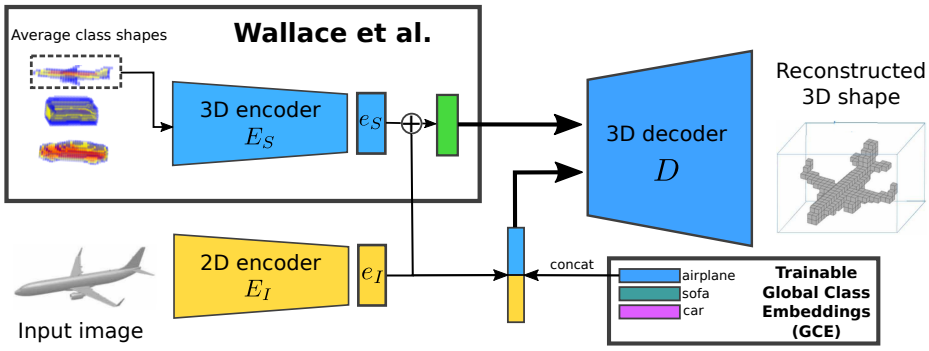


Fig. 2: Comparison of [32] to GCE. The former collapses variability of new classes by averaging. GCE is able to obtain a global shape representation for each class. Note that we combine e_I and e_S by concatenation, instead of element-wise sum.

$\bar{S}$. The model should also be able to leverage the limited data in $\mathcal{D}_n^K$ to successfully generalize to novel categories. Similar to previous works employing an encoder-decoder architecture, we choose voxels as our 3D shape representation (facilitating comparison to [32]) and propose *three* strategies to achieve this.

3.1 Shape Encoding and Global Class Embedding

Consider an encoder-decoder framework involving

- an encoder E_I that takes a 2D image, I, and outputs its embedding, e_I;
- a category-specific shape embedding, e_S;
- a decoder D that takes the image and shape embeddings and outputs the reconstructed 3D shape in the form of a voxelized 3D grid, $\bar{S}$:

$$\bar{S} = D\left(e_I, e_S\right) = D\left(E_I(I), e_S\right). \tag{1}$$

This model can be trained using a binary cross-entropy loss between the predicted occupancy confidence p_i at voxel i, and the respective label $y_i \in \{0, 1\}$ from a ground truth shape S with N_v voxels:

$$\mathcal{L}(S, \bar{S}) = -\frac{1}{N_v} \sum_{i}^{N_v} y_i \log(p_i) + (1 - y_i) \log(1 - p_i). \tag{2}$$

In the rest of the text, we drop S for notational simplicity.

Figure 2 (top) illustrates the pipeline of [32]. e_S^i is computed with a shape encoder E_S that takes a category-specific shape prior, S_i^p, as input; i.e., $e_S = E_S(S_i^p)$. For base class training, E_I, E_S, and D are learned by minimizing (2). For inference on new classes, the 3D shape is recovered simply by feeding the image and class specific prior (e_S) to the trained network. The shape prior S_i^p is defined either as a randomly selected shape from the training set $\mathcal{D}_n^K$ associated with class i, or the average, in voxel space, of all training shapes of class i.

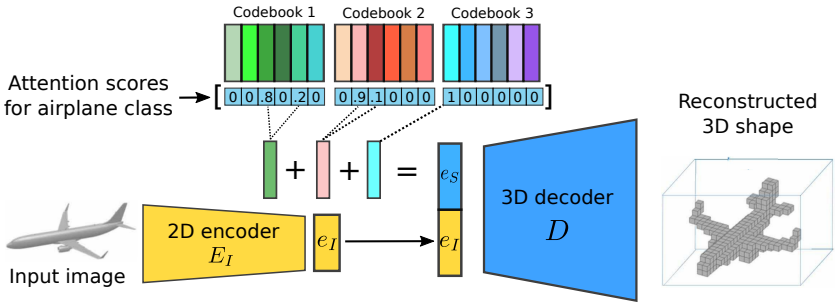


Fig. 3: Compositional GCE constructs a code by a composition of codes from different codebooks, applying a different attention to each codebook based on the class.

Both choices have severe limitations: they cannot account for intra-class variability and are, therefore, intrinsically sub-optimal when more than one training examples are available. To address this limitation, we propose to *learn* a global class embedding (GCE), e_S^i, that conditions the network for object class-i, but is dependent non-linearly on all available shapes.We expect this conditioning vector that is a derived from all shapes to capture nuances (like intra-class variability) more accurately than simple shape averaging.

Our framework is illustrated in Figure 2 (bottom). We first train the model on base classes, jointly optimizing the parameters of the encoder E_I, the decoder D, and the base class embeddings e_S^i, by minimizing the objective in Eq. (2). For novel classes with a small training set $\{(I_i^n, S_i^n)\}_{i=1}^K$, all model parameters of E_I and D are fixed, and class specific embeddings e_S^i are obtained by solving

$$\hat{e}_S^i = \arg\min_{e_S^i} \sum_{j=1}^{K} \mathcal{L}(D(E_I(I_j), e_S^i)). \tag{3}$$

Our approach enjoys the following practical advantages: First, the optimization problem in Eq. (3) can be solved in just a few iterations since it only involves a small set of parameters (e_S^i) and a small number of novel category samples. Second, the model can continually learn implicit shape priors for novel classes, *without compromising performance on the base classes*, by construction, since the weights of E_I and D are frozen. Finally, we note that we combine e_I and e_S by concatenation, instead of the element-wise sum used in [32].

3.2 Compositional Global Class Embeddings

GCE allows us to exploit all available training shapes to learn a representative shape prior for a specific object class. However, the learned global embeddings do not explicitly exploit similarities across different classes, which may result in sub-optimal, and potentially redundant representations. As a result, exploring inductive biases for sharing representations across classes has the potential to increase robustness in the lowest data regimes. To this end, we introduce an

extension of the GCE model, which we call Compositional Global Class Embeddings (CGCE), aiming to learn compositional representations between classes. This model is illustrated in Figure 3.

Our objective is to explicitly encourage the model to discover "concepts", representing geometric or semantic parts, that are shared across different object categories. Taking inspiration from work on compressing word embeddings [28], we propose to decompose our class representation into a linear combination of learned vectors that are shared across classes. More specifically, we learn a set of M codebooks (or embedding tables), with each codebook $\mathbf{C}_j$ containing m individual embedding vectors (*codes*) $\mathbf{C}_j = \{\mathbf{c}_{j,1}, \ldots, \mathbf{c}_{j,m}\}$, where $\mathbf{c}_{j,m} \in \mathbb{R}^D$. Intuitively, each codebook can be interpreted as the representation of an abstract concept which can be shared across multiple classes.

For each class i, we learn an attention vector $\boldsymbol{\alpha_i}$ that selects the most relevant code(s) from each codebook. A weighted sum of all codes yields the final embedding: $e_S^i = \sum_{k=1}^{M} \sum_{j=1}^{m} a_i^{k,j} c_{k,j}$, where $a_i^{j,k}$ is the scalar attention on code j at codebook k, while $c_{j,k}$ corresponds to the j^{th} code of codebook k. During base class training we learn both $\boldsymbol{\alpha}_i$ and $c_{j,k}$. We highlight that codebooks are shared across classes and therefore need only be trained on base classes. As a result, $\boldsymbol{\alpha}_i$ is the only class-specific variable we need to infer for novel classes:

$$\hat{\boldsymbol{\alpha}}_i = \arg\min_{\boldsymbol{\alpha}_i} \sum_{j=1}^{N} \mathcal{L}(D(E_I(I_j), e_S^i)).$$

A desirable property is that codebooks capture distinct and diverse class attributes and that they contain meaningful codes, with minimal redundancies. We encourage such behavior by having the model select only a sparse subset of codes from each codebook, using a form of attention that relies on the sparsemax operator [15]. Specifically, the attention vector for class i at codebook j is given by $\boldsymbol{a}_i^j = \textsc{sparsemax}(\boldsymbol{w}_i^j)$, where $\boldsymbol{w}_i^j$ are learned parameters; sparsemax produces an output that sums to 1, but will typically attend to just a few outputs.

3.3 Multi-scale Conditional Class Embeddings

The strategies proposed so far influence the shape reconstruction stage at the input level by combining the 2D image embedding with a learned shape prior. Another approach we propose to investigate is multi-scale conditioning throughout the decoding process. An elegant way to do this is by applying the conditional batch normalization technique [22] to the 3D decoder model. Conditional batch normalization replaces the affine parameters in all batch-normalization layers with layer-specific learned embeddings. Since 3D decoders have an inherently multi-scale structure with layers producing features at progressively higher resolutions, each layer's batch-norm parameters can be seen as conditioning/constraining the reconstruction process at different scales. Similarly to GCE, class-specific conditional batch normalization parameters are learned by fine-tuning the model on novel classes, keeping the encoder and decoder frozen. We refer to this approach as Multi-scale Conditional Class Embeddings (MCCE).

3.4 Nearest Neighbor Oracle, Zero-Shot and All-Shot Baselines

We introduce three simple baselines we use in our experiments. First, we consider an *oracle nearest neighbor (ONN)* [31] baseline. Given a query 3D shape, ONN exhaustively searches a shape database for the most similar entry with respect to a given metric (Intersection over Union in this case). Although this method cannot be applied in practice, it provides an upper bound on how well a retrieval method can perform on the task.

We also consider a zero-shot (ZS), and all-shot (AS) baseline. For the *ZS baseline*, we train the encoder-decoder model as described in Eq. (1) and use it to infer 3D shapes for novel classes, *without* using the category-specific shape prior e_S. We expect this to give a lower bound of performance, since it does not make any use of shape prior information. For the *AS baseline*, we merge the base class and novel class datasets, train the model on this joint dataset, and then test only on novel class examples. We expect that this baseline will set an upper bound on the performance of the vanilla encoder-decoder architecture, since the model also has access to the examples from the novel classes in $\mathcal{D}_n^K$.

4 Experiments

4.1 Dataset and Evaluation Protocol

For our experiments we use the ShapeNetCore_v1.0 [1] dataset and the few-shot generalization benchmark of [32]. As in [32], we use 7 categories as our *base classes*: **plane, car, chair, display, phone, speaker, table**; and 10 categories as our *novel classes*: **bench, cabinet, lamp, rifle, sofa, watercraft, knife, bathtub, guitar, laptop**. Note that we have added additional categories to the standard benchmark, for a more extensive evaluation. Out data comes in the form of pairs of 128×128 images rendered using Blender [4], and $32 \times 32 \times 32$ voxelized representations obtained using Binvox [20,19]. Each 3D model has 24 associated images, rendered from random viewpoints. For evaluation, we use the standard Intersection over Union (IoU) score to compare predicted shapes $\bar{S}$ to ground truth shapes S: IoU $= |S \cap \bar{S}|/|S \cup \tilde{S}|$.

4.2 Implementation Details

All methods are trained on the 7 base classes except for the AS-baseline which is trained on all 17 categories. All methods share the same 2D encoder and 3D decoder architectures. We use the same 2D encoder as in [26,32], a ResNet [9] that takes a 128×128 image as input, and outputs a 128-dimensional embedding. Our 3D decoder consists of 7 convolutional layers, followed by batch-normalization, and ReLU activations. For training, we use the same 80-20 train-test split as in R2N2 [3,32]. Unless otherwise stated, we use $l_r = 0.0001$ as the learning rate and ADAM [12] as the optimizer. All networks are trained with binary cross entropy on the predicted voxel presence probabilities in the output 3D grid.

ZS-Baseline is trained on the 7 base categories for 25 epochs. We use the trained model to make predictions for novel classes without further adaptations.
AS-baseline is trained on *all* 17 categories for 25 epochs. We do not use any pre-trained weights, but train this baseline model from a random initialization.
Wallace et al. [32]. To ensure a fair comparison, we re-implemented this framework, using the exact same settings reported in [32]. In the supplemental material we include a comparison only on the subset of classes used in [32], validating that our implementation yields practically identical results.
GCE. We use the same architecture as in the baseline models in [32]. Contrary to the element-wise addition used in [32], we concatenate the 128-d embeddings from the 2D encoder and the conditional branch, and we feed the resulting 256-d embedding into the 3D decoder. The class conditioning vectors are initialized randomly from a normal distribution $\sim \mathcal{N}(0, 1)$. After training the GCE on the base classes, we freeze the parameters of E_I and D and initialize the *novel class* embeddings as the average of the learned base class encodings. We then optimize them using stochastic gradient descent (SGD) with momentum set to 0.9.
CGCE. The conditional branch is composed of 5 codebooks, each containing 6 codes of dimension 128, and an attention array of size $17 \times 5 \times 6$; i.e., one attention value per $(class, codebook, code)$ triplet. The codes and attention values are initialized using a uniform distribution $U(-0.4, 0.4)$. During training, we push the attention array to focus on meaningful codes by employing *sparsemax* [15]. After training the CGCE on the base classes, we freeze the parameters of E_I and D, as well as the codebook entries $\mathbf{c}_{j,k}$. We initialize the *novel class* attentions $\boldsymbol{\alpha}_i$ from a uniform distribution $U(-0.4, 0.4)$. We then optimize $\boldsymbol{\alpha}_i$ using stochastic gradient descent (SGD) with momentum set to 0.9.
MCCE We replace all batch normalization (bnorm) layers in the 3D decoder with *conditional* batch normalization (cond-bnorm) [22]. More precisely, the affine parameters γ_i and β_i are initialized from a normal distribution $\sim \mathcal{N}(1, 0.2)$, and conditioned on the class i. For novel class adaptation only the aforementioned γ_i and β_i for new classes are learned. We use SGD as optimizer with momentum set to 0.9 for this novel class adaptation.

4.3 Comparing Baselines in the Few-Shot Regime

Tatarchenko et al. [31] showed that naive 3D reconstruction baselines not only perform well, but manage to surpass in performance more complicated, state of the art approaches. We show that such baselines, however, perform poorly in a few-shot learning setup [32], where a more nuanced understanding of 3D shape is required for generalization to novel examples. In Table 1 we compare the ONN, ZS, and AS baselines, described in Section 3.4. We consider several versions of ONN, with access to varying numbers of examples in the few-shot spectrum, ranging from a "1-shot" (ONN-1) to "full-shot" (ONN-full - access to all shapes for that class). We observe that ONN-full outperforms AS, which has been trained on all available data, supporting the findings of [32]. However, once the number of shots decreases, performance for ONN quickly deteriorates, and drops below that of even the ZS baseline.

cat	ZS-Baseline	AS-baseline	ONN-1	ONN-2	ONN-3	ONN-4	ONN-5	ONN-10	ONN-25	ONN-full
bench	0.366	0.524	0.238	0.240	0.245	0.271	0.276	0.360	0.420	0.708
cabinet	0.686	0.753	0.400	0.458	0.460	0.461	0.480	0.495	0.631	0.842
lamp	0.186	0.368	0.153	0.162	0.177	0.189	0.194	0.223	0.282	0.515
firearm	0.133	0.561	0.377	0.396	0.420	0.425	0.434	0.510	0.550	0.707
sofa	0.519	0.692	0.445	0.458	0.459	0.530	0.534	0.579	0.616	0.791
watercraft	0.283	0.560	0.259	0.286	0.317	0.354	0.372	0.479	0.527	0.697
mean_novel	0.362	0.576	0.312	0.333	0.346	0.371	0.381	0.441	0.504	0.710

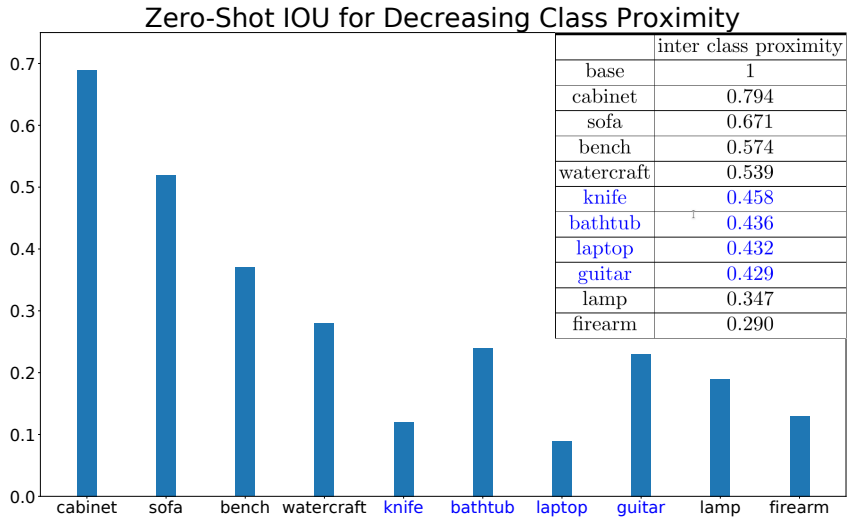
Table 1: Zero-shot (ZS), All-shot (AS), and Oracle Nearest Neighbor (ONN-K) IoU results for different number of shots, K. ONN outperforms an encoder-decoder model when the full dataset is available. However, in the low-shot regime, even the zero-shot variant shows better generalization, outperforming ONN.

Note that the ZS baseline already achieves relatively high performance on select classes (`sofa` and `cabinet`). We hypothesize that this is due to the similarity of these classes to some of the base categories. To test the validity of our hypothesis, we compute a similarity score between each novel class and the base class set. Let $\mathcal{C}$ be the set of all shapes S in a novel class. We compute its nearest neighbor with respect to *all* base classes: $\text{IoU}(S, \mathcal{D}_b) = \max_{S_b \in \mathcal{D}_b} \text{IoU}(S, S_b)$. We then compute an *inter-class proximity* between $\mathcal{C}$ and all base classes as the average of these IoU scores: $\text{P}(\mathcal{C}, \mathcal{D}_b) = \frac{1}{|\mathcal{C}|} \sum_{S_i \in \mathcal{C}} \text{IoU}(S_i, \mathcal{D}_b)$.

Figure 4 shows the IoU scores of novel classes, sorted by decreasing proximity scores to the base set. To better study the effect of proximity to IoU performance, we have included four (4) additional novel classes from ShapeNet, highlighted in blue. Note that ZS performs better for classes with higher proximity to base classes, supporting our original hypothesis. This also means that novel classes with low proximity have much higher potential for improvement using few-shot learning, with respect to ZS.

4.4 Evaluating Few Shot-Generalization

In Table 2 we evaluate the three methods described in Sec. 3, on 1-shot reconstruction. We report both the IOU as well as the relative improvement over the ZS baseline. Note that as the ZS-baseline provides strong performance for easy classes, the average IOU is dominated by these, thus relative improvement is a more meaningful metric for aggregation across classes. Please note that GCE improves performance over [32], particularly for classes with low proximity to the base set, obtaining 45% relative improvement over ZS, overall, compared to [32]. The compositional and multiscale priors lead to further improvements of 54% and 52%, respectively, compared to the simple shape prior of [32].



	inter class proximity
base	1
cabinet	0.794
sofa	0.671
bench	0.574
watercraft	0.539
knife	0.458
bathtub	0.436
laptop	0.432
guitar	0.429
lamp	0.347
firearm	0.290

Fig. 4: Zero-shot IoU for decreasing proximity between ot the base set. The higher the proximity of a novel class to the base set, the better ZS performs. To make this point clear, we add more classes of low proximity (blue color) to our evaluation.

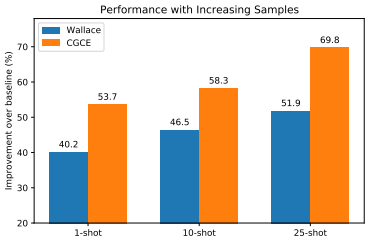


Fig. 5: Percentage gains for 1, 10 and 25 shot over ZS baseline. The gains of our method increase, relative to [32], with greater number of shots (larger intra-class variability).

cat	ZS	AS	GCE	GCE_rand
plane	0.580	0.572	0.582	0.198
car	0.835	0.830	0.837	0.412
chair	0.504	0.500	0.510	0.284
monitor	0.516	0.508	0.520	0.346
cellphone	0.704	0.689	0.710	0.497
speaker	0.648	0.659	0.670	0.505
table	0.536	0.537	0.540	0.376

Table 4: Performance drops significantly when the class embedding is randomly selected, validating that the class conditioning is being used by the GCE model.

In Table 3 we evaluate the CGCE variant (which performs best in 1-shot evaluation) on the 10- and 25-shot settings, and compare to [32]. We observe that, similarly to the 1-shot case, most methods do not significantly improve the performance for classes with high proximity to the base set. For distant classes, on the other hand, we see substantial performance improvements (sometimes 200%+ in IoU). Table 3 also shows the increased gap in performance between CGCE and [32], as the number of shots increases, supporting our argument that the global conditional embedding can better capture intra-class variability and thus remains effective beyond the 1-shot setting.

	Zero-shot	All-shot	1 shot			
cat	ZS	AS	Wallace [32]	GCE	CGCE	MCCE
cabinet	0.69	0.75	0.69 (0.00)	0.69 (0.01)	**0.71 (0.03)**	0.69 (0.01)
sofa	0.52	0.69	0.54 (0.04)	0.52 (0.00)	**0.54 (0.04)**	**0.54 (0.03)**
bench	0.37	0.52	0.37 (0.00)	0.37 (0.00)	**0.37 (0.00)**	**0.37 (0.00)**
watercraft	0.28	0.56	0.33 (0.16)	0.34 (0.19)	**0.39 (0.39)**	0.37 (0.29)
knife	0.12	0.60	0.30 (1.47)	0.26 (1.13)	**0.31 (1.5)**	0.27 (1.19)
bathtub	0.24	0.46	0.26 (0.05)	0.27 (0.09)	**0.28 (0.13)**	0.27 (0.11)
laptop	0.09	0.56	0.21 (1.30)	0.27 (1.85)	**0.29 (2.10)**	0.27 (1.87)
guitar	0.23	0.69	0.31 (0.38)	0.30 (0.31)	**0.32 (0.42)**	0.30 (0.31)
lamp	0.19	0.37	0.20 (0.05)	0.20 (0.07)	0.20 (0.05)	**0.22 (0.16)**
firearm	0.13	0.56	0.21 (0.58)	0.24 (0.83)	0.23 (0.70)	**0.30 (1.26)**
mean (relative to ZS)			40.2%	44.7%	**53.7**%	52.2%

Table 2: IoU scores for single-image 3D reconstruction in the 1-shot setting. Numbers in parentheses indicate *relative* performance gains over ZS. Note the marked improvement, especially for novel classes with low proximity to the base set, indicating much better generalization of our method.

	Zero-shot	All-shot	10 shot		25 shot	
cat	ZS	AS	Wallace	CGCE	Wallace	CGCE
cabinet	0.69	0.75	0.69 (0.00)	**0.71 (0.03)**	0.69 (0.01)	**0.71 (0.04)**
sofa	0.52	0.69	0.54 (0.04)	0.54 (0.04)	0.54(0.04)	**0.55 (0.06)**
bench	0.37	0.52	0.36 (-0.01)	**0.37 (0.03)**	0.36 (-0.01)	**0.38 (0.04)**
watercraft	0.28	0.56	0.36 (0.26)	**0.41 (0.45)**	0.37 (0.29)	**0.43 (0.53)**
knife	0.12	0.60	0.31 (1.52)	**0.32 (1.62)**	0.31 (1.57)	**0.35 (1.87)**
bathtub	0.24	0.46	0.26 (0.05)	**0.28 (0.16)**	0.26 (0.06)	**0.30 (0.23)**
laptop	0.09	0.56	0.24 (1.53)	**0.30 (2.24)**	0.27 (1.85)	**0.32 (2.45)**
guitar	0.23	0.69	0.32 (0.39)	**0.33 (0.47)**	0.32 (0.42)	**0.37 (0.62)**
lamp	0.19	0.37	0.19 (0.04)	**0.20 (0.05)**	0.19 (0.03)	**0.20 (0.07)**
firearm	0.13	0.56	**0.24 (0.83)**	0.23 (0.75)	0.26 (0.95)	**0.28 (1.08)**
mean (relative to ZS)			46.5%	**58.3**%	51.9%	**69.8**%

Table 3: IoU scores for K-shot evaluation ($K \in \{10, 25\}$). Numbers in parentheses are performance gains over ZS. Improvements for CGCE widen as K increases.

Validating the contribution of the shape prior. To validate that our GCE framework (and by extension, CGCE and MCCE) does not simply ignore the conditioning on the shape prior, we perform a simple ablation in which we randomly select the class of the corresponding global embedding for a given input; we call this variant GCE-rand. As shown in Table 4, performance drops drastically, validating that the model learns to use the class-specific shape priors.

Analysis of the Compositional GCE. We analyze the CGCE codes learned by our model through visualizations that unveil associations of codebook entries with object parts. Given a 2D input image, we generate its 3D reconstruction, after randomly removing the contribution of selected codebook entries in the compositional shape prior. Figure 6 shows the results. We observe that removing

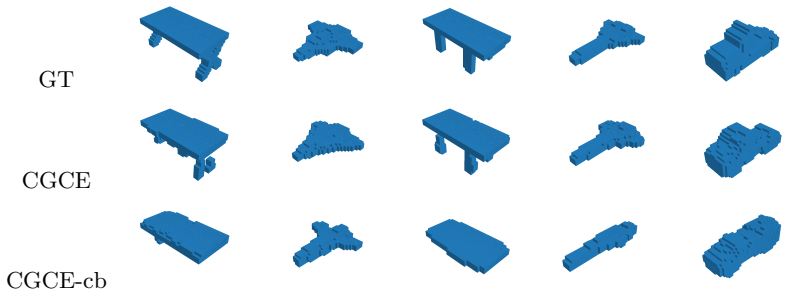


Fig. 6: 3D reconstructions with our compositional GCE (CGCE) model. Eliminating the contribution of a selected codebook in the shape prior (CGCE-cb) deletes object parts, such as table legs or plane wings, from the reconstructed shape, indicating that the learned codebooks capture meaningful semantic attributes.

certain codes results in the removal of semantically meaningful portions of the reconstructed object, such as table legs or plane wings.

We also explicitly analyze the learned attention over the codebook entries. We start by using the IOU-based class proximity metric described in Sec. 4.3 to associate each novel class to its closest base classes (see Figure 7). We observe a positive correlation of high proximity scores and alignment of the attention distribution over codes for novel and base classes. Finally, in Figure 8 we visually compare CGCE reconstructions to those of [32], in the 25-shot case, confirming that numerical performance gains translate into higher reconstruction quality for our approach. For more visualizations we refer to the supplementary material.

5 Conclusions

We have identified few-shot 3D reconsruction as an ideal benchmark for studying 3D deep learning models and their ability to reason about object shapes and generalize to new categories. We have addressed several key weaknesses of previously proposed models in this setting, particularly in capturing intra-class variability, and have proposed compositional and multi-scale shape priors that improve performance and interpretability. Plans for future work in this area include whether incorporating alternative shape representations can further improve generalization, especially for higher resolution shapes.

Acknowledgements This work has been funded by the Australian Research Council through grant FT170100072. Authors would like to thank Ming Xu for constructive feedback. EB acknowledges funding from IVADO.

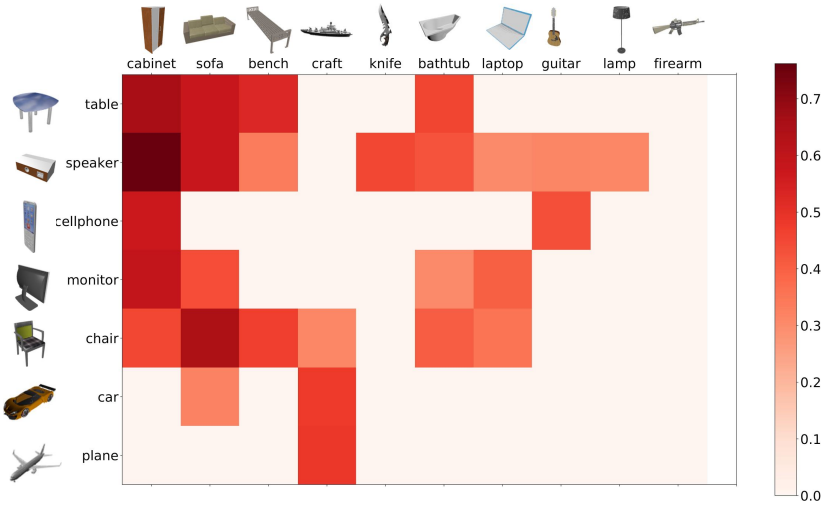


Fig. 7: Proximity between base classes (y-axis) and novel classes (x-axis). Distance is measured, as a mean IoU of nearest neighbors.

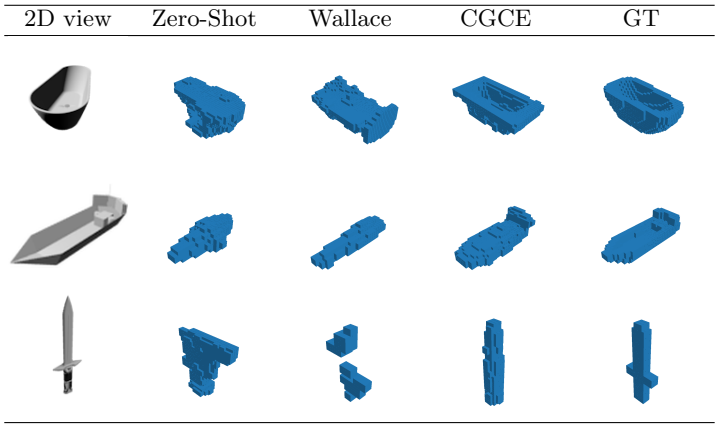


Fig. 8: Qualitative analysis on 3 different examples using novel classes with 25-shots. We show predictions by different models and the ground truth (GT). Our model exhibits qualitatively better reconstructions than [32] and the Zero-Shot baseline.

References

1. 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) 8
2. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5939–5948 (2019) 4
3. 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: Proceedings of the European Conference on Computer Vision (ECCV) (2016) 3, 4, 8
4. Community, B.O.: Blender - a 3D modelling and rendering package. Blender Foundation, Stichting Blender Foundation, Amsterdam (2018), http://www.blender.org 8
5. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. vol. 2, p. 6 (2017) 4
6. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1126–1135. JMLR. org (2017) 4
7. Gidaris, S., Komodakis, N.: Dynamic few-shot visual learning without forgetting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4367–4375 (2018) 4
8. Girdhar, R., Fouhey, D.F., Rodriguez, M., Gupta, A.: Learning a predictable and generative vector representation for objects. In: European Conference on Computer Vision. pp. 484–499. Springer (2016) 4
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016) 8
10. Hoiem, D., Efros, A.A., Hebert, M.: Automatic photo pop-up. In: ACM SIGGRAPH 2005 Papers, pp. 577–584. ACM (2005) 2
11. Horn, B.K.: Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. Tech. rep., CSAIL, USA (1970) 2
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 8
13. Kong, C., Lin, C.H., Lucey, S.: Using locally corresponding cad models for dense 3d reconstructions from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4857–4865 (2017) 2
14. Kutulakos, K.N., Seitz, S.M.: A theory of shape by space carving. International journal of computer vision **38**(3), 199–218 (2000) 2
15. Martins, A., Astudillo, R.: From softmax to sparsemax: A sparse model of attention and multi-label classification. In: International Conference on Machine Learning. pp. 1614–1623 (2016) 7, 9
16. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space (2018) 3
17. Michalkiewicz, M., Belilovsky, E., Baktashmotlagh, M., Eriksson, A.: A simple and scalable shape representation for 3d reconstruction. arXiv preprint arXiv:2005.04623 (2020) 3, 4
18. Michalkiewicz, M., Pontes, J.K., Jack, D., Baktashmotlagh, M., Eriksson, A.P.: Deep level sets: Implicit surface representations for 3d shape inference. In: Proceedings of the International Conference on Computer Vision (ICCV) (2019) 4

19. Min, P.: binvox. `http://www.patrickmin.com/binvox` or `https://www.google.com/search?q=binvox` (2004 - 2019), accessed: 2020-03-05 8
20. Nooruddin, F.S., Turk, G.: Simplification and repair of polygonal models using volumetric techniques. IEEE Transactions on Visualization and Computer Graphics **9**(2), 191–205 (2003) 8
21. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 165–174 (2019) 4
22. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018) 4, 7, 9
23. Qi, H., Brown, M., Lowe, D.G.: Low-shot learning with imprinted weights. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5822–5830 (2018) 4
24. Ramakrishna, V., Kanade, T., Sheikh, Y.: Reconstructing 3d human pose from 2d image landmarks. In: European conference on computer vision. pp. 573–586. Springer (2012) 2
25. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: ICLR (2017) 4
26. Richter, S.R., Roth, S.: Matryoshka networks: Predicting 3d geometry via nested shape layers. In: CVPR. pp. 1936–1944. IEEE Computer Society (2018) 3, 8
27. Savarese, S., Andreetto, M., Rushmeier, H., Bernardini, F., Perona, P.: 3d reconstruction by shadow carving: Theory and practical evaluation. International journal of computer vision **71**(3), 305–336 (2007) 2
28. Shu, R., Nakayama, H.: Compressing word embeddings via deep compositional code learning. arXiv preprint arXiv:1711.01068 (2017) 7
29. Siam, M., Oreshkin, B., Jagersand, M.: Adaptive masked proxies for few-shot segmentation. arXiv preprint arXiv:1902.11123 (2019) 4
30. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in neural information processing systems. pp. 4077–4087 (2017) 4
31. 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 Conference on Computer Vision and Pattern Recognition. pp. 3405–3414 (2019) 2, 3, 4, 8, 9
32. Wallace, B., Hariharan, B.: Few-shot generalization for single-image 3d reconstruction via priors. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3818–3827 (2019) 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14
33. Wang, C., Wang, Y., Lin, Z., Yuille, A.L., Gao, W.: Robust estimation of 3d human poses from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2361–2368 (2014) 2
34. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In: ECCV (2018) 4
35. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly simple few-shot object detection. arXiv preprint arXiv:2003.06957 (2020) 4
36. Witkin, A.P.: Recovering surface shape and orientation from texture. Artificial intelligence **17**(1-3), 17–45 (1981) 2
37. Wu, J., Xue, T., Lim, J.J., Tian, Y., Tenenbaum, J.B., Torralba, A., Freeman, W.T.: Single image 3d interpreter network. In: European Conference on Computer Vision. pp. 365–382. Springer (2016) 2

38. Wu, J., Zhang, C., Xue, T., Freeman, B., Tenenbaum, J.: Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: Advances in neural information processing systems. pp. 82–90 (2016) 4
39. Wu, X., Sahoo, D., Hoi, S.C.: Meta-rcnn: Meta learning for few-shot object detection. arXiv preprint arXiv:1909.13032 (2019) 4
40. Xu, B., Chang, W., Sheffer, A., Bousseau, A., McCrae, J., Singh, K.: True2form: 3d curve networks from 2d sketches via selective regularization. ACM Transactions on Graphics (TOG) **33**(4), 1–13 (2014) 2
41. Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In: Advances in Neural Information Processing Systems. pp. 1696–1704 (2016) 4
42. Zhu, R., Kiani Galoogahi, H., Wang, C., Lucey, S.: Rethinking reprojection: Closing the loop for pose-aware shape reconstruction from a single image. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017) 4