# Using Shape to Categorize: Low-Shot Learning with an Explicit Shape Bias

Stefan Stojanov, Anh Thai, James M. Rehg
Georgia Institute of Technology
{sstojanov, athai6, rehg}@gatech.edu

## Abstract

*It is widely accepted that reasoning about object shape is important for object recognition. However, the most powerful object recognition methods today do not explicitly make use of object shape during learning. In this work, motivated by recent developments in low-shot learning, findings in developmental psychology, and the increased use of synthetic data in computer vision research, we investigate how reasoning about 3D shape can be used to improve low-shot learning methods' generalization performance. We propose a new way to improve existing low-shot learning approaches by learning a discriminative embedding space using 3D object shape, and using this embedding by learning how to map images into it. Our new approach improves the performance of image-only low-shot learning approaches on multiple datasets. We also introduce Toys4K, a 3D object dataset with the largest number of object categories currently available, which supports low-shot learning.* [1]

## 1. Introduction

Understanding the role of 3D object shape in categorizing objects from images is a classical topic in computer vision [29, 9, 51], and the early history of object recognition was dominated by considerations of object shape. For example, David Marr's influential theory [27] posits that image-based recognition should be formulated as a sequence of information extraction steps culminating in a 3D representation to be used for recognition. The difficulty of reliably extracting 3D shape from images, combined with the availability of large-scale image datasets [6, 22], motivated the modern development of purely appearance-based approaches to recognition and categorization. This has culminated in current approaches such as CNNs that learn feature representations directly from images. Moreover, a study by Geirhos et al. [14] of the inductive biases of CNNs trained on ImageNet suggests that categorization performance is driven primarily by a bias towards image texture rather than object shape.[2]

However, studies of infant learning [24, 7, 23, 15] suggest that shape does play a significant role in the ability to rapidly learn object categories from a small number of examples, a task which is analogous to few-shot learning. Both young children and adults who are forced to categorize novel objects based on a few examples display a *shape bias*, meaning that shape cues seem to play a dominant role in comparison to color and texture when inferring category membership. These studies beg the question of whether information about 3D object shape could be useful in learning to perform few-shot categorization from images. While prior work has demonstrated effective approaches to object categorization using 3D shapes as input [34, 36, 55, 56, 4], and there is a large literature on few-shot learning from images alone [44, 53, 18, 38, 59, 49, 11], the question of how shape cues could be used to learn effective representations for image-based low-shot categorization has not been investigated previously.

The goal of this paper is to explore the incorporation of a shape bias in SOTA approaches to few-shot object categorization and thereby investigate the utility of shape information in category learning. We leverage the recent availability of datasets of 3D object models with category labels, such as ModelNet40 [56] and ShapeNet [2]. By sampling surface point clouds and rendering images of these models, we can construct datasets that combine 3D shape and image cues. Unfortunately, however, ShapeNet and ModelNet contain a relatively small number of object categories (55 and 40 respectively), making it difficult to test categorization at a sufficient scale. To resolve this limitation, we introduce a new 3D object dataset, *Toys4K* consisting of 4,179 3D objects from 105 object categories, designed to contain categories of objects that are commonly encountered by infants and children during their development.

We report on two sets of investigations. First, we examine the relative effectiveness of purely image-based and purely shape-based approaches to few-shot categorization.

[1] The code and data for this paper are available at our project page https://rehg-lab.github.io/publication-pages/lowshot-shapebias/

[2] This study does not speak to the possibility of whether shape could be used more effectively, and it is unclear how much of the bias stems from the composition of the ImageNet dataset itself.

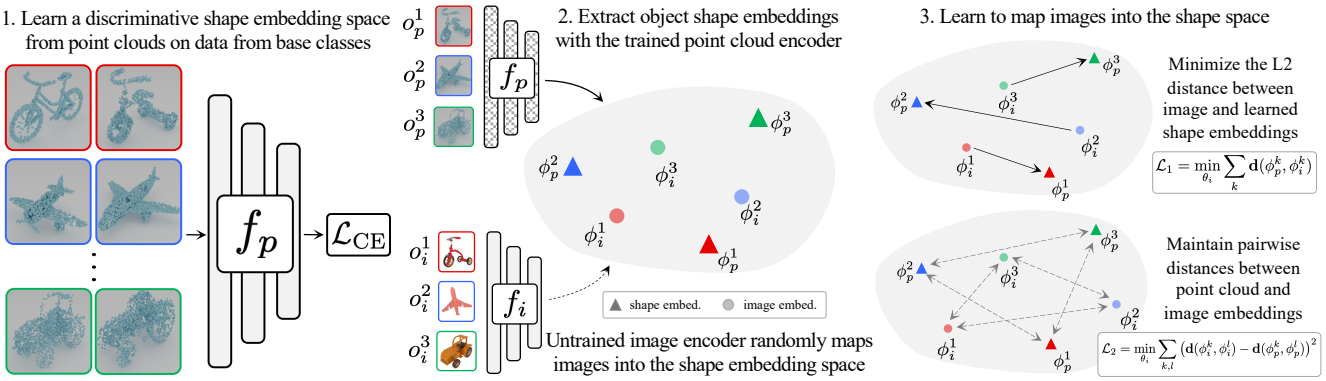


Figure 1. To perform low-shot learning with shape-bias, we train a embedding space defined by a point cloud encoder $f_p$ trained with cross-entropy. Shape based embedding spaces are more discriminative than image-based ones (see Tbl. 1). We extract object shape embeddings, and train an image encoder $f_i$ to map images into the shape space. If trained successfully, $f_i$ will have the discriminative properties of $f_p$.

We demonstrate that purely shape-based few-shot learning outperforms image-based approaches, and establish an empirical upper bound on the effectiveness of a shape bias. Second, we develop a novel approach for training an image embedding representation for low-shot categorization which incorporates an explicit shape bias, which we outline in Figure 1. We benchmark this approach on a representative set of SOTA few-shot learning architectures and demonstrate that the incorporation of shape bias results in increased generalization accuracy over image-based training alone. In summary, this paper makes the following contributions:

- A new approach to add explicit shape-bias to existing low-shot image classification methods, utilizing 3D object shape to learn similarity relationships between objects, which leads to improved low-shot performance.
- The first evidence that shape information can enable image-based low-shot classifiers to generalize with higher accuracy to novel object categories.
- Toys4K - new 3D object dataset containing approximately twice the number of object categories as previous datasets which can be used for low-shot learning.

## 2. Related Work

**Object Recognition from Synthetic Data**

A large body of work focuses on appearance [46, 28, 35, 10, 17], point cloud [34, 55, 36, 4] and voxel [56, 35] based recognition of synthetic object data with category taxonomies based on object shape such as ModelNet40 [56]. The trade-offs between learning using point clouds, depth maps, voxels, or images have been studied by [46, 35] but their study focuses on standard supervised classification and does not extend to low-shot classification of novel object categories or on combining shape and appearance information during learning.

**Low-Shot Learning**

Low-shot learning algorithms can be categorized into two broad sets. Optimization-based algorithms such as MAML [11, 12], LEO [38], and Reptile [30], which during the base-classes training stage, attempt to learn a representation that can quickly be adapted using small amounts of information with gradient-based learning in the low-shot learning stage. Metric learning-based methods such as Prototypical [44], Matching [53], and Relation [47] networks, as well as the more recent SimpleShot [54], FEAT [59], and RFS [49] aim to use the base class data to learn a similarity metric that will also be discriminative for novel classes during the low-shot phase. Despite their simplicity, metric-based approaches have superior performance on low-shot learning benchmarks [54, 59]. Our approach of adding shape bias belongs to the latter category, and compared to both is the first approach to combine both appearance and shape information for low-shot learning.

**3D Object Shape Datasets**

Other related works focus on building datasets of 3D object models for recognition, single image object shape reconstruction and shape segmentation [42, 48, 60, 21, 56, 2, 45]. The most widely used 3D shape datasets with category labels are ModelNet40 [56] with 12K object instances of 40 categories with no object surface material properties, ShapeNetCore.v2 [2] with 52K objects of 55 object categories with basic surface texture properties (basic shading and UV mapping, but no physically based materials). The ShapeNetSem split of ShapeNet consists of over 100 categories but is unsuitable for recognition since individual object instances are assigned to multiple categories. Datasets such as ABC [21] and Thingi10k [60] claim higher mesh quality than previous datasets but lack object category annotation, making them more suitable for low-level tasks like surface normal estimation and category agnostic shape reconstruction. The ModelNet40 and ShapeNet datasets were scraped from online repositories and have categories largely based on the data that was available in these repositories. In contrast, our new Toys4K is curated specifically for testing the generalization ability of learned representations to new

classes. Compared to the aforementioned datasets, Toys4K consists of highly diverse object instances within a category (evident in Figure 3, detailed composition is included in the supplement) and has the highest number of individual object categories despite its smaller total size.

**Multi-modal Learning**

Aligning representations from different data modalities has been extensively studied in vision and language works on zero-shot learning [57, 19, 39, 13]. More recently, Schwartz et al. [41] and Xing et al. [58] improve low shot image classification performance on standard low-shot datasets by combining the representation learned through the appearance modality (images) with language model word vector embeddings. In comparison, we combine appearance (images) and shape (point clouds) to learn a representation for low shot object recognition that is biased to object shape and leads to better low-shot generalization. It is important to note that these works use multi-modal information for the low-shot queries at test time, whereas our approach only uses multi-modal information for the low-shot support set.

Another category of multi-modal learning works focuses on learning joint embedding spaces of 3D meshes and images for image-based 3D shape retrieval [25, 26]. While these works focus on retrieval for the same object categories at training and testing time, our work focuses on combining appearance and shape information for low-shot generalization to *novel* object categories.

## 3. Using Shape for Low-Shot Classification

In principle, 3D shape is an attractive representation for object recognition [27, 31, 26, 25] due to its invariance to the effects of viewpoint, illumination, and background, which can be challenging for appearance-based approaches. While appearance-based methods may be able to model these sources of variation given sufficient training images, there is always a question of how well such models can generalize to novel categories and objects [14].

Despite its potential advantages, no previous work on low-shot learning has utilized 3D shape, for at least two reasons: 1) It is unclear how to leverage 3D shape in improving *image-based* low shot learning;[3] 2) There is a lack of 3D shape datasets that contain a sufficient number of object categories to support effective experimentation. This is due to the additional data requirements of few-shot learning: The training/validation/testing split is over different classes and not data points of the same class [37, 53] in order to effectively test generalization to unseen classes.

To explain this issue more formally, let $\mathcal{D}^{\text{train}}$ denote the base classes, and $\mathcal{D}^{\text{val}}$ and $\mathcal{D}^{\text{test}}$ denote the validation and testing sets, respectively, where these sets comprise a disjoint partition of the total available classes. The base classes must be sufficiently large and diverse to learn an effective feature representation in the training phase, and the $\mathcal{D}^{\text{val}}$ set must similarly support the accurate assessment of low-shot generalization ability during hyperparameter tuning (i.e. model selection while training on the base classes). The $\mathcal{D}^{\text{test}}$ set is used to generate labeled low-shot training examples (supports), and unlabelled low-shot testing examples (queries), which are used to evaluate the generalization performance of the model at testing time, which we refer to as the *low-shot phase*. As a result of these constraints, the standard 3D shape datasets ModelNet40 [56] and ShapeNet55 [2] can only support 10-way and 20-way testing, respectively. If the number of testing classes is insufficient, the estimation of the generalization performance of the method may be inaccurate.

In this section, we describe our two primary contributions which address the limitations described above. In § 3.1 we present our novel method for introducing *shape bias* in learning a low-shot image representation. In § 3.2, we introduce a novel 3D object category dataset, *Toys4K*, consisting of 4,179 object instances organized into 105 categories, with an average of 35 objects (3D meshes) per category. Toys4K supports up to 50-way classification, expanding well beyond ModelNet40 and ShapeNet55 (see Fig. 4).

### 3.1. Low-Shot Learning with Shape Bias

We begin by describing the *problem formulation:* We assume that shape data in the form of 3D point clouds is available for each RGB image in a dataset. We achieve this by rendering RGB images from the 3D models. 3D shape information is used directly during training and validation, in order to construct a representation with an explicit shape bias. In addition, during the low-shot phase, episodes are generated so that point clouds are available for the support objects, but *not* for the query objects. This assumption allows for both appearance and shape information to be used in building class prototypes, but *inference is done using images only*. The distinction between image only low-shot learning and our new setting is illustrated in Figure 2.

In this work, we adopt a low-shot learning approach based on a *metric embedding space*. In this approach, $\mathcal{D}^{\text{train}}$ is used to learn a function $f_i$ that maps the input data into an embedding space where object instances of the same category are close and instances of different categories are far apart, according to some distance metric. This mapping can be fixed after being learned from $\mathcal{D}^{\text{train}}$ or fine-tuned further, depending upon the algorithm design. During the low-shot phase, the supports and queries are mapped into the embedding space (see Figure 2), and the queries are classified according to a nearest neighbor or nearest class prototype (e.g. support centroid) rule. Metric-based low-shot learning has high accuracy [54] and is significantly more computation-

[3] Our focus is on few-shot methods in which the queries are *images*, with no 3D shape information available, as this is the most general and useful paradigm.

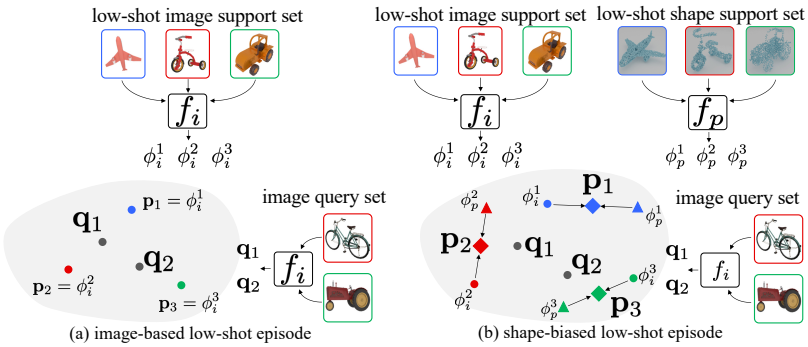


Figure 2. (a) The standard setting: Prototypes are formed from images. (b) Our novel shape-biased setting: Image and shape embeddings are averaged. In both cases, the image-only queries $q_i$ can be classified by identifying the closest prototype $p_j$. The training process for the mapping functions $f_i$ and $f_p$ is illustrated in Figure 1.

| | 5-way | | 10-way | |
|---|---|---|---|---|
| | 1-shot | 5-shot | 1-shot | 5-shot |
| Image | 58.99 | 74.29 | 45.82 | 62.73 |
| Point Cloud | 66.02 | 83.61 | 54.44 | 75.26 |
| Img + Ptcld Oracle | 68.04 | 82.07 | 57.03 | 73.11 |

Table 1. On ModelNet40-LS, low-shot generalization is higher for point-cloud based learning than image based learning, justifying our approach in combining the modalities. Oracle model has access to both image and point cloud information. See text for details.

ally efficient than approaches that fine-tune on the low-shot supports. We first demonstrate that shape-based low-shot learning allows for better generalization than image-based low-shot learning, and then show how a shape-based embedding with high generalization ability can be used to improve image-based low-shot classification.

**Shape-based low-shot learning outperforms image-based low-shot learning**

We perform a simple empirical study to determine whether shape has an advantage for low-shot generalization. We train two embedding spaces, one using image data and one using point cloud data. For each type of data, we follow the SimpleShot [54] approach, meaning that we train a classifier using cross-entropy on $\mathcal{D}^{\text{train}}$ and use the learned feature space (output of the last pooling layer) to perform nearest centroid-based low-shot classification in normalized Euclidean space. We use a ResNet18 [16] for image learning and a DGCNN [55] for point cloud learning on the ModelNet40-LS dataset (see § 4).

We present the results in Tbl. 1, and as might be expected, see significantly higher low-shot performance for the point cloud model relative to the image model. This quantifies the improvement in generalization to novel categories as as result of using a 3D shape-based representation and suggests that 3D shape can yield a more discriminative embedding space. The question then is *how can this benefit be retained when testing the model on image data alone?*

**Combining Appearance and Shape**

Figure 1 illustrates our approach to using the 3D shape information available at training time in order to learn how to embed the image-only queries. First, we train a low-shot point-cloud based classifier on the set of base-classes $\mathcal{D}^{\text{train}}$, resulting in an a highly discriminative embedding space for both seen and novel categories. We then extract point cloud embeddings for each object in the training set and train a CNN to map images into the shape embedding space.

Let $\mathcal{D}$ be a dataset of paired object point clouds $o_p$ and images $o_i$, partitioned into $\mathcal{D}^{\text{train}}$, $\mathcal{D}^{\text{val}}$, and $\mathcal{D}^{\text{test}}$. Let $f_p(x)\colon N \times \mathbb{R}^3 \to \mathbb{R}^d$ denote the trained function for mapping point clouds of size $N$ into an embedding space of dimension $d$. This embedding space is optimized to yield favorable metric properties for low shot classification, using the labelled point cloud data in $\mathcal{D}^{\text{train}}$. Our goal is then to learn a second mapping, $f_i(x)\colon \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^d$, where $H, W$ are the image height and width, from images into the shape embedding space defined by $f_p(x)$. We denote point cloud embeddings as $f_p(o_p) = \phi_p$ and image embeddings as $f_i(o_i) = \phi_i$.

We train a model that learns the mapping from images to shape embeddings by minimizing two loss functions (see part 3 of Figure 1). For a mini-batch $\mathcal{B} \subset \mathcal{D}^{\text{train}}$ the first loss minimizes the squared Euclidean distance (which we denote as $\mathbf{d}(x, y)$) between the learned point cloud embeddings, and the image based embeddings

$$\mathcal{L}_1 = \sum_{(o_i, o_p)\in\mathcal{B}} \mathbf{d}\big(\phi_i, \phi_p\big).$$

The second loss constrains the pairwise distances between the image embeddings of different object instances to be the same as the pairwise distances of the learned shape embeddings. Let $\mathcal{I}$ denote the set of all $(k, l) = \big((o_i^k, o_p^k), (o_i^l, o_p^l)\big)$ object instance data pairs in a mini-batch. We define the second loss as

$$\mathcal{L}_2 = \sum_{(k,l)\in\mathcal{I}} \big(\mathbf{d}(\phi_p^k, \phi_p^l) - \mathbf{d}(\phi_i^k, \phi_i^l)\big)^2.$$

During training, both losses are minimized with equal weight. Validation for choosing $f_i$ is done by nearest centroid classification on $\mathcal{D}^{\text{val}}$. In Section 4 we show that minimizing only $\mathcal{L}_1$ results in convergence without learning to match the distribution of the shape embedding well on the training set, resulting in poor performance.

**Inference:** During the low-shot phase, as shown in Figure 2, class prototypes are built by averaging the shape $\phi_p$ and image $\phi_i$ embeddings for each support object, whereas only image information is used to map the query objects via

Figure 3. Approximately one third of the objects in Toys4K, a new dataset of 3D assets for low-shot object learning using object appearance and shape information.

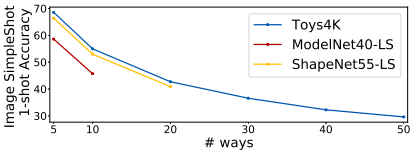


Figure 4. The high number of categories in Toys makes low-shot learning on Toys4K a challenging task.

| Dataset | Instances | Categories |
|---|---|---|
| Toys4K | 4,179 | 105 |
| ModelNet40 [56] | 12,311 | 40 |
| ShapeNet [2] | 52,000 | 55 |
| Thingi10K [60] | 10,000 | N/A |
| ABC [21] | 750K | N/A |

Table 2. Toys4K has the most categories of any available dataset of 3D objects.

$f_i$. The queries are classified based on the nearest centroid to the query embedding. This inference procedure is used for all algorithms in this paper that combine both image and shape information, with the exception of FEAT [59], which uses an additional set-to-set mapping.

It is important to understand how the shape-biased encoder performs when there is no explicit shape information available in the low-shot phase, and what is the gain in accuracy by making shape available for building class prototypes. To this end, in § 4.3 we also evaluate the setting where there are no point clouds available in the low-shot phase.

**Why is mapping images to shape embeddings difficult?** If the mapping $f_i(x)$ is learned exactly, it would map images to their corresponding point cloud embeddings so that

$$\forall(o_i, o_p) \in \{\mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{val}} \cup \mathcal{D}^{\text{test}}\}, ||\phi_i - \phi_p||_2 = 0.$$

This is challenging, however, since $f_i$ can only be trained on the base classes in $\mathcal{D}^{\text{train}}$, requiring it to correctly extrapolate to the metric properties of objects from novel classes.

We perform a simple test to validate the feasibility of mapping images to shape embeddings in general and establish an empirical upper bound. We perform this by simply minimizing the $L_2$ distance between the images and their corresponding shape embeddings on combined data from base classes, validation and test classes ($\mathcal{D}^{\text{train}} = \{\mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{val}} \cup \mathcal{D}^{\text{test}}\}$). This model is referred to as Image + Point Cloud Oracle in Table 1 and provides empirical evidence that it is possible to learn how to map images into a shape embedding space with high accuracy when all of the data is available. This model's performance closely matches that of the shape-only model, and significantly outperforms the image-based approach, providing further evidence that extrapolating the metric properties of the shape-embedding space to novel categories is the key challenge in learning to map images to shape embeddings.

### 3.2. Toys4K Dataset

An object dataset with a high number of diverse categories and high-quality 3D meshes is essential to study whether leveraging 3D object shape can enable improved low-shot generalization. We satisfy this requirement with our new *Toys4K dataset*. While it is possible to use existing datasets such as ModelNet40 and ShapeNet (which we include in our experiments), the limited number of categories is an obstacle to few-shot learning. For example, applying standard training/validation/test ratios (e.g. from mini-ImageNet [53]) to the 40 categories in ModelNet40 results in a 20-10-10 split, which limits the possibilities for many-way testing. A comparison of Toys4K to prior datasets is available in Table 2. In Figure 4 we demonstrate that many-way low-shot classification on Toys4K is a challenging task in comparison to ModelNet40 and ShapeNet.

Toys4K consists of 4,179 object instances in 105 categories, with an average of 35 object instances per category with no less than 15 instances per category, allowing for 5 support 10 query low-shot episodes to be formed. Fig. 3 provides an example of the quality and variety of the models. Further details on the dataset composition are available in the supplement. Toys4K was collected by selecting freely-available objects from Blendswap [1], Sketchfab [43], Poly [32] and Turbosquid [50] under Creative Commons and royalty-free licenses. Our list of object categories was developed in collaboration with experts in developmental psychology to include categories of objects available and relevant to children in their infancy. We manually selected each object and manually aligned the objects within each category to a canonical coordinate system that is consistent across all instances in that category.

## 4. Experiments

In this section, we perform an empirical evaluation of the benefit of explicit shape bias on multiple datasets and image-only low-shot learning algorithms.

### 4.1. Datasets

In addition to our new dataset, Toys4K, we use the 3D object category datasets ModelNet40 [56], ShapeNet [2]. For descriptions of the datasets please refer to § 2. We render images using the Cycles ray tracing renderer in Blender [33] using uniform lighting on white backgrounds. For all datasets, camera pose is randomly sampled for 25 views of each object with azimuth $\psi \in [0, 360]$ and elevation $\theta \in [-50, 50]$ degrees. Object surface point clouds are sampled from the 3D object meshes.

**Toys4K** is our new low-shot learning dataset is described in detail in § 3.2. We use a split of 40, 10, 55 for base, low-shot validation, and testing classes, respectively. For Toys and all other datasets, the split is designed such that the categories with most classes are in the training set, and the validation and testing classes are randomly chosen from the remainder of the data.

**ModelNet40-LS** is the existing ModelNet40 [56] dataset, with a 20, 10, 10 split for base, low-shot validation and testing classes respectively.

**ShapeNet-LS** is the existing ShapeNetCore.v2 [56] dataset, with a 25, 10, 20 split for base, low-shot validation and testing classes respectively, using a reduced subset of object samples per category to reduce training time due to the high data imbalance.

### 4.2. Baselines

Regarding low-shot learning, we compare with the classical low-shot learning method Prototypical Networks [44], and the state-of-the-art algorithms FEAT [59], RFS [49], and SimpleShot [54]. With respect to learning joint embeddings, we compare with a simple triplet loss-based approach that learns joint embeddings of images and point clouds. All baselines use a standard ResNet18 [16] as a backbone for image encoding and a DGCNN [55] to encode point clouds. In the supplement we perform an ablation study over different point cloud architectures including PointNet [34] and PointNet++ [36]. Our low-shot learning baseline implementations were all validated by re-creating the results from the original papers.[4]

**SimpleShot** [54] is a simple low-shot learning baseline algorithm that outperforms many recent methods. It makes use of an embedding space learned by a CNN by training on the base training classes for a standard classification task using cross-entropy loss. Validation and testing are done using a nearest neighbor classifier in the learned embedding space, with feature normalization and training set mean subtraction resulting in improved performance.

**RFS** [49] is another simple low-shot learning algorithm that is competitive with many recent approaches. Training the embedding space is done using cross-entropy on the training set, but at testing time, a simple logistic regression classifier is learned for each low-shot episode. In the original work, the authors show that training a set of embedding models with distillation slightly improves performance. We omit this for a fair comparison with all metric-based works since this addition would likely lead to performance improvements across the board.

**Prototypical Networks** [44] is a standard metric-based low shot learning approach, which uses the base class set to create low-shot episodes and learn a feature space that embeds object instances close or far based on visual similarity.

**FEAT** [59] builds on Prototypical Networks by learning an additional set-to-set function implemented as a Transformer [52] on top of a cross-entropy pre-trained embedding space to refine the class prototypes used for low-shot classification. FEAT achieves state of the art performance for inductive low-shot learning. Note that FEAT requires separate retraining for each $n$-way $m$-shot configuration [5].

**Triplet** We use a simple triplet loss-based approach as a baseline algorithm with access to both image and shape information during training, similar to prior approaches in shape retrieval [25]. A joint embedding is learned by using triplet loss [3, 40], creating positive pairs between image and point cloud features of same objects, and negative pairs between image and shape features from different object instances. Empirically we found that this performs better than using category labels. Inference is done by nearest centroid classification, building class prototypes that contain both appearance and shape information by averaging the individual support features.

### 4.3. Explicit Shape-Bias Improves Image-Based Generalization

We evaluate our method of adding shape bias to low-shot learning algorithms with state of the art low-shot image-only classification algorithms and show that shape bias improves performance in a low data regime. We present results on multiple datasets in Tables 3, 4, and 5 where we refer to models as Shape Bias (w/pc) if the shape-biased image encoder uses point cloud information to build prototypes (see Fig 2(b)) and (wo/pc) if there are no point clouds used to build prototypes for both validation and testing (see Fig 2(a)). Our approach of introducing shape bias, when trained with $\mathcal{L}_1$ and $\mathcal{L}_2$ losses improves the performance of image-only low-shot recognition algorithms in the

[4] Experiment implementation details included in the supplement

[5] Since none of the datasets have more than 10 classes for validation, the 20 and 30-way evaluations are done using a model trained for 10-way classification.

| Episode Setup → | 1-shot 5-way | 5-shot 5-way | 1-shot 10-way | 5-shot 10-way |
|---|---|---|---|---|
| RFS [49] | 56.67 ±0.30 | 72.64 ±0.26 | 43.79 ±0.16 | 60.61 ±0.11 |
| ProtoNet [44] | 50.11 ±0.31 | 64.44 ±0.24 | 36.44 ±0.17 | 46.70 ±0.26 |
| Triplet | 52.53 ±0.66 | 63.07 ±0.59 | 37.24 ±0.37 | 49.79 ±0.26 |
| SimpleShot [54] | 58.99 ±0.29 | 74.29 ±0.24 | 45.82 ±0.17 | 62.73 ±0.11 |
| Shape Bias (w/ pc) - SimpleShot - $\mathcal{L}_1$ only | 59.81 ±0.31 | 71.61 ±0.26 | 47.89 ±0.15 | 59.48 ±0.11 |
| Shape Bias (w/o pc) - SimpleShot | 60.23 ±0.30 | **75.59 ±0.24** | 47.92 ±0.15 | **64.88 ±0.11** |
| Shape Bias (w/ pc) - SimpleShot | **61.91 ±0.31** | 75.39 ±0.24 | **49.84 ±0.16** | 64.21 ±0.11 |
| FEAT [59] | 58.30 ±0.29 | 71.54 ±0.23 | 45.41 ±0.16 | 60.44 ±0.11 |
| Shape Bias (w/o pc) - FEAT | 60.19 ±0.31 | 74.66 ±0.25 | 48.6 ±0.16 | **64.08 ±0.11** |
| Shape Bias (w pc) - FEAT | **62.84 ±0.30** | **74.84 ±0.24** | **51.49 ±0.15** | 63.80 ±0.11 |

Table 3. Results on image-only and shape-biased low-shot recognition on **ModelNet40-LS**. Parenthesis show confidence intervals based on 5K low shot episodes. Bold indicates best performance between a low-shot learning approach with and without shape bias; underline indicates best overall. Adding shape bias improves performance in the 1-shot learning setting and has competitive performance otherwise.

| Episode Setup → | 1-shot 5-way | 5-shot 5-way | 1-shot 10-way | 5-shot 10-way | 1-shot 20-way | 5-shot 20-way | 1-shot 30-way | 5-shot 30-way |
|---|---|---|---|---|---|---|---|---|
| RFS [49] | 67.10 ±0.71 | 81.76 ±0.54 | 52.94 ±0.51 | 71.30 ±0.45 | 40.97 ±0.32 | 59.53 ±0.30 | 34.34 ±0.26 | 53.46 ±0.24 |
| ProtoNet [44] | 62.48 ±0.34 | 79.69 ±0.25 | 48.27 ±0.24 | 68.03 ±0.21 | 36.38 ±0.15 | 56.25 ±0.15 | 30.62 ±0.11 | 49.58 ±0.11 |
| Triplet | 63.87 ±0.34 | 73.95 ±0.62 | 48.78 ±0.54 | 60.44 ±0.48 | 36.34 ±0.35 | 47.28 ±0.32 | 30.09 ±0.25 | 40.08 ±0.24 |
| SimpleShot [54] | 68.87 ±0.32 | **83.69 ±0.23** | 55.22 ±0.24 | **73.58 ±0.19** | 43.05 ±0.16 | **62.64 ±0.14** | 36.78 ±0.12 | 56.22 ±0.12 |
| Shape Bias (w/o pc) - SimpleShot | 68.74 ±0.34 | 82.57 ±0.25 | 56.12 ±0.25 | 72.80 ±0.25 | 44.83 ±0.17 | 62.41 ±0.14 | 38.94 ±0.13 | **56.38 ±0.11** |
| Shape Bias (w/ pc) - SimpleShot | **70.96 ±0.33** | 81.33 ±0.24 | **58.47 ±0.25** | 70.81 ±0.20 | **46.96 ±0.17** | 60.3 ±0.14 | **40.59 ±0.14** | 54.00 ±0.11 |
| FEAT [59] | 70.66 ±0.33 | **84.13 ±0.23** | 57.15 ±0.24 | **74.29 ±0.19** | 44.84 ±0.16 | **63.65 ±0.14** | 38.43 ±0.12 | **57.42 ±0.11** |
| Shape Bias (w/o pc) - FEAT | 69.21 ±0.32 | 82.56 ±0.25 | 56.76 ±0.24 | 72.95 ±0.20 | 45.15 ±0.16 | 62.58 ±0.15 | 39.24 ±0.12 | 56.60 ±0.11 |
| Shape Bias (w/ pc) - FEAT | **71.58 ±0.34** | 81.45 ±0.25 | **59.09 ±0.25** | 71.00 ±0.20 | **47.45 ±0.17** | 59.98 ±0.15 | **41.38 ±0.12** | 53.64 ±0.11 |

Table 4. Results on image-only and shape-biased low-shot recognition on **Toys4K**. Parenthesis show 95% confidence intervals based on 5K low shot episodes. Bold indicates best performance for a low-shot approach with and without shape bias; underline indicates best overall. Adding shape-bias improves 1-shot performance when the number of low-shot ways is higher.

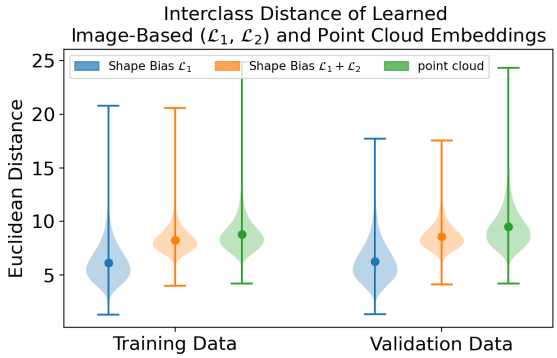


Figure 5. Examining the distribution of interclass distances in the mappings learned by minimizing either $\mathcal{L}_1$ or $\mathcal{L}_1 + \mathcal{L}_2$ relative to the reference point cloud embedding shows that adding $\mathcal{L}_2$ in a better approximation of the shape embedding space on both novel categories and categories seen during training.

low-data, one-shot learning regime for the SimpleShot and FEAT algorithms by up to 6%-points. For the (w/o pc) models that do not have any explicit shape information in the low-shot phase, we see a smaller one-shot improvement, but good five-shot performance. This indicates that shape bias is useful *without* any explicit shape information in the low-shot phase, and suggests possible future improvements by using strategies other than averaging to combine image and shape information in the low-shot phase.

We add shape-bias to SimpleShot by directly using the learned image to shape-mapping $f_i$ for nearest class mean classification, whereas for FEAT we train the set-to-set Transformer module on top of $f_i$, fine-tuning the model end-to-end as in the original FEAT design. The object shape embeddings for the low-shot supports are fixed and not trained further. Notice that as the total number of categories (the number of low-shot ways) increases, *the improvement in one-shot performance increases.* Further, our approach of learning shape bias significantly outperforms the triplet-loss based approach, indicating that first learning an embedding space with point clouds only is a better strategy than joint training with images and point clouds. All experiments for SimpleShot are averaged over 5 runs and for FEAT are averaged over 3 runs, indicating consistent performance improvements. To ensure statistical significance, for all experiments we perform 5K low-shot episodes and report results with 95% confidence intervals.[6]

### 4.4. Analysis of Pairwise Loss

We perform an analysis to determine the benefit of including the pairwise distance loss $\mathcal{L}_2$. In Figure 5, we plot the pairwise interclass distances of object instances from categories in the validation set for the learned mapping $f_i$ trained either with one loss or both losses (blue and orange respectively), along with the interclass distances in the point cloud embedding that $f_i$ is trained to learn. The greater the overall interclass distance, the better, and ideally the pairwise distance distributions are the same between the learned mapping and the point cloud mapping. Just optimizing $\mathcal{L}_1$ results in learning a poor mapping on both the training set

[6]For further qualitative and quantitative analysis please refer to the supplement.

| Episode Setup → | 1-shot 5-way | 5-shot 5-way | 1-shot 10-way | 5-shot 10-way | 1-shot 20-way | 5-shot 20-way |
|---|---|---|---|---|---|---|
| RFS [49] | 65.79 ±0.32 | 80.51 ±0.23 | 52.16 ±0.20 | 69.92 ±0.10 | 40.25 ±0.10 | 58.44 ±0.08 |
| ProtoNet [44] | 52.00 ±0.31 | 69.65 ±0.24 | 37.75 ±0.19 | 55.87 ±0.16 | 27.00 ±0.11 | 43.16 ±0.09 |
| Triplet | 61.07 ±0.34 | 71.43 ±0.28 | 46.89 ±0.22 | 58.37 ±0.18 | 35.09 ±0.12 | 46.20 ±0.08 |
| SimpleShot [54] | 66.73 ±0.32 | 80.93 ±0.22 | 53.37 ±0.21 | 70.32 ±0.16 | 41.09 ±0.12 | 59.09 ±0.08 |
| Shape Bias (w/o pc) - SimpleShot | 67.5 ±0.34 | **81.30 ±0.23** | 54.99 ±0.23 | **71.24 ±0.17** | 43.60 ±0.13 | **61.03 ±0.08** |
| Shape Bias (w/ pc) - SimpleShot | **69.72 ±0.32** | 80.93 ±0.24 | **57.49 ±0.21** | 70.75 ±0.16 | **46.24 ±0.12** | 60.21 ±0.08 |
| FEAT [59] | 67.81 ±0.32 | 80.25 ±0.23 | 54.35 ±0.22 | 70.18 ±0.16 | 42.12 ±0.12 | 59.01 ±0.08 |
| Shape Bias (w/o pc)- FEAT | 67.78 ±0.32 | **81.45 ±0.22** | 55.69 ±0.22 | **71.74 ±0.16** | 44.44 ±0.13 | **61.46 ±0.08** |
| Shape Bias (w/ pc) - FEAT | **70.24 ±0.32** | 80.95 ±0.22 | **58.45 ±0.22** | 70.95 ±0.16 | **47.03 ±0.13** | 60.43 ±0.08 |

Table 5. Results on image-only and shape-biased low-shot recognition on ShapeNet55-LS. Parenthesis show confidence intervals based on 5K low shot episodes. Bold indicates best performance between a low-shot approach with and without shape bias and underline indicates best overall. Adding shape bias leads to consistent improvement for both FEAT and SimpleShot.

and the novel classes in the validation set, whereas adding the pairwise term $\mathcal{L}_2$ leads to a better approximation of the point cloud embedding. The utility of $\mathcal{L}_2$ is also shown in Table 3, with the significant improvement over just $\mathcal{L}_1$ on SimpleShot with shape bias.

### 4.5. Shape Bias and Failure Analysis

To better understand the distinctions between the purely image-based low-shot classifier and the shape-biased low shot classifier, we compute the Pearson correlation ($p < 0.05$) between the accuracy achieved on the same 5K low-shot episodes for the point cloud model and the shape-biased and image-only classifiers (Figure 6). The shape-biased low-shot classifier correlates more strongly with the point cloud model across multiple datasets. This is evidence for a qualitative difference beyond classification accuracy between the shape biased and purely image low-shot classifiers. This would not be possible if the image data was such that it could not be classified differently as a result of introducing shape bias. Furthermore, in Table 6 we see that shape-biased SimpleShot misclassifies similarly to the point cloud SimpleShot, and that there is significant room for improvement by learning to map images into shape embeddings more accurately.

| | ModelNet | ShapeNet | Toys |
|---|---|---|---|
| 5-way | 38.73% | 44.81% | 57.59% |
| 10-way | 30.88% | 38.17% | 50.53% |

Table 6. Percent of queries misclassified by shape-biased SimpleShot but not misclassified by point cloud model (over 5K episodes). This indicates there is significant room for improvement by learning better maps from images to shape embeddings.

## 5. Discussion and Conclusion

This paper takes the first step in investigating the utility of shape bias for low-shot object categorization. Through extensive empirical analysis of our novel approach for adding shape bias to image-only low-shot learning algorithms, we demonstrate improved generalization. We also introduce Toys4K, a diverse and challenging dataset for object learning with the largest number of categories available

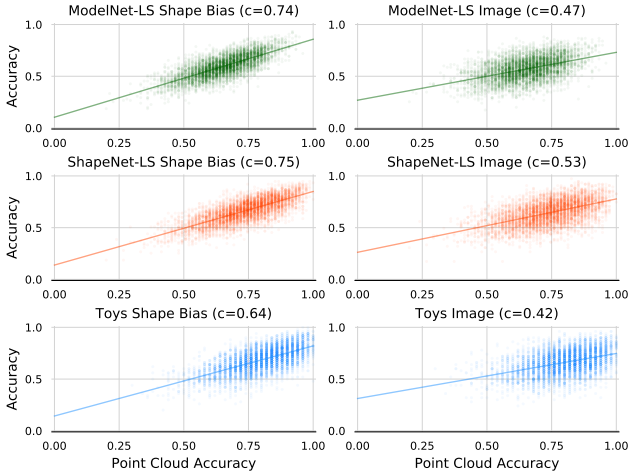


Figure 6. Accuracy of point cloud SimpleShot model vs. shape bias and image SimpleShot models over the same 5K episodes shows higher correlation between the point cloud model and shape bias model, indicating that the shape-biased model classifies more similarly to the point cloud model than the image only model.

to date. While dependence of our findings on synthetic object data limits our ability to draw conclusions about shape bias under more general conditions, it is essential since it is currently the only feasible way to obtain matched 2D and 3D data at a large enough scale. Moreover, synthetic data has been widely adopted for other vision tasks [5, 8, 20].

Progress in few-shot learning is crucial in order to overcome the need for large amounts of labeled training data. This work constitutes a step in a new direction: the exploitation of the natural biases of the visual world, such as object shape, in the design of few-shot architectures. Building on this approach by exploiting other sources of bias is a logical and exciting direction for future work.

## 6. Acknowledgement

We would like to thank Rohit Gajawada and Jiayuan Chen for their help with initial data collection. We also thank Zixuan Huang and Miao Liu for their helpful discussion of the paper draft. This work was supported by NSF award 1936970 and NIH award R01-MH114999.

## References


[1] blendswap.com. https://blendswap.com. 5

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 2, 3, 5, 6

[3] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 11–14. Springer, 2009. 6

[4] Yunlu Chen, Vincent Tao Hu, Efstratios Gavves, Thomas Mensink, Pascal Mettes, Pengwan Yang, and Cees GM Snoek. Pointmixup: Augmentation for point clouds. In *Proceedings of the European Conference on Computer Vision (ECCV) 2020*. 1, 2

[5] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3164–3174, 2020. 8

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[7] Gil Diesendruck and Paul Bloom. How specific is the shape bias? *Child development*, 74(1):168–178, 2003. 1

[8] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 8

[9] Shimon Edelman. *Representation and recognition in vision*. 1999. 1

[10] Yifan Feng, Zizhao Zhang, Xibin Zhao, Rongrong Ji, and Yue Gao. Gvcnn: Group-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 264–272, 2018. 2

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135, 2017. 1, 2

[12] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, pages 9516–9527, 2018. 2

[13] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 3

[14] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2018. 1, 3

[15] Lisa Gershkoff-Stowe and Linda B Smith. Shape and the first hundred nouns. *Child development*, 75(4):1098–1114, 2004. 1

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 6

[17] Chih-Hui Ho, Bo Liu, Tz-Ying Wu, and Nuno Vasconcelos. Exploit clues from views: Self-supervised and regularized learning for multiview object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9090–9100, 2020. 2

[18] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018. 1

[19] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. Learning robust visual-semantic embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3571–3580, 2017. 3

[20] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 8

[21] Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. Abc: A big cad model dataset for geometric deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9601–9611, 2019. 2, 5

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[23] Barbara Landau, Linda Smith, and Susan Jones. Object shape, object function, and object name. *Journal of memory and language*, 38(1):1–27, 1998. 1

[24] Barbara Landau, Linda B Smith, and Susan S Jones. The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321, 1988. 1

[25] Tang Lee, Yen-Liang Lin, HungYueh Chiang, Ming-Wei Chiu, Winston Hsu, and Polly Huang. Cross-domain image-based 3d shape retrieval by view sequence learning. In *2018 International Conference on 3D Vision (3DV)*, pages 258–266. IEEE, 2018. 3, 6

[26] Yangyan Li, Hao Su, Charles Ruizhongtai Qi, Noa Fish, Daniel Cohen-Or, and Leonidas J Guibas. Joint embeddings of shapes and images via cnn image purification. *ACM transactions on graphics (TOG)*, 34(6):1–12, 2015. 3

[27] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman and Company, 1982. 1, 3

[28] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5115–5124, 2017. 2

[29] Joseph L. Mundy. Object Recognition in the Geometric Era: A Retrospective. In Jean Ponce, Martial Hebert, Cordelia Schmid, and Andrew Zisserman, editors, *Toward Category Level Object Recognition*, pages 3–28. Springer, 2006. 1

[30] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018. 2

[31] Robert Osada, Thomas Funkhouser, Bernard Chazelle, and David Dobkin. Shape distributions. *ACM Transactions on Graphics (TOG)*, 21(4):807–832, 2002. 3

[32] poly.google.com/. https://poly.google.com/. 5

[33] Blender Proejct. https://blender.org. 6

[34] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 1, 2, 6

[35] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2016. 2

[36] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017. 1, 2, 6

[37] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018. 3

[38] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018. 1, 2

[39] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8247–8255, 2019. 3

[40] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 6

[41] Eli Schwartz, Leonid Karlinsky, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Baby steps towards few-shot learning with multiple semantics. *arXiv preprint arXiv:1906.01905*, 2019. 3

[42] Philip Shilane, Patrick Min, Michael Kazhdan, and Thomas Funkhouser. The princeton shape benchmark. In *Proceedings Shape Modeling Applications, 2004.*, pages 167–178. IEEE, 2004. 2

[43] sketchfab.com. https://sketchfab.com. 5

[44] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 1, 2, 6, 7, 8

[45] Stefan Stojanov, Samarth Mishra, Ngoc Anh Thai, Nikhil Dhanda, Ahmad Humayun, Chen Yu, Linda B Smith, and James M Rehg. Incremental object learning from contiguous views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8777–8786, 2019. 2

[46] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 2

[47] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. 2

[48] Atsushi Tatsuma, Hitoshi Koyanagi, and Masaki Aono. A large-scale shape benchmark for 3d object retrieval: Toyohashi shape benchmark. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–10. IEEE, 2012. 2

[49] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision (ECCV) 2020*, August 2020. 1, 2, 6, 7, 8

[50] turbosquid.com. https://turbosquid.com. 5

[51] Shimon Ullman. *High Level Vision: Object Recognition and Visual Cognition*. MIT Press, 1996. 1

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 6

[53] 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. 1, 2, 3, 5

[54] Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019. 2, 3, 4, 6, 7, 8

[55] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019. 1, 2, 4, 6

[56] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 1, 2, 3, 5, 6

[57] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018. 3

[58] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro OO Pinheiro. Adaptive cross-modal few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4847–4857, 2019. 3

[59] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 5, 6, 7, 8

[60] Qingnan Zhou and Alec Jacobson. Thingi10k: A dataset of 10,000 3d-printing models. *arXiv preprint arXiv:1605.04797*, 2016. 2, 5