Bridging the Gap between Humans and Machines in 3D Object Perception

Jasmine Collins

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-147
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-147.html

May 12, 2023

Bridging the Gap between Humans and Machines in 3D Object Perception

by

Jasmine Collins

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jitendra Malik, Chair
Professor Alexei A. Efros
Professor Alison Gopnik

Spring 2023

Bridging the Gap between Humans and Machines in 3D Object Perception

Abstract

Bridging the Gap between Humans and Machines in 3D Object Perception

by

Jasmine Collins

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Jitendra Malik, Chair

Humans possess a remarkable ability to extract general object representations from a single image, capturing not only shape and texture, but also 3D form. In contrast, 3D reasoning in many computer vision systems is often limited. This thesis present three efforts aimed towards bridging this gap in 3D object perception. First we introduce a new dataset that focuses on real-world, object-centered 3D understanding. The dataset provides a diverse set of objects corresponding to real household objects, with varying geometries and physically-based rendering materials. It also includes additional annotations describing each object, making it a valuable resource for training and evaluating computer vision models. Next, we design a method for automatically inferring the articulation of 3D objects. The method enables the interaction of 3D objects and can be used to generate more realistic and dynamic scenes. By understanding how different parts of an object move and interact, computer vision systems can better model and reason about complex 3D scenes in simulation. Finally, we investigate the effectiveness of contrastive learning with 3D data augmentation to generate multiple views of objects, a departure from the typical method of training single view images. We show that generating multiple views of objects can help computer vision systems learn better representations and improve their overall object understanding in terms of classification and shape perception. These contributions represent efforts towards bridging the gap between human and machine 3D object perception, ultimately enabling them to understand 3D objects from single images in ways that are more aligned with human perception.

To my parents.

Contents

List of Figures

List of Tables

Acknowledgments

I want to thank my advisor, Jitendra Malik, for inspiring and supporting me for the past 6 years. I also want to thank Bruno Olshausen and everyone in the Redwood Center. I spent the first few years of my Ph.D. in and out of Redwood and attended many thought-provoking talks and lab meetings there. I'd like to thank the members of my qualification exam and dissertation committee: Alexei Efros, Angjoo Kanazawa and Alison Gopnik. Alison, I've had a great time collaborating with you and your students. It's been a pleasure to be an honorary member of your lab!

I've learned so much from the people who have mentored me and that I've collaborated with over the past several years, both at Berkeley and during industry internships: Andrew Owens, David Sussillo, Jonathon Shlens, Johannes Balle, David Chan, Jessica Hamrick, Sandy Huang, Deepak Pathak, Pulkit Agarwal, Eunice Yiu, Shiry Ginosar, Matthieu Guillaumin, Anqi Liang, Hao Zhang, Frederic Devernay, Georgia Gkioxari, and Ari Morcos. Andrew, getting results the night before the conference deadline and staying up all night to write a paper with you was one of my Ph.D. highlights. Shiry, thanks for your wisdom and all the thought provoking chats about cognitive science.

I would like to thank the BAIR community, specifically Roxana Infante and Angie Abbatecola, who make sure things actually run smoothly behind the scenes! Angie, it has been so great getting to know you over the past 6 years. Your card decorating activities have always been such a treat. I want to thank everyone from lab, Ashish Kumar, Allan Jabri, Sasha Sax, Shubham Goel, Vickie Ye, Evonne Ng, Karttikeya Mangalam, Vongani Maluleke, Anastasios Angelopoulos, Haozhi Qi, Hang Gao, Andrea Bajcsy, Boyi Li, Ilija Radosavovic, Jathushan Rajasegaran, Antonio Loquercio, Georgios Pavlakos, Neerja Thakkar, Toru Lin, Dave Epstein, and Tim Brooks, to a name few. Allan, thanks for being a great friend and always being down for a good vent session. Ashish, thanks for building a computer with me during our first year. I'm glad we didn't both electrocute ourselves. I'll miss our tea chats!

I want to express deep thanks to my parents (Yanxi Liu and Robert Collins) and my grandparents. Thanks for giving me a life of opportunities. I also want to thank my cats, Jerry and Patricia. Being a human chair for the both of them day in and day out has definitely lowered my productivity, but has maximally increased my joy. Kyle Bilton, thanks for offering me encouragement and all the adventures we've been on together. To my close friends, I am eternally grateful. Eliza Kosoy, it has been so fun having you as a friend and collaborator. Sam Schoenholz and Jenny Marsh, my favorite couple, someday we will become pool sharks. Brian Cheung, your genuine curiosity and love for science is truly inspirational. Kelvin Xu, I feel so lucky we got to do our Ph.D. journeys at the same time! Thanks for being such a strong source of emotional support, advice, hot gossip, and being my best bud to get beers with. Sam Ritzer, it was so nice having a fellow yinzer out in the bay. Thanks for being there during my times of need. Brittany Quinn and Trevor Read, I'm so glad we've maintained our friendship despite the distance between us. Finally I want to thank my oldest and closest friend, Delaney Del Vecchio. Thanks for always being there for me, I couldn't have done it without you.

Chapter 1

Introduction

The field of artificial intelligence (AI) has made remarkable progress in recent years with rapid progress across a wide range of applications, from computer vision and natural language processing to robotics and healthcare. These advancements have been driven by the development of new deep learning architectures, the availability of large-scale datasets, and improvements in computing power, leading to a surge of interest and investment in the field.

In many areas, this technology has caught up to or even surpassed the ability of humans. Large language models trained on massive amounts of text data are able to generate coherent and grammatically correct sentences that are often indistinguishable from those written by humans [14, 21, 149, 116, 152]. This has led to exciting new applications in chatbots, language translation services, and content generation tools. Within computer vision, the advancement of generative models has enabled us to generate convincingly photorealistic images of almost anything, revolutionizing the field of image generation [13, 70, 60, 127].

Despite this impressive progress, there is still a large gap in the ability to represent images of arbitrary 3D objects. Methods for single-view 3D reconstruction [179, 103, 46] and neural rendering [104, 40, 5] have come a long way, yet currently only work well on a limited set of objects or are category-specific. The same is true for models trained on 2D object-centric data. When probed on out-of-distribution evaluation sets, state-of-the-art object recognition models still display large performance gaps compared to humans [42, 58]. We have yet to develop a computational model with object-centric representations that are aligned with the types of representations humans have.

This thesis makes an effort to study precisely some of those discrepancies between human and neural network-based object representations, and takes steps towards bridging that gap. In Part I we discuss major differences in human and machine perception, as well as efforts to compare the two in unified environments. Part I ends with proposing possible reasons for the differences between the visual representations of humans and neural network models. Part II dives into three efforts towards closing the gap between human and neural network performance, specifically on object perception tasks.

Part I: Human vs. Machine Perception This part highlights a set of results from cognitive and developmental psychology, measuring how humans (adults and children) perform on various reasoning and perceptual tasks. In many cases, we can directly compare the performance of computational models trained for a specific domain with that of humans. In Chapter 2, we discuss the tradition of using humans (typically adults) as baselines in AI research, and how this has led to algorithmic advancements that sometimes even surpass human performance. We then consider the merits of using children (rather than adults) as comparison points for computational models - namely because they reflect a human checkpoint with less explicit supervision, and often have more creative and exploratory behavior. The chapter reviews some recent work comparing children and AI agents in unified virtual environments. This approach has the benefit of being a 1:1 comparison, where computational baselines can actually be used to better understand and characterize the children's elusive behaviors. Chapter 3 focuses specifically on 3D object perception, the main topic of this thesis. This chapter covers relevant literature studying 3D object perception infants, young children, and adults. We also introduce a new study on object completion, replicating the results of the previous literature as well as extending them to measure preference for bilateral symmetry in 3D objects. We are also the first, to our knowledge, to compare directly to ImageNet trained ResNet models on the task of 3D object completion, and find that while ImageNet trained models have similar preferences to humans on volumetric completion, they do not share the same 3D symmetry biases. This raises the question of where and how these biases from, which leads into Chapter 4.

Chapter 4 is a brief discussion on some of the differences between the kinds of "input data" infant and toddler visual systems are developed on, compared to the typical kind of training datasets that are used for training neural networks. Namely, we highlight three key components in how children learn that are lacking in our typical training of computational systems: realistic training data, interaction with objects, and multiple views of objects with mostly self-directed play and sparse labels. The second part of this thesis introduces work that is inspired by making progress in each of these aspects.

Part II: Towards Bridging the Gap In Part II, we focus on efforts to bridge the gap between human and machine perception, making progress in more realistic 3D datasets, understanding articulated objects, and generating multi-view images for contrastive feature learning. Chapter 5 introduces ABO, a more realistic dataset for 3D object understanding based on product listings. Compared to existing datasets, ABO contains more realistic meshes that actually correspond to real objects that can be purchased on Amazon, and and physically based rendering materials so they can be rendered photorealistically. In this work we also introduce a set of benchmarks in single-view 3D reconstruction, material prediction, and multi-view cross-domain object retrieval. We show that ABO can be used to better understand the performance of state-of-the-art single-view 3D reconstruction methods, as well as be used as a training set for material prediction and image retrieval that generalizes better to real-world images.

In Chapter 6 we introduce an approach to solve the problem of *single-view articulation transfer*. That is, given a mesh in its rest-state and an image example of how it can articulate, automatically predict a 3D part segmentation and motion parameters such that the rest-state mesh can be articulated to match the input image. This work can be thought of as introducing a new shape understanding task, but also as a method for extending 3D assets with motion annotations without requiring expert artists (i.e. automatic dataset creation). Realistic 3D mesh models that can articulate could be valuable for downstream robotics applications and learning how to interact with and better represent household objects.

Chapter 7 is inspired by the idea that children learn robust object representations from observing many diverse viewpoints of the same object, with little supervision. In Chapter 7 we introduce a way to generate novel views of ImageNet images that can be used for training contrastive learning models. This 3D data augmentation leads to better overall performance on ImageNet accuracy (via fitting a linear probe) and higher accuracy on out-of-distribution downstream evaluations.

Taken together, the work in this thesis represents progress towards understanding the types of object-centric 3D representations humans have and takes steps towards building these ideals into computational systems.

Part I

Human vs. Machine Perception

Chapter 2

Comparing Humans and Machines with Unified Evaluations

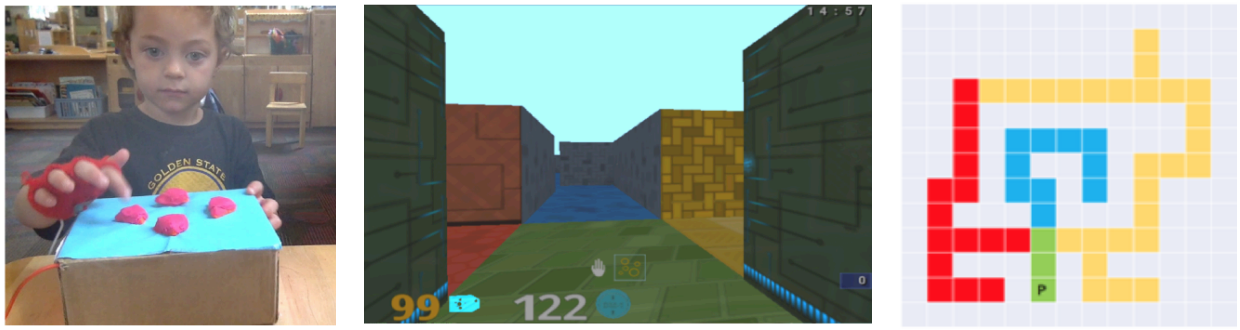

Figure 2.1: **Comparing children and AI agents in first-person maze environments.** Participant using the Arduino-based controller to explore a maze in DeepMind Lab (left). The maze that the child sees on the screen (middle). Top-down view of maze layout (right).

In this chapter we discuss experiments to compare the abilities of AI agents and pretrained models with that of humans. We cover literature that uses adults as baselines for model performance, as well as computational tools to better understand the behavior of children.

Human Performance as a Baseline

Prior work has shown that using human behavior as a comparison point for agents is useful in various settings. For example, human baselines on Atari games [108] have been a key component in progressing deep reinforcement learning (RL). Setting such baselines in deep RL serves as a North Star goal and has led to the development of algorithms with super-human performance. In computer vision, ImageNet [29] pretrained models are commonly tested against human recognition performance subject to various image distortions [41] and

cue-conflict stimuli [42]. Even within the distribution of natural training images, ImageNet-A [58] contains an evaluation set of *natural adversarial examples* for which humans can classify effortlessly, but ImageNet trained models have high error on.

Rosenfeld et al. [128] measure retrieval performance of state-of-the-art vision models on the "totally-looks-like" dataset, where humans typically agree on pairs of images with surprising levels perceptual similarity. They find that measuring image similarity with neural network-based features does not reproduce human rankings. The Winoground dataset [150] is another example of a task that humans perform extremely well on, correctly matching captions describing different spatial and compositional relationships with their respective image. Even the highest performing vision and language model shows a performance gap of $\sim$70% compared to human performance.

Children vs. Adults

While most existing human baselines are established using adults [108, 42, 128, 150], *children* can also serve as direct inspiration for AI model capabilities. Most human learning takes place during childhood, and young children explore new environments widely and effectively with little direct training [134, 27, 87, 133, 132], making their behavior an interesting benchmark for navigation and reasoning tasks. Interestingly, Bambach et al. [4] even found that training a neural network with training data derived from a head camera placed on toddlers outperformed data derived from adults.

Comparing Children and Machines

In work with Kosoy et al. [77], we present a methodology based on DeepMind Lab [7] for directly comparing child and agent behavior in a simulated maze exploration task, allowing us to precisely test questions about how children explore, how agents explore, and how and why they differ. An overview of the 3D first-person maze environment is shown in Figure 2.1.

We use this environment to first show differences in exploration behavior based on "no-goal" and "goal" conditioning. In the no-goal condition, the children are simply told to explore freely, whereas in the goal condition they are told to search for a "gummy" (the goal). We find that children exhibit a wide range of variability in how much they explore in the no-goal setting, and that those who explore more in the no-goal setting are able to find the goal faster in the goal setting condition. This may suggest that the "high explorers" build a mental model of the maze in the no-goal condition that enables them to perform more efficient navigation during the goal condition. We also find that children's search strategies between the two conditions differ. We compared children's behavior to that of a depth-first search (DFS) agent and found that in the no-goal condition, children made choices consistent with DFS 90% of the time, whereas in the goal condition children made choices consistent with DFS 96% of the time. This suggests that children's behavior becomes more systematic and efficient in the goal condition.

In a follow up study [78] we compare the ability of children and various RL agents to reason about causal overhypotheses in a unified, simulated environment based on the classical "blicket machine" [47]. The environment setup consists of a hierarchical casual structure where the higher-level structure determines the number of objects (i.e. blickets) that are needed to light-up the detector, and the lower-level structure describes which of the objects in this task correspond to blickets. Similar to the setup in Lucas et al. [102], we consider two possible higher-level causal hypotheses: conjunctive and disjunctive. In the conjunctive condition, a pair of blickets placed on the machine together cause it to activate. In the disjunctive case, any individual object that is a blicket can light the machine on its own (or in combination with other blickets). Children were given demonstrations depicting two blicket detectors (conjunctive and disjunctive) and asked to figure out how a third detector works, by trying out different object combinations. We compared the action sequences of children with various optimal policies based on information gain [114, 113, 24] and found that while children often take many more actions than is optimal, they are largely able to disambiguate between conjuntive and disjunctive blicket detectors. Further, they appeared to be optimizing for a diverse set of possible overhypotheses, beyond the two that were specified. How to construct a computational model that can also be creative and expand on the possible space of overhypotheses remains and interesting question for future work.

These works are first attempts to better understand children's behavior in comparison to computational models using unified environments, for maze exploration and navigation, as well as causal reasoning with a virtual blicket detector. In the next chapter, we compare the performance of children (and adults) to computational models in the domain of visual object perception using a novel 3D object completion task [176].

Chapter 3

Three-Dimensional Object Perception

Three-dimensional objects pose a challenge for our visual system, since we can only view objects from a single limited perspective at a given moment. Previous work found that given a limited perspective, infants represent 3D objects as complete volumes. Our study replicated this finding in 4- to 7-year-olds and adults, using an explicit prediction measure rather than looking times. We also explored whether humans have a bias to represent visually limited 3D objects as symmetrical rather than asymmetrical across shape, size, texture, and color. Overall, there was an above-chance preference for full volumetric and symmetrical object completion that increased with age. Low-level perceptual similarity of choices did not predict participants' choices. Moreover, we evaluated ImageNet trained ResNet-50 models on the same tasks: they represented objects as complete volumes, but did not show substantial preference for symmetrical 3D representations. This raises the possibility that incorporating human symmetry biases could improve computer vision.

3.1 Motivation

Humans can extract a general representation of a 3D object from a single perspective or image. This ability helps us to recognize objects, reason about object affordances, interact with objects, and understand scenes. However, it raises a puzzle. Given the limited information from a single perspective, how do we infer what the unrevealed portions of objects will look like? In this chapter, we aim to, first, investigate if the finding that infants prefer complete volumetric representations of objects from limited viewpoints [141, 140] applies to older children and adults viewing more diverse novel 3D objects, and second, examine preference for symmetrical completions of these novel objects in children and adults. We further compare participants' responses against the predictions made by ResNet-50 [55] models trained on ImageNet [29]. We tested three forms of the neural network: one that is supervised and

This chapter contains work that was first published as *Three-Dimensional Object Completion in Humans and Computational Models* in CogSci, 2022 [176].

trained on ImageNet, one that is self-supervised and trained on ImageNet as well, and one that is untrained serving as a random feature baseline.

Understanding human priors that underlie the 3D object completion from limited perspectives may also be relevant to computer vision. In neural network training, the ability to recognize objects from multiple viewpoints is acquired through an abundance of 2D images displayed from different viewpoints for each object category. There is often neither a direct transfer of 3D competence across categories, nor extracted or built-in priors about what novel views of objects should look like. Further, state-of-the-art detection and segmentation methods are only capable of recognizing and localizing visible object parts [55, 74].

We hypothesize that older humans, like infants, prefer complete volumes, and may likewise prefer symmetry to asymmetry in their object completion. In particular, they may care more about symmetry in geometry than in material, since many studies have shown that both children and adults have a shape bias. The shape bias refers to the inclination to classify, sort, and name objects on the basis of shape rather than other object elements such as color or texture [85, 139]. In contrast, an ImageNet trained neural network may not necessarily show preferences for complete volumes and geometric symmetry in 3D object completion, but it may show preference for material symmetry: like many standard convolutional neural networks, the training of the model may lead it to be more texture-biased than shape-biased [42, 126].

3.2 Background

Humans are able to conceive general representations of objects, and can make amodal completions of objects even when they are partially occluded [15]. Previous work showed that when presented with occluded two-dimensional surfaces [95] or self-occluded three-dimensional objects [94], adults prefer global as opposed to local completions. Furthermore, young infants represent simple 3D objects as complete and solid volumes instead of incomplete and hollow volumes even when they see a limited perspective that is compatible with either interpretation [141, 140]. Soska, Adolph, and Johnson [140] further suggested that the visuo-manual exploratory skills and self-sitting experience of infants facilitate their ability to complete 3D objects as solid volumes.

All these findings across different paradigms provide evidence that humans hold certain prior expectations about objects whose forms are not fully revealed, and those expectations influence their inferences about the unseen parts of objects. However, it is unclear what other sorts of predictions underpin human 3D object completion, beyond global completions for 2D surfaces and solid volumes for 3D objects. One potential perceptual bias that has not been explored is symmetry.

Objects in our visual world, natural or manmade, commonly exhibit mirror and/or rotational symmetries [28, 154]. It has been argued that much of our understanding of objects is guided by the perception and recognition of repeated or common patterns [148]. In fact, symmetry is a salient cue in human perception from early development. 4-month-old infants

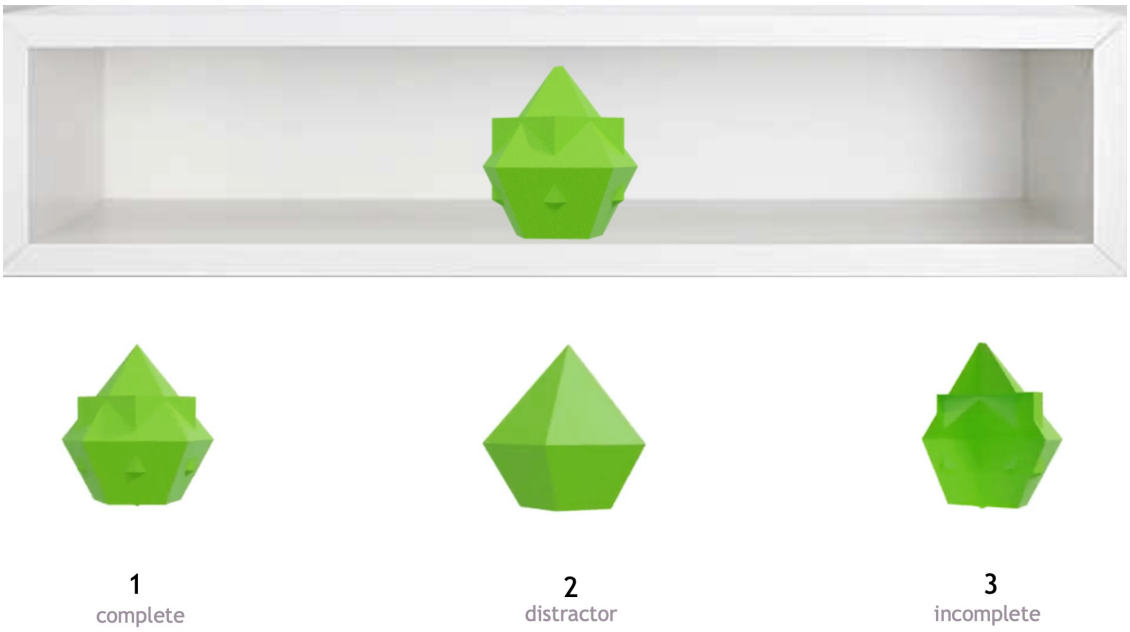

Figure 3.1: **Object completion task.** A test trial involving two possible options (1 and 3), and one impossible option (2).

can discriminate bilaterally symmetrical patterns from asymmetrical forms [36, 120]. Not only can infants process bilaterally symmetrical patterns more immediately than asymmetrical patterns [10], but they also develop faster and more accurate recognition memory for the former [11]. There have been theories suggesting that this ability to appreciate symmetry confers cognitive and evolutionary advantages. Symmetry in physical ornaments and motor patterns of living organisms also serves as an indicator of fitness in mate selection [33, 177]. It is thus interesting to explore the role symmetry plays in human 3D object completion and representation. When we approach a novel object from a single viewpoint, do we generally expect it to be symmetrical rather than asymmetrical? Does this expectation vary across development? What kinds of symmetry, such as symmetry in material (color and texture) and geometry (shape and size), do humans incorporate in their object completion? Are they equally favored, or are some kinds of symmetry more preferred than others?

3.3 Experiment 1

In Experiment 1, we expanded Soska and Johnson's infant study [141, 140] in three ways: one, we evaluated preferences for 3D solid volumetric completion in older children and adults; two, we tested more diverse and complex 3D stimuli; three, we tested explicit predictions about the object's appearance, rather than the more implicit infant looking-time measures. In addition to presenting participants with two possible options as in the original study,

we introduced a physically impossible distractor option that conflicted with the limited viewpoint to ensure participants were not merely making random guesses.

Methods

Participants 38 child participants aged between 4 years old and 7 years old (M_{age} = 5.94 years, SD = 1.16, 20 females) were recruited and tested at children's museums. More specifically, the sample comprises 10 4-year-olds (M_{age} = 4.53 years, SD = .36), 10 5-year-olds (M_{age} = 5.44 years, SD = .36), 9 6-year-olds (M_{age} = 6.48 years, SD = .32) and 9 7-year-olds (M_{age} = 7.49 years, SD = .25). An additional 4-year-old was tested but excluded from the sample analysis due to selecting impossible distractors in 1/3 of the test trials. In addition, 40 adult participants (M_{age} = 28.70, SD = 6.88; 20 females) were recruited on Prolific to complete the same task. The same experimental stimuli were also tested on a self-supervised, supervised, and untrained ResNet-50.

Stimuli and Procedure The study was performed on a computer screen. Participants were introduced to two virtual characters exploring a toy store, and were told that the toys were located on a shelf that was too high for them to reach, and so these objects could only be viewed from a limited perspective with their back parts being occluded. Participants were asked by the experimenter to help predict what 14 novel and abstract 3D toys would look like if they were taken off the shelf and turned around. The 14 objects were divided between 2 practice trials and 12 test trials. All objects were downloaded from Thingi10K, a large dataset of 3D printing models [180], and were further edited in Blender (an open-source 3D computer graphics software) for adaptation to the experiment.

Practice Trials The practice trials were designed to ensure that the participants understood the basic object completion task. In each of the 2 practice trials, participants saw a novel object on the shelf from a limited viewpoint followed by two 15° pivoting options representing what the two characters respectively thought the object would look like if it was turned around. Only one of the two options was physically possible and did not conflict with the limited viewpoint in terms of shape, size, texture, and color. Critically, the experimenter asked the participant, "See this object on the shelf? If you take if off the shelf and turn it around, will it look like [pointing to the two options] this or this?" After they selected one view, participants saw a full 360° rotation video of the object and were told whether their response was correct. All participants went through both practice trials before proceeding to the test trials.

Test Trials In each of the 12 test trials, participants were shown a novel object on the shelf from a limited viewpoint as in the practice trials. This time, the experimenter asked the same question, but participants had to decide among three instead of two different options, "See this object on the shelf? If you take it off the shelf and turn it around, will it look

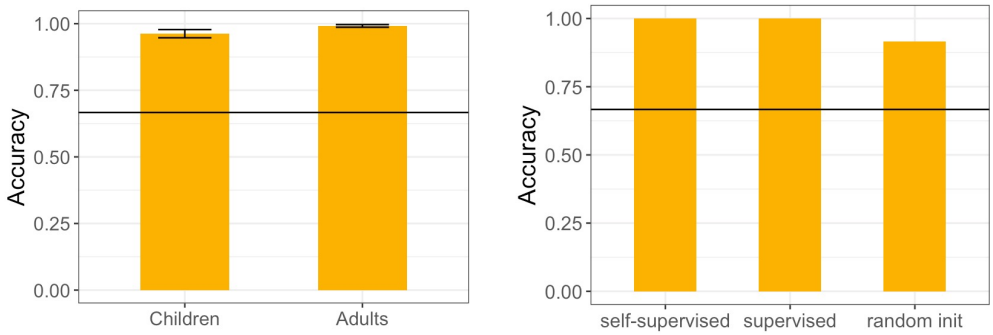

Figure 3.2: **Object completion accuracy of humans and ResNets.** Mean object completion accuracy of humans (left) and ResNets (right). Error bars show 1 standard error. Horizontal lines indicate chance-level accuracy (66.7%).

like 1, 2, or 3?" The choices included a possible complete volumetric option, a possible incomplete volumetric option, and an impossible distractor option that conflicted with the limited viewpoint of the object. They were presented in a randomized, counterbalanced order. The test trial setup can be seen in Figure 3.1.

Similar to the practice trials, these options in the test trials pivoted by 15° to facilitate the perception of the depth and three-dimensionality of the object, but the full rotation was not revealed. Once participants made a choice, they were rewarded with encouragement irrespective of what they chose. The goal was to motivate younger children to continue with the task without shaping their responses.

While all children, adults, and neural network were tested on the same stimuli, the task was administered in slightly different formats. Child participants were guided through the experiment by a human experimenter, whereas adult participants finished the task in the form of a Qualtrics survey (the experimenter's questions and instructions were written out in words). Since the neural network models were trained with static 2D images, they were fed with the center frame of each 15-degree pivoting option for evaluation. In each trial, we extracted the feature representation from each neural network for the limited viewpoint image and each option image, and considered the option with the highest cosine similarity with the limited viewpoint as the model's choice.

Results and Discussion

Object Completion Accuracy First, we determined participants' object completion accuracy to evaluate their ability to understand our task and to complete the rest of each object with a form that was not physically impossible. This was computed as the proportion of test trials in which participants did not choose the distractor options. In other words, we tested whether participants preferred the possible choices to the impossible distractor option. Children scored a mean of 97.4% (SE = .89%); age effects were not observed from 4 to 7

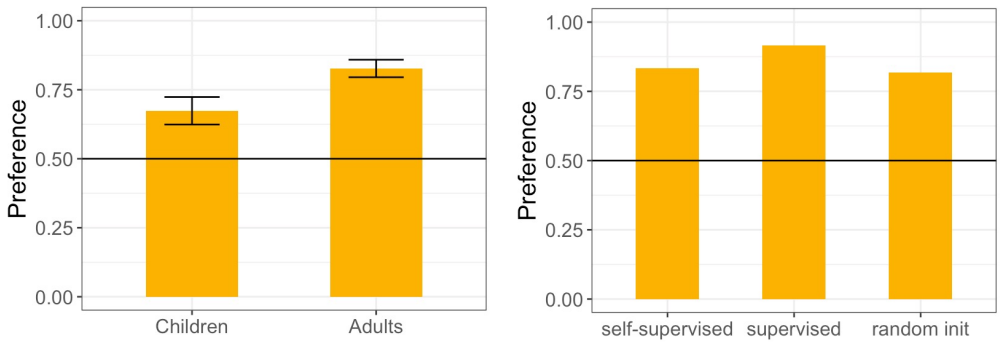

Figure 3.3: **Mean complete volumetric preference of humans and ResNets.** Mean complete volumetric preference of humans (left) and ResNets (right). Error bars show 1 standard error. Horizontal lines indicate no preference (50%).

years old. Adults scored a mean of 99.2% (SE = .50%) (Figure 3.2, left). The high accuracy scores in objection completion confirm that both children and adults could demonstrate object completion and the task was appropriate for both age groups. The Welch's two-sample, two-tailed t-test showed that adults' scores were marginally higher than children's, $t(77) = 1.75$, $p = .084$; Cohen's effect size $d = .40$ suggested small significance. The self-supervised and supervised ResNet-50 scored 100%; even the untrained ResNet-50 (random baseline) scored 91.7% (Figure 3.2, right), suggesting that even random pixel-level features capture the information needed to solve the task.

Complete Volumetric Preference Complete volumetric preference refers to the proportion of test trials in which the complete volumetric option was selected out of the total number of trials in which the distractor option was not chosen. We dropped one child who selected the distractor options in 4 out of the 12 trials. All adults passed this critical requirement and were considered in the analysis.

Both children and adults chose the complete volumetric option significantly above the chance level of 50% (symmetrical option vs. asymmetrical option). Children had a mean complete volumetric preference score of 58.6% (SE = 4.27%), while adults had a mean score of 82.7% (SE = 3.18%) (Figure 3.3, left). While age effects were not observed from 4 to 7 years old, the Welch's two-sample, two-tailed t-test on children and adults revealed significantly stronger preference in adults, $t(77)$, $p < .001$. Cohen's effect size $d = 1.03$ suggested high practical significance. Thus, our complete volumetric preference continues to strengthen past the age of 7, potentially through increasing exposure to the visual statistics of objects (for instance, noticing there are more objects with complete volumes in the environment).

Like human participants, all three forms of ResNet-50 also showed preference for complete volumes over incomplete volumes: the self-supervised network showed a complete volumetric preference of 83.3%, the supervised network 91.7%, and the untrained network 81.8%. The incomplete objects possessed convex surfaces which often had darker shadings than the

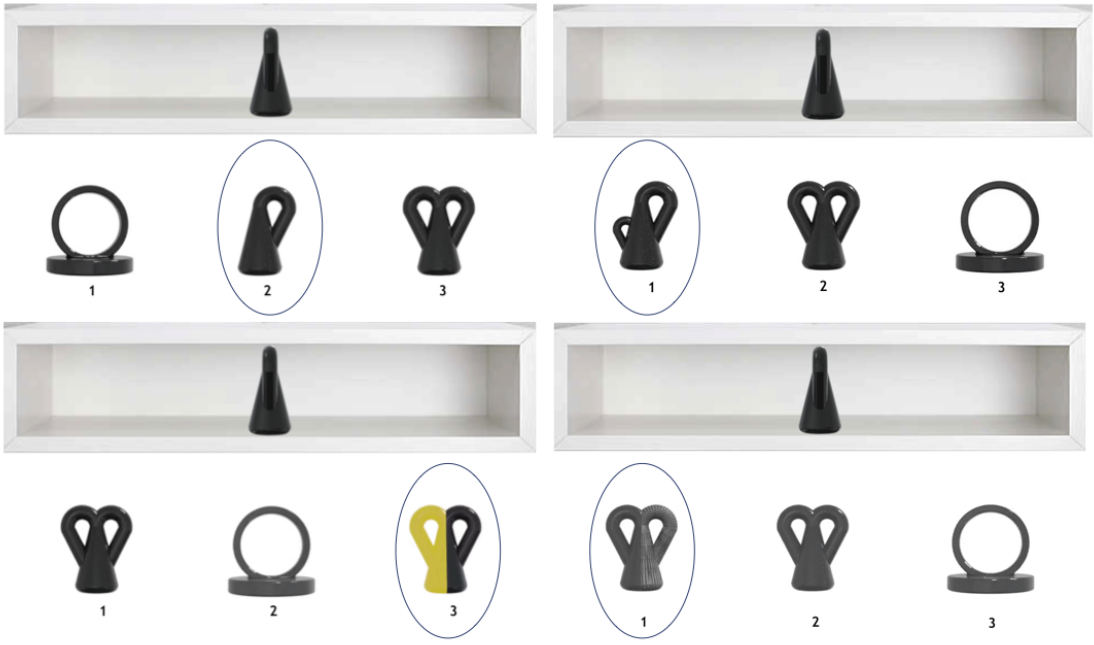

Figure 3.4: **Sample stimuli for studying bilateral symmetry preference.** Four types of asymmetrical edits (circled) for the same object (from left to right, top to bottom): shape, size, color, and texture, as presented in the test trials.

surfaces of the reference object in the shelf. This might have prompted ResNet-50, even in its untrained form, to eliminate the incomplete option.

3.4 Experiment 2

We adopted the same paradigm in Experiment 1 to study another potential preference in 3D object completion: bilateral symmetry.

Methods

Participants 82 child participants aged between 4 years old and 7 years old (M_{age} = 5.90 years, SD = 1.13, 40 females) were recruited and tested at children's museums. More specifically, the sample comprised 16 4-year-olds (M_{age} = 4.29 years, SD = .27), 22 5-year-olds (M_{age} = 5.25 years, SD = .29), 22 6-year-olds (M_{age} = 6.31 years, SD = .22) and 22 7-year-olds (M_{age} = 7.29 years, SD = .28). Sixteen additional children were tested but excluded from the sample analysis as they selected the incorrect distractors in at least 1/3 of the test trials. This could be due to inattention, language comprehension issues, or simply inability

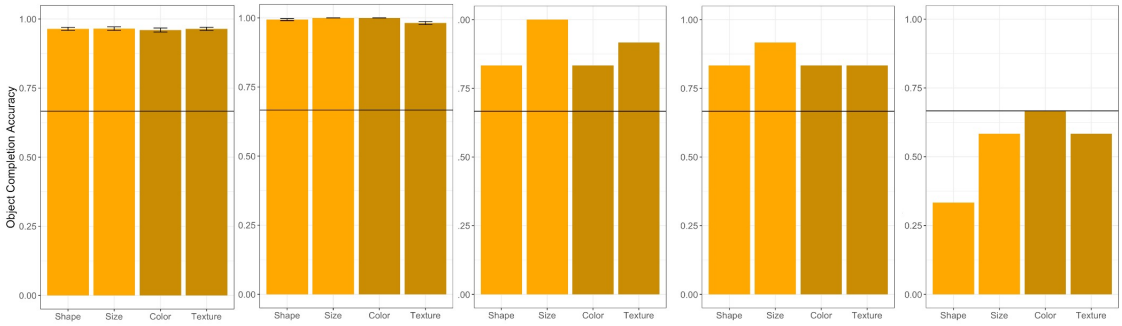

Figure 3.5: **Mean object completion accuracy on symmetry stimuli.** Mean object completion accuracy of (from left to right) children, adults, self-supervised, supervised, and untrained ResNet-50 in geometry and material conditions. Error bars show 1 standard error. Horizontal lines indicate chance-level accuracy (66.7%).

to understand the task. Again, the same task was also tested on 40 adult participants (M_{age} = 23.65, SD = 4.94; 31 females) who were recruited on Prolific and on ResNet-50.

Stimuli and Procedure The task design and objects used in Experiment 2 were identical to those used in Experiment 1, but the available options were different. In Experiment 2, the choices included a possible symmetrical option, a possible asymmetrical option, and an impossible distractor option. The asymmetrical option could be asymmetrical in its geometry (shape or size) or material (color or texture). To prevent the inherent differences between objects from confounding with the type of asymmetry presented, each object was edited for asymmetry in all four conditions (shape, size, color, and texture) (Figure 3.4). These four types of edits of the same object were then allocated to four separate test sets. Participants were randomly assigned to one of the test sets and thus only saw each object once. This enabled preferences for the different types of asymmetry to be assessed and compared across the same set of objects. In other words, the 12 test trials were comprised of 3 trials with asymmetrical shape options, 3 trials with asymmetrical size options, 3 trials with asymmetrical color options, and 3 trials with asymmetrical texture options. The order of options was randomized.

Results and Discussion

Object Completion Accuracy One-sample, two-tailed t-tests on children and adults respectively showed significantly above-chance accuracy (66.7%) across all conditions (p ¡ .001). Children had mean accuracy scores 96.4% in shape (SE = .58%), 96.5% in size (SE = .63%), 96.0% in color (SE =.71%), and 96.4% in texture (SE = .59%). Adults also scored highly above chance for shape (M = 99.4%, SE = .35%), size (M = 100%, SE = 0%), color (M = 100%, SE = 0%), and texture (M = 98.1%, SE = .56%) as well (Figure 3.5). In

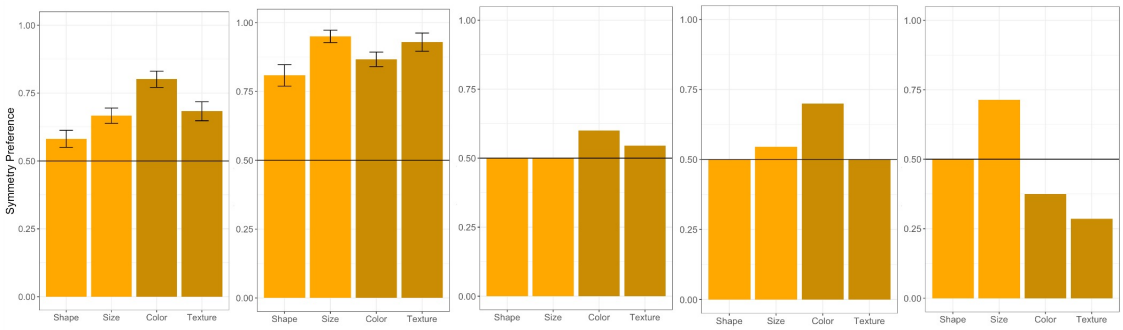

Figure 3.6: **Mean symmetry preference of humans and ResNets.** Mean symmetry preference of (from left to right) children, adults, self-supervised, supervised, and untrained ResNet-50 in geometry and material conditions. Error bars show 1 standard error. Horizontal lines indicate no preference (50%).

contrast, the neural network models demonstrated poorer performance than humans on the task. The self-supervised ResNet-50 scored a mean accuracy of 83.3% in shape and color, 100% in size, and 91.7% in texture; the supervised ResNet-50 scored 83.3% in shape, color, and texture, and 91.7% in size; the untrained ResNet-50 did not show any capability of completing 3D objects: it scored 33.0% in shape, 58.4% in size, 66.7% in color, and 58.4% in texture (Figure 3.5).

Symmetry Preference across Conditions We measured participants' symmetry preference by computing the proportion of test trials in which the symmetrical option was selected out of the total number of trials in which the distractor option was not chosen. We dropped sixteen children who selected the distractor options beyond chance level in at least 4 out of the 12 trials, retaining only those with an accuracy score above 66.7% for our analysis of symmetry preference. Again, all adults passed this critical requirement and were considered in the analysis. Both children and adults chose the symmetrical option significantly above the chance level of 50% (Figure 3.6). Children attained mean symmetry scores of 58.1% in shape (SE = 3.13%), 66.7% in size (SE = 2.82%), 80.1% in color (SE = 3.00%), and 68.3% in texture (SE = 3.49%). Meanwhile, adults had mean symmetry scores of 80.8% in shape (SE = 3.89%), 95.0% in size (SE = 2.25%), 86.7% in color (SE = 2.61%), and 92.9% in texture (SE = 3.32%).

Children and adults respectively revealed significantly above-chance symmetry preference across all conditions ($p < .001$), but there was an exception: children as a group did not show symmetry preference for shape beyond chance level, $p = .12$. Unlike human participants, the neural network models showed much weaker symmetry preference: the self-supervised ResNet-50 scored 50.0% in shape and size, 60.0% in color, and 54.5% in texture; the supervised ResNet-50 scored 50.0% in shape and texture, 54.5% in size and 70.0% in color; the untrained ResNet-50 scored 50.0% in shape, 71.4% in size, 37.5% in color, and 28.6% in tex-

ture (Figure 3.6). Overall, none of the neural networks showed a substantial preference for completing objects in a symmetrical fashion as our human participants did. Perhaps understanding symmetry in objects requires 3D reasoning that cannot be acquired from training on ImageNet. Unlike in Experiment 1, the symmetric option cannot be deduced from 2D cues alone.

A within-subjects ANOVA with condition (shape, size, color, texture) as the independent variable and symmetry preference as the dependent variable yielded a main effect of condition in both children, $F(3,324) = 8.82$, $p < .001$, and adults, $F(3,117) = 7.36$, $p < .001$. We used Bonferroni corrections, leading to an adjusted alpha of .0125 in our multiple comparisons. We found that children had greater symmetrical completion preference for color than for shape $t(81) = 5.06$, $p < .001$, and size, $t(81) = 3.26$, $p < .01$. Adults on the other hand, exhibited significantly stronger symmetrical completion preference for size than for shape, $t(39) = 3.15$, $p < .05$. The differences in preferences between the types of symmetries are more flattened out in adulthood than in childhood, as adults preferences approach ceiling.

Symmetry Preference and Age Overall, the results of the study suggest a general increase in preference for symmetrical object completions across age. Mean symmetrical preference across conditions increased from a mean of 61.5% (SE = 2.92%) among the 4-year-olds (M_{age} = 4.29 years, SD = .27) to a mean of 77.3% (SE = 3.07%) among the 7-year-olds (M_{age} = 7.29 years, SD = .28). In an exploratory analysis, given the few number of trials per condition, we compared symmetry preferences for geometry (shape and size combined) (M = 61.7%, SE = 2.03%) and material (color and texture combined) (M = 73.5%, SE = 2.57%). We found that there is a significantly positive correlation between age and symmetry preference for both geometry, $r(81) = .34$, $p < .01$, and material, $r(81) = .35$, $p < .01$, respectively. The developmental progression of symmetry preferences in both the material and geometry conditions appear to be similar, with material symmetry preference being stronger than geometry symmetry, $t(81) = 3.58$, $p < .001$, Cohen's effect size $d = .56$ suggested medium significance. When comparing the symmetry preferences in each of these four symmetry conditions between all child participants and all adult participants (with Bonferroni corrections applied, leading to an adjusted alpha of .025), we found that adults showed a higher symmetry preference than children for shape, $t(120) = 4.54$, $p < .001$), size, $t(120) = 7.86$, $p < .001$, and texture, $t(120) = 5.11$, $p < .001$, but not for color, $t(120) = 1.66$, $p = .10$. This suggests that from childhood to adulthood, there is further strengthening of symmetry preferences in object completion with regards to shape, size, and texture. The older we get, the more biased towards symmetry we are when we imagine the occluded surfaces of 3D objects.

3.5 Discussion

The present study provides initial evidence that both children aged 4-7 years old and adults incorporate a preference for complete volumes and symmetry into their completion of novel

objects from limited perspectives. Our task asks participants to select what is most plausible to them within a given set of possibilities, so it does not probe what participants might expect outside these options. Nevertheless, this finding complements the existing work on object completion and further demonstrates how priors or expectations concerning the occluded parts of objects may shift across development. The preference for completing novel objects as solid, complete volumes is evident in the looking times of infants [141], and we showed via an explicit prediction measure that this preference continues to strengthen with increasing age. Similarly, completing objects in a symmetrical fashion is observed in early childhood, but adults show an even stronger preference for doing so. This implies that complete volumetric and symmetrical completion preferences are not rigid, built-in priors. Infants' self-sitting experience and visual-manual exploration of objects predicted their looking longer at a volumetrically incomplete object than a volumetrically completion object in 3D object completion [140]. Preferences may continue to develop through dynamic interactions with objects and scenes in the environment over time. More empirical work is needed to investigate this possibility at a later age.

Some may argue that instead of actually demonstrating complete volumetric and symmetrical object completion preferences, participants were choosing options based on low-level perceptual similarity (e.g., an option was chosen because it looked most perceptually similar to the limited viewpoint); or based on whether or not the 2D limited viewpoint was symmetrical (e.g., the symmetrical option was chosen because the 2D limited viewpoint presented was symmetrical). To test the first possibility, we used F-score to evaluate the perceptual similarity of each option relative to the limited viewpoint and relative to the other options in terms of 2D and 3D geometry. The F-score evaluates the distance between object surfaces and counts the percentage points that lie within a certain distance on another object under comparison. We also used SSIM [159] to evaluate perceptual differences in terms of material. We did not find any statistically significant correlation between the 2D or 3D F-scores (whether it be relative to the limited viewpoint or relative to other options), or SSIM, and how often an option was chosen by children and adults in both experiments. To test for the second possibility, we evaluated whether the limited viewpoints were symmetrical or asymmetrical along the vertical axis (since we evaluated preferences for bilateral symmetry across the trials). Again, performance was similar whether the limited viewpoint was symmetrical or not.

Our finding of a robust human expectation for symmetry in object completion aligns with the vast literature on human sensitivity to and preference for symmetry [117, 130]. The increasing tendency to complete objects in a symmetrical manner with age also resonates with existing developmental studies that human symmetry preference increases with age [64, 63]. However, contrary to our initial hypothesis, and in spite of children's well-established shape bias, children exhibited the strongest preference for symmetrical object completion for color and the weakest preference for shape. One possible reason is that bilateral asymmetries in shape (one half of the object is one shape and the other half is a different shape) may occur more frequently than bilateral asymmetries in color (one half of the object is one color and the other half is a different color) in the artefacts in our everyday environments, such as kitchen

tools or toys. From infancy, humans have a strong capacity for visual statistical learning [35, 75], and may have come to expect fewer color asymmetries than shape asymmetries. This could explain the relatively lower shape symmetry preference in both children and adults. Another possibility stems from the fact that young children between 3.5 and 6 years of age also tend to be faster and more accurate at naming the color of an object when the object was presented in an abstract shape [121]. The abstractly shaped objects in our study may therefore have increased children's attention to color and driven their preference for color symmetry.

Like children and adults, pretrained ResNet-50 networks completed objects at above-chance accuracy in both experiments, while the untrained network failed to do so in Experiment 2, potentially because the distractor options were harder to visually distinguish in that experiment. Critically, in their object completion, the neural networks preferred volumetrically complete objects in Experiment 1, but not symmetrical objects in Experiment 2. Compared to the reference objects on the shelf and the volumetrically complete options, the volumetrically incomplete options had convex surfaces reflected through darker shadings. Darker shadings might contribute to a lower cosine similarity, thereby causing the neural networks to avoid the incomplete options. This also lends some support to the hypothesis that neural networks are texture-biased and attend less to shape [42, 44]. That said, SSIM, which includes luminance and contrast masking terms, did not predict the choices of the neural networks. By contrast, the neural networks cannot use 2D cues such as shading in Experiment 2. Trained or untrained, they demonstrated little to no preference for symmetrical completions, suggesting that 3D symmetrical object completion may not be achieved solely through cosine similarities following training on 2D images. The ImageNet trained ResNet-50 models did not necessarily develop 3D understanding of objects the way humans do, and thus may not further develop preferences for symmetry in 3D objects.

From a broader perspective, the present study may shed light on understudied aspects of object completion, an important problem in both human cognition and computer vision. We examined preferences for solid volumes and bilateral symmetry, the developmental trend from childhood to adulthood, as well as the differences between humans and ResNet-50 neural networks in completing occluded parts of an object. Given the small number of trials (n = 12) in these experiments, we hope to scale up the object dataset to generate more reliable findings. Further studies may also explore other variables such as the axis of symmetry and the animacy of objects to further understand this important aspect of perception. Manmade structures tend to possess more segments of straight lines, longer linear lines, and coterminations than animates or natural objects [65]. Hence, it is possible that symmetry preferences may differ between these categories. On the computational side, we plan to train neural networks on 3D (as opposed to 2D) object data and test the preferences of single-view 3D reconstruction networks.

Chapter 4

The Gap in Training Data

Why might there be such a discrepancy in the way humans and neural networks represent visual data? While there are several factors at play, including the biological differences between humans and artificial systems, the specificities of the neural network architecture, and the optimization algorithm for neural networks not being entirely biologically plausible, this thesis proposes that an important and potentially understudied culprit is the training data that is used. Many aspects of the training data that we use to train neural networks is drastically different from the visual inputs that children are "trained" on. In this chapter, we discuss a few of those key differences.

4.1 Realism of Training Data

One of the most significant differences is the realism of the training data used to trained deep networks. Children learn to see and reason about 3D objects in the real world, exposed to a wide range of visual inputs, including variations in lighting, texture, and color. In contrast, neural networks are often trained on images that are artificially generated [17], or captured in the real world but contain category or photographer bias [29]. This lack of realism in the training data can significantly impact the ability of neural networks to accurately represent visual data.

To bridge this gap, one potential solution is to use a dataset with realistic-looking 3D models with physically-based rendering materials. These 3D models can then be rendered photorealistically, under many different viewpoints and lighting conditions, exposing the neural network to a much wider range of visual inputs that are closer to what humans encounter in the real world. In addition to providing a more diverse set of visual inputs, a dataset with realistic-looking 3D models can be used to render images in a controlled and systematic way, allowing researchers to study the impact of specific factors, such as lighting or texture, on the performance of neural networks. Second, using such a dataset can help to overcome limitations in collecting real-world training data, such as the cost and difficulty of capturing images under diverse enough lighting and material conditions.

4.2 Interaction with Objects

Children learn about the world around them not only by passive observation, but also by actively exploring and interacting with it [137, 140]. They manipulate objects, move around the environment, and experiment with cause and effect relationships [47]. This active exploration and interaction plays a critical role in shaping their understanding of the world. In contrast, current neural network models are often trained on static images and videos, without the ability to interact with the objects and environment. This lack of interaction can result in models that are poorly equipped to represent the dynamic and interactive nature of objects in the world.

One potential solution to help address this limitation is to have better 3D model datasets that can articulate. By incorporating the ability to manipulate and interact with objects during training, we can create more dynamic and realistic representations of objects. Mesh models with motion assets can be used to generate large amounts of training data that include variations in object articulation, as well as potentially object-object and object-environment interactions. This type of data can help deep networks learn more robust and accurate representations of objects.

4.3 Multiple Views and Sparse Labels

Clerkin et al.[23] conducted an intriguing study using head-mounted cameras to record the visual inputs that infants typically receive in their daily lives. One of the key findings of the study was that infants are frequently exposed to a few specific objects, such as a parent's face or a special cup, viewed from many different angles, distances, backgrounds, and lighting environments. Further, their learning about objects is largely driven by self-play interspersed with sparse labeling events from a parent or caretaker [142]. This type of visual input is vastly different from the training data that we typically use to train neural networks, which usually consists of single views of objects from relatively stereotyped viewpoints[29, 97].

In this thesis we consider how incorporating the types of visual inputs that infants typically receive into the training data of neural networks can change the learned representations. One approach is to train on multiple viewpoints of objects, rather than only use single view images. In addition, rather than relying solely on labeled data, which may not always be available or representative of real-world scenarios, we can use self-supervised learning techniques that leverage the natural structure and regularities in the data to learn more robust and generalizable visual representations of objects.

In Part II we discuss efforts targeting each of these key differences. In Chapter 5 we introduce a more realistic 3D dataset with physically-based rendering materials for photorealistic rendering. In Chapter 6 we introduce a methology to infer motion assets for rigid 3D models. Finally, in Chapter 7 we study the effect of training with multiple views (vs. single views) on self-supervised feature learning.

Part II

Towards Bridging the Gap

Chapter 5

A Realistic Dataset for 3D Object Understanding

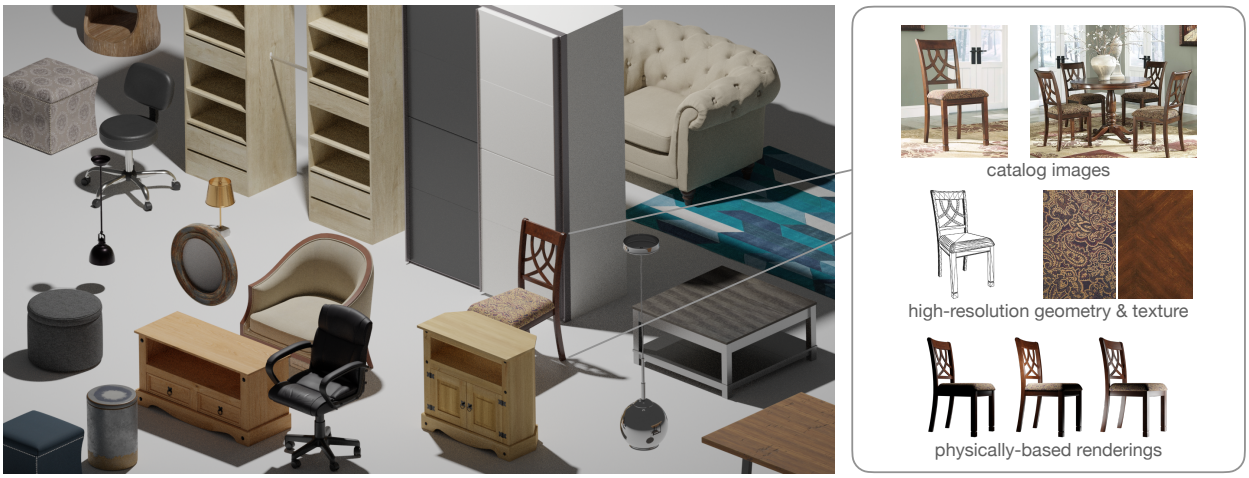

Figure 5.1: **ABO is a dataset of product images and realistic, high-resolution, physically-based 3D models of household objects.**

We introduce ABO, a new large-scale dataset designed to help bridge the gap between real and virtual 3D worlds. ABO contains product catalog images, metadata, and artist-created 3D models with complex geometries and physically-based materials that correspond to real, household objects. We derive challenging benchmarks that exploit the unique properties of ABO and measure the current limits of the state-of-the-art on three open problems for 3D object understanding: single-view 3D reconstruction of real world objects, material estimation, and cross-domain multi-view object retrieval.

This work was first published as *ABO: Dataset and Benchmarks for Real-World 3D Object Understanding* in CVPR, 2022 [25].

Dataset	# Models	# Classes	Real images	Full 3D	PBR
ShapeNet [17]	51.3K	55	✗	✓	✗
3D-Future [38]	16.6K	8	✗	✓	✗
Google Scans [32]	1K	-	✗	✓	✗
CO3D [125]	18.6K	50	✓	✗	✗
IKEA [96]	219	11	✓	✓	✗
Pix3D [144]	395	9	✓	✓	✗
PhotoShape [119]	5.8K	1	✗	✓	✓
ABO (Ours)	8K	63	✓	✓	✓

Table 5.1: **A comparison of the 3D models in ABO and other commonly used object-centric 3D datasets.** ABO contains nearly 8K 3D models with physically-based rendering (PBR) materials and corresponding real-world catalog images.

5.1 Motivation

Progress in 2D image recognition has been driven by large-scale datasets [80, 29, 97, 143, 52]. The ease of collecting 2D annotations (such as class labels or segmentation masks) has led to the scale of these diverse, in-the-wild datasets, which in turn has enabled the development of 2D computer vision systems that work in the real world. Theoretically, progress in 3D computer vision should follow from equally large-scale datasets of 3D objects. However, collecting large amounts of high-quality 3D annotations (such as voxels or meshes) for individual real-world objects poses a challenge. One way around the challenging problem of getting 3D annotations for real images is to focus only on synthetic, computer-aided design (CAD) models [17, 180, 76]. This has the advantage that the data is large in scale (as there are many 3D CAD models available for download online) but many of the models are low quality or untextured and do not exist in the real world. This has led to a variety of 3D reconstruction methods that work well on clear-background renderings of synthetic objects [22, 49, 168, 103] but do not necessarily generalize to real images, new categories, or more complex object geometries [147, 6, 8].

To enable better real-world transfer, another class of 3D datasets aims to link existing 3D models with real-world images [165, 166]. These datasets find the closest matching CAD model for the objects in an image and have human annotators align the pose of the model to best match the image. While this has enabled the evaluation of 3D reconstruction methods in-the-wild, the shape (and thus pose) matches are approximate. Further, because this approach relies on matching CAD models to images, it inherits the limitations of the existing CAD model datasets (i.e. poor coverage of real-world objects, basic geometries and textures).

The IKEA [96] and Pix3D [144] datasets sought to improve upon this by annotating real images with exact, pixel-aligned 3D models. The exact nature of such datasets has allowed them to be used as training data for single-view reconstruction [46] and has bridged some

of the synthetic-to-real domain gap. However, the size of the datasets are relatively small (90 and 395 unique 3D models, respectively), likely due to the difficulty of finding images that exactly match 3D models. Further, the larger of the two datasets [144] only contains 9 categories of objects. The provided 3D models are also untextured, thus the annotations in these datasets are typically used for shape or pose-based tasks, rather than tasks such as material prediction.

Rather than trying to match images to synthetic 3D models, another approach to collecting 3D datasets is to start with real images (or video) and reconstruct the scene by classical reconstruction techniques such as structure from motion, multi-view stereo and texture mapping [20, 136, 32]. The benefit of these methods is that the reconstructed geometry faithfully represents an object of the real world. However, the collection process requires a great deal of manual effort and thus datasets of this nature tend to also be quite small (398, 125, and 1032 unique 3D models, respectively). The objects are also typically imaged in a controlled lab setting and do not have corresponding real images of the object "in context". Further, included textured surfaces are assumed to be Lambertian and thus do not display realistic reflectance properties.

Motivated by the lack of large-scale datasets with realistic 3D objects from a diverse set of categories and corresponding real-world multi-view images, we introduce Amazon Berkeley Objects (ABO). Overall, ABO contains 147,702 product listings associated with 398,212 unique catalog images, and up to 18 unique metadata attributes (category, color, material, weight, dimensions, etc.) per product. ABO also includes "360º View" turntable-style images for $8,222$ products and 7,953 products with corresponding artist-designed 3D meshes. In contrast to existing 3D computer vision datasets, the 3D models in ABO have complex geometries and high-resolution, physically-based materials that allow for photorealistic rendering. A sample of the kinds of real-world images associated with a 3D model from ABO can be found in Figure 5.1. The dataset is released under CC BY-NC 4.0 license and can be downloaded at `https://amazon-berkeley-objects.s3.amazonaws.com/index.html`.

To facilitate future research, we benchmark the performance of various methods on three computer vision tasks that can benefit from more realistic 3D datasets: (i) single-view shape reconstruction, where we measure the domain gap for networks trained on synthetic objects, (ii) material estimation, where we introduce a baseline for spatially-varying BRDF from single- and multi-view images of complex real world objects, and (iii) image-based multi-view object retrieval, where we leverage the 3D nature of ABO to evaluate the robustness of deep metric learning algorithms to object viewpoint and scenes.

5.2 Background

3D Object Datasets ShapeNet [17] is a large-scale database of synthetic 3D CAD models commonly used for training single- and multi-view reconstruction models. IKEA Objects [96] and Pix3D [144] are image collections with 2D-3D alignment between CAD models and real images, however these images are limited to objects for which there is an exact CAD model

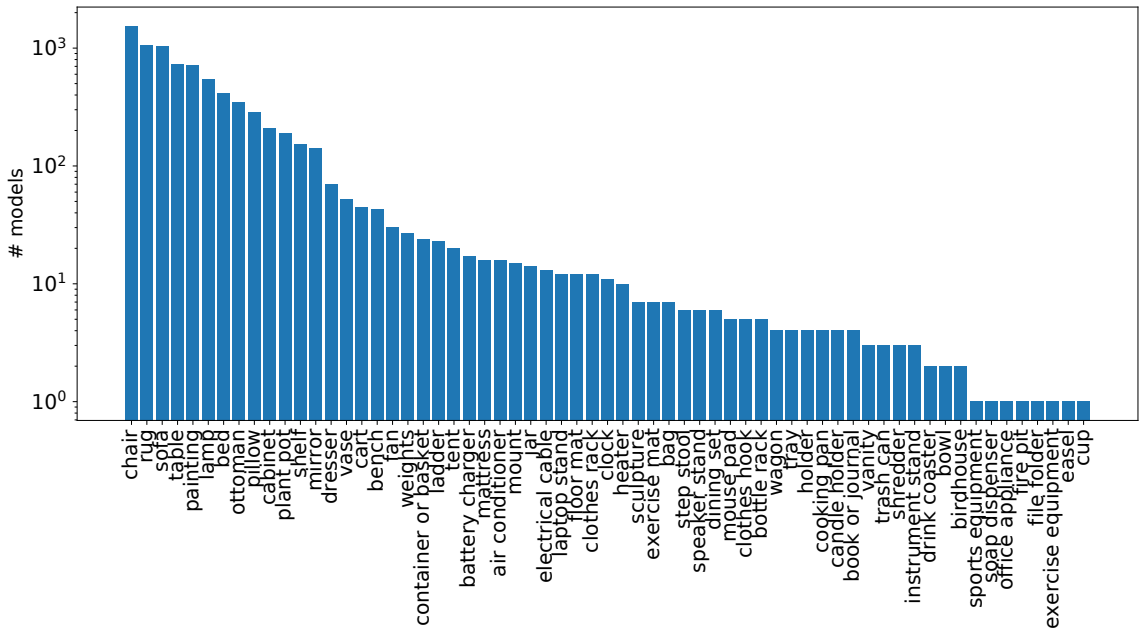

Figure 5.2: **3D model categories.** Each category is also mapped to a synset in the WordNet hierarchy. Note the y-axis is in log scale.

match. Similarly, Pascal3D+ [165] and ObjectNet3D [166] provide 2D-3D alignment for images and provide more instances and categories, however the 3D annotations are only approximate matches. The Object Scans dataset [20] and Objectron [2] are both video datasets that have the camera operator walk around various objects, but are limited in the number of categories represented. CO3D [125] also offers videos of common objects from 50 different categories, however they do not provide full 3D mesh reconstructions.

Existing 3D datasets typically assume very simplistic texture models that are not physically realistic. To improve on this, PhotoShapes [119] augmented ShapeNet CAD models by automatically mapping spatially varying (SV-) bidirectional reflectance distribution functions (BRDFs) to meshes, yet the dataset consists only of chairs. The works in [30, 39] provide high-quality SV-BRDF maps, but only for planar surfaces. The dataset used in [71] contains only homogenous BRDFs for various objects. [93] and [9] introduce datasets containing full SV-BRDFs, however their models are procedurally generated shapes that do not correspond to real objects. In contrast, ABO provides shapes and SV-BRDFs created by professional artists for real-life objects that can be directly used for photorealistic rendering.

Table 5.1 compares the 3D subset of ABO with other commonly used 3D datasets in terms of size (number of objects and classes) and properties such as the presence of real images, full 3D meshes and physically-based rendering (PBR) materials. ABO is the only dataset that contains all of these properties and is much more diverse in number of categories than existing 3D datasets.

Figure 5.3: **Posed 3D models in catalog images.** We use instance masks to automatically generate 6-DOF pose annotations for the dataset.

3D Shape Reconstruction Recent methods for single-view 3D reconstruction differ mainly in the type of supervision and 3D representation used, whether it be voxels, point clouds, meshes, or implicit functions. Methods that require full shape supervision in the single-view [34, 179, 144, 103, 46] and multi-view [69, 22, 168] case are often trained using ShapeNet. There are other approaches that use more natural forms of multi-view supervision such as images, depth maps, and silhouettes [170, 162, 69, 153], with known cameras. Of course, multi-view 3D reconstruction has long been studied with classical computer vision techniques [54] like multi-view stereo and visual hull reconstruction. Learning-based methods are trained typically in a category-specific way and evaluated on new instances from the same category. Out of the works mentioned, only [179] claims to be category-agnostic. In this work we are interested in how well these ShapeNet-trained networks [22, 179, 103, 46] generalize to more realistic objects.

Material Estimation Several works have focused on modeling object appearance from a single image, however realistic datasets available for this task are relatively scarce and small in size. [89] use two networks to estimate a homogeneous BRDF and an SV-BRDF of a flat surface from a single image, using a self-augmentation scheme to alleviate the need for a large training set. However, their work is limited to a specific family of materials, and each separate material requires another trained network. [173] extend the idea of self-augmentation to train with unlabeled data, but their work is limited to the same constraints. [31] use a modified U-Net and rendering loss to predict the SV-BRDFs of flash-lit photographs consisting of only a flat surface. To enable prediction for arbitrary shapes, [93] propose a cascaded CNN

architecture with a single encoder and separate decoder for each SV-BRDF parameter. While the method achieves good results on semi-uncontrolled lighting environments, it requires using the intermediate bounces of global illumination rendering as supervision.

More recent works have turned towards using multiple images to improve SV-BRDF estimation, but still only with simplistic object geometries. For instance, [30] and [39] use multiple input images with a flash lit light source, but only for a single planar surface. [9] and [12] both use procedurally generated shapes to estimate SV-BRDFs from multi-view images. ABO addresses the lack of sufficient realistic data for the material estimation, and in this work we propose a simple baseline method that can estimate materials from single or multi-view images of complex, real-world shapes.

2D/3D Image Retrieval Learning to represent 3D shapes and natural images of products in a single embedding space has been tackled by [91]. They consider various relevant tasks, including cross-view image retrieval, shape-based image retrieval and image-based shape retrieval, but all are inherently constrained by the limitations of ShapeNet [17] (cross-view image retrieval is only considered for chairs and cars). [79] introduced 3D object representations for fine-grained recognition and a dataset of cars with real-world 2D imagery (CARS-196), which is now widely used for deep metric learning (DML) evaluation. Likewise, other datasets for DML focus on instances/fine categories of few object types, such as birds [155], clothes [100], or a few object categories [115].

Due to the limited diversity and the similar nature of query and target images in existing retrieval benchmarks, the performance of state-of-the-art DML algorithms are near saturation. Moreover, since these datasets come with little structure, the opportunities to analyze failure cases and improve algorithms are limited. Motivated by this, we derive a challenging large-scale benchmark dataset from ABO with hundreds of diverse categories and a proper validation set. We also leverage the 3D nature of ABO to measure and improve the robustness of representations with respect to changes in viewpoint and scene. A comparison of ABO and existing benchmarks for DML can be found in Table 5.4.

5.3 The ABO Dataset

Dataset Properties The ABO dataset originates from worldwide product listings, metadata, images and 3D models provided by Amazon.com. This data consists of 147,702 listings of products from 576 product types sold by various Amazon-owned stores and websites (e.g. Amazon, PrimeNow, Whole Foods). Each listing is identified by an item ID and is provided with structured metadata corresponding to information that is publicly available on the listing's main webpage (such as product type, material, color, and dimensions) as well as the media available for that product. This includes $398,212$ high-resolution catalog images, and, when available, the turntable images that are used for the "360° View" feature that shows the product imaged at 5° or 15° azimuth intervals ($8,222$ products).

3D Models ABO also includes $7,953$ artist-created high-quality 3D models in glTF 2.0

format. The 3D models are oriented in a canonical coordinate system where the "front" (when well defined) of all objects are aligned and each have a scale corresponding to real world units. To enable these meshes to easily be used for comparison with existing methods trained on 3D datasets such as ShapeNet, we have collected category annotations for each 3D model and mapped them to noun synsets under the WordNet [105] taxonomy. Figure 5.2 shows a histogram of the 3D model categories.

Catalog Image Pose Annotations We additionally provide 6-DOF pose annotations for $6,334$ of the catalog images. To achieve this, we develop a fully automatic pipeline for pose estimation based on the knowledge of the 3D model in the image, off-the-shelf instance masks [56, 74], and differentiable rendering. For each mask $\mathbf{M}$, we estimate $\mathbf{R} \in SO(3)$ and $\mathbf{T} \in \mathbb{R}^3$ such that the following silhouette loss is minimized

$$\mathbf{R}^*, \mathbf{T}^* = \underset{\mathbf{R},\mathbf{T}}{\operatorname{argmin}} \|DR(\mathbf{R}, \mathbf{T}) - \mathbf{M}\|$$

where $DR(\cdot)$ is a differentiable renderer implemented in PyTorch3D [124]. Examples of results from this approach can be found in Figure 5.3. Unlike previous approaches to CAD-to-image alignment [166, 144] that use human annotators in-the-loop to provide pose or correspondences, our approach is fully automatic except for a final human verification step.

Material Estimation Dataset To perform material estimation from images, we use the Disney [16] base color, metallic, roughness parameterization given in glTF 2.0 specification [50]. We render 512x512 images from 91 camera positions along an upper icosphere of the object with a 60° field-of-view using Blender's [26] Cycles path-tracer. To ensure diverse realistic lighting conditions, we illuminate the scene using 3 random environment maps out of 108 indoor HDRIs [48]. For these rendered images, we generate the corresponding ground truth base color, metallicness, roughness, and normal maps along with the object depth map and segmentation mask. The resulting dataset consists of 2.1 million rendered images and corresponding camera intrinsics and extrinsics.

5.4 Experiments

Evaluating Single-View 3D Reconstruction

As existing methods are largely trained in a fully supervised manner using ShapeNet [17], we are interested in how well they will transfer to more real-world objects. To measure how well these models transfer to real object instances, we evaluate the performance of a variety of these methods on objects from ABO. Specifically we evaluate 3D-R2N2 [22], GenRe [179], Occupancy Networks [103], and Mesh R-CNN [46] pre-trained on ShapeNet. We selected these methods because they capture some of the top-performing single-view 3D reconstruction methods from the past few years and are varied in the type of 3D representation that they use (voxels in [22], spherical maps in [179], implicit functions in [103], and meshes

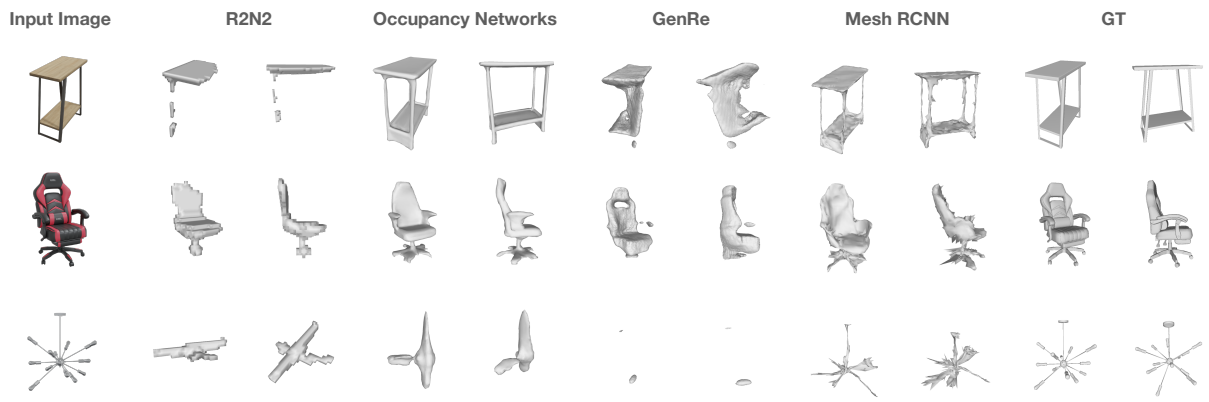

Figure 5.4: **Qualitative 3D reconstruction results for R2N2, Occupancy Networks, GenRe, and Mesh-RCNN on ABO.** All methods are pre-trained on ShapeNet and show a decrease in performance on objects from ABO.

in [46]) and the coordinate system used (canonical vs. view-space). While all the models we consider are pre-trained on ShapeNet, GenRe trains on a different set of classes and takes as input a silhouette mask at train and test time.

To study this question (irrespective of the question of cross-category generalization), we consider only the subset of ABO models objects that fall into ShapeNet training categories. Out of the 63 categories in ABO with 3D models, we consider 6 classes that intersect with commonly used ShapeNet classes, capturing 4,170 of the 7,953 3D models. Some common ShapeNet classes, such as "airplane", have no matching ABO category; similarly, some categories in ABO like "air conditioner" and "weights" do not map well to ShapeNet classes.

Unlike models in ShapeNet, the 3D models in ABO can be photo-realistically rendered due to their detailed textures and non-lambertian BRDFs. We render 30 viewpoints of each mesh using Blender [26], each with a 40° field-of-view and such that the entire object is visible. Camera azimuth and elevation are sampled uniformly on the surface of a unit sphere with a $-10°$ lower limit on elevations to avoid uncommon bottom views. We used a publicly available HDRI [48] for scene lighting.

GenRe and Mesh-RCNN make their predictions in "view-space" (i.e. pose aligned to the image view), whereas R2N2 and Occupancy Networks perform predictions in canonical space (predictions are made in the same category-specific, canonical pose despite the pose of the object in an image). For each method we evaluate Chamfer Distance and Absolute Normal Consistency and largely follow the evaluation protocol of [46].

Results A quantitative comparison of the four methods we considered on ABO objects can be found in Table 5.2. We also re-evaluated each method's predictions on the ShapeNet test set from R2N2 [22] with our evaluation protocol and report those metrics. We observe that Mesh R-CNN [46] outperforms all other methods across the board on both ABO and ShapeNet in terms of Chamfer Distance, whereas Occupancy Networks performs the best in

	Chamfer Distance (↓)						Absolute Normal Consistency (↑)					
	bench	chair	couch	cabinet	lamp	table	bench	chair	couch	cabinet	lamp	table
[22]	2.46/0.85	1.46/0.77	1.15/0.59	1.88/0.25	3.79/2.02	2.83/0.66	0.51/0.55	0.59/0.61	0.57/0.62	0.53/0.67	0.51/0.54	0.51/0.65
[103]	1.72/0.51	0.72/0.39	0.86/0.30	0.80/0.23	2.53/1.66	1.79/0.41	0.66/0.68	0.67/0.76	0.70/0.77	0.71/0.77	0.65/0.69	0.67/0.78
[179]	1.54/2.86	0.89/0.79	1.08/2.18	1.40/2.03	3.72/2.47	2.26/2.37	0.63/0.56	0.69/0.67	0.66/0.60	0.62/0.59	0.59/0.57	0.61/0.59
[46]	1.05/0.09	0.78/0.13	0.45/0.10	0.80/0.11	1.97/0.24	1.15/0.12	0.62/0.65	0.62/0.70	0.62/0.72	0.65/0.74	0.57/0.66	0.62/0.74

Table 5.2: **Single-view 3D reconstruction generalization from ShapeNet to ABO.** Chamfer distance and absolute normal consistency of predictions made on ABO objects from common ShapeNet classes. We report the same metrics for ShapeNet objects (denoted in gray), following the same evaluation protocol. All methods, with the exception of GenRe, are trained on all of the ShapeNet categories listed.

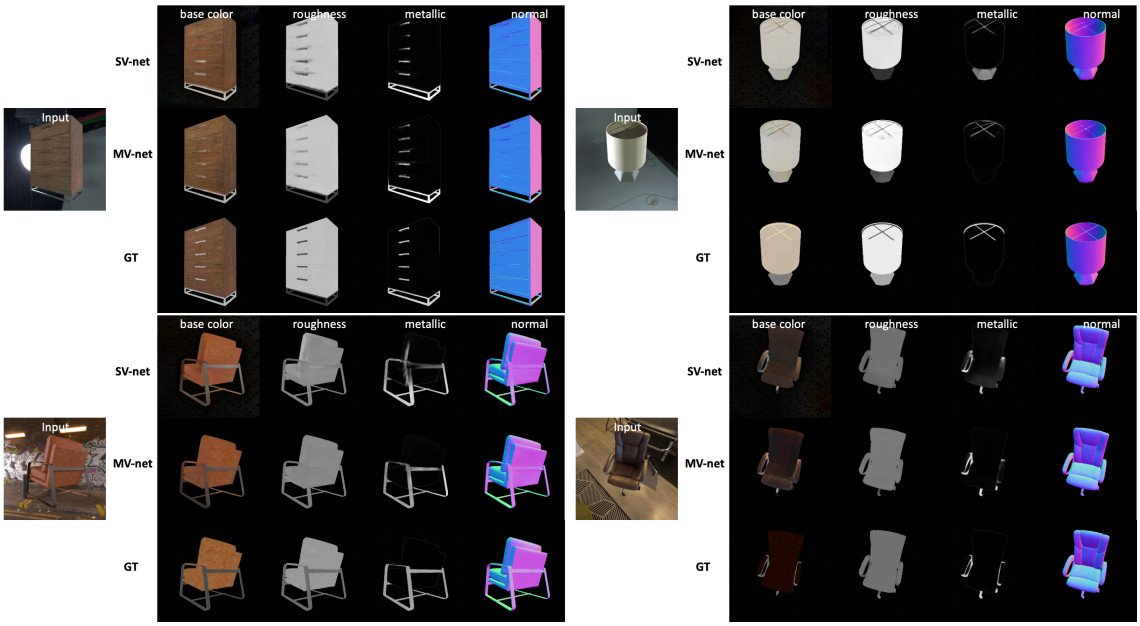

Figure 5.5: **Qualitative material estimation results for single-view (SV-net) and multi-view (MV-net) networks.** We show estimated SV-BRDF properties (base color, roughness, metallicness, surface normals) for each input view of an object compared to the ground truth.

terms of Absolute Normal Consistency. As can be seen, there is a large performance gap between all ShapeNet and ABO predictions. This suggests that shapes and textures from ABO, while derived from the same categories but from the real world, are out of distribution and more challenging for the models trained on ShapeNet. Further, we notice that the *lamp* category has a particularly large performance drop from ShapeNet to ABO. Qualitative results suggest that this is likely due to the difficulty in reconstructing thin structures. We highlight some qualitative results in Figure 5.4, including one particularly challenging lamp instance.

Material Prediction

While these existing material estimation approaches show good performance on their respective datasets, the datasets that they use are quite simplistic (planar objects [30]) or not realistic (procedurally generated shape and material mappings [9]). To date, there are not many available datasets tailored for the material prediction task. Most publicly available datasets with large collections of 3D objects [17, 20, 38] don't contain physically-accurate reflectance parameters that can be used for physically-based rendering to generate photorealistic images. Datasets like PhotoShape [119] do contain such parameters but are limited to a single category.

In contrast, the realistic 3D models in ABO are artist-created and have highly varied

	SV-net	MV-net (no proj.)	MV-net
Base Color (↓)	0.129	0.132	**0.127**
Roughness (↓)	0.163	0.155	**0.129**
Metallicness (↓)	0.170	0.167	**0.162**
Normals (↑)	0.970	0.949	**0.976**
Render (↓)	0.096	0.090	**0.086**

Table 5.3: **ABO material estimation results for the single-view, multi-view, and multi-view network without projection (MV-net no proj.) ablation.** Base color, roughness, metallicness and rendering loss are measured using RMSE (lower is better) - normal similarity is measured using cosine similarity (higher is better).

shapes and SV-BRDFs. We leverage this unique property to derive a benchmark for material prediction with large amounts of photorealistic synthetic data. We also present a simple baseline approach for both single- and multi-view material estimation of complex geometries.

Dataset Curation We use the material estimation dataset outlined in Section 5.3, but filter out objects with transparencies, resulting in 7,679 models. We split the 3D models into a non-overlapping train/test set of 6,897 and 782 models, respectively. To test generalization to new lighting conditions, we reserve 10 out of 108 HDRI environment maps for the test set only.

Method To evaluate single-view and multi-view material prediction and establish a baseline approach, we use a U-Net-based model with a ResNet-34 backbone to estimate SV-BRDFs from a single viewpoint. The U-Net has a common encoder that takes an RGB image as input and has a multi-head decoder to output each component of the SV-BRDF separately. Inspired by recent networks in [9, 30], we align images from multiple viewpoints by projection using the depth data and bundle original image/projected image pairs as input data to enable an analogous approach for the multi-view network. We reuse the single-view architecture for the multi-view network and use global max pooling to handle an arbitrary number of input images. Ground truth material maps are used for direct supervision. Similar to [31], we utilize a differential rendering layer to render the flash illuminated ground truth and compare it to similarly rendered images from our predictions to better regularize network and guide the training process.

Our model takes as input 256x256 rendered images. For training, we randomly subsample 40 views on the icosphere for each object. For the multi-view network, for each reference view we select its immediate 4 adjacent views as neighboring views. We use mean squared error as loss function for base color, roughness, metallicness, normal and render losses. Each network is trained using the AdamW optimizer [101] with a learning rate of 1e-3 and weight decay of 1e-4, and trained for 17 epochs.

Results Results for the single-view network (SV-net) and multi-view network (MV-net) can

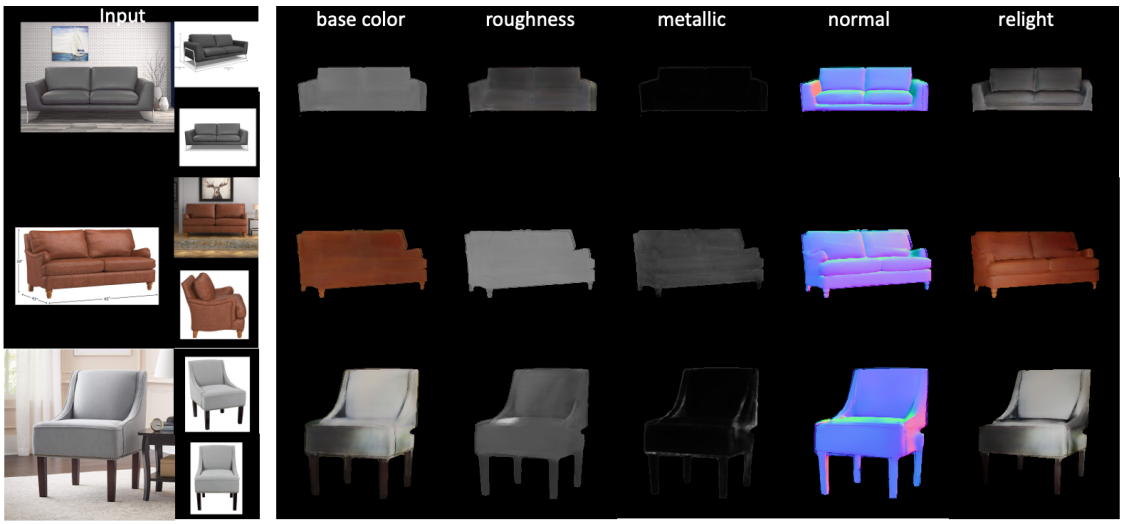

Figure 5.6: **Qualitative multi-view material estimation results on real catalog images.** Each of the multiple views is aligned to the reference view using the catalog image pose annotations.

be found in Table 5.3. The multi-view network has better performance compared to single-view network in terms of the base color, roughness, metallicness, and normal prediction tasks. The multi-view network is especially better at predicting properties that effect view-dependent specular components like roughness and metallicness.

We also run an ablation study on our multi-view network without using 3D structure to align neighboring views to reference view (denoted as *MV-net: no projection*). First, we observe that even without 3D structure-based alignment, the network still outperforms the single-view network on roughness and metallic predictions. Comparing to the multi-view network, which uses 3D structure-based alignment, we can see structure information leads to better performance for all parameters. We show some qualitative results from the test set in Figure 5.5. As a focus of ABO is enabling real-world transfer, we also test our multi-view network on catalog images of objects from the test set using the pose annotations gathered by the methodology in Section 5.3. We use the inferred material parameters to relight the object. The results can be found in Figure 5.6. Despite the domain gap in lighting and background, our network trained on synthetic images makes reasonable predictions on the catalog images. In one case, the network fails to accurately infer the true base color, likely due to the presence of self-shadow.

Multi-View Cross-Domain Object Retrieval

Merging the available catalog images and 3D models in ABO, we derive a novel benchmark for object retrieval with the unique ability to measure the robustness of algorithms with respect to viewpoint changes. Specifically, we leverage the renderings described in Section 5.3,

Benchmark	Domain	Classes	Instances			Images				Structure	Recall@1	
			train	val	test	train	val	test-target	test-query			
CUB-200-2011	Birds	200	-	-	-	5994	0	-	5794	15 parts	79.2%	[68]
Cars-196	Cars	196	-	-	-	8144	0	-	8041	-	94.8%	[68]
In-Shop	Clothes	25	3997	0	3985	25882	0	12612	14218	Landmarks, poses, masks	92.6%	[72]
SOP	Ebay	12	11318	0	11316	59551	0	-	60502	-	84.2%	[68]
ABO (MVR)	Amazon	562	49066	854	836	298840	26235	4313	23328	Subset with 3D models	30.0%	

Table 5.4: **Common image retrieval benchmarks for deep metric learning and their statistics.** Our proposed multi-view retrieval (MVR) benchmark based on ABO is significantly larger, more diverse and challenging than existing benchmarks, and exploits 3D models.

with known azimuth and elevation, to provide more diverse views and scenes for training deep metric learning (DML) algorithms. We also use these renderings to and evaluate the retrieval performance with respect to a large gallery of catalog images from ABO. This new benchmark is very challenging because the rendered images have complex and cluttered indoor backgrounds (compared to the cleaner catalog images) and display products with viewpoints that are not typically present in the catalog images. These two sources of images are indeed two separate image domains, making the test scenario a multi-view cross-domain retrieval task.

Dataset Curation For this task, we used product type annotations to focus only on rigid objects, removing items such as garments, home linens, and some accessories (cellphone accessories, animal harnesses, cables, etc.). As the set of products beyond just those with 3D models are likely to contain near-duplicates (i.e. different sizes of the same shoe), we then applied a hierarchical Union-Find algorithm for near-duplicate detection and product grouping, based on shared imagery as a heuristic. We considered near-duplicates as correct matches of a single instance and thus assigned a unique *instance id* to all near-duplicate listings. Product groups are sets of such instances that are from product lines that may share design details, materials, patterns and thus may have common images (close-up detail of fabric, fact sheet image, ...). Consequently, we ensured that all instances in a group are assigned to the same data split (train, val or test). The val and test sets contain only instances with 3D models and, while their catalog images compose their respective target set (val-target and test-target), we use rendered images as queries (val-query and test-query). Table 5.4 summarizes the statistics of this new retrieval benchmark in comparison to the most common benchmark datasets for DML in the literature.

The validation performance, which is used to optimize model hyperparameters (HPO), is measured using val-query queries against the val-target set. After HPO, the test performance is measured using the test-query images against the union of test-target *and* the $187,912$ catalog images of the train set. We used standard instance-level retrieval metrics (recall at 1, 2, 4 and 8) and report them as an aggregate over all test queries and also as a function of azimuth and elevation. We split the azimuth angle into 30 bins of 12° each and elevation in 3 bins ($< 21^\circ$, $[21^\circ - 50^\circ]$ and $> 50^\circ$), such that each bin had roughly equal amounts of images.

Method To compare the performance of state-of-the-art DML methods on our multi-view cross-domain retrieval benchmark, we use PyTorch Metric Learning implementations that cover the main approaches to DML: NormSoftmax [178] (classification-based), ProxyNCA [111] (proxy-based) and Contrastive, TripletMargin, NTXent [18] and Multi-similarity [158] (tuple-based). We leveraged the Powerful Benchmarker framework to run fair and controlled comparisons as in [112], including Bayesian HPO.

We opted for ResNet-50 [55] as the backbone, projected it to 128D after a LayerNorm [3] layer, did not freeze the batch-norm parameters and added an image padding transformation to obtain undistorted square images before resizing to 256x256. We used batches of 256 samples with 4 samples per class, except for NormSoftmax and ProxyNCA where we obtained

	Rendered images				Catalog
Recall@k (%)	$k=1$	$k=2$	$k=4$	$k=8$	$k=1$
Pre-trained	5.0	8.1	11.4	15.3	18.0
Constrastive	28.6	38.3	48.9	59.1	**39.7**
Multi-similarity	23.1	32.2	41.9	52.1	38.0
NormSoftmax	**30.0**	**40.3**	**50.2**	60.0	35.5
NTXent	23.9	33.0	42.6	52.0	37.5
ProxyNCA	29.4	39.5	50.0	**60.1**	35.6
TripletMargin	22.1	31.1	41.3	51.9	36.9

Table 5.5: **Test performance of state-of-the-art deep metric learning methods on the ABO retrieval benchmark.** Retrieving products from rendered images highlights performance gaps that are not as apparent when using catalog images.

better results with a batch size of 32 and 1 sample per class. After HPO, we trained all losses for 1000 epochs and chose the best epoch based on the validation Recall@1 metric, computing it only every other epoch.

Importantly, whereas catalog and rendered images in the training set are balanced (188K vs 111K), classes with and without renderings are not (4K vs. 45K). Balancing them in each batch proved necessary to obtain good performance: not only do we want to exploit the novel viewpoints and scenes provided by the renderings to improve the retrieval performance, but there are otherwise simply not sufficiently many negative pairs of rendered images being sampled.

Results As shown in Table 5.5, the ResNet-50 baseline trained on ImageNet largely fails at the task (5% recall@1). This confirms the challenging nature of our novel benchmark. DML is thus key to obtain significant improvements. In our experiments, NormSoftmax, ProxyNCA and Contrastive performed better ($\approx 29\%$) than the Multi-similarity, NTXent or TripletMargin losses ($\approx 23\%$), a gap which was not apparent in other datasets, and is not as large when using cleaner catalog images as queries. Moreover, it is worth noting that the overall performance is significantly lower than for existing common benchmarks (see Table 5.4). This confirms their likely saturation [112], the value in new and more challenging retrieval tasks, and the need for novel metric learning approaches to handle the large scale, the unique challenges and the unique properties of our new benchmark.

Further, the azimuth (θ) and elevation (φ) angles available for rendered test queries allow us to measure how performance degrades as these parameters diverge from typical product viewpoints in ABO's catalog images. Figure 5.7 highlights two main regimes for both azimuth and elevation: azimuths beyond $|\theta| = 75^\circ$ and elevations above $\varphi = 50^\circ$ are significantly more challenging to match, consistently for all approaches. Closing this gap is an interesting direction of future research on DML for multi-view object retrieval. For one, the current losses do not explicitly model the geometric information in training data.

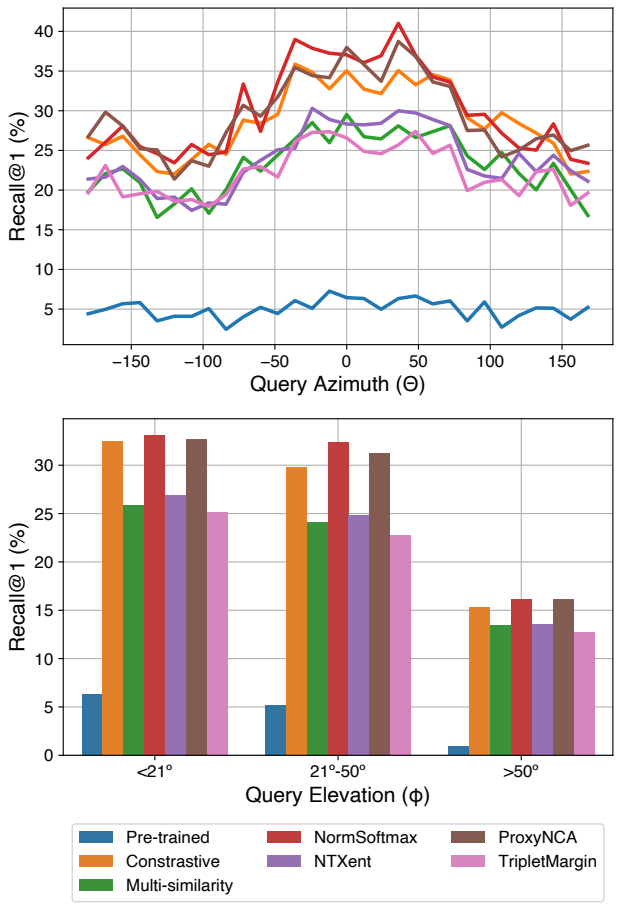

Figure 5.7: **Recall@1 as a function of the azimuth and elevation of the product view.** For all methods, retrieval performance degrades rapidly beyond azimuth $|\theta| > 75°$ and elevation $\varphi > 50°$.

5.5 Discussion

In this work we introduced ABO, a new dataset to help bridge the gap between real and synthetic 3D worlds. We demonstrated that the set of real-world derived 3D models in ABO are a challenging test set for ShapeNet-trained 3D reconstruction approaches, and that both view- and canonical-space methods do not generalize well to ABO meshes despite sampling them from the same distribution of training classes. We also trained both single-view and multi-view networks for SV-BRDF material estimation of complex, real-world geometries - a task that is uniquely enabled by the nature of our 3D dataset. We found that incorporating multiple views leads to more accurate disentanglement of SV-BRDF properties. Finally, joining the larger set of products images with synthetic renders from ABO 3D models, we proposed a challenging multi-view retrieval task that alleviates some of the limitations in

diversity and structure of existing datasets, which are close to performance saturation. The 3D models in ABO allowed us to exploit novel viewpoints and scenes during training and benchmark the performance of deep metric learning algorithms with respect to the azimuth and elevation of query images.

While not considered in this work, the large amounts of text annotations (product descriptions and keywords) and non-rigid products (apparel, home linens) enable a wide array of possible language and vision tasks, such as predicting styles, patterns, captions or keywords from product images. Furthermore, the 3D objects in ABO correspond to items that naturally occur in a home, and have associated object weight and dimensions. This can benefit robotics research and support simulations of manipulation and navigation.

Chapter 6

Inferring How Objects Can Articulate

In this chapter, we present a neural network approach to transfer the motion from a single image of an articulated object to a *rest-state* (i.e., unarticulated) 3D model. Our network learns to predict the object's pose, part segmentation, and corresponding motion parameters to reproduce the articulation shown in the input image. The network is composed of three distinct branches that take a shared joint image-shape embedding and is trained end-to-end. Unlike previous methods, our approach is independent of the topology of the object and can work with objects from arbitrary categories. Our method, trained with only synthetic data, can be used to automatically animate a mesh, infer motion from real images, and transfer articulation to functionally similar but geometrically distinct 3D models at test time.

6.1 Motivation

We live in a 3D world where interacting with objects in our environment is necessary to carry out daily activities. Therefore, the objects around us have been designed with the appropriate structures and part motions to afford various actions to both humans and robotic agents. Whether for general visual perception tasks such as functionality inference or robotic action planning, it is a valuable skill to be able to predict the part motions or articulations of everyday objects. Recently, there has been much interest in training deep neural networks to learn such predictions, especially in the context of embodied AI [164, 110]. However, the availability of 3D models with associated part articulations [157, 164] for training these methods is still limited.

Creating articulated 3D models relies on human effort which can be expensive in terms of time, cost and expertise. As a result, most 3D model datasets do not contain objects with any kind of functional part and motion annotations. For example while most of the 3D models in ABO represent objects that can be interacted with, such as desks with openable drawers or chairs that recline, the 3D models have all been constructed in their "rest" (i.e., unarticulated) poses that do not expose their real functionality. The same is true for the vast majority of 3D models in well-established repositories such as ShapeNet [17] (3M models) and

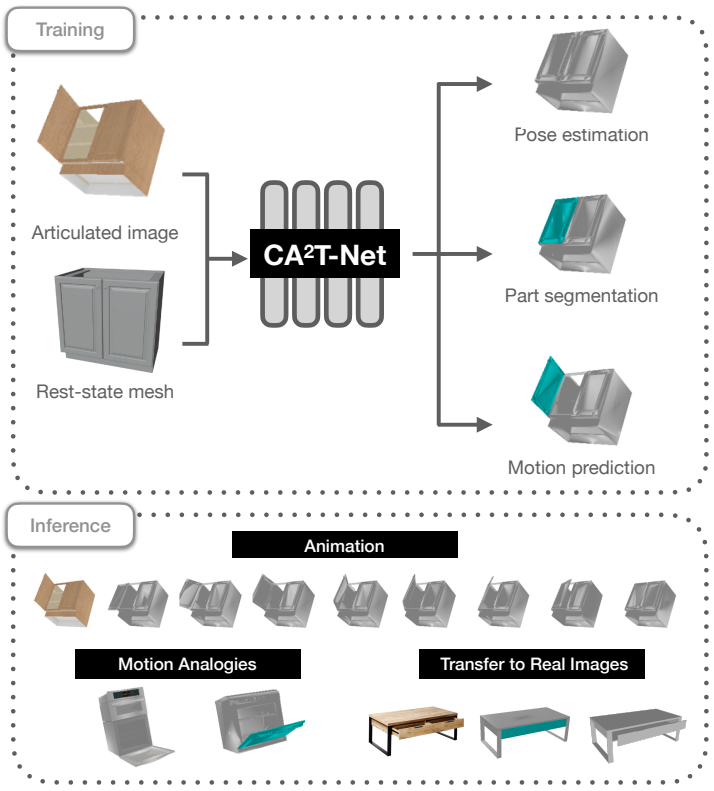

Figure 6.1: **Predicting pose, part, and motion annotations for synthetic meshes given an exemplar image.** At test time, we can perform 3D animation, transfer motion to functionally similar objects, and generalize to real {image, mesh} pairs.

PartNet [109] (27K models). The largest 3D datasets with part articulations, SAPIEN [164] and Shape2Motion [157], only contain 2K *manually* annotated synthetic models.

In this chapter, we introduce *single-view 3D articulation transfer*, a learning-based approach aimed at endowing the many existing 3D models, in particular, those of real products such as ABO, with part articulations. Specifically, given a single RGB image, I, showing how an object can be articulated, and a 3D mesh model M in its rest state, our goal is to infer the pose, motion parameters, and 3D part segmentation of M such that we can re-pose and transform it to match the articulation in the input image. Our method is structure-agnostic, as it makes no assumption on the topology of the input model. Unlike many existing methods for articulated pose estimation, our approach is also *category-agnostic*, meaning no further training is necessary to run our model on objects from novel categories and it can make predictions for objects from arbitrary categories. Further, the object represented in the input image I and the mesh M need not be exactly the same, but should simply share a functional similarity to allow a plausible articulation transfer. To our knowledge, our work is the first to pose the problem of single-view articulation transfer onto a target 3D object. As indicated

in Table 6.1, the most relevant works [175, 157, 90, 67, 171] differ from our problem setting in one way or another.

We call our network CA^{2}T-Net for category-agnostic articulation transfer, and demonstrate that our method has superior performance in single-view 3D articulation transfer compared to baseline methods and the current state-of-the-art in terms of motion prediction. Our approach takes as input an articulated image and rest-state 3D point cloud and uses three branches to predict global pose, part segmentation, and 3D motion parameters. The network is trained end-to-end using synthetic data. We also show three downstream applications of our trained approach: automatic mesh animation, generalization to functionally similar objects, and transfer to real objects and images in ABO.

Our work is a first step towards replacing manual part and motion annotations [157, 164] with a fully automated inference of part articulations and motion parameters of arbitrary 3D shapes. With the wide availability of images depicting object articulations (e.g., catalog images found in the product metadata within ABO), our image-to-3D transfer presents a promising step towards articulated 3D shape creation at scale.

6.2 Background

Part Segmentation In computer vision, most research on 2D or 3D part segmentation focuses on semantic [174, 109] or functional [73, 61, 157, 67] meanings of the parts. Jiang et al. [67] extended 2D semantic segmentation networks (e.g. Mask R-CNN) to predict the location of openable parts from a single-view image. Gelfand et al. [45] defined a part segmentation as a kinematic surface composed of points that undergo the same type of motion. Other work has relied on a category-level canonical container or a fixed kinematic structure. For example, follow-up work [90, 171] learned a fully supervised canonicalization to map instance observations of different position, orientation, scale and articulation into a canonical container. Xu et al. [169] estimated the part segmentation from a depth image in a known category of a fixed kinematic chain. In contrast, our approach requires neither semantic annotation nor category-level functional information or kinematic structure. Further, our work is not about learning semantic or functional part segmentation, but rather image-guided part segmentation. Our method finds the correspondence between an input image and a 3D shape under different articulation configurations.

Articulated Pose Estimation Articulated pose estimation from a 2D image is another focus of our work. Recently, Wang et al. [156] introduced the Normalized Object Coordinate Space (NOCS), allowing category-level rigid pose estimation. Follow-ups [90, 99] have used this representation to perform articulated pose estimation. For example, Li et al. [90] proposed a part-level canonical reference frame and performed part pose fitting using RANSAC. Liu et al. [99] extended NOCS to Real-World Articulation NOCS which focuses on pose estimation for articulated objects with varied kinematic structures in real-world settings. Kulkarni et al., [81] introduced an approach for articulating a template mesh given

Method	Task	Input	Pose	3D Part Seg.	Cat-Specific
DeepPartInduction [175]	Transfer articulation from PC → PC	2 PCs	×	✓	No
Shape2Motion [157]	Predict motion for all moveable parts	PC	×	✓	No
RPM-Net [171]	Predict motion for all moveable parts	PC	×	✓	No
ANCSH [90]	Articulated pose estimation	PC	✓	✓	Yes
NASAM [160]	Reconstruct and animate objects	RGBs + pose	×	✓	Yes
OPD [67]	Predict motion for all openable parts	RGB	✓	×	No
CA^{2}T-Net (ours)	Transfer articulation from image → mesh	RGB, PC	✓	✓	No

Table 6.1: **CA^{2}T-Net addresses the problem of *single-view articulation transfer*, from an image to a mesh.** Existing work targets related problems, but differs in terms of task and inputs of the model.

a target image segmentation mask from a category-specific image collection, but requires hand-annotated part segmentations. Other work [171, 161, 172] relies on a series of frames displaying an object actively articulating. In contrast, our method does not require learning the canonical representation with a fixed kinematic structure for each category. Without the need of sequences of images of an object's different articulation states as inputs, we require only an articulated single-view image and its unarticulated 3D model to perform articulated pose estimation.

Motion Estimation & Transfer In 3D understanding, considerable work has focused on the problem of motion prediction. Methods largely differ in the types of input(s) used for motion estimation and transfer. Initial work [45, 107, 171] utilized only the geometry of an object to infer part segmentation and motion. Li et al. [88] focused on predicting part mobility of an object from a sequence of scans displaying the the dynamic motion of articulated models. More recently, motion estimation has been performed from depth images [1, 90]. Mo et al., [110] learned to make per-point motion predictions from a single RGB or depth image by interacting with articulated objects in simulation. Other approaches use a sequence of images or video clips to infer 3D part movement [62, 175, 66]. Wei et al., [160] trained an implicit representation of the geometry, appearance, and motion of an object in a self-supervised manner from image observations, however their approach is limited to certain classes of articulated objects. Recently, Jiang et al., [67] introduced a method for estimating 3D motion parameters from single image. Their approach aims to predict motion parameters for all "openable" parts, rather than focusing on transferring arbitrary motion from one object to another.

Our work makes no assumptions on the input category and can both estimate and transfer 3D motion. We formulate the problem as learning the articulated segmentation of the input geometry, estimating the pose and predicting motion parameters of the moveable part (driven by the input image). Unlike many existing approaches, our work is both category-agnostic and structure agnostic.

6.3 Method

We propose a model for the *single-view articulation transfer* task. Given an articulated RGB image and corresponding 3D model, our method infers the pose, part segmentation and motion parameters that can be used to deform the mesh to match the input image.

Architecture

Our architecture is composed of 3 distinct branches: a pose prediction branch, a motion prediction branch, and a 3D part segmentation branch. Our model takes two inputs: an image $I \in \mathbb{R}^{h \times w \times 3}$ and point cloud shape $S \in \mathbb{R}^{n \times 3}$, where n is the number of points in a point cloud sampled from the normalized input mesh. Image features, $\mathbf{h} = f_\theta(I)$, are

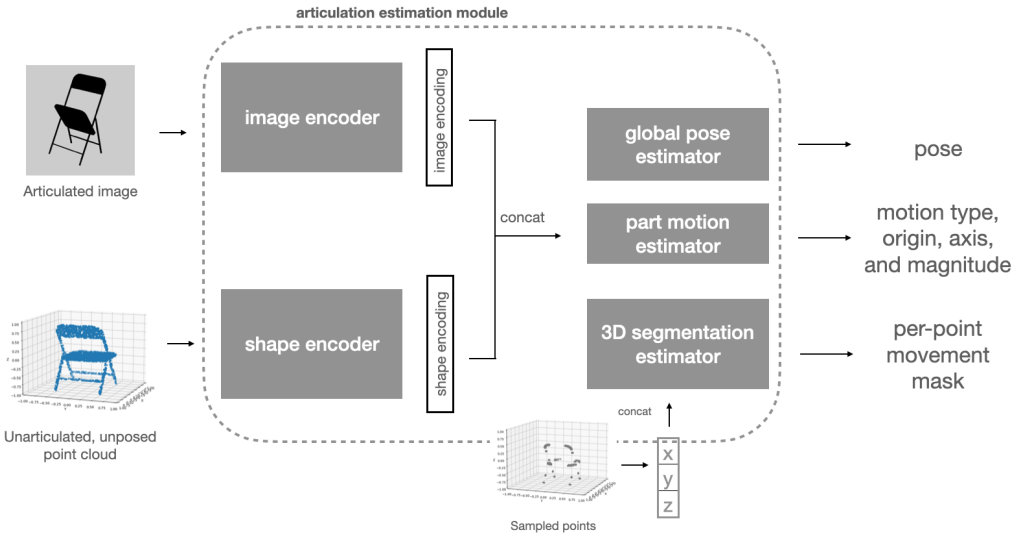

Figure 6.2: **Diagram of training architecture.** CA^{2}T-Net takes in an RGB image and 3D model and predicts the global pose of the object in the image, segmentation of the part that is articulated in the image, as well as part-specific motion parameters. Each motion attribute is predicted from a spearate head.

extracted with a ResNet-18 [146], and shape features, $\mathbf{s} = g_\theta(S)$, are extracted with a PointNet-like [122] architecture, without input and feature transforms for pose-invariance. Both the image and shape encoder are trained from scratch. The shape and image features are concatenated and passed through a set of fully connected layers resulting in a joint image and shape embedding, $\mathbf{z} \in \mathbb{R}^{b\times 512}$. Each of the three branches take $\mathbf{z}$ as input. An overview of the full architecture can be found in Figure 6.2.

Pose Prediction The pose prediction branch of our architecture is inspired by [167], a method that also takes as input an RGB image (however unarticulated) and a 3D shape, and predicts the pose of the 3D model in the image. As in [167] we predict pose in terms of azimuth, elevation, and in-plane rotation. The network first classifies the angle bin and the regresses an offset. Let $\hat{l}_{i,j}$ be the predicted bin for the j-th Euler angle (azimuth, elevation, in-plane rotation) of the i-th datapoint, and $\hat{\delta}_{i,j}$ be the predicted relative offset within the bins. For bin prediction, we have 24 azimuth classes, 12 elevation classes, and 24 in-plane rotation classes. We train the model with a cross-entropy loss ($\mathcal{L}_{CE}$) for bin classification and Huber loss ($\mathcal{L}_{huber}$) for regression offsets.

$$\mathcal{L}_p = \sum_{i=1}^{N} \sum_{j\in\mathcal{E}} \mathcal{L}_{CE}(l_{i,j}, \hat{l}_{i,j}) + \mathcal{L}_{huber}(\hat{\delta}_{i,j}, \delta_{i,j}), \tag{6.1}$$

where $\mathcal{E} = \{\text{azimuth}, \text{elevation}, \text{in-plane}\}$ and $l_{i,j}$ and $\delta_{i,j}$ are the ground truth pose bins and offsets, accordingly.

Motion Estimation In this work we consider two distinct motion types, revolute (rotation) and prismatic (translation). We parameterize the revolute motion in terms of motion axis, origin, and magnitude, and prismatic in terms of motion axis and magnitude. Each motion attribute is predicted by a separate head, for origin, axis, magnitude, and class. Let $\mathcal{L}_{mc}$ be the classification loss on motion type, defined as:

$$\mathcal{L}_{mc} = \sum_{i=1}^{N} \mathcal{L}_{CE}(m_i, \hat{m}_i), \tag{6.2}$$

where m_i is the ground-truth motion class represented as a one-hot vector and $\hat{m}_i$ is the corresponding predicted class.

Motion axis $\hat{a} \in \mathbb{R}^3$, origin $\hat{o} \in \mathbb{R}^3$, and magnitude $\hat{z} \in \mathbb{R}$ are regressed and optimized using Huber loss:

$$\mathcal{L}_m = \sum_{i=1}^{N} \mathcal{L}_{huber}(a_i, \hat{a}_i) + \mathcal{L}_{huber}(o_i, \hat{o}_i) + \mathcal{L}_{huber}(z_i, \hat{z}_i). \tag{6.3}$$

We predict motion parameters in the world coordinate space (i.e., object's canonical coordinate frame). Motion origin and magnitude for prismatic motion is given in terms of normalized object size, and magnitude for revolute motion is given in terms of radians. As motion origin is not relevant for prismatic joints, we only apply the motion origin loss, $\mathcal{L}_{huber}(o_i, \hat{o}_i)$, in the cases of predicted revolute motion.

3D Part Segmentation Our 3D part segmentation branch takes in the transformed global shape and image features, $\mathbf{z}$, as well as a 3D point, $x \in \mathbb{R}^3$, and predicts whether or not that point (from the pointcloud) corresponds to the moveable part. Our architecture is inspired by DeepSDF [118] which also takes as input a set of features and a 3D point. After processing these inputs with 2 fully connected layers, the initial inputs are again concatenated with the intermediate features and further processed by 3 more fully connected layers. Let $\hat{p}$ be the predicted confidence of a certain point x's movability - we define the 3D part segmentation loss as:

$$\mathcal{L}_s = \sum_{i=1}^{N} \sum_{j=1}^{M} \mathcal{L}_{CE}(p_{i,j}, \hat{p}_{i,j}), \tag{6.4}$$

where i indexes over training datapoints and j over the M sampled 3D points from the input shape. For training we use $M = 1000$.

Full Training Objective Our full training loss is a weighted combination of losses applied to each of the three branches:

$$\mathcal{L} = \lambda_p \mathcal{L}_p + \lambda_{mc} \mathcal{L}_{mc} + \lambda_m \mathcal{L}_m + \lambda_s L_s. \tag{6.5}$$

The coefficients on the loss terms are $\lambda_p = 2$, $\lambda_{mc} = 1$, $\lambda_m = 8$, and $\lambda_s = 1$. We train the network using the Adam optimizer with a learning rate of 1e-3 and batch size of 128 for 50 epochs and decrease the learning rate to 1e-4 for the final 10 epochs.

Inference During inference, we randomly sample 100 points on each distinct part of the input model and compute the network's 3D part prediction for each point. We apply the network's predicted motion only to parts that have >50% points considered "moveable".

6.4 Evaluation and Results

We train and evaluate our method using images and ground-truth data generated from the SAPIEN PartNet-Mobility [164] dataset.

Dataset

The original SAPIEN PartNet-Mobility dataset consists of 2,346 CAD models from 46 categories with part segmentations and motion annotations. As we are interested in articulation transfer from a single-image, we focus on objects with large moveable parts such as the door or drawer of a cabinet, rather than the individual keys on a keyboard. To do so, we filter out CAD models that only contain moveable parts that consist of <5% of the overall surface area of the object, as well as objects with missing or incomplete motion data. After this filtering process, 1,717 objects remain.

From the remaining objects, we render 256 images per object using the SAPIEN Vulkan Renderer, each with a random pose, randomly sampled part, and randomly sampled motion parameters within the valid range. For simplicity, we only consider one moving part at a time. The pose distribution ranges from azimuth $\in [-90°, 90°]$, elevation $\in [-45°, 45°]$, and in-plane rotation $\in [-20°, 20°]$. Such ranges were chosen so that the articulated part on the 3D model is usually visible (most moveable parts are on the front of the object, rather than the back). For each image, we also randomly sample the camera's field-of-view $\in [20°, 60°]$. In total we render approximately 280K images for training, and 10K images for evaluation.

For each image we also write out the rest-state mesh, the ground truth pose of the object, 3D part segmentation (of the articulated part from the rest of the object), and motion parameters. The motion parameters consist of a motion type $\in \{revolute, prismatic\}$, axis, origin, and magnitude. In the case of revolute motion, the magnitude corresponds to a rotation amount, whereas it corresponds to translation distance for prismatic motion. From the rest-state mesh we sample a point cloud that is the input to our network. For training, we

		Motion					3D	
		Pose (↑)	Type (↑)	Axis (↓)	Origin (↓)	Mag-R (↓)	Mag-P (↓)	Seg. (↑)	Chamf. (↓)	F1@0.1 (↑)
I-split	RandMot	12.6%	56.6%	78.8°	1.09	62.0°	0.746	54.5%	5.18	7.50
	FreqMot	13.0%	69.0%	66.2°	0.689	51.7°	0.335	78.1%	3.51	9.09
	CA^{2}T-Net	89.0%	99.2%	6.23°	0.274	36.7°	0.166	92.2%	1.05	42.5
O-split	RandMot	13.6%	58.6%	79.1°	1.13	63.2°	0.734	54.5%	4.79	7.57
	FreqMot	13.2%	74.4%	67.3°	0.792	54.7°	0.333	76.6%	3.38	8.84
	CA^{2}T-Net	74.0%	91.5%	58.7°	0.605	51.9°	0.230	81.8%	1.54	33.3

Table 6.2: **Quantitative performance of baselines and CA^{2}T-Net (ours) on SAPIEN validation split.** Results for per-image (*I-split*) and per-object (*O-split*) dataset splits.

resize images to 224×224 and sample a pointcloud of 2500 points for each mesh. Pointclouds are further normalized to fit in a unit box.

Baselines

We consider three baseline approaches to single-view articulation transfer: selecting a random motion and part (RandMot), selecting the most frequent motion parameters (FreqMot), and OPD-O [67], the current state-of-the-art approach for 3D motion prediction. RandMot and FreqMot are inspired by similar baselines used by OPD [67], however modified to predict additional outputs such as object pose and motion magnitude. RandMot gives a lower bound on performance whereas FreqMot is a relatively strong heuristic-based baseline approach.

RandMot For the random motion (RandMot) baseline, we sample a random set of ground truth labels from the training dataset, including random pose, motion axis, motion origin, type, and magnitude. For the predicted part segmentation, we randomly sample a valid part from the set of parts contained in the input mesh.

FreqMot In contrast to the RandMot baseline, the frequent motion (FreqMot) baseline method predicts the most frequent motion parameters from the train dataset. We still randomly sample a global pose, but sample the predicted part segmentation based on statistics from the training dataset. We compute the average size of the most frequently sampled part in the training set and select the part that is most similar in size at inference time. For the remaining categorical variables, we sample the most frequent value from the training set. For real-valued parameters such as motion magnitude, motion axis, and motion origin, we first cluster the values from the training set and sample the mean value of the most frequent cluster.

Openable Part Detection [67] While designed for a different problem formulation, Openable Part Detection (OPD) is the most applicable to our articulation transfer task in terms of model output. OPD takes a single RGB image and predicts 2D segmentation masks and 3D motion parameters for all parts identified as "openable". While their method cannot be directly used to animate a 3D model, we can still compare to them in terms of 3D motion prediction and 2D part segmentation. Specifically, we compare our method to OPD-O, the best-performing variant of OPD in terms of real image performance. OPD-O also predicts motion parameters in the world coordinate system by using an additional branch to predict extrinsic parameters.

Metrics

We consider evaluation metrics for pose, motion, and segmentation performance, as well as 3D reconstruction metrics on the final deformed mesh. Our *Pose* metric is the percentage

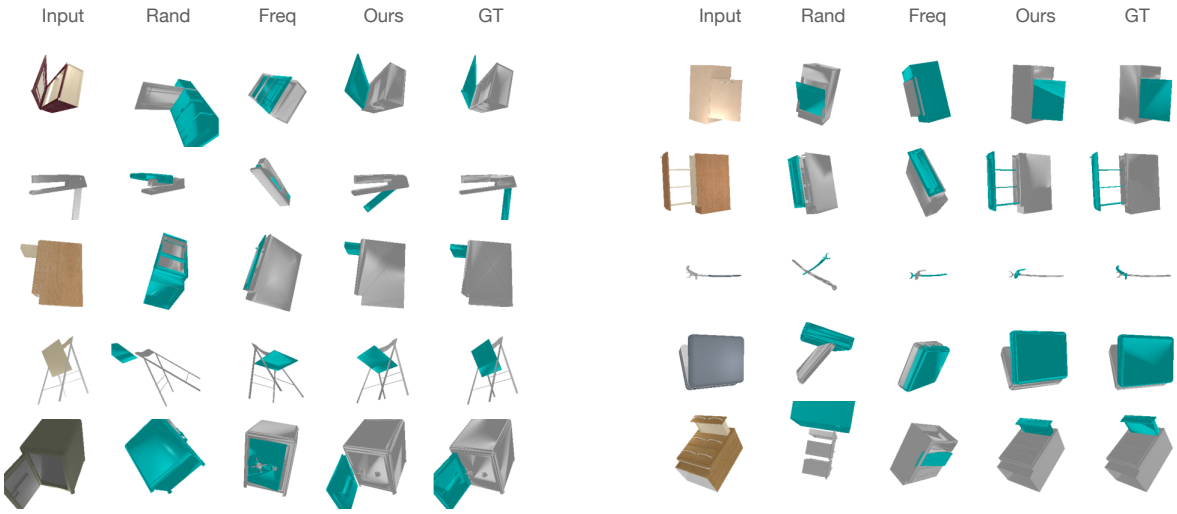

Figure 6.3: **Qualitative results on SAPIEN validation set.** Our method's predictions can be used to deform the input mesh to match the articulated image.

of pose estimations with rotation error less than 30°. *Motion-Type* refers to the 2-way classification accuracy of revolute or prismatic motion. *Motion-Axis* is the average angular error (in degrees) between the normalized predicted and ground-truth motion axes, and *Motion-Origin* is the average L_1 distance between the predicted and ground-truth motion origin. *Mag-R* and *Mag-P* refer to the revolute and prismatic magnitude errors, respectively, measured in degrees for the case of revolute motion and L_1 distance for prismatic motion. Because motion origin is not applicable for prismatic motion, we only evaluate motion origin for predicted revolute parts. *Seg.* refers to the 3D segmentation accuracy, measured across 1000 sampled points. Finally for 3D reconstruction metrics, we report Chamfer distance and F1@0.1.

Experimental Results

Performance on Validation Set We report the performance of our model trained on the SAPIEN dataset in Table 6.2. As no neural network-based approach currently exists for the full single-view articulation transfer task, we compare against our two heuristic-based baselines: RandMot and FreqMot. We consider two train/validation splits of varying difficult, *I-split* (per-image: validation set contains unseen images) and *O-split* (per-object: validation set contains unseen images of novel objects).

Compared to the RandMot and FreqMot baselines, we find that our method achieves the best performance across all metrics. Understandably, the validation performance is better on the I-split than the O-split. Visualizations of each method's predictions can be found in Figure 6.3.

	Motion			
	Type (↑)	Axis (↓)	Origin (↓)	mAP (↑)
OPD-O [67]	94.0%	72.5°	0.912	41.0%
CA^{2}T-Net	98.9%	50.5°	0.757	49.2%

Table 6.3: **Comparison to OPD for 3D motion prediction and 2D segmentation**. We compare to OPD, a method that estimates motion parameters for *all* movable parts given an RGB image. For segmentation, we report mAP@IoU=0.5.

Comparison to OPD We compare our approach to the recently introduced OPD method [67], which represents the state-of-the-art in terms of 3D motion estimation. OPD predicts 2D segmentation masks, motion type, and 3D motion parameters (axis, origin) for *all* moveable parts (based on priors in the training dataset) given a single image, whereas CA^{2}T-Net takes a 3D model and a driving input image, and predicts 3D segmentation, motion type, and 3D motion parameters (axis, origin, *and* magnitude). To compare the two approaches on a somewhat level playing field, we evaluate the performance of CA^{2}T-Net and OPD on the motion parameter predictions that the methods share in common. Because OPD makes multiple predictions for all moveable parts in an image but in our problem setting we are only interested in the part that has actually moved in the image, we take only the OPD predictions that have ≥ 0.5 IOU with the ground-truth segmented part. Further, as OPD is limited to predicting only *openable* part motion, we only consider datapoints that come from the subset of the 11 categories that OPD was trained on. This reduces the size of our valiation set from 10,000 images to 1,039 images. We utilize the CA^{2}T-Net model trained on the more difficult per-object training split (O-split).

Results can be found in Table 6.3. We also compare to OPD in terms of 2D segmentation ability, by projecting the CA^{2}T-Net -predicted 3D part segmentation to 2D and measuring mAP at IoU = 0.5. In general we find that CA^{2}T-Net outperforms OPD in terms of motion type, origin, and axis prediction, as well as 2D segmentation ability. Intuitively, learning 3D features from the input model in CA^{2}T-Net should support better prediction of 3D motion parameters. A common failure case that we observe with OPD is predicting a motion axis that is 180 degrees rotated from the correct axis.

Multi-Task vs Single-Branch Learning We study the contribution of each of the loss terms corresponding to each branch of our architecture by training three variants of our model: pose prediction only, motion prediction only, and 3D segmentation prediction only. The results can be found in Table 6.4. We find that the multi-task objective with all three loss terms leads to the best performance across most metrics, specifically for pose and 3D segmentation estimation. The performance gains for motion estimation are less pronounced. Intuitively, accurate pose prediction should aid the alignment between the image and shape

		Motion					
	Pose (↑)	Type (↑)	Axis (↓)	Origin (↓)	Mag-R (↓)	Mag-P (↓)	Seg (↑)
Pose Only	69.6%	-	-	-	-	-	-
Motion Only	-	92.5%	52.2°	0.599	52.3°	0.26	-
Seg Only	-	-	-	-	-	-	78.9%
All	74.0 %	91.5%	58.7°	0.605	51.9°	0.23	81.8%

Table 6.4: **Training for individual objectives vs multi-task learning.** We train three separate networks for pose, motion and segmentation only and compare their performance to the proposed multi-task objective. We find that optimizing all losses together yields the best results.

inputs, which should improve part segmentation ability. Such an effect may explain the boost in performance by doing multi-task training.

Failure Cases The most common failure cases of our method include pose prediction failure (e.g., predicting a pose that is 180-rotated from the ground-truth pose) as well as failure to segment any part of the object when the motion magnitude is small. Visualizations of these failure cases can be found in Figure 6.6.

6.5 Applications

We present three exciting applications of our model: generalization to functionally similar objects (i.e., motion analogies), 3D model animation, and transfer to real images and meshes in ABO.

Motion Analogies

During training, our method takes as input a 3D model that exactly matches the geometry of the object pictured in the 2D image. However, at test time we can swap the 3D input with a functionally similar 3D model that does not exactly match the input image to perform a *motion analogy*. To achieve this, we swap the input model with a different model from the same category as the original input 3D model, and run our method with no additional modifications. The articulated outputs predicted by our method can be seen in Figure 6.4. We note that in this setting our model has to not only segment and predict motion for the corresponding articulated part on the swapped shape, it also has to accurately predict the pose in this setting that was never seen during training. In general we find that our method can segment the functionally similar part and produce motion parameters that lead

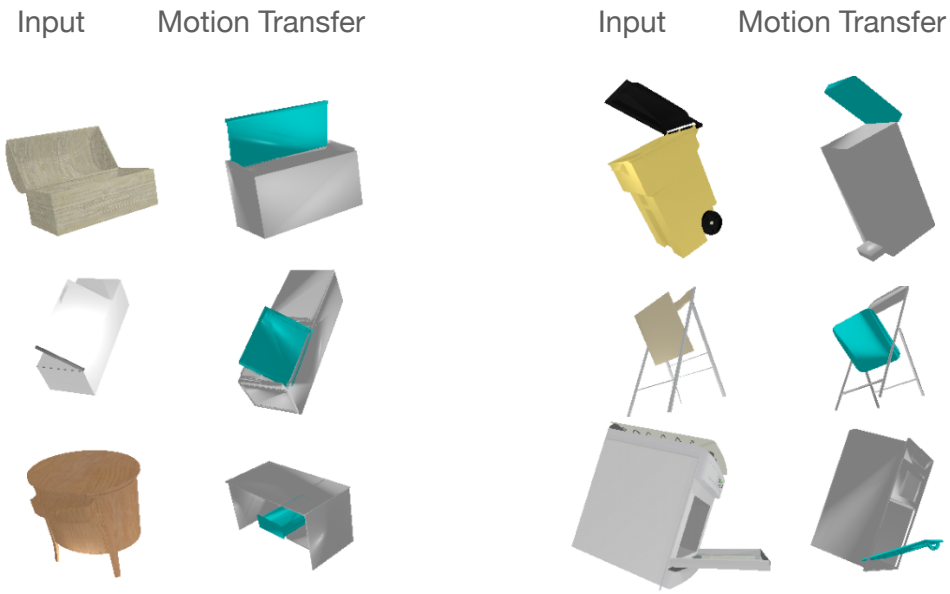

Figure 6.4: **Motion analogies.** By presenting a functionally similar but geometrically distinct 3D model at test time, we can transfer motion across analogous 3D parts.

to a plausible deformation of the non-matching input model, even when the input model is drastically different from the articulated image (see Figure 6.4 left column, third row).

Animation: Interpolation and Extrapolation

A convenience of our motion parameterization is that animation of the input 3D model can be achieved simply by varying the predicted motion magnitude. As a result, the predicted segmented part will move along or around the predicted axis. We can interpolate between the rest-state magnitude (0) and predicted magnitude to simulate the motion of a part opening and closing, or even consider magnitude settings beyond the predicted value to get motion extrapolation. See Figure 6.5 for an animation example.

Transfer to Real Objects & Images

A natural source of real-world articulated objects, including rest-state mesh and image pairs, comes from the ABO dataset [25]. The dataset contains catalog images of Amazon products, as well as a subset of 3D models corresponding to the object in the images. The catalog images range from stock-like photos to close-up detail shots, and also sometimes contain an articulated picture of the object. We gathered a subset of approximately 50 catalog images that show articulated objects (with a single moving part) for which there also exist

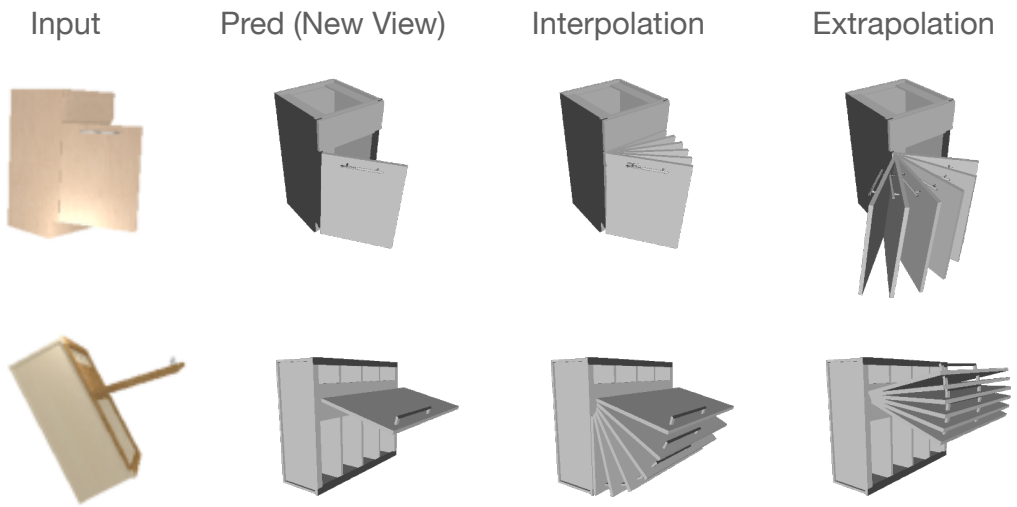

Figure 6.5: **Interpolation and extrapolation in 3D from a single input image.** Our motion parameterization can easily be used to interpolate between the rest-state and articulated object pose, or even extrapolate beyond the motion that is seen in the input image.

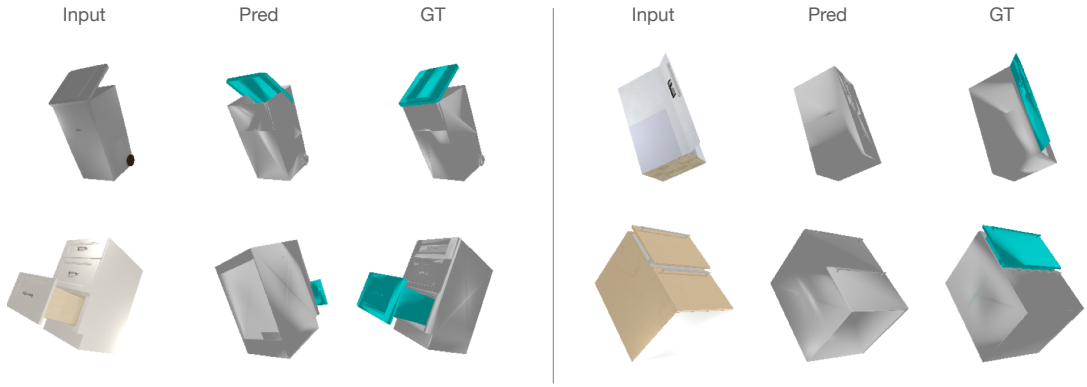

Figure 6.6: **Typical failure cases in articulation transfer.** Pose prediction failures (180 degrees off, left column), and failure to predict a segmented part when motion magnitude in image is small (right column).

corresponding 3D models as a real-world use-case of the single-view articulation transfer setup.

Unlike the training objects in the SAPIEN dataset that have annotated part segmentations, meshes from the ABO dataset are unsegmented. Further, articulated image instances in ABO have no ground-truth in terms of pose, part segmentation, and motion annotations.

To achieve a candidate part segmentation we leverage the fact that the meshes in ABO happen to be designed by artists, part by part, and thus different parts of the object tend to be represented by mesh pieces that are disconnected from the rest of the object. As a result, running connected components on the ABO meshes tends to give us an over-segmentation of the object's functionally relevant parts. We treat these connected components as pseudo-part segmentations and can then run CA^{2}T-Net on this data with no modifications.

Qualitative results showing our model applied, without any finetuning, to real images from the ABO dataset can be seen in Figure 6.7. Note that the meshes in ABO were designed to simply show the product in its rest-state, rather than with articulation in mind (like in SAPIEN). This is shown by the fact that drawers are simply external panels rather than a full drawer part enclosed by the rest of the object. Further, our model was trained only with inputs that show a single articulated part, whereas realistic catalog images typically show multiple articulated parts at once. We include a challenging instance of this in Figure 6.7 (row 3), where the input image contains a recliner chair, which is a category *not seen during training*, with multiple complex motions, including a reclined headrest and raised footrest. As our method is category-agnostic, it can handle inputs like this at test-time.

6.6 Discussion

We proposed a method to transfer the articulation from a single-view image to a 3D model in such a way that the model can be deformed to match the input image. To our knowledge, we are the first to solve this task in the setting of household objects from arbitrary categories. In this work we only consider articulated exemplar images with a single moving part displayed. Of course, articulated images can commonly show multiple moving parts. Extending our method to predict multiple motion types and part segmentations is a practical next step and can be achieved by combining Mask RCNN-style detections (as used in OPD) with our method. Further, our method requires a 3D model as input which can limit its practical use in say, an embodied perception system navigating the world. Recent advances in CAD model retrieval from a single image [82, 83, 51] can be used to retrieve a similar 3D CAD model, and our results on functionally similar shape swapping suggest that our approach will be able to handle this case. Thus combining CA^{2}T-Net with CAD retrieval methods could be a promising path forward to allowing our method to run on single image inputs only. Finally, we consider only rigid motion (further, just prismatic and revolute motion) in this work, but considering how this approach can be extended to more complicated motion types as well as non-rigid deformation is an interesting avenue for future work.

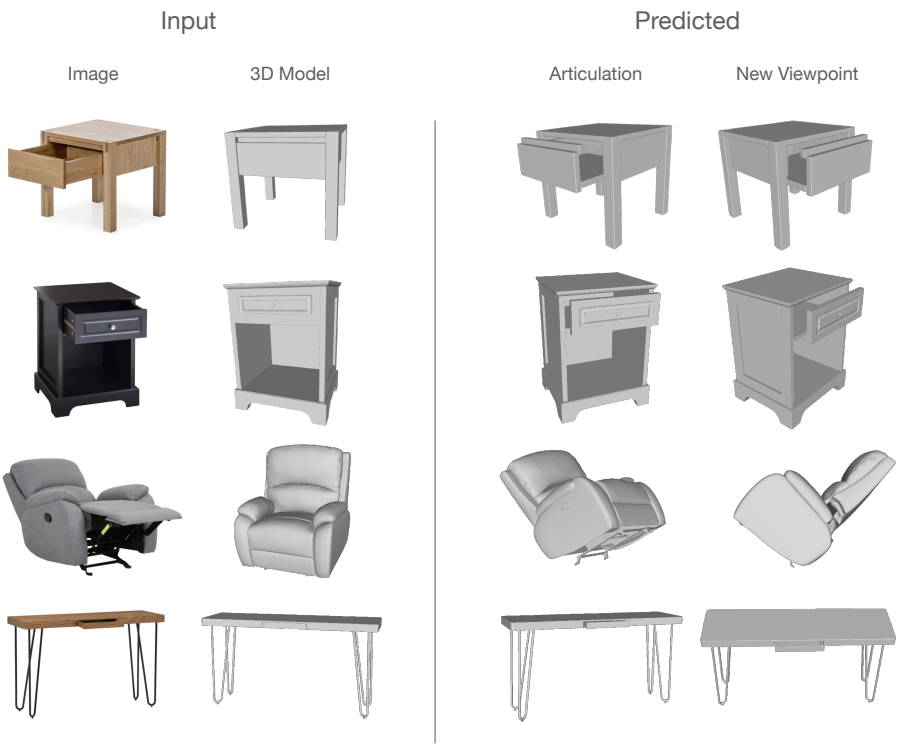

Figure 6.7: **Transfer results to real articulated objects.** We run our model on instances of ABO objects with no finetuning and are able to make reasonable predictions despite the domain gap.

Chapter 7

Multi-View Data for Self-Supervised Learning

Figure 7.1: **Generating 3D data augmentations on ImageNet.** We explore the possibility of generating 3D rotations of ImageNet images, in addition to the standard 2D data augmentations such as color jitter and cropping.

This chapter studies the impact of learning unsupervised contrastive features with multiple viewpoints of objects. This is in contrast to the standard setting, where learning occurs only with single image views that are then 2D augmented with various image-level transformations. We first render a synthetic dataset from CAD models that allows for training with multiple viewpoints of the same object, and find that the learned representations outperform training with single views in terms of nearest neighbors classification. We then introduce an image-to-mesh generation pipeline that allows generating novel 3D views from any RGB input image ("3D data augmentation"). We apply this pipeline to the entire ImageNet [29] dataset, and find that training with 3D data augmentation slightly improves downstream ImageNet accuracy, as well as performance on various out-of-distribution test sets.

This chapter includes joint work with Ari Morcos, Georgia Gkioxari, Shiry Ginosar, and Jitendra Malik

7.1 Motivation

Humans are remarkably able to extract general representations of objects from a single image. These representations capture both the shape and texture of objects and enable us to reason about category membership and object affordances. We learn this implicit shape reasoning capability in childhood by viewing objects from multiple angles—first passively and eventually actively by directly interacting with them [4]. In fact, extensive multi-view visual input from just a small set of objects makes up the bulk of visual data that we receive as infants [23]. This experience with object-centric views from a diverse set of poses [4] bootstraps children's ability to recognize objects from novel categories, eventually with as little as one label [85, 23]. Critically, object labels play a minor role in this learning process.

In machine learning, self-supervised methods such as contrastive learning [86, 53, 37, 163, 151, 106, 19, 18, 57], have done away with the need for labels in object-representation learning. However, in contrast with the multi-view paradigm of human object-learning, these methods rely on only seeing single instances, and therefore single views, of different objects. This is perhaps due to the current dominant trend of training neural networks with readily-available large datasets of single images.

In this paper, we present a scientific study to understand the benefits of learning from multiple views of the same object, compared to the traditional single-image approach [29, 56, 74]. Taking note of the ability of infants to learn shape representations from visual input with minimal category labels [85, 23], we study this question in a self-supervised, contrastive framework where *no category labels are required for learning*. Rather, we use a naturally available cue—the preservation of object identity when viewed from different orientations. This form of supervision is natural for an active agent collecting its ego-centric datastream.

As real-world object-centered multi-view datasets are not abundant, we first demonstrate a system that learns from rendered views of synthetic objects. Nevertheless, we show that our model trained with synthetic multi-view inputs is better able to capture class and shape-based information that generalizes to novel instances and categories. Using advances in off-the-shelf tools [129, 145, 135], we then introduce a 3D data augmentation pipeline that can extend this idea to arbitrary single-view image datasets, such as ImageNet [29], when no multi-view data is actually available.

7.2 Background

Contrastive learning Contrastive learning and triplet losses have a long history in representation learning [86, 53, 37]. More recently, contrastive losses have seen a resurgence as a very successful method for unsupervised pretraining for ImageNet classification [163, 151, 106, 19, 18, 57]. These methods typically rely on some type of data augmentation on a single image in order to learn representations that are invariant to augmentations at the image instance-level. Rather, we are interested in learning representations that are invariant to viewpoint changes at the object-level. Most related to our efforts are Lin et al. [98], who

learn from multiple views of rendered objects and demonstrate that their learned metric is better aligned with human perception than AlexNet features. While their experiments consider only synthetic images, we provide evidence that multi-view images (synthetic and real) are beneficial for tasks on real-world data.

Learned pose- and illumination-invariance has been well-studied in the context of face recognition. For example, the seminal work of [131] employs metric learning on multi-view face images. Here, we show the value of multi-view representation learning for recognizing and generalizing to arbitrary classes of objects.

Learning 3D from multiple views Learning explicit 3D representations from multiple views [22, 103, 168, 179, 46] is a problem that is related to ours. However, in addition to multi-view images, these methods often require 3D supervision and/or camera intrinsics and extrinsics. Moreover, they operate either on a single object or on a few classes of objects that were seen during training. In contrast, we do not require any extra forms of supervision and learn an implicit representation that is invariant to pose.

Learning classes from multiple views Bambach et al. [4] collect a multi-view dataset for supervised learning based on images collected from a head-mounted camera on both toddlers or adults. They found that the views from toddlers contained a more diverse range of poses and object scales than adults, and networks trained on toddler-generated multi-view data generalized better than that of adults. We are interested in this problem but in an unsupervised setting. Further, while [4] only showed generalization results on different views of single objects, we show generalization across objects, as well as to novel object classes.

Texture bias in CNNs A common finding in both ImageNet-supervised [42] and self-supervised neural networks [43] is that they are biased towards texture, rather than shape. Hermann et al. [59] have shown that data augmentations such as random cropping can increase texture bias in self-supervised networks. Existing approaches have either used artificial stylized images [42, 92] or non-cropping data augmentation [59] to reduce texture bias. In this work, we show that training on naturalistic, multi-view data can reduce the texture bias in CNNs.

7.3 Learning View-Invariant Representations with ShapeNet

To gain an intuition about the effects of training with multiple 3D views in contrastive learning, we set up a toy dataset based off of synthetic mesh models for which we can render arbitrary novel viewpoints. We then can leverage this multi-view data as supervision to learn visual representations that are invariant to pose. To understand the benefits of multi-view imagery, we build off of the contrastive learning framework of SimCLR [18]. We train a

single-view network, where two images, x_i and x_j, that originate from the same underlying image and differ only by randomly sampled data augmentations are passed through encoders with shared weights to obtain $h_i = f(x_i)$ and corresponding h_j. These representations are further passed through a small multi-layer perceptron (MLP) projection head, $g(\cdot)$ to obtain z_i and z_j. We apply normalized temperature-scaled cross entropy loss [18] to push the learned representations together, and away from those corresponding to other image instances. In practice, we train in a batch setting and negatives are obtained from the other examples in the batch. We also train a **multi-view network**, where instead of augmented versions of the same image, we propose to use two viewpoints of the same object as input to the model. Over training, this model learns to push the learned representations of the same object in different poses together, and away from representations of other objects. In other words, under perfect training conditions, this model should learn to become invariant to object pose. Finally, to get a feel for the difficulty of each dataset and task, we train a variant that uses class labels (**class-supervised network**). In this setup, rather than sampling two augmented versions of the same image, or two viewpoints of the same object, positives are drawn from the same class and negative examples come from different classes.

Experimental Setup

Dataset We utilize ShapeNet [17], a collection of 3D CAD models commonly used for benchmarking the performance of 3D reconstruction methods. Because ShapeNet objects are organized into 55 distinct categories, we can leverage this category information to create an effective train/test split for testing generalization. We train on only 20 categories (decided as those with $\geq$ 500 models per category). For evaluation, we hold out object instances both from training classes, as well as evaluate on instances from the 35 held-out classes. For training, we use rendered images from 500 different models (24 poses per model) per class. We render viewpoints evenly spaced from 0°-360° in azimuth and elevation sampled randomly from the range [-10°, 10°]. The rendered dataset is divided into an 80%/10%/10% train/val/test split at the object-level (views from the same object always fall within the same split).

Implementation Details All models use a ResNet-18 [55] backbone. For these experiments we use a 512-D h-representation and 256-D z-representation. We train with a batch size of 256 and Adam optimizer with a learning rate of 5e-4 that decays with a multiplier of 0.1 after 30,000 training steps. Since we are interested in studying *only* the effect of training on single vs. multi-view images, we train both models in precisely the same way using the same data augmentation (from [18]).

Results

Pose-Invariant Representations Better Capture Class Structure To quantify how well each learned representation captures class structure, we set up a 1-NN classification

	Seen Classes		Unseen Classes	
	$k = 1$	$k = 5$	$k = 1$	$k = 5$
R	9.2 ± 2.0	19.8 ± 2.8	12.0 ± 2.3	25.2 ± 3.1
SV	34.8 ± 3.1	53.4 ± 3.2	24.5 ± 3.0	40.4 ± 3.5
MV	**48.4 ± 3.0**	**62.0 ± 3.1**	**49.8 ± 2.7**	**68.4 ± 2.5**
CS	58.8 ± 2.9	69.9 ± 2.8	28.2 ± 3.2	45.7 ± 3.5

Table 7.1: **Multi-view networks outperform single-view networks on ShapeNet NN classification.** Classification accuracy (± SEM) for random weights (R), single-view (SV), multi-view (MV), and class-supervised (CS) networks using a support set of either $k = 1$ or $k = 5$ examples per class.

task. We extract features for all images in the test set, and create either a one-shot or five-shot classification similar to [84]. That is, we sample a query image in addition to a support set of either 1 or 5 examples from each class, then use the label of the instance from the support set with highest cosine similarity to the query image as the class prediction. We conducted 10,000 runs total and average results within-class and finally across classes.

In this experiment, we perform this classification on *unseen* instances from classes *seen* during training. Because of the different number of classes for seen and unseen on ShapeNet (n = 20 seen classes and n = 35 unseen classes) we perform a 55-way classification on all classes at once. This is done so that later when we also look at the performance on *unseen* classes, the final accuracy numbers can be directly compared without being conflated by the difficulty of the classification problem. However, to understand the generalization performance of each condition, we report seen classification separately from unseen.

Classification accuracy on seen classes for the single- and multi-view networks can be found in Table 7.1 (left). Additionally, we report a baseline of a randomly initialized network (to get a feel for how identifiable each class is by random features alone) and a second "baseline" of the class-supervised variant of the model. This gives us an indication of the level of performance we should expect when labels are available for supervision. We find that training with multi-view images gives a significant boost over single images from ShapeNet: +13.6% in the case of 1 example per class and +8.6% for 5 examples (Table 7.1, left). As expected, the class-supervised model outperforms all other methods.

Pose-Invariant Representations Generalize to Novel Classes Next, we test how well single-view and pose-invariant features trained on ShapeNet, generalize to both unseen classes from ShapeNet. The results in Table 7.1 (right) demonstrate that representations trained on multi-view images outperform those trained on single views. Interestingly enough, we find that class-supervised features generalize *worse* than multi-view features on the 35 unseen ShapeNet classes. We take this as a positive indication that multi-view object-level supervision leads to features that generalize better to new classes of object shapes than those

Figure 7.2: **Shape vs. texture task**. s_1 represents the cosine similarity between the query object x and a different object with the same texture x_t. s_2 represents the cosine similarity between the query object x and the same object with a different texture x_s. We call the decision *texture-biased* if $s_1 > s_2$ and shape-biased otherwise.

obtained from class labels.

Multi-View Training Increases Shape Bias

Having models that are biased more towards shape (over texture) is desired as this bias is more aligned with human perception [42]. We wondered if perhaps training with multi-view images could increase shape bias. To study this, we formulate a novel test-time task to investigate the shape or texture bias of each model's learned representations. Our setup is inspired by [42, 43] but unlike these prior works, is independent of class and class-specific texture choices. We use this task to probe the representations of the different self-supervised models directly in a *class-agnostic* manner.

Our task setup relies on access to synthetic 3D models, for which we can easily swap out the surface textures and render the models from different views. For this task, we generate sets of triplets of the form $\{x, x_s, x_t\}$. Here, x is some reference object with some underlying shape s and texture t. x_s is an object of the same shape as x, but with a different texture. x_t is an object with the same texture as x, but a different shape. In other words, x_s is a *shape match* to x and x_t is a *texture match* (see Figure 7.2). To prevent shape-matching from being trivially easy, we sample images of x_s and x_t from different poses than x. We performed this task on models trained on the 20 ShapeNet training categories. In total, we rendered 105 objects (3 models from each of the 35 unseen classes) with 5 different textures from 5 different viewpoints and tested each model on 10,000 triplets. We opt to use objects from unseen classes to prevent any possibility of the networks having associated certain textures with certain training classes. For each triplet, if a network finds the shape-matched pair to be more similar than the texture-matched pair, we call it a "shape decision" on that sample. The more shape decisions a model makes, the mode shape biased we assume it is.

It is reasonably well documented that single-view networks display a bias towards texture [42, 59, 43], but we find that networks trained with multi-view images (and either no data augmentation, or color distortion augmentation only) have a tendency to make many more shape decisions on our Shape vs. Texture task. However, this bias is diminished by the specific data augmentation used in [18], which includes a combination of cropping and color distortion. We suspected that the cropping might be the culprit behind this reduction

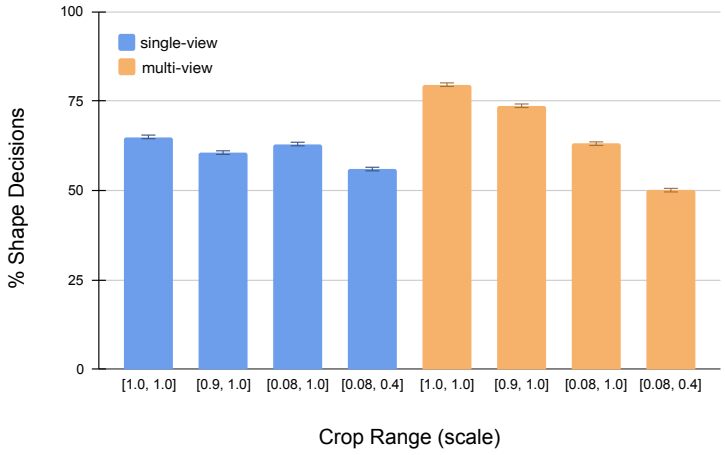

Figure 7.3: **Multi-view networks are more shape-biased than single-view ones, and cropping increases texture-bias**. X-axis labels represent the range from which random crop scales are sampled ([1.0, 1.0] corresponds to no cropping). While multi-view networks trained without cropping are more shape-biased than single-view ones, cropping reduces this tendency. The same is not true for single-view networks.

in shape bias and set up a controlled study to find out. We train single- and multi-view networks with color augmentation and cropping, but modify only how much random cropping is performed by controlling the range from which the scale of the crop was sampled. We then measure the outcome of these modifications on shape bias. As can be seen in Figure 7.3, we find that 1) multi-view networks are capable of being more shape biased than single-view networks, and 2) cropping has a direct impact on the amount of shape bias for multi-view networks (and less of an effect for single-view networks). To the best of our knowledge, this is the first demonstration that training with naturalistic multi-view images, akin to the visual input that humans receive, can increase the shape bias of CNNs.

7.4 Scaling up to ImageNet

Given promising results on the rendered ShapeNet dataset, we wondered if such a result could scale to larger and more realistic datasets such as ImageNet [29]. Unlike the case with ShapeNet, we do not have corresponding 3D models in the dataset and thus develop a new pipeline for generating novel views from each image. In theory, this approach can be applied to any image dataset. As we are now working with single images rather than rendering new views from a CAD model, we refer to this new-view generation as a "3D data augmentation".

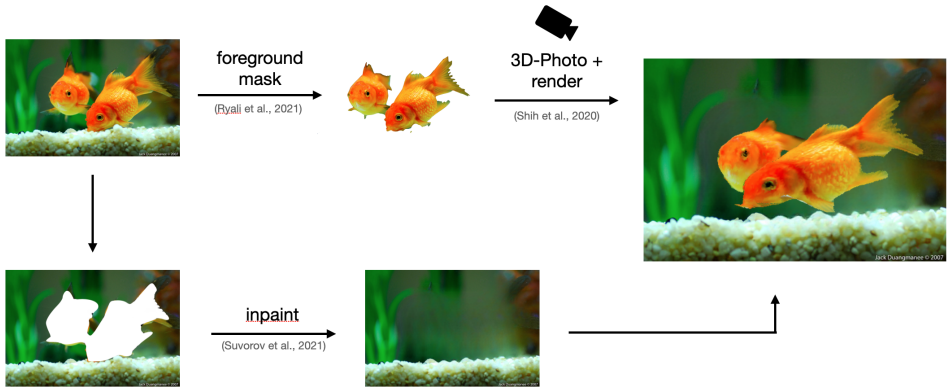

Figure 7.4: **Overview of 3D data augmentation pipeline.** For each ImageNet image, we separate the foreground and background. We inpaint the background image and use the foreground to generate a mesh for which we can render at a novel view.

7.5 Method

3D Data Augmentation

Our mesh generation pipeline consists of a few steps including foreground-background splitting, background inpainting, foreground mesh generation, novel viewpoint rendering, and image quality filtering. A high-level summary of the mesh generation process is visualized in Figure 7.4.

Background Inpainting We generate background images by inpainting the foreground object according to saliency masks from [129]. In general, masking out the foreground object leaves a trace of foreground-colored pixels around the edge, and the inpainting algorithm tries to reconstruct the foreground object. To avoid this, we dilate the foreground mask before inpainting. We use the LaMa inpainting algorithm [145].

Mesh Generation We based our mesh generation pipeline largely on 3D Photo Inpainting [135], however we found that their inpainting algorithm caused large stretching artifacts when applying any kind of large rotation, and thus skip that major step. We also only generate a mesh for the foreground portion of the image, where foreground pixels are determined by masks from [129]. For depth estimation, we use MiDaS v3 [123]. Again, with noisy foreground masks we find that some depth values at the edge of the foreground can be noisy and cause stretching artifacts in the mesh. To avoid this, we remove outlier depth values and additionally remove those pixels from the foreground mask. Each generated mesh typi-

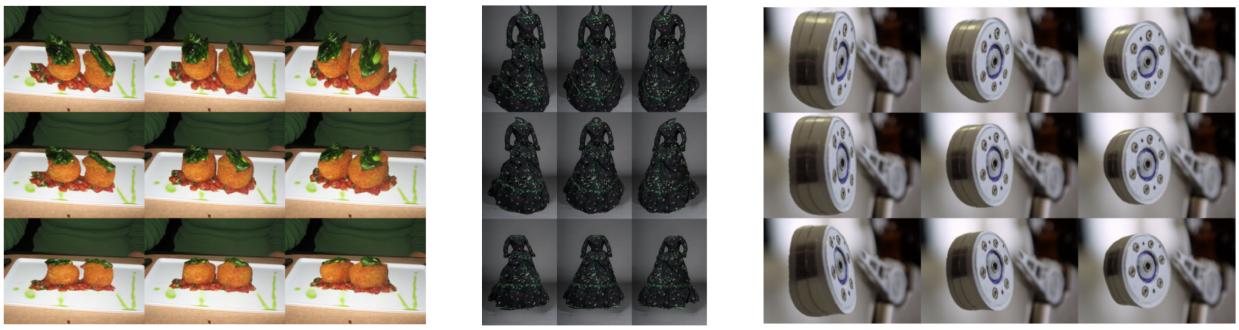

Figure 7.5: **Sample images rendered from mesh generation pipeline.** Center image represents original ImageNet image viewpoint. Left to right is varying azimuth, top to bottom is varying elevation.

cally contains hundreds of thousands of vertices, thus we downsample each mesh to contain <50,000 vertices. Lastly, we center and scale each mesh to the origin and with unit scale.

Novel Viewpoint Rendering We render novel views of our generated foreground meshes using Pytorch3D [124]. We randomly rotate each mesh by sampling an azimuth and elevation uniformly in [-15.0, 15.0] and render a new view at 512x512 resolution, for up to 10 images per ImageNet instance. We then calculate the bounding box around the rendered foreground object and resize it such that the width of the rendered image matches the width of the original foreground object. This resized rendering then gets pasted onto the inpainted background image at the same location as the previous foreground object.

Image Quality Filtering Some rendered images look unrealistic due to artifacts in the mesh, poor foreground masks, or too much rotation. We are able to filter out some of these instances by passing each newly generated image into a ResNet-50 pretrained on ImageNet, and seeing if the predicted class label changes between the rendered image and the original ImageNet image. We do not keep renderings where the predicted class changes. After the filtering criteria described, we are able to generate 3D augmentations for 79% of the ImageNet training images. At the end of this process we are left with 6,186,264 images (roughly 5× the size of ImageNet). Sample images that pass the image quality filtering step can be found in Figure 7.5.

Training

We train MoCo v2 [19] with a ResNet-50 backbone using distributed data parallel across 4 nodes with 8 GPUs each. We use a batch size of 256 per node (per-gpu batch size of 32) and learning rate of 0.12. We train for a total of 220 epochs, with a linear warmup for the first 20 epochs. We evaluate the learned representations using a linear probe, finetuning the final

Task	Top-1	Top-5
2D Aug.	67.6%	88.0%
3D Aug.	68.3% (+0.7%)	88.4% (+0.4%)

Table 7.2: **Linear probe ImageNet accuracy.** Top-1 and top-5 ImageNet accuracy for models trained with 2D and 3D data augmentation.

linear layer weights on ImageNet classification labels. This finetuning is done on 8 GPUs for 100 epochs using a total batch size of 1024, and a learning rate of 120 (divided by 10 at 60 and 80 epochs).

We experiment with incorporating the 3D augmented images in different ways. In all cases we have p_{aug} which is the probability a 3D augmented image gets sampled. Otherwise, 3D augmentation is not applied. In all experiments we also apply the standard 2D data augmentation recipe, regardless of if the image is 3D augmented or not. When sampling a 3D augmented image, we pass it as a query image only. In other words, we do not add the 3D augmented image to the momentum encoder. We found that doing so causes a decrease in performance. We experimented with different schedules for p_{aug}, but found that starting at $p_{aug} = 1.0$ and linearly annealing down to $p_{aug} = 0.05$ over the entire course of training worked best.

7.6 Results

We found that incorporating 3D augmented images into training improves linear probe ImageNet top-1 and top-5 accuracy by 0.7% and 0.4%, respectively (Table 7.2). Results are an average of 3 runs. We also evaluated the fine-tuned model on the Cue-Conflict, Edge, and Silhouette downstream tasks that favor shape features over texture-based features [41, 42]. We find that training with 3D data augmentation improves accuracy on all of these tasks, with the most notable gain on the Edge dataset (+5.8%).

7.7 Discussion

In this chapter we studied the benefit of multiple views in object-centric contrastive learning for a variety of synthetic and real world datasets. We discovered that training with multi-view images can lead to pose-invariant representations that outperform those learned from single views for the downstream task of classification. We further found that training on multi-view images can increase the shape bias of CNNs, when trained without aggressive cropping. Finally, we scaled the multi-view dataset creation idea to a large-scale, real-world dataset, ImageNet. We found that incorporating multiple views further can increase accuracy

Task	Cue Conflict	Edge	Silhouette
2D Aug.	19.5%	17.7%	39.4%
3D Aug.	20.2% (+0.7%)	23.5% (+5.8%)	41.0% (+1.6%)

Table 7.3: **Performance on out-of-distribution evaluation sets.** Accuracy of model trained with 2D and 3D augmentation on cue-conflict, edge, and silhouette test sets.

on downstream classification and shape-based tasks. We hope these findings will help future researchers in building machine vision systems which are more capable—need fewer labels, generalize better, and be closer to human performance than the systems of today.

Chapter 8

Conclusion

In this thesis we presented three efforts towards bridging the gap between humans and machines in 3D object perception. Inspired by the ability of infants and toddlers to rapidly learn about objects in the world, we highlighted three key features of their visual inputs: realism, interaction (object articulation), and diverse, multiple views with sparse labels. We then went on to propose a dataset-driven solution in each of these directions.

In *The Development of Embodied Cognition: Six Lessons from Babies*, Smith and Gasser [138] offer advice from years of research experience in developmental psychology for building embodied, intelligent agents. Some of the directions explored in this thesis overlap with the lessons proposed by Smith and Gasser, but there are many, still untapped, fascinating areas of research. For example, multi-modal models are another promising direction for improving the ability of neural networks to perceive and represent objects. By marring vision and language, we can enable machines to learn more complex associations between objects and their attributes, such as their functions, relationships, and categories. In the future, incorporating additional modalities such as touch, sound, and proprioception could further enrich the learning experience of machines and enable them to better understand the physical properties and affordances of objects. For example, haptic feedback can provide information about the texture, shape, and weight of an object, which can be useful for manipulation, grasping, and exploration. Similarly, auditory feedback can indicate the position, movement, and interactions of objects in the environment, which can help machines to navigate, localize, and recognize objects.

In summary, while the efforts presented in this thesis are important steps towards improving the 3D object perception of machines, there are many other exciting avenues of research that can contribute to building more embodied and intelligent agents. By drawing inspiration from the learning strategies of infants and toddlers, we can push the frontiers of machine perception and enable machines to better understand and interact with the world around us.

Bibliography

[1] Ben Abbatematteo, Stefanie Tellex, and George Konidaris. "Learning to generalize kinematic models to novel objects". In: *Proceedings of the 3rd Conference on Robot Learning*. 2019.

[2] Adel Ahmadyan et al. "Objectron: A Large Scale Dataset of Object-Centric Videos in the Wild with Pose Annotations". In: *arXiv preprint arXiv:2012.09988* (2020).

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. "Layer normalization". In: *arXiv preprint arXiv:1607.06450* (2016).

[4] Sven Bambach et al. "Toddler-inspired visual object learning". In: *Advances in neural information processing systems* 31 (2018).

[5] Jonathan T Barron et al. "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 5855–5864.

[6] Miguel Angel Bautista et al. "On the generalization of learning-based 3D reconstruction". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 2180–2189.

[7] Charles Beattie et al. "Deepmind lab". In: *arXiv preprint arXiv:1612.03801* (2016).

[8] Jan Bechtold et al. "Fostering Generalization in Single-view 3D Reconstruction by Learning a Hierarchy of Local and Global Shape Priors". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 15880–15889.

[9] Sai Bi et al. "Deep 3d capture: Geometry and reflectance from sparse multi-view images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 5960–5969.

[10] Marc H Bornstein, Kay Ferdinandsen, and Charles G Gross. "Perception of symmetry in infancy." In: *Developmental psychology* 17.1 (1981), p. 82.

[11] Marc H Bornstein and Joan Stiles-Davis. "Discrimination and memory for symmetry in young children." In: *Developmental Psychology* 20.4 (1984), p. 637.

[12] Mark Boss et al. "Two-shot spatially-varying brdf and shape estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 3982–3991.

[13] Andrew Brock, Jeff Donahue, and Karen Simonyan. "Large scale GAN training for high fidelity natural image synthesis". In: *arXiv preprint arXiv:1809.11096* (2018).

[14] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[15] Nicola Bruno, Marco Bertamini, and Fulvio Domini. "Amodal completion of partly occluded surfaces: Is there a mosaic stage?" In: *Journal of Experimental Psychology: Human Perception and Performance* 23.5 (1997), p. 1412.

[16] Brent Burley and Walt Disney Animation Studios. "Physically-based shading at disney". In: *ACM SIGGRAPH*. Vol. 2012. vol. 2012. 2012, pp. 1–7.

[17] Angel X Chang et al. "ShapeNet: An information-rich 3D model repository". In: *arXiv preprint arXiv:1512.03012* (2015).

[18] Ting Chen et al. "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[19] Xinlei Chen et al. "Improved baselines with momentum contrastive learning". In: *arXiv preprint arXiv:2003.04297* (2020).

[20] Sungjoon Choi et al. "A large dataset of object scans". In: *arXiv preprint arXiv:1602.02481* (2016).

[21] Aakanksha Chowdhery et al. "Palm: Scaling language modeling with pathways". In: *arXiv preprint arXiv:2204.02311* (2022).

[22] Christopher B Choy et al. "3D-R2N2: A unified approach for single and multi-view 3D object reconstruction". In: *European conference on computer vision*. Springer. 2016, pp. 628–644.

[23] Elizabeth M Clerkin et al. "Real-world visual statistics and infants' first-learned object names". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1711 (2017), p. 20160055.

[24] Anna Coenen, Jonathan D Nelson, and Todd M Gureckis. "Asking the right questions about the psychology of human inquiry: Nine open challenges". In: *Psychonomic Bulletin & Review* 26 (2019), pp. 1548–1587.

[25] Jasmine Collins et al. "ABO: Dataset and benchmarks for real-world 3D object understanding". In: *CVPR*. 2022, pp. 21126–21136.

[26] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation. Stichting Blender Foundation, Amsterdam, 2018. URL: `http://www.blender.org`.

[27] C. Cook, N. D. Goodman, and L. E. Schulz. "Where science starts: Spontaneous experiments in preschoolers' exploratory play." In: *Cognition* (2011).

[28] György Darvas. *Symmetry: Cultural-historical and ontological aspects of science-arts relations; the natural and man-made world in an interdisciplinary approach*. Springer Science & Business Media, 2007.

[29] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[30] Valentin Deschaintre et al. "Flexible SVBRDF Capture with a Multi-Image Deep Network". In: *Computer Graphics Forum*. Vol. 38. 4. Wiley Online Library. 2019, pp. 1–13.

[31] Valentin Deschaintre et al. "Single-image SVBRDF capture with a rendering-aware deep network". In: *ACM Transactions on Graphics (TOG)* 37.4 (2018), p. 128.

[32] Laura Downs et al. "Google scanned objects: A high-quality dataset of 3d scanned household items". In: *2022 International Conference on Robotics and Automation (ICRA)*. IEEE. 2022, pp. 2553–2560.

[33] Magnus Enquist and Anthony Arak. "Symmetry, beauty and evolution". In: *Nature* 372.6502 (1994), pp. 169–172.

[34] Haoqiang Fan, Hao Su, and Leonidas J Guibas. "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*. 2017, pp. 605–613.

[35] József Fiser and Richard N Aslin. "Statistical learning of new visual feature combinations by infants". In: *Proceedings of the National Academy of Sciences* 99.24 (2002), pp. 15822–15826.

[36] Celia B Fisher, Kay Ferdinandsen, and Marc H Bornstein. "The role of symmetry in infant form discrimination". In: *Child development* (1981), pp. 457–462.

[37] Andrea Frome et al. "Learning globally-consistent local distance functions for shape-based image retrieval and classification". In: *2007 IEEE 11th International Conference on Computer Vision*. IEEE. 2007, pp. 1–8.

[38] Huan Fu et al. "3d-future: 3d furniture shape with texture". In: *arXiv preprint arXiv:2009.09633* (2020).

[39] Duan Gao et al. "Deep inverse rendering for high-resolution SVBRDF estimation from an arbitrary number of images". In: *ACM Transactions on Graphics (TOG)* 38.4 (2019), p. 134.

[40] Kyle Gao et al. "Nerf: Neural radiance field in 3d vision, a comprehensive review". In: *arXiv preprint arXiv:2210.00379* (2022).

[41] Robert Geirhos et al. "Generalisation in humans and deep neural networks". In: *Advances in neural information processing systems* 31 (2018).

[42] Robert Geirhos et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness". In: *arXiv preprint arXiv:1811.12231* (2018).

[43] Robert Geirhos et al. "On the surprising similarities between supervised and self-supervised models". In: *arXiv preprint arXiv:2010.08377* (2020).

[44] Robert Geirhos et al. "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673.

[45] Natasha Gelfand and Leonidas J Guibas. "Shape segmentation using local slippage analysis". In: *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*. 2004, pp. 214–223.

[46] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. "Mesh R-CNN". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 9785–9795.

[47] Alison Gopnik and David M Sobel. "Detecting blickets: How young children use information about novel causal powers in categorization and induction". In: *Child development* 71.5 (2000), pp. 1205–1222.

[48] Andreas Mischok Greg Zaal Sergej Majboroda. *HDRIHaven*. `https://hdrihaven.com/`. Accessed: 2020-11-16.

[49] Thibault Groueix et al. "A papier-mâché approach to learning 3d surface generation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 216–224.

[50] Khronos Group. *glTF 2.0 Specification*. `https://github.com/KhronosGroup/glTF`. Accessed: 2020-11-16.

[51] Can Gümeli, Angela Dai, and Matthias Nießner. "ROCA: Robust CAD Model Retrieval and Alignment from a Single Image". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 4022–4031.

[52] Agrim Gupta, Piotr Dollar, and Ross Girshick. "LVIS: A dataset for large vocabulary instance segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5356–5364.

[53] Raia Hadsell, Sumit Chopra, and Yann LeCun. "Dimensionality reduction by learning an invariant mapping". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2. IEEE. 2006, pp. 1735–1742.

[54] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[55] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[56] Kaiming He et al. "Mask R-CNN". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2961–2969.

[57] Kaiming He et al. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9729–9738.

[58] Dan Hendrycks et al. "Natural adversarial examples". In: *arXiv preprint cs.LG 1907.07174* 5.6 (2019).

[59] Katherine L Hermann, Ting Chen, and Simon Kornblith. "The origins and prevalence of texture bias in convolutional neural networks". In: *arXiv preprint arXiv:1911.09071* (2019).

[60] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[61] Ruizhen Hu, Manolis Savva, and Oliver van Kaick. "Functionality representations and applications for shape analysis". In: *Computer Graphics Forum*. Vol. 37. 2. Wiley Online Library. 2018, pp. 603–624.

[62] Ruizhen Hu et al. "Learning to predict part mobility from a single static snapshot". In: *ACM Trans. Graph. (SIGGRAPH Asia)* 36.6 (2017), Article 227.

[63] Yi Huang et al. "Cognitive basis for the development of aesthetic preference: Findings from symmetry preference". In: *Plos one* 15.10 (2020), e0239973.

[64] Diane Humphrey. "Preferences in symmetries and symmetries in drawings: Asymmetries between ages and sexes". In: *Empirical Studies of the Arts* 15.1 (1997), pp. 41–60.

[65] Qasim Iqbal and Jake K Aggarwal. "Retrieval by classification of images containing large manmade objects using perceptual grouping". In: *Pattern recognition* 35.7 (2002), pp. 1463–1479.

[66] Ajinkya Jain et al. "ScrewNet: Category-independent articulation model estimation from depth images using screw theory". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 13670–13677.

[67] Hanxiao Jiang et al. "OPD: Single-view 3D Openable Part Detection". In: *ECCV*. 2022.

[68] HeeJae Jun et al. "Combination of Multiple Global Descriptors for Image Retrieval". In: *arXiv preprint arXiv:1903.10663* (2019).

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

[70] Tero Karras, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4401–4410.

[71] Kihwan Kim et al. "A lightweight approach for on-the-fly reflectance estimation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 20–28.

[72] Sungyeon Kim et al. "Proxy Anchor Loss for Deep Metric Learning". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[73] Vladimir G Kim et al. "Shape2Pose: Human-centric shape analysis". In: *ACM Transactions on Graphics (TOG)* 33.4 (2014), pp. 1–12.

[74] Alexander Kirillov et al. "Pointrend: Image segmentation as rendering". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9799–9808.

[75] Natasha Z Kirkham, Jonathan A Slemmer, and Scott P Johnson. "Visual statistical learning in infancy: Evidence for a domain general learning mechanism". In: *Cognition* 83.2 (2002), B35–B42.

[76] Sebastian Koch et al. "Abc: A big cad model dataset for geometric deep learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 9601–9611.

[77] Eliza Kosoy et al. "Exploring exploration: Comparing children with RL agents in unified environments". In: *arXiv preprint arXiv:2005.02880* (2020).

[78] Eliza Kosoy et al. "Learning Causal Overhypotheses through Exploration in Children and Computational Models". In: *Conference on Causal Learning and Reasoning*. PMLR. 2022, pp. 390–406.

[79] Jonathan Krause et al. "3D Object Representations for Fine-Grained Categorization". In: *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*. Sydney, Australia, 2013.

[80] Alex Krizhevsky et al. "Learning multiple layers of features from tiny images". In: (2009).

[81] Nilesh Kulkarni et al. "Articulation-aware Canonical Surface Mapping". In: *CVPR*. 2020.

[82] Weicheng Kuo et al. "Mask2CAD: 3D shape prediction by learning to segment and retrieve". In: *European Conference on Computer Vision*. Springer. 2020, pp. 260–277.

[83] Weicheng Kuo et al. "Patch2CAD: Patchwise Embedding Learning for In-the-Wild Shape Retrieval from a Single Image". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 12589–12599.

[84] Brenden Lake et al. "One shot learning of simple visual concepts". In: *Proceedings of the annual meeting of the cognitive science society*. Vol. 33. 33. 2011.

[85] Barbara Landau, Linda B Smith, and Susan S Jones. "The importance of shape in early lexical learning". In: *Cognitive development* 3.3 (1988), pp. 299–321.

[86] Yann LeCun, Fu Jie Huang, and Leon Bottou. "Learning methods for generic object recognition with invariance to pose and lighting". In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* Vol. 2. IEEE. 2004, pp. II–104.

[87] C. H. Legare. "Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children." In: *Child development* (2012).

[88] Hao Li et al. "Mobility fitting using 4D RANSAC". In: *Computer Graphics Forum.* Vol. 35. 5. Wiley Online Library. 2016, pp. 79–88.

[89] Xiao Li et al. "Modeling surface appearance from a single photograph using self-augmented convolutional neural networks". In: *ACM Transactions on Graphics (TOG)* 36.4 (2017), p. 45.

[90] Xiaolong Li et al. "Category-level articulated object pose estimation". In: *CVPR.* 2020, pp. 3706–3715.

[91] Yangyan Li et al. "Joint Embeddings of Shapes and Images via CNN Image Purification". In: *ACM Trans. Graph.* (2015).

[92] Yingwei Li et al. "Shape-Texture Debiased Neural Network Training". In: *arXiv preprint arXiv:2010.05981* (2020).

[93] Zhengqin Li et al. "Learning to reconstruct shape and spatially-varying reflectance from a single image". In: *SIGGRAPH Asia 2018 Technical Papers.* ACM. 2018, p. 269.

[94] Rob van Lier and Johan Wagemans. "From images to objects: Global and local completions of self-occluded parts." In: *Journal of Experimental Psychology: Human Perception and Performance* 25.6 (1999), p. 1721.

[95] Rob J van Lier, Emanuel LJ Leeuwenberg, and Peter A van der Helm. "Multiple completions primed by occlusion patterns". In: *Perception* 24.7 (1995), pp. 727–740.

[96] Joseph J Lim, Hamed Pirsiavash, and Antonio Torralba. "Parsing ikea objects: Fine pose estimation". In: *Proceedings of the IEEE International Conference on Computer Vision.* 2013, pp. 2992–2999.

[97] Tsung-Yi Lin et al. "Microsoft COCO: Common objects in context". In: *European conference on computer vision.* Springer. 2014, pp. 740–755.

[98] Xingyu Lin et al. "Transfer of view-manifold learning to similarity perception of novel objects". In: *arXiv preprint arXiv:1704.00033* (2017).

[99] Liu Liu et al. "Toward Real-World Category-Level Articulation Pose Estimation". In: *IEEE Transactions on Image Processing* 31 (2022), pp. 1072–1083.

[100] Ziwei Liu et al. "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations". In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* June 2016.

[101] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017).

[102] Christopher G Lucas et al. "When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships". In: *Cognition* 131.2 (2014), pp. 284–299.

[103] Lars Mescheder et al. "Occupancy networks: Learning 3D reconstruction in function space". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4460–4470.

[104] Ben Mildenhall et al. "Nerf: Representing scenes as neural radiance fields for view synthesis". In: *Communications of the ACM* 65.1 (2021), pp. 99–106.

[105] George A Miller. "WordNet: a lexical database for English". In: *Communications of the ACM* 38.11 (1995), pp. 39–41.

[106] Ishan Misra and Laurens van der Maaten. "Self-supervised learning of pretext-invariant representations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6707–6717.

[107] Niloy J Mitra et al. "Structure-aware shape processing". In: *ACM SIGGRAPH 2014 Courses*. 2014, pp. 1–21.

[108] Volodymyr Mnih et al. "Human-level control through deep reinforcement learning". In: *nature* 518.7540 (2015), pp. 529–533.

[109] Kaichun Mo et al. "PartNet: A Large-Scale Benchmark for Fine-Grained and Hierarchical Part-Level 3D Object Understanding". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[110] Kaichun Mo et al. "Where2Act: From pixels to actions for articulated 3D objects". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 6813–6823.

[111] Yair Movshovitz-Attias et al. "No Fuss Distance Metric Learning Using Proxies". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017.

[112] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. "A metric learning reality check". In: *European Conference on Computer Vision*. Springer. 2020, pp. 681–699.

[113] Jonathan D Nelson et al. "Experience matters: Information acquisition optimizes probability gain". In: *Psychological science* 21.7 (2010), pp. 960–969.

[114] Mike Oaksford and Nick Chater. "A rational analysis of the selection task as optimal data selection." In: *Psychological review* 101.4 (1994), p. 608.

[115] Hyun Oh Song et al. "Deep Metric Learning via Lifted Structured Feature Embedding". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[116] Long Ouyang et al. "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27730–27744.

[117] Stephen E Palmer, Karen B Schloss, and Jonathan Sammartino. "Visual aesthetics and human preference". In: *Annual review of psychology* 64 (2013), pp. 77–107.

[118] Jeong Joon Park et al. "DeepSDF: Learning continuous signed distance functions for shape representation". In: *CVPR*. 2019, pp. 165–174.

[119] Keunhong Park et al. "PhotoShape: Photorealistic materials for large-scale shape collections". In: *arXiv preprint arXiv:1809.09761* (2018).

[120] Marc H Pornstein and Sharon J Krinsky. "Perception of symmetry in infancy: The salience of vertical symmetry and the perception of pattern wholes". In: *Journal of experimental child psychology* 39.1 (1985), pp. 1–19.

[121] Meredith B Prevor and Adele Diamond. "Color–object interference in young children: A Stroop effect in children 31/2–61/2 years old". In: *Cognitive development* 20.2 (2005), pp. 256–278.

[122] Charles R Qi et al. "PointNet: Deep learning on point sets for 3D classification and segmentation". In: *CVPR*. 2017, pp. 652–660.

[123] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. "Vision transformers for dense prediction". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 12179–12188.

[124] Nikhila Ravi et al. "Accelerating 3D deep learning with Pytorch3D". In: *arXiv preprint arXiv:2007.08501* (2020).

[125] Jeremy Reizenstein et al. "Common Objects in 3D: Large-Scale Learning and Evaluation of Real-life 3D Category Reconstruction". In: *arXiv preprint arXiv:2109.00512* (2021).

[126] Sam Ringer et al. "Texture bias of CNNs limits few-shot classification performance". In: *arXiv preprint arXiv:1910.08519* (2019).

[127] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.

[128] Amir Rosenfeld, Markus D Solbach, and John K Tsotsos. "Totally looks like-how humans compare, compared to machines". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 1961–1964.

[129] Chaitanya K Ryali, David J Schwab, and Ari S Morcos. "Characterizing and improving the robustness of self-supervised learning through background augmentations". In: *arXiv preprint arXiv:2103.12719* (2021).

[130] Joanna E Scheib, Steven W Gangestad, and Randy Thornhill. "Facial attractiveness, symmetry and cues of good genes". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 266.1431 (1999), pp. 1913–1917.

[131] 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*. 2015, pp. 815–823.

[132] E. Schulz et al. "Searching for rewards like a child means less generalization and more directed exploration." In: *Psychological science* 30 (2019), pp. 1561–1572.

[133] L.E. Schulz. "The Origins of inquiry: Inductive inference and exploration in early childhood." In: *Trends in Cognitive Sciences* (2012).

[134] L.E. Schulz and E. B. Bonawitz. "Serious fun: Preschoolers play more when evidence is confounded." In: *Developmental Psychology,* (2007).

[135] Meng-Li Shih et al. "3d photography using context-aware layered depth inpainting". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8028–8038.

[136] Arjun Singh et al. "BigBird: A large-scale 3D database of object instances". In: *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2014, pp. 509–516.

[137] Lauren K Slone, David S Moore, and Scott P Johnson. "Object exploration facilitates 4-month-olds' mental rotation performance". In: *PLoS One* 13.8 (2018), e0200468.

[138] Linda Smith and Michael Gasser. "The development of embodied cognition: Six lessons from babies". In: *Artificial life* 11.1-2 (2005), pp. 13–29.

[139] Linda B Smith et al. "Object name learning provides on-the-job training for attention". In: *Psychological science* 13.1 (2002), pp. 13–19.

[140] Kasey C Soska, Karen E Adolph, and Scott P Johnson. "Systems in development: motor skill acquisition facilitates three-dimensional object completion." In: *Developmental psychology* 46.1 (2010), p. 129.

[141] Kasey C Soska and Scott P Johnson. "Development of three-dimensional object completion in infancy". In: *Child development* 79.5 (2008), pp. 1230–1236.

[142] Stefan Stojanov et al. "Incremental object learning from contiguous views". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 8777–8786.

[143] Chen Sun et al. "Revisiting unreasonable effectiveness of data in deep learning era". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 843–852.

[144] Xingyuan Sun et al. "Pix3D: Dataset and methods for single-image 3D shape modeling". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 2974–2983.

[145] Roman Suvorov et al. "Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2021. Resolution-robust Large Mask Inpainting with Fourier Convolutions". In: *arXiv preprint arXiv:2109.07161* (2021).

[146] Sasha Targ, Diogo Almeida, and Kevin Lyman. "Resnet in resnet: Generalizing residual architectures". In: *arXiv preprint arXiv:1603.08029* (2016).

[147] Maxim Tatarchenko et al. "What Do Single-view 3D Reconstruction Networks Learn?" In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3405–3414.

[148] J Arthur Thomson. *On growth and form*. 1917.

[149] Romal Thoppilan et al. "Lamda: Language models for dialog applications". In: *arXiv preprint arXiv:2201.08239* (2022).

[150] Tristan Thrush et al. "Winoground: Probing vision and language models for visio-linguistic compositionality". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5238–5248.

[151] Yonglong Tian, Dilip Krishnan, and Phillip Isola. "Contrastive multiview coding". In: *arXiv preprint arXiv:1906.05849* (2019).

[152] Hugo Touvron et al. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971* (2023).

[153] Shubham Tulsiani et al. "Multi-view supervision for single-view reconstruction via differentiable ray consistency". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2626–2634.

[154] Christopher W Tyler. "Empirical aspects of symmetry perception". In: *Spatial Vision* 9.1 (1995), pp. 1–8.

[155] C. Wah et al. *The Caltech-UCSD Birds-200-2011 Dataset*. Tech. rep. CNS-TR-2011-001. California Institute of Technology, 2011.

[156] He Wang et al. "Normalized object coordinate space for category-level 6D object pose and size estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2642–2651.

[157] Xiaogang Wang et al. "Shape2Motion: Joint analysis of motion parts and attributes from 3D shapes". In: *CVPR*. 2019, pp. 8876–8884.

[158] Xun Wang et al. "Multi-similarity loss with general pair weighting for deep metric learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5022–5030.

[159] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[160] Fangyin Wei et al. "Self-supervised Neural Articulated Shape and Appearance Models". In: *CVPR*. 2022, pp. 15816–15826.

[161] Yijia Weng et al. "CAPTRA: Category-level pose tracking for rigid and articulated objects from point clouds". In: *CVPR*. 2021, pp. 13209–13218.

[162] Olivia Wiles and Andrew Zisserman. "Silnet: Single-and multi-view reconstruction by learning from silhouettes". In: *arXiv preprint arXiv:1711.07888* (2017).

[163] Zhirong Wu et al. "Unsupervised feature learning via non-parametric instance-level discrimination". In: *arXiv preprint arXiv:1805.01978* (2018).

[164] Fanbo Xiang et al. "SAPIEN: A SimulAted Part-based Interactive ENvironment". In: *CVPR*. 2020, pp. 11097–11107.

[165] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. "Beyond pascal: A benchmark for 3D object detection in the wild". In: *IEEE winter conference on applications of computer vision*. IEEE. 2014, pp. 75–82.

[166] Yu Xiang et al. "Objectnet3D: A large scale database for 3D object recognition". In: *European Conference on Computer Vision*. Springer. 2016, pp. 160–176.

[167] Yang Xiao et al. "Pose from shape: Deep pose estimation for arbitrary 3D objects". In: *British Machine Vision Conference (BMVC)*. 2019.

[168] Haozhe Xie et al. "Pix2Vox: Context-aware 3D reconstruction from single and multi-view images". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 2690–2698.

[169] Xianghao Xu et al. "Unsupervised Kinematic Motion Detection for Part-segmented 3D Shape Collections". In: *ACM Trans. Graph. (SIGGRAPH)* (2022), pp. 1–9.

[170] Xinchen Yan et al. "Perspective Transformer Nets: Learning single-view 3D object reconstruction without 3D supervision". In: *Advances in neural information processing systems*. 2016, pp. 1696–1704.

[171] Zihao Yan et al. "RPM-Net: Recurrent prediction of motion and parts from point cloud". In: *ACM Trans. Graph. (SIGGRAPH Asia)* 38.6 (2019).

[172] Gengshan Yang et al. "LASR: Learning articulated shape reconstruction from a monocular video". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 15980–15989.

[173] Wenjie Ye et al. "Single Image Surface Appearance Modeling with Self-augmented CNNs and Inexact Supervision". In: *Computer Graphics Forum*. Vol. 37. 7. Wiley Online Library. 2018, pp. 201–211.

[174] Li Yi et al. "A scalable active framework for region annotation in 3D shape collections". In: *ACM Transactions on Graphics (ToG)* 35.6 (2016), pp. 1–12.

[175] Li Yi et al. "Deep part induction from articulated object pairs". In: *ACM Trans. Graph. (SIGGRAPH Asia)* (2018).

[176] Eunice Yiu, Jasmine Collins, and Alison Gopnik. "Three-Dimensional Object Completion in Humans and Computational Models". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 44. 44. 2022.

[177] Dahlia W Zaidel, Shawn M Aarde, and Kiran Baig. "Appearance of symmetry, beauty, and health in human faces". In: *Brain and cognition* 57.3 (2005), pp. 261–263.

[178] Andrew Zhai and Hao-Yu Wu. "Classification is a Strong Baseline for Deep Metric Learning". In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (BMVC)*. 2019.

[179] Xiuming Zhang et al. "Learning to reconstruct shapes from unseen classes". In: *Advances in Neural Information Processing Systems*. 2018, pp. 2257–2268.

[180] Qingnan Zhou and Alec Jacobson. "Thingi10k: A dataset of 10,000 3D-printing models". In: *arXiv preprint arXiv:1605.04797* (2016).