

# Physical Primitive Decomposition

Zhijian Liu<sup>1</sup>, William T. Freeman<sup>1,2</sup>, Joshua B. Tenenbaum<sup>1</sup>, and Jiajun Wu<sup>1</sup>

<sup>1</sup> Massachusetts Institute of Technology

<sup>2</sup> Google Research

**Abstract.** Objects are made of parts, each with distinct geometry, physics, functionality, and affordances. Developing such a distributed, physical, interpretable representation of objects will facilitate intelligent agents to better explore and interact with the world. In this paper, we study physical primitive decomposition—understanding an object through its components, each with physical and geometric attributes. As annotated data for object parts and physics are rare, we propose a novel formulation that learns physical primitives by explaining both an object’s appearance and its behaviors in physical events. Our model performs well on block towers and tools in both synthetic and real scenarios; we also demonstrate that visual and physical observations often provide complementary signals. We further present ablation and behavioral studies to better understand our model and contrast it with human performance.

## 1 Introduction

Humans use a hammer by holding its handle and striking its head, not vice versa. In this simple action, people demonstrate their understanding of functional parts [33, 39]: a tool, or any object, can be decomposed into primitive-based components, each with distinct physics, functionality, and affordances [17].

How to build a machine of such competency? In this paper, we tackle the problem of physical primitive decomposition (PPD)—explaining the shape and the physics of an object with a few shape primitives with physical parameters. Given the hammer in Figure 1, our goal is to build a model that recovers its two major components: a tall, wooden cylinder for its handle, and a smaller, metal cylinder for its head.

For this task, we need a physical, part-based object shape representation that models both object geometry and physics. Ground-truth annotations for such representations are however challenging to obtain: large-scale shape repositories like ShapeNet [8] often have limited annotations on object parts, let alone physics. This is mostly due to two reasons. First, annotating object parts and physics is labor-intensive and requires strong domain expertise, neither of which can be offered by current crowdsourcing platforms. Second, there exist intrinsic ambiguity in the ground truth: it is impossible to precisely label underlying physical object properties like densities from only images or videos.

Let’s think more about what these representations are for. We want our object representation to faithfully encode its geometry; therefore, it should be able to explain our visual observation of the object’s appearance. Further, as



































