Learning Generalizable Final-State Dynamics of 3D Rigid Objects

Davis Rempe Srinath Sridhar He Wang Leonidas J. Guibas
Stanford University

Abstract

Humans have a remarkable ability to predict the effect of physical interactions on the dynamics of objects. Endowing machines with this ability would allow important applications in areas like robotics and autonomous vehicles. In this work, we focus on predicting the final-state dynamics of 3D rigid objects, in particular an object's final resting position and total rotation when subjected to an impulsive force. Our approach generalizes to unseen object shapes—an important requirement for real-world applications. To achieve this, we represent object shape as a 3D point cloud that is used as input to a neural network, making our approach agnostic to appearance variation. The design of our network is informed by an understanding of physical laws. We train our model with data from a physics engine that simulates the dynamics of a large number of shapes. Experiments show that we can accurately predict the resting position and total rotation for unseen object geometries.

1. Introduction

Humans have a fundamental intuitive understanding of the dynamics of the physical world. Even at a young age, we are able to understand and predict the effect of physical interactions with objects. This *intuitive knowledge of dynamics* allows us to operate in previously unseen environments, and interact with and manipulate objects encountered for the first time. Endowing machines with the same ability would allow new applications in autonomous driving, home robotics, and augmented reality (AR) scenarios.

The 3D dynamics of objects can be predicted using well-studied physical laws given precise properties and system parameters (*e.g.*, mass, moment of inertia, applied force). In practice however, it is impossible to estimate all system parameters, especially from non-contact sensory data. Furthermore, simulating the physics of complex environments requires exact specification of a partially-observed system, and can be computationally expensive and imprecise.

Inspired by the generalizable ability of humans to intuit object dynamics, we develop a deep learning approach to predict the physical dynamics of unseen 3D rigid objects. Learned dynamics has advantages over traditional simulation as it offers differentiable predictions useful for optimization. There has recently been a lot of interest in learning to predict object dynamics, but a number of limitations remain. First, most prior work lacks the ability to generalize to shapes unseen during training time [4], or lacks scalability [9]. Second, many methods are limited to 2D objects and environments [3, 6] and cannot generalize well to 3D objects. Finally, many methods use images as input [7, 2] which provide only partial shape information and entangle variations in object appearance with physical motion.

Our goal is to learn to predict the final-state dynamics of 3D rigid objects and generalize these predictions to previously unseen object geometries. To this end, we focus on the problem of accurately predicting the final rest state (position and total rotation) of an object (initially stationary on a plane) that has been subjected to an *impulse*—a force causing an instantaneous change in velocity. As a result of this impulse, the object moves along the plane but friction eventually brings it to rest (see Figure 1). This problem formulation has surprisingly many nuances. The motion of an object after an applied impulse depends non-linearly on factors such as its moment of inertia, amplitude of the force, and surface friction. Furthermore, sliding objects could wobble resulting in unpredictable motions. Learning these subtleties in a generalizable way requires a deep understanding of the connection between object shape, mass, and dynamics. Since this problem formulation is well-defined, it allows us to better evaluate shape generalization without worrying about other complex dynamics like collisions. Yet, there are many practical applications, for instance, in tabletop robotics where a differentiable final state prediction is useful for planning actions (*i.e.* applied impulses) [8].

To solve this problem, we present a neural network that takes the shape of an object and additional information about the applied impulse as the input, and predicts the final rest position and *total rotation* undergone throughout the entire motion of the object. We use a 3D point cloud to represent shape and use features extracted by PointNet [10]. This decouples object motion from appearance variation making our method more robust. To train this network, we simulate a large number of household objects

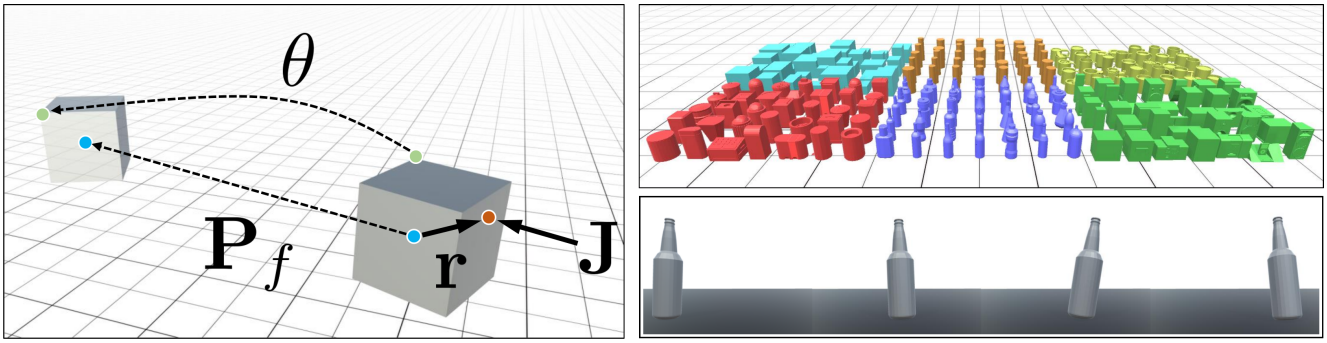

Figure 1. We study the problem of predicting the *position* $\mathbf{P_f}$ and total *rotation* θ of an object initially resting on a plane subjected to an impulse $\mathbf{J}$ at position $\mathbf{r}$ (left). Our method can predict the dynamics of a variety of different shapes (right top) and generalizes to previously unseen object shapes and impulses. One challenge of this is unpredictable 3D motion such as *wobbling* (right bottom).

from the ShapeNet repository [5]. Our network learns to extract salient shape features from these examples. This allows it to make accurate predictions not just for impulses and object shapes seen during training, but also for unseen objects in novel shape categories subjected to new impulses.

2. Problem Formulation

We investigate the problem of predicting the dynamics of an initially stationary rigid object subjected to an impulse. We assume the following **inputs**: (1) the shape of the object in the form of a point cloud ($\mathbf{O} \in \mathbb{R}^{N\times 3}$), and (2) the applied impulse vector and its position. We further assume that the object moves on a plane under standard gravity, the applied impulse is parallel to the plane at the same height as the center of mass, and that all objects have the same friction coefficient, density, and restitution.

Our goal is to accurately **predict** the final rest position ($\mathbf{P}_f \in \mathbb{R}^2$) and the total rotation ($\theta \in \mathbb{R}$) (about the vertical axis) of an object subjected to an impulse. Since the object could undergo multiple 360° rotations before coming to rest, the total rotation θ is often different from the rigid rotation. While we parametrize the final object state with 2 translational and 1 rotational parameters, we *do not* restrict the object motion to 2D. As shown in Figure 1, the object is free to move in 3D as long as it does not topple over. We use a point cloud to encode object geometry since it depends on the surface geometry, making it agnostic to appearance, and can be readily captured in the real-world.

Instead of solving the highly challenging unconstrained 3D dynamics prediction problem, we choose to specifically model 3D motion along a plane and predict final rest state (as opposed to multi-step [9]). This allows us to work on a well-defined problem and to focus on gaining insight and evaluating generalization to unseen object shapes without complex interactions such as collisions. We do not allow the object to roll on its side or to topple over, but varying contact surface area from 3D *wobbling* (see Figure 1) results in complex trajectories. Unobserved quantities (*e.g.*, mass, volume, moment of inertia, and contact surface) additionally contribute to the difficulty of this problem.

3. Data Simulation

We use 3D simulation data from the Bullet physics engine [1]. In each simulation, an object is placed at rest on a flat plane and a random impulsive force is applied parallel to the ground. The object eventually comes to rest. We record the point cloud (1024 points), the magnitude, direction and position of the applied impulse, and the final resting position and total rotation. Friction coefficients and densities are the same for all objects. We only record simulations where the object does not fall over, but motion is not explicitly constrained in any way. Training objects are simulated with a non-uniform random scale to increase shape diversity. Simulated objects travel between 0.5 and 5 meters, and can rotate from 0° to more than 2000° (5–6 rotations).

We synthesized multiple categories of datasets. There are two primitive object datasets: the `Box` and `Cylinders` datasets. There are four datasets which contain everyday objects taken from the ShapeNet [5] repository (`Mugs`, `Bottles`, `Trashcans`, and `Speakers`). These exhibit wide shape diversity and offer a more challenging task. Lastly, we have a `Combined` dataset which combines all of the objects from the previous six to create a large and extremely diverse set of shapes. In total, we use **793** distinct object shapes and ran **98826** simulations.

4. Method

To predict final rest position and total rotation after an impulse, we use a neural network trained on simulated data. We take a principled approach and inform the design of our network based on our understanding of physical laws and priors. We observe that the linear and angular velocities depend on: (1) the applied impulse ($\mathbf{J}$) magnitude, direction,

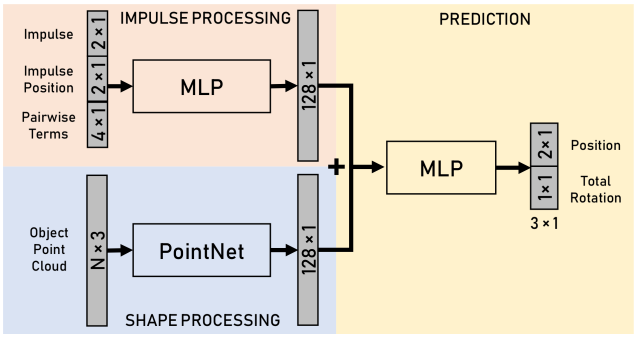

Figure 2. Model architecture. Our network predicts the final resting position and total rotation for a sliding object. + is concatenation and MLP indicates multilayer perceptron.

position ($\mathbf{r}$), and its angular impulse ($\mathbf{r} \times \mathbf{J}$), and (2) the shape of the object which affects its mass and moment of inertia. We therefore base our network design on learning important information related to the applied impulse and shape of the object. Our network (see Figure 2) is composed of two main branches whose output features are jointly used to make a final position and rotation prediction.

Impulse Processing: The top branch in our network is the impulse processing branch which takes the applied impulse, its position, and 4 pairwise terms as input, and outputs an *impulse feature*. The 4 pairwise terms are the products of the components of the impulse with those of the impulse position $\mathbf{r}$; this helps the network learn the difficult cross product $\mathbf{r} \times \mathbf{J}$. The aim of this branch is to learn the effect of the impulse and the angular impulse on the motion of the object producing a final *impulse feature*.

Shape Processing: The bottom shape processing branch is designed to extract salient shape features that are crucial to making accurate predictions. Object geometry affects both linear and angular velocities so the network must develop notions of volume, mass, and inertia from a point cloud. We use PointNet [10] to effectively learn this. As shown in Figure 2, the object point cloud is fed to the PointNet classification network which outputs a global feature that is further processed to output our final *shape feature*.

Prediction: After concatenating the impulse and shape features, we jointly predict final position and total rotation with a 6-layer multilayer perceptron (MLP).

Loss Functions: The goal of the network is to minimize the error between the predicted and ground truth position and rotation. We propose a relative loss: for translation we penalize the relative distance between the predicted final position $\hat{\mathbf{P}}_\mathbf{f}$ and ground truth $\mathbf{P}_\mathbf{f}$: $\mathcal{L}_p = \|\mathbf{P}_\mathbf{f} - \hat{\mathbf{P}}_\mathbf{f}\|/\|\mathbf{P}_\mathbf{f}\|$. For rotation we use a relative L^1 error between the predicted total rotation $\hat{\theta}$ and the ground truth θ (and sum the denominator for numerical stability): $\mathcal{L}_\theta = (|\hat{\theta} - \theta|)/(|\hat{\theta}| + |\theta|)$. Our final loss is the sum of the two $\mathcal{L} = \mathcal{L}_p + \mathcal{L}_\theta$.

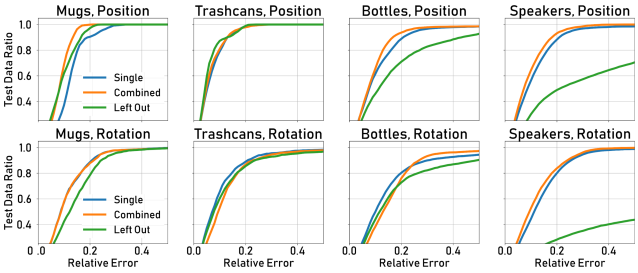

Figure 3. Comparison of performance on single object categories (blue), the full `Combined` dataset (orange), and `Combined` dataset with the evaluated category left out (green). Curves show cumulative fraction of test examples under a certain error.

5. Experiments

Evaluation Metrics: For all experiments, we report mean relative errors for position and total rotation. For position, we use the same relative error used for the loss. For rotation, we report a *binned* relative error $\eta_\theta = (|\hat{\theta} - \theta|/b)/(\lceil|\theta|/b\rceil)$ where θ is the ground truth total rotation and $\hat{\theta}$ is the prediction. For all results we use a bin of $b = 30°$. This metric prevents relative rotation error unfairly increasing when ground truth rotation is near zero.

5.1. Object Generalization

We perform object generalization experiments to evaluate whether the learned model is able to apply accurate dynamics predictions to unseen objects—a crucial ability for autonomous systems in unseen environments. For these experiments, we split datasets based on unique objects such that **no test objects are seen during training**. Furthermore, the impulses applied for test objects are disjoint from those in the training set. We evaluate models trained on both single and combined categories.

Single Category: Results when a separate network is trained for each object category are summarized by the blue curves in Figure 3. For position, around 90% of predictions for all object categories fall under 20% relative error, while for rotation this number falls closer to 80-85%. This indicates that the network is able to generalize to unseen objects within the same shape category.

Combined Categories: Performance when training on the `Combined` dataset then evaluating on all individual datasets is shown by the orange curves in Figure 3. Performance is similar to training on individual datasets and even improves errors in some cases. This indicates that exposing the network to larger shape diversity at training time can focus learning on underlying physical relationships rather than properties of a single or small group of objects. In order to maintain this high performance, the network is likely learning a general approach to extract salient physical features from the diverse objects in the `Combined` dataset

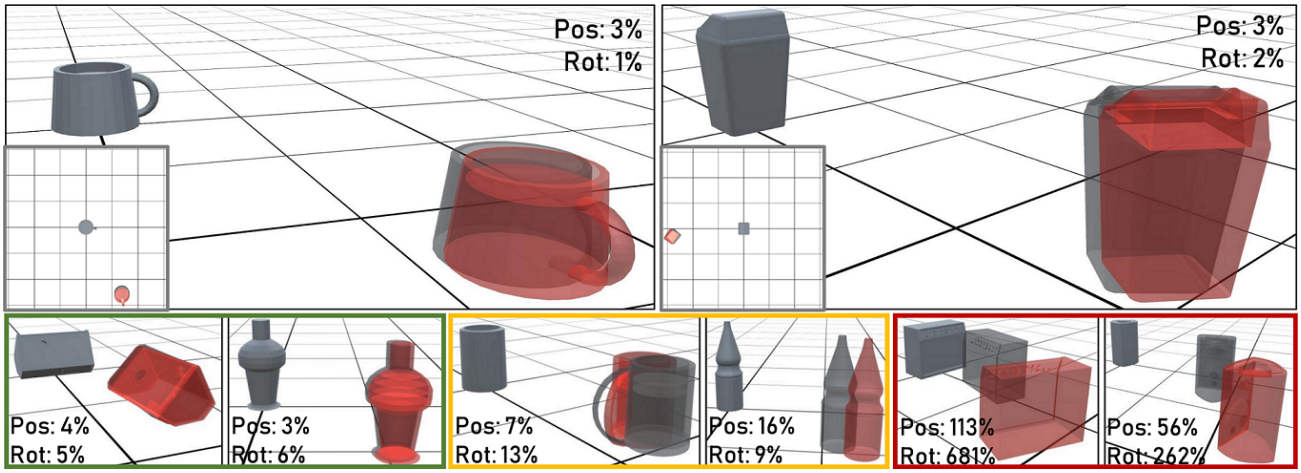

Figure 4. Sample predictions from models trained on the `Combined` dataset with one category left out. Initial object state is shown in shaded grey, ground truth final state is in transparent grey, and network prediction is in transparent red. Relative errors are shown.

rather than just memorizing how specific shapes behave.

Out of Category: Lastly, we evaluate performance on the extreme task of generalizing **outside of trained object categories**. For this, we create a new combined datasets each with one object category left out of the training set. We then evaluate its performance on objects from the left out category. Results for these experiments are shown by the green curves in Figure 3. The network is able to achieve good results on all left-out object categories except for `Speakers`. `Speakers` contributes the most unique objects to the `Combined` dataset by far (368 unique objects); without them, the network may not see enough diversity in training to perform well. Overall, this result shows that we can still make accurate predictions for objects from completely different categories in spite of their shape not being close to the trained objects. Some predictions from leave-one-out trained models are visualized in Figure 4.

6. Limitations and Future Work

In this work, we took a different approach by predicting the final state of a 3D rigid object instead of multi-step predictions. Future work should consider closing the loop by predicting both the final state as well as multiple intermediate states. We ignore the physical parameter estimation problem and assume constant friction and density. We also ignore free 3D dynamics and complex phenomena such as collisions which are important directions for future work. We believe that our approach provides a strong foundation for developing methods for these complex motions.

7. Conclusion

We presented a method for learning to predict the final position and total rotation of a 3D rigid object subjected to an impulse and moving along a plane. Our neural network model is capable of generalizing to previously unseen object shapes by operating directly on 3D point clouds.

References

[1] Bullet physics engine. https://pybullet.org. 2

[2] P. Agrawal, A. Nair, P. Abbeel, J. Malik, and S. Levine. Learning to poke by poking: Experiential learning of intuitive physics. In *NIPS*, 2016. 1

[3] P. Battaglia, R. Pascanu, M. Lai, D. J. Rezende, and K. kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *NIPS*, 2016. 1

[4] A. Byravan and D. Fox. Se3-nets: Learning rigid body motion using deep neural networks. In *ICRA*, 2017. 1

[5] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[6] M. B. Chang, T. Ullman, A. Torralba, and J. B. Tenenbaum. A compositional object-based approach to learning physical dynamics. In *ICLR*, 2017. 1

[7] C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In *NIPS*, 2016. 1

[8] M. Janner, S. Levine, W. T. Freeman, J. B. Tenenbaum, C. Finn, and J. Wu. Reasoning about physical interactions with object-oriented prediction and planning. In *ICLR*, 2019. 1

[9] D. Mrowca, C. Zhuang, E. Wang, N. Haber, L. Fei-Fei, J. B. Tenenbaum, and D. L. K. Yamins. Flexible neural representation for physics prediction. In *NIPS*, 2018. 1, 2

[10] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CVPR*, 2017. 1, 3