Visual Assessment for Non-Disruptive Object Extraction

Sarthak Ahuja

CMU-RI-TR-20-28

July 2020

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Henny Admoni (Co-Advisor)
Aaron Steinfeld (Co-Advisor)
Oliver Kroemer
Senthil Purushwalkam Shiva Prakash

*Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Robotics.*

Keywords: Intuitive Physics, Visual Assessment, Non-Disruptive Object Extraction

To the year 2020, for showing me how unexpected life can be; and to my family, friends, and mentors for reminding me that everything will still be okay.

Abstract

Robots operating in human environments need to perform a variety of dexterous manipulation tasks on object arrangements that have complex physical support relationships, e.g. procuring utensils from a large pile of dishes, grabbing a bottle from a stuffed fridge, or fetching a book from a loaded shelf. The cost of a misjudged extraction in these situations can be very high (e.g., other objects falling) and therefore robots must be careful not to disturb other objects when executing manipulation skills. This requires robots to reason about the effect of their manipulation choices by accounting for the support relationships among objects in the scene. Humans do this in part by visually assessing the scene and using physics intuition to infer how likely it is that a particular object can be safely moved. Inspired by this human capability, we explore how robots can emulate similar vision-based physics intuition using deep learning based data-driven models.

We formulate our research problem as a scene understanding task for visually assessing the feasibility of extraction from an arrangement of objects. We focus on data-driven approaches that assess possible object interactions with only a few glimpses of the scene. Ongoing work has shown that deep convolutional neural networks can learn intuitive physics over images generated in simulation and determine the stability of an arrangement of objects in the real world. We extend these physics intuition models to the task of assessing safe object extraction by conditioning the visual images on specific objects in the scene using object masks. Our method identifies which objects can be safely extracted, from which direction to extract them, and the potential impact such extraction will have on nearby objects. Our results, in both simulation and real-world settings, show that physics intuition models using our proposed method can successfully inform a robot's actions during object extraction. We compare the performance of our method against simulation-based and geometry-based assessment methods and highlight their pros and cons for their application to the task of assessing safe object-extraction. Furthermore, we show that using aggregation techniques to combine multiple views, we can obtain a unified visual assessment that improves the model's predictive performance.

Acknowledgments

I would like to thank my advisors - Prof. Henny Admoni and Prof. Aaron Steinfeld, for giving me the opportunity to work with them. I am grateful for their unwavering dedication towards teaching me the ins and outs of academic research and concurrently optimizing for my professional growth and personal well-being. I am also thankful to my other committee members Prof. Oliver Kroemer and Senthil, for taking the time to give me helpful feedback on my research.

I consider myself fortunate for having worked in two labs among some incredibly smart and kind researchers. I am thankful to my labmates Abhijat, Ada, Allan, Ben, Fern, Kirabo, Liz, Maggie, Mike, Oscar, Pallavi, Reuben, Sam, Stephanie, Tesca, Yaxin and Zhi for their regular and useful feedback on the work herein and for just being amazing people to have shared the lab with. On the same note, I would like to thank my close friend and colleague Aditya Agarwal for helping me collect real-world data and run the simulation-based baselines used in this thesis.

It has been a very challenging year for everyone in the world because of the global pandemic. It took many support relationships in real life to help me complete this thesis on assessing physical support relationships among objects. I am indebted to my family for their continual encouragement and apologize for my limited availability. I want to thank my friends back home Aditi, Amol, Anchita, Manan, Swastic, Sharan, and Shreya for motivating me to remain positive and productive. Equally important were my self-effacing friends here in Pittsburgh Hitesh, Neha, Talha, and Tanmay, who actually made sure that I do. Finally, I am grateful to Silvia for helping me in every way possible a person can from half-way across the world.

This work was supported by the Office of Naval Research under award number N00014-18-1-2503.

Contents

List of Figures

List of Tables

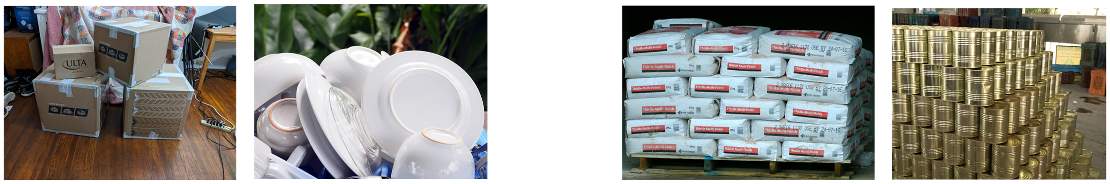
Chapter 1

Introduction

1.1 Motivation

Robots operating in human environments need to perform a variety of dexterous manipulation tasks on object arrangements that have complex support relationships, e.g. procuring utensils from a large pile of dishes, grabbing a bottle from a stuffed fridge, and fetching a book from a loaded shelf. Such arrangements are common occurrences in human spaces and can be either accidental (Figure 1.1a) or deliberate (Figure 1.1b). The cost of a misjudged extraction in these situations can be very high as it may lead to the other objects falling and being damaged. Therefore, robots must be capable of assessing whether an object can be extracted without changing the configuration of adjacent objects.

Humans are proficient at safe manipulation in part because they can visually assess their surroundings and use physics intuition to judge how likely it is that a particular object can be manipulated safely, i.e., without causing disruption to the configuration of other objects in the scene. This allows humans to make efficient, yet safe judgment calls. For example, instead of clearing the entire stack of cans in Figure 1.1b to get to a desired lower level can, humans can infer the likelihood of success for directly extracting the target can, thereby enabling a cost-benefit assessment on different approaches. Similarly, in the presence of a multitude of options around which can to extract, humans can selectively avoid cans that are critical to supporting the arrangement. We may be hopeless at explaining our exact reasoning as to how the center of mass of the pile shifts after the extraction, but we can make remarkably accurate judgments with just a few glances at a scene.



(a) Accidental arrangements of objects [6].

(b) Deliberate arrangements of objects [4, 5].

Figure 1.1: Can robots **visually assess** which objects from these common object arrangements can be **safely extracted**, i.e., without disturbing the configuration of the remaining scene?

Existing research in psychology explores this phenomenon and points towards an intuitive physics engine that each human implicitly develops over time as we experience and understand interactions in the world [11]. Human intuitions are initially imperfect but significantly improve with experience and acquired knowledge [21]. While this is largely subjective, our daily interactions with the real world help us develop a basic physics intuition that makes us remarkably efficient at common manipulation tasks. In particular, this human generalized physics model of the world helps us in quickly filtering out unsafe and dangerous manipulation choices, particularly for assessing the stability of a scene. The game of Jenga [7] is a classic example where we witness a playful display of this skill as human players carefully pick only a few potentially safe candidate blocks to extract. This model, in turn, makes humans smart responders by making their actions more targeted, yet safe at the same time.

As robots encounter dynamic object arrangements in human spaces and are increasingly expected to be capable of common-sense physics reasoning during their decision making, their ability to assess a scene, reason about which parts of the scene can be safely manipulated using current skills, and communicate this assessment to a human becomes an important part of their autonomy. In this work, we develop and show that vision-based physics intuition improves robot attempts to safely extract objects using only a few initial glimpses of a scene.

1.2 Problem Description

Object extraction has been explored in great depth in the robotics community [16, 25, 49]. Interestingly, a common assumption made in many of these works is that the state of the environment can be changed or even ignored while extracting the target object; i.e., obstacles in the path of a robot to the target object can be cleared or disregarded as static obstacles. This is because manipulation tasks here assume either objects are present in physical isolation or lying on a planar surface where there is no need to prevent objects from falling. As discussed in the previous section, this assumption does not necessarily hold when operating in human spaces and disrupting the scene arbitrarily is not always efficient or safe. Hence we seek to answer the question: "*How can robots safely perform efficient manipulation skills on object arrangements displaying complex support relationships?*" We particularly seek to find solutions that are generalizable across a range of manipulation tasks and can be easily adopted by existing robots.

In the domain of safe manipulation, there exists research that uses specialized policies learned by robots equipped with multiple sensors to actively evaluate the stability of the scene for such non-disruptive object extraction. Early work [36, 65] uses force sensors and cameras to register large motions in a Jenga tower and evaluate if applying a pulling action on a random block is disruptive. Fazeli et al. [19] takes this idea one step further and learns over these visual and tactile measurements in the real world to identify which blocks are hard or easy to move while performing the block extraction. While these approaches show remarkable proficiency at object extraction, the interactive policy used is very specific to the robot and sensors and not easily transferable to other robots and tasks. It would be extremely useful if these robots could still assess a scenario using commonsense physics reasoning irrespective of their existing capabilities and across different tasks. As RGB-D cameras are one of the most common sensors available in most robotic systems, we seek methods that can use RGB-D images to assess the stability of an arrangement of objects to provide a judgment about questions like: *"Will the scene remain stable upon extracting this object?"*, *"In which direction should we extract the object?"*, and *"Which objects will move if I extract the object?"*. We refer to the act of making these judgments as "**Visual Assessment**".

We assume a risk-averse and conservative safe object extraction setting where a robot has two objectives: (1) procuring an object and (2) not disturbing the configuration of other objects in the environment. An interesting approach is to separate the models for each of these objectives, independently assessing the primary objective of procuring an object and the secondary objective of adhering to environmentally imposed safety constraints. [10]. While the primary objective can be carried out in multiple ways depending on the capabilities of the robot, it is the role of these independent models to guide the robot in selecting a way that also adheres to the secondary objective of safety. We refer to these independent models that capture the inherent physics constraints of the system as "**Physics Intuition (PI)**" models. This is in contrast to the previously described approaches that incorporate the safety constraint directly in their learned policy.

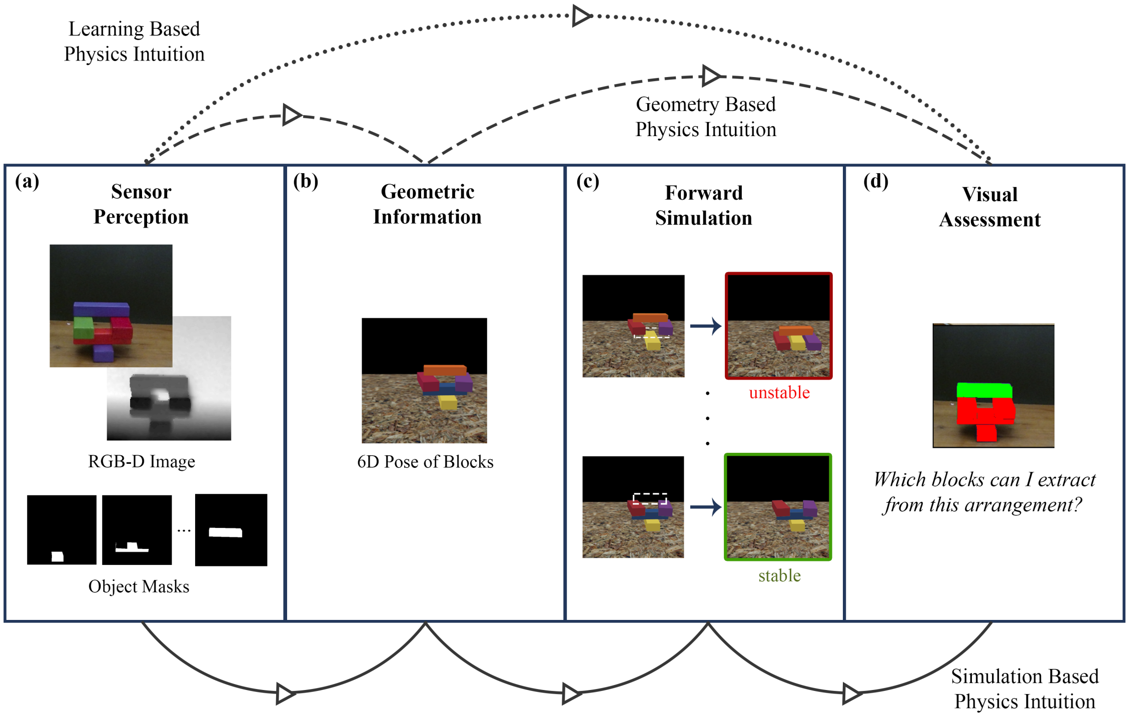


Figure 1.2: Visual assessment can be represented as a sequence of steps (boxes): (a) perceive the sensor data, (b) obtain the geometric information of the objects, (c) simulate the extraction and (d) make the assessment. Approaches towards creating physics intuition models can be classified on the basis of the subsequence of steps (arrows) explicitly carried out in the pipeline.

Exploring independent physics intuition models for assessing safe manipulation is currently an active avenue of robotics research. We can organize existing approaches into three broad sequential categories, as depicted in Figure 1.2.

1. **Simulation-Based Physics Intuition**: A number of methods use physics simulators. A physics simulator is a computer algorithm to solve dynamic equations and predict the future states of objects by performing the computations on discretized real-world quantities [18]. at test time to model the interaction and transition of states among objects. By performing a roll-out of their intended manipulation skill, robots can infer the cost associated with its execution. While these methods are highly accurate, they rely on having a robust perception pipeline to gain full geometric information of a scene (6-D poses of objects from a fixed frame of reference) and further require access to physics simulators to perform repeated roll-outs during test time.
2. **Geometry-Based Physics Intuition**: Instead of performing roll-outs using a physics simulator, some methods perform a static analysis of the perceived arrangement of objects using geometric reasoning. These geometric rules are defined by hand and are often specific to the characteristics of the objects in the scene. Similar to simulation-based assessment, these methods rely on having full geometric scene information. In contrast, these methods do not need to perform a test-time simulation.
3. **Learning-Based Physics Intuition**: Contrary to the previous assessments, learning-based methods are data-driven and try to directly map observations to predictions around stability or spatial relationships. Although these methods often require manually labeled datasets and/or hand-crafted features from a perception pipeline to obtain a sufficient level of accuracy and generalizability, they show great potential to provide the quick and robust inference that is necessary for real-world assessment. This is primarily because these methods are capable of operating without any explicit physical simulation or precise geometric information of all of the objects in the scene at test time.

In this work, we explore a learning-based approach for creating physics intuition models using deep neural networks trained in simulation. More recently, due to the emergence of realistic physics engines and simulators, generating labeled synthetic training data has become increasingly convenient. Simultaneously, due to the rising effectiveness of deep neural networks, learning robust representations of physical scenes that bolster higher-level understanding of object relationships has become an active area of research. In the last few years, researchers have employed deep neural networks to approximate physics simulators and reason about the stability of a scene directly from visual inputs [23, 37, 38]. These data-driven approaches greatly alleviate the need for explicit object modeling by using the richness of simulated passive observations to approximate the physical properties of complex scenes. In particular, [23, 38] showed that the learned physics intuition models went beyond a passive assessment of stacked towers as stable or unstable. Their approach was actively able to guide stacking tasks and even derive meaningful affordances, such as *stackability* of individual blocks. In contrast, we build deep learning-based physics intuition models that can visually assess complex object arrangements from multiple views and provide a physics intuition around safe object extraction (or *extractability* of each object).

1.3 Contributions

The key contributions of our work are:

1. We introduce a pipeline to implement learning-based physics intuition models for non-disruptive object extraction by adding a target conditioning variable in the form of an object mask while training the models.
2. We specifically explore performing visual assessment around: (i) Which blocks can be safely extracted? (ii) In which direction can they each be extracted? and (iii) Which blocks will be affected if we extract a particular block?
3. We demonstrate the effectiveness of our method in both simulation and real-world settings on a curated multi-view dataset of table-top block scenes consisting of homogeneous rigid cuboid blocks of known mass and dimensions.
4. We compare the performance of our method against simulation-based and geometry-based methods for application to the task of assessing safe object-extraction.
5. We further analyze various aggregation techniques to combine physics intuitions over multiple views to obtain a unified visual assessment.

1.4 Outline

The rest of this thesis is structured as follows: Chapter 2 presents an overview of previous work on physics intuition models using simulation-based, geometry-based, and learning-based techniques. We especially focus on deep learning based methods that operate directly on simulator generated images. In Chapter 3, we formalize our scene understanding problem and sequentially walk through the design of our physics intuition models for performing skill-agnostic ("Which blocks can be safely extracted?") and skill-specific ("In which direction can they be extracted?") visual assessment from multi-views of a scene. We briefly talk about our effect-mask prediction model that provides an intuition around "Which blocks will be affected if we extract a particular block?". In Chapter 4 we dive deep into the methodology around curating our datasets and training our models and in Chapter 5 we present our experimental results evaluating the aforementioned methods and describe the baselines and metrics used. Finally, in Chapter 6 we discuss the limitations of our methods and discuss future work to address them.

Chapter 2

Related Work

2.1 Geometry-Based Methods

Traditionally, researchers have used explicit geometry and kinematics-based techniques to infer support-order among objects in the scene. Kartmann et al. [31] recreates real-world scenes in simulation by fitting primitive objects onto 3D point clouds. Their approach computes support relationships on the obtained primitive shape arrangements using rules defined by Static Equilibrium Assessment (SEA), as proposed by Mojtahedzadeh et al. [41]. Along similar lines, there exist other explicit, rule-based approaches for safe deconstruction of object piles [32, 46]. One drawback of these methods is that the hand-crafted rules are often highly specific to a collection of objects, the range of their physical properties and the expected support relationships. For example, Kartmann et al. [31] adds a rule-based support-polygon assessment on top of the act relations [41] to account for top-down object support relationships. This makes them difficult to scale across different domains without explicitly changing the underlying rules manually. Another obvious drawback of these methods is their reliance on the availability of complete geometric information of the scene (6-D pose of each object) in order to provide accurate inferences.

2.2 Simulation-Based Methods

Another body of work uses on-board physics simulators to roll-out a scene in time to infer the future state of objects after the target object has been manipulated. These methods often need to trade accuracy for inference time [70] and require complete geometric information of the scene. Battaglia et al. [11] use noisy approximate simulations and Bayesian modeling for doing fast physical inference. Another work by Wagner et al. [64] deals with a question quite similar to ours in that it tries to answer "what if" questions for physical scenes. They parse the questions and generate answers using a data-driven model, although the inference of future states directly uses an internal physics engine. Recent work Bejjani et al. [12] employs a physics simulator for look-ahead planning in the image space for constrained manipulation of simple, isolated, planar objects. Moll et al. [42] uses a physics engine to evaluate possible complex multi-body dynamical interactions interactively during test time.

2.3 Learning-Based Methods

2.3.1 Learning Spatial Relationships

Another line of research tries to find the support relationships among objects in a scene directly from images through supervised learning or non-monotonic reasoning over hand-crafted features of individual objects, such as centroid distance, depth boundary, proximity, and containment [41, 47, 57]; 3-D planes [55]; base width, number of blocks, and presence of balanced top [50]; and contact points [51]. Likewise, previous work [35] learns spatial preconditions for manipulation skills by learning computational models using objects, parts and their interactions (represented by their mean 3D position). Other work [24, 30, 53] builds physically plausible scene representations by modeling the world as cuboids and reasoning about the support structure and occluded regions. In contrast to these approaches, we explore learning features directly from simulator-generated images without assuming any intermediate predetermined form.

2.3.2 Image-Based Intuitive Physics

Early work by Lerer et al. [37] uses a feed-forward visual model to predict the stability and falling trajectories for simple block towers from images. Some [23, 38] propose a similar model, but with the purpose of guiding block stacking. Groth et al. [23] sample candidate positions on the surface of an object and guides the construction of the tower by picking the candidate that leads to the highest stability score over the “hallucinated” scene from their learned physics intuition model. Similarly, [38] hallucinates sample candidate positions on the images themselves instead of sampling candidate positions in simulation. This makes exploiting the physics intuition models in the real world much more viable. However, they perform the training as well as candidate sampling on binary-valued foreground masks of the scene, which limits the generalizability of the method to complex, real-world scenes.

Our work differs from [23, 38] in that we try to learn physics intuition models that capture a notion of safe object extraction instead of stacking. Similar to [38], we try to sample candidate objects to remove by directly hallucinating object extractions in the images. However, our approach accounts for the visual complexity of the cluttered scenes by adding a separate conditioning variable in the form of a single object mask alongside the RGB images during training. As with [23], we use multiple views of the scene to make predictions but perform a more comprehensive analysis of various methods to aggregate these predictions in a significantly more occlusion sensitive setting.

There exists a plethora of ongoing research that aims to accurately model the physics dynamics of a scene by framing it as a future object state [14, 20, 66, 67], image frame [29, 69], or object trajectory [13, 33, 43, 44] problem. Our work focuses on using passive observations in the form of images with no access to previous frames. This approach excludes object supervision beyond object masks, and makes object-level predictions of a high-level property (extractability) of the system. Although it is worth noting that [44] also uses object masks to make predictions about the trajectory of an object in the image space. We make predictions about the stability of the remaining scene after removing the target object, rather than predicting the trajectory of the object of interest.

Chapter 3

Approach

3.1 Predicting Plausible Object Extractions

Learning physics intuition models is a supervised learning task. By changing the traditional image class labels to stability labels, we learn a physics intuition model over a large number of images of scenes (i.e., arrangements of objects). We obtain the stability labels (*stable* or *unstable*) for these images by running simulations of an object extraction scenario in a physics simulator. Formally put, the objective is to learn a mapping f that, given an image I of the initial configuration of a scene S (consisting of n objects defined as $\{s_1, s_2 \ldots s_n\}$), can provide the stability prediction $P(S)$, which is a probability value between 0 (unstable) and 1 (stable).

$$f : I(S) \rightarrow P(S) \tag{3.1}$$

3.1.1 Target Object Conditioning

We note that in the above formulation, the model is unable to naturally provide inferences about individual objects in the scene, i.e., it is unable to answer "*What is the stability of the scene after removing object s_i?*". One way to answer this question in the current setting is to remove the corresponding object from the scene and get the inference from the same function mapping f.

$$f : I(S \setminus s_i) \rightarrow P(S \setminus s_i) \tag{3.2}$$

Computing $I(S \setminus s_i)$ at test time is non-trivial as it requires a robot to hallucinate the removal of an object from an image. As a simple solution to resolve this issue and adapt the model for object extraction tasks, we propose to generate only stable scenes to train our physics intuition models and, in each scene, remove a block and obtain the stability label for the resulting configuration. The corresponding segmentation masks of object s_i in image I are defined as $\phi(s_i)$. We add these masks to the above mapping function, thereby conditioning the obtained probability value on the object (i) to which the mask corresponds, as depicted in Figure 3.1.

$$f : I(S|\phi(s_i)) \rightarrow P(S \setminus s_i) \tag{3.3}$$

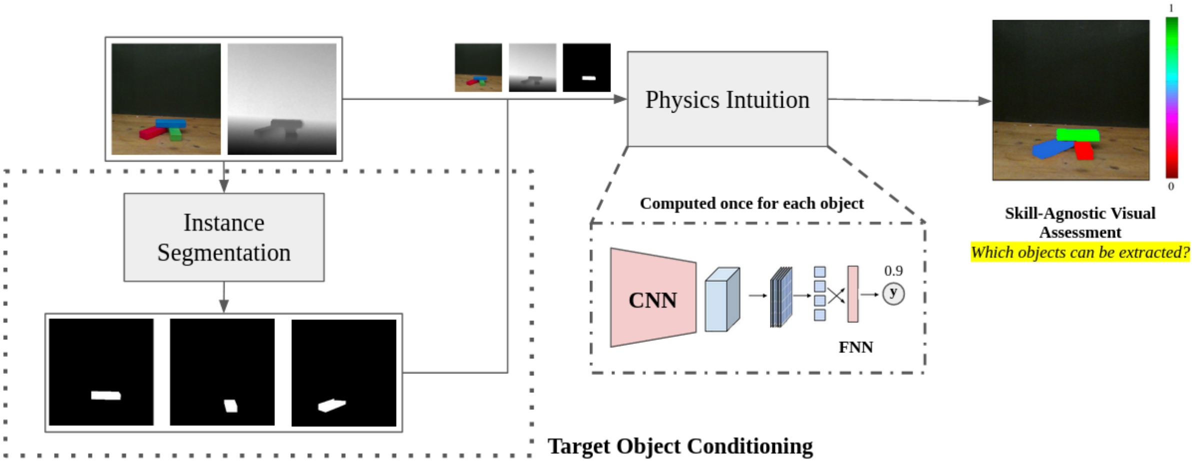

Instance Segmentation
Physics Intuition
Computed once for each object
CNN
0.9
y
FNN
1
0
Skill-Agnostic Visual Assessment
Which objects can be extracted?
Target Object Conditioning

Figure 3.1: The binary mask of a target object is used as an another channel to the RGB-D input. The output corresponds to whether or not the scene remains stable after removing the object.

A target object mask is the only external input passed as an additional channel alongside the RGB-D image to the physics intuition model (making it RGB-DM). Due to the availability of large datasets, high performance GPUs and modern deep learning approaches for instance segmentation [27, 48], state-of-the-art models have become remarkably accurate and easily accessible [72]. In practice, we can obtain the segmentation masks over target objects in real time by fine-tuning existing models on our class of objects. However, in this work we obtain these masks directly from the simulator during training and use Mask R-CNN [27] and color thresholding [39] to aid annotating our real-world dataset in an offline manner.

3.1.2 Aggregation Over Multiple Views

We may obtain different predictions for the same scenario from different camera angles because of occlusion from objects in the scene and the inability of a single 2D image to capture all relevant 3D information in the scene. For camera angle k,

$$f : I_k(S|\phi_k(s_i)) \rightarrow P_k(S \setminus s_i) \tag{3.4}$$

Therefore, it is important to account for multiple views of a scenario and obtain an accurate assessment of the scene in order to generate a single prediction. A common choice for capturing this mapping, f, has been deep convolutional neural networks [23, 37, 38], which consist of a feature extractor module (multiple convolution layers, CNN) followed by a classifier module (multiple fully connected layers, FNN). We explore two ways of performing aggregation over K views of a scenario in the context of these deep convolutional neural networks.

- **Pre-training**: As first proposed in [58], we modify our model architecture during training to compute the feature representations over all available views (regardless of their order) and use view pooling to get an aggregated representation of the scene. This representation is then passed on to the classifier module to make a single prediction. See Figure 3.2.

$$f : \{I_0(S|\phi_0(s_i))...I_K(S|\phi_k(s_i))\} \rightarrow P(S \setminus s_i) \tag{3.5}$$

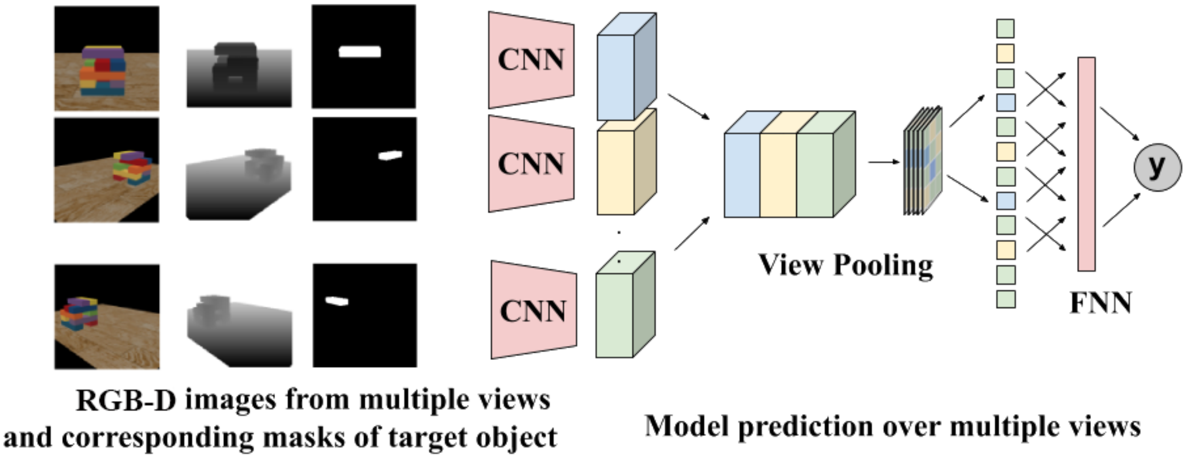

CNN
CNN
CNN
View Pooling
FNN
y
RGB-D images from multiple views
and corresponding masks of target object
Model prediction over multiple views

Figure 3.2: Pre-training multi-view aggregation: A single inference y is made on all views of the scenario. View-pooling is an element-wise maximum operation across the views in the layer.

- **Post-training**: In contrast to pre-training, aggregation takes place on the image features; the other alternative is to aggregate the individual predictions. We use a function mapping g that combines the predictions obtained over multiple views from the existing model f (refer to eq. 3.4) using an aggregation method Ψ. In our case, we evaluate mean, median, mode, maximum and minimum as potential candidates for this aggregation function. See Figure 3.3.

$$g : \Psi(\{P_0(S \setminus s_i)...P_k(S \setminus s_i)\}) \rightarrow P(S \setminus s_i) \tag{3.6}$$

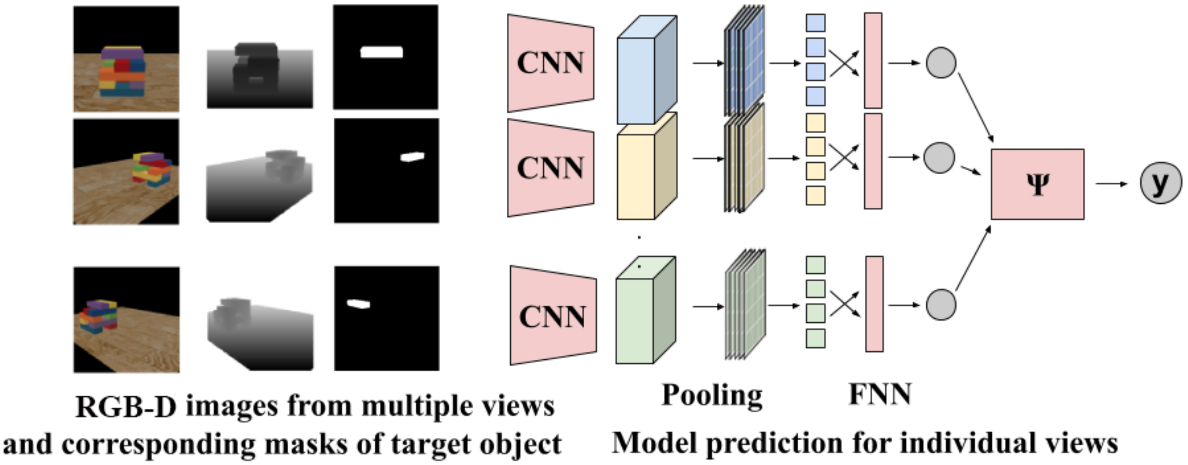

CNN
CNN
CNN
Ψ
y
RGB-D images from multiple views
and corresponding masks of target object
Pooling
FNN
Model prediction for individual views

Figure 3.3: Post-training multi-view aggregation: Inferences for each view are made on the scenario and are aggregated to provide a single inference y.

3.2 Predicting the Extraction Direction

Up to this point, our formulation only accounts for predicting whether a particular object can be safely removed from the scene and does not account for a robot's skill/capabilities; i.e., our inference function f is **skill-agnostic**. We propose to extend this assessment and enable our physics intuition model to answer *"How should the object be extracted from the scene?"*. In order to obtain this **skill-specific** model, we account for the robot's skill during the generation of the stability labels. We parameterize a robot's skill in an object-centric manner; that is, for object extraction, we parameterize a robot's skill as a set of discrete extraction directions. We define 5 discrete skills as *[Extract Up (UP), Extract Forward (FW), Extract Backward (BK), Extract Left (LF), Extract Right (RT)]* with respect to the table in the scene. To avoid visual ambiguity, we choose camera views from only one side of the table that uniquely identify each extraction direction. For learning the skill-specific models, we reformulate our problem from a logistic regression problem with a single label (in the skill-agnostic case) to a multi-variate regression problem with five labels, as shown in Equation 3.7.

$$f : \{I_0(S|\phi_0(s_i))...I_K(S|\phi_k(s_i))\} \rightarrow \begin{bmatrix} P^{UP}(S \setminus s_i) \\ P^{FW}(S \setminus s_i) \\ P^{BK}(S \setminus s_i) \\ P^{LF}(S \setminus s_i) \\ P^{RT}(S \setminus s_i) \end{bmatrix} \tag{3.7}$$

In our simulation experiments, we move the target object (canceling out all forces) in each of the aforementioned 5 directions to obtain the ground truth labels for each skill. The configuration of remaining objects in the scene at the end of simulated extraction will also depend on physical properties of the objects that are not visually perceivable, such as mass and material friction. We assume knowledge of the mass of the blocks and collect the stability labels for each scene using 3 different values of friction coefficients. We then examine two cases: (1) where we assume the knowledge of the friction coefficient and use the ground-truth labels as binary variables, and (2) where we assume a uniform prior over three sampled friction settings and use the ground-truth labels as continuous values between 0 and 1 (Figure 3.4). For example, when evaluating the probability of extracting a block in the forward direction, we obtain the continuous label, as depicted in equation 3.8.

$$y^{FW}(S \setminus s_i) = \sum_{f \in \{low, medium, high\}} y^{FW}(S \setminus s_i | F = f) P(F = f) \tag{3.8}$$

We expect this visual assessment will enable robots to select safe object extraction strategies by re-orienting themselves with respect to the object arrangement or assessing if they are capable of performing the extraction at all. For example, a robot like a Rethink Robotics Baxter may be able to pick up objects robustly but have limited capability to horizontally manipulate objects. In contrast, a Kinova arm may be robust at pushing and pulling objects horizontally but may not be able to pick them effectively from the top of the tower.

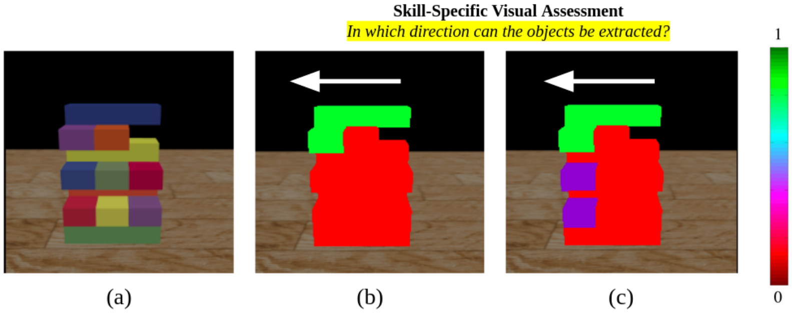

Skill-Specific Visual Assessment
In which direction can the objects be extracted?
1
0
(a)
(b)
(c)

Figure 3.4: Ground truth labels of a Jenga scene for blocks being *extractable* in a particular direction (white arrow) when the material friction of the blocks is (b) known and (c) unknown. We can observe that the ground-truth labels are binary in (b) and continuous in (c).

3.3 Predicting Effect Masks

There may be instances where an object will need to be removed from an arrangement even if it is supporting other objects. In such cases, a robot will need to safely clear or secure the supported objects to make the extraction of the target object safe and feasible. It will be useful for a robot to be able to get physics intuition around *"Which objects will move if this object is removed?"*. We formulate this prediction as being able to predict a single mask (referred to as **Effect-Mask**) around objects that will be disturbed because of an extraction (Figure 3.5). Prior work has looked at generating support graphs [31, 47] that determine the order in which the supported objects should be removed by identifying pair-wise support relationships among objects. As future work, we foresee that predicting such effect-masks can enable generating similar support-graphs as well while operating in the image-space. We present a preliminary analysis using Fully Convolutional Networks [40] for this semantic segmentation task on our simulation dataset.

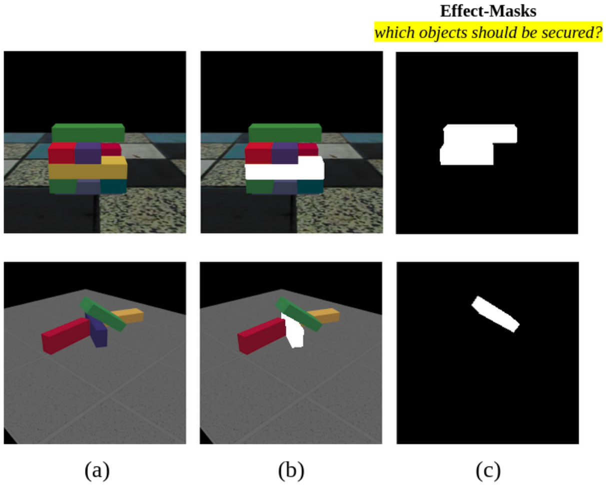


Figure 3.5: Visualizing ground truth effect-masks (c) for object arrangements (a) when a particular target object, highlighted in white, is removed from the scene (b).

Chapter 4

Experiment Setup

4.1 Dataset Generation

A large amount of data is required to train deep convolutional neural networks to learn effectively. Since collecting data in the real-world is an expensive process, a common approach is to use domain knowledge and synthesize data in simulation. Publicly available datasets of stable and unstable scenes [23, 37] generated in simulation are limited to block towers, as they primarily focus on stacking tasks and do not possess some of the characteristic features of a cluttered scene (e.g., objects being supported by multiple objects, objects supporting each other along the plane, etc.). Therefore, to evaluate our approach in a more principled way, we propose a taxonomy of object arrangements and choose a scene type from each proposed category to generate data and report results.

4.1.1 Taxonomy

We propose a categorization of object arrangements based on two factors that we believe affect the learning capacity of physics intuition models:

- **Homogeneous Structure**: This refers to the diversity of support relationships in a scene. An inherent homogeneous structure aids the learning process while multiple kinds of inter-object support relationships in the scene make the learning process difficult. This can be understood by observing that in the presence of inherent pattern, information learned from one part of the scene can be extended to parts that exhibit a similar pattern. We observe this in assumptions made in prior work around predicting specific kinds of object relationships (such as bottom-up, side-support) to contain the complexity created by this factor.
- **Tractability**: This refers to the number of objects sharing a support-relationship in a scene. Learning over scenarios with a large number of individual objects requires the physics intuition model to capture a larger uncertainty around the object interactions as well as handle the increased occlusions in the images. Since the model capacity for the physics intuition models is limited (based on the number of parameters), adding more objects should make learning difficult. This has been shown to be true in the case of block stacking tasks [38], and we expect this relationship to extend to object extraction tasks as well.

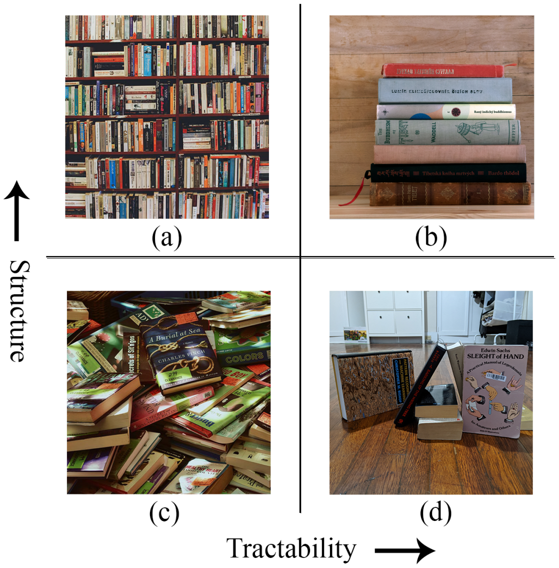


Figure 4.1: (a) A large, organized collection of books on a shelf is representative of an object arrangement with inherent structure and low tractability [2]. (b) A small pile of stacked books is representative of an object arrangement with both inherent structure and tractability [1]. (c) A large pile of books randomly jumbled around is representative of an object arrangement which lacks both inherent structure and tractability [3]. (d) A small pile of delicately balanced books displaying a myriad of support relations is representative of an object arrangement which is tractable but lacks inherent structure.

Figure 4.1 illustrates this taxonomy and the resulting categories using a single object. We do not consider the category where an arrangement of objects can be both intractable and lacking homogeneous structure as it becomes increasingly impractical to reason about disruption of individual objects. In such cases, one must instead reason about the effect of manipulating a collection of objects, which we leave for future work. In the following subsection, we go into more detail about the three selected categories and their corresponding data generation methods.

4.1.2 Synthetic Data Generation

To generate data corresponding to each category, we use the V-REP simulation platform [17] enabled with the Bullet physics engine [15]. For each category, we choose a representative scene type as depicted in Figure 4.2. In order to aid the network in identifying individual objects, similar to previous work [23, 37], we generate randomly colored objects. The dimensions of the base plane upon which a scenario is generated are constant across our data generation pipeline. We generate scenes using rigid homogeneous cuboids across all of the scene types to aid the generation of stable structures. Evaluation of the method on more complex shapes will be addressed in future work. The specific details about the methods used to generate data for each scene type are given below. The size of the blocks across all three datatypes is based on standard Jenga block dimensions and is constant throughout the scene.

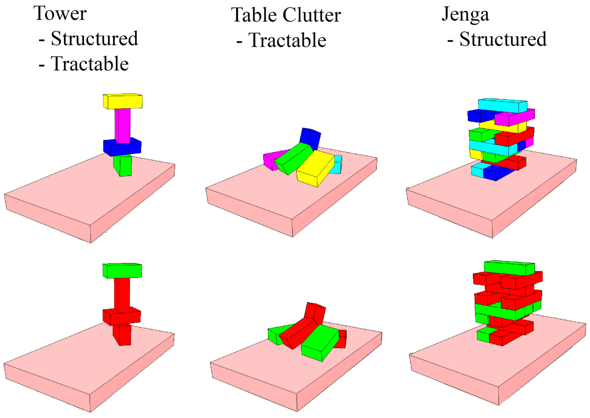


Figure 4.2: Dataset visualization for skill-agnostic models. Top row depicts sample scenes and the bottom row visualizes the ground truth - **green** indicates blocks that can be removed and **red** indicates otherwise.

- **Tower**: For each sample, we simulate a tower of 1 to 4 objects on a uniformly sampled base location on the table. We simulate 500 towers for each height. While the size of each block is kept constant, we uniformly sample the orientation of the sides on which each block will rest and the orientation around the table normal. This scene type is an example of a configuration which displays both tractability and inherent structure. It merely serves as a sanity check for our models, and we exclude it when reporting our experimental results (we expect only the top block to be predicted as extractable).
- **Table Clutter**: For each sample, we simulate a tabletop cluttered scene with 2 to 5 objects (500 arrangements for each). The size of the objects is kept constant and their positions are sampled uniformly across the table. Blocks are generated at a small distance above the table with a normally sampled orientation. They are then allowed to fall freely in the simulator and a sample is saved from all of the blocks that remain at rest on the table.
- **Jenga Tower**: For each sample, we generate a stable Jenga tower at the center of the table with the tower height ranging between 5 and 8 (500 towers each). We ensure that each row is supported by either ≥ 2 blocks from below or by one block placed along the center. Simultaneously we ensure that the top layer does not contain a single block positioned along an edge.

In our experiments we note that adding slight perturbations in the position of a block in the case of Jenga tower and in the dimension of the blocks in the Clutter scenario leads to a better performance on our real-world dataset for skill-agnostic models. We use this noisy dataset for reporting our real-world experiments for the skill-agnostic models.

Obtaining Ground Truth Labels

We run the target object extraction in our simulator to obtain the ground-truth label for each scenario. We enable surface friction and gravity as we let the objects first reach a static equilibrium state. We then step through the simulation for a fixed number of steps (0.25 seconds for skill-agnostic models and 2 seconds for skill-specific models) and record the position and velocities for the remaining object. This change in velocity is used to label a scenario as stable or unstable according to empirically determined threshold. For the skill-agnostic case, we simply delete the target object from the scene as depicted in Figure 4.3. We split the dataset to have an equal proportion of positive and negative samples across each data split (Table 4.1).

Table 4.1: Simulation Dataset Statistics For Skill-Agnostic Visual Assessment

Type	**Scenarios At Generation**			**Balanced Data Split**		
	Total	Stable	Unstable	Train	Validation	Test
Clutter	6978	4419	2559	4886	1046	1046
Jenga	13053	5473	7580	9138	1958	1957

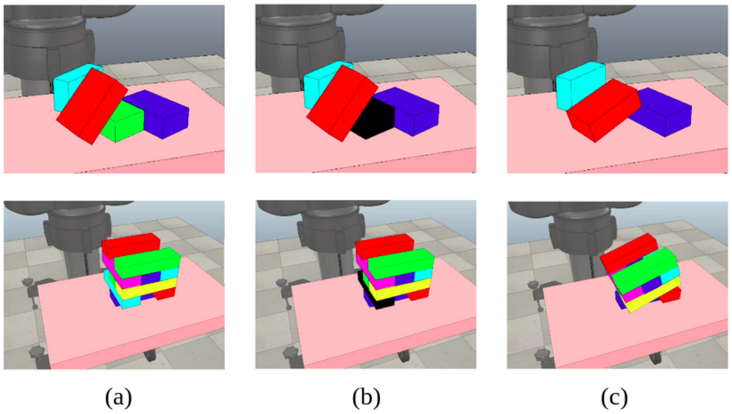


Figure 4.3: Ground truth labels for the skill-agnostic scenarios are generated by deleting the target object from the scene. Here (a) depicts the initial scene, (b) shows a specific scenario highlighting the target object (black) and (c) shows the effect of deleting the target object from the scene.

For the skill-specific case, we use a smaller subset (200 each) of the generated dataset and remove the target object using the 5 discrete skills described in the previous section (Section 3.2). For the skill-specific labels we use the change in position of the remaining blocks in the scene, instead of change in velocity, as the velocity of the system may not change instantly in this case. For each extraction direction, we move the target object by 0.2 m assuming a maximum acceleration to 0.1 m/sec^2. We determine these values as being reasonable for a robot to complete the extraction after grasping an object in 2 seconds. The configuration of remaining objects in the scene at the end of simulated extraction will also depend on physical properties of the objects which are not visually perceivable such as mass of the blocks and material friction. In our experiments we collect the stability labels for each scenario using 3 different material friction settings using the predefined dynamic material types from V-REP.

Table 4.2: Simulation Dataset Statistics For Skill-Specific Visual Assessment

Type	Split	UP	FW	BK	LF	RT	Total
Clutter	Train	1475	645	663	796	816	1552
	Validation	320	145	143	172	165	330
	Test	323	146	141	187	192	338
Jenga	Train	1541	1005	1013	1134	1084	3095
	Validation	348	215	228	229	233	651
	Test	333	205	241	244	248	644

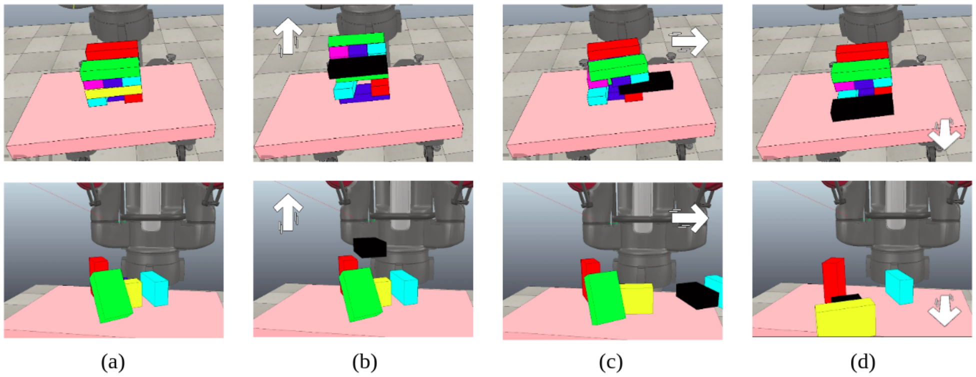


Figure 4.4: Ground truth labels for the skill-specific scenarios are generated by extracting the target object in a specific direction. Here (a) depicts the initial scene, and (b,c, d) depict specific scenario highlighting the target object (black), the direction of extraction (white arrow) and the effect of the extraction on the scene.

Generating Images

While V-REP is a convenient option for generating ground truth labels on our data, the rendered images from V-REP are not realistic and the model trained on these images do not transfer well to the real world images. We re-render our collected dataset in MuJoCo [61] which is another physics engine with better rendering capabilities. We capture RGB-D images and object masks of individual blocks from the scene from 5 uniformly spaced camera angles (3 for Jenga) on one side of the table at a resolution of 224x224 (Figure 4.5). From each rendered scene, we obtain multiple scenarios based on the number of objects in the scene (where a *scenario* is comprised of the sampled scene and one of the target objects). We render each scenario randomly varying the texture of the table and color of the blocks. We do this as a domain randomization step [60] to make our learned models transfer better to real-world images.

4.1.3 Real-World Data Collection

Our real-world dataset consists of RGB-D images taken from 3 different angles for 25 configurations of Jenga and Clutter each. We crop the images to center align the blocks in the scene and smoothen the depth images using the depth-toolbox from [55]. We annotate the dataset with segmentation masks for each object using Mask R-CNN [27] after fine-tuning the pre-trained model

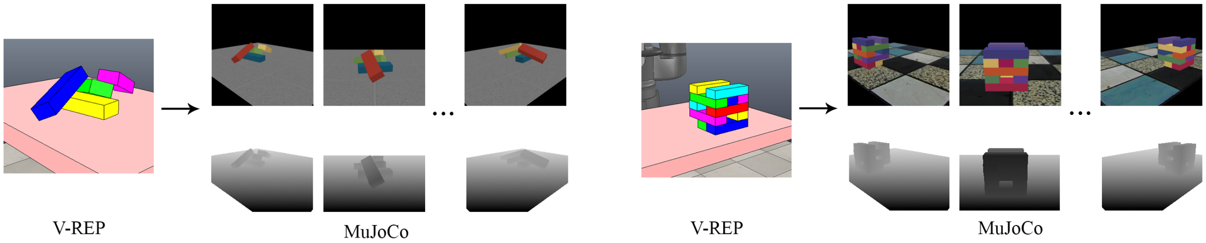


Figure 4.5: Rendering RGB-D images of object arrangements from multiple-views in MuJoCo.

on our simulation dataset. This model however is not efficient enough to detect all the blocks in our training dataset, as depicted in Figure 4.6, and we manually correct the annotations using RectLabel [8]. Ideally, in the presence of more real-world data we expect no manual annotation to be necessary. To label the dataset with skill-agnostic and skill-specific ground truth labels we reconstruct each scene in V-REP by obtaining the 6-D pose of each block. For the clutter scenes we use PERCH 2.0 [9], a multi-object pose estimation method that takes a 3-D point cloud, 2-D object masks and a 3-D model of the constituent scene objects as input to generate 6-D poses of each object in the frame of the table. The generated poses are not perfect and we further correct them manually in V-REP (Figure 4.7). For Jenga scenes, we observe significantly worse pose estimation from the pose estimation method and we choose to manually create abstract representations of the configurations and regenerate the samples in V-REP. We then use the same method as explained in Section 4.1.2 to generate the ground truth labels for both skill-agnostic and skill-specific conditions. The dataset is summarized in Table 4.3.

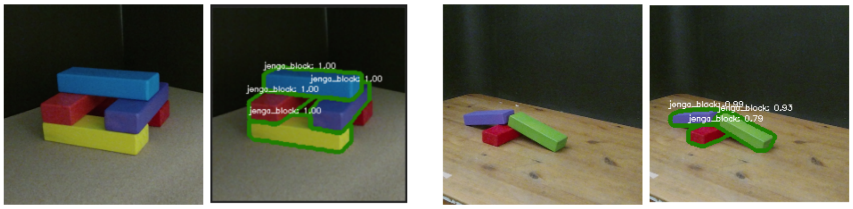


Figure 4.6: Visualizing instance segmentation results from Mask R-CNN on real-world images.

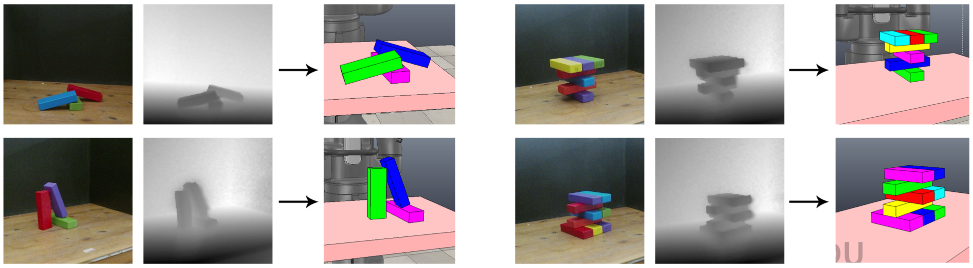


Figure 4.7: Sample Clutter (left) and Jenga (right) scenes from the real world dataset and their respective reconstructions in V-REP.

Table 4.3: Real-World Dataset Statistics

	Skill-Agnostic			Skill-Specific				
	Positive	**Negative**	**Total**	**UP**	**FW**	**BK**	**LF**	**RT**
Clutter	46	30	76	45	30	35	25	25
Jenga	80	110	190	48	40	41	32	30

4.2 Model Training

4.2.1 Skill-Agnostic Physics Intuition

We tried three state-of-the-art deep network architectures, AlexNet [34], ResNet-50 [26], and Inception-V4 [59], in a pilot study and found that AlexNet consistently gave similar performance to the other two on our simulation dataset while requiring significantly fewer parameters. Hence, we report our results on AlexNet across our experiments. We implement the modified Multi-View Convolutional Neural Networks (MVCNN) architecture proposed in [58] to extend the AlexNet to multi-view inputs of the scene. As depicted in Figure 3.2, each RGB-DM view of a scene is passed through the shared representation part of the network (CNN) and then aggregated at a view-pooling layer before being flattened and sent to the classification part (FNN - a single 256 hidden unit fully-connected layer). View-pooling is similar to max-pooling, with the only difference being the dimension that the pooling operation is performed on, as it simply performs element-wise maximum operation across the CNN representations of the multiple-views. Next, similar to [23], we optimize the standard logistic regression loss function for the binary classification task for the skill-agnostic models:

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log\left(p_i\right) + \left(1 - y_i\right)\log\left(1 - p_i\right)\right] \tag{4.1}$$

We train over 80 epochs with a batch size of 32, learning rate 0.0001, using the Adam optimizer and keep the model with best weighted F1-score. It is worth noting that we assume access to object correspondences across masks from multiple views. We acknowledge that this is a separate active area of current computer vision research although is beyond the scope of this work [28].

4.2.2 Skill-Specific Physics Intuition

For the skill-specific models, we train only on scenarios that are classified as stable by the skill-agnostic model. This is because the scenarios that are classified as unstable by the skill-agnostic model will always be labeled as unstable by the skill-specific model.

We start by considering the case where we assume knowledge about the material friction value of the blocks in the scene, assuming a high friction value. In this case, the loss function used for the skill-specific models is the same as above, with the only difference being that it comprises the multi-dimensional loss.

$$\mathcal{L}^j = -\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{y}_i^j \log\left(\mathbf{p}_i^j\right) + \left(1-\mathbf{y}_i^j\right)\log\left(1-\mathbf{p}_i^j\right)\right] \tag{4.2}$$

Here, $\mathbf{p}_i^j$ is the output of the logit function at the end of the network, $\mathbf{y}_i^j$ is the ground truth binary label for input image x_i and j is the skill identifier, $j \in$ [UP, FW, BK, LF, RT]. We evaluate the model on a validation set and keep the model with the lowest Hamming loss [63].

Next, we consider the case where we assume a uniform prior over the material friction value of the blocks and generate a continuous ground-truth label. We use Huber loss (Smooth L1 loss) as the loss function to optimize over during the model training. Huber loss is a preferred choice for regression tasks as it is less sensitive to outliers than simple mean-squared error and resistant to exploding gradients [22].

$$\mathcal{L}^j = \frac{1}{n}\sum_{i=1}^{n} z_i, z_i = \begin{cases} 0.5\left(\mathbf{p}_i^j - \mathbf{y}_i^j\right)^2, & \text{if } \left|\mathbf{p}_i^j - \mathbf{y}_i^j\right| < 1 \\ \left|\mathbf{p}_i^j - \mathbf{y}_i^j\right| - 0.5, & \text{otherwise} \end{cases} \tag{4.3}$$

Here, $\mathbf{p}_i^j$ is the clamped output (between 0 and 1) at the end of the network, $\mathbf{y}_i^j$ is the ground truth continuous label for input image x_i and j is the skill identifier, $j \in$ [UP, FW, BK, LF, RT]. We evaluate the model on a validation set keep the model with the lowest mean square error.

For both the skill-specific and skill-agnostic cases we train two models: (1) one trained over single-view images of the scene and (2) another trained over multiple views of a scene. We test both of these models on our test set for each scene type. We also perform testing using multiple aggregation techniques (Section 3.1.2) on predictions obtained over multiple camera views of the scene. For the purposes of this paper, a model that uses only a single view of the scene to make a prediction is called **single-view** and a model that uses multiple views is called **multi-view**. For skill-specific cases we present results only for the multi-view models.

4.2.3 Effect-Mask Prediction

Generating effect-masks is a dense prediction problem where, given a RGB-M image, our model predicts a corresponding mask over regions that will potentially be disturbed because of an object extraction. We use Fully-Convolution Networks (FCN) [40] that fundamentally adapt the fully-connected layers of the previously mentioned classification models by transforming them into convolutional layers. The network is then able to act as a spatial filter and output classification maps. These maps are up-sampled using deconvolution layers and made finer by adding links that combine the final prediction layer with lower layers with finer strides. In our implementation, we use a FCN-16 variant[1] of VGGNet [56] for end-to-end learning optimizing the pixel-wise cross-entropy loss. We evaluate the model on a validation set over 20 epochs and batch size 10 during training and keep the model with best precision score.

[1] https://github.com/pochih/FCN-pytorch

Chapter 5

Evaluation

5.1 Metrics

Since we are dealing with a binary prediction problem when evaluating skill-agnostic physics intuition, we compare all models using weighted F1-scores. F1-score is the harmonic mean of precision and recall, while weighted F1-score is a modified version that further accounts for class imbalance. It is calculated by averaging the F1-scores for each label weighted by their support. This measure lies between 0 and 1 and a higher value indicates a better model.

For skill-specific models in the case we assume knowledge of the material friction value, we continue using weighted F1-score as a measure for each extraction direction class. When we assume a uniform prior over friction values, we are evaluating a regression model, and we use Mean Square Error between the actual and predicted values as our metric. This measure lies between 0 and 1 and a lower value indicates a better model.

For effect-mask models, we use two metrics, Intersection-Over-Union (IoU) and Precision. IoU computes the number of pixels in the intersection of the predicted and ground truth segmentation maps, divided by the number of pixels in their union. Precision is a slightly modified version of IoU where we divide correct positive predictions over all positive predictions. This metric is used to evaluate how well our model can avoid false positives. This is important in our case because when we carry out manipulation using this effect-mask, we do not want the robot to manipulate a wrong object. We compute IoU and Precision scores for each data point and report their average in our evaluation. Both these measure lies between 0 and 1 and a higher value indicates a better model.

5.2 Baselines

We evaluate our learning-based physics intuition models on a set of baseline methods in simulation and on real-world data. For our real-world evaluation, we specifically choose baselines that take as inputs the RGB-D images of a scene from multiple views, 2-D segmentation masks for all the objects in each view, and 3-D models of the blocks in the scene. This allows for a fair comparison with our learning-based approach as it uses the same input.

5.2.1 Simulation-Based Physics Intuition

We implement a perception pipeline to obtain 6-D poses of all the blocks in the frame of the table. We then use a physics engine (V-REP) to provide skill-agnostic and skill-specific visual assessment by running roll-outs on each object of the inferred scene. Before performing the roll-outs we let the detected objects settle down under gravity and reach a static equilibrium. We apply the same perception algorithm, PERCH 2.0 [9], that we used to assist in obtaining the ground truth poses of objects in each real-world scene. We choose PERCH over other learning based methods such as Deep Object Pose Estimation (DOPE) [62] because these require a large training dataset to achieve a desirable level of accuracy while the former is a model-based approach. PERCH 2.0 builds upon the deliberative pose estimation method proposed by Narayanan and Likhachev [45] and takes advantage of GPU acceleration to provide accurate object pose-estimates quickly. It uses the available object masks to label the point cloud of each scene and renders sample 6-D poses of the object for each object point cloud using the available 3-D object model. It then performs Generalized Iterative Closest Point (ICP) to best align each sample with the original object point cloud [52]. PERCH defines a cost function to select the best aligned sample pose that penalizes points in rendered and observed point clouds for being distant in the euclidean space. This sampling and matching procedure happens in parallel for each object and reduces the runtime of the approach. It should be noted here that in the absence of sufficient points in the object point cloud, PERCH does not detect an object at all. We observe that even though PERCH is sufficiently accurate on Clutter scenes, its performance suffers on Jenga due to significant occlusions in the scenes (Figure 5.1). In our experiments, among the multiple views available, we choose the view from which we get the best reconstructed scene in terms of number of object detections.

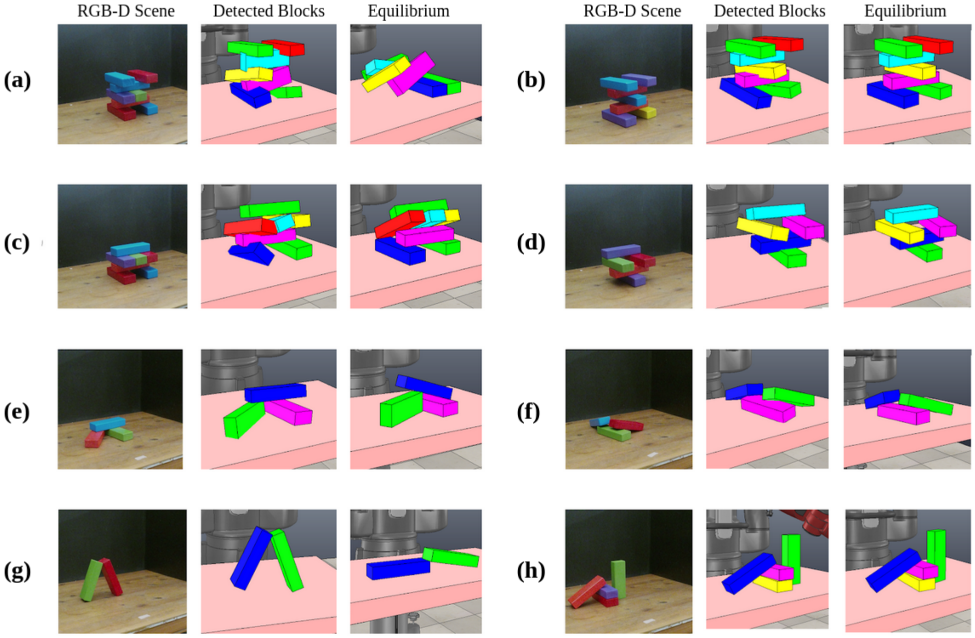


Figure 5.1: Simulation-Based Physics Intuition: (a, c, e, g) depict noisy scene reconstructions due to the perception pipeline either failing at detecting a block or because of having noisy pose estimates for some objects. (b, d, f, h) reflect accurate scene reconstructions.

5.2.2 Geometry-Based Physics Intuition

We further compare our skill-agnostic learning-based physics intuition models against recent work by Kartmann et al. [31]. They propose a geometry-based approach towards extracting physically plausible binary support relations between each pair of objects in a scene. Following the Static Equilibrium Analysis (SEA) proposed in [41], the authors compute ACT relationships among all pairs of objects in the scene. To compute the existence of such a relationship, they create a plane P separating A and B using contact points and normals. If A is above P and B is above it, they mark a support relationship between A and B. The authors add additional rules to capture 2 more relationships. To account for objects located horizontally next to each other i.e. when P is vertical, they introduce a threshold α_{max}. If the rotation angle of the separating plane relative to the gravity vector is below this threshold, they change the ACT relationship to an "unknown" support relationship and later empirically suggest to not include these in the support graph. To account for top-down support, they introduce a support area ratio r_s that is calculated as the ratio of the supported area of an object (by objects below it) and its total area. If r_s is below a threshold, a top-down relationship is added to the supporting blocks from below.

We use the recommended settings for the algorithm as mentioned in [41] and annotate both the ground truth and PERCH detected poses of all the scenes in our real world dataset. We adjust the contact margin parameter to account for perceptual inaccuracies of PERCH and set it to 0.5 cm. The approach suffers from similar challenges as the simulation-based model because of the imperfect pose estimation of blocks. Furthermore, we observe imperfect support-graphs even when we provide the method ground-truth poses. This is due to the method being very sensitive to its hyper-parameters (α_{max}, r_s) and not being able to capture all the object relationships reflected by our dataset (Figure 5.2).

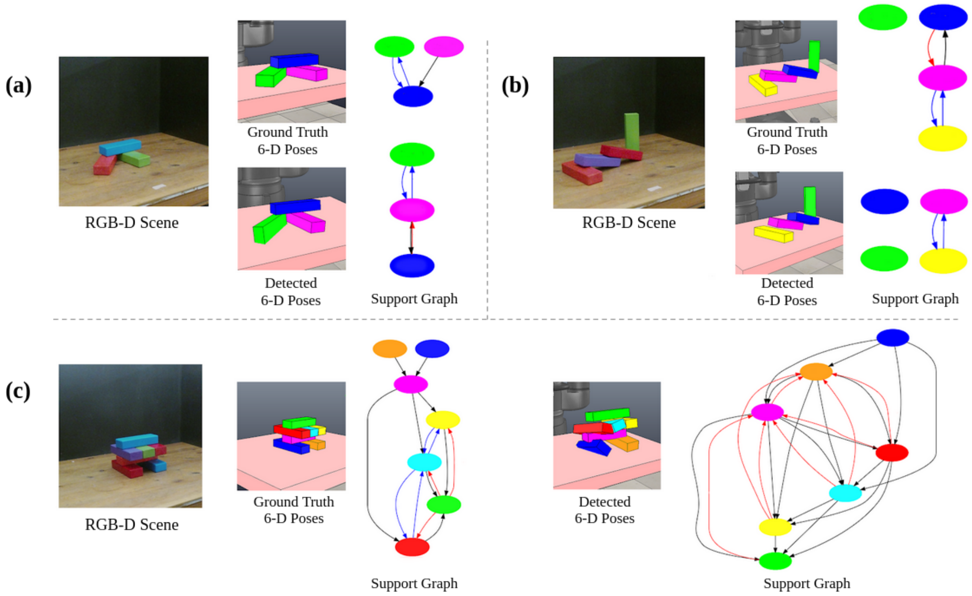


Figure 5.2: Geometry-Based Physics Intuition: Support graphs generated from ground truth and detected poses. **Red** edges indicate top-down and **black** edges indicate bottom-up support. **Blue** edges indicate "unknown" support and are not considered when parsing the support graph.

5.2.3 Heuristics

We also evaluate the performance of our proposed approach against 2 heuristics:

1. **Random Assignment Baseline**: We randomly label an object in a scene as being safe to extract or not. We create 100 such sample assignments and report the average performance.
2. **Pick Any Baseline**: We label all objects in a scene as being safe to extract regardless of their spatial configuration. This mimics the case where do not use any visual assessment.

Aside from comparing our approach to the aforementioned baselines, we also evaluate our claim that aggregating inferences over multiple views of a scene improves the performance of the system as compared to using a single view. This is done by comparing the **Multi-View** models to the *best* performance obtained from any of the **Single-View** models in the skill-agnostic case. For multi-view models, we perform both pre-training and post-training aggregation. For pre-training aggregation, we use the MVCNN architecture explained in Section 3.1.2, and for post-training, we evaluated 5 different aggregation techniques: mean, median, mode, max and min. We consistently saw better performance with the mean aggregation and report on this post-training aggregation method. It should be noted that the single-view models train on the entire multiple-view dataset and only make a prediction based on a single view of the scene.

5.3 Results

5.3.1 Skill-Agnostic Physics Intuition

On the simulation test dataset, our learning-based skill-agnostic physics intuition models demonstrate significantly higher performance as compared to the heuristic baselines (Table 5.1) across both multi-view and single-view models. Multi-view methods perform marginally better on both metrics compared to single-view methods on the simulation dataset, but this gap is more significant on the real-world dataset. We present examples from our simulation and real-world test dataset in Figure 5.3 and Figure 5.4, respectively.

		Learning-Based Physics Intuition			**Random Assignment**	**Pick Any**
		Multi-View (MVCNN)	*Multi-View (Mean)*	*Single-View*		
Clutter	*Simulation*	**0.85**	0.84	0.83	0.51	0.49
	Real-World	**0.86**	0.80	0.80	0.49	0.44
Jenga	*Simulation*	0.981	**0.983**	0.981	0.50	0.25
	Real-World	**0.83**	0.73	0.78	0.50	0.25

Table 5.1: Comparing learning-based physics intuition models against heuristic baselines (weighted F1-scores).

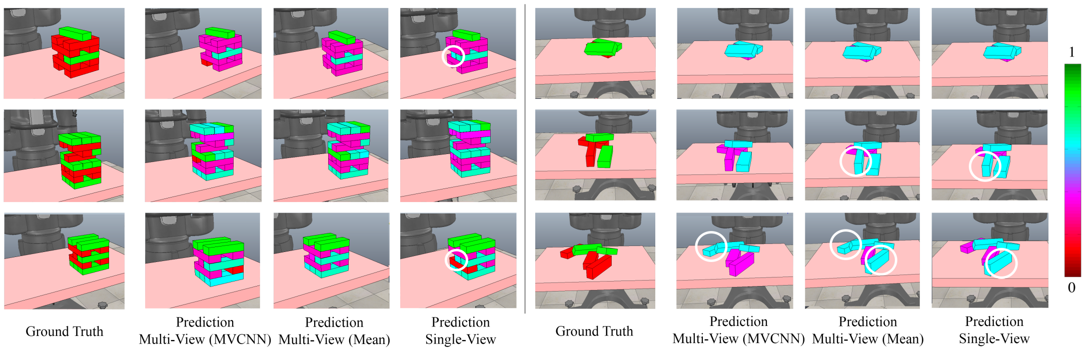


Figure 5.3: Visualizing predictions made by multi-view skill-agnostic models in simulation. Single-view models often misclassify objects (white circles) that are partially occluded or where the object's pose is not clearly identifiable.

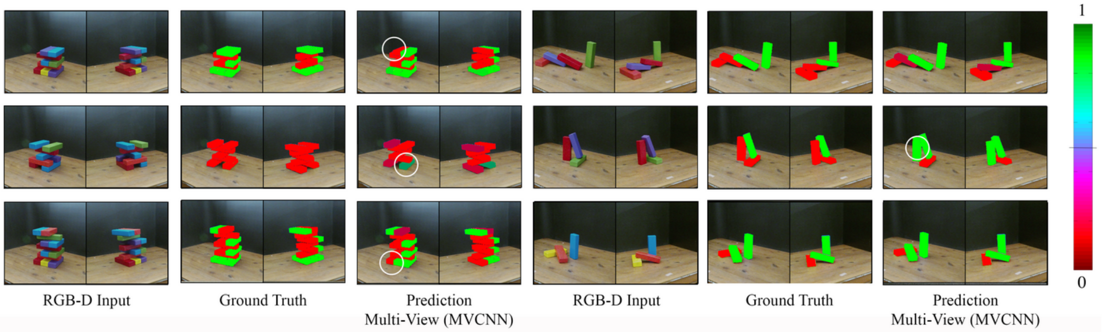


Figure 5.4: Visualizing predictions made by multi-view (MVCNN) skill-agnostic models on real-world data. For each scene, images taken from multiple camera views are passed to our trained physics intuition model and the probabilistic predictions are recorded. The misclassifications are highlighted with a white circle.

We do not observe a clear winner among different aggregation methods for multiple views on the simulation dataset but observe an interesting trend on the Jenga real-world data where the best single-view model performs better than the multi-view mean aggregation model but worse than the multi-view (MVCNN) model. For Jenga towers, a correct prediction for a particular arrangement can often be reached only from a few corresponding selected angles (while other angles remain ambiguous or under-represented in the training data), and MVCNN proves to be a better choice of model than both multi-view mean aggregation in making predictions over the multi-view input. On the contrary, for Clutter scenes, a correct prediction for a particular arrangement can be reached from multiple camera angles, and taking the mean prediction still provides a comparable performance to single-view models (Figure 5.5) albeit worse than MVCNN models.

Next, we compare the multi-view (MVCNN) method against the simulation-based and geometry-based baselines. For the geometry-based method, we parse the support-graph to obtain extractability predictions for each detected object. An object that has an outward bottom-up or top-town support edge is labeled as negative while all other objects are labeled as positive. Some blocks (22 out of 190) in the Jenga dataset are not detected by our perception method and therefore we report the performance for the simulation-based method in two conditions: (a) labeling the missing blocks as negative and (b) removing them from the evaluation. This is reported in the format "$a(b)$" across our evaluations, where a represents the result when operating under

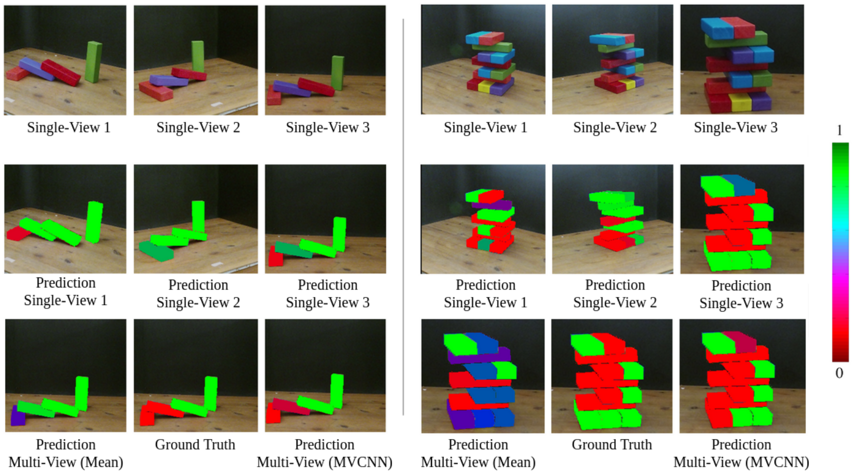


Figure 5.5: Visualizing predictions made by learning-based skill-agnostic models on real-world data. The multi-view (mean) predictions are obtained after aggregating the single-view predictions (second row). The third row compares the predictions made by the multi-view (mean) model and multi-view (MVCNN) model against the ground truth.

condition (a) and likewise b represents the results when operating under condition (b).

We observe that for the Clutter scenes, the simulation-based approach has the best performance (Table 5.2). This validates that our perception pipeline is robust and is able to detect most blocks in the clutter scenes with sufficient accuracy. It is worth noting that our learning-based approach has a comparable performance to the simulation-based counterpart while being quicker and more computationally efficient. The geometry-based approach shows a better-performance on Clutter scenes when given the ground-truth object poses as compared to when given the detected poses. This is expected as the method is based on rules that are highly sensitive to the object poses. However, even with the ground-truth poses, the performance of the geometric method is below that of the simulator-based method. This may be because the method is governed by a combination of rules that do not entirely reflect the characteristics of our dataset. For example, we frequently observe spurious top-down edges in the support-graphs generated even from ground truth poses (Figure 5.2). We additionally often see that potential bottom-up relationships get conflated with adjacent-support relationships (“unknown” edges) and lead to inaccurate support graphs.

For the Jenga scenes, the learning-based approach comes out superior to the simulation-based method as the latter suffers significantly more from perceptual inaccuracies in perceiving arrangements that feature a large number of occlusions. The reconstructed scenes often collapse when allowed to reach static equilibrium due to the imperfect pose estimation of blocks and inaccurate assessment of their extractability (Figure 5.1). The geometry-based approach too performs poorly in the case of Jenga. This is in part for the same reason, but also because the support graphs generated by the approach do not address cases where an object can be supported by more than 2 other objects from the bottom (3 in the case of Jenga) and can stay stable if one of them is removed. This introduces many spurious bottom-up edges in the support graph and explains the poor performance of geometry-based methods on our dataset.

		Learning-Based Physics Intuition (MVCNN)	Geometry-Based Physics Intuition	Simulation-Based Physics Intuition
Clutter	*RGBD Images*	0.86	-	-
	Ground Truth Poses	-	0.74	-
	Perceived Poses	-	0.65	**0.91**
Jenga	*RGBD Images*	**0.83**	-	-
	Ground Truth Poses	-	0.53 (0.53)	-
	Perceived Poses	-	0.55 (0.58)	0.58 (0.61)

Table 5.2: Real-world evaluation for skill-agnostic models (weighted F1-scores).

5.3.2 Skill-Specific Physics Intuition

We use the multi-view (MVCNN) model as our learning-based physics intuition model for evaluating the skill-specific physics intuition models. Assuming access to the material friction value, we observe a similar trend in skill-specific models as seen earlier in the skill-agnostic case. On the simulation test dataset, the skill-specific physics intuition models demonstrate a better performance as compared to the heuristic baselines (Table 5.3). On the real-world dataset, similar to the skill-agnostic case, the simulation-based models are best for Clutter scenes but suffer with the Jenga ones (Table 5.4). We visualize a few samples from our simulation and real-world test dataset in Figure 5.6 and Figure 5.7 respectively.

		UP	FW	BK	LF	RT	Average
Clutter	*Learning-Based PI (MVCNN)*	0.93	0.87	0.89	0.82	0.83	**0.87**
	Random Assignment	0.62	0.50	0.50	0.50	0.50	0.53
	Pick Any	0.93	0.26	0.25	0.40	0.41	0.45
Jenga	*Learning-Based PI (MVCNN)*	0.93	0.93	0.94	0.93	0.93	**0.93**
	Random Assignment	0.50	0.52	0.51	0.51	0.51	0.51
	Pick Any	0.16	0.15	0.20	0.20	0.21	0.18

Table 5.3: Simulation test set evaluation for skill-specific models (weighted F1-scores). Here we assume knowledge of the material friction value of the blocks in the scene.

When assuming a uniform prior over material friction values, we obtain a continuous value prediction between 0 and 1 from our skill-specific models. To generate the ground truth values, for both our simulation and real-world datasets, we assume 3 different material friction values and labeled each scenario as being stable or unstable after performing the target-object extraction. For the clutter dataset, the objects in the scene slip against each other on lower friction values and have very few interactions in their initial equilibrium state. This leads to even objects that are safe to extract be labeled as unstable under certain friction conditions because other object

		UP	FW	BK	LF	RT	Average
Clutter	*Learning-Based PI (MVCNN)*	0.97	0.74	0.45	0.87	0.81	0.77
	Simulation-Based PI (Perception)	0.98	0.87	0.85	0.91	0.87	**0.90**
	Random Assignment	0.67	0.50	0.52	0.51	0.51	0.54
	Pick Any	0.97	0.51	0.66	0.38	0.38	0.59
Jenga	*Learning-Based PI (MVCNN)*	0.92	0.86	0.79	0.88	0.80	**0.85**
	Simulation-Based PI (Perception)	0.60 (0.60)	0.72 (0.73)	0.60 (0.58)	0.74 (0.76)	0.74 (0.77)	0.68 (0.69)
	Random Assignment	0.50	0.50	0.50	0.50	0.51	0.50
	Pick Any	0.36	0.33	0.35	0.23	0.20	0.30

Table 5.4: Real-world evaluation for skill-specific models (weighted F1-scores). Here we assume knowledge of the material friction value of the blocks in the scene.

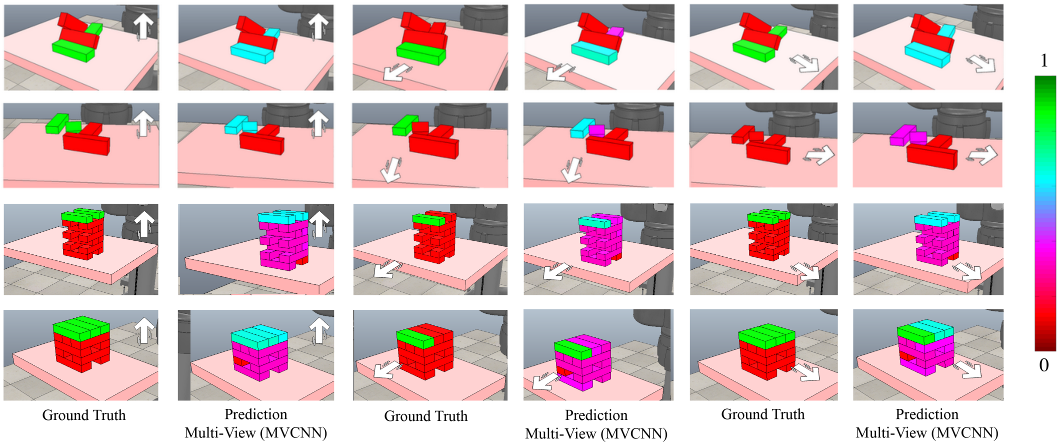


Figure 5.6: Visualizing predictions made by multi-view (MVCNN) skill-specific models in simulation. Here we assume knowledge of the material friction value of the blocks in the scene. Which objects can be safely removed depends on which skill we try to remove them with (white arrow). We observe that skill-specific models make very few misclassifications and are able to accurately predict this skill-specific notion of object-extraction.

interactions in the scene are unstable. For this reason, we only use the Jenga scenario for this analysis as the scene remains stable in the equilibrium state for all three friction values. On the real-world dataset, we compare the predictions of the learning-based models to the ones obtained from the simulation-based models. The results are summarized in Table 5.5. Despite the drop in performance due to the implicit sim-to-real gap, we see promising results as our learning-based physics intuition models perform reasonably well on real-world images compared the simulation-based model. We visualize a few predictions from our model in Figure 5.8 and 5.9.

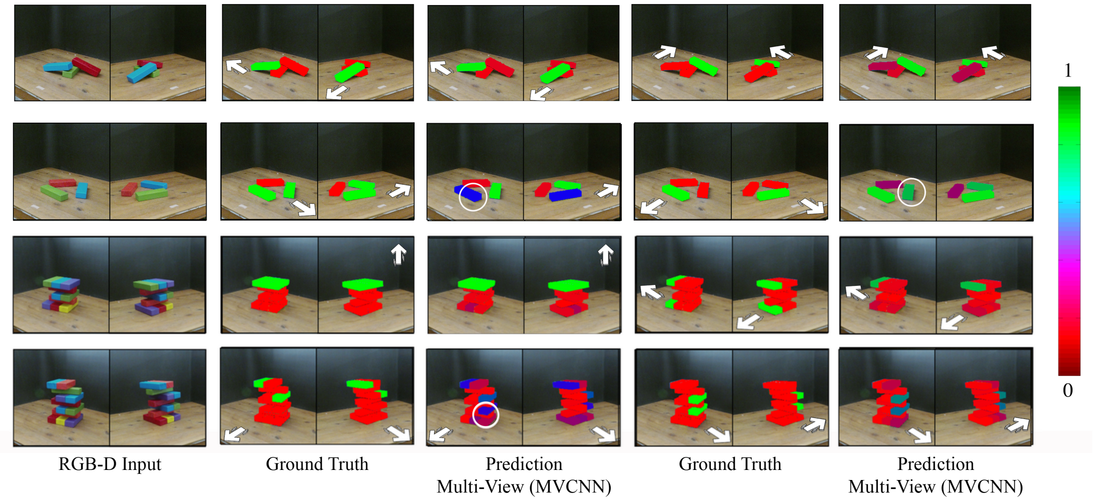


Figure 5.7: Visualizing predictions made by multi-view skill-specific models on real-world data. Here we assume knowledge of the material friction value of the blocks in the scene. The extraction direction is indicated with a white arrow, and misclassifications are highlighted with a white circle. Only blocks that are judged as extractable by the skill-agnostic models are classified by the skill-specific models.

	Jenga	**UP**	**FW**	**BK**	**LF**	**RT**	**Average**
Simulation	*Learning-Based PI (MVCNN)*	0.02	0.02	0.03	0.03	0.03	**0.02**
Real-World	*Learning-Based PI (MVCNN)*	0.08	0.10	0.11	0.13	0.12	**0.11**
	Simulation-Based PI (Perception)	0.21 (0.22)	0.28 (0.26)	0.35 (0.38)	0.23 (0.22)	0.23 (0.20)	0.26 (0.25)

Table 5.5: Real-world evaluation for skill-specific models (mean squared error). Here we assume a uniform prior over the material friction value of the blocks in the scene. Lower value is better.

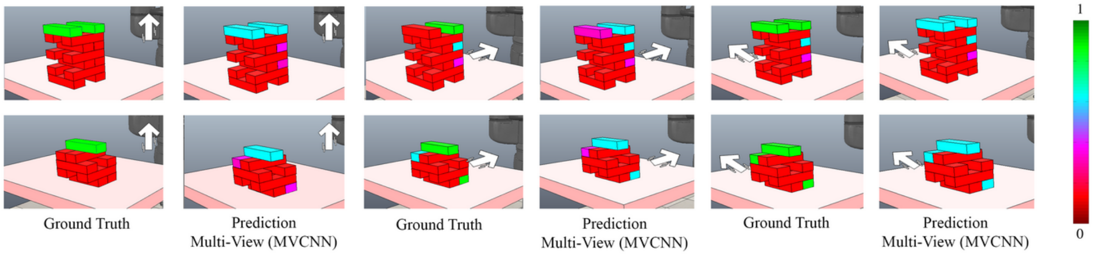


Figure 5.8: Visualizing predictions made by multi-view skill-specific models in simulation. Here we assume a uniform prior over the material friction value of the blocks in the scene because of which we can observe that the ground-truth labels are continuous instead of binary. The extraction direction is indicated with a white arrow.

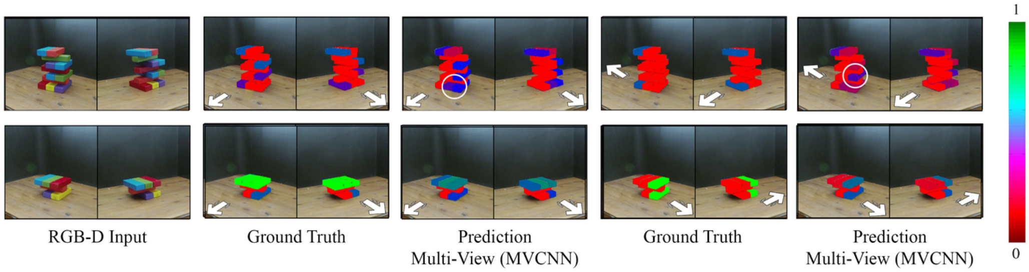


Figure 5.9: Visualizing predictions made by multi-view skill-specific models on real-world data. Here we assume a uniform prior over the material friction value of the blocks. The extraction direction is indicated with a white arrow and misclassifications are highlighted with a circle.

5.3.3 Effect-Mask Prediction

The effect-mask model makes per-pixel predictions on a single RGB image input around which part of the scene will move on removing a target object from the scene. These predictions take the form of a probabilistic mask on the input image that we refer to as effect-mask. Specific object can be deduced from this effect-mask by finding its intersection with specific object masks. We expect this intuition around "*which objects will move?*" to be very useful as it will allow robots to secure a set of blocks when attempting an unsafe extraction. Our evaluation of the effect-mask predictor is preliminary, yet promising. In Figure 5.10, we visualize the predicted effect-masks as heatmaps and can observe that the model is reasonably accurate in identifying the objects that will move if the target-object is removed from the scene. We report our results on the simulation validation set in Table 5.6. The model learns to predict significantly better effect-masks on Jenga scenes than on Clutter. This maybe due to the relatively consistent structure of the masks across Jenga scenes that make it a relatively easier segmentation task as compared to clutter scenes.

	Intersection over Union	**Precision**
Clutter	0.36	0.67
Jenga	0.62	0.80

Table 5.6: Simulation test set evaluation for effect-mask prediction.

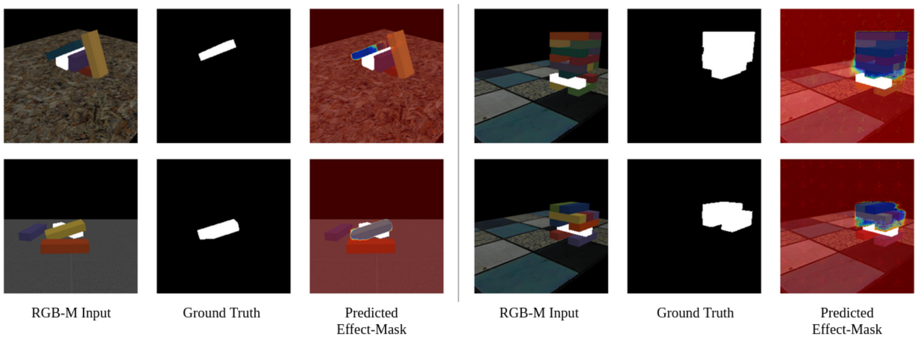


Figure 5.10: Visualizing predicted effect-masks as overlaid heatmaps on simulation data. The target object is highlighted in white in the RGB-M input and predicted effect-mask.

5.4 Analysis

In this section, we analyze our learning-based, skill-agnostic models to gain further insight into their internal properties and understand their limitations to inform future work.

5.4.1 Generalizability

A drawback of learning-based approaches is their limited generalizability on data points that are significantly different from the ones in their training data. For example, in Figure 5.11 we observe a Clutter scene where two blocks are almost vertical and mutually support each other. We do not witness many such cases during our data generation process since these arrangements require more deliberate construction. Expectedly, our model is confused in this case and makes a wrong prediction on both the blocks. It is also worth noting that due to the large data imbalance along which objects can be picked "UP" on the Clutter dataset, our model performs as good as the Pick Any baseline on both the simulation (Table 5.3) and real-world dataset (Table 5.4).

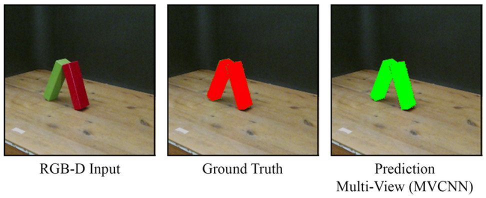


Figure 5.11: An example scene where learning-based skill-agnostic models make wrong predictions due to not seeing enough, if any, such arrangements in the training dataset.

However despite a few such aforementioned cases, our models are quite generalizable. We analyze the predictions made by our skill-agnostic (MVCNN) model on our real-world Clutter scenes and observe that the model is robust against minor inconsequential changes in the scene and, at the same time, is well-aware of important ones. We can see in Figure 5.12 that our learning-based, skill-agnostic model continues to give accurate predictions as we modify a scene along various dimensions.

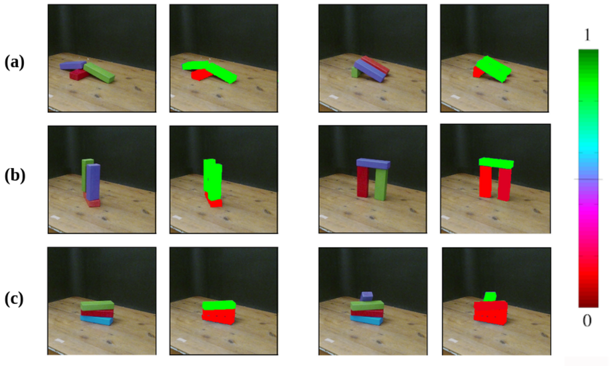


Figure 5.12: Visualizing changes in predictions made by the skill-agnostic (MVCNN) model in response to modifications to the scene in terms of (a) object orientation, (b) scene inversion (c) new object addition.

	Training	Testing	Weighted F1-Score
	Num. Blocks		
Clutter	2, 3	4	0.81
	2, 4	3	0.87
	3, 4	2	0.96
	Tower Height		
Jenga	5, 6	7	0.95
	5, 7	6	0.99
	6, 7	5	0.99

Table 5.7: Evaluation of generalizability about number of objects in the scene for skill-agnostic multi-view (MVCNN) models.

We further evaluate the performance of skill-agnostic multi-view (MVCNN) models when trained on scenes consisting of different numbers of blocks than the testing scenes. Results from this experiment are summarized in Table 5.7. We observe that models trained on a larger number of blocks extend well to scenes with fewer blocks but not always vice versa. In the case of Jenga, this gap is much smaller, as the bigger towers include many similar scenarios to the smaller tower scenarios in their training data. This difference is larger in Clutter scenarios as some unique interactions among a larger number of blocks are not be well-represented when trained on a smaller number. The results show that even though our models are able to learn certain rules that determine safe extraction and extend them to images with different number of objects than in the training set.

5.4.2 Performance

Fast visual assessment can enable robots to continually gain physics intuition about an arrangement of objects and account for a change in the configuration of objects in real-time. We note that an elaborate perception pipeline can introduce significant barriers towards achieving this goal. PERCH uses a 8GB GPU and takes on average 8 seconds to reconstruct a scene. This time is independent of the number of objects in the scene as PERCH performs pose estimation in parallel across all objects. V-REP further takes approximately 3-5 seconds to run a single simulation roll-out for each object in fast simulation mode on a 3.5 GHz CPU. We perform these roll-outs sequentially but in practice this can be done in parallel across all the objects. In contrast, our learning-based approach uses the same compute setup and takes less than 1 second to visually assess all the objects in the scene, even when done sequentially for each object. Although, the geometry-based models take negligible time in creating support graphs from a reconstructed scene, they rely on the same perception-pipeline as the simulation-based approach. Therefore, we can see that learning-based physics intuition models are both fast and accurate, making them suitable for real-time visual assessment.

5.4.3 Network Visualization

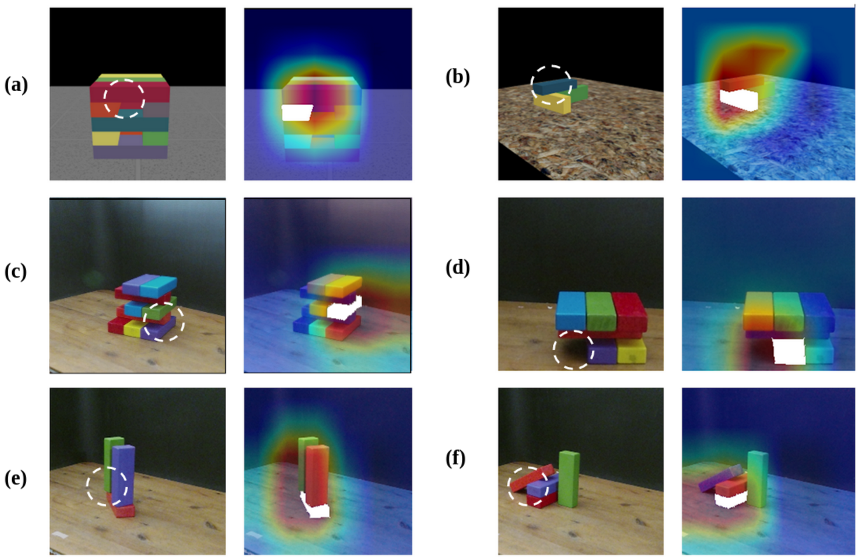


Figure 5.13: Class activation maps for simulated (a, b) and real-world (c, d, e, f) scenes. Target object is highlighted in white. Dotted circles in the original images (left) highlight the region corresponding to the activation maps (right).

To inspect which properties of the scene our models focus on when making predictions, we visualize the learned discriminative image regions from the CNN layer of the models. As proposed in [71], introducing a Global Average Pooling layer between the last convolution layer and the final fully connected layer of our model allows us to back-project the weights from the fully-connected layer and obtain Class Activation Maps (CAMs).

We inspect these discriminative regions in our test set for the skill-agnostic models. Figure 5.13 displays visualizations across two views of the Jenga and Clutter scenes both in the simulated and real image settings. The regions that contribute to the models' predictions are proximal to the target object. We can also see that for Jenga scene images that have a clear view of the objects in the target object row, the network either focuses on that row or on the blocks below it. This may be because having an alternate supporting block in the row is critical for the stability of the tower, as is having supporting blocks underneath to counter the change in weight due to the removal of a block. These properties may be best assessed independently from separate views rather than from a single view alone, and may further explain how multi-view models may have an edge over single view models. In the table Clutter scenario, we observe that the network focuses on the contact points of the target object and the object's support dependencies. This may be because having contact with another block and the arrangement of nearby non-target objects (such as presence of alternate support) are important factors that determine a scene's stability.

Chapter 6

Conclusion

6.1 Summary

Existing research has shown that robots can use learning-based physics intuition models to predict the stability of a scene directly from images and exploit this reasoning to create stable stacks of objects in the real-world. Here, we demonstrated how robots can use similar visual assessment to perform the inverse process of predicting which objects can be safely extracted from a configuration (skill-agnostic) and in which direction (skill-specific). We also touched on predicting which objects will move on extracting a particular object from a configuration (effect-masks).

We extended the existing physics intuition training methodology by conditioning the training images on specific objects using an object mask alongside the image of the scene. We further showed that aggregating multiple views can increase the model's performance for scenes with multiple objects and unstructured object arrangements. For our evaluations we curated a diverse dataset of table-top Clutter and Jenga tower scenes in simulation and in the real-world. We demonstrated on these datasets that our learning-based visual assessment can provide a skill-agnostic and skill-specific physics intuition around safe *extractability* of an object and therefore effectively reduce the probability of a robot disrupting the scene.

We also compared our proposed learning-based approach to simulation-based and geometry-based visual-assessment and saw that it achieved comparable, if not better, performance on both Clutter and Jenga real-world datasets. We noted that since the learning-based approach operates directly on images, it can operate well in scenarios where gaining access to complete and accurate geometric information of a scene is not feasible. This also made learning-based physics intuition models capable of providing real-time visual assessment.

In analyzing the discriminative image regions found by the model, we observed that discriminative regions correlated with regions that were critical to the stability of the scene, such as objects being directly supported by the target object or regions where alternate support must be present in order to avoid disruption. This analysis suggests that the system has learned meaningful intuitive physics features of the scenes.

6.2 Limitations and Future Work

Our proposed approach employs deep neural networks and is data intensive by nature. Both the quality and quantity of data are important to learn physics intuition models that would be sufficiently generalizable to more complex scenes with objects of different shapes and sizes as well as different environment conditions. For example, even though we vary the lighting conditions as we generate training data for our models, the variation in simulation is limited and our models remain sensitive to ambient light in the real-world. To tackle this challenge we would need to use more photo-realistic renders and generate images in a variety of lighting conditions along with employing significantly more domain randomization. Alternately, to scale our approach to diverse scenes we can to extract a better representation of scenes. For example instead of only using a mask over the target-object, we can create representations of each object in the image using their respective masks and use graph neural networks to encode the scene configuration as in [54, 68].

On similar lines, an important thing to note is that our approach currently assume access to good quality object masks as well correspondences across them from multiple views of a scene. Since we directly use simulator-generated object masks in our model training and testing, our models are sensitive to quality of these masks. By introducing random noise in these masks during training, the models can be made more robust to noisy masks at test time.

Any kind of visual prediction will suffer from uncertainty around unperceived features such as material properties and forces acting between objects (e.g., friction). In this work, we assume scenes comprised of rigid homogeneous objects and explore how assuming a prior over friction values can allow our model to incorporate the uncertainty around these unknown parameters. An interesting direction to explore would be to interact separately with a few objects from the scene to capture an intuition about their physical properties that can be combined with the vision-based physics intuition to provide more precise assessment about the entire scene. This would be similar to system identification, but instead of learning specific physical features individually, the robot learns representations that are meaningful to assist the vision-based physics intuition. Furthermore, the robot can also learn meaningful interaction policies to learn useful representations from interactions with the subset of objects.

We independently evaluate single-view and multi-view models and find that multiple views can increase a model's performance. Although obtaining multiple views of a scene every time may not be necessary. We do not explicitly model the uncertainty in our predictions and that can be a useful addition to our approach. A future direction to explore would be a resolution for the dilemma of choosing between the single-view and multi-view models in real-time. For example, non-optimal viewing angles may report a high uncertainty in their predictions and signal a need to obtain another view of the scene. Furthermore, a more detailed analysis of the relation between the number and similarity of views used over training and testing can be valuable future work.

Another limitation of our work is that for skill-specific visual assessment we currently restrict our model to 4 directions with respect to the table in the image. Future work can look at encoding continuous direction action vectors directly in images perhaps using colored masks similar to [44]. Another alternative would be to include the robot's actuator in the image as it views the scene and use it as the reference object to be always available in the image.

Bibliography

[1] *Image - Books - Structured Tractable*, . URL https://images.app.goo.gl/eKaeeSyKAB9WkQ2BA. (document), 4.1

[2] *Image - Books - Structured Intractable*, . URL https://images.app.goo.gl/tgvqHAGwX9gVE67X9. (document), 4.1

[3] *Image - Books - Unstructured Intractable*, . URL https://images.app.goo.gl/oHHTD2ZSNs6jj4VS8. (document), 4.1

[4] *Image - Arrangement of Cans.* URL https://images.app.goo.gl/A4PMdw9VJCu6DfGr8. 1.1b

[5] *Image - Arrangement of Cement Bags.* URL https://images.app.goo.gl/WCwyZYpb5CBSCajBA. 1.1b

[6] *Image - Arrangement of Dishes.* URL https://images.app.goo.gl/n7doFGNRjA39x9126. 1.1a

[7] *Jenga Official Website.* URL https://jenga.com/. 1.1

[8] *RectLabel.* URL https://rectlabel.com/. 4.1.3

[9] Aditya Agarwal. Fast and high-quality gpu-based deliberative perception for object pose estimation. Master's thesis, Pittsburgh, PA, June 2020. 4.1.3, 5.2.1

[10] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. 1.2

[11] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110 (45):18327–18332, 2013. 1.1, 2.2

[12] Wissam Bejjani, Mehmet R Dogar, and Matteo Leonetti. Learning physics-based manipulation in clutter: Combining image-based generalization and look-ahead planning. *arXiv preprint arXiv:1904.02223*, 2019. 2.2

[13] Arunkumar Byravan and Dieter Fox. Se3-nets: Learning rigid body motion using deep neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 173–180. IEEE, 2017. 2.3.2

[14] Michael B Chang, Tomer Ullman, Antonio Torralba, and Joshua B Tenenbaum. A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*, 2016. 2.3.2

[15] Erwin Coumans. Bullet physics engine. *Open Source Software: http://bulletphysics. org*, 1 (3):84, 2010. 4.1.2

[16] Mehmet Dogar and Siddhartha Srinivasa. A framework for push-grasping in clutter. *Robotics: Science and systems VII*, 1, 2011. 1.2

[17] M. Freese E. Rohmer, S. P. N. Singh. V-rep: a versatile and scalable robot simulation framework. In *Proc. of The International Conference on Intelligent Robots and Systems (IROS)*, 2013. 4.1.2

[18] David Eberly. *3D game engine architecture: engineering real-time applications with wild magic*. CRC Press, 2004. 1

[19] N Fazeli, M Oller, J Wu, Z Wu, JB Tenenbaum, and A Rodriguez. See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion. *Science Robotics*, 4 (26):eaav3123, 2019. 1.2

[20] Katerina Fragkiadaki, Pulkit Agrawal, Sergey Levine, and Jitendra Malik. Learning visual predictive models of physics for playing billiards. *arXiv preprint arXiv:1511.07404*, 2015. 2.3.2

[21] Tobias Gerstenberg and Joshua B Tenenbaum. Intuitive theories. 1.1

[22] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 4.2.2

[23] Oliver Groth, Fabian B Fuchs, Ingmar Posner, and Andrea Vedaldi. Shapestacks: Learning vision-based physical intuition for generalised object stacking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 702–717, 2018. 1.2, 2.3.2, 3.1.2, 4.1, 4.1.2, 4.2.1

[24] Abhinav Gupta, Alexei A Efros, and Martial Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *European Conference on Computer Vision*, pages 482–496. Springer, 2010. 2.3.1

[25] K Hauser. Cutting through the clutter: Identifying minimally disturbed subsets. In *RSS Workshop on Robots in Clutter: Manipulation, Perception and Navigation in Human Environments*, 2012. 1.2

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4.2.1

[27] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3.1.1, 4.1.3

[28] Weiming Hu, Min Hu, Xue Zhou, Tieniu Tan, Jianguang Lou, and Steve Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):663–671, 2006. 4.2.1

[29] Michael Janner, Sergey Levine, William T. Freeman, Joshua B. Tenenbaum, Chelsea Finn, and Jiajun Wu. Reasoning about physical interactions with object-oriented prediction and planning. In *International Conference on Learning Representations*, 2019. 2.3.2

[30] Zhaoyin Jia, Andrew C Gallagher, Ashutosh Saxena, and Tsuhan Chen. 3d reasoning from blocks to stability. *IEEE transactions on pattern analysis and machine intelligence*, 37(5): 905–918, 2014. 2.3.1

[31] Rainer Kartmann, Fabian Paus, Markus Grotz, and Tamim Asfour. Extraction of physically plausible support relations to predict and validate manipulation action effects. *IEEE Robotics and Automation Letters*, 3(4):3991–3998, 2018. 2.1, 3.3, 5.2.2

[32] Shinya Kimura, Tsutomu Watanabe, and Yasumichi Aiyama. Force based manipulation of jenga blocks. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4287–4292. IEEE, 2010. 2.1

[33] Kris M Kitani, De-An Huang, and Wei-Chiu Ma. Activity forecasting: An invitation to predictive perception. In *Group and Crowd Behavior for Computer Vision*, pages 273–294. Elsevier, 2017. 2.3.2

[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 4.2.1

[35] Oliver Kroemer and Gaurav S Sukhatme. Learning spatial preconditions of manipulation skills using random forests. In *Humanoid Robots (Humanoids), 2016 IEEE-RAS 16th International Conference on*, pages 676–683. IEEE, 2016. 2.3.1

[36] Torsten Kröger, Bernd Finkemeyer, Simon Winkelbach, and Friedrich M Wahl. Demonstration of multi-sensor integration in industrial manipulation (poster). 1.2

[37] Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. *arXiv preprint arXiv:1603.01312*, 2016. 1.2, 2.3.2, 3.1.2, 4.1, 4.1.2

[38] Wenbin Li, Aleš Leonardis, Jeannette Bohg, and Mario Fritz. Learning manipulation under physics constraints with visual perception. *arXiv preprint arXiv:1904.09860*, 2019. 1.2, 2.3.2, 3.1.2, 4.1.1

[39] Young Won Lim and Sang Uk Lee. On the color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques. *Pattern recognition*, 23(9):935–952, 1990. 3.1.1

[40] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3.3, 4.2.3

[41] Rasoul Mojtahedzadeh, Abdelbaki Bouguerra, Erik Schaffernicht, and Achim J Lilienthal. Support relation analysis and decision making for safe robotic manipulation tasks. *Robotics and Autonomous Systems*, 71:99–117, 2015. 2.1, 2.3.1, 5.2.2

[42] Mark Moll, Lydia Kavraki, Jan Rosell, et al. Randomized physics-based motion planning for grasping in cluttered and uncertain environments. *IEEE Robotics and Automation Letters*, 3(2):712–719, 2017. 2.2

[43] Roozbeh Mottaghi, Hessam Bagherinezhad, Mohammad Rastegari, and Ali Farhadi. Newtonian scene understanding: Unfolding the dynamics of objects in static images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages

3521–3529, 2016. 2.3.2

[44] Roozbeh Mottaghi, Mohammad Rastegari, Abhinav Gupta, and Ali Farhadi. “what happens if...” learning to predict the effect of forces in images. In *European conference on computer vision*, pages 269–285. Springer, 2016. 2.3.2, 6.2

[45] Venkatraman Narayanan and Maxim Likhachev. Perch: Perception via search for multi-object recognition and localization. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5052–5059. IEEE, 2016. 5.2.1

[46] Oni Ornan and Amir Degani. Toward autonomous disassembling of randomly piled objects with minimal perturbation. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4983–4989. IEEE, 2013. 2.1

[47] Swagatika Panda, AH Abdul Hafez, and CV Jawahar. Learning support order for manipulation in clutter. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 809–815. IEEE, 2013. 2.3.1, 3.3

[48] Pedro OO Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *Advances in Neural Information Processing Systems*, pages 1990–1998, 2015. 3.1.1

[49] Babak Rasolzadeh, Mårten Bj¨orkman, Kai Huebner, and Danica Kragic. An active vision system for detecting, fixating and manipulating objects in the real world. *The International Journal of Robotics Research*, 29(2-3):133–154, 2010. 1.2

[50] Heather Riley and Mohan Sridharan. Non-monotonic logical reasoning and deep learning for explainable visual question answering. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, pages 11–19, 2018. 2.3.1

[51] Benjamin Rosman and Subramanian Ramamoorthy. Learning spatial relationships between objects. *The International Journal of Robotics Research*, 30(11):1328–1342, 2011. 2.3.1

[52] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-icp. 5.2.1

[53] Tianjia Shao, Aron Monszpart, Youyi Zheng, Bongjin Koo, Weiwei Xu, Kun Zhou, and Niloy J Mitra. Imagining the unseen: Stability-based cuboid arrangements for scene understanding. *ACM Trans. on Graphics*, 33(6), 2014. 2.3.1

[54] Maximilian Sieb, Zhou Xian, Audrey Huang, Oliver Kroemer, and Katerina Fragkiadaki. Graph-structured visual imitation. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 979–989. PMLR, 30 Oct–01 Nov 2020. URL `http://proceedings.mlr.press/v100/sieb20a.html`. 6.2

[55] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012. 2.3.1, 4.1.3

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4.2.3

[57] Kristoffer Sjöö and Patric Jensfelt. Learning spatial relations from functional simulation. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1513–

1519. IEEE, 2011. 2.3.1

[58] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 3.1.2, 4.2.1

[59] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 4.2.1

[60] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ Int. Conference on Intelligent Robots and Systems (IROS)*, pages 23–30. IEEE, 2017. 4.1.2

[61] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012. 4.1.2

[62] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. In *Conference on Robot Learning (CoRL)*, 2018. URL `https://arxiv.org/abs/1809.10790`. 5.2.1

[63] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007. 4.2.2

[64] Misha Wagner, Hector Basevi, Rakshith Shetty, Wenbin Li, Mateusz Malinowski, Mario Fritz, and Ales Leonardis. Answering visual what-if questions: From actions to predicted scene descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2.2

[65] Jiuguang Wang, Philip Rogers, Lonnie Parker, Douglas Brooks, and Mike Stilman. Robot jenga: Autonomous and strategic block extraction. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5248–5253. IEEE, 2009. 1.2

[66] Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual interaction networks: Learning a physics simulator from video. In *Advances in neural information processing systems*, pages 4539–4547, 2017. 2.3.2

[67] Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. Learning to see physics via visual de-animation. In *Advances in Neural Information Processing Systems*, pages 153–164, 2017. 2.3.2

[68] Victoria Xia, Zi Wang, and Leslie Pack Kaelbling. Learning sparse relational transition models. *arXiv preprint arXiv:1810.11177*, 2018. 6.2

[69] Yufei Ye, Maneesh Singh, Abhinav Gupta, and Shubham Tulsiani. Compositional video prediction. 2019. 2.3.2

[70] Renqiao Zhang, Jiajun Wu, Chengkai Zhang, William T Freeman, and Joshua B Tenenbaum. A comparative evaluation of approximate probabilistic simulation and deep neural networks as accounts of human physical scene understanding. *arXiv preprint*

arXiv:1605.01138, 2016. 2.2

[71] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 5.4.3

[72] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019. 3.1.1