Sim-to-Real via Sim-to-Sim: Data-efficient Robotic Grasping via Randomized-to-Canonical Adaptation Networks

Stephen James¹, Paul Wohlhart², Mrinal Kalakrishnan², Dmitry Kalashnikov³, Alex Irpan³, Julian Ibarz³, Sergey Levine³,⁵, Raia Hadsell⁴, Konstantinos Bousmalis⁴

slj12@imperial.ac.uk, {wohlhart, kalakris}@x.team, {dkalashnikov, alexirpan, julianibarz, slevine, raia, konstantinos}@google.com,

Abstract

Real world data, especially in the domain of robotics, is notoriously costly to collect. One way to circumvent this can be to leverage the power of simulation to produce large amounts of labelled data. However, training models on simulated images does not readily transfer to real-world ones. Using domain adaptation methods to cross this “reality gap” requires a large amount of unlabelled real-world data, whilst domain randomization alone can waste modeling power. In this paper, we present Randomized-to-Canonical Adaptation Networks (RCANs), a novel approach to crossing the visual reality gap that uses no real-world data. Our method learns to translate randomized rendered images into their equivalent non-randomized, canonical versions. This in turn allows for real images to also be translated into canonical sim images. We demonstrate the effectiveness of this sim-to-real approach by training a vision-based closed-loop grasping reinforcement learning agent in simulation, and then transferring it to the real world to attain 70% zero-shot grasp success on unseen objects, a result that almost doubles the success of learning the same task directly on domain randomization alone. Additionally, by joint finetuning in the real-world with only 5,000 real-world grasps, our method achieves 91%, attaining comparable performance to a state-of-the-art system trained with 580,000 real-world grasps, resulting in a reduction of real-world data by more than 99%.

1. Introduction

Deep learning for vision-based robotics tasks is a promising research direction [58]. However, it necessitates large amounts of real-world data, which is a severe limitation, since real-robot data collection is expensive and cumbersome, often requiring days or even months for a single task [34, 44]. Due to the availability of affordable cloud computing services, it is becoming more attractive to leverage large-scale simulations to collect experience from a large number of agents in parallel. But with this comes the issue of transferring gained experience from simulation to the real world — a non-trivial task given the usually large domain shift.

Reducing the reality gap between simulation and reality is possible with recent advances in visual domain adaptation [14, 36, 5, 55, 4, 66, 71, 30, 54, 59, 21]. Such techniques usually require large amounts of unlabelled images from the real world. Although such unlabelled images are easier to capture than labelled, they can still be costly to collect in
robotics tasks. Domain randomization [51, 61, 25, 38, 3, 24] is another technique that is particularly popular in robotics, where an agent is trained on a wide range of variations of sensory inputs, with the intention that this forces the input processing layers of the network to extract semantically relevant features in a way that is agnostic to the superficial properties of the image (such as particular textures or particular ways shadows are cast from a constant light source). The intuition is that this leads to a network that extracts the same information from real-world images, featuring yet another variation of the input. However, performing randomization directly on the input of a learning algorithm, as done in related work, makes the task potentially harder than necessary, as the algorithm has to model both the arbitrary changes in the visual domain, while at the same time trying to decipher the dynamics of the task. Moreover, although randomization has been successful in the supervised learning setting, there is evidence that some popular reinforcement learning (RL) algorithms, such as DDPG [35] and A3C [39], can be destabilized by this transfer method [38, 70].

In this paper, we investigate learning vision-based robotic closed-loop grasping, where a robotic arm is tasked with picking up a diverse range of unseen objects, with the help of simulation and the use of as little real-world data as possible. Robotic grasping is an important application area in robotics, but also an exceptionally challenging problem: since a grasping system must successfully pick up previously unseen objects, it is not enough simply to memorize grasps that work well for individual instances, but to generalize and extrapolate from an internal understanding of geometry and physics. This presents a particularly difficult challenge for simulation-to-real-world transfer: besides the distributional shift from simulated images and physics, the system must also handle domain shift in the distribution of objects themselves.

To that end, we propose Randomized-to-Canonical Adaptation Networks (RCAN), a novel approach to crossing the reality gap that translates real-world images into their equivalent simulated versions, but makes use of no real-world data. This is achieved by leveraging domain randomization in a unique way, where we learn to adapt from one heavily randomized scene to an equivalent non-randomized, canonical version. We are then able to train a robotic grasping algorithm in a pre-defined canonical version of our simulator, and then use our RCAN model to convert the real-world images to the canonical domain our grasping algorithm was trained on.

Using RCAN along with a grasping algorithm that uses QT-Opt, a recent reinforcement learning algorithm, we achieve almost double the performance in comparison to alternative methods of using randomization. Bootstrapping from this performance, and with the addition of only 5,000 real-world grasps, we are able to achieve higher performance than a system trained with 580,000 real-world grasps. In our particular experiment, none of the objects used during testing are seen during either simulated training or real-world joint finetuning.

Our results also show that RCAN (summarized in Figure 1) is superior to learning a grasping network directly with domain randomization. RCAN has additional advantages compared to other simulation-to-real-world transfer methods. Firstly, unlike domain adaptation methods, it does not need any real-world data in order to learn our reality-to-simulation translation function. Secondly, RCAN gives an interpretable intermediate output that would otherwise not be available when performing domain randomization directly on the policy. Finally, as our method is trained in a supervised manner and preprocesses the input to the downstream task, it enables the use of RL methods that currently suffer from the stability issues when learning a policy directly from domain randomization [38, 70].

In summary, our contributions are as follows:

- We present a novel approach of crossing the reality gap by using an image-conditioned generative adversarial network (cGAN) [23] to transform randomized simulation images into their non-randomized, canonical versions, which in turn enables real-world images to also be transformed to canonical simulation versions.

- We show that by using this approach, we are able to train a state-of-the-art vision-based grasping reinforcement learning algorithm (QT-Opt) purely in simulation and achieve 70% success on the challenging task of grasping previously unseen objects in the real world, almost double the performance obtained by naively using domain randomization on the input of the learning algorithm.

- We also show that by using RCAN and joint finetuning in the real-world with only 5,000 additional grasping episodes we are able to increase grasping performance to 91%, outperforming QT-Opt when trained from scratch in the real-world with 580,000 grasps — a reduction of over 99% of required real-world samples.

2. Related Work

Robotic grasping is a well studied problem [2]. Traditionally, grasping was usually solved analytically, where 3D meshes of objects would be used to compute the stability of a grasp against external wrenches [45, 47] or constrain the object’s motion [47]. These solutions often assume that the same, or similar objects will be seen during testing, such that point clouds of the test objects can be matched
with stored objects based on visual and geometric similarity [6, 11, 19, 20, 29]. Due to this limitation, data-driven methods have become the dominant way to solve grasping [33, 37]. These methods commonly make use of either hand-labeled grasp positions [33, 28], self-supervision [44], or predicting grasp outcomes [34]. State-of-the-art grasping systems typically either operate in an open-loop style, where grasping locations are chosen, and then a motion is executed to complete the grasp [69, 41, 37, 60], or in a closed-loop manner, where grasp prediction is continuously run during motion, either explicitly [65], or implicitly [27].

Simulation-to-real-world transfer concerns itself with learning skills in simulation and then transferring them to the real world, which reduces the need for expensive real-data collection. However, it is often not possible to naively transfer such skills directly due to the visual and dynamics differences between the two domains [26]. Numerous works have looked into enabling such transfer both in computer vision and robotics. In the context of robotic manipulation in particular, Saxena et al. [53] used rendered objects to learn a vision-based grasping model. Rusu et al. [50] introduced progressive neural networks that help adapt an existing deep reinforcement learning policy trained from pixels in simulation to the real world for a reaching task. Other works have considered simulation-to-real-world transfer using only depth images [64, 18]. Although this may be an attractive option, using depth cameras alone is not suitable for all situations, and coupled with the low cost of simple RGB cameras, there is considerable value in studying transfer in systems that solely use monocular RGB images. Although in this work we use depth estimation from RGB input as an auxiliary task to aid with our randomized-to-canonical image translation model, we neither use depth sensors in the real world, nor do we use our estimated depth during training.

Data augmentation has been a standard tool in computer vision for decades. More recently, and as a way to avoid overfitting, the random application of cropping, flipping, and photometric variations to input images were used to train AlexNet [31] and many more subsequent deep learning models. In robotics, a number of recent works have examined using randomized simulated environments [61, 25, 38, 3, 24, 52] specifically for simulation-to-real world transfer for grasping and other similar manipulation tasks, extending on prior work on randomization for collision-free robotic indoor flight [51]. These works apply randomization in the form of random textures, lighting, and camera position, allowing the resulting algorithm to become invariant to domain differences and applicable to the real world. There have been more robotics works that do not use vision, but that apply domain randomization on physical properties of the simulator to aid transferability [40, 46, 1, 68, 43]. Recently, Chebotar et al. [9] have specifically looked into learning, from few real-world trajectories, the optimal distribution of such simulation properties, for transfer of policies learned in simulation to the real world. All of these methods learn a policy directly on randomization, whilst our method instead utilizes domain randomization in a novel way in order to learn a randomized-to-canonical adaption function to gain an interpretable intermediate representation and achieve superior results in comparison to learning directly on randomization.

Visual domain adaptation [42, 13] is a process that allows a machine learning model trained with samples from a source domain to generalize to a target domain, by utilizing existing but (mostly) unlabeled target data. In simulation-to-reality transfer, the source domain is usually the simulation, whereas the target is the real world. Prior methods can be split into: (1) feature-level adaptation, where domain-invariant features are learned between source and target domains [17, 15, 57, 7, 14, 36, 5, 55], or (2) pixel-level adaptation, which focuses on re-stylizing images from the source domain to make them look like images from the target domain [4, 66, 71, 30, 54, 59, 21]. Pixel-level domain adaptation differs from image-to-image translation techniques [23, 10, 67], which deal with the easier task of learning such a re-stylization from matching pairs of examples from both domains. Our technique can be seen as an image-to-image translation model that transforms randomized renderings from our simulator to their equivalent non-randomized, canonical ones.

In the context of robotics, visual domain adaptation has also been used for simulation-to-real-world transfer [62, 56, 3]. Bousmalis et al. [3], introduced the GraspGAN method, which combines pixel-level with feature-level domain adaptation to limit the amount of real data needed for learning grasping. Although the task is similar to ours, GraspGAN required significant amounts of unlabeled real-world data that were previously collected by a variety of pre-existing grasping networks. Our method can be viewed as orthogonal to existing domain adaptation methods and GraspGAN: the process of training the adapter could make use of unlabeled real-world data by incorporating ideas from domain adaptation in the form of additional auxiliary losses to improve performance further. Although in this work we do explore using our simulation-trained policy to collect labeled real-world data for joint finetuning, the combination with domain adaptation techniques is proposed as a promising future research direction.

The reverse, i.e. reality-to-simulation transfer, has been examined recently by Zhang et al. [70] in the context of a simple robotic driving task. The approach has certain advantages, namely the learning algorithm is trained only in simulation, and during inference the real-world images are adapted to look like simulated ones. This decouples adaptation from training and if the real-world environment
changes, it is only the adaptation model that needs to be re-learned. We also explore reality-to-simulation transfer, but unlike [70], which uses CyCaDA [21] and unlabeled real-world data, we do so only in simulation, by learning to adapt randomized images from our simulator to their equivalent non-randomized versions, which allows data-efficient transfer of our model to the real-world.

3. Background

We demonstrate our approach by using a recent reinforcement algorithm, *Q-function Targets via Optimization (QT-Opt)* [27], though our method is compatible with any reinforcement learning or imitation learning algorithm, as we are only adapting the input. QT-Opt is a state-of-the-art method for vision base grasping, which made it an ideal choice as a baseline for a direct comparison. Below, we will cover the fundamentals of Q-learning and then provide an overview of QT-Opt.

In reinforcement learning, we assume an agent interacting with an environment consisting of states $s \in S$, actions $a \in A$, and a reward function $r(s_t, a_t)$, where $s_t$ and $a_t$ are the state and action at time step $t$ respectively. The goal of the agent is then to discover a policy that results in maximizing the total expected reward. One way to achieve such a policy is to use the recently proposed QT-Opt [27] algorithm. QT-Opt is an off-policy, continuous-action generalization of Q-learning, where the goal is to learn a parametrized Q-function (or state-action value function). This can be learned by minimizing the Bellman error:

$$E(\theta) = \mathbb{E}_{(s, a, s') \sim \mathcal{D}} \left[ D(Q_{\theta}(s, a), Q_T(s, a, s')) \right],$$

where $Q_T(s, a, s') = r(s, a) + \gamma V(s')$ is a target value, and $D$ is a divergence metric, defined as the cross-entropy function in this case. Much like other works in RL, stability was improved by the introduction of two target networks. The target value $V(s')$ was computed via a combination of Polyak averaging and clipped double Q-learning to give $V(s') = \min_{i=1,2} Q_{\bar{\theta}}(s', \arg \max_a Q_{\bar{\theta}}(s', a'))$. QT-Opt differs from other methods primarily with regards to action selection. Rather than selecting actions based on the argmax: $\pi_{\bar{\theta}}(s) = \arg \max_a Q_{\bar{\theta}}(s, a)$, QT-Opt instead evaluates the argmax via a stochastic optimization algorithm over $a$; in this case, the cross-entropy method (CEM) [49].

4. Method

Our method, Randomized-to-Canonical Adaptation Networks (RCAN), consists of an image-conditioned generative adversarial network (cGAN) [23] that transforms images from randomized simulated environments (an example is show in Figure 2a) into images that seem similar to those obtained from a non-randomized, canonical one (Figure 2b). Once trained, the cGAN generator is also able to transform real-world images into images that seem as if they were obtained from the canonical simulation environment. We are then able to train a reinforcement learning algorithm (in this case QT-Opt) fully in simulation, and use such a generator to enable the trained policy to act in the real-world.

The approach assumes 3 domains: the randomized simulation domain, the canonical simulation domain, and the real-world domain. Let $\mathcal{D} = \{(x_s, x_c, m_c, d_c)\}_{j=1}^N$ be a dataset of $N$ training samples, where each sample is a tuple containing an RGB image $x_s$ from the randomization (source) domain, an RGB image $x_c$ from the canonical (target) domain (with semantic content, i.e. scene configuration, matching that of $x_s$), a segmentation mask $m_c$, and a depth image $d_c$. Both the segmentation mask and depth mask are only used as auxiliary tasks during the training of our generator. The RCAN generator function $G(x) \rightarrow \{x_a, m_a, d_a\}$, maps an image $x$ from any domain to an adapted image $x_a$, segmentation mask $m_a$, and depth image $d_a$, such that they appear to belong to the canonical domain.

4.1. RCAN Data Generation

In order to learn this translation $G$, we need pairs of observations capturing the robot in interaction with the scene, with one observation showing the scene in its canonical version and the other one showing the same scene but with randomization applied, as shown in image (a) and (b) of Figure 2. Our simulated environments are based on the Bullet physics engine and use the default renderer [12]. They are built to roughly correspond to the real word, and include a Kuka IIWA, a tray, an over-the-shoulder camera aimed
at the tray, and a set of graspable objects. Graspable objects consist of a combination of 1,000 procedurally generated objects (consisting of randomly merged geometric shapes), and 51,300 realistic objects from 55 categories obtained from the ShapeNet repository [8].

We create the trajectories from which we sample paired snapshots by running training of QT-Opt in simulation. At the beginning of each episode, the position of the divider in the tray is randomly sampled, and 5 randomly selected objects are dropped into the tray. Then, at each timestep we freeze the scene, apply a new arbitrary randomization (described below) to capture the randomized observation, reset to and capture an observation of the canonical version, and let QT-Opt proceed. In our case, observations consist of RGB images, depth, and segmentation masks, labeling each pixel with one of 5 categories: graspable objects, tray, tray snapshots by running training of shapes), and 51,300 realistic objects from 55 categories objects consist of a combination of 1,000 procedurally generated objects (consisting of randomly merged geometric shapes), and 51,300 realistic objects from 55 categories obtained from the ShapeNet repository [8].

The randomization includes applying at each timestep randomly selected textures from a set of over 5,000 images to all models, which includes the tray, graspable objects, arm segments, and floor. Additionally we randomize the position, direction and color of the lighting. To further increase the diversity of scene configurations beyond those that the normal robot operation during QT-Opt training gives us, we also slightly randomize the position and size of the arm and tray (sampling from a uniform distribution), applying the same transformation to both the canonical and the randomized scene when creating the snapshot, such that the semantics between the two still match.

One important question is: what should the canonical environment look like? In practice, the canonical environment can be defined in a number of ways. We opt for applying uniform colors to the background, tray and arm, while leaving the textures for the objects from the randomized version in-place, as this preserves the objects’ identity and thus opens up the potential for instance-specific grasping in future works. Each link of the arm is colored independently to aid tracking of individual links of the arm. We opt for fixing the light source in the canonical version, requiring the network to learn some aspect of geometry in order to re-render any shadows in the correct shape and direction.

4.2. RCAN Training Method

We aim to learn \( G(x_a) \rightarrow \{x_r, m_a, d_a\} \), which transforms randomized sim images into canonical sim images with matching semantics, with the intuition that the generator will generalize to accept an image from the real world \( x_r \), and produce a canonical RGB image, segmentation mask, and depth image: \( G(x_r) \rightarrow \{x_a, m_a, d_a\} \). To train the generator, we encourage visual equality between the generated \( x_a \) and target \( x_c \) through a loss function \( l_{eq,a} \), semantic equality between \( m_c \) and \( m_a \) through a function \( l_{eq,m} \), and depth equality between \( d_c \) and \( d_a \) through a function \( l_{eq,d} \). Having experimented with L1, L2, and the mean pairwise squared error (MPSE), our solution uses MPSE for \( l_{eq,d} \), which was found to converge faster with no loss in performance [5], along with the L2 distance for our auxiliary losses \( l_{eq,m} \) and \( l_{eq,d} \). This results in the following loss:

\[
L_{eq}(G) = \mathbb{E}_{(x_a, x_c, m_c, d_c)}[\lambda_x l_{eq,x}(G(x_a), x_c) + \lambda_m l_{eq,m}(G(x_a), m_c) + \lambda_d l_{eq,d}(G(x_a), d_c)],
\]

where \( G_x, G_m, \) and \( G_d \) denotes the image, mask, and depth element of the generator output respectively. In addition, \( \lambda_x, \lambda_m \) and \( \lambda_d \) represent the respective weightings.

It is well known that these equality losses can lead to blurry images [32], and so we employ a sigmoid-cross entropy generative adversarial (GAN) objective [16] to encourage high-frequency sharpness. Let \( D(x) \) be a discriminator that outputs the likelihood that a given image \( x \) is from the canonical domain. With this, the GAN is trained with the following objective:

\[
L_{GAN}(G, D) = \mathbb{E}_x[\log D(x)] + \mathbb{E}_x[\log(1 - D(G(x))]),
\]

where \( G_x \) denotes the image element of the generator output. The final objective for the generator then becomes:

\[
\hat{G} = \arg \min_G \max_D L_{GAN}(G, D) + \lambda l_{eq}(G).
\]

The generator \( G \) and discriminator \( D \) are parameterized by weights of a convolutional neural network; details of which are presented in Appendix A. Qualitative results of our generator can be seen in Figure 3 and on the project web-page\(^6\).

4.3. Real World Grasping with QT-Opt

We use QT-Opt for our grasping algorithm, and follow the same state and action definition as Kalashnikov et al. [27], where the state is defined as \( s_t = (x_t, g_{apt,t}, g_{height,t}) \) at each timestep \( t \), which includes a 472 × 472 image \( x_t \) taken from a mounted over-the-shoulder camera overlooking the work space, a binary open/close indicator of gripper aperture \( g_{apt,t} \), and the scalar height of the gripper above the bottom of the tray \( g_{height,t} \).

In our case, rather than sending the image directly to the RL algorithm, the image \( x_t \) is instead passed through the generator \( G \), and the resulting generated image \( x_a \) is extracted and concatenated, channel-wise, with the original source image \( x_t \). This results in the state \( s_t = [(G(x_t) + x_t), g_{apt,t}, g_{height,t}] \), where \( [G(x_t) + x_t] \) represents the concatenation. Note that we do not use the generated depth and segmentation masks of \( G \) as input to QT-Opt in order to make a fair comparison to Kalashnikov et al. [27], though these could also be added in practice. The action space of Kalashnikov et al. [27], which consists of gripper

\(^6\)https://sites.google.com/view/rcan/
(a) Randomized-to-canonical samples.  
(b) Real-to-canonical samples.

Figure 3: Sample outputs of our trained generator $G$ when given randomized sim images (3a) and real images (3b). Note the accuracy of the reconstruction of the canonical images from real-world images in complex and cluttered scenes, along with shadows being re-rendered into the canonical representation. However, also note that randomized-to-canonical adaptation performs a noticeably better reconstruction of the gripper in comparison to the real-to-canonical adaptation. This leads to the failure cases discussed in Section 5. The generated depth and segmentation masks are used as auxiliaries during training of the generator. Further examples can be seen in Figure 8 of the Appendix.

5. Experiments

Our experimental section aims to answer the following questions: (1) Can we train an agent to grasp arbitrary unseen objects without having seen any real-world images?

(2) How does $QT$-Opt perform with standard domain randomization, and can our method perform better than this?

(3) Does the addition of real-world on-policy training of our method lead to higher grasping performance while still drastically reducing the amount of real-world data required? We answer these questions through a series of rigorous real-world vision-based grasping experiments across multiple Kuka IIWA robots.

5.1. Evaluation Protocol

During evaluation, each robot attempts 102 grasps on its own set of 5 to 6 previously unseen test objects (shown in Figure 5 of the Appendix) which are deposited into each robots’ respective tray and remain constant across all evaluations. Each grasp attempt (episode) consists of at most 20 time steps. If after 20 time steps no object has been grasped, the attempt is regarded as a failure. Following a grasp attempt, the object is deposited back into the tray at a random location. Although grasping was done with replacement, in practice, $QT$-Opt was not found attempting a grasp on the same object multiple times in a row. All observations come from an over-the-shoulder RGB camera.

5.2. Results

We first focus on the first 4 columns of Table 1. The first row of this section shows the results of $QT$-Opt reported in Kalashnikov et al. [27]; where following 580,000 off-policy real-world grasps, a performance of 87% was achieved.
The **Canonical Sim** data source (second row) takes **QT-Opt** trained in the canonical simulation environment and then runs this directly in the real-world. The low success rate of 21% shows the existence of the reality gap. The following three rows show the result of training **QT-Opt** directly on varying degrees of randomization: mild, medium and heavy. **Mild randomization** consists of varying tray texture, object texture and color, robot arm color, lighting direction and brightness, and a background image consisting of 6 different images from the view of the real-world camera. **Medium randomization** adds a diverse mix of background images to the floor. Finally, **heavy randomization** uses the same scheme used to train **RCAN**, explained in Section 4.1.

Surprisingly, an unexpected discovery was that **QT-Opt** responds well to heavy domain randomization during training (i.e. is not destabilized). This is contrary to other RL methods, such as DDPG [35] and A3C [39], where heavy domain randomization has been shown to cause training to fail [38, 70]. Although **QT-Opt** was able to train stably with randomization, the results show that this does not lead to a successful transfer, achieving between 33% and 37% zero-shot grasping performance, whereas **RCAN** achieves **70%**: over **double** the success in the real world. This success highlights that **RCAN** better utilizes domain randomization to achieve sim-to-real transfer, rather than training a policy directly on domain randomization.

We now focus on the remaining 2 columns, that is, the ability to jointly finetune on a small amount of real-world on-policy grasps. We chose to use 5,000 to represent “small”, which is less than 1% of the 580,000 grasps used in Kalashnikov et al. [27] for the off-policy training and takes only a day to collect, instead of months. To make comparison easier, in addition to reporting the 28,000 on-policy grasps for joint finetuning from [27], we also report the performance after 5,000 grasps. This baseline result of 85% suggests that 5,000 real-world grasps for joint finetuning a system already trained with 580,000 does not improve performance. For the next joint finetuning experiment, we take each of the agents that were trained directly on domain randomization, and jointly finetune them on 5,000 real grasps, achieving between 77% and 85% grasping success. The rapid increase of ∼50p.p. is very surprising, and to the best of our knowledge, no other related works have shown such a dramatic performance increase from pre-training on domain randomization.

Finally, we look at joint finetuning **RCAN** with 5,000 and 28,000 real grasps, where the real images are adapted by the generator and then both the source and adapted image are passed to the grasping network; in this case, the gradients are only applied to the grasping network and not the generator network. The result of 91% for 5,000 shows that the improvement over learning directly on domain randomization holds, though for this result the difference is much smaller. What we believe is incredibly encouraging for the robotics community, is that with **91% RCAN outperforms** a version of **QT-Opt** that was trained on 580,000 real-world grasps, while using **less than 1% of the data**. Moreover, following joint finetuning with with the same number of online grasps as Kalashnikov et al. [27] (28,000), we are able to achieve an almost equal grasp performance of 94%.

<table>
<thead>
<tr>
<th><strong>QT-Opt Data Source</strong></th>
<th><strong>Offline Real Grasps</strong></th>
<th><strong>Performance In Sim</strong></th>
<th><strong>Performance In Real</strong></th>
<th><strong>Online Real Grasps</strong></th>
<th><strong>Performance In Real</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>580,000</td>
<td>-</td>
<td>87%</td>
<td>+5,000</td>
<td>85%</td>
</tr>
<tr>
<td>Canonical Sim</td>
<td>0</td>
<td>99%</td>
<td>21%</td>
<td>+5,000</td>
<td>30%</td>
</tr>
<tr>
<td>Mild Randomization</td>
<td>0</td>
<td>98%</td>
<td>37%</td>
<td>+5,000</td>
<td>85%</td>
</tr>
<tr>
<td>Medium Randomization</td>
<td>0</td>
<td>98%</td>
<td>35%</td>
<td>+5,000</td>
<td>77%</td>
</tr>
<tr>
<td>Heavy Randomization</td>
<td>0</td>
<td>98%</td>
<td>33%</td>
<td>+5,000</td>
<td>85%</td>
</tr>
<tr>
<td><strong>RCAN</strong></td>
<td>0</td>
<td>99%</td>
<td><strong>70%</strong></td>
<td>+5,000</td>
<td><strong>91%</strong></td>
</tr>
</tbody>
</table>

Table 1: Average grasp success rate on test objects after 102 grasp attempts on each of the multiple Kuka II WA robots. The first 4 columns of the table highlight the performance after training on a specified number of real world grasps. Zero grasps implies that all training was done in simulation. The last 2 columns highlight the results of on-policy joint finetuning on a small amount of real-world grasps.
to the generator network, thus simplifying the task for the

size that our method allows offloading visual complexity

to a policy directly with domain randomization? W e hypoth-

ample: why does our method perform better than learning

5.4. Discussion

cimplified, yet potentially incomplete adapted view and the

al image to the agent offers the best combination of a

data, feeding both the output of

we believe that, with a limited amount of labeled real-world

to extract information from for each part of the image as the

grasping algorithm the option to choose which data source

to translate randomized simulation images into a canoni-

sion Networks (RCAN), a sim-to-real method that learns

to translate randomized simulation images into a canonical

representation, which in turn allows for real-world im-

gages to also be translated to this canonical representation.

Given that our grasping algorithm (QT-Opt) is trained in

this canonical environment, it is possible to run policies

trained in simulation in the real world. W e show that this

approach is superior to the common domain randomization

approach, and argue that it is a much more meaningful use

of domain randomization. This general style of transfer has

applications beyond just grasping, and can be used in other

settings where real world data is expensive to collect, for ex-

ample, producing segmentation masks for self-driving cars.

For future work, we wish to explore further ways of intro-

ducing unlabelled real-world data in order to improve the

real-to-canonical translation. Moreover, we are interested

in exploring the effect of using the auxiliary outputs as ad-

ditional inputs to the grasping network.

6. Conclusion

We have presented Randomized-to-Canonical Adapta-

tion Networks (RCAN), a sim-to-real method that learns
to translate randomized simulation images into a canoni-
cal representation, which in turn allows for real-world im-

gages to also be translated to this canonical representation.

Given that our grasping algorithm (QT-Opt) is trained in

this canonical environment, it is possible to run policies

trained in simulation in the real world. W e show that this

approach is superior to the common domain randomization

approach, and argue that it is a much more meaningful use

of domain randomization. This general style of transfer has

applications beyond just grasping, and can be used in other

settings where real world data is expensive to collect, for ex-

ample, producing segmentation masks for self-driving cars.

For future work, we wish to explore further ways of intro-

ducing unlabelled real-world data in order to improve the

real-to-canonical translation. Moreover, we are interested

in exploring the effect of using the auxiliary outputs as ad-

ditional inputs to the grasping network.

Acknowledgments

W e would like to give special thanks to Ivonne Fajardo,
Peter Pastor, Iñaki Gonzalo and Benjamin Swanson for
overseeing the robot operations, Yunfei Bai for discussion
on PyBullet, and Serkan Cabi for valuable comments on the

paper.

References

[1] Rika Antonova, Silvia Cruciani, Christian Smith, and Dan-
ica Kragic. Reinforcement learning for pivoting task.

Figure 4: A graph showing how the performance of RCAN
and directly learning a policy on domain randomization
varies with the number of real-world on-policy grasps.

encouraging, as we ultimately wish to limit the amount of
real-world data that we are reliant on.

5.3. Failure cases

A large contributing factor to QT-Opts 96% grasp suc-

cess, was its ability to perform corrective behaviors, re-

grasping, probing motions to ascertain the best grasp, and

non-prehensile repositioning of objects. Much of this abili-

ty remained with our approach, except for the regrasping

ability. This powerful ability allows the policy to detect
when there is no object in the closed gripper, and thus, it can
decide to re-open it in an attempt to try and re-grasp. Given
that our method is not perfect at translating real-world im-
gees into simulation ones, artifacts may arise. As objects
that we grasp are often small, it can be very difficult for
the agent to differentiate between artifacts in the image or if
there is indeed an object in the gripper. W e observe this to
be detrimental to the agents ability to perform regrasping,
resulting in only a small amount of regrasps. The main ob-

ervation from joint finetuning our method with 5,000 real-

world grasps, is the re-emergence of the regrasping. W e be-

lieve that this is contributed by our decision to concatenate
the source image to the generated ones, and thus giving the
grasping algorithm the option to choose which data source
to extract information from for each part of the image as the
joint finetuning continues. W e hypothesize, that as the num-

ber of joint finetuning grasps increase, the network would
eventually learn to solely rely on the source (real-world)
image, rather than the adapted simulation image. However,
we believe that, with a limited amount of labeled real-world
data, feeding both the output of RCAN as well as the origi-
nal image to the agent offers the best combination of a
simplified, yet potentially incomplete adapted view and the
complex, but complete original real-world view.

5.4. Discussion

A number of questions arise from these results. For ex-

ample: why does our method perform better than learning

a policy directly with domain randomization? W e hypoth-
esize that our method allows offloading visual complexity
to the generator network, thus simplifying the task for the

grasping network and in turn, leading to a higher grasping
success. Moreover, having a chosen canonical environment
allows us to impose structure on the task which may be
beneficial for training the grasping network. Despite our
method achieving over double the zero-shot performance in
the real world in comparison to domain randomization, with
5,000 additional real-world grasps, the performance of di-
rect domain randomization also achieves a surprisingly high
performance. This leads us to the hypothesis that learning
a policy directly on domain randomization can act as a very
powerful pre-training regime, where the network is forced
to learn a very general feature extractor that can be easily
jointly finetuned to a new environment. Having said that,
our method outperforms this and has the added benefit of
giving us an interpretable output for sim-to-real transfer.

Another question for future work would be: is there a
way to better utilize the data collected during the 5,000 on-
policy grasps? Given this real-world data, it is now possible
to consider fusing ideas from other transfer methods that
require some real-world data, such as PixelIDA [5].


