

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 Levine^{3,5}, 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 in order 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 at best a large amount of unlabelled real-world data, whilst domain randomization alone can waste modeling power, rendering certain reinforcement learning (RL) methods unable to learn the task of interest. 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%, outperforming 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 rather promising research direction [57]. However, it necessitates

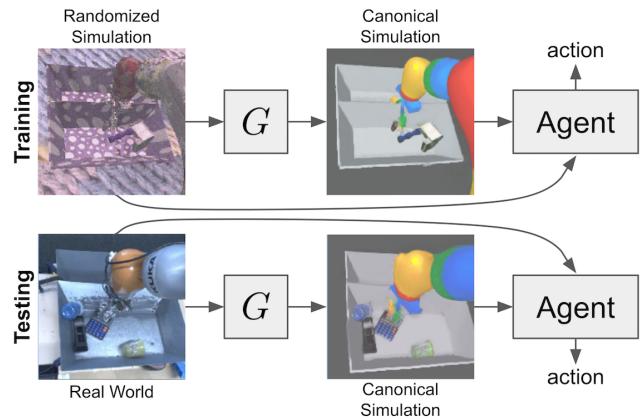


Figure 1: We learn a generator that translates randomized simulation images to a chosen canonical simulation version which are then used to train a robot grasping agent (top). The system can then be used to translate real-world images to canonical images, and consequently allow for Sim-to-Real transfer of the agent (bottom). Feeding both source and target images to the agent allows for joint finetuning of the agent in the real world.

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, 54, 4, 65, 70, 30, 53, 58, 21]. Such techniques usually require large amounts of unlabelled images from the real world. Although such unlabelled images are easier to

¹Imperial College London. Work done while Stephen James was at X

²X, Mountain View, California, United States

³Google Brain, United States

⁴DeepMind, London

⁵University of California Berkeley, Berkeley, California, United States

capture than labelled, they can still be costly to collect in robotics tasks. Domain randomization [51, 60, 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, 69].

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. Bootstrap-

ping 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, 69].

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 location are chosen, and then a motion is executed to complete the grasp [68, 41, 37, 59], or in a closed-loop manner, where grasp prediction is continuously run during motion, either explicitly [64], 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.* [52] 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 [63, 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 samples horizontally, 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 [60, 25, 38, 3, 24] 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, 67, 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, 56, 7, 14, 36, 5, 54], 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, 65, 70, 30, 53, 58, 21]. Pixel-level domain adaptation differs from image-to-image translation techniques [23, 10, 66], 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 [61, 55, 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: any available unlabeled and labeled real-world data can be trivially exploited to improve performance even 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.* [69] 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 [69], 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. 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 $\mathbf{s} \in \mathcal{S}$, actions $\mathbf{a} \in \mathcal{A}$, and a reward function $r(\mathbf{s}_t, \mathbf{a}_t)$, where \mathbf{s}_t and \mathbf{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:

$$\mathcal{E}(\theta) = \mathbb{E}_{(\mathbf{s}, \mathbf{a}, \mathbf{s}') \sim p(\mathbf{s}, \mathbf{a}, \mathbf{s}')} [D(Q_\theta(\mathbf{s}, \mathbf{a}), Q_T(\mathbf{s}, \mathbf{a}, \mathbf{s}'))], \quad (1)$$

where $Q_T(\mathbf{s}, \mathbf{a}, \mathbf{s}') = r(\mathbf{s}, \mathbf{a}) + \gamma V(\mathbf{s}')$ 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 $V(\mathbf{s}')$ was computed via a combination of Polyak averaging and clipped double Q-learning to give $V(\mathbf{s}') = \min_{i=1,2} Q_{\bar{\theta}_i}(\mathbf{s}', \arg \max_{\mathbf{a}'} Q_{\bar{\theta}_1}(\mathbf{s}', \mathbf{a}'))$. *QT-Opt* differs to other methods primarily with regards to action selection. Rather than selecting actions based on the argmax: $\pi_{\bar{\theta}_1}(\mathbf{s}) = \arg \max_{\mathbf{a}} Q_{\bar{\theta}_1}(\mathbf{s}, \mathbf{a})$, *QT-Opt* instead evaluates the argmax via a stochastic optimization algorithm over \mathbf{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 shown 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

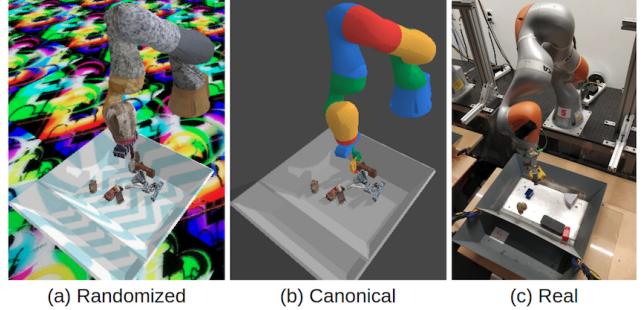


Figure 2: The setup used in our approach. A dataset of observations from a randomized version of a simulated environment (a) are paired with observations from a canonical version of the same environment (b) in order to learn an adaptation function and allow observations from the real-world (c) to be transformed into observations looking as if they came from the canonical simulation environment.

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 $\mathbb{D} = \{(x_s, x_c, m_c, d_c)_j\}_{j=1}^N$ 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 showing the scene in its canonical version and the other one showing the same scene but with randomization applied, as shown in 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 world, 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

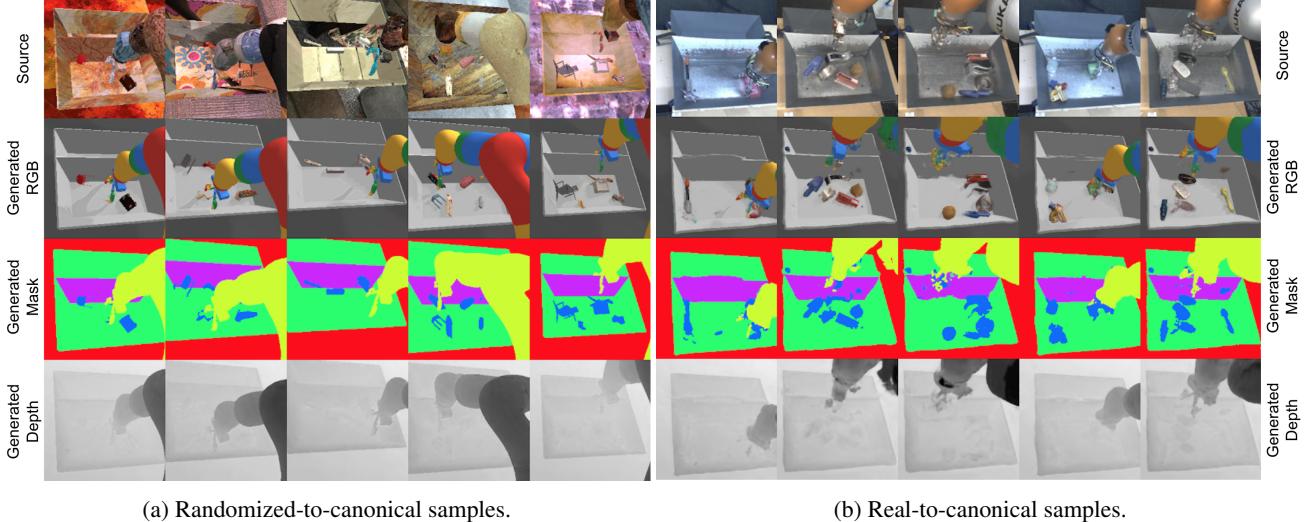


Figure 3: Sample outputs of our trained generator G when given randomized sim images (8a) and real images (8b). Note the accuracy of the reconstruction of the canonical images from real-world images in complex and cluttered scenes, along with shadows being re-rendered into the canonical representation. 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.

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 divider, robot arm, and background.

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 ver-

sion 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, in order to enforce the network to learn tracking of individual link 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 $G(x_s) \rightarrow \{x_a, m_a, d_a\}$, 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_x} , 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_x} 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:

$$\mathcal{L}_{eq}(G) = \mathbb{E}_{(x_s, x_c, m_c, d_c)} [\lambda_x l_{eq_x}(G(x_s), x_c) + \lambda_m l_{eq_m}(G_m(x_s), m_c) + \lambda_d l_{eq_d}(G_d(x_s), d_c)], \quad (2)$$

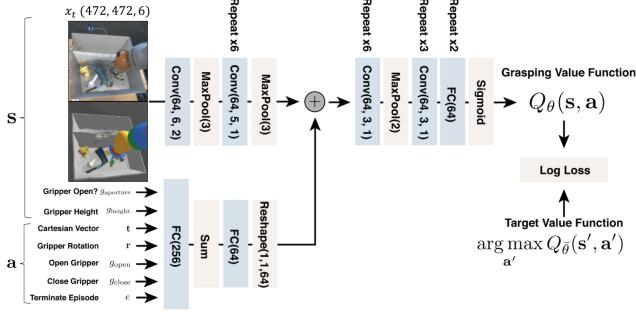


Figure 4: The Q-function of the grasping algorithm. The source image x (either from the randomized domain or real-world domain) and generated canonical image x_a are concatenated (channel-wise) and processed by a convolutional neural network (and fused with action and state variables) to produce a scalar representing the Q value $Q_\theta(s, a)$.

where G_x , G_m , and G_d denotes the image, mask, and depth element of the generator output respectively. In addition, λ_x , λ_m and λ_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:

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_x[\log D(x)] + \mathbb{E}_x[\log(1 - D(G_x(x)))] \quad (3)$$

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

$$\hat{G} = \arg \min_G \max_D \mathcal{L}_{GAN}(G, D) + \mathcal{L}_{eq}(G) \quad (4)$$

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-page⁶.

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 $s_t = (x_t, g_{aperture,t}, g_{height,t})$ 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 $g_{aperture,t}$, 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 $s_t = ([G(x_t) + x_t], g_{aperture,t}, g_{height,t})$, 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 pose displacement and an open/close command, remains unchanged. A summary of the Q-function is shown in Figure 4, and further details of the action space and architecture can be found in Appendix B.

In Kalashnikov *et al.* [27], the authors take their agent that was trained with 580,000 off-policy real-world grasps, and jointly finetune with an additional 28,000 on-policy grasps. During this joint finetuning process, *QT-Opt* asynchronously updates target values, collects real on-policy data, reloads real off-policy (offline) data from past experiences, and then trains the Q-network on both the on and off policy data streams within a distributed optimization framework. In the case of jointly finetuning *RCAN*, we also collect real on-policy data, but rather than using real-world past experiences (which we assume we do not have), we instead leverage the power of our simulation to continuously generate on-policy simulation data, and instead train on these streams of data. During the real world on-policy collection of both approaches, a selection of about 1,000 diverse training objects are used; a sample of which are shown in Figure 5. Between 5 and 10 objects are randomly chosen every few hours to be placed in each of the trays until the desired number of joint finetuning grasps are reached.

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) 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

⁶<https://sites.google.com/view/rcan/>

<i>QT</i> -Opt Data Source	Offline Real Grasps	Performance In Sim	Performance In Real	Online Real Grasps	Performance In Real
Real	580,000	-	87%	+5,000 +28,000	85% 96%
Canonical Sim	0	99%	21%	+5,000	30%
Mild Randomization	0	98%	37%	+5,000	85%
Medium Randomization	0	98%	35%	+5,000	77%
Heavy Randomization	0	98%	33%	+5,000	85%
<i>RCAN</i>	0	99%	70%	+5,000 +28,000	91% 94%

Table 1: Average grasp success rate on test objects after 102 grasp attempts on each of the multiple Kuka IIWA 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.



Figure 5: Real-world grasping objects that range greatly in size and appearance. *Left*: about 1000 visually and physically diverse training objects used for joint finetuning. *Right*: the unseen test objects.

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, 69]. 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

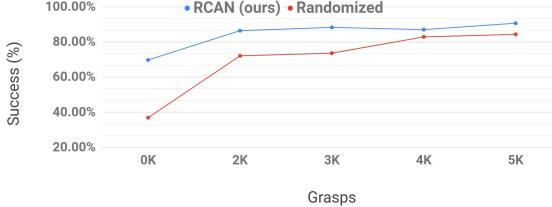


Figure 6: A graph showing how the performance of *RCAN* and directly learning a policy on domain randomization varies with the number of real world on-policy grasps.

85% suggest 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 $\sim 50\%$ 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 results 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 the same number of on-line grasps as Kalashnikov *et al.* [27] (28,000), we are able to achieve an almost equal grasp performance of 94%.

In order to understand how performance varies as we progress from 0 to 5,000 on-policy grasps, we repeat the evaluation protocol set above for intermediate checkpoints. We re-evaluate both agents at every 1,000 grasps for both *RCAN* and *Mild Randomization*. The results, presented in Figure 6, show that the majority of the success is gained within the first 2,000 grasps for both approaches. This is 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-Opt*'s 96% grasp success, was its ability to perform corrective behaviors, regrasping, probing motions to ascertain the best grasp, and non-prehensile repositioning of objects. Much of this ability 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 images 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. We observe this to be detrimental to the agents ability to perform regrasping, resulting in only a small amount of regrasps. The main observation from joint finetuning our method with 5,000 real-world grasps, is the re-emergence of the regrasping. We believe 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. We hypothesize, that as the number 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 original 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 example: why does our method perform better than learning a policy directly with domain randomization? We hypothesize 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 direct 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 PixelDA [5].

6. Conclusion

We have presented Randomized-to-Canonical Adaptation 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 images 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. We 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 example, producing segmentation masks for self-driving cars. For future work, we wish to explore further ways of introducing 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 additional inputs to the grasping network.

Acknowledgments

We would like to give special thanks to Ivonne Fajardo and Inaki Gonzalo for overseeing the robot operations, Yunfei Bai for discussion on PyBullet, and Serkan Cabi for valuable comments on the paper.

References

- [1] R. Antonova, S. Cruciani, C. Smith, and D. Kragic. Reinforcement learning for pivoting task. *arXiv:1703.00472*, 2017.
- [2] J. Bohg, A. Morales, T. Asfour, and D. Kragic. Data-driven grasp synthesis—a survey. *IEEE Transactions on Robotics*, 30(2):289–309, 2014.
- [3] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, S. Levine, and V. Vanhoucke. Using Simulation and Domain Adaptation to Improve Efficiency of Deep Robotic Grasping. *ICRA*, 2018.
- [4] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial neural networks. In *CVPR*, 2017.
- [5] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. *NIPS*, 2016.
- [6] P. Brook, M. Ciocarlie, and K. Hsiao. Collaborative grasp planning with multiple object representations. *ICRA*, 2011.
- [7] R. Caseiro, J. F. Henriques, P. Martins, and J. Batista. Beyond the shortest path: Unsupervised Domain Adaptation by Sampling Subspaces Along the Spline Flow. In *CVPR*, 2015.
- [8] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An information-rich 3D model repository. *arXiv:1512.03012*, 2015.
- [9] Y. Chebotar, A. Handa, V. Makoviychuk, M. Macklin, J. Isaac, N. Ratliff, and D. Fox. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. *arXiv:1810.05687*, 2018.
- [10] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. *arxiv:1707.09405*, 2017.
- [11] M. Ciocarlie, K. Hsiao, E. G. Jones, S. Chitta, R. B. Rusu, and I. A. Sucan. Towards reliable grasping and manipulation in household environments. In *Experimental Robotics*, pages 241–252. Springer, 2014.
- [12] E. Coumans and Y. Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. <http://pybullet.org>, 2016–2018.
- [13] G. Csurka. Domain adaptation for visual applications: A comprehensive survey. *arxiv:1702.05374*, 2017.
- [14] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 2016.
- [15] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.
- [16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *NIPS*, 2014.
- [17] R. Gopalan, R. Li, and R. Chellappa. Domain Adaptation for Object Recognition: An Unsupervised Approach. In *ICCV*, 2011.
- [18] M. Gualtieri, A. ten Pas, K. Saenko, and R. Platt. High precision grasp pose detection in dense clutter. In *IROS*, pages 598–605, 2016.
- [19] C. Hernandez, M. Bharatheesha, W. Ko, H. Gaiser, J. Tan, K. van Deurzen, M. de Vries, B. Van Mil, J. van Egmond, R. Burger, et al. Team delfts robot winner of the amazon picking challenge 2016. In *Robot World Cup*, pages 613–624. Springer, 2016.
- [20] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. *ICCV*, 2011.
- [21] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, 2018.
- [22] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*, 2015.
- [23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.
- [24] S. James, M. Bloesch, and A. J. Davison. Task-embedded control networks for few-shot imitation learning. *CoRL*, 2018.
- [25] S. James, A. J. Davison, and E. Johns. Transferring end-to-end visuomotor control from simulation to real world for a multi-stage task. *CoRL*, 2017.
- [26] S. James and E. Johns. 3d simulation for robot arm control with deep q-learning. *NIPS Workshop: Deep Learning for Action and Interaction*, 2016.

- [27] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine. QT-Opt: Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation. jun 2018.
- [28] D. Kappler, J. Bohg, and S. Schaal. Leveraging big data for grasp planning. *ICRA*, 2015.
- [29] B. Kehoe, A. Matsukawa, S. Candido, J. Kuffner, and K. Goldberg. Cloud-based robot grasping with the google object recognition engine. 2013.
- [30] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. *ICML*, 2017.
- [31] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [32] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv:1512.09300*, 2015.
- [33] I. Lenz, H. Lee, and A. Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.
- [34] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen. Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection. *International Symposium on Experimental Robotics*, 2016.
- [35] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv:1509.02971*, 2015.
- [36] M. Long and J. Wang. Learning transferable features with deep adaptation networks. *ICML*, 2015.
- [37] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *RSS*, 2017.
- [38] J. Matas, S. James, and A. J. Davison. Sim-to-real reinforcement learning for deformable object manipulation. *CoRL*, 2018.
- [39] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *ICML*, 2016.
- [40] I. Mordatch, K. Lowrey, and E. Todorov. Ensemble-cio: Full-body dynamic motion planning that transfers to physical humanoids. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 5307–5314. IEEE, 2015.
- [41] D. Morrison, A. W. Tow, M. McTaggart, R. Smith, N. Kelly-Boxall, S. Wade-McCue, J. Erskine, R. Grinover, A. Gurman, T. Hunn, et al. Cartman: The low-cost cartesian manipulator that won the amazon robotics challenge. *arXiv:1709.06283*, 2017.
- [42] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015.
- [43] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *ICRA*, pages 1–8. IEEE, 2018.
- [44] L. Pinto and A. Gupta. Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours. In *International Conference on Robotics and Automation*, 2016.
- [45] D. Prattichizzo and J. C. Trinkle. Grasping. In *Springer handbook of robotics*, pages 671–700. Springer, 2008.
- [46] A. Rajeswaran, S. Ghotra, B. Ravindran, and S. Levine. Epopt: Learning robust neural network policies using model ensembles. *arXiv:1610.01283*, 2016.
- [47] A. Rodriguez, M. T. Mason, and S. Ferry. From caging to grasping. *The International Journal of Robotics Research*, 31(7):886–900, 2012.
- [48] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015.
- [49] R. Y. Rubinstein and D. P. Kroese. The cross-entropy method: A unified approach to monte carlo simulation, randomized optimization and machine learning. *Information Science & Statistics, Springer Verlag, NY*, 2004.
- [50] A. A. Rusu, M. Vecerik, T. Rothörl, N. Heess, R. Pascanu, and R. Hadsell. Sim-to-real robot learning from pixels with progressive nets. *CoRL*, 2017.
- [51] F. Sadeghi and S. Levine. CAD2RL: Real single-image flight without a single real image. In *RSS*, 2017.
- [52] A. Saxena, J. Driemeyer, and A. Y. Ng. Robotic grasping of novel objects using vision. *IJRR*, 2008.
- [53] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, 2017.
- [54] R. Shu, H. Bui, H. Narui, and S. Ermon. A DIRT-t approach to unsupervised domain adaptation. In *ICLR*.
- [55] G. J. Stein and N. Roy. Genesis-rt: Generating synthetic images for training secondary real-world tasks. In *ICRA*, pages 7151–7158. IEEE, 2018.
- [56] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *AAAI*. 2016.
- [57] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, et al. The limits and potentials of deep learning for robotics. *IJRR*, 37(4-5):405–420, 2018.
- [58] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. *ICLR*, 2017.
- [59] A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt. Grasp pose detection in point clouds. *The International Journal of Robotics Research*, 36(13-14):1455–1473, 2017.
- [60] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IROS*, 2017.
- [61] E. Tzeng, C. Devin, J. Hoffman, C. Finn, P. Abbeel, S. Levine, K. Saenko, and T. Darrell. Adapting deep visuomotor representations with weak pairwise constraints. *WAFR*, 2016.
- [62] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv:1607.08022*, 2016.

- [63] U. Viereck, A. t. Pas, K. Saenko, and R. Platt. Learning a visuomotor controller for real world robotic grasping using easily simulated depth images. In *CoRL*, 2017.
- [64] U. Viereck, A. ten Pas, K. Saenko, and R. Platt. Learning a visuomotor controller for real world robotic grasping using simulated depth images. *CoRL*, 2017.
- [65] Z. Yi, H. R. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. *ICCV*, 2017.
- [66] D. Yoo, N. Kim, S. Park, A. S. Paek, and I. S. Kweon. Pixel-Level Domain Transfer. *arxiv:1603.07442*, 2016.
- [67] W. Yu, J. Tan, C. K. Liu, and G. Turk. Preparing for the unknown: Learning a universal policy with online system identification. In *RSS*, 2017.
- [68] A. Zeng, S. Song, S. Welker, J. Lee, A. Rodriguez, and T. Funkhouser. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. *arXiv:1803.09956*, 2018.
- [69] J. Zhang, L. Tai, Y. Xiong, M. Liu, J. Boedecker, and W. Burgard. Vr goggles for robots: Real-to-sim domain adaptation for visual control. *arXiv:1802.00265*, 2018.
- [70] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV*, 2017.

A. RCAN Architecture

The generator G is parameterized by weights of a convolutional neural network, summarized in Figure 7, and follows a U-Net style architecture [48] with downsampling performed via 3×3 convolutions with stride 2 for the first 2 layers, and average pooling with 3×3 convolution of stride 1 for the remaining layers. Upsampling was performed via bilinear upsampling, followed by a 3×3 convolutions of stride 1, and skip connections were fused back into the network via channel-wise concatenation, followed by a 1×1 convolution. All layers were followed by instance normalization [62] and ReLU non-linearities. The discriminator D is also parameterized by weights of a convolutional neural network with 2 layers of 32, 3×3 filters, followed by a layer of 64, 3×3 filters, and finally a layer of 128, 3×3 filters. The network follows a multi-scale patch-based design [3], where 3 scales of 472×472 , 236×236 , and 118×118 , are used to produce domain estimates for all patches which are then combined to compute the joint discriminator loss.

B. QT-Opt Architecture

The action space of [27], which consists of gripper pose displacement and an open/close command, remains unchanged in our paper, and is defined as $\mathbf{a}_t = (\mathbf{t}_t, \mathbf{r}_t, g_{\text{close},t}, g_{\text{open},t}, e_t)$, containing Cartesian translation $\mathbf{t}_t \in \mathbb{R}^3$, sine-cosine rotation encoding $\mathbf{r}_t \in \mathbb{R}^2$, a one-hot vector gripper open/close command $[g_{\text{close},t}, g_{\text{open},t}] \in \{0, 1\}^2$, and a learned stopping criterion e_t . The reward function is sparse, consisting of a reward of 1 following a successful grasp, or 0 for an unsuccessful grasp, and -0.05 on all other transitions. Summarized in Figure 4, the Q-function follows the same architecture as [27] (originally inspired by [34]).

Rather than a single RGB image input, our network takes in a 6 channel image, consisting of channel-wise concatenation of the source image x (either from the randomized domain or real-world domain) and generated image x_a . Features are extracted from these images via 7 convolutional layers and then merged with a transformed action and state vector (which have passed through 2 fully-connected layers) via element-wise addition. The merged streams are then processed by a further 9 convolution layers and 2 fully-connected layers, resulting in a scalar output representing the Q value $Q_\theta(s, a)$. Each layer, excluding the final, uses batch normalization [22] and ReLU non-linearities.

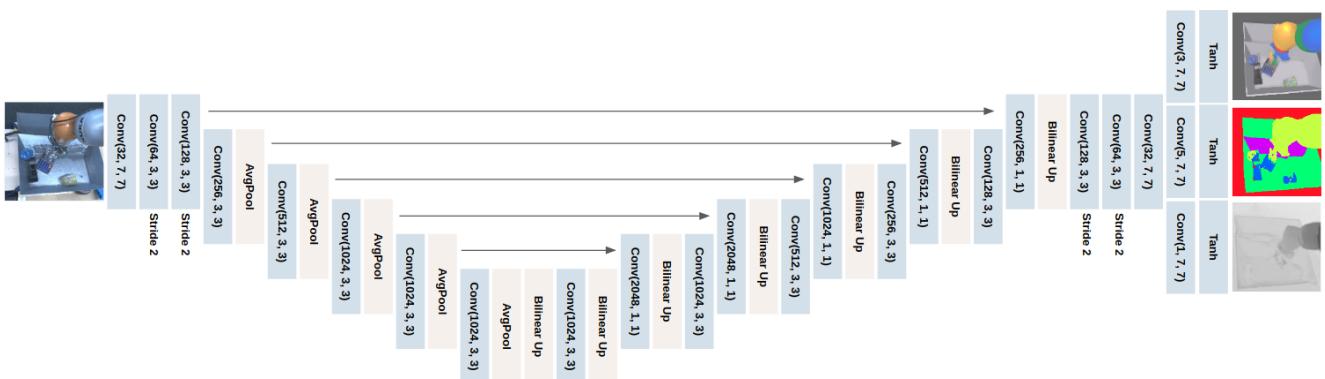
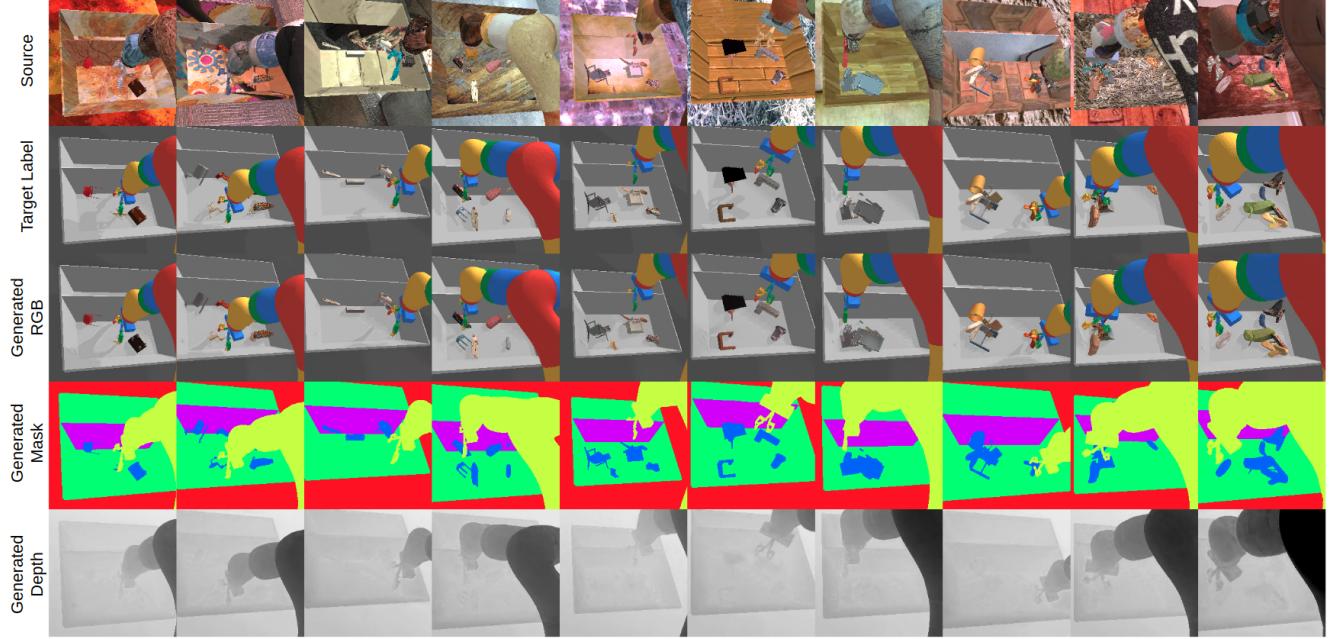
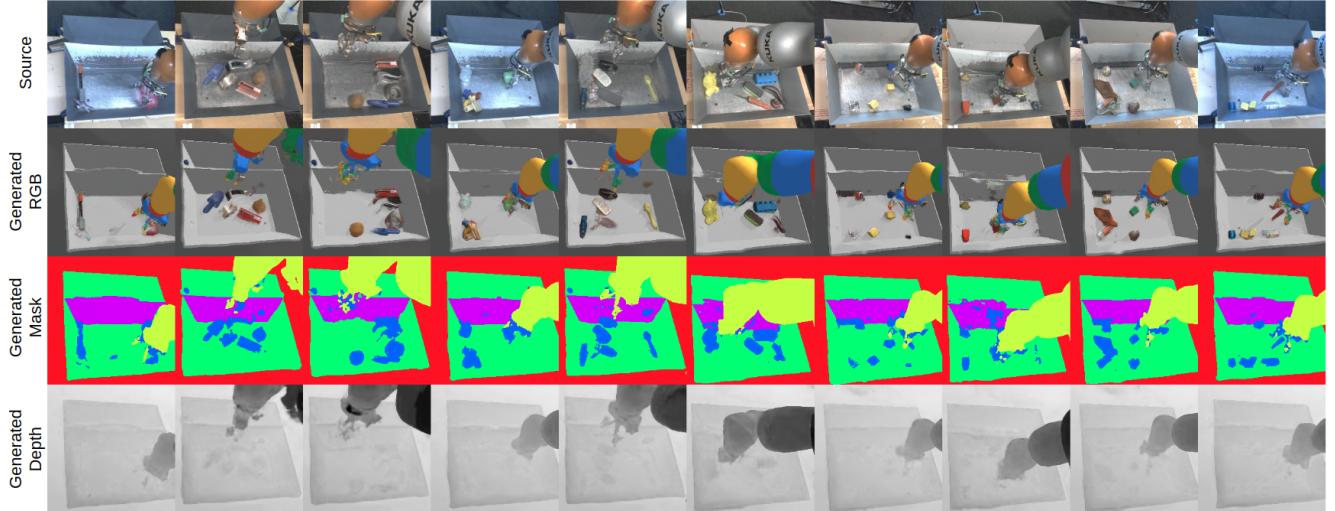


Figure 7: Network architecture of the generator function G . An RGB image from the source domain (either from the randomized domain or real-world domain) is processed via a U-Net style architecture [48] to produce a generated RGB image x_a , and auxiliaries that includes a segmentation mask m_a and depth image d_a . These auxiliaries forces the generator to extract semantic and depth information about the scene and encode them in the intermediate latent representation, which is then available during the generation of the output image.



(a) Randomized-to-canonical samples.



(b) Real-to-canonical samples.

Figure 8: Additional sample outputs of our trained generator G when given randomized sim images (8a) and real images (8b).