Digital Twin (DT)-CycleGAN: Enabling Zero-Shot Sim-to-Real Transfer of Visual Grasping Models

David Liu, Yuzhong Chen, and Zihao Wu

***Abstract*—Deep learning has revolutionized the field of robotics. To deal with the lack of annotated training samples for learning deep models in robotics, Sim-to-Real transfer has been invented and widely used. However, such deep models trained in simulation environment typically do not transfer very well to the real world due to the challenging problem of "reality gap". In response, this letter presents a conceptually new Digital Twin (DT)-CycleGAN framework by integrating the advantages of both DT methodology and the CycleGAN model so that the reality gap can be effectively bridged. Our core innovation is that real and virtual DT robots are forced to mimic each other in a way that the gaps or differences between simulated and realistic robotic behaviors are minimized. To effectively realize this innovation, visual grasping is employed as an exemplar robotic task, and the reality gap in zero-shot Sim-to-Real transfer of visual grasping models is defined as grasping action consistency losses and intrinsically penalized during the DT-CycleGAN training process in realistic simulation environments. Specifically, first, cycle consistency losses between real visual images and simulation images are defined and minimized to reduce the reality gaps in visual appearances during visual grasping tasks. Second, the grasping agent's action consistency losses are defined and penalized to minimize the inconsistency of the grasping agent's actions between the virtual states generated by the DT-CycleGAN generator and the real visual states. Extensive experiments demonstrated the effectiveness and efficiency of our novel DT-CycleGAN framework for zero-shot Sim-to-Real transfer.**

***Index Terms*—Grasping, continual learning, computer vision for automation.**

I. Introduction

DEEP learning has made major impacts on robotics [1]. To achieve desirable predictive performance, deep learning models often require large amounts of annotated data. However, in many robotic tasks such as visual grasping, obtaining annotated datasets to train end-to-end deep learning models in the real-world can be remarkably time-consuming and expensive [2], [3], [4], [5], sometimes even impossible. To solve this challenging problem, Sim-to-Real approaches [4], [6] have been proposed to train the robotic task performance models, such as visual grasping models, by simulations and then transfer these models to a real environment by domain randomization, domain adaptation, or other methods [5]. Typically, domain randomization mainly relies on randomizing various visual aspects, such as texture and color, to generate many random samples, such that the models are forced to focus on the fundamental aspects of image content which is invariant in real-world scenarios [7], [8], [9], [10]. In comparison, domain adaption aims to generalize the model trained in the source domain (i.e., simulations) to the target domain (i.e., real-world) by exploiting the unlabeled data in the target domain [4], [11]. Notably, generative adversarial networks (GANs) [12] were recently adopted for producing adapted images that are close to the realistic ones, thus significantly reducing the needed number of real-world samples [4], [13], [14]. However, the deflections and artifacts in the generated images by GANs may still cause the "reality gap" [5], negatively affecting the accuracy in both model training and real-world deploying.

To bridge the above-mentioned reality gap and enable zero-shot Sim-to-Real transfer, in this letter, we present a conceptually novel Digital Twin (DT)-CycleGAN framework by integrating the advantages of both DT approaches [15] and CycleGAN models [16]. Our core innovation is that real and virtual DT robots are forced to mimic each other such that the gaps between simulated and realistic robotic behaviors are minimized. That is, the similarity between the physical and virtual DT robots is maximized so that their behaviors tend to be identical, and thus the reality gap during zero-shot Sim-to-Real transfer is minimized at a fundamental level. Furthermore, to create much more realistic visual backgrounds for the DT robot's camera views, we randomize the background texture in the simulated environment during visual grasping and DT-CycleGAN model training, and as a result, the reality gap between virtual and physical environments is further significantly reduced. To effectively implement the above-mentioned innovation, the reality gaps between real/DT robots are defined as two types of consistency losses that are intrinsically minimized during the DT-CycleGAN model training in simulated environments. More specifically, the gaps between real visual images and simulated images seen by two cameras mounted on real/DT robots are defined as cycle consistency loss in the DT-CycleGAN model. By minimizing this loss, the reality gaps in visual appearances perceived by the robots' cameras during vision-based grasping tasks are significantly reduced. More importantly, in the DT-CycleGAN model,

Manuscript received 18 October 2022; accepted 17 February 2023. Date of publication 8 March 2023; date of current version 16 March 2023. This letter was recommended for publication by Associate Editor H. Wang and Editor H. Liu upon evaluation of the reviewers' comments. *(David Liu and Yuzhong Chen are co-first authors.) (Corresponding author: Zihao Wu.)*

David Liu is with Athens Academy, Athens, GA 30606 USA (e-mail: david.weizhong.liu@gmail.com).

Yuzhong Chen is with the University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: chenyuzhong211@gmail.com).

Zihao Wu is with the University of Georgia, Athens, GA 30602 USA (e-mail: zw63397@uga.edu).

https://github.com/YuzhongChen-98/DigitalTwin-CycleGAN

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2023.3254460, provided by the authors.

Digital Object Identifier 10.1109/LRA.2023.3254460

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

the inconsistency of the robotic grasping agent's actions between the virtual states and the real states are defined and minimized as the grasping agent's action consistency loss, which is our core algorithmic innovation. Through the joint minimization of both cycle consistency loss and grasping agent's action consistency loss, the reality gaps in both visual appearance and grasping action states for DT/real robots are minimized, thus contributing to a much more effective and robust zero-shot Sim-to-Real transfer of the visual grasping models trained by simulations to real environments.

II. Related Works

A. General Grasping and Visual Grasping

Robotic grasping is one of the most fundamental problems in the field of robotics and has been actively studied for decades [17], [18]. To accomplish a grasping task, vision and tactile sensors have played instrumental roles in locating, positioning and picking up the target [18]. Recently, vision-based robotic grasping has attracted increasing attention as a central research theme due to the fast advancements of deep learning and computer vision, and has achieved remarkable progress [5], [19]. In general, visual grasping approaches can be categorized into two main streams: object reconstruction based methods and end-to-end methods [19], [20]. In general, 3D reconstruction of free-form objects can enable accurate grasp planning, while end-to-end methods can generate grasp proposals directly from the camera sensor [20]. By sampling the grasp candidates or generating suitable grasps, end-to-end methods demonstrated promising results in generalizing to even unseen and novel objects/backgrounds, and have gained increasing popularity [5]. However, training those end-to-end models, especially deep learning ones, requires large-scale annotated visual grasping datasets, which could be very time-consuming and expensive to be instrumented and collected in robotics [2], [3], [4], [5]. End-to-end visual grasping methods can be further categorized into open-loop or closed-loop grasp executions [2]. In this work, two global and local cameras are mounted on the robot's base and gripper arm for robot-centric "action-view" sensing of the grasping target identification and closed-loop grasping action state estimation.

B. Visual Grasping in Complex Background

In prior works [5], [17], [18], the visual grasping task was typically conducted in relatively plain backgrounds. However, image backgrounds could significantly affect the accuracy of object detection and recognition in real-world grasping. For instance, the authors in [21] discovered that object recognition models are capable of making accurate classifications based on only the image backgrounds. The studies in [21] emphasized the importance of visual grasping models' capability of being able to identify the correct target object using foreground information despite complex backgrounds. In this direction, the authors in [14] used object-detection consistency in visual grasping models. Inspired by the studies in [14], we randomized the background texture in the simulated environment for grasping agent model initialization to create more complex and realistic backgrounds from the robotic cameras' views. It is demonstrated that our framework enables effective zero-shot transfer of trained visual grasping models in virtual environment to unseen physical environment with complex background.

C. Digital Twin for Robotic Visual Grasping

Digital twin (DT) has been regarded as a transformative paradigm [15]. Recently, DT technologies have been explored in robotics [15], [22]. For instance, the authors in [22] developed a DT-enabled approach for achieving the effective transfer of deep reinforcement learning (DRL) algorithms to a physical robot. In this letter, our robotics setup and visual grasping scenario are quite different and more challenging than those in [22]. Specifically, we employed two cameras mounted on the physical robot's base and grasping arm, respectively. Both cameras are robot-centric with "action-views", instead of being static as that in [22], and they acquire both global and local scene information for grasping target identification and closed-loop grasping action estimation. The two cameras on the virtual DT robot function in a similar way, and they acquire large-scale visual grasping datasets with ground truth annotations in virtual environment to train the DT-CycleGAN model. Therefore, our designed cycle consistency losses and novel grasping agent's action consistency losses in the DT-CycleGAN model can simultaneously penalize visual appearance reality gap and grasping action reality gap in complex backgrounds in realistic virtual world that mirrors the physical world, making zero-shot Sim-to-Real transfer possible and robust.

D. GAN Models for Bridging Reality Gap in Visual Grasping

Sim-to-Real transfer has been widely adopted in robotics [5], and a variety of approaches have been developed to deal with the reality gap problem [5], [23], [24], [25]. Recently, in the milestone work of RetinaGAN [14], a novel GAN model was proposed to effectively adapt simulated images to realistic ones by enforcing an object-detection consistency for visual grasping. In addition to RetinaGAN [14], another interesting literature work RL-CycleGAN model [26] applied the similarity of total expected future reward (Q-value) of Q-learning as an additional reinforcement learning (RL)-scene consistency loss, which facilitates effective Sim-to-Real transfer for reinforcement learning. These GAN-based enhancements allowed the visual grasping model to take the full advantage of virtual data and to achieve better results with less training on real data. In our DT-CycleGAN model, the novel grasping agent's action consistency losses are specifically designed to minimize the inconsistency of the agent's actions between the visual states generated by the DT-CycleGAN generator and the original visual states within a DT robot's context. Also, to better reduce the reality gap induced by the virtual grasping agent model, we enforce the agent to continuously learn during the training of the DT-CycleGAN model. In this way, the reality gap during zero-shot Sim-to-Real transfer of visual grasping models is significantly reduced, which is one of our core innovations and contributions compared to other prior methods [14], [26].

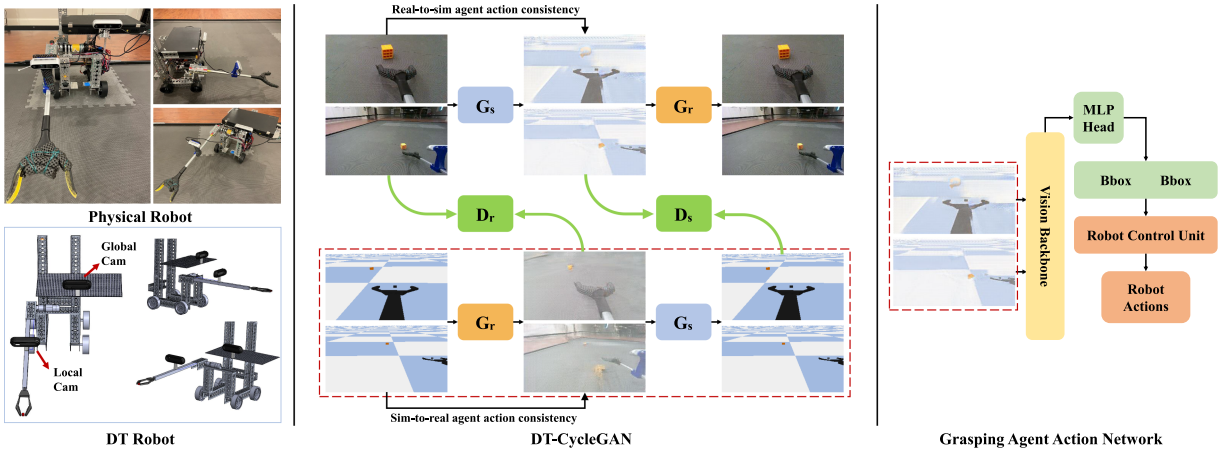

Fig. 1. Illustration of the physical/DT robots, DT-CycleGAN model and grasping agent action network. The annotations are referred to the main text.

III. Methods

Our framework consists of three components: the physical/DT robots, grasping agent action network, and the DT-CycleGAN model, as illustrated in Fig. 1.

A. Physical/DT Robots

Similar materials and tools in our prior work [27] were used to build our physical robot. The main differences and improvements over our prior robot [27] in this work are that two RealSense RGB-D cameras are installed on the robot base and gripper arm, respectively, for both global and local views of the scene, and a laptop is used to process and analyze real-time videos/images to inform and guide the robotic arm and gripper to perform the data-driven visual grasping, as illustrated in Fig. 1. The virtual DT robot was constructed using the Solidworks tool. When building the DT robot, we made a major effort to maximize the similarity between the physical and virtual DT robots such that their behaviors tend to be identical, thus bridging the reality gap during zero-shot Sim-to-Real transfer at a fundamental level.

Based on the 3D virtual DT robot, visual grasping simulations are run within the PyBullet environment. In PyBullet simulations, an analytic method [19], [28] is employed for target grasping due to its ease with known 3D geometries and locations of both the DT robot and the grasping target. Importantly, during the analytic grasping process, virtual cameras mounted on the DT robot simultaneously acquire realistic videos via effective CycleGAN-style [16] image synthesis techniques in the entire grasping task. These robot-centric views of videos during grasping tasks are automatically annotated with ground truth actions and are then used to train the DT-CycleGAN model in virtual environment.

B. Grasping Agent Network

The grasping *Agent* network accepts the images of two cameras' views on the robot as input and generates the position of the target to be grasped. As shown in Fig. 1, the visual input consists of two images: a global image with a wide field of view on the robot base and a local image with a focused field of view on the gripper arm. Therefore, the input image state can be described as $\mathbf{x} \in \mathbb{R}^{B2CHW}$ where $B2$ represent the batch size and 2 views of input images, and CHW represent the channel (C), height (H) and width (W) of images. For a robust zero-shot Sim-to-Real transfer, the grasping agent network outputs the location information of the target to be grasped on the input image, represented by the Bbox in Fig. 1. Therefore, the output of the gripper *Agent* model is $a = [x^l_{min}, y^l_{min}, x^l_{max}, y^l_{max}, x^g_{min}, y^g_{min}, x^g_{max}, y^g_{max}] \in \mathbb{R}^8$, where the first four parameters record the target Bbox on the local (annotated by the superscript l) image, while the last four parameters are on the global (annotated by the superscript g) image. Based on the predicted Bbox, the actual action of the robot is generated via the closed-loop grasping estimation. More specifically, the grasping *Agent* network consists of three main parts, as illustrated in Fig. 1. The first part is the visual backbone network (e.g., Swin-Ti [29], ViT-Ti [30] and ResNet-34d [31] explored in this work) which extracts visual features from multi-view input images, and the second one is the multilayer perceptron (MLP) head which predicts the position of the grasp target (represented by the Bbox) from extracted visual features. To facilitate easier transfer from the pre-trained visual network model, the input images are resized from 1080×1920 to 224×224. The third part is the robot control unit that generates the actual actions of the robot, including both the grasping action and the robot's motion, based on the predicted Bbox of the grasping target and the current input image. More specifically, the robot control unit aims to minimize the grasping error, which can be mathematically formulated and minimized as:

$$Error = \alpha * (\mathbf{y}^l_{pred} - \mathbf{y}^l_{gt}) + (1 - \alpha) * (\mathbf{y}^g_{pred} - \mathbf{y}^g_{gt}) \quad (1)$$

where $\mathbf{y}^*_{pred}$ is the center of the predicted Bbox and $\mathbf{y}^*_{gt}$ is the ground truth of the center of the Bbox where the robot can grasp the target. Here, the local (l) and global (g) superscripts represent images from both cameras on the robot. When the Bbox of the predicted target overlaps with the graspable position, i.e., the

error tends to be 0, the robot performs the closed-loop grasping action.

To enable zero-shot Sim-to-Real transfer of the grasping $Agent$ network trained in the virtual environment, the above-mentioned grasping $Agent$ network itself, the input images acquired through the DT robot's cameras, and the predicted Bboxes on image states (representing the robot's corresponding actions) are holistically modeled by the DT-CycleGAN.

C. DT-CycleGAN: Cycle Consistency Losses

CycleGAN [16] was originally proposed for style transfer for unpaired images, and it is adopted here and integrated into our DT-CycleGAN model to bridge the visual appearance gaps in simulations and real scene images, as illustrated by the Real-to-Sim and Sim-to-Real arrows in Fig. 1. Suppose $\mathbf{x}_r \in \mathbf{R}$ represents real visual images input and $\mathbf{x}_s \in \mathbf{S}$ represents simulation visual images, as illustrated in Fig. 1. The simulated visual images are collected from the virtual environment via analytic grasping models [19], [28] (called expert model here) and the virtual cameras on the DT robot with ground truth actions, while the real images are collected from the real environment in random actions without performing the grasping task, as illustrated by the DT-CycleGAN panel in the middle of Fig. 1. The analytic expert model is generated in the virtual environment by extracting the actual known position of the target, which provides the grasp object's Bbox. The generator $G_r(\cdot)$ will transform images to real style, while the generator $G_s(\cdot)$ will transform input images to simulation style, which are similar to the generators $F(*)$ and $G(*)$ in the original CycleGAN model [16]. The real discriminator $D_r(\cdot)$ will determine if the input images are real style and the simulation discriminator $D_s(\cdot)$ will determine if the input images are simulation style. The cycle loss function of DT-CycleGAN can be written as:

$$\begin{aligned} &L(G_s, G_r, D_s, D_r) \\ &= L_{GAN}(G_s, D_s, \mathbf{R}, \mathbf{S}) + L_{GAN}(G_r, D_r, \mathbf{S}, \mathbf{R}) \\ &\quad + L_{cyc}(G_r, G_s) + L_{identity}(G_r, G_s) \end{aligned} \tag{2}$$

where $L_{GAN}(G_s, D_s, \mathbf{R}, \mathbf{S})$, $L_{GAN}(G_r, D_r, \mathbf{S}, \mathbf{R})$ are the adversarial losses, $L_{cyc}(G_r, G_s)$ is the cycle consistency loss and $L_{identity}(G_r, G_s)$ is the identity mapping loss.

The adversarial losses are for matching the generated image's style to the target domain style (e.g., real-world environment), which are:

$$\begin{aligned} L_{GAN}(G_s, D_s, \mathbf{R}, \mathbf{S}) &= \mathbb{E}_{\mathbf{x}_s \sim p_{data}(S)}[logD_s(\mathbf{x}_s)] \\ &\quad + \mathbb{E}_{\mathbf{x}_r \sim p_{data}(R)}[log(1 - D_s(G_s(\mathbf{x}_r))] \end{aligned} \tag{3}$$

$$\begin{aligned} L_{GAN}(G_r, D_r, \mathbf{S}, \mathbf{R}) &= \mathbb{E}_{\mathbf{x}_r \sim p_{data}(R)}[logD_r(\mathbf{x}_r)] \\ &\quad + \mathbb{E}_{\mathbf{x}_s \sim p_{data}(S)}[log(1 - D_r(G_r(\mathbf{x}_s))] \end{aligned} \tag{4}$$

The cycle consistency loss is applied to avoid mapping input image to any random permutation of images in the target domain style. The images should be able to transform back to the original space (e.g., virtual environment). Therefore, it can be written as:

$$\begin{aligned} L_{cyc}(G_r, G_s) &= \mathbb{E}_{\mathbf{x}_s \sim p_{data}(S)}[\|G_s(G_r(\mathbf{x}_s)) - \mathbf{x}_s\|_1] \\ &\quad + \mathbb{E}_{\mathbf{x}_r \sim p_{data}(R)}[\|G_r(G_s(\mathbf{x}_r)) - \mathbf{x}_r\|_1] \end{aligned} \tag{5}$$

The identity mapping loss is used to keep the image content consistent, which can be described as:

$$\begin{aligned} L_{identity}(G_r, G_s) &= \mathbb{E}_{\mathbf{x}_s \sim p_{data}(S)}[\|G_s(\mathbf{x}_s) - \mathbf{x}_s\|_1] \\ &\quad + \mathbb{E}_{\mathbf{x}_r \sim p_{data}(R)}[\|G_r(\mathbf{x}_r) - \mathbf{x}_r\|_1] \end{aligned} \tag{6}$$

D. DT-CycleGAN: Agent Action Consistency Losses

In our DT-CycleGAN model, the novel grasping Agent's action consistency losses are designed to penalize the inconsistency of the agent actions (represented by the Bboxes on image states that determine the robot's actual actions) between the raw input image states and the style-transformed image states by the generator. Suppose an image state $\mathbf{x} \in \mathbf{S}$, the robot's action $\mathbf{a}_1 = Agent(\mathbf{x})$ should be the same as $\mathbf{a}_2 = Agent(G_r(\mathbf{x}))$. Therefore, the agent consistency losses (equivalent to the discrepancies of the predicted Bboxes in Fig. 1) can be written as:

$$\begin{aligned} &L_{agent}(G_r, G_s, Agent) \\ &= \mathbb{E}_{\mathbf{x}_s \sim p_{data}(S)}[\|Agent(G_r(\mathbf{x}_s)) - Agent(\mathbf{x}_s)\|_1] \\ &\quad + \mathbb{E}_{\mathbf{x}_r \sim p_{data}(R)}[\|Agent(G_s(\mathbf{x}_r)) - Agent(\mathbf{x}_r)\|_1] \end{aligned} \tag{7}$$

Therefore, the entire loss function of DT-CycleGAN is:

$$\begin{aligned} &L(G_s, G_r, D_s, D_r, Agent) \\ &= \lambda_{gan} * L_{GAN}(G_s, D_s, \mathbf{R}, \mathbf{S}) \\ &\quad + \lambda_{gan} * L_{GAN}(G_r, D_r, \mathbf{S}, \mathbf{R}) \\ &\quad + \lambda_{cyc} * L_{cyc}(G_r, G_s) + \lambda_{identity} * L_{identity}(G_r, G_s) \\ &\quad + \lambda_{agent} * L_{agent}(G_r, G_s, Agent) \end{aligned} \tag{8}$$

where λ_{gan}, λ_{cyc}, $\lambda_{identity}$ and λ_{agent} are the hyper-parameters. Here, we set λ_{gan} to 1, λ_{cyc} to 5, the $\lambda_{identity}$ was set to 2 and λ_{agent} was 10.

E. Learning Strategies

To effectively train the entire robotic visual grasping model in Fig. 1, we divided the training into two steps. Notably, both the grasping Agent network and the DT-CycleGAN model were purely trained in simulation environment for the purpose of zero-shot Sim-to-Real transfer.

Learning Grasping Agent in Simulation Environment: The first step is to learn the grasping $Agent$ model's parameters (including those in the vision backbone and the MLP head) in the simulation environment. The visual backbones first used the pre-trained weights on ImageNet-1 k as initialization. The grasping target was randomly placed in the scene within the scopes of both cameras, and the images with automatically labeled Bboxes (with ground truth) were collected for the fine-tuning of the vision backbones and the supervised training of the MLP head.

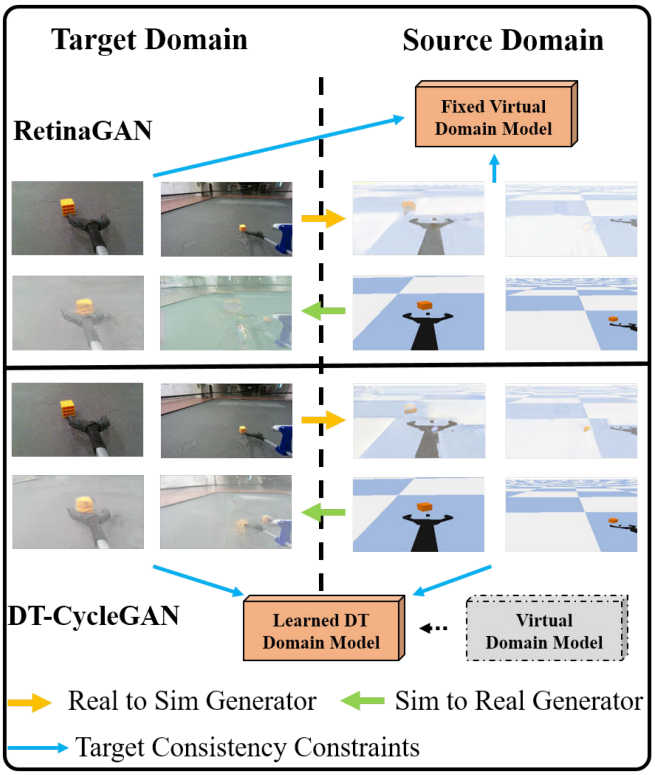

Fig. 2. Illustration of different methods. Compared with RetinaGAN, the target consistency constrained model is trained together with the CycleGAN component in our DT-CycleGAN.

Learning DT-CycleGAN in Simulation Environment: When training DT-CycleGAN by simulations, it can significantly reduce the reality gap in zero-shot Sim-to-Real transfer by using real images of random actions (not the human expert's grasping demonstration data). About 2 k images with random positions of grasping target (without either expert's grasping demonstration or Bbox labels) were collected for the training. The training setting was consistent with [16]. The training algorithm for DT-CycleGAN is shown in Algorithm 1. Compared with RetinaGAN, we directly used the grasping Agent's action consistency rather than the object detection consistency loss. Besides, as shown in Fig. 2, we allowed the grasping Agent model and the DT-CycleGAN model to be trained in parallel in Algorithm 1 (rows: 12-13), rather than being only considered as a data augmentation method to train the grasping Agent. Later experiments will demonstrate that our new approach significantly outperforms the RetinaGAN model.

IV. Evaluation

A. Experiments in Simulation Environment

By using the methods described in Section III-A, we collected 8 k paired images of the initial states generated by random positions of the grasping target and randomly colored textures in the virtual environment as the training dataset, which was conveniently facilitated by the DT robot and the PyBullet environment. The image textures were obtained from the Describable Textures Dataset (DTD). All the visual backbone models were based on the pre-trained weights on ImageNet-1 k as initialization. We set the learning rate to 1e-4 and weight decay to 1e-4 with the OneCycleLR learning rate scheduler for 1500 steps of training. The batch size was set to 32 and SmoothL1 was applied as the loss function. We compared the results of different visual backbones (Swin-Ti [29], ViT-Ti [30] and ResNet-34d [31]), and reported the success rates and the errors of predicted Bboxes compared to the expert model (via analytic methods) in the virtual environment.

The experiment results were shown in Table I for 30 trials of testing, and the Swin-Ti backbone achieved the highest accuracy of 90%, which is quite promising. To further evaluate the impact of using different numbers of training samples, we provided the experimental results of using the corresponding obtained models in the table.

B. Experiments in Real Environment With Zero-Shot Sim-to-Real Transfer

Real-world visual grasping experiments were performed by our physical robot equipped with the DT-CycleGAN model and

Algorithm 1: Training Algorithm for DT-CycleGAN.

Require: $Agent$ trained in the simulation environment, G_r, G_s, D_r, D_s with initialization parameters. $\mathbf{R}$, $\mathbf{S}$ from real and simulation state. $Expert$ model in simulation.

1: Initialize empty buffer $\mathbf{B}_s, \mathbf{B}_r$ to store simulation and real environment training data
2: **while** Training **do**
3: Sample batch real and simulation state $\mathbf{x}_r \sim \mathbf{R}, \mathbf{x}_s \sim \mathbf{S}$;
4: Compute $\mathbf{B}_s = \mathbf{B}_s \cup G_s(\mathbf{x}_r), \mathbf{B}_r = \mathbf{B}_r \cup G_r(\mathbf{x}_s)$
5: Compute $Loss(G_s, G_r, D_s, D_r, Agent)$
6: Update parameter of G_s, G_r
7: Sample batch from Buffer $\mathbf{x}_s = \mathbf{x}_s \cup G_s(\mathbf{x}_r) \sim \mathbf{B}_s$, $\mathbf{x}_r = \mathbf{x}_r \cup G_r(\mathbf{x}_s) \sim \mathbf{B}_r$
8: Compute $Loss(D_s, \mathbf{x}_s, G_s(\mathbf{x}_r)) = \mathbb{E}_{\mathbf{x}_s}[log (D_s(\mathbf{x}_s))] + \mathbb{E}_{G_s(\mathbf{x}_r)}[log(1 - D_s(G_s(\mathbf{x}_r)))]$
9: Update parameter of D_s
10: Compute $Loss(D_r, \mathbf{x}_r, G_r(\mathbf{x}_s)) = \mathbb{E}_{\mathbf{x}_r}[log (D_r(\mathbf{x}_r))] + \mathbb{E}_{G_r(\mathbf{x}_s)}[log(1 - D_r(G_r(\mathbf{x}_s)))]$
11: Update parameter of D_r
12: Compute $Loss(Agent, Expert, \mathbf{x}_s, G_r(\mathbf{x}_s)) = \mathbb{E}_{\mathbf{x} \sim x_s \cup G_r(\mathbf{x}_s)}[\|Agent(\mathbf{x}) - Expert(\mathbf{x})\|_1]$
13: Update parameter of $Agent$
14: **end while**

TABLE I
The Predictive Error and Success Rate for Grasping of 30 Trials With Different Visual Backbones in the Simulation Environment

Visual Backbone	Test error	Success rate
Swin-Ti	**0.0016**±0.0024	**90.0**%
ViT-Ti	0.0035±0.0038	63.3%
ResNet-34d	0.0022±0.0018	66.7%

Here, 0.001 equals 0.224 pixels in an image.

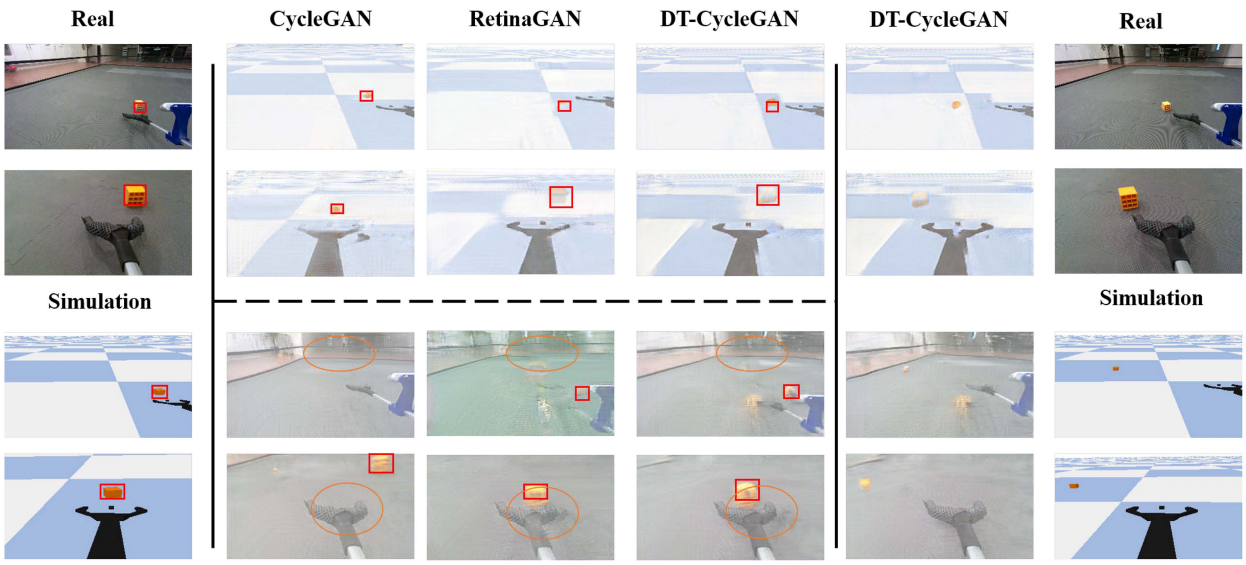

Fig. 3. Images in the visual grasping task generated by either the Sim-to-Real (bottom) or Real-to-Sim (top) generator. The grasping target and the background are highlighted by the red box and the yellow oval.

TABLE II
THE SUCCESS RATES OF GRASPING IN 20 TRIALS FOR DIFFERENT METHODS WITH TWO TYPES OF BACKGROUNDS

Methods	Success rates	
	Plain background	Complex background
Sim-Only	0	0
Randomized Sim	60%	20%
CycleGAN [16]	0	0
RetinaGAN [14]	55%	25%
DT-CycleGAN	**85%**	**65%**

grasping Agent action network trained in virtual environment. Here, we selected the Swin-Ti as the visual backbone for the grasping $Agent$ model as it performed significantly better than other models in simulated learning experiments. We compared different methods for zero-shot Sim-to-Real transfer, including Sim-Only, Randomized Sim, CycleGAN and RetinaGAN, with our DT-CycleGAN. That is, all of these compared visual grasping models were purely trained in a simulation environment, and there is no fine-tuning step performed (truly zero-shot Sim-to-Real transfer) when transferring them to the real environment. Other methods' settings are as follows:

- The Sim-Only was the model that was only trained with randomly initialized positions.
- The Randomized Sim was the same model and setting of Swin-Ti as in Table I.
- The CycleGAN was trained on the simulation images and real images, and then the simulation images and real-style images generated by CycleGAN were mixed together to train the agent action network.
- Similar to CycleGAN, the RetinaGAN was trained as a data augmentation method with an object detection model to preserve the perception consistency.

The quantitative comparison results were provided in Table II for 20 trials of testing, and our DT-CycleGAN model clearly achieved the highest grasping accuracy, which was much better than other methods in both plain and complex backgrounds. It is noted that the Sim-Only and CycleGAN methods completely failed in the zero-shot Sim-to-Real transfer to a real-world environment. To interpret this result, a visual comparison among CycleGAN, RetinaGAN and our DT-CycleGAN is shown in Fig. 3. While it is clear that CycleGAN has a good ability to transfer the background information of the image, the grasping target information is severely lost. RetinaGAN retains the target information relatively well after increasing the consistency loss of target detection, however, its transformation of background is flawed, which may limit effective zero-shot Sim-to-Real transfer. One possible reason is that RetinaGAN fixed the detection model parameters, which also fixed the model in the source domain (e.g., simulation environment). When moving to the target domain (e.g., real environment), the fixed detection model is not adaptive to the novel domain, thus resulting in an incomplete transfer of the background style. In comparison, our DT-CycleGAN demonstrated much better results by allowing for synchronous training of the Agent action network model and the DT-CycleGAN model, suggesting the superiority of our framework to bridge reality gaps in zero-shot Sim-to-Real transfer.

Notably, in the real-world experiments, the complex backgrounds (examples shown in Fig. 4) were never seen before by the physical robot or the DT robot in simulation environment. Our physical robot empowered by DT-CycleGAN still achieved a promising success rate of 65%, demonstrating the possibility and feasibility of zero-shot Sim-to-Real transfer of visual grasping models in completely unseen complex backgrounds. In particular, the success rate in real-world experiments by our DT-CycleGAN model is much higher than other compared methods, suggesting the effectiveness of our strategy of using the digital twin methodology and the DT-CycleGAN model to reduce the reality gap during zero-shot Sim-to-Real transfer.

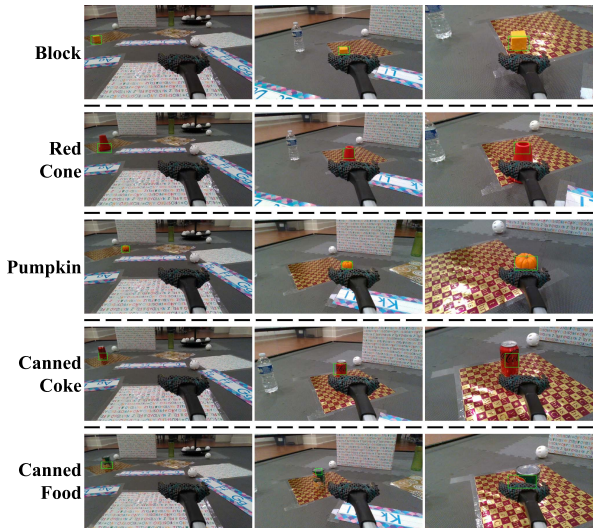

Fig. 4. Illustrations of visual grasping process in complex backgrounds with patterns and other objects scattered around. Snapshots of three steps (each row) from local camera's view are shown here.

TABLE III
THE SUCCESS RATES OF 20 TRIALS WITH DIFFERENT HYPER-PARAMETERS OF λ_{agent} IN REARRANGED COMPLEX BACKGROUND

Method	Action Consistency	Agent Network	SR
CycleGAN	None	None	0%
RetinaGAN	10	Frozen	30%
DT-CycleGAN	1	Trainable	70%
	2		60%
	5		75%
	10		80%
	20		85%

SR = SUCCESS RATE.

C. Ablation Studies

Effect of Agent Action Consistency Loss: Compared with CycleGAN and RetinaGAN, we designed a new grasping agent network and the novel agent action consistency losses in DT-CycleGAN. To explore the effect of agent action consistency losses on DT-CycleGAN, we experimented with different parameter settings of λ_{agent} in (8). The training setting is consistent with Section IV-B. We evaluated the performance of models on complex backgrounds with 20 trials and recorded the success rate of grasping for each model. The result in Table III indicated that the DT-CycleGAN can achieve promising results in a wide range of λ_{agent}. Also, the result further proved the phenomenon in Fig. 3.

Effect of Identity Mapping Loss: The identity mapping loss was explored in CycleGAN [16], which was used to keep the image content consistent when the same-style images were fed into the generator. To further explore the influence of the identity mapping loss on our DT-CycleGAN, we experimented with/without different hyper-parameters on identity mapping loss and reported the results in Table IV. It is clear that the identify mapping loss is very important for the DT-CycleGAN model to accurately grasp.

TABLE IV
THE SUCCESS RATE OF 20 TRIALS WITH DIFFERENT HYPER-PARAMETERS OF $\lambda_{identity}$ IN A REARRANGED COMPLEX BACKGROUND

Identity Mapping Loss	Success rate
0.0	60%
2.0	80%
5.0	90%

TABLE V
THE SUCCESS RATES OF GRASPING IN 20 TRIALS IN COMPLEX BACKGROUND FOR DIFFERENT SAMPLE SIZES

Methods	Simulation		Real Environment	
	Size	SR	Size	SR
Randomized Sim	8k	90%	0	20%
Randomized Sim	12k	100%	0	90%
DT-CycleGAN	0k	None	2k	0%
DT-CycleGAN	8k	100%	2k	65%
DT-CycleGAN	12k	100%	2k	100%

SR = SUCCESS RATE.

TABLE VI
THE SUCCESS RATES OF GRASPING IN 40 TRIALS IN COMPLEX BACKGROUND WITH DIFFERENT TARGET OBJECTS

Methods	Training Object		Test Object	SR
	Sim	Real		
DT-CycleGAN	Block	Block	Block	97.5%
DT-CycleGAN	Block	Red Cone	Red Cone	80%
DT-CycleGAN	Block	Block	Pumpkin	82.5%
DT-CycleGAN	Block	Block	Canned Coke	75%
DT-CycleGAN	Block	Block	Canned Food	77.5%

SR = SUCCESS RATE.

D. Experiments on Different Data Sizes and Unseen Objects

Evaluation on Different Training Data Sizes: The digital twin methodology in our proposed DT-CycleGAN model can significantly reduce the human effort in collecting training data in real-world environment and thus allows us to adopt large amounts of simulation data to facilitate zero-shot Sim-to-Real transfer. To quantitatively evaluate the impact of the simulation training data size, we trained the DT-CycleGAN model with different training samples, and reported the comparison results in Table V. In the first two rows, without utilizing any real images, the grasping success rate in real environment improved significantly from 20% to 90% by solely increasing the simulation data size from 8 k to 12 k, which is easily achievable in a simulation environment. Moreover, by incorporating 2 k real images in addition to the simulation images, the grasping success rate further increased from 90% (second row, using 12 k simulation data only) to 100% (last row, using 12 k simulation data and 2 k real data).

Generalization to Unseen Objects: To evaluate the generalizability of our DT-CycleGAN model to unseen objects in real world, we reported the results of using different training and testing grasping objects in Table VI, which shows that our model can achieve a promising success rate in generalizing to unseen objects with different colors and shapes in complex background. Examples of the visual grasping process for different objects

were shown in Fig. 4. Overall, these experiments demonstrated the effectiveness and robustness of our DT-CycleGAN model in zero-shot Sim-to-Real transfer of visual grasping models.

V. Conclusions and Discussions

This letter contributes a conceptually novel framework of DT-CycleGAN to enable zero-shot Sim-to-Real transfer of visual grasping models. Our core methodological innovations and contributions are the definitions and implementations of the cycle consistency losses and the grasping agent's action consistency losses in DT-CycleGAN within a digital twin context. In the future, the DT-CycleGAN can be extended in various directions. For instance, in addition to the two-finger gripper used in this letter, a variety of other gripper configurations/factors, such as gripper types, soft grippers, gripper freedoms, gripper arm freedoms, and grasp target numbers, should be explored and evaluated. Also, zero-shot Sim-to-Real transfer of DT-CycleGAN models in 3D unstructured environments, such as robotic apple harvesting, will be investigated in the near future. Finally, in addition to Swin-Ti [29], ViT-Ti [30] and ResNet-34d [31] used in this work, more vision backbone models and their fine-tuning schemes will be examined to further potentially improve the performance of the DT-CycleGAN framework.

Acknowledgment

David Liu would like to thank Professor Charlie Li, Dr. Javier Rodriguez, Daniel Petti and Mike Callinan for their guidance. The authors would like to thank Haixing Dai, Enze Shi, Huawen Hu, Steven Xu, Yiheng Liu, Lin Zhao, and Tianming Liu for their help on various aspects of this project including laptop-robot communication, construction of digital twin robot, setup of simulation environment, preparation of background images, and editorial assistance.

References

[1] A. I. Károly, P. Galambos, J. Kuti, and I. J. Rudas, "Deep learning in robotics: Survey on model structures and training strategies," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 51, no. 1, pp. 266–279, Jan. 2021.

[2] S. Song, A. Zeng, J. Lee, and T. Funkhouser, "Grasping in the wild: Learning 6DoF closed-loop grasping from low-cost demonstrations," *IEEE Robot. Automat. Lett.*, vol. 5, no. 3, pp. 4978–4985, 2020.

[3] D. Kalashnikov et al., "Qt-Opt: Scalable deep reinforcement learning for vision-based robotic manipulation," 2018, *arXiv:1806.10293*.

[4] K. Bousmalis et al., "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 4243–4250.

[5] K. Kleeberger, R. Bormann, W. Kraus, and M. F. Huber, "A survey on learning-based robotic grasping," *Curr. Robot. Rep.*, vol. 1, no. 4, pp. 239–249, 2020.

[6] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," 2018, *arXiv:1809.10790*.

[7] 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 *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 23–30.

[8] S. James, A. J. Davison, and E. Johns, "Transferring end-to-end visuomotor control from simulation to real world for a multi-stage task," in *Proc. Conf. Robot Learn.*, 2017, pp. 334–343.

[9] J. Tobin et al., "Domain randomization and generative models for robotic grasping," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 3482–3489.

[10] Y. Chebotar et al., "Closing the sim-to-real loop: Adapting simulation randomization with real world experience," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 8973–8979.

[11] S. James et al., "Sim-To-Real via Sim-To-Sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12627–12637.

[12] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, pp. 139–144, 2020.

[13] Y. Tang, J. Chen, Z. Yang, Z. Lin, Q. Li, and W. Liu, "DepthGrasp: Depth completion of transparent objects using self-attentive adversarial network with spectral residual for grasping," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 5710–5716.

[14] D. Ho, K. Rao, Z. Xu, E. Jang, M. Khansari, and Y. Bai, "RetinaGAN: An object-aware approach to sim-to-real transfer," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 10920–10926.

[15] M. Liu, S. Fang, H. Dong, and C. Xu, "Review of digital twin about concepts, technologies, and industrial applications," *J. Manuf. Syst.*, vol. 58, pp. 346–361, 2021.

[16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[17] A. Bicchi and V. Kumar, "Robotic grasping and contact: A review," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2000, vol. 1, pp. 348–353.

[18] Q. Marwan, S. C. Chua, and L. C. Kwek, "Comprehensive review on reaching and grasping of objects in robotics," *Robotica*, vol. 39, pp. 1849–1882, 2021.

[19] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *Int. J. Robot. Res.*, vol. 27, no. 2, pp. 157–173, 2008.

[20] N. Chavan-Dafle, S. Popovych, S. Agrawal, D. D. Lee, and V. Isler, "Simultaneous object reconstruction and grasp prediction using a cameracentric object shell representation," 2021, in *IROS. IEEE*, 2022, pp. 1396–1403.

[21] K. Xiao, L. Engstrom, A. Ilyas, and A. Madry, "Noise or signal: The role of image backgrounds in object recognition," 2020, *arXiv:2006.09994*.

[22] Y. Liu et al., "A digital twin-based sim-to-real transfer for deep reinforcement learning-enabled industrial robot grasping," *Robot. Comput.- Integr. Manuf.*, vol. 78, 2022, Art. no. 102365.

[23] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3722–3731.

[24] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 3803–3810.

[25] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, May 2015.

[26] K. Rao, C. Harris, A. Irpan, S. Levine, J. Ibarz, and M. Khansari, "RL-CycleGAN: Reinforcement learning aware simulation-to-real," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11157–11166.

[27] D. Liu and M. Callinan, "Few-shot object detection for robotics," *Int. J. High Sch. Res.*, vol. 5, no. 2, 2023.

[28] A. Sahbani, S. El-Khoury, and P. Bidaud, "An overview of 3D object grasp synthesis algorithms," *Robot. Auton. Syst.*, vol. 60, no. 3, pp. 326–336, 2012.

[29] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10012–10022.

[30] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.