Torwards Unpaired Human-to-Robot Demonstration Translation Learning Novel Tasks

Ruisen Liu
Georgia Institute of Technology
Atlanta, USA
ruisenericliu@gatech.edu

Matthew C. Gombolay
Georgia Institute of Technology
Atlanta, USA
matthew.gombolay@cc.gatech.edu

Stephen Balakirsky
Georgia Institute of Technology
Atlanta, USA
stephen.balakirsky@gtri.gatech.edu

***Abstract*—Advancements in autonomy can enhance space flight and exploration by enabling robots as cost-efficient agents when humans are unavailable [1]. However, long-term mission success may require continuous maintenance and the ability to adapt on the fly. When encountering a novel scenario that is outside expected robot capabilities, it becomes valuable for a non-robotics expert to be able to visually demonstrate the intended task execution to the robot. Relying on visual demonstration introduces ambiguity in mapping from human to robot execution. One mapping approach is to learn unpaired image translations from human demonstrations and unrelated robot motions. In this paper, we target extensions to image translation to enable robust conveyance of desired task execution. We propose methods to ground generated images with truth in kinematic feasibility, without imposing additional data collection or computational requirements on the demonstrator.**

***Index Terms*—learning from demonstration, adversarial networks, unpaired human-robot task translation**

I. Introduction

Remote exploration in space requires coordination and efficient deployment between human and robots. In scenarios where continuous human presence may both be unsustainable and costly, semi-autonomous robots can provide a viable alternative for mission execution [1]. However, independent robot deployment requires robustness to uncertainty, and the capability to adapt to novel scenarios. In particular, sustain operational habitats requires systematic maintenance, and thus the ability for a supervising robot to carry out novel repair tasks as anomalies arise [2].

As pre-programmed rule based logic may not be able to anticipate all possible repair scenarios, it becomes necessary for the robot to learn novel skills to adapt to new scenarios. With time-critical efficiency in mind, it would be advantageous to enable a human to teach the requisite skills to the robot remotely through demonstration, rather than writing and testing a new program from scratch. Learning from demonstration can be applied via kinesthetic teaching, teleoperation, or passive demonstration [3]. Kinesthetic teaching involves physically moving the robot through a task trajectory, and therefore requires the physical presence of the human. Similarly, teleoperation requires remote manipulation of the robot, a task which can be challenging in many space deployments due to severe latency and incomplete state information [4]. Furthermore, the demonstrator may be an expert in the repair task, but not necessarily be an expert in robot manipulation. As such, we would like to enable non-experts in robotics to be able to convey their intended task execution via passive demonstration. Within passive demonstration, a demonstrator completes the task independently and without involving the robot, and the robot maps the demonstrator execution to its own state-action space. [3]. This setup enables the robot to learn to complete tasks simply by observing a demonstrator.

However, mapping between demonstrator and robot is difficult due to differences in anatomy. Smith et al. was the first to leverage image stylization via CycleGAN [5] as a method for demonstration translation [6]. CycleGAN looks to translate demonstrations between unpaired images, where there does not exist a 1-1 correspondence label between source and target. Unpaired correspondence is valuable, as it only requires paired collection of human demonstrations with random robot motions. However, excess differences between source and target domains results in increased visual artifacts during translation [5], [6]. Visual artifacts reduce the clarity of the demonstration to the robot learner, which compounds the difficulty of policy learning for task execution thereafter.

Our objective is to retain asymmetrical mapping to preserve demonstration efficiency, but reduce visual artifact generation by centering the translation with ground truth about robot kinematics. We propose a new angle of exploration for unpaired human-to-robot demonstration transfer to convey the intended course of action. We 1) illustrate how current state of the art pixel-to-pixel demonstration translation lacks a grounding in physical reality, 2) propose methods to ground the demonstration transfer with robot joint angle information, and 3) discuss avenues to learn a correspondence between images of human demonstrations and joint angle space.

II. Related Work

Prior methods have sought a variety of other approaches to tackle the correspondence issue between human and robot anatomy when learning from passive demonstration. Some approaches looked to center the demonstration on objects in the scene, teaching the robot to place objects in relative proximity to each other [7], [8]. However, these approaches assume the robot has been pre-programmed to interact with the object, and does not scale well with number of tasks.

NASA's Space Technology Research Grants Program.

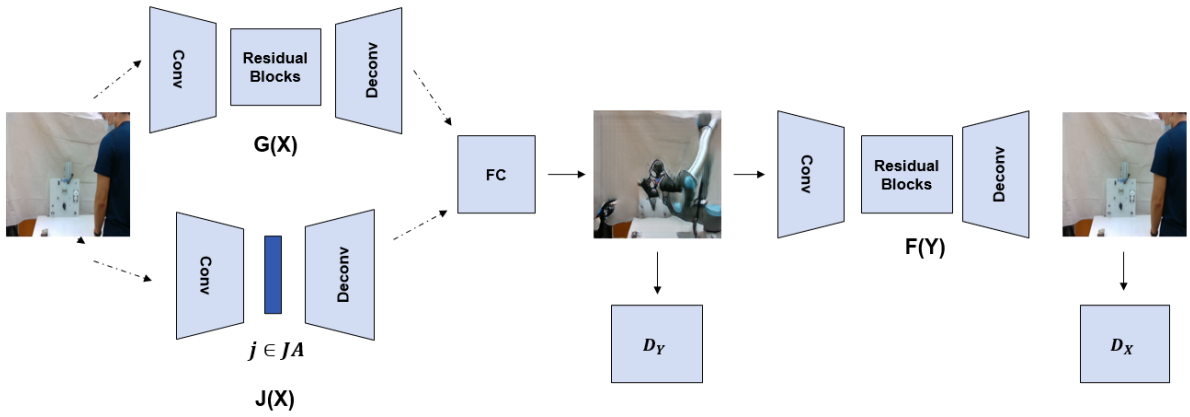

Fig. 1: Augmentation of CycleGAN network architecture from human image to robot image to reconstructed human image. We introduce an additional human to robot mapping directed through joint angle space j.

Alternatively, other methods seek to redefine the correspondence issue by prioritizing end effector location to match the hand location of the human demonstration via hand detection [9] and trajectory planning [10], [11]. This correspondence criterion grounds learned robot actions in viable trajectories, but imposes additional costs in sensing and motion planning. Critically, object interaction is also implicitly assumed to be pre-programmed. As such, we consider the direction of unpaired image translation as most promising in trade-off efficiency between limiting engineering and demonstrator workload, and ability to adapt to generalized scenarios.

III. Preliminaries

In the context of human-to-robot demonstration, we look to learn a mapping with images of the human giving the demonstration, and images of the robot generating random motions within the designated workspace. Discovering the mapping amounts to learning an unpaired translation in the absence of labeled correspondence between image-to-image pairs. Previous work utilized CycleGAN [5], which learns a mapping ($G : X \; ! \; Y$) from source to target domain, and another from target to source ($F : Y \; ! \; X$). Each translation is also taught to fool their respective domain discriminators, D_X and D_Y. These discriminators are trained on real and generated images, leading to paired loss functions detailed in Equation 1 and 2, In addition, CycleGAN works using an additional loss function for cycle consistency of reconstructed images, given in Equation 3, leading to the overall loss function in Equation 4, where is a hyper-parameter controlling the importance of the cycle-consistency objective.

$$L_{GAN}(G; D_Y) = \mathbb{E}[log D_Y(y) + log(1 \quad D_Y(G(x))] \quad (1)$$

$$L_{GAN}(F; D_X) = \mathbb{E}[log D_X(x) + log(1 \quad D_X(F(y))] \quad (2)$$

$$L_{rec}(G; F) = \mathbb{E}jjx \quad F(G(x))jj_1 + \mathbb{E}jjy \quad G(F(y))jj_1 \quad (3)$$

$$L_{CYC} = L_{GAN}(G; D_Y) + L_{GAN}(F; D_X) + \quad L_{rec}(G; F) \quad (4)$$

As noted in previous literature, CycleGAN does not exploit implicit temporal information from demonstration videos and translated images often include visual artifacts, which makes the translated demonstration ambiguous. Further, these artifacts are more frequent when the source and target domains are dissimilar. For instance, a mapping between horses and zebras may yield artifacts infrequently, while a mapping between cats and dogs yields artifacts regularly [5]. However, we can create a better image translations by leveraging implicit pose information, as robots exhibit rigid transforms between joints, whereas animals will have much more deformable poses.

IV. Formulation

Our approach is to augment CycleGAN with a third mapping generator ($J : X \; ! \; Y$), which has the distinct property of utilizing an embedding vector (see Figure 1) regularized to generate exact robot joint angles. Our objective in training generator $J : X \; ! \; Y$ is to create a mapping grounded in feasible robot kinematics, and thus generate a separate robot image representation that captures implicit pose information. In particular, we seek to train the generator $J : X \; ! \; Y$ as a disjoint encoder (J_E) and decoder (J_D) pair: the encoder seeks to propose joint angles based on the image of the human demonstrator, and the decoder seeks to map the proposed joint angles to a robot image within the scene (see Figure 1).

Separation of $J : X \; ! \; Y$ into disjoint pair J_E and J_D is advantageous as the decoder can be pre-trained via supervised learning without incurring additional costs to the human demonstrator. That is we can obtain robot joint-angles cost-free during the collection of images of random robot motion. We can construct a supervised loss (Equation 5) between the expected robot image and the decoded robot image using the joint angles as input.

$$L_2 = jjy \quad J_D(j)jj \quad (5)$$

Consequently, the challenge of learning generator $J : X \; ! \; Y$ lies with the encoder, which requires matching human im-

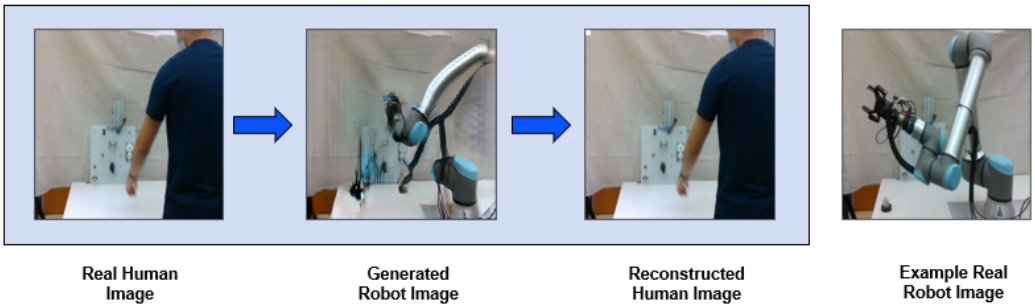

Fig. 2: Example reconstruction from cycleGAN. While the re-projected human image matches the original, the generated robot image exhibits high visual artifacts.

ages to real robot joint angles without an explicit supervisory signal. To tackle this challenge, we learn both human to robot mappings concurrently, with $G : X \; ! \; Y$ looking to learn image features and $J : X \; ! \; Y$ looking to update the encoder. We utilize both outputs as a weak supervisory signal to each other using the same loss function in Equation 6, where we penalized the pixel difference between both decoder outputs.

$$L_{diff} = jjG(x) \quad J(x)jj \tag{6}$$

In addition, to regularize the encoding output in real joint angle space, we apply the following criterion:

1) The output of the encoder is a sigmoid function. When training the decoder, we apply a linear transform to compress the range of the each joint angle to [0, 1].
2) We incentivize joint angle feasibility by penalizing the minimal distance between the output and the set of collected joint angles $fJAg$ from random motion.

$$L_{real} = min \; jjJ_E(x) \quad jjj; j \; 2 \; JA \tag{7}$$

3) We leverage temporal adjacency from the demonstration, by penalizing the distance in embedding space for encoding of sequential human frames x_i.

$$L_{seq} = jjJ_E(x_i) \quad J_E(x_{i+1})jj \tag{8}$$

In summary we train the overall network by combining cycleGAN losses with the generator loss in Equation 9 and a separate back-propagation for the encoder in Equation 10.

$$L_{full} = L_{CYC} + L_{diff} \tag{9}$$

$$L_{J_E} = L_{real} + L_{seq} \tag{10}$$

V. Evaluation

A. Data Collection

To evaluate the viability of our approach, we collect demonstrations and random robot motions on a UR10 on pick and place tasks similar to Smith et al. [6]. For each task, we used 20 human demonstrations of the pick and place action to obtain 1,000 images across demonstrations, along with 500 images of random human data. The robot data consisted of 12,000 images of random robot movements across six different environment settings modified by moving the starting location of the block. Crucially, we differ by allowing full body motion during the pick and place demonstration, and a larger robot.

B. Image Artifacts from CycleGAN

We first train cycleGAN using the default parameters set by the repository associated from the original paper [5]. From the gradient descent, we note that the algorithm successfully minimizes reprojection loss (see Figure 2), but the discriminator is never fooled by generated robot images.

We highlight two key examples of image artifacts in Figure 3. In the human robot image pair on the left, the human has minimal reach into the scene, and the corresponding robot image has a high amount of local features, with some level of distortion. Conversely, the image pair on the left shows a scene where the human is placing an object, and leaning forward into the scene. Consequently, the generated robot image struggles to replace the entire region of the human demonstrator with a robot, and the scene is increasingly distorted with artifacts. Thus we believe that the dissimilarity between human anatomy and UR10 anatomy is different enough that an alternative mapping grounded in joint angle space is necessary for construction of realistic robot images.

C. Joint Mapping Training

We then split the dataset of random robot motion images and corresponding joint angles in a training and test set with a 80/20 ratio. We train the joint decoder J_D via the supervised loss in Equation 5, and the discriminator in Equation 1. In Figure 4, we see that although we do not achieve complete convergence on image reconstruction from unknown joint angles, the overall robot pose alignment is retained.

We then tested training the human to robot joint encoder J_E via the loss functions indicated in Equations 6, 7, and 8 for optimizing a human to robot joint angle correspondence. However, we encountered difficulties in concurrent convergence, where upon decomposition, it was found that gradient descent could occur for Equation 7 and 8 separately, but not at the same time. Thus we desire alternative approaches to Equation 8 to leverage temporal information from human