Sim2real Image Translation Enables Viewpoint-Robust Policies from Fixed-Camera Datasets

Jeremiah Coholich[1], Justin Wit[1], Zsolt Kira[1]

***Abstract*—Vision-based policies for robot manipulation have achieved significant recent success, but are still brittle to distribution shifts such as camera viewpoint variations. One reason is that robot demonstration data used to train such policies often lacks appropriate variation in camera viewpoints. Simulation offers a way to collect robot demonstrations at scale with comprehensive coverage of different viewpoints, but presents a visual sim2real challenge. To bridge this gap, we propose an unpaired image translation method with a novel segmentation-conditioned InfoNCE loss, a highly-regularized discriminator design, and a modified PatchNCE loss. We find that these elements are crucial for maintaining viewpoint consistency during translation. For image translator training, we use only real-world robot play data from a single fixed camera but show that our method can generate diverse unseen viewpoints. We observe up to a 46% absolute improvement in manipulation success rates under viewpoint shift when we augment real data with our sim2real translated data.**

I. Introduction

There has been significant progress in learning vision-based policies for manipulation tasks based on demonstrations. This includes recent Transformer-based architectures trained via behavior cloning methods [1] as well as large-scale vision-language-action (VLA) models [2], [3]. Unlike pure vision or language training, which can be scraped from the web, robot demonstration data is scarce and lacks diversity; even pre-trained models must be fine-tuned to achieve significant performance under environment or viewpoint shifts. As a result, there remains a significant challenge in obtaining diverse, high-quality demonstration data for downstream tasks.

In order to achieve generalization, a comprehensive robot demonstration dataset must cover all scene and task variations the robot may encounter during deployment. Practically however, it is onerous to collect tabletop manipulation datasets that have a range of camera viewpoints since they often employ fixed, third-person cameras [4]–[6]. Cameras may be fixed to provide consistent environments for visual policy evaluation or due to the cost of recalibrating cameras, especially in conjunction with other sensors such as depth or motion capture sensors. Indeed, we have observed empirically that when robot polices are trained on fixed-camera datasets, changes in camera viewpoint during deployment causes drastic decrease in performance (Table II). Adapting or generalizing to unseen viewpoints is difficult because the viewpoint affects all objects in the scene, and the model must implicitly estimate the robot's position relative to the new camera position. Therefore, demonstration data with diverse viewpoints must be incorporated into the training dataset.

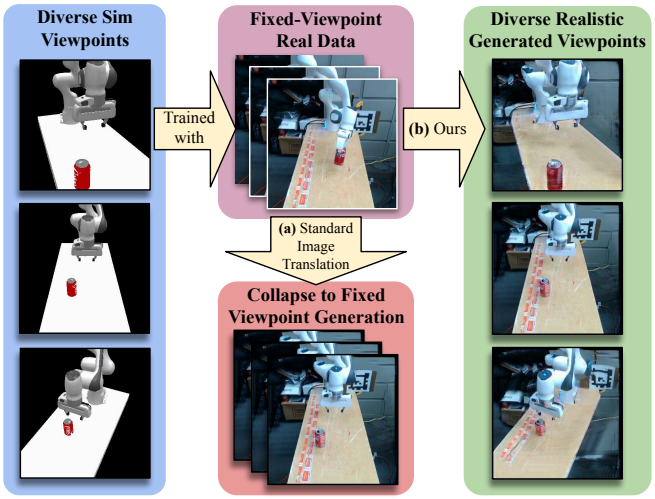

Fig. 1. (a) Standard image translation methods fail to generalize to new viewpoints when trained on fixed-viewpoint target domain dataset. (b) Our method enables realistic generation of unseen viewpoints

To this end, we propose to instead *generate* simulated demonstrations covering diverse viewpoints and propose a method for bridging the visual sim2real gap. One option is to add visual domain randomization in simulation for lighting, textures, and colors [8]–[10]. Other researchers opt instead to improve simulator realism [11] – a labor-intensive engineering effort which must be done for every scene. In leiu of simulation engineering, we propose learning a generative, segmentation-conditioned unpaired image-to-image translation model mapping from sim to real. Our model is trained on a small, real-world play dataset collected from a single fixed camera that is then able to generate a range of diverse viewpoints (Figure 1). The output of our entire pipeline is a dataset of synthetic demonstrations which can be combined with a small amount of real demonstrations and fed into any downstream learning algorithm.

Our core contributions are as follows:

1) A sim2real image translation method incorporating a novel, segmentation-informed contrastive InfoNCE loss capable of preserving unseen simulation viewpoints during translation
2) Experimental proof that demonstrations generated with our simulation plus translation pipeline improves

[1]Institute of Robotics and Intelligent Machine, Georgia Institute of Technology, Atlanta, GA, USA. Emails: {jcoholich, jwit3, zkira}@gatech.edu.

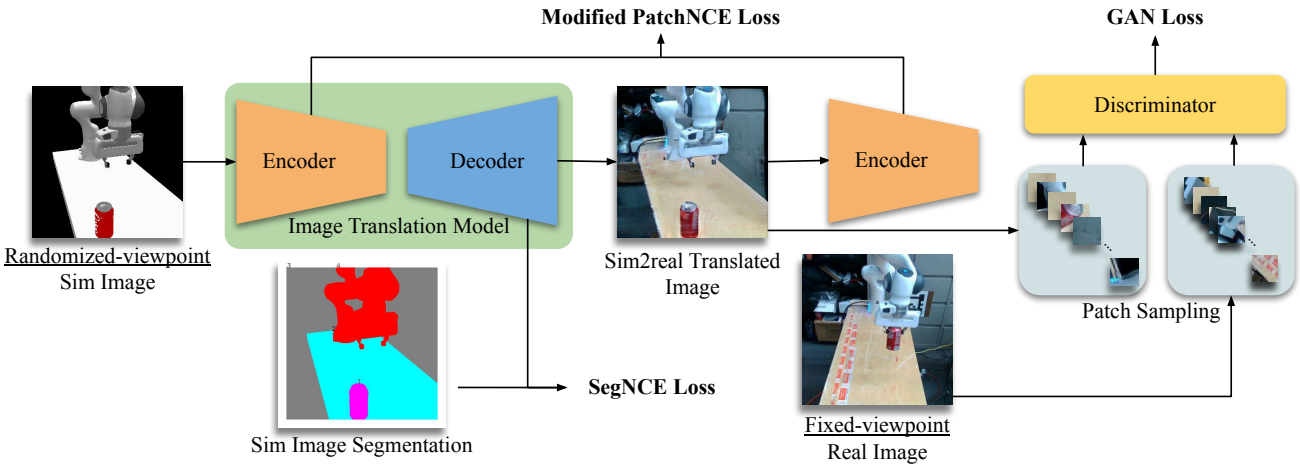

Fig. 2. Our proposed image translation method is trained on unpaired real and sim images, specifically a real dataset obtained with a fixed camera and a simulation dataset with diverse camera viewpoints. To ensure the simulation viewpoint is preserved during translation, we employ a novel segmentation-based InfoNCE loss, a modified PatchNCE loss from [7], and a novel patch-sampling process to regularize the discriminator D.

downstream robot policy robustness to shifts in camera position

3) Analysis of why our method is successful on the domain of robot demonstration datasets in comparison to the many other approaches developed for the generic problem of unpaired image translation

Code and videos can be found at www.sites.google.com/view/sim2real-viewpoints.

II. Related Work

A. Visual Sim2Real Translation for Robotics

A wide variety of unpaired image-to-image translation architectures exist. Among the most prominent are CycleGAN [12], CUT [7], EGSDE [13], StarGAN [14], DualGAN [15], OASIS [16], and ILVR [17]. To translate from sim2real, robotics practitioners can train these models on datasets collected from simulation and real-world robot observations. For example, in [18] and [19], the authors train an unmodified CycleGAN to translate visual observations for grasping and navigation.

Others have tailored these methods to the specific robotics and policy-learning applications. DigitalTwin-CycleGAN adds an action cycle-consistency loss to CycleGAN for a sim2real visual grasping task [20]. This loss means their image translation model depends on learning a successful grasping policy concurrently. RL-CycleGAN incorporates Q-function consistency on translated images [21], where the Q-function is obtained while learning a task-specific RL policy. RetinaGAN enforces cycle consistency with an object detector which requires thousands of labeled images to train beforehand [22]. GraspGAN trains an image translation model without cycle-consistency and instead enforces accurate image content translation through a grasp success predictor [23]. Additionally, they include an auxiliary generator objective of reproducing the ground-truth sim image segmentation.

CyCADA places these methods within a more general framework with the concept of a "task loss" [24]. The authors train image translation models which enable source-domain images to be segmented or classified with models trained on the target-domain. In contrast, our proposed method is agnostic to the downstream learning algorithm, enabling us to train one image translation model for many tasks.

Diffusion models [25] have emerged as the primary architecture for image generation over generative adversarial networks (GANs) [26], with some exceptions [27]–[29]. However, we find that for the specific domain of unpaired image-to-image translation, GANs obtain results competitive with the best diffusion approaches [13]. We hypothesize that that multimodal capabilities that enable diffusion models to generate diverse outputs are not an advantage when the style and content of the generated image are tightly-specified by the input image and target domain dataset, respectively. Our proposed method uses a GAN; however in theory our novel segmentation-based InfoNCE loss could be applied to any image translation architecture containing a generator network with spatially-indexed latent feature maps.

B. Robot Viewpoint Invariance

RoboNet offered early proof that training a robot policy on multiple views helps it generalize across these views [30]. Multi-view Masked World Models (MV-MWM) [31] demonstrates impressive robustness to camera viewpoints by training a viewpoint-invariant visual encoder and task-specific world model. They showcase real-world robot experiments, relying on small sim2real gap for sim2real transfer. In contrast to the MV-MWM evaluation environment with a solid black background a white tabletop, we evaluate our downstream polices in a messy lab environment.

MoVie [32] achieves view generalization by adapting the policy's image encoder to the novel views encountered during test-time. We do not do any test-time adaptation or

require any real-world images from viewpoints other than a single fixed-camera. [33] trains an RL policy that is robust to single-camera viewpoint changes after learning from a teacher policy trained with a multi-view observation. Impressively, VISTA leverages pretrained models with 3D priors to generate novel viewpoints given a single real-world image observation [34]. However, since they do not use simulation-generated demonstrations they are unable to generate new robot trajectories and must rely on human demonstration collection. 3D representations are inherently viewpoint invariant, but building good representations typically requires more data than a single 2D image. For example, GROOT [35] achieves viewpoint invariance but requires depth imaging.

There are real world collection efforts with custom-built platforms incorporating easily-movable cameras. Real-world datasets that are sufficiently large to cover a large range of viewpoint variations should train policies that serve as the upper-bound of viewpoint robustness. Viewpoint variations further increase the amount of data required and labor per demonstration [36] [5].

III. Method

We propose a novel unpaired image-to-image translation method to translate image observations from the simulation domain to the real world domain (Figure 2). The objective of unpaired image-to-image translation is to translate images from domain A to domain B without access to a paired dataset of images $\mathcal{D}_{paired} = \{d_A,\ d_B | d_A \in A,\ d_b \in B\}_{i=0}^{N}$. Instead, we learn from two separate datasets $\mathcal{D}_A$ and $\mathcal{D}_B$. In our problem setting, domain A is simulation, domain B is the real-world, and $|\mathcal{D}_A| > |\mathcal{D}_B|$. This problem is considered unsupervised because there is no label, or ground-truth image, in $\mathcal{D}_B$ that images in $\mathcal{D}_A$ should map to.

Image translators must change the style of the input image while preserving its content. We employ the GAN architecture with a novel, highly-regularized discriminator to learn the target domain style. For accurate content preservation, we use the InfoNCE [37] loss between input and output image features in a similar style as CUT [7], but with a modified scoring function. Additionally, we propose a novel segmentation-based InfoNCE loss on generator features. With our method, we can train a model on a small, fixed-viewpoint dataset $\mathcal{D}_B$ which is capable of accurate geometric content translation to diverse, unseen viewpoints.

A. Style Loss

We use the standard GAN loss [26] to enforce target domain style on the generated images, given by Equation 1. G is the generator network, and D is the discriminator network.

$$\mathcal{L}_{\text{GAN}}(G, D, \mathcal{D}_A, \mathcal{D}_B) = \mathbb{E}_{x \sim \mathcal{D}_B} \log(D(x)) + \mathbb{E}_{y \sim \mathcal{D}_A} \log(1 - D(G(y))) \quad (1)$$

One assumption underlying image-translation GANs, such as CycleGAN [12] and CUT [7], is that the shared attributes among all images in D_B constitute the target domain "style". However, our real-world robot image observations in $\mathcal{D}_B$ only differ from one another in robot and object poses. Much of the image content, such as the background and tabletop, is nearly identical in all images in $\mathcal{D}_B$. A naïve discriminator will memorize the repetitive details and force the generator to recreate them. To mitigate this problem, Pix2pix [38] proposed a "PatchGAN", where the discriminator only receives local image patches and cannot therefore enforce global image elements. We take this a step further and randomly sample patch locations and apply per-patch random rotations. This process is shown in Figure 3. The result is a highly-regularized discriminator capable of enforcing the style of $\mathcal{D}_B$ on images with viewpoints not seen in $\mathcal{D}_B$.

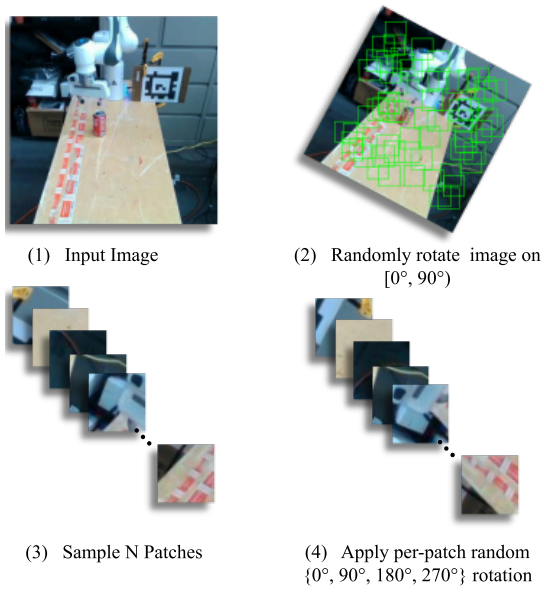

Fig. 3. Discriminator D Patch Sampling Process

B. Content Loss

Our content translation loss consists of two parts: a modified version of the PatchNCE loss from CUT [7] and a novel segmentation-based NCE loss.

1) Modified PatchNCE Loss: The PatchNCE loss from CUT [7] applies an info noise-contrastive estimation [39] (NCE) loss across encoder features generated by a source domain image $d_A \in D_A$ and its corresponding translated output image $G(d_A)$. For an input image d, we randomly sample N latent features from the encoder's feature map at L different layers. We call the set of features at layer l $\mathcal{Z}_l$ and $|\mathcal{Z}_l| = N\ \forall\ l \in \{l_0, ..., l_L\}$. The translated image $\hat{d}$ is passed through the encoder again to obtain $|\hat{\mathcal{Z}}_l|\ \forall\ l \in \{l_0, ..., l_L\}$. All $\hat{\mathcal{Z}}_l$ are obtained from the same feature map indices as in $\mathcal{Z}_l$.

The InfoNCE loss for feature i in encoder layer l is given by Equation 2. This is the categorical cross-entropy loss on the probability that a feature $\mathbf{z} \in \mathcal{Z}$ will be correctly classified as the corresponding feature in $\hat{\mathcal{Z}}$, based on a

scoring function $\rho_l(\cdot)$. See [7] for further details. τ is a temperature hyperparameter.

$$\ell_{\text{NCE}}(l, \mathbf{z}, \hat{\mathcal{Z}}, i) = -\log\left[\frac{\exp\left(\rho_l(\mathbf{z}, \hat{\mathbf{z}}_i)/\tau\right)}{\sum_{\hat{\mathbf{z}}\in\hat{\mathcal{Z}}}\exp\left(\rho_l(\mathbf{z}, \hat{\mathbf{z}})/\tau\right)}\right] \tag{2}$$

$\rho_l(\cdot)$ is defined in Equation 3. Features $\mathbf{z}_i$ and $\mathbf{z}_j$ are passed through a function H_l then scored with cosine similarity.

$$\rho_l(\mathbf{z}_i, \mathbf{z}_j) = \frac{H_l(\mathbf{z}_i)\cdot H_l(\mathbf{z}_j)}{\|H_l(\mathbf{z}_i)\|\|H_l(\mathbf{z}_j)\|} \tag{3}$$

The full loss is given in Equation 4

$$\mathcal{L}_{\text{PatchNCE}}(G, H, \mathcal{D}) = \mathbb{E}_{d\sim\mathcal{D}}\sum_{l=1}^{L}\sum_{i=1}^{|\mathcal{Z}_l|}\ell_{\text{NCE}}(\ l,\ \mathbf{z}_{l,i},\ \hat{\mathcal{Z}}_l,\ i\ ) \tag{4}$$

The assumption behind Equation 2 is that input and output features from the same feature map locations are "positive" samples and should have high similarity scores. All other features are "negative" samples and should be repelled. However, we observe that many different input image patches are highly similar due to repeated patterns or textures in robot image datasets, which include background elements, the tabletop, etc. Furthermore, in the simulated image dataset $\mathcal{D}_A$, these regions all have *identical* pixel values. Therefore, Equation 2 will repel many false negative features.

To mitigate this, we modify the scoring function. If the cosine similarity of a negative sample exceeds a threshold θ, we scale the value down by a factor $0 \leq \alpha < 1$. The modified scoring function is given by Equation 5.

$$\tilde{\rho}_l(\mathbf{z}_i, \mathbf{z}_j) = \begin{cases}\alpha\,\rho_l(\mathbf{z}_i, \mathbf{z}_j) & \text{if } \rho_l(\mathbf{z}_i, \mathbf{z}_j) > \theta \text{ and } i \neq j,\\ \rho_l(\mathbf{z}_i, \mathbf{z}_j) & \text{otherwise.}\end{cases} \tag{5}$$

We have empirically found this to be more effective than increasing τ. We denote the modified NCE loss which uses the scoring function in Equation 5 as $\tilde{\mathcal{L}}_{\text{PatchNCE}}$.

2) Segmentation NCE Loss: We leverage the simulator used to generate $\mathcal{D}_A$ to obtain ground-truth segmentation maps for each image. We propose an InfoNCE loss which clusters generator features based on segmentation category in order to ensure that semantic segmentation boundaries are preserved during translation. We note that for object-centric manipulation, this is a crucial aspect to preserve when training policies.

Each image in $\mathcal{D}_A$ contains C segmentation classes and each feature $\mathbf{z}_i \in \mathcal{Z}$ generated from $d \sim D_A$ has an associated class label $y_i \in \mathcal{Y}$. In the case when a layer's feature map is of a lower resolution than the input image, we scale the image segmentation with nearest-neighbors downsampling to obtain $\mathcal{Y}$.

The Segmentation NCE (SegNCE) loss is defined in Equation 6. In contrast to ℓ_{NCE} as shown in Equation 2, there are multiple positive samples for the query feature $\mathbf{z}_i$. All features that are members of the same segmentation class as the query feature are positive samples assigned the index j.

In Equation 2, the target distribution for the cross-entropy loss is a one-hot vector. In Equation 6 the target distribution is a uniform distribution over features from the same segmentation class and zero elsewhere.

Here, we use the original scoring function $\rho_l(\cdot)$ defined in Equation 3; since we are operating with ground-truth image segmentations, there are no false negatives.

$$\ell_{\text{SegNCE}}(l, \mathcal{Z}, i, \mathcal{Y}) = \frac{1}{\left\{j\middle|\substack{j\in 1..|\mathcal{Z}|\\ y_j=y_i\\ i\neq j}\right\}}\sum_{\left\{j\middle|\substack{j\in 1..|\mathcal{Z}|\\ y_j=y_i\\ i\neq j}\right\}}\ell_{\text{NCE}}(l, \mathbf{z}_i, \mathcal{Z}, j) \tag{6}$$

The full loss term is given in Equation 7.

$$\mathcal{L}_{\text{SegNCE}}(G, H, \mathcal{D}_A) = \mathbb{E}_{d\sim\mathcal{D}_A}\sum_{l=1}^{L}\sum_{i=1}^{S}\ell_{\text{SegNCE}}(l, \mathcal{Z}_l, i, \mathcal{Y}_l) \tag{7}$$

The SegNCE loss is computed from input image generator features only.

C. Model training

The total loss function for G is given in equation 8. We include an identity PatchNCE loss following [7]. The full discriminator loss is given in Equation 9 and is simply the GAN objective.

$$\begin{aligned}\mathcal{L}_G = \tilde{\mathcal{L}}_{\text{PatchNCE}}(G, H, \mathcal{D}_A)\\ +\tilde{\mathcal{L}}_{\text{PatchNCE}}(G, H, D_B)\\ +\mathcal{L}_{\text{SegNCE}}(G, H, \mathcal{D}_A)\\ +\mathcal{L}_{\text{GAN}}(G, D, \mathcal{D}_A, \mathcal{D}_B)\end{aligned} \tag{8}$$

$$\mathcal{L}_D = -\mathcal{L}_{\text{GAN}}(G, D, \mathcal{D}_A, \mathcal{D}_B) \tag{9}$$

IV. Experiments

Our experiments are designed to answer the following questions:

1) How well can the proposed image translation method generalize to unseen real-world viewpoints?
2) Does the synthetic data generated with the proposed method and used for imitation learning make a downstream real-world robot policy more robust to shifts in camera position?
3) How does robot policy performance scale with sim data generated with our proposed method vs real world data?

For all experiments, we use a Coke can grasping task we call `pick coke` with a Franka arm.

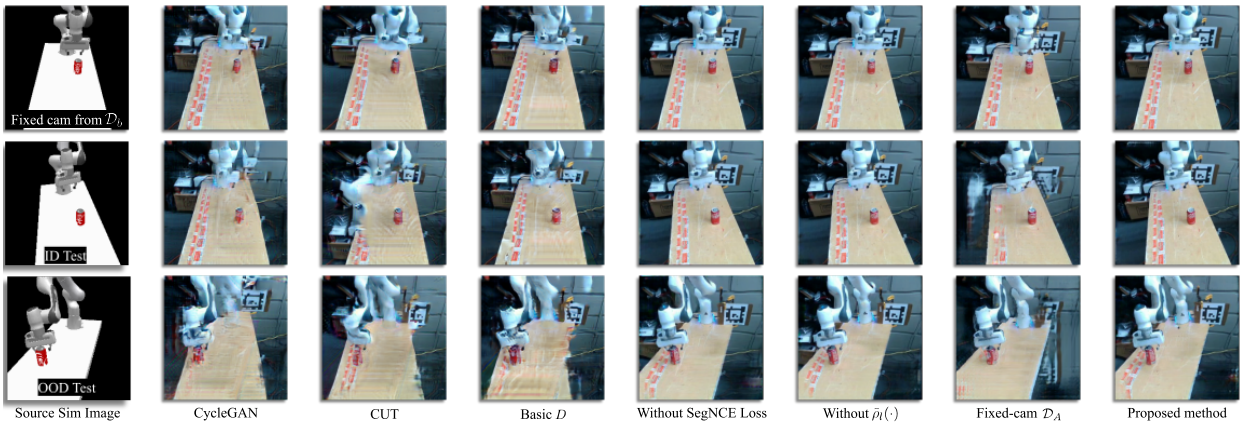

Fig. 4. Samples of translated images. Our proposed method achieves realistic results and preserves object locations and boundaries.

A. Image Translation

1) Training Details: Our generator G is a 12M parameter ResNet-based network. The discriminator D is a wider three-layer CNN with 11M parameters. Additionally, we parameterize the H_l in Equations 3 and 5 as a 2-layer MLP with 700k parameters.

For benchmarking image translation results, we use a dataset of 256×256 images. $\mathcal{D}_A$ is a dataset of 1,775 simulated `pick coke` demos with camera viewpoints randomized within a box of dimensions (50, 50, 20) cm (L×W×H). $\mathcal{D}_B$ contains 745 images obtained from roughly 2 and a half minutes of coke can play data. Training a single model takes approximately 20 hours on an RTX 2080 Ti GPU.

We curate two test datasets: In-Distribution (ID) Camera and Out-of-Distribution (OOD) Camera. Each test set contains 25 pairs of sim and real images from unseen viewpoints. ID Camera contains real and sim images with the same camera randomization range as the training sim images in $\mathcal{D}_A$: (50, 50, 20) cm. OOD Camera tests viewpoints randomized within a box of size (100, 100, 84) cm but excludes viewpoints contained in ID Camera. In order to recreate real images in sim, we use AprilTags and robot proprioception.

2) Metrics: We report on two metrics for image translation: segmentation mean intersection over union (mIOU) and Frénchet Inception Distance (FID) score. The former measures how well image content is translated, while the FID Score measures how semantically similar the generated images are to the real images.

To find mIOU, we compare a segmentation of the generated image with the ground-truth segmentation from the simulation. To segment generated images, we use the open-world segmentation model lang-SAM [40]. For the `pick coke` scenes, we query three segmentation categories with prompts ("Franka arm", "coke", "table"), while a fourth "background" category consists of the remaining pixels.

3) Results: Table I gives the scores of our proposed method against baselines and ablations. Our proposed method obtains the lowest FID score by a wide margin on ID Camera.

TABLE I

UNPAIRED IMAGE TRANSLATION FID(↓) AND MIOU(↑) SCORES. METRICS ARE AVERAGED ACROSS TWO RANDOM SEEDS. OUR NOVEL DISCRIMINATOR DESIGN HAS THE HIGHEST IMPACT. OUR PROPOSED METHOD SHOWS IMPRESSIVE GENERALIZATION EVEN WHEN $\mathcal{D}_A$ IS FIXED-CAMERA.

Method	**ID Camera**		**OOD Camera**	
	FID(↓)	mIOU(↑)	FID(↓)	mIOU(↑)
No Translation	411.2	0.88	379.7	0.85
CUT [7]	312.3	0.54	344.3	0.52
CycleGAN [12]	323.7	0.53	360.0	0.51
Basic D	331.5	0.54	362.0	0.43
Without SegNCE loss	177.6	0.79	226.5	0.70
Without $\tilde{\rho}_l(\cdot)$	171.6	0.82	**212.6**	0.70
Fixed-cam $\mathcal{D}_A$	214.7	**0.84**	280.3	**0.83**
Proposed Method	**167.9**	0.82	219.7	0.70

The novel patch discriminator D has the largest impact on all metrics, highlighting its importance in this domain. "No Translation" represents the upper bound for the mIOU and FID Score. An mIOU values less than 1.0 here reflects errors made by lang-SAM and small discrepancies in camera and object pose estimation obtained from AprilTags.

Translated image results are given in Figure 4. Methods without the novel D design are not able to maintain segmentation boundaries, as can be seen by the generated table positions on the OOD camera translations. We suspect that the SegNCE loss and the thresholding done in $\tilde{\rho}_l(\cdot)$ fulfill similar functions given the high image quality when these are ablated separately.

Impressively, when our proposed method is trained on fixed-cam simulated data (as opposed to sim data with ID camera randomization), it still produces competitive metrics. From Figure 4, one can observe visual aberrations in the background and on the arm, but the major segmentation boundaries are preserved.

Note that while the relative FID scores correlate with image quality, the numbers are high compared to results reported in other literature. We posit that this is due to the

TABLE II

SUCESS RATES ON `PICK COKE` TASK. WITHOUT DATA PRODUCED FROM OUR METHOD, POLICIES FAIL WHEN CAMERA POSITION IS SHIFTED. WE ALSO OBSERVE PERFORMANCE SCALES WITH THE AMOUNT OF SIMULATED TRANSLATED DATA

Policy Training Demonstrations				**Evaluation Viewpoints**		
Fixed-cam real demos	Simulated demos	Sim Camera Randomization	Visual Sim2real Method	Fixed	ID	OOD
100	0	-	-	5/15	0/15	0/15
100	500	ID	None	9/15	0/15	1/15
100	500	ID + OOD	None	2/15	0/15	0/15
100	500	ID	Domain Randomization	1/15	1/15	0/15
100	500	ID + OOD	Domain Randomization	4/15	0/15	0/15
100	500	ID	Ours	12/15	4/15	1/15
100	500	ID + OOD	Ours	9/15	2/15	2/15
100	1000	ID	Ours	13/15	2/15	1/15
100	1000	ID + OOD	Ours	8/15	**7/15**	3/15
100	1500	ID	Ours	**14/15**	6/15	**4/15**

small size of our test sets (25 images each) and that our robotics lab scene may be OOD for the Inception network used for FID Score.

We hypothesize that off-the-shelf methods like CUT and CycleGAN struggle on robot data due to lack of diversity. Typically, unpaired image-to-image translation methods are tested on computer vision benchmark datasets containing diverse images scraped from the internet. In comparison, robotics datasets contain limited diversity. To support this claim, we computed average pairwise LPIPS [41] on various image datasets and on our own dataset, shown in Table III. Our $\mathcal{D}_B$ shows the lowest score for diversity.

TABLE III

AVERAGE PAIRWISE LPIPS ON NATURAL IMAGE DATASETS. A CORE CHALLENGE IN ROBOT LEARNING IS LACK OF DATASET DIVERSITY AS COMPARED TO WEB DATA.
† AVERAGE PAIRWISE LPIPS COMPUTED ON 1000 RANDOMLY SAMPLED IMAGES.

Dataset	Average Pariwise LPIPS (↑)
Laion-5B†	0.725
ImageNet†	**0.819**
Cifar-10†	0.221
Cifar-100†	0.250
Horse2zebra $\mathcal{D}_A$ (Horses)	0.747
Horse2zebra $\mathcal{D}_B$ (Zebras)	0.765
Seg2Cityscapes $\mathcal{D}_B$ (Real)	0.548
`pick coke` $\mathcal{D}_B$ (Real)	0.155

B. Robot Grasping Experiments

We train `pick coke` grasping policies on a combined set of fixed-viewpoint real-world human demonstrations and variable viewpoint translated sim demonstrations. To collect human demonstrations, we use a modified OPEN TEACH VR teleoperation platform [42]. Our robot environment, sample observations, and policy rollouts are shown in Figure 6.

1) Policy Training and Evaluation Details: We train action chunking transformer (ACT) [1] policies on our generated data. ACT was chosen to isolate the effects of our image data as it does not incorporate any pretraining or language conditioning. Additionally, policies are trained without proprioception inputs so that they must rely entirely upon image observations to predict actions.

We train each ACT policy for 8,000 epochs with a chunk size of 10. Rollouts are done with temporal aggregation turned on. Success is defined as a successful grasp of the Coke can without dropping.

In addition to our own sim2real method, we compare against policies trained with simulated demos with no translation applied and with visual domain randomization (Figure 5). For domain randomization we follow [8] and randomize color, texture, and lighting for all objects in the scene.

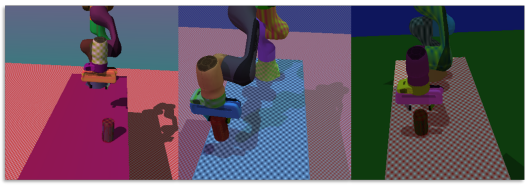

Fig. 5. Samples of domain randomized visual observations

2) Results: Results for the `pick coke` task are shown in Table II.

We observe that our method is necessary for viewpoint robustness, as compared to ACT models trained on fixed-cam human demonstrations only, which obtain a success rate of 0 under any camera variation. We also observe that our method is necessary to bridge the sim2real gap, since the policies trained on sim demos without translation obtain a similarly negative result on ID and OOD cameras. Finally, we can see that performance scales positively with the number of translated demonstrations produced by our method. The highest success rate for the challenging OOD camera positions was achieved by an ACT model trained on 1500 of our translated demos, the highest amount we tested.

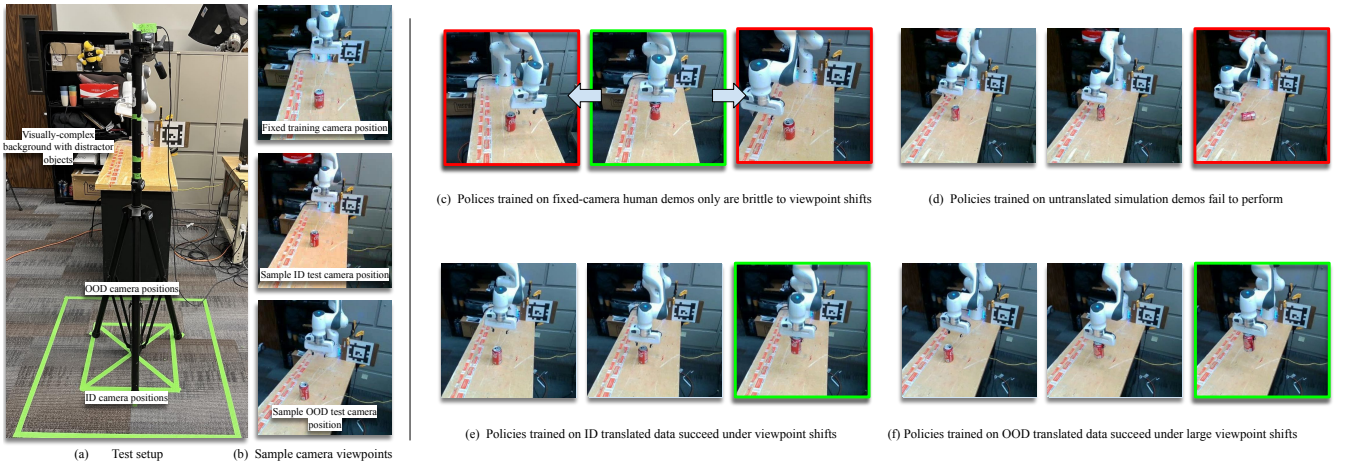

Fig. 6. (a) Our test setup, showing the 0.5m ID viewpoint box and the 1m OOD viewpoint box. (b) Sample ACT observations (c) Policies trained on fixed-viewpoint human demos only are brittle to small camera shifts. (d) Polices struggle to learn viewpoint robustness on untranslated sim demos. (e-f) Our translation method is robust to ID and OOD camera viewpoint shifts.

3) Failure Cases: We observe two common failure cases of trained ACT policies. The robot lacks depth information, leading to grasp attempts that occur when the gripper is aligned with the coke can from the camera's perspective but mispositioned along the camera's z-axis. The second failure case occurs when the Franka end-effector is framed by the background instead of the table, which typically occurs when testing an OOD viewpoint. The ACT policy begins to output repetitive actions which do not seem to correlate with any trajectory in the demonstration dataset. We hypothesize that this is due to degraded or unrealistic background generation in translated OOD image observations.

V. Conclusions and Limitations

We train a novel image generation method on simulated and real robot data. We observe that with only fixed-camera real data, our novel SegNCE loss, discriminator design, and modified PatchNCE loss enable generation of novel views. Our method improves the robustness of downstream imitation learning policies to camera shift, as demonstrated by success rates on a manipulation task. We observe that our method is necessary for sim2real transfer and that robot performance scales with the number of translated demos and their range of camera randomization.

There are several limitations to this work. When translating sim images with large camera shifts, our method fills the background by repeating textures and scene elements (a result of the fixed-camera training dataset). This can create a sim2real gap for the downstream imitation learning policy. Additionally, our generated demonstrations are dependent on a small amount of human demonstrations to enable viewpoint robustness in robot policies.

Future work includes further investigation of scaling trends and training multi-task or multi-scene translation models. Ultimately, we believe visual sim2real is a powerful tool for scaling robot learning datasets.

References

[1] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[2] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[3] A. Szot, B. Mazoure, H. Agrawal, R. D. Hjelm, Z. Kira, and A. Toshev, "Grounding multimodal large language models in actions," *Advances in Neural Information Processing Systems*, vol. 37, pp. 20 198–20 224, 2025.

[4] S. Haldar, V. Mathur, D. Yarats, and L. Pinto, "Watch and match: Supercharging imitation with regularized optimal transport," in *Conference on Robot Learning*. PMLR, 2023, pp. 32–43.

[5] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning (CoRL)*, 2023.

[6] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *2019 International conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 8943–8950.

[7] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020, pp. 319–345.

[8] L. Pinto, M. Andrychowicz, P. Welinder, W. Zaremba, and P. Abbeel, "Asymmetric actor critic for image-based robot learning," *arXiv preprint arXiv:1710.06542*, 2017.

[9] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray *et al.*, "Learning dexterous in-hand manipulation," *The International Journal of Robotics Research*, vol. 39, no. 1, pp. 3–20, 2020.

[10] R. Garcia, R. Strudel, S. Chen, E. Arlaud, I. Laptev, and C. Schmid, "Robust visual sim-to-real transfer for robotic manipulation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. ieee, 2023, pp. 992–999.

[11] X. Li, K. Hsu, J. Gu, K. Pertsch, O. Mees, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani *et al.*, "Evaluating real-world robot manipulation policies in simulation," *arXiv preprint arXiv:2405.05941*, 2024.

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[13] M. Zhao, F. Bao, C. Li, and J. Zhu, "Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3609–3623, 2022.

[14] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.

[15] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2849–2857.

[16] V. Sushko, E. Schönfeld, D. Zhang, J. Gall, B. Schiele, and A. Khoreva, "You only need adversarial supervision for semantic image synthesis," *arXiv preprint arXiv:2012.04781*, 2020.

[17] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "Ilvr: Conditioning method for denoising diffusion probabilistic models," *arXiv preprint arXiv:2108.02938*, 2021.

[18] H. Zhang, H. Liang, L. Cong, J. Lyu, L. Zeng, P. Feng, and J. Zhang, "Reinforcement learning based pushing and grasping objects from ungraspable poses," *arXiv preprint arXiv:2302.13328*, 2023.

[19] J. Truong, S. Chernova, and D. Batra, "Bi-directional domain adaptation for sim2real transfer of embodied navigation agents," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2634–2641, 2021.

[20] D. Liu, Y. Chen, and Z. Wu, "Digital twin (dt)-cyclegan: Enabling zero-shot sim-to-real transfer of visual grasping models," *IEEE Robotics and Automation Letters*, vol. 8, no. 5, pp. 2421–2428, 2023.

[21] K. Rao, C. Harris, A. Irpan, S. Levine, J. Ibarz, and M. Khansari, "Rl-cyclegan: Reinforcement learning aware simulation-to-real," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 157–11 166.

[22] D. Ho, K. Rao, Z. Xu, E. Jang, M. Khansari, and Y. Bai, "Retinagan: An object-aware approach to sim-to-real transfer," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 10 920–10 926.

[23] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige *et al.*, "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 4243–4250.

[24] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. Pmlr, 2018, pp. 1989–1998.

[25] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[27] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan *et al.*, "Scaling autoregressive models for content-rich text-to-image generation," *arXiv preprint arXiv:2206.10789*, vol. 2, no. 3, p. 5, 2022.

[28] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park, "Scaling up gans for text-to-image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 124–10 134.

[29] A. Sauer, T. Karras, S. Laine, A. Geiger, and T. Aila, "Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis," in *International conference on machine learning*. PMLR, 2023, pp. 30 105–30 118.

[30] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn, "Robonet: Large-scale multi-robot learning," *arXiv preprint arXiv:1910.11215*, 2019.

[31] Y. Seo, J. Kim, S. James, K. Lee, J. Shin, and P. Abbeel, "Multi-view masked world models for visual robotic manipulation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 30 613–30 632.

[32] S. Yang, Y. Ze, and H. Xu, "Movie: Visual model-based policy adaptation for view generalization," *Advances in Neural Information Processing Systems*, vol. 36, pp. 21 507–21 523, 2023.

[33] C. Acar, K. Binici, A. Tekirdağ, and Y. Wu, "Visual-policy learning through multi-camera view to single-camera view knowledge distillation for robot manipulation tasks," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 691–698, 2023.

[34] S. Tian, B. Wulfe, K. Sargent, K. Liu, S. Zakharov, V. Guizilini, and J. Wu, "View-invariant policy learning via zero-shot novel view synthesis," *arXiv preprint arXiv:2409.03685*, 2024.

[35] Y. Zhu, Z. Jiang, P. Stone, and Y. Zhu, "Learning generalizable manipulation policies with object-centric 3d representations," *arXiv preprint arXiv:2310.14386*, 2023.

[36] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024.

[37] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[38] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[39] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[40] L. Medeiros, "Lang segment anything," 2024, gitHub repository. [Online]. Available: https://github.com/luca-medeiros/lang-segment-anything

[41] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[42] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto, "Open teach: A versatile teleoperation system for robotic manipulation," *arXiv preprint arXiv:2403.07870*, 2024.