MoDA: Map style transfer for self-supervised Domain Adaptation of embodied agents

Eun Sun Lee[1[0000−0003−1731−5714]], Junho Kim[1[0000−0002−5947−2147]], SangWon Park[1[0000−0002−9735−1303]], and Young Min Kim[1[0000−0002−6735−8539]]

Seoul National University, Seoul, 08826, Republic of Korea
{ eunsunlee, 82magnolia, paulmoguri, youngmin.kim } @snu.ac.kr

Abstract. We propose a domain adaptation method, MoDA, which adapts a pretrained embodied agent to a new, noisy environment without ground-truth supervision. Map-based memory provides important contextual information for visual navigation, and exhibits unique spatial structure mainly composed of flat walls and rectangular obstacles. Our adaptation approach encourages the inherent regularities on the estimated maps to guide the agent to overcome the prevalent domain discrepancy in a novel environment. Specifically, we propose an efficient learning curriculum to handle the visual and dynamics corruptions in an online manner, self-supervised with pseudo clean maps generated by style transfer networks. Because the map-based representation provides spatial knowledge for the agent's policy, our formulation can deploy the pretrained policy networks from simulators in a new setting. We evaluate MoDA in various practical scenarios and show that our proposed method quickly enhances the agent's performance in downstream tasks including localization, mapping, exploration, and point-goal navigation.

Keywords: Domain Adaptation, Self-Supervised Learning, Image Translation, Embodied Agent, Visual Navigation

1 Introduction

The absence of ground-truth labels is a critical bottleneck for training embodied agents in complex 3D world. A widely-used alternative is to train the agents in interactive simulators [41, 44] which can load various 3D indoor scenes [7, 41, 44]. Yet, when the agent optimized for a simulator is deployed in the real world, it fails to persist its performance due to the various unseen environmental noises [31]. The domain gap between simulators and the real world may be diminished by modeling the environmental noises [9,30], but it is not possible to obtain the correct noise model for the countless combinations of practical set-ups. Nonetheless, the visual agents collect an enormous amount of unlabelled data from 3D scenes over spatial movement. If an agent can transfer its performance utilizing such data, the adaptation scheme can serve as a generic solution for an embodied agent to serve in diverse real environments.

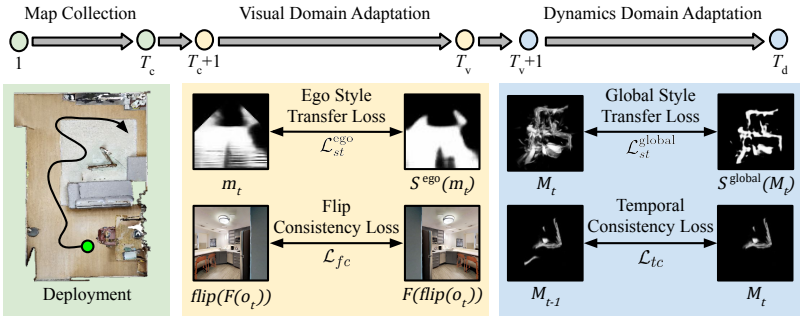

Fig. 1: MoDA suggests an integrated domain adaptation method for visual and dynamics corruptions, which fine-tunes an agent in an efficient learning curriculum. The agent collects a new map dataset to learn map style transfer networks (*green*). The agent is then trained for the new visual domain with ego style transfer loss and flip consistency loss (*yellow*). Lastly, the dynamics domain adaptation transfers the agent with global style transfer loss and temporal consistency loss (*blue*).

Many studies in embodied agents have shown how the map-based memory aids an agent for robust visual navigation [10, 32, 48]. The agent aligns its egocentric visual observations to generate a top-down map representing the environment's layout. The allocentric understanding helps the agent to localize itself and plan for various navigation tasks efficiently. Furthermore, the map provides a domain-agnostic representation, disentangling the agent's perceptual module from its planning. In a scenario where the pretrained agent is transferred to a new, noisy environment with various visual and dynamics corruptions, we suggest a domain adaptation method which only fine-tunes the domain-agnostic map memory, rather than the overall pipeline of the embodied agent.

To compensate for the absence of ground-truth in the real world, we propose a self-supervised domain adaptation method, MoDA. The proposed method transfers the pretrained agent to the new, noisy environment by learning style transfer networks on maps. Our objective is to learn the structural regularities of indoor scenes from the clean maps obtained from the noiseless simulator. We then transfer the style onto the new maps generated amidst visual and dynamics noises. Our self-supervision loss compares the generated maps with the style-transferred maps. More specifically, our method suggests a learning curriculum as shown in Fig. 1. First, the agent is deployed to collect the noisy maps and learns two style transfer networks for egocentric and global maps. We then transfer the agent for the visual corruptions through the ego style loss, followed by compensating for the dynamics corruptions with the global style loss. To stabilize the training, we additionally encourage the flip consistency on RGB observations and temporal consistency over the agent's movement. MoDA provides an integrated self-supervised solution for both visual and dynamics corruptions and enables online adaptation in an environment where the ground-truth is unavailable.

We analyze MoDA in various domain adaptation scenarios where the pre-trained agent is transferred to the novel environments with both visual and dynamics corruptions. We investigate multiple types of visual corruptions along with the two main dynamics corruptions, which are the odometry sensor noise and actuation noise. Our experiments show that the proposed adaptation method effectively enhances the embodied agents' performance in localization, mapping, and the downstream navigation tasks. To summarize, our main contributions are as follows: i) we propose a self-supervised domain adaptation method using map style transfer, ii) we suggest an efficient curriculum to learn an adaptation integrated for visual and dynamics corruptions, and iii) we demonstrate that the proposed approach enhances the agent's performance in localization, mapping, and the final downstream visual navigation tasks in novel, unseen environments.

2 Related Work

In this section, we describe existing approaches for our main task, visual navigation and simulation-to-reality adaptation (Sim2Real), along with methods for image translation which is the key technique of our work.

Visual Navigation The objective of visual navigation is to devise vision-based mapping and planning policies for solving a designated task. Classical approaches generate a map of the environment using SLAM techniques [5, 18, 19], and apply planning algorithms [26, 35, 46] using the generated map. On the other hand, recent learning-based approaches often train agents end-to-end with integrated mapping and planning [9–11, 39]. These agents have shown competitive performance in a wide variety of tasks such as embodied QA [15] and goal-oriented navigation [10, 39].

Maintaining a dedicated spatial memory unit is a key to such learning-based navigation agents. The spatial structure of the surrounding environment is often implicitly encoded with LSTM or GRU [14, 27], or using graph structures that embed keyframes as graph nodes [11, 13, 16, 40]. Nevertheless, map-based memories that depict spatial information on occupancy grid maps [8–10, 39] efficiently aid embodied agents for tasks requiring long-range tracking and spatially-grounded planning. We mainly retain our focus on map-based spatial memory and propose a self-supervised task formulated on grid maps for effective domain adaptation.

Sim2Real Simulators enable training an embodied agent with ground-truth poses or labels. While recent simulators [41, 44] can realistically model the world to a certain degree, there are non-idealities in agent and object dynamics. More importantly, domain gap is inevitable when deploying an agent trained in simulation to real-world environments. As a result, agents trained on simulators often fail to generalize in real-word settings [31]. To alleviate Sim2Real gap, domain randomization [12, 45] proposes to train the agent in various dynamics and visual simulations, which in turn allows the agent to observe a wide range of domains

before actual deployment. Furthermore, representation learning techniques are used to improve the generalization performance in the context of embodied agent studies [17, 20, 25, 42, 43]. Alternatively, Hansen et al. [24] proposes to adapt the policy during real-world deployment, using self-supervised objectives such as inverse dynamics and rotation prediction. Recently, Lee et al. [36] introduced a self-supervised domain adaptation algorithm that is formulated upon occupancy maps, which showed performance enhancement in various deployment scenarios. However, Lee et al. [36] requires multiple agent round trips for successful adaptation, which limits the practical usage of the algorithm.

We compare MoDA against existing approaches for adapting agents to Sim2Real deployment and demonstrate that MoDA can perform effective adaptation without mandating fixed agent trajectories such as round trips.

Image-to-Image Translation The goal of image-to-image translation is to transfer an image from a source domain into the style of target domain but to maintain its key contents. Early approaches such as Pix2Pix [29] propose to use generative adversarial networks (GANs) [22] for paired image-to-image translation. CycleGAN [50] aims to solve a more challenging problem of unpaired translation, where only a group of source and target domain images are provided for training. To accommodate for the lack of paired data, CycleGAN [50] proposes a novel cycle consistency loss that learns a forward and backward mapping simultaneously, leading to realistic image transfer. We leverage CycleGAN for transforming occupancy grid maps to the target domain, which allows for effective map generation under new, unseen environments. While recent advances in image-to-image translation enable high-resolution or multi-modal synthesis [28, 47, 51], we find that CycleGAN [50] is sufficient for transferring between occupancy grid maps that is not as diverse as real-world images.

3 Method

Given a pretrained agent from a noiseless simulator, MoDA transfers the agent to unseen, noisy environments with visual and dynamics corruptions. We first describe the overall pipeline of visual navigation with map-based memory in Sec. 3.1. Then Sec. 3.2 describes our map-to-map style transfer network which serves as the self-supervision signal. Lastly, Sec. 3.3 provides the learning curriculum of our online domain adaptation.

3.1 Visual Navigation with Spatial Map Memory

MoDA builds on conventional map-based navigation agents, where the action policy is planned based on map representation as shown in Fig. 2. The global map is estimated from the mapping and localization models. At each step, the RGB observation o_t is given as an input to the mapping model f_M which predicts the egocentric map m_t:

$$f_M(o_t) = m_t. \tag{1}$$

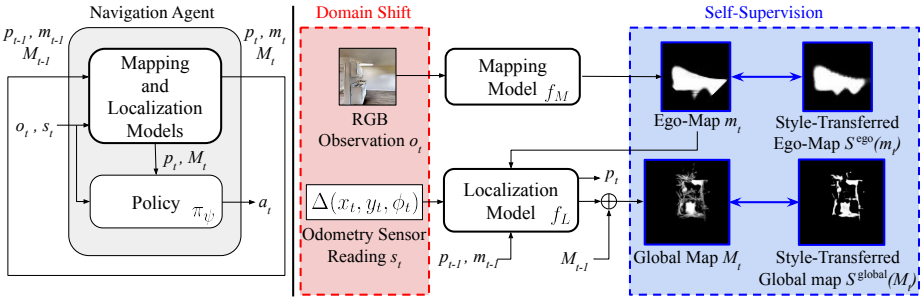

Fig. 2: The structure of map-based navigation agent decouples the mapping and localization models from the policy with the intermediate map memory (*left*). We suggests a domain adaptation method which transfers an agent to a shifted domain with visual and dynamics corruptions (*right*). Given the corrupted sensory inputs, o_t and s_t, our method only fine-tunes the agent's mapping f_M and localization f_L models with the self-supervision loss, encouraging the generated ego-map m_t and global map M_t to be similar to the style-transferred maps.

Given the odometry sensor measurement s_t, the localization model f_L predicts the 2D pose, $p_t = (x, y, \phi)$, where (x, y) indicates the 2D coordinate and ϕ denotes the 1D orientation. The localization model f_L is represented as

$$f_L(p_{t-1}, s_t, \{m_{t-1}, m_t\}) = p_t. \tag{2}$$

Note that the predicted egocentric maps from the previous and current timesteps $\{m_{t-1}, m_t\}$ are given to obtain the pose prediction $\Delta p_t = p_t - p_{t-1}$. Then the egocentric maps from the mapping model are transformed by the estimated poses from the localization model and accumulated as a global map M_t:

$$M_{t-1} \oplus \mathcal{T}_{p_t}(m_t) = M_t, \tag{3}$$

where $\oplus$ represents the fusion of 2D grid maps. The representation of transformation is simplified in Fig. 2. Then the policy module π_ψ plans an action $a_t = (u_x, u_y, u_\phi) \in \mathcal{A}$. The policy is derived from the sensory inputs o_t, s_t, the agent's current pose p_t and the global map M_t,

$$\pi_\psi(o_t, s_t, p_t, M_t) = a_t. \tag{4}$$

Nonetheless, many studies have shown that the policy is most dependent on the map-based memory M_t, which is useful for long-range or complex navigation, rather than the sensory inputs providing partial observability [4, 23, 48].

The agent performs various tasks including mapping and localization using the policy module trained in an ideal environment with ground-truth poses and maps. However, in realistic environments, the overall architecture is disturbed by two main types of corruption: visual and dynamics corruptions. The visual corruptions on RGB observations make it difficult to predict egocentric maps

while the dynamics corruptions on odometry readings and actuation degrade the pose estimation. As the modular pipeline of map-based agents decouples the policy from mapping and localization models with the intermediate spatial memory, MoDA only fine-tunes the perceptual models, namely localization and mapping, to learn the perturbations in measurements and generate a domain-agnostic map for the subsequent policy.

Visual Corruptions Visual corruptions affect the RGB observation, which is the input of the mapping model shown in Fig. 2. As a result, the egocentric perception of the embodied agent suffers from the domain discrepancy. While our adaptation does not assume any particular form of visual corruption, we test our adaptation in scenarios that properly represent the wide varieties of possible visual variations in the real world [12]. The tested scenarios are described in Sec. 4.

Dynamics Corruptions Dynamics corruptions degrade the accuracy of the agent's pose estimated by the localization model, as shown in Fig. 2. The two main sources of dynamics corruptions are the actuation and odometry sensor noises. The actuation noise interrupts an agent from reaching the target location provided by the control commands. After an action, the agent's ground-truth movement $\Delta p_t = \Delta(x_t, y_t, \phi_t)$ is defined as

$$\Delta(x_t, y_t, \phi_t) = (u_x, u_y, u_\phi) + \epsilon_{\text{act}}, \tag{5}$$

where (u_x, u_y, u_ϕ) and ϵ_{act} indicate the intended action control and the actuation noise, respectively. Additionally, the odometry sensor noise ϵ_{sen} disturbs the agent from accurately perceiving its own movement. The final pose reading with the odometry sensor noise ϵ_{sen} is erroneously measured as

$$s_t = \Delta(x_t, y_t, \phi_t) + \epsilon_{\text{sen}}. \tag{6}$$

The actuation and odometry sensor noises are expected in all realistic settings and many studies present realistic models for both types of dynamics corruptions [12, 21, 33, 41].

3.2 Unpaired Map-to-map Translation Network

The main objective of our self-supervised domain adaptation is to translate maps observed from the new, noisy domain such that it recovers the noiseless structure in the absence of paired data. The neural network learns to capture the structural regularities of indoor scenes from the collection of ground-truth maps D_{gt} and translate the learned characteristics into the collection of noisy maps D_{noisy}. The set of ground-truth maps D_{gt} is collected from a noiseless simulator. When the pretrained agent is deployed in a new environment, it collects another set of map data D_{noisy} which is generated amidst visual and dynamics corruptions. We adopt the unpaired image-to-image translation network suggested in [50] on our maps. Since the formulation does not require one-to-one correspondences

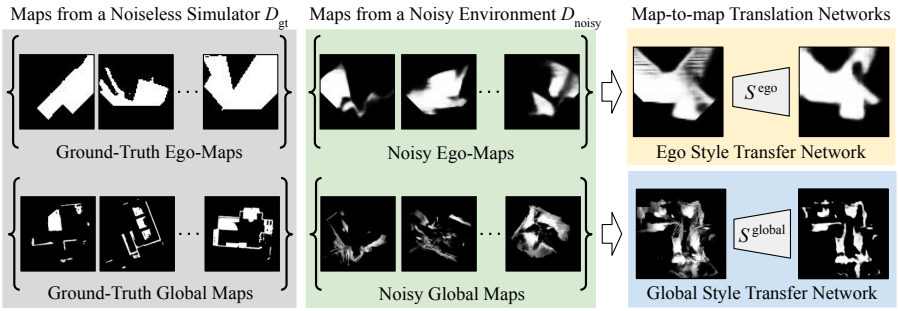

Fig. 3: Given the set of ground-truth maps $D_{\text{gt}}(grey)$ obtained from the noiseless simulator, the set of noisy maps $D_{\text{noisy}}(green)$ is collected by the pretrained agent deployed in a novel environment amidst visual and dynamics corruptions. We then learn two map-to-map translation networks for the egocentric map S^{ego} $(yellow)$ and global map S^{global} $(blue)$ to translate the maps in D_{noisy} into the style of the maps in D_{gt}

between the two sets, our domain adaptation is completely self-supervised and successfully reasons about the stylistic difference between the two collections.

More specifically, we learn two map-to-map translation networks for the egocentric map S^{ego} and the global map S^{global} as shown in Fig. 3. The egocentric maps observe magnified views of the environment, and the style transfer network S^{ego} can help the agent to learn the detailed visual structure of the indoor scene to train the mapping model f_M against visual corruptions. The global style transfer network S^{global}, on the other hand, enforces a globally coherent structure over long-horizon navigation. The self-supervised loss on global maps restricts the localization model f_L from making erroneous predictions due to dynamics corruptions.

The success of the style transfer loss is tightly coupled with the clear structural regularities within the desired set of map data, which are represented as simple gray-scale images. We use different representations that contain more information in respective scenarios; the network learns over the explored area for egocentric maps, whereas the obstacle maps are used for global maps. While the two different representations are easily converted from one to the other given the pose of the agent, the obstacle maps are much more sparse than the explored area. Because the ego-maps cover smaller region, the obstacle maps cannot produce meaningful prior, consisting only about 6% of nonzero values on average within the images. On the other hand, the large overlapping regions in explored areas can be challenging in noisy global maps, whereas obstacle maps exhibit clearly distinguished structure.

3.3 Curriculum Learning for Domain Adaptation

We design a sequential curriculum using the hierarchical structure of map-based models as shown in Fig. 1. Once the agent collects the map data during its

initial deployment up to time T_c, it then learns the two style transfer networks for the new environment. In the next step, our self-supervised adaptation method fine-tunes the perceptual module to robustly handle unknown corruptions. The style transfer loss on the egocentric maps provides signals to adapt the mapping model f_M against visual corruptions. Additionally, we enforce flip consistency on the RGB observations. Next, the global style transfer loss fine-tunes the localization model f_L up to time T_d against dynamics corruptions, in addition to the temporal consistency in the global map. The transferred agent then stably performs various navigation tasks in the new, noisy environment.

Visual Domain Adaptation According to the learning curriculum, we first adapt the mapping model for unknown visual changes in the new environments. The style transfer network for the egocentric map S^{ego} converts the predicted egocentric map m_t into noiseless style. The ego style transfer loss minimizes the discrepancy between the two maps:

$$\mathcal{L}_{st}^{\text{ego}} = \sum_{t=T_c+1}^{T_v} \|m_t - S^{\text{ego}}(m_t)\|_2, \tag{7}$$

with T_v indicating the ending time of visual domain adaptation. In addition, we fine-tune the feature extractor F of the mapping model f_M along with a consistency loss to stabilize the training. The flip consistency loss $\mathcal{L}_{fc}$ assumes that the feature extractor should make consistent predictions over the flipped observations [3, 37]. Specifically, when a horizontally flipped RGB observation, $\text{flip}(o_t)$, and a non-flipped original observation, o_t, are given as inputs, the estimated egocentric maps should be equal but flipped. We, therefore, define our flip consistency loss as

$$\mathcal{L}_{fc} = \sum_{t=T_c+1}^{T_v} \|flip(F(o_t)) - F(flip(o_t))\|_2, \tag{8}$$

Together the visual domain adaptation transfers the mapping model of the pretrained agent with the visual domain loss

$$\mathcal{L}_V = \lambda_{st}^{\text{ego}} \mathcal{L}_{st}^{\text{ego}} + \lambda_{fc} \mathcal{L}_{fc}. \tag{9}$$

The values for hyper-parameters are provided in the supplementary material.

Dynamics Domain Adaptation In the next stage, we adapt the localization model f_L to the dynamics corruptions that are present in the new environment. The agent generates a global map over its trajectory and encodes the predicted pose information onto the map. Thus, by learning the structural priors from the style transfer network, the agent can inversely learn to estimate more accurate pose. Given the estimated and style-transferred global maps, the global style transfer loss, $\mathcal{L}_{st}^{\text{global}}$ is formulated as

$$\mathcal{L}_{st}^{\text{global}} = \sum_{t=T_v+1}^{T_d} \|M_t - S^{\text{global}}(M_t)\|_2. \tag{10}$$

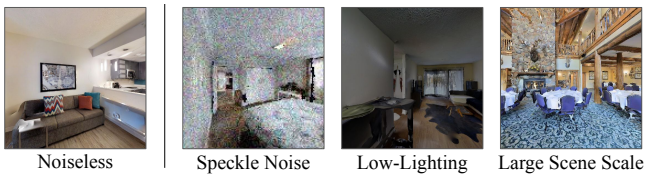

Fig. 4: Visualization of RGB observations with visual corruptions: speckle noise, low-lighting and large scene scale

where T_d denotes the ending time of dynamics domain adaptation. Moreover, we encourage the agent to generate consistent global maps over time. As the pose error accumulates over time, the global map generated in the earlier step encodes more accurate pose information over the global map generated in the later step. The temporal consistency loss $\mathcal{L}_{tc}$ compares the generated global map to the map from the previous time-step, and it is defined as

$$\mathcal{L}_{tc} = \sum_{t=T_v+1}^{T_d} \|M_t - M_{t-1}\|_2. \tag{11}$$

Therefore, the full objective of dynamics domain adaptation becomes

$$\mathcal{L}_{\mathcal{D}} = \lambda_{st}^{\text{global}} \mathcal{L}_{st}^{\text{global}} + \lambda_{tc} \mathcal{L}_{tc}, \tag{12}$$

completing the final stage of suggested learning curriculum.

4 Experiments

We show the validity of MoDA using the navigation agent from Active Neural SLAM [9] which is widely adapted for other navigation models [10,39]. Nonetheless, MoDA is applicable to various navigation agents with map-like memory. We use the Habitat simulator [41]. The pretrained agent is trained on the standard train split [41] of Gibson dataset [49] with ground-truth supervision. We split the unseen scenes of Gibson and Matterport3D [7] for adaptation and evaluation. The scenes for each split are listed in the supplementary material. MoDA is implemented using Pytorch [38] and accelerated with an RTX 2080 GPU.

Visual and Dynamics Corruptions We evaluate the proposed method in three environments where visual and dynamics corruptions are present. Each environment is distinguished by the three visual variations: speckle noise, low-lighting, and scene scale change. Specifically, our experiments transfer the pretrained agent in three types of variation visualized in Fig. 4. First, we apply image quality degradation, which may be caused by the physical condition of the mounted camera. We generate low-quality RGB observations with additive speckle noises. The second type of perturbation is the low-lighting condition to reflect the common light variations in the real world. We show if our agent can be transferred

to low-lighting scenes by adjusting the contrast and brightness of the input RGB image. Lastly, we evaluate if our pretrained agent from Gibson scenes can generalize to different scene scale by transferring the agent to the scenes in Matterport3D. While the Gibson scenes consist of scans collected from offices, the scenes in Matterport3D generally consist of large-scale homes. As the dynamics corruptions are seen in all realistic environments, we add the odometry sensor and actuation noise models to all three scenarios. In our experiment, we use the noise parameters generated from the actual physical deployment of LoCobot [1] in previous work [9, 31] and draw from a Gaussian Mixture Model at each step.

Baselines We extensively compare our agent, referred as "**MoDA**" to various baselines. "No adaptation (**NA**)" reflects the performance degradation of the un-adapted, pretrained model due to the domain gap. "Domain Randomization (**DR**)" adapts the pretrained model with ground-truth supervision in a randomized domain with various combinations of visual and dynamics corruptions. "Policy Adaptation during Deployment (**PAD**)" proposed by Hansen et al. [24] performs visual domain adaptation using an auxiliary task, namely rotation prediction. As the original method mainly targets visual adaptation, we further extend PAD for dynamics corruption by additionally training with our dynamics adaptation method. Lastly, "Global Map Consistency (**GMC**)" from Lee et al. [36] imposes global map consistency loss on round trip trajectories to adapt to dynamics corruptions. Further details of baseline implementation are explained in the supplementary material.

Tasks and Evaluation Metric We report the transferred agent's localization and mapping performance in the new, noisy environment. For fair comparison, we evaluate each adaptation method on an identical set of trajectories obtained from the un-adapted agent in each environment. Following [6, 34], localization is evaluated with the median translation (x, y) and rotation (ϕ) error. Mapping performance is evaluated with the mean squared error (MSE) of the generated occupancy grid maps compared to the ground-truth.

We also demonstrate our adapted agent's performance in downstream navigation tasks. Following Chaplot et al. [9], we report exploration performance using the explored area and explored area ratio after letting the agent to explore for a fixed number of steps. In addition, we report the collision ratio, which is the percentage of collisions from the agent's total steps. We include the collision ratio to distinguish simple random policy, which often results in undesirable collisions coupled with sliding along the walls, and eventually explore large areas. We further evaluate our agent on point-goal navigation(PointNav) as suggested in [2]. Here, we report the success rate and Success weighted by Path Length(SPL) [2].

We investigate MoDA in two settings, generalization and specialization, following [8]. In our main experiment, *generalization* adapts the agent in a set of unseen scenes with unknown noises (Sec. 4.1). The trained agent is then evaluated in a different set of novel scenes but with the same visual and dynamics corrup-

Table 1: Generalization performance in the three new environments with speckle noise, low-lighting, and scene scale change. All three scenes contain dynamic corruptions not present in the original training setup of the pretrained agent

(a) Gibson Scenes Containing Speckle Noise and Dynamics Corruptions

	Pose		Map(MSE)		Exploration			PointNav	
	x, y(m)	$\theta(^\circ)$	ego	global	area	ratio	collision	success	SPL
NA	0.16	15.02	1.11	0.32	28.45	0.82	0.40	0.12	0.10
DR	0.13	8.74	1.14	0.27	29.35	**0.88**	0.43	0.20	0.13
PAD	**0.04**	**1.08**	1.35	0.32	22.83	0.66	0.51	0.08	0.07
GMC	0.06	5.10	1.11	0.28	**29.73**	0.85	**0.35**	0.36	0.29
MoDA	**0.04**	2.61	**1.08**	**0.25**	28.63	0.82	0.36	**0.56**	**0.47**

(b) Gibson Scenes under Low-Lighting and Dynamics Corruptions

	Pose		Map(MSE)		Exploration			PointNav	
	x, y(m)	$\theta(^\circ)$	ego	global	area	ratio	collision	success	SPL
NA	0.18	15.78	0.90	0.32	30.31	0.87	0.34	0.22	0.17
DR	0.14	7.60	0.96	0.26	29.95	0.88	0.41	0.20	0.14
PAD	**0.05**	**2.17**	1.02	0.27	26.88	0.78	0.37	0.22	0.18
GMC	0.06	5.39	0.90	0.28	**32.45**	**0.91**	0.32	0.46	0.37
MoDA	**0.05**	2.87	**0.89**	**0.25**	31.56	**0.91**	**0.26**	**0.56**	**0.45**

(c) MatterPort3D Scenes with Large Scene Scale and Dynamics Corruptions

	Pose		Map(MSE)		Exploration			PointNav	
	x, y(m)	$\theta(^\circ)$	ego	global	area	ratio	collision	success	SPL
NA	0.17	14.42	1.07	0.39	52.16	0.45	0.44	0.04	0.02
DR	0.14	9.17	1.10	0.35	59.22	0.50	0.43	0.08	0.06
PAD	**0.05**	3.12	1.26	0.41	41.66	0.35	0.49	0.02	0.02
GMC	0.10	6.05	1.07	0.40	54.73	0.47	0.36	0.10	0.08
MoDA	**0.05**	**2.34**	**1.02**	**0.31**	**63.68**	**0.54**	**0.28**	**0.22**	**0.18**

tions. For *specialization*, the agent is fine-tuned and evaluated in the same set of unseen scenes, but it starts from a different initial pose for evaluation (Sec. 4.2).

4.1 Task Adaptation to Noisy Environments: Generalization

In generalization, we test whether the adaptation method properly transfers the agent to the existing visual and dynamics corruptions, and avoids over-fitting the agents to the particular scenes. We compare our agent transferred to the three new environments with MoDA, as shown in Table 1. MoDA shows significant performance improvements in all three environments. By effectively fine-tuning pretrained agents using style-transfer in the map domain, MoDA improves localization and mapping, further enhancing the downstream task performance in pointNav and exploration. We additionally report the collision ratio to investigate our agent's stability in exploration. Compared to the baselines, agent trained by MoDA distinctively reports the low number of collision ratio during its exploration steps. While NA and DR agents also show competitive exploration performance, they also exhibit high collision ratios, which indicate instability.

Table 2: Specialization result in the three new environments with speckle noise, low-lighting, and scene scale change. All scenes contain dynamics corruptions

	Gibson Speckle Noise						Gibson Low-Lighting						Matterport3D Large Scene Scale					
	Pose		Map(MSE)		PointNav		Pose		Map(MSE)		PointNav		Pose		Map(MSE)		PointNav	
	x, y(m)	$\theta(^\circ)$	ego	global	success	SPL	x, y(m)	$\theta(^\circ)$	ego	global	success	SPL	x, y(m)	$\theta(^\circ)$	ego	global	success	SPL
NA	0.17	15.25	1.10	0.30	0.18	0.16	0.17	14.44	0.87	0.29	0.18	0.17	0.16	15.04	1.09	0.38	0.04	0.03
DR	0.13	9.17	1.14	0.28	0.20	0.16	0.14	9.11	0.92	0.27	0.22	0.17	0.13	9.54	1.11	0.34	0.02	0.01
PAD	**0.03**	**1.09**	1.33	0.31	0.06	0.05	**0.04**	**2.88**	1.01	0.26	0.32	0.28	0.05	3.21	1.25	0.40	0.06	0.05
GMC	0.06	5.66	1.09	0.28	0.44	0.38	0.07	7.82	0.87	0.28	0.42	0.37	0.09	6.51	1.09	0.38	0.12	0.08
MoDA	0.04	2.54	**1.08**	**0.25**	**0.56**	**0.47**	0.06	3.24	**0.85**	**0.25**	**0.54**	**0.47**	**0.04**	**2.21**	**1.03**	**0.31**	**0.22**	**0.17**

Further, while GMC shows competitive performance against MoDA, it mandates the agent to navigate in round trip trajectories for adaptation. MoDA performs successful adaptation without such constraints, thus more practical than GMC. As a result, our agent's performance across all evaluation metrics confirms the effectiveness of MoDA which successfully adapts to a new, shifted domain with visual and dynamics corruptions.

In Fig. 5, we show the visualization of the estimated pose trajectory and reconstructed global maps generated by the agents observing the same sequence of RGB observations and odometry sensor readings. Our model better aligns with the ground-truth compared to the baselines. MoDA compensates for both visual and dynamics corruptions online without using ground-truth data.

4.2 Task Adaptation to Noisy Environments: Specialization

The specialization setting reflects the practical scenario where the agent is continuously deployed in the same scenes. In Table 2, we compare our agent's performance to baselines in localization, mapping, and PointNav. As in generalization, we evaluate the models in three different environments where the dynamics corruptions and one of each visual corruptions are present.

MoDA successfully adapts in the specialization scenario by showing coherent performance in localization, mapping, and PointNav. In localization, our model outperforms the baselines except for PAD. Although PAD exhibits the lowest pose estimation error in localization metric when evaluated on logged trajectories, it fails to outperform our model in mapping or PointNav. GMC shows the performance enhancement in all metrics over the other baselines, yet underperforms compared to our model. While adapted and deployed in the same scenes, our agent stably adapts to the visual corruption as shown from the ego-map prediction result in all three environments. We also observe that our agent estimates distinctively accurate poses, leading to generating an accurate global map. The performance improvement in mapping and localization, which is tightly coupled to the corruptions added to the observation, aids our agent to generate a more accurate intermediate spatial map. Then, the enhanced domain-agnostic representation further leads our model to outperform the baselines for the PointNav task. Therefore we conclude that MoDA provides a powerful, integrated adaption method for closing the domain gap from visual and dynamics corruptions present in a realistic environment.

Table 3: Ablation study on various loss functions employed in MoDA

	Pose		Map(MSE)		PointNav	
	x, y(m)	$\theta(^\circ)$	ego	global	success	SPL
NA	0.16	15.02	1.11	0.32	0.12	0.10
NA + $\mathcal{L}_{fc}$	0.16	15.01	1.12	0.32	0.20	0.16
NA + $\mathcal{L}_{fc}$ + $\mathcal{L}_{st}^{\text{ego}}$	0.16	14.98	**1.08**	0.31	0.20	0.17
NA + $\mathcal{L}_{fc}$ + $\mathcal{L}_{st}^{\text{ego}}$ + $\mathcal{L}_{tc}$	**0.04**	3.16	**1.08**	0.25	0.42	0.34
NA + $\mathcal{L}_{fc}$ + $\mathcal{L}_{st}^{\text{ego}}$ + $\mathcal{L}_{tc}$ + $\mathcal{L}_{st}^{\text{global}}$	**0.04**	**2.61**	**1.08**	**0.25**	**0.56**	**0.47**

4.3 Ablation study

In this section, we verify the effectiveness of each loss function in visual and dynamics domain adaptation. In Table 3, we report the adaptation result of the agent transferred to the environment with speckle noise visual corruption and dynamics corruptions. The overall experiment setup is the same as generalization. Beginning from the pretrained agent, referred as "NA", we gradually add the four losses mentioned in Sec. 3.3. We first train the pretrained agent only with the flip consistency loss $\mathcal{L}_{fc}$, which improves the agent's performance in localization and PointNav, but not in mapping. However, the model trained with both flip consistency loss $\mathcal{L}_{fc}$ and ego style transfer loss $\mathcal{L}_{st}^{\text{ego}}$ results in the performance enhancement in all evaluation metrics. This ablated model with $\mathcal{L}_{fc}$ and $\mathcal{L}_{st}^{\text{ego}}$ indicates that style transfer loss is more effective in adapting the agent during the visual domain adaptation stage. The adaptation for visual perturbations, which targets at predicting more accurate egocentric maps, also leads to the improvement in the subsequent evaluation tasks, localization, global map prediction and PointNav performance. The visually adapted agent is then trained with the temporal consistency loss $\mathcal{L}_{tc}$. The addition of $\mathcal{L}_{tc}$ effectively transfers the agent for the dynamics corruptions. Nonetheless, our full model, jointly trained with the global style transfer loss $\mathcal{L}_{st}^{\text{global}}$, exhibits the best performance in all metrics compared to all versions of ablated model.

5 Conclusion

In conclusion, we propose MoDA, a self-supervised domain adaptation method which provides an integrated solution to adapt the pretrained embodied agents to visual and dynamics corruptions. By transferring the noisy maps into clean-style maps, the agent can successfully adapt to the new environment with additional assistance with consistency loss. Our evaluation in generalization and specialization proves that MoDA is a powerful and practical domain adaptation method, showing its applicability in the noisy real world in the absence of ground-truth.

Acknowledgements: This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2020R1C1C1008195), Creative-Pioneering Researchers Program through Seoul National University, and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub).

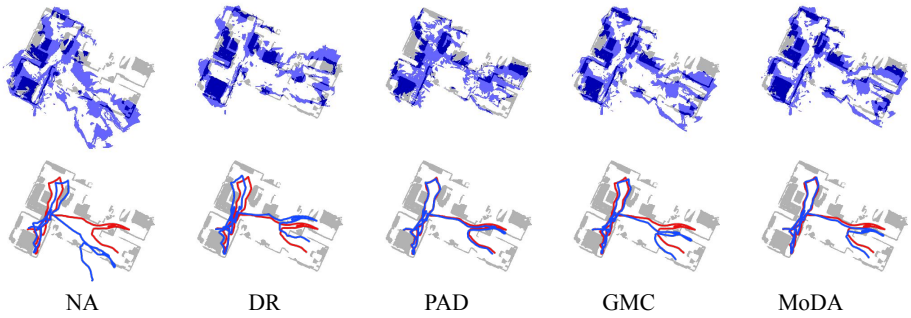

Fig. 5: Qualitative result of mapping (*top*) and localization (*bottom*) obtained from agents observing the identical sequence of RGB observations and odometry sensor readings. The reconstructed maps (*blue*) are aligned on the ground-truth maps (*grey*), and the estimated pose trajectories (*blue line*) are compared to the ground-truth trajectories (*red line*)

References

1. "LoCoBot: An open source low cost robot", http://www.locobot.org/, 2021
2. Anderson, P., Chang, A., Chaplot, D.S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., et al.: On evaluation of embodied navigation agents. arXiv preprint arXiv:1807.06757 (2018)
3. Araslanov, N., Roth, S.: Self-supervised augmentation consistency for adapting semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15384–15394 (June 2021)
4. Bhatti, S., Desmaison, A., Miksik, O., Nardelli, N., Siddharth, N., Torr, P.H.: Playing doom with slam-augmented deep reinforcement learning. arXiv preprint arXiv:1612.00380 (2016)
5. Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., Leonard, J.J.: Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE Transactions on robotics **32**(6), 1309–1332 (2016)
6. Campbell, D., Petersson, L., Kneip, L., Li, H.: Globally-optimal inlier set maximisation for camera pose and correspondence estimation. IEEE transactions on pattern analysis and machine intelligence **42**(2), 328–342 (2018)
7. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. arXiv preprint arXiv:1709.06158 (2017)
8. Chaplot, D.S., Dalal, M., Gupta, S., Malik, J., Salakhutdinov, R.R.: Seal: Self-supervised embodied active learning using exploration and 3d consistency. Advances in Neural Information Processing Systems **34** (2021)
9. Chaplot, D.S., Gandhi, D., Gupta, S., Gupta, A., Salakhutdinov, R.: Learning to explore using active neural slam. arXiv preprint arXiv:2004.05155 (2020)
10. Chaplot, D.S., Gandhi, D.P., Gupta, A., Salakhutdinov, R.R.: Object goal navigation using goal-oriented semantic exploration. Advances in Neural Information Processing Systems **33** (2020)
11. Chaplot, D.S., Salakhutdinov, R., Gupta, A., Gupta, S.: Neural topological slam for visual navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12875–12884 (2020)
12. Chattopadhyay, P., Hoffman, J., Mottaghi, R., Kembhavi, A.: Robustnav: Towards benchmarking robustness in embodied navigation. ArXiv **abs/2106.04531** (2021)
13. Chen, K., de Vicente, J.P., Sepulveda, G., Xia, F., Soto, A., Vázquez, M., Savarese, S.: A behavioral approach to visual navigation with graph localization networks. arXiv preprint arXiv:1903.00445 (2019)
14. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling (2014)
15. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., Batra, D.: Embodied question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–10 (2018)
16. Deng, Z., Narasimhan, K., Russakovsky, O.: Evolving graphical planner: Contextual global planning for vision-and-language navigation. arXiv preprint arXiv:2007.05655 (2020)
17. Du, Y., Gan, C., Isola, P.: Curious representation learning for embodied intelligence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10408–10417 (2021)

18. Durrant-Whyte, H., Bailey, T.: Simultaneous localization and mapping: part i. IEEE robotics & automation magazine **13**(2), 99–110 (2006)
19. Engel, J., Schöps, T., Cremers, D.: Lsd-slam: Large-scale direct monocular slam. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 834–849. Springer International Publishing, Cham (2014)
20. Fan, L., Wang, G., Huang, D.A., Yu, Z., Fei-Fei, L., Zhu, Y., Anandkumar, A.: Secant: Self-expert cloning for zero-shot generalization of visual policies. arXiv preprint arXiv:2106.09678 (2021)
21. Gonçalves, J., Lima, J., Oliveira, H., Costa, P.: Sensor and actuator modeling of a realistic wheeled mobile robot simulator. In: 2008 IEEE International Conference on Emerging Technologies and Factory Automation. pp. 980–985. IEEE (2008)
22. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014), https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf
23. Gupta, S., Davidson, J., Levine, S., Sukthankar, R., Malik, J.: Cognitive mapping and planning for visual navigation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2616–2625 (2017)
24. Hansen, N., Jangir, R., Sun, Y., Alenyà, G., Abbeel, P., Efros, A.A., Pinto, L., Wang, X.: Self-supervised policy adaptation during deployment. arXiv preprint arXiv:2007.04309 (2020)
25. Hansen, N., Wang, X.: Generalization in reinforcement learning by soft data augmentation. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 13611–13617. IEEE (2021)
26. Hart, P., Nilsson, N., Raphael, B.: A formal basis for the heuristic determination of minimum cost paths. IEEE Transactions on Systems Science and Cybernetics **4**(2), 100–107 (1968). https://doi.org/10.1109/tssc.1968.300136, https://doi.org/10.1109/tssc.1968.300136
27. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
28. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018)
29. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017)
30. Kadian, A., Truong, J., Gokaslan, A., Clegg, A., Wijmans, E., Lee, S., Savva, M., Chernova, S., Batra, D.: Are we making real progress in simulated environments? measuring the sim2real gap in embodied visual navigation (2019)
31. Kadian, A., Truong, J., Gokaslan, A., Clegg, A., Wijmans, E., Lee, S., Savva, M., Chernova, S., Batra, D.: Sim2real predictivity: Does evaluation in simulation predict real-world performance? IEEE Robotics and Automation Letters **5**(4), 6670–6677 (2020)
32. Karkus, P., Cai, S., Hsu, D.: Differentiable slam-net: Learning particle slam for visual navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2815–2825 (2021)
33. Khosla, P.K.: Categorization of parameters in the dynamic robot model. IEEE Transactions on Robotics and Automation **5**(3), 261–268 (1989)
34. Kim, J., Choi, C., Jang, H., Kim, Y.M.: Piccolo: Point cloud-centric omnidirectional localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3313–3323 (October 2021)

35. Koenig, S., Likhachev, M.: Dˆ* lite. Aaai/iaai **15** (2002)
36. Lee, E.S., Kim, J., Kim, Y.M.: Self-supervised domain adaptation for visual navigation with global map consistency. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1707–1716 (January 2022)
37. Li, B., Hu, M., Wang, S., Wang, L., Gong, X.: Self-supervised visual-lidar odometry with flip consistency. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3844–3852 (2021)
38. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32**, 8026–8037 (2019)
39. Ramakrishnan, S.K., Al-Halah, Z., Grauman, K.: Occupancy anticipation for efficient exploration and navigation. In: European Conference on Computer Vision. pp. 400–418. Springer (2020)
40. Savinov, N., Dosovitskiy, A., Koltun, V.: Semi-parametric topological memory for navigation. arXiv preprint arXiv:1803.00653 (2018)
41. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
42. Shah, R., Kumar, V.: Rrl: Resnet as representation for reinforcement learning. arXiv preprint arXiv:2107.03380 (2021)
43. Srinivas, A., Laskin, M., Abbeel, P.: Curl: Contrastive unsupervised representations for reinforcement learning. arXiv preprint arXiv:2004.04136 (2020)
44. Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D., Maksymets, O., Gokaslan, A., Vondrus, V., Dharur, S., Meier, F., Galuba, W., Chang, A., Kira, Z., Koltun, V., Malik, J., Savva, M., Batra, D.: Habitat 2.0: Training home assistants to rearrange their habitat. arXiv preprint arXiv:2106.14405 (2021)
45. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 23–30 (2017). https://doi.org/10.1109/IROS.2017.8202133
46. Wang, H., Yu, Y., Yuan, Q.: Application of dijkstra algorithm in robot path-planning. In: 2011 second international conference on mechanic automation and control engineering. pp. 1067–1069. IEEE (2011)
47. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
48. Wani, S., Patel, S., Jain, U., Chang, A.X., Savva, M.: Multion: Benchmarking semantic map memory using multi-object navigation. arXiv preprint arXiv:2012.03912 (2020)
49. Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson env: Real-world perception for embodied agents. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9068–9079 (2018)
50. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on (2017)

51. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 465–476. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/6650-toward-multimodal-image-to-image-translation.pdf