GAIT: Generating Aesthetic Indoor Tours with Deep Reinforcement Learning

Desai Xie
Stony Brook University

Ping Hu
Stony Brook University

Xin Sun
Adobe Research

Sóren Pirk
Adobe Research

Jianming Zhang
Adobe Research

Radomír Měch
Adobe Research

Arie E. Kaufman
Stony Brook University

Abstract

Placing and orienting a camera to compose aesthetically meaningful shots of a scene is not only a key objective in real-world photography and cinematography but also for virtual content creation. The framing of a camera often significantly contributes to the story telling in movies, games, and mixed reality applications. Generating single camera poses or even contiguous trajectories either requires a significant amount of manual labor or requires solving high-dimensional optimization problems, which can be computationally demanding and error-prone. In this paper, we introduce GAIT, a Deep Reinforcement Learning (DRL) agent, that learns to automatically control a camera to generate a sequence of aesthetically meaningful views for synthetic 3D indoor scenes. To generate sequences of frames with high aesthetic value, GAIT relies on a neural aesthetics estimator, which is trained on a crowed-sourced dataset. Additionally, we introduce regularization techniques for diversity and smoothness to generate visually interesting trajectories for a 3D environment, and to constrain agent acceleration in the reward function to generate a smooth sequence of camera frames. We validated our method by comparing it to baseline algorithms, based on a perceptual user study, and through ablation studies. The source code of our method will be released with the final version of our paper.

1. Introduction

Composing a shot by framing a scene with a camera plays an integral part in photography and cinematography. A carefully composed frame does not only provide the information of a scene, but also serves to define the visual style, to instill a desired emotion in the viewer, and to carry forward the story the artist wants to convey [36]. Photographers and movie directors commonly spend a significant amount of time to perfect the camera framing, which – consequently – often leads to extensive cost footprints. While framing scenes in virtual setups, such as games or mixed reality applications, is arguably less involved, defining aesthetically valuable camera poses and trajectories still requires a considerable amount of manual work. An artist has to position key frames in space, define the orientation of the camera, and specify the temporal profile for the interpolation between keyframes. While this provides a high degree of control, for many applications, it would be desirable to frame a scene automatically by computing camera poses.

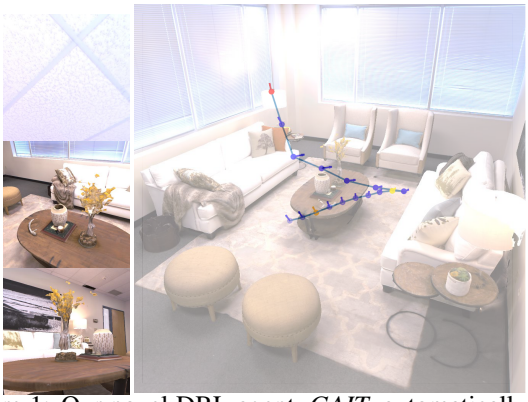


Figure 1: Our novel DRL agent, *GAIT*, automatically generates camera poses so as to obtain aesthetic views of 3D indoor scenes. Left: Views of three generated sequences of camera poses. Right: the 3D indoor environment with three highlighted camera poses. Red, yellow, orange corresponds to the three frames on the left respectively, where the red dot is the initial pose. It maintains high aesthetic views throughout the sequence, while satisfying the initial pose, diversity, smoothness, boundary constraints.

Existing methods for automatically computing single camera poses or camera trajectories either rely on handcrafted methods with heuristics, which commonly do not generalize [37], are limited to specific targets, such as a trip between two cities [19], or specifically focus on moving targets [34]. Recently, it has been recognized that neural aesthetics estimators can serve as more generic solution for

generating camera poses for the purpose of obtaining aesthetically meaningful camera frames [30] – a 2D image is mapped to a score that quantifies the aesthetics. An aesthetics estimator can be used to compare any two images with different contents or styles, which makes them applicable for finding views in lower dimensional spaces such as images and videos [46], or in constrained robotics settings [1]. In a 3D environment, finding a desired camera pose requires searching in a continuous R^6 space (*i.e.*, position R^3 and orientation in R^3) while also considering obstacles or even dynamically moving objects. Computing camera paths in a flexible and versatile manner therefore is a challenging and open problem.

In this paper, we propose a novel method for automatically generating trajectories to aesthetically frame synthetic 3D indoor scenes. We introduce *GAIT*, a framework for training a Deep Reinforcement Learning (DRL) network that learns to move the camera so as to generate trajectories that show the most aesthetic views while also satisfying smoothness constraints. Our method is able to robustly generate diverse trajectories with varying start and end camera poses. Camera poses are optimized with a neural aesthetic metric [46] without any pre-determined targets in the view.

GAIT computes camera transformations – the translation in 3D Euclidean space and the rotation (defined as yaw and pitch) – in a continuous 5D space for each step of the sequence. We define *diversity regularization* to provide control for either generating diverse sets of trajectories with varying start and end poses or to generate more uniform trajectories that always converge to the same final pose. To constrain the agent from taking actions that would create discontinuities in camera pose, we define *smoothness regularization*. Smooth trajectories tend to be more pleasing visually, which is important when the generated trajectories are used for video tours. To obtain a *GAIT* agent, we introduce a flexible framework to leverage existing RL methods for policy training, such as DrQ-v2 [47] and CURL [25].

Based on a number of experiments we show that our method is able to generate trajectories of camera poses that frame scenes in an aesthetically meaningful manner. We show that our method is able to generate camera trajectories for a variety of complex 3D indoor scenes, which can be used to automatically create aesthetic video tours. Moreover, we show that the learned policies are robust against random initial camera poses – independently of the starting pose of the agent, it can converge to the same target pose.

In summary, our contributions are: (1) We propose *GAIT*, the first DRL-based framework for generating sequences of camera poses with constrained globally optimal aesthetics in 3D synthetic indoor scenes; (2) We allow for user control based on diversity regularization and use smoothness regularization to constrain the agent to generate smooth and visually pleasing camera poses; (3) We show that employing image augmentation techniques facilitates learning representation of 3D scene aesthetics from a high-dimensional pixel-space, which is commonly considered challenging for DRL algorithms; (4) We implemented our algorithm to efficiently utilizes multiple GPUs: on a 8-GPU compute node, it can finish training in 3.5 hours; (5) We show that the generated camera poses can be interpolated to generate high-quality video tours of a scene; (6) Finally, we perform an extensive set of experiments and carefully validate our method based on quantitative and qualitative visual evaluations, via comparison with baseline method in a user study, and ablation studies to validate our algorithm design.

2. Related Work

Automatically generating views based on optimized camera control has received a considerable amount of research attention in the past decades. Existing approaches range from scientific visualization [5], surveillance [38, 6, 9] and robot photography [23] to 3D reconstruction [51, 29], virtual cinematography [21] – even focusing on characters [21, 20] – and tracking dynamics objects [14]. The breadths of these approaches is a testament for the importance of automatically generating camera poses, which remains challenging as finding solutions commonly requires solving highly complex solution spaces. Finding optimal views can be accomplished by defining metrics for information measurement [5], heuristics and rules of photography [8], or data-driven aesthetics assessment [9, 46].

Deep aesthetic assessment: Methods for deep aesthetic assessment enable to categorize images based on aesthetic quality [33, 30]. To this end, Lu *et al*. [30] employ learned style attributes, while the approach of Kao *et al*. [22] relies on multi-task convolutional neural networks (CNN). Deng et al. [10] provide a survey of different techniques for aesthetic scoring and report that deep aesthetic assessment methods provide superior performance compared to handcrafted methods. Liu *et al*. [28] evaluate aesthetic quality based on graphs, where nodes represent local information of the different parts of an image. Finally, it has been recognized that defining uniform rules to assess aesthetics of motion is challenging [37]. Our method employs the deep aesthetics model developed by Wei *et al*. [46] because of its robustness and lightweight inference cost. They built a large photo pair dataset including more than 1M comparisons between different views from the same image that enables robust training for estimating aesthetic quality.

Aesthetic view finding: A number of methods address the problem of automatically generating aesthetics and finding views. View finding techniques can be categorized based on various metrics, such as information metrics [5, 41, 12, 38, 19, 39, 21] or aesthetic metrics [7, 8, 44, 50, 46]. Chang and Chen [7] present a method for finding views in a panoramic image, while Cheng *et al*. [8] and Wei *et al*. [46]

aim to select cropped views in a larger input image to compute 2D image compositions. Aesthetics models are also used for view recommendation that can even be computed in real time [44]. Yeh *et al*. [50] incorporate both image and motion aesthetic attributes into their video aesthetic assessment and Ma *et al*. [31] developed an instant photo tool based on the view proposal network of Wei *et al*. [46]. Hong *et al*. [18] introduce key composition maps to encode rules for composition-aware image cropping.

For finding viewpoints in 3D space, Zhu *et al*. [52] trained a robot as a reinforcement learning agent to search a given target view in a room. Fang et al. [11] extended this framework to not only support reinforcement but also imitation learning. Gschwindt *et al*. [15] control drone movement with DRL and Bonatti *et al*. [4] define an aesthetic metric of short video clips based on crowd sourcing data. In AutoPhoto [1], the authors traine a reinforcement learning agent for a ground robot mounted camera to find views according to the aesthetics model from [46]. In their work, however, the agent is only tasked to find aesthetic views in the vicinity of its initial location. Unlike these methods, *GAIT* generates a sequence of camera poses in a 3D synthetic scene and moves the camera in a 5D continuous space. The aesthetics of all frames are optimized globally, while satisfying the diversity and smoothness constraints.

Reinforcement learning: Learning directly on image observations is commonly considered challenging for standard RL algorithms. In this setting, the underlying state information has to be extracted from the high-dimensional space of image observations. The RL loss is not sufficient for both policy learning and this implicit representation learning [48], which often leads to sample inefficiency or learning stagnation. In Data Regularized Q (DrQ) [48], Yarats *et al*. employ regularization on task invariant image augmentation to facilitate the representation learning in RL [48]. Contrastive Unsupervised Representations for Reinforcement Learning (CURL) [25] further extend this augmentation strategy by also adopting contrastive learning on the augmented images.

Actor-critic RL methods tend to perform well in continuous action space domains where the policy needs to be explicitly expressed, which differs from the ϵ-greedy policy in DQN [32]. The actor and the critic models represent the agent's policy $\pi(a_t|s_t)$ and the state-action value function $Q(s_t, a_t)$ respectively. In Deep Deterministic Policy Gradient (DDPG) [27], the actor and the critic are represented as neural networks. The critic is updated according to Q-learning [32], and the actor is updated using its action value from the critic and the chain rule, following DPG [42]. DDPG also integrates techniques including experience replay, target networks, and exploration noise, which make it one of the most widely adopted algorithms for the continuous action domain. Soft Actor Critic (SAC) differs from DDPG in that it has a policy entropy term in addition to the RL objective for a better exploration and multi-modal behavior. The two Visual DRL algorithms we employ, DrQ-v2 [47] and CURL [25], use DDPG [27] and SAC [16] as their base actor-critic RL algorithms respectively.

3. *GAIT* Agent

To generate aesthetic indoor tours we introduce *GAIT*, a framework for training a DRL agent based on existing policy training approaches. Specifically, we show that a *GAIT* agent can be trained with DrQ-v2 [47], which is an efficient visual RL method that leverages task invariant image augmentation to help representation learning from pixels, as well as with CURL [25], which uses contrastive learning with extended data augmentation. To predict camera pose trajectories, we unify both approaches into a single framework and extend by introducing diversity and smoothness regularization terms. Futhermore, we evaluate aesthetics based on images instead of parametric models or heuristics.

In the following sections we describe our framework. The formulation of our aesthetic camera RL framework is discussed in Section 3.1; the design of our reward function based on a neural aesthetics score, the out-of-bound penalty, and constrains for diversity and smoothness is described in Section 3.2; and finally, in Section 3.3 we describe the two DRL algorithms we implement with our framework, DrQv2 and CURL. An overview of the agent is shown in Figure 2.

3.1. Formulation

Markov Decision Process (MDP): We consider the standard Reinforcement learning (RL) setting, formulated by the Markov Decision Processes (MDP) [3]. In MDP, an agent interacts with the environment in discrete time steps. At step t, given state s_t, the agent selects an action a_t according to its policy $\pi(a_t|s_t)$, and receives a scalar reward r_t along with the next state s_{t+1} from the environment. This process repeats until a terminal state s_T is reached.

Objective: We aim at generating view sequences with superior aesthetic quality, smoothness and controlled diversity. Given an arbitrary initial camera pose x_0^P and the corresponding view x_0^I in a 3D indoor scene, we move the camera for $\mathcal{T} = 15$ time steps consecutively to produce an aesthetic sequence of 16 images. Start at time step $t = 1$, our agent transforms the camera pose to optimize the expected return of the following steps

$$\mathbb{E}[R_1] = \mathbb{E}[\sum_{t=1}^{\mathcal{T}} \gamma^{t-1} r_t], \tag{1}$$

which is the sum of discounted reward. The reward r_t for time step t will be detailed in Section 3.2.

Observation: At time step t, the observation of the agent includes the time step x_t^T, camera pose x_t^P, view image

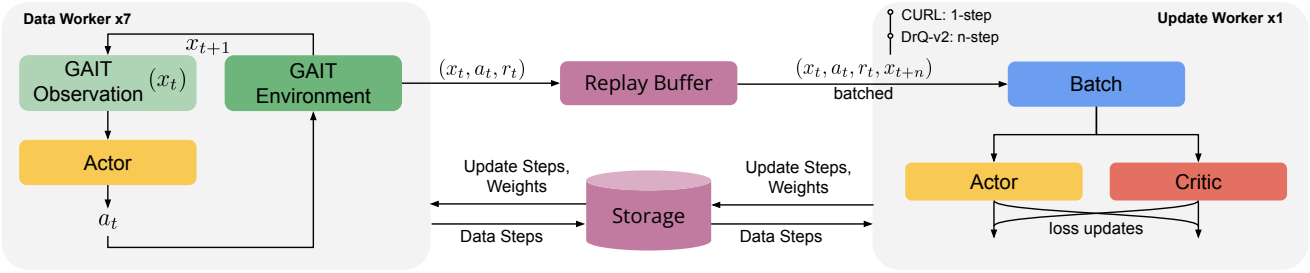


Figure 2: Overview of the GAIT framework in the multi-GPU setting: we describe the shared components for the two actor-critic algorithms, DrQ-v2 and CURL. In the Data loop, the actor interacts with the environment repeatedly and saves each transition to the Replay Buffer. In the Update loop, a batch of transitions is sampled from the Replay Buffer. Then, it is used to update the actor and the critic networks based on the RL loss of DrQ-v2 or CURL. Updated weights of the Actor network and the current Update Step are pushed to the Storage and pulled by Data Workers. Data Workers also pushes the current Data Step to be received by the Update Worker. To maintain an data-update step ratio of $2:1$, Data Workers and the Update Worker look at each others' step number and pauses if it runs faster than the other.

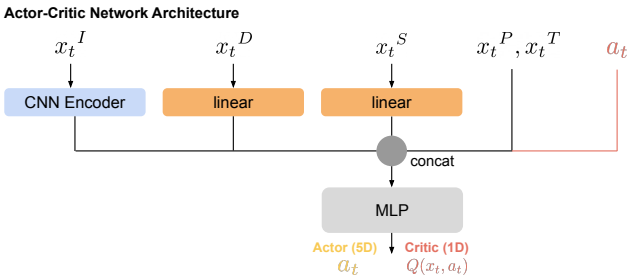


Figure 3: Actor-critic network architecture: while the network architecture for our actor and critic networks are similar, they do not share layers except for the encoder.

x_t^I, diversity regularization observations x_t^D and temporal smoothness regularization observations x_t^S. The time step number x_t^T is normalized to $[0, 1]$. A camera pose is represented as a 5D vector, i.e. the position $\{x, y, z\}$, and the rotation of yaw ψ and pitch θ. The camera pose is limited in an axis-aligned bounding box, within which its position $\{x, y, z\}$ is normalized to $[-1, 1]$. The angle of yaw θ is limited in $[-\pi, \pi]$ and the angle of pitch ψ is limited in $[-\pi/2, \pi/2]$, both of them are also normalized to $[-1, 1]$. A view image is rendered with the camera pose in a resolution of 84×84 with 3 color channels. The evaluations of diversity and smoothness will be detailed in Section 3.2.

Action: An action taken by our agent is also a 5D vector in the same space of camera pose. For each time step, the camera pose is transformed by adding the action vector,

$$x_{t+1}^P = x_t^P \oplus a_t, \tag{2}$$

where $\oplus$ means the result of the addition is round to $[-1, 1]$.

Actor-critic RL: For our *GAIT* framework, we employ two state-of-the-art visual actor-critic DRL algorithms DrQ-v2 [47] and CURL [25]. For both, the actor network represents the policy $\pi(a_t|x_t)$, while the critic network represents the state-action value function $Q(s_t, a_t)$. The critic network is optimized to approximate the expected return given state-action pairs,

$$Q^\pi(s_t, a_t) = \mathbb{E}_{a_{i>t} \sim \pi}[R_t | s_t, a_t], \tag{3}$$

where the actions are selected by the actor network. The actor network is optimized on the expected return over all possible initial states in the environment:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}_{a_i \sim \pi}[R_1]. \tag{4}$$

For DrQv2, we employ Q-learning to update the critic as described in DDPG [27]. We use the action value from the critic and the chain rule following DPG [42], to update the actor. For CURL, the actor and the critic update targets also include the entropy of policy as introduced in SAC [16].

3.2. Reward Function

The reward function consist of several components: first, we introduce an out-of-boundary penalty to prevent the camera being placed outside the indoor scene. For each scene, we set an axis-aligned bounding-box as the domain of the camera position. The position within the domain is normalized as $[-1, 1]^3$. If the camera is placed out-of-boundary, the reward is a negative constant $\mathcal{B}$:

$$r_t = \begin{cases} \mathcal{A}_t \mathcal{S}_t \mathcal{D}_t, & \text{if inside boundary,} \\ \mathcal{B}, & \text{otherwise.} \end{cases} \tag{5}$$

$\mathcal{B}$ is set to be -10 for all experiments in this paper. When the camera is placed inside the boundary, the reward is the product of the evaluations of view aesthetics $\mathcal{A}_t$, temporal smoothness $\mathcal{S}_t$ and diversity regularization $\mathcal{D}_t$.

View Aesthetics: We employ the neural aesthetic model [46] to evaluate the aesthetics of a view. The model performs generic aesthetics assessment based on crowd sourcing data, which also works well 3D indoor scenes. Instead of using the view image x_t^I in the observation with

the resolution of 84×84, we render an image with higher resolution of 240×240 to evaluate the aesthetics $\mathcal{A}_t$ – the higher resolution is required by the neural aesthetic model.

Temporal Smoothness: Obtaining aesthetics assessment for temporal sequences of frames is an open research problem. Therefore, we rely on an aesthetic model [46] for single frames to assess aesthetics. However, if we would only use single frames instead of an entire sequence to quantify the aesthetics, the agent would diverge, which would then lead to degenerated results. Therefore, we introduce the temporal smoothness term $\mathcal{S}_t$ in the reward function (Equation 5) to penalize the agent when it takes an action which is too different from the its preceding time steps.

Adding the smoothness term generates smoother camera trajectories and it forces the camera poses of a sequence to be moderately different from each other. Consequently, the *GAIT* agent will generate a sequence of aesthetic views instead of converging to a single view with high aesthetic value. On the contrary, without the smoothness term, our agent tends to take abrupt actions and also stays at the same aesthetic view until an episode terminates (Section 4.5).

The 5D vector of action is composed of two parts, including a 3D translation and 2D rotation $a_t = \{\tilde{a}_t, \hat{a}_t\}$. The temporal smoothness term compares the action a_t with the actions at the last three time steps $\{a_{t-3}, a_{t-2}, a_{t-1}\}$,

$$\mathcal{S}_t = \frac{1}{2}\left(\min_{i=1}^{3} \mathcal{S}_{t,i} + \frac{1}{3}\sum_{i=1}^{3} \mathcal{S}_{t,i}\right), \mathcal{S}_{t,i} = \tilde{\mathcal{S}}_{t,i} + \hat{\mathcal{S}}_{t,i}, \quad (6)$$

where $S_{t,i}$ is evaluated with a_t and the action taken i time steps earlier a_{t-i}. It consists of two parts, $\tilde{\mathcal{S}}_{t,i}$ and $\hat{\mathcal{S}}_{t,i}$, for translation and rotation respectively. $\tilde{\mathcal{S}}_{t,i}$ is a 1D Gaussian function with the amplitude of 1, the mean of $\tilde{a}_{t-i}$ and the standard deviation of $\max\left(\frac{1}{2}\left|\tilde{a}_{t-i}\right|, 0.1\right)$. $\hat{\mathcal{S}}_{t,i}$ is evaluated similarly except the standard deviation is $\max\left(\left|\hat{a}_{t-i}\right|, 0.1\right)$. Therefore, S_t will penalize the agent when a_t is close to zero or a_t is very different from $\{a_{t-i}\}$. Translation is more penalized than rotation. The Smoothness observation x_t^S is set to the recent actions $\{a_{t-i}\}$, while the Critic additionally observes the distance between its current action and recent actions $\{\|a_t - a_{t-i}\|^2\}$, where $i = 1, 2, 3$. See Supplementary Material for a visualization of the Gaussian smoothness function for Temporal Smoothness.

Diversity Regularization: Without Diversity Regularization, *GAIT* generates sequences in a robust manner – varying initial camera poses lead to trajectories toward the globally most aesthetic target pose. While this may be wanted in many situations, $GAIT$ also allows for generating diverse aesthetic trajectories. We enable this by specifying a diversity regularization term $\mathcal{D}_t$ in the reward function (Equation 5). Specifically, we define up to 4 camera poses along with distances, $\{\bar{x}_j^P, d_j\}_{0 \le j < 4}$ that we refer to as *exclusion poses*. The diversity regularization term penalizes the agent if the distance between its pose and any exclusion poses $\bar{x}_j^P$ is less than the corresponding distance d_j,

$$\mathcal{D}_t = \min_{j=0}^{3}\left(\min\left(\frac{\left\|x_t^P - \bar{x}_j^P\right\|^2}{d_j}, 1\right)\right). \quad (7)$$

To train the *GAIT* agent with exclusion poses we want to randomly define them in the spatial domain. However, as most random camera poses in the scene only have low aesthetics scores (e.g. the camera is placed inside an object or the view may be occluded) selecting exclusion poses randomly would be inefficient. Therefore, we follow another strategy: for each episode, we set the exclusion poses to be the ending camera poses of the last 4 episodes with a random excluding distance in $[0.3, 1.3]$. Because our *GAIT* agent tends to end at camera poses with high aesthetics scores to maximize the reward, selecting the end poses as the exclusion poses converges faster compared to selecting the exclusion poses randomly.

Once trained, a user can define the exclusion camera poses $\{\bar{x}_j^P, d_j\}$ at runtime, which then produces trajectories of camera poses that avoid the provided exclusion poses. This way, our agent can produce sets of diverse, yet aesthetically meaningful, trajectories with diversity regularization, while without diversity regularization it is possible to generate aesthetic trajectories from various different initial conditions that all converge to the same globally most aesthetic camera pose. Diversity Regularization can be disabled by defining out-of-boundary excluding poses.

3.3. Visual Deep Reinforcement Learning

Sample inefficiency or learning stagnation are known to occur in the Visual DRL setting. This is because the supervision from the RL loss is not sufficient to support both policy learning and representation learning [48]. For the Aesthetic Tour problem, learning the complex representation of 3D scene appearance and aesthetics calls for a Visual DRL algorithm. Model-free Visual RL algorithms addresses the representation learning challenge mainly in three ways: image augmentation as task-invariant perturbation, image augmentation for contrastive learning, or self-supervised learning [26].We adopt DrQ-v2 and CURL in *GAIT*, which are the first and the second category respectively.

The DrQ-v2 image augmentation strategy includes randomly shifting the original observation image x_t^I by 4 pixels as well as bilinearly interpolating the shifted image [47]. Both operations act as a task-invariant perturbation to regularize the Q function, which is first introduced in DrQ [48]. CURL relies on the random crop augmentation for contrastive learning, cropping from original image of resolution 100×100 to x_t^I with resolution 84×84. The anchor and the positive images are generated from two different random crops of the same image while the negatives are obtained by cropping other images [25]. Anchors, postives,

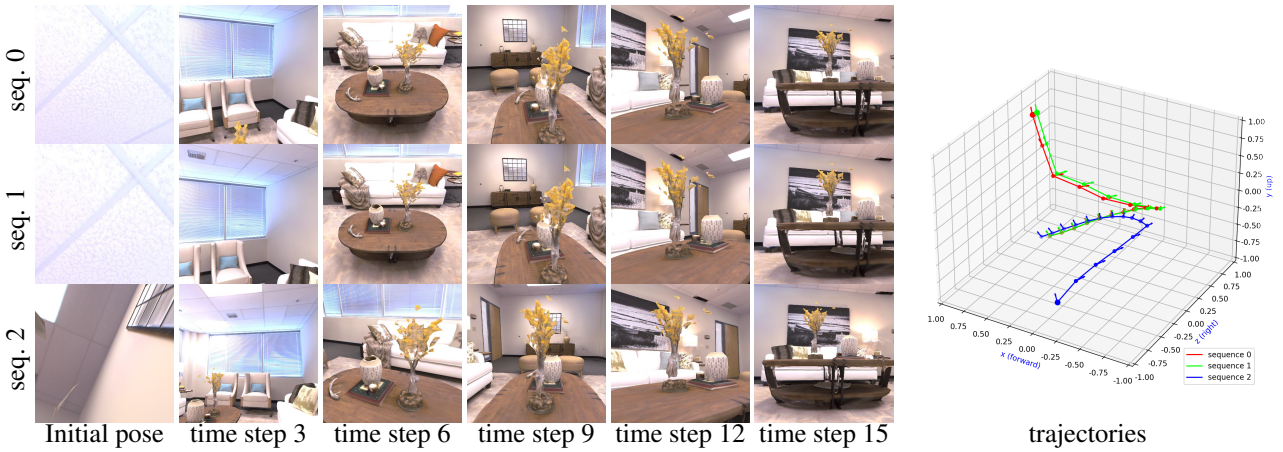


Figure 4: Camera trajectories generated with three different initial camera poses. Left: selected view images including the initial and ending camera poses. Right: visualization of the trajectories whose initial camera poses are located at large dots. All sequences start showing aesthetically pleasing views since time step 3 and ending with similar views.

and negatives are then used to compute the additional InfoNCE loss [35] sto train the encoder.

Other than representation learning, DrQ-v2 and CURL share common components. Both methods are actor-critic, use a replay buffer for off-policy learning, rely on the network architecture introduced in SAC-AE [49], and use clipped double Q learning introduced in TD3 [13]. We modify the shared network architecture by adding inputs as detailed in Section 3.1. DrQ-v2 additionally integrates a linear exploration noise decay schedule, n-step TD, introduced in D4PG [2], and fast implementations of the replay buffer and the image augmentation module. Futhermore, we enables higher throughput for CURL, by adopting DrQ-v2's fast replay buffer and image augmentation.

3.4. Implementation Details

We conducted the training and testing experiments in the Habitat-Sim simulation framework [40, 45] that provides support for rendering and the realistic indoor dataset Replica [43]. We use the Adam [24] adaptive gradient descent as our network optimizer. The discount factor γ is set to 0.99. It is very close to 1 as our objective is the aesthetics of the whole sequence instead of the last view.

Our multiple-GPU implementation boosts the training performance significantly. As shown in Figure 2, on a 8-GPU compute node, we run 7 Data Workers and 1 Update Worker, with each worker running on one GPU. 8-GPU implementation brings us much more speedup compared with 1-GPU, as shown in Figure 8 The actor and critic each has three fully connected layers with hidden sizes 1024.

Our view sequences can contribute to the indoor tour video generation. To generate tour videos for evaluation, 9 intermediate camera poses are interpolated based on the spherical interpolation between the adjacent two frames in each view sequence, producing a 5-second video clip in 30 FPS play mode.

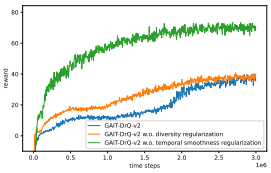
4. Experimental Results

We have conducted an extensive set of quantitative and qualitative experiments to validate *GAIT*. In Section 4.1, we show that *GAIT* stably converges to the same global-near-optimal ending pose from different initial poses. In Section 4.3, we compare results of *GAIT* trained with or without Temporal Smoothness. In Section 4.2, we compare results of *GAIT* training with or without Diversity Regularization. In Section 4.4, we compare GAIT implementations, *GAIT-DrQ-v2*, *GAIT-CURL*, with CMA-ES. In Section 4.5, we identify important components of DrQ-v2 and CURL that significantly contribute to the converged performance. The experiments in this section are conducted with three scenes of Room, Apartment and Office. The images are rendered in Room. We will show more comprehensive experimental results in the supplementary material.

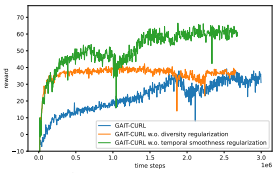
4.1. Aesthetic Camera Sequence Generation

A camera sequence of 16 frames is generated with an arbitrary initial camera pose. Because the initial camera pose is set randomly, its view image is usually with a low aesthetics score. *GAIT* agent transforms the camera pose effectively that the view aesthetics has been substantially improved at the time step 2 or 3, and the rest frames are all with good aesthetics.

GAIT is robust in that the camera trajectory generation is not sensitive to the initial camera pose, as shown in Figure 4. Similar initial camera poses introduce similar trajectories. The global optimal views are always explored even with very different initial camera poses. The ending camera poses are similar if diversity regularization is not applied.



(a) *GAIT-DrQ-v2*



(b) *GAIT-CURL*

Figure 5: The comparisons of training with and without the regularization. The reward is lower with diversity regularization and temporal smoothness regularization because the corresponding terms of $\mathcal{D}_t$ and $\mathcal{S}_t$ is introduced into the reward function in Equation 5. But they converge with similar time steps. This is tested with the scene of Room.

		time (hours) ↓	training steps	throughput (FPS) ↑
GAIT-DrQ-v2	1 GPU	16.33	$1.5M$	25.52
	8 GPUs	3.38	$3M$	246
GAIT-CURL	1 GPU	36.27	$1.5M$	11.49
	8 GPUs	23.03	$1.5M$	17.63

Table 1: The statistics of training. Note that *GAIT-CURL* only need 1.5M steps to converge on 8-GPU.

scene	Room	Apartment	Office
GAIT-DrQ-v2	38.34	20.10	42.10
GAIT-CURL	33.95	32.40	32.67
CMA-ES	32.06	32.29	31.78

Table 2: The comparison of the averaged reward over 9 sequences. The 9 sequences start with 3 different initial positions and generated with 0, 1 and 2 exclusion regions.

4.2. Diversity Regularization

Although *GAIT* is robust to explore global optimal views, users would like to have more control of the generation for different trajectories. The diversity regularization module helps explore the areas with sub-optimal but still plausible aesthetics scores. As shown in Figure 6, the camera poses are effectively moved out of the exclusion regions. The training with diversity regularization is stable and converges with similar time steps, as shown in Figure 5. Although the reward is slightly declined as expected, the quality of sequences are still plausible with the camera exploring nice views outside the exclusions.

4.3. Temporal Smoothness Regularization

Temporal smoothness regularization is critical to produce aesthetic camera sequence because the neural aesthetics model [46] is for individual images. Without the temporal smoothness regularization, the accumulated aesthetics of all images is still good but the camera trajectory is trivial, such as the sequence 1 shown in Figure 7, the camera transform little after the time step 6. The temporal smoothness regularization reduces the reward of training, as shown in Figure 5, which prevents being stuck in optimal poses.

4.4. Comparisons

We use Evaluation Episode Reward over the training process to compare the performance of DrQ-v2 [47], CURL [25], and CMA-ES [17] Evaluation Episode Reward is the average sum of rewards per episode. For every 300 training episodes, a Evaluation of 10 episodes are run with random initial poses x_0^P, random excluding distances d_j, and ending poses of up to 4 previous Evaluation episodes as excluding poses. During Evaluation, the *GAIT* agent acts deterministically, i.e. the noises used for training are turned off. The training processes shown in Figure 8 illustrate similar pattern across different scenes. *GAIT-DrQ-v2* and *GAIT-CURL* take similar time steps to converge, but the latter is $2-8$ times slower according to GPU numbers.

We compare the averaged reward of 9 camera sequences in Table 2. *GAIT-DrQ-v2* and *GAIT-CURL* performs generally better than CMA-ES, except the reward with CMA-ES is better than *GAIT-DrQ-v2* in the scene of Apartment. But both *GAIT-DrQ-v2* and *GAIT-CURL* are better in the user study, which will be shown in Section 4.6 and supplementary material. The view sequences are generated instantly with *GAIT-DrQ-v2* and *GAIT-CURL*. CMA-ES takes about 1 hours to generate a sequence, which is not scalable nor generalizable to interactive application.

4.5. Ablation Study

GAIT can be trained with DrQ-v2 [47] or CURL [25]. DrQ-v2 is improved upon DDPG [27], the linear decayed exploration noise and image augmentation are critical for the learning. As shown in Figure 9 (a), the return will be significantly lower in training without image augmentation or linear decayed exploration noise. CURL is based on SAC [16] and use image augmentation for contrastive learning. As shown in Figure 9 (b), *GAIT-CURL* can hardly converge without image augmentation.

4.6. User Study

To further evaluate the aesthetics in the generated tours, we conduct a user study. We recruited 10 participants (2 females and 8 males, with the age range of 25-35) to rate the video smoothness, video aesthetics, and subjective general preference, of the tours generated using three methods (*i.e.*, CMA-ES, *GAIT-DrQ-v2* and *GAIT-CURL*). Specifically, we prepared the video clips in three indoor scenes using the video generation described in Section 3.4.

We asked for participants' feedback via 2-alternative forced choice. For each initial pose with the same exclusion level, every participant watch two video pairs, {*GAIT-DrQ-v2 vs.* CMA-ES}, and {*GAIT-DrQ-v2 vs. GAIT-CURL*}. In each pair, participants chose the preferred video clip based on (1) video smoothness and (2) video content aesthetics. After comparing the video clips rooted from the same initial

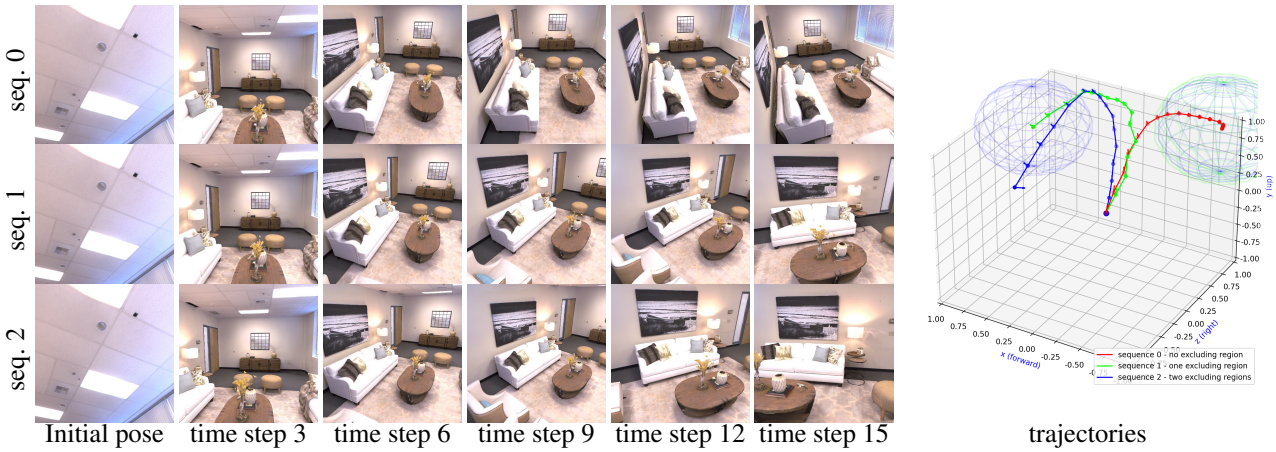
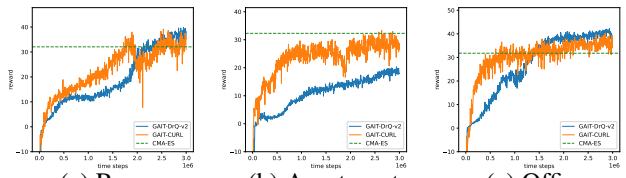


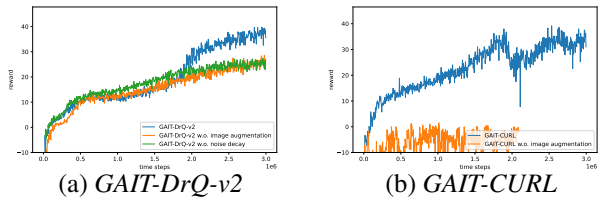
Figure 6: Camera trajectories generated with different diversity regularization. Three trajectories from the same initial camera pose diverge according to the exclusion regions and ending with very different views.

Figure 7: Comparison of temporal smoothness regularization. Without the smoothness, the camera is transformed to the aesthetically optimal pose in a few time steps and stuck there until the end.



(a) Room (b) Apartment (c) Office

Figure 8: The comparisons between the training our method and CMA-ES. Our method is trained with DrQ-v2 and CURL. CMA-ES does not need to train. The dashed line is only for reference. It is achieved by averaging the reward of 9 trajectories it generates with up to 2 exclusion regions.



(a) *GAIT-DrQ-v2* (b) *GAIT-CURL*

Figure 9: The ablation experiments of the training algorithms. The image augmentation is critical for both *GAIT-DrQ-v2* and *GAIT-CURL*. The linear decayed exploration noise significantly helps the training with *GAIT-DrQ-v2*. This is tested with the scene of Room.

pose, participants are also asked to rate "which method provides higher video content diversity" and "which method is the overall preferred aesthetic video generation method".

When asked about "overall which method provides better aesthetic videos" in {*GAIT-DrQ-v2 vs.* CMA-ES} in the three scenes, 93.3%, 96.7%, and 66.7% users favor *GAIT-DrQ-v2*. With the same question in {*GAIT-DrQ-v2 vs. GAIT-CURL*}, 60.0%, 53.3%, and 43.3% users favor *GAIT-DrQ-v2*. We have learned from the users' feedback that majority participants favor the video clips generated in *GAIT-DrQ-v2* over CMA-ES but *GAIT-DrQ-v2* and *GAIT-CURL* generate the view sequences with different capabilities. *GAIT-DrQ-v2* is better at video aesthetics while *GAIT-CURL* performs better in video smoothness. Please refer to our supplementary materials for the collected data and the comprehensive data analysis.

5. Conclusions

We propose *GAIT*, a novel DRL agent to generate aesthetically meaningful sequences of camera poses in 3D indoor scenes. Started with an arbitrary initial camera pose without any pre-determined targets, a view sequence is transformed sequentially by the $GAIT$ agent according to the neural aesthetics model. Trajectories' diversity regular-

ization term is user-controlled to generate diverse, yet aesthetically meaningful, trajectories. Temporal smoothness regularization is introduced to avoid discontinuities camera poses for more visually pleasing trajectories. Our framework enables training *GAIT* agents with existing RL methods, including Drq-v2 and CURL. We evaluate our method extensively. The conducted user study indicates *GAIT* is robust for high quality view sequences and it is easy to control for diverse outputs.

References

[1] Hadi AlZayer, Hubert Lin, and Kavita Bala. Autophoto: Aesthetic photo capture using reinforcement learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 944–951, 2021. 2, 3

[2] Gabriel Barth-Maron, Matthew W Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva Tb, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributed distributional deterministic policy gradients. *arXiv preprint arXiv:1804.08617*, 2018. 6

[3] Richard Bellman. A markovian decision process. *Journal of Mathematics and Mechanics*, 6(5):679–684, 1957. 3

[4] Rogerio Bonatti, Arthur Bucker, Sebastian Scherer, Mustafa Mukadam, and Jessica Hodgins. Batteries, camera, action! learning a semantic control space for expressive robot cinematography. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 7302–7308, 2021. 3

[5] Udeepta D Bordoloi and H-W Shen. View selection for volume rendering. In *IEEE Visualization*, pages 487–494, 2005. 2

[6] Jason Campbell and Padmanabhan Pillai. Leveraging limited autonomous mobility to frame attractive group photos. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3396–3401, 2005. 2

[7] Yuan-Yang Chang and Hwann-Tzong Chen. Finding good composition in panoramic scenes. In *IEEE 12th International Conference on Computer Vision*, pages 2225–2231, 2009. 2

[8] Bin Cheng, Bingbing Ni, Shuicheng Yan, and Qi Tian. Learning to photograph. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 291–300, 2010. 2

[9] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, pages 288–301. Springer, 2006. 2

[10] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34:80–106, 2017. 2

[11] Qiang Fang, Xin Xu, Xitong Wang, and Yujun Zeng. Target-driven visual navigation in indoor scenes using reinforcement learning and imitation learning. *CAAI Transactions on Intelligence Technology*, 7(2):167–176, 2022. 3

[12] Miquel Feixas, Mateu Sbert, and Francisco González. A unified information-theoretic framework for viewpoint selection and mesh saliency. *ACM Transactions on Applied Perception (TAP)*, 6(1):1–23, 2009. 2

[13] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018. 6

[14] Quentin Galvane, Marc Christie, Rémi Ronfard, Chen-Kim Lim, and Marie-Paule Cani. Steering behaviors for autonomous cameras. In *Proceedings of Motion on Games*, MIG '13, page 93–102, New York, NY, USA, 2013. Association for Computing Machinery. 2

[15] Mirko Gschwindt, Efe Camci, Rogerio Bonatti, Wenshan Wang, Erdal Kayacan, and Sebastian Scherer. Can a robot become a movie director? learning artistic principles for aerial cinematography. *arXiv preprint arXiv:1904.02579*, 2019. 3

[16] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018. 3, 4, 7

[17] Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001. 7

[18] Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. Composing photos like a photographer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7057–7066, 2021. 3

[19] Hui Huang, Dani Lischinski, Zhuming Hao, Minglun Gong, Marc Christie, and Daniel Cohen-Or. Trip synopsis: 60km in 60sec. In *Computer Graphics Forum*, volume 35, pages 107–116. Wiley Online Library, 2016. 1, 2

[20] Hongda Jiang, Marc Christie, Xi Wang, Libin Liu, Bin Wang, and Baoquan Chen. Camera keyframing with style and control. *ACM Transactions on Graphics (TOG)*, 40(6):1–13, 2021. 2

[21] Hongda Jiang, Bin Wang, Xi Wang, Marc Christie, and Baoquan Chen. Example-driven virtual cinematography by learning camera behaviors. *ACM Transactions on Graphics (TOG)*, 39(4):45–1, 2020. 2

[22] Yueying Kao, Ran He, and Kaiqi Huang. Deep aesthetic quality assessment with semantic information. *IEEE Transactions on Image Processing*, 26(3):1482–1495, 2017. 2

[23] Myung-Jin Kim, Tae-Hoon Song, Seung-Hun Jin, Soon-Mook Jung, Gi-Hoon Go, Key-Ho Kwon, and Jae-Wook Jeon. Automatically available photographer robot for controlling composition and taking pictures. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 6010–6015, 2010. 2

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[25] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pages 5639–5650. PMLR, 2020. 2, 3, 4, 5, 7

[26] Xiang Li, Jinghuan Shang, Srijan Das, and Michael S Ryoo. Does self-supervised learning really improve reinforcement

learning from pixels? In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 5

[27] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015. 3, 4, 7

[28] Dong Liu, Rohit Puri, Nagendra Kamath, and Subhabrata Bhattacharya. Composition-aware image aesthetics assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3569–3578, 2020. 2

[29] Yilin Liu, Ruiqi Cui, Ke Xie, Minglun Gong, and Hui Huang. Aerial path planning for online real-time exploration and offline high-quality reconstruction of large-scale urban scenes. *ACM Transactions on Graphics (TOG)*, 40(6):1–16, 2021. 2

[30] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 457–466, 2014. 2

[31] Shuai Ma, Zijun Wei, Feng Tian, Xiangmin Fan, Jianming Zhang, Xiaohui Shen, Zhe Lin, Jin Huang, Radomír Měch, Dimitris Samaras, et al. Smarteye: Assisting instant photo taking via integrating user preference with deep view proposal network. In *Proceedings of CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019. 3

[32] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, Feb 2015. 3

[33] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 2408–2415, 2012. 2

[34] Tobias Nägeli, Lukas Meier, Alexander Domahidi, Javier Alonso-Mora, and Otmar Hilliges. Real-time planning for automated multi-view drone cinematography. *ACM Transactions on Graphics (TOG)*, 36(4):1–10, 2017. 1

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6

[36] Des O'Rawe. Towards a poetics of the cinematographic frame. *Journal of Aesthetics & Culture*, 3(1):5378, 2011. 1

[37] Madhura V Phatak, Manasi S Patwardhan, and Meenakshi S Arya. Deep learning for motion based video aesthetics. In *2019 IEEE Bombay Section Signature Conference (IBSSC)*, pages 1–6, 2019. 1, 2

[38] Faisal Z Qureshi and Demetri Terzopoulos. Surveillance camera scheduling: A virtual vision approach. *Multimedia Systems*, 12(3):269–283, 2006. 2

[39] Mike Roberts, Debadeepta Dey, Anh Truong, Sudipta Sinha, Shital Shah, Ashish Kapoor, Pat Hanrahan, and Neel Joshi. Submodular trajectory optimization for aerial 3d scanning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5324–5333, 2017. 2

[40] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 6

[41] Adrian Secord, Jingwan Lu, Adam Finkelstein, Manish Singh, and Andrew Nealen. Perceptual models of viewpoint preference. *ACM Transactions on Graphics (TOG)*, 30(5):1–12, 2011. 2

[42] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. Pmlr, 2014. 3, 4

[43] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 6

[44] Hsiao-Hang Su, Tse-Wei Chen, Chieh-Chi Kao, Winston H Hsu, and Shao-Yi Chien. Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features. *IEEE Transactions on Multimedia*, 14(3):833–843, 2012. 2, 3

[45] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 6

[46] Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomir Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5437–5446, 2018. 2, 3, 4, 5, 7

[47] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. *arXiv preprint arXiv:2107.09645*, 2021. 2, 3, 4, 5, 7

[48] Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations*, 2020. 3, 5

[49] Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. Improving sample efficiency in model-free reinforcement learning from images, 2019. 6

[50] Hsin-Ho Yeh, Chun-Yu Yang, Ming-Sui Lee, and Chu-Song Chen. Video aesthetic quality assessment by temporal integration of photo-and motion-based features. *IEEE transactions on multimedia*, 15(8):1944–1957, 2013. 2, 3

[51] Han Zhang, Yucong Yao, Ke Xie, Chi-Wing Fu, Hao Zhang, and Hui Huang. Continuous aerial path planning for 3d urban scene reconstruction. *ACM Transactions on Graphics (TOG)*, 40(6):1–15, 2021. 2

[52] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3357–3364, 2017. 3