arXiv:2109.09923v1 [cs.CV] 21 Sep 2021

AutoPhoto: Aesthetic Photo Capture using Reinforcement Learning

Hadi AlZayer[1], Hubert Lin[1], and Kavita Bala[1]

***Abstract*— The process of capturing a well-composed photo is difficult and it takes years of experience to master. We propose a novel pipeline for an autonomous agent to automatically capture an aesthetic photograph by navigating within a local region in a scene. Instead of classical optimization over heuristics such as the rule-of-thirds, we adopt a data-driven aesthetics estimator to assess photo quality. A reinforcement learning framework is used to optimize the model with respect to the learned aesthetics metric. We train our model in simulation with indoor scenes, and we demonstrate that our system can capture aesthetic photos in both simulation and real world environments on a ground robot. To our knowledge, this is the first system that can automatically explore an environment to capture an aesthetic photo with respect to a learned aesthetic estimator. *Source code is at* https://github.com/HadiZayer/AutoPhoto**

I. Introduction

Cameras are now widely accessible to most people, but taking a well-composed photo is a difficult task that requires significant practice and experience. With advances in autonomous agents, there is an increasing interest in leveraging drones or robots to reduce human effort in various domains. For example, automatic camera planning can be used to capture sports events [1], and drones can be used to create cinematographic videos [2]–[4]. Autonomous agents are also well-suited for use in remote, dangerous, or otherwise difficult-to-access locations (like caves, forests, or, in an extreme case, other planets like Mars). In real estate, marketing properties requires carefully composed photographs that showcase indoor architecture and layouts. The process of cataloguing different properties is time intensive for a human agent, and physical access to properties may be limited (due to, for example, the recent COVID-19 pandemic). Finally, an autonomous photography system can also be used to guide novice photographers towards better composed photos.

Our goal is to build an autonomous system that can capture aesthetically pleasing photographs. Relying on heuristics is one way to compose aesthetic photographs. For example, the rule of thirds is a heuristic in which important objects of interest are aligned with imaginary lines that divide an image into thirds along horizontal and/or vertical axes. However, heuristics do not fully capture human aesthetics preferences. Indeed, recent work [5]–[10] has focused on learning subjective preferences for aesthetics directly from humans. These aesthetics models can implicitly understand what makes well-composed photographs beautiful, but this understanding cannot be easily decomposed into explicit rules like the rule of thirds.

[1]All of the authors are with the Department of Computer Science, Cornell University, Ithaca NY 14850, USA {ha366,hl2247,kb97}@cornell.edu.

Fig. 1: **Photos Captured By AutoPhoto.** Left: The AutoPhoto system deployed on Clearpath Jackal robot. Right: Photos autonomously captured by AutoPhoto. See Fig. 4 for comparisons against initial environment views.

In this work, we present AutoPhoto, a system that sequentially takes actions to explore an environment, with the end goal of capturing an aesthetic photograph. Two important branches of work in autonomous aesthetic composition are (a) image cropping [5], [6] and (b) drone cinematography [2], [11]. Image cropping is a limited form of "pseudo-photography" with a fixed viewpoint and known environment (i.e., the original uncropped image is the environment from which the cropped photo should be captured). In this work, we are interested in a more general setting in which the photographer must explore an unknown environment via different viewpoints. Although autonomous cinematography is also concerned with varying viewpoints in unknown environments, it is constrained by tradeoffs between smooth motion planning, temporal constraints, and the final aesthetics of the film. Further, cinematography is not strictly concerned with aesthetic view capture; instead, it is focused on capturing events or telling a story [12], [13]. As a result, some existing work simplify aesthetic estimation by leveraging heuristics like shot templates or optimizing rule of thirds with respect to actors in a scene [2], [14]–[16]. In this paper, we are interested in optimizing image aesthetics with respect to data-driven aesthetics models to better capture human preferences.

A challenge with photo viewpoint optimization with respect to a learned aesthetic function is that it is difficult to formulate analytically (contrast this with the rule of thirds which can be directly optimized given an object of interest, e.g. [14], [16]). As such, we use reinforcement learning to learn a controller that can navigate to views that are aesthetic with respect to a learned aesthetic model. A second challenge lies in the aesthetic function itself. Existing work

in learning image aesthetics has focused primarily on its application in cropping [5], [6] where models are trained to compare crops from within the same image. However, the task of photo capture requires views from different parts of an environment to be compared, and the aesthetics function should be robust to variations that may naturally arise such as minimal camera translation, and a change in camera exposure. We propose an aesthetics model which is better suited for this task than existing cropping-based aesthetics models. Given this aesthetics model, we demonstrate that we can learn to navigate an environment to capture aesthetic photos. Our system is trained in simulation with realistic indoors reconstructed scenes using the Gibson dataset [17] with AI Habitat [18]. Our experiments demonstrate generalization to unseen scenes across simulation and real life.

This paper is organized as follows. First, we review related work in Section II. We describe our problem formulation in Section III. Then, we describe our reinforcement learning pipeline in Section IV. We propose an improved aesthetics estimator in Section V suited for the task of photo capture. We cover implementation details in Section VI. Finally, we present the results for our AutoPhoto system in Section VII, with quantitative evaluations in simulation and real life, including evaluation against human preferences.

Our contributions are:

- A novel reinforcement learning pipeline for a generalized photo capture problem which includes (a) an unknown environment, (b) photographer movement and rotation, and (c) optimization against a learned aesthetics estimator that models human preferences better than heuristics.
- An aesthetics model that is consistent with human preferences across diverse viewpoints, and is robust to variances encountered while taking photographs.
- Experimental validation that demonstrates successful navigation in unseen environments across both simulation and real life to capture aesthetically pleasing photos. We deploy the system autonomously on a Clearpath Jackal UGV.

II. Related Work

a) Image Aesthetics: Understanding human judgements of image aesthetics has a long history in psychology and neuroscience, and photographers follow well-known rules to capture aesthetic images [19]. For automatic view selection, such rules can be captured by heuristics like the rule of thirds and template matching [2], [14]–[16]. Computer vision models have also been developed to estimate the aesthetics of images [19]. Modern aesthetics models are trained on large numbers of human judgements, and can generalize better to scenes where no clear heuristics can be applied. Recent work learns to rank the aesthetics of pairs of images [5], [6]. In [6], the model is trained to rank images against random crops of the same image. The assumption is that the composition of professional images are well-balanced while a random crop is less balanced. Alternatively, learning directly from human preferences is possible given datasets of aesthetic score judgements of crops or images [5], [9], [10]. Aesthetic models can also learn personal preferences, where the scores can vary from user to user [7], [8].

b) Automatic Photo Composition: Existing work on automatic photo composition primarily focuses on image cropping, where the composition of photos are improved post-capture. Both common heuristics like the rule of thirds and visual balance [16] and learned aesthetic scores [5], [6] are used for automatic image cropping.

To be able to use the rule of thirds and visual balance for image composition, one needs to specify an object of interest. [16] extracts salient objects from the image, then automatically composes the photo by using the saliency mask to compute scores for the rule of thirds, visual balance, and diagonal dominance. Creatism [20] mimics the pipeline of a professional landscape photographer by cropping and post-processing panoramas from Google Street View.

On the other hand, a learned aesthetics model is able to capture more nuanced composition rules that are not captured by explicit heuristics. To compose photos using a learned aesthetic model, [6] uses a sliding window with various aspect ratios to select the crop with the highest aesthetic score. Since a sliding window approach is computationally expensive, [21] uses a reinforcement learning model to sequentially adjust the crop window.

c) Drone Cinematography: Autonomous cinematography and camera planning have received increasing attention in recent years [1]–[4], [22]. One way to compose videos is to optimize paths between key frames. These key frames can be predefined by a user [23], [24] or sampled intelligently from a set of template shot types [2], [22]. [25] automatically records videos of two subjects by matching the view with predefined templates for scenes with two actors. Instead of defining key frames, imitation learning [26] can be used to learn camera trajectories directly from films. [4], [11] apply imitation learning to learn from human-created films while [3] focuses on learning from examples for computer-animated cinematography. Recent work has also explored semantic control over the emotions evoked by clips by learning from annotated video clips [27].

A key challenge in cinematography lies in temporal consistency – the aesthetics of the final video depends on frames captured in real-time. This challenge limits the ease of integrating learned aesthetics models into drone cinematography. As such, a common property of the many autonomous cinematography works is that the captured video focuses on tracking human subjects [1], [2], [4], [11], [22], [25]. Commercial drones (such as Skydio) also focus on autonomous cinematography for a single user by filming them during activities like skiing or mountain biking. By focusing on human subjects, heuristics for framing human actors can be readily applied instead of optimizing for aesthetics in a general setting.

III. Problem Setup and Pipeline

Our objective is to autonomously explore a local region within an environment to capture an aesthetic photo.

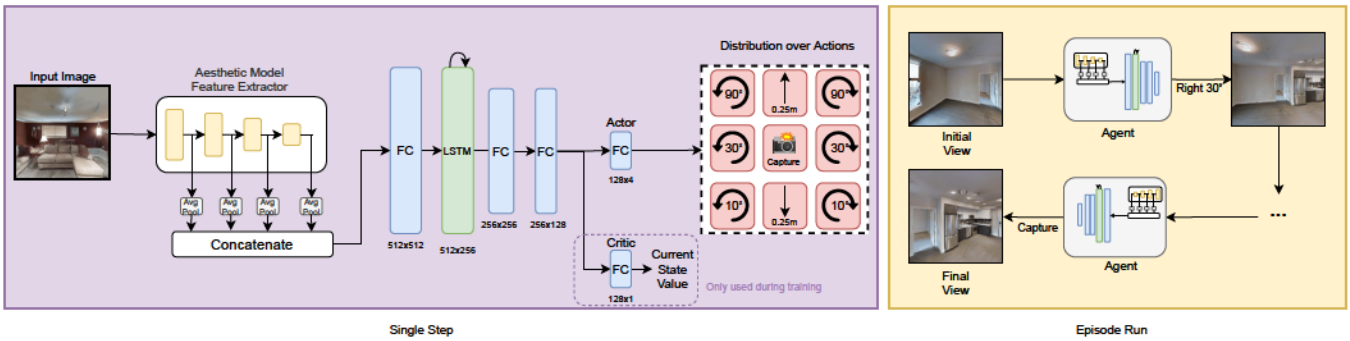

Fig. 2: **Illustration of the Pipeline and Runtime Execution.** Left: AutoPhoto is composed of an aesthetics model that extracts features from the current view, a common MLP+LSTM backbone that processes these features, and two separate layers that parameterize the actor and critic. The actor selects an action to take and the critic estimates the current state value. We iteratively run multiple episodes to sample action and state value pairs to optimize the model parameters. Right: During inference, the model takes a sequence of movement actions until the agent (actor) selects the capture action.

To take a photo, a photographer assesses a view through the camera viewport, and then iteratively adjusts the composition through movement and rotation of the camera. When a balanced composition is achieved, the photographer captures the photo. Our goal is to create an agent that can achieve this functionality. This process can be described as a series of actions: (a) observe the current view, (b) move the camera if the view can be improved and repeat from (a), or (c) capture the photo if the view is sufficiently well-composed and terminate. To judge the quality of the image, we assume that we have oracle access to an aesthetic value function ϕ that maps an image to its aesthetic score. Our objective is to optimize the aesthetic score of the final captured image, so that the aesthetic score of the final image achieves some threshold. To generalize to a variety of scenes, we introduce an aesthetic score threshold that is scene dependent and invariant to the scene size. This threshold is based on the estimate of the mean of aesthetic scores and score variations in the local region to make the threshold independent from the scene size. We discuss how an aesthetics model can be learned (Section V), and how it can be used in practice to set the threshold (Section VI-B).

We formulate the problem as a partially observable Markov decision process (POMDP) – the environment is not fully observable as the agent only has access to partial views of the environment through the camera viewport. The actions are camera movements and photo capture. In Section IV, we discuss this formulation in detail.

In Fig. 2 we demonstrate our RL-based pipeline. On the left we show the architecture details to perform a single step, and on the right we show the action sequence taken by the agent at runtime. To train our model, we run multiple episodes and compute the reward function after every step to sample state-action reward pairs, then we train the Actor based on the current Critic state value estimates, and train the Critic using the collected state-action reward pairs.

IV. RL for Aesthetic Image Capture

We describe the objective, state space, action space, and reward functions such that maximizing the objective corresponds to capturing an aesthetically pleasing photo. We will detail our architecture and implementation in Section VI.

The objective is to find a policy that maximizes the expected sum of discounted rewards. Specifically, denote π to be a policy that maps to a distribution over actions given a state s_t, and let ρ^π be the probability distribution over possible reward realizations $R = r_0, r_1, ..., r_{T-1}$ based on the state-action trajectories given π. Formally, the objective is to find an optimal policy π^* that satisfies

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \underset{R\sim\rho^\pi}{\mathbb{E}}[\sum_{t=0}^{T-1} \gamma^t r_t] \tag{1}$$

where γ is a discount factor in $[0, 1]$ that reduces the weight of rewards far in the future.

a) State and Action Space: We set the state to be the current camera view and an LSTM memory cell. Memory enables RL-based solutions for a POMDP [28] by allowing the agent to utilize information from its exploration history to make decisions. A memoryless agent makes decisions based only on the current view, which can lead to sub-optimal decision making as it might re-explore regions with low aesthetics, and potentially lead to a non-terminating loop.

Our action space consists of the following actions: forward and backward for 0.25m, and turning right and left by 10°, 30°, and 90°, and finally the CAPTURE action to take the photo and terminate the episode. A composition of these actions allows the agent to navigate to any position and orientation in the search space, modulo the regions which lie between the discretizations of the actions defined above. The values of 0.25m translation and 30° rotations are the default in [18]. We introduced more turning angles to allow better control between fine turns to adjust the view composition and large turns to quickly explore the scene.

b) Objective and Rewards: The objective is to capture a final view s_T that has a high aesthetic score according to an aesthetic estimator ϕ (i.e., $\phi(s_T) > \tau_{aes}$). τ_{aes} is set according to the aesthetics of the local region (see Section

VI-B). We set the reward for the CAPTURE action to be:

$$r(s_T, \text{CAPTURE}) = \begin{cases} +1 & \phi(s_T) > \tau_{aes} \\ -1 & \text{otherwise} \end{cases} \quad (2)$$

However, a final reward is not sufficient to train the RL agent on its own since it is sparse, and does not take into consideration the number of steps taken to capture the photo. It is important to learn an efficient policy which takes as few steps as possible. If efficiency is not of concern, one could run an exhaustive grid search to find a view that maximizes the aesthetic score, but this is impractical in general. For movement actions where $a \neq$ CAPTURE, we define the step reward as:

$$r(s_t, a) = \phi(s_{t+1}) - \phi(s_t) + 0.1\Gamma(\zeta) - \beta t \quad (3)$$

$(\phi(s_{t+1}) - \phi(s_t))$ is the score difference between the current and next view to encourage the agent to move towards regions with increasing aesthetic score. $\Gamma(\zeta)$ is an exponentially decaying exploration reward where ζ is the number of steps since training has started (in contrast, t is the number of steps in the current episode). The exploration reward encourages the model to explore and avoid terminating during the early stages of training; this is similar to intrinsic curiosity rewards as in [29]. Finally, β is a time step penalty to discourage the model from taking too many steps. In our experiments, the exploration reward $\Gamma(\zeta)$ is 0.9999^{ζ} and we set the time step penalty β to be 0.005. As we show in our ablation studies in section VII-E, the step reward for non-terminal actions is critical for good performance.

V. Aesthetics Model

We now describe the aesthetics model we use to model human aesthetics preferences. This model will be used to generate rewards for the RL agent during training.

a) Ranking Views: The task of photo capture requires estimating the aesthetics of images across different viewpoints. Existing aesthetics models are typically trained for cropping, and do not learn to rank images with different viewpoints [5], [6]. The Aesthetic Visual Analysis (AVA) dataset [9] contains scores for images of different content and view points. However, existing work [5] has shown that models trained solely on AVA struggle to perform well on cropping benchmark datasets such as Comparative Photo Composition (CPC), a dataset that contains rankings of different crops of each image. This suggests that AVA and CPC may contain complementary information about human aesthetics preferences, so we use both datasets. For AVA, photos are categorized into different genres, such as landscape and portrait. Because it is difficult to meaningfully rank images from different genres, the model is trained to rank pairs of images from the same genre. We use the standard pairwise ranking loss used in [5], [6]. Let s_1 and s_2 be two images where s_1 should have a higher aesthetics score than s_2. The loss is:

$$\ell_{rank}(s_1, s_2) = \max(0, \phi(s_2) - \phi(s_1) + 1) \quad (4)$$

b) Improving Robustness: We also consider losses to increase the robustness of the model to camera translation and exposure. Small translations are common as noise in camera movements is inevitable during real world deployment, and images should not be scored differently based on very small translations. To train the model to generate similar scores for similar images, we minimally crop images and minimize the Mean Square Error (MSE) between the score of the original image and the minimally cropped image. It is also useful for the model to rank well-exposed images better than over-/under-exposed images. Camera exposure can vary as lighting changes when a photographer navigates an environment. Since CPC and AVA only include well-exposed images, we introduce over-/under-exposed images by increasing/decreasing the brightness of the images. For an image s, the robustness loss is:

$$\ell_{robust}(s) = \lambda_{sim}\ell_{sim}(s) + \lambda_{expo}\ell_{expo}(s) \quad (5)$$
$$\text{where } \ell_{sim}(s) = \frac{1}{2}(\phi(s) - \phi(s_{\text{min crop}}))^2$$
$$\ell_{expo}(s) = \ell_{rank}(s, s_{\text{poorly exposed}})$$

Our full loss function for the aesthetic model is:

$$\mathcal{L}_{aes}(s_1, s_2) = \lambda\ell_{rank}(s_1, s_2) + (1 - \lambda)\ell_{robust}(s) \quad (6)$$

s is selected from $\{s_1, s_2\}$ uniformly, and we set $\lambda = 0.6$, $\lambda_{sim} = 0.875$, and $\lambda_{expo} = 0.125$.

VI. Implementation

In this section, we describe the implementation of our system. Training in simulation is necessary since the model has to interact with the environment for a large number of steps. There are different simulations that can be used such as AirSim [30] and AI Habitat [18]. We choose AI Habitat in our experiments due to its high frame rate and support for realistic indoor datasets like Gibson [17] and Replica [31].

In Section VI-A, we describe the aesthetics model. In Section VI-B we describe the actor-critic RL model.

A. Aesthetic Model Implementation

a) Architecture: We use ResNet18 with a single scalar aesthetic score output. Modern CNNs are inherently sensitive to small pixel translations, so we adopt an antialiasing solution [32] by adding a blur layer during max pooling to increase robustness to small translations.

b) Training Details: Each image batch is drawn from CPC or AVA with equal probability. Minimally cropped images are created by randomly cropping images between 1 and 5 pixels for each side. Over-/under-exposure is generated by multiplying brightness by 4 and 0.5 respectively. The model is trained for 210,000 iterations with batch size 32.

B. RL Model Implementation

a) Architecture: We use an actor-critic setup [33] for the RL agent, illustrated in Fig. 2. The aesthetic model is used as a feature extractor for the camera view. The camera view features are computed by average pooling the output

of each of the four residual blocks and concatenating them together. These features are given to an MLP with an LSTM layer which serves as a common backbone for the actor and critic layers. The output from the MLP+LSTM backbone is a combination of the current view and the memory state in the LSTM, and thus forms a representation of the current state. We use one classification layer for the actor to output a distribution over actions, and one layer for the critic to estimate the current state value. To optimize the architecture parameters, we use PPO [34] using the stable-baselines implementation [35] with default hyper parameters.

b) Aesthetic Score Threshold: The terminal reward depends on an aesthetic score threshold τ_{aes} that captured images should overcome. This threshold is set adaptively based on local regions within each scene, as aesthetics can vary across scenes and across sub-regions within a scene.

We define a local region in an environment by a set of points near the agent's starting location. First, assume we have uniformly sampled some set of N points (and their corresponding views) across the environment. For any starting location, the K nearest neighbors define a local region. The physical neighborhood defined by the KNN views depends on the density of the N sampled points. The target aesthetics threshold for a local region is set by considering the scores of the KNN views to the starting camera position. Specifically, we use the mean aesthetic score of the KNN views μ and the standard deviation of their scores σ and set the threshold to be:

$$\tau_{aes} = \mu + \sigma \tag{7}$$

The objective is to capture an image s_T such that $\phi(s_T) > \tau_{aes}$. This corresponds to the top 16% views in the local region assuming the scores follow a normal distribution.

c) Training Details: We train the model using realistic reconstructions of indoor scenes from Gibson [17] with Habitat to run the simulation. We use a subset of the Gibson dataset that was filtered by [18] to include only high quality reconstructions, and we remove any scenes that include reconstruction artifacts that affect the scene aesthetics. We split the subset of the Gibson dataset into 61 environments for training and 20 environments for evaluation. For each scene, we sample 2,000 random views, and we compute the aesthetic threshold using nearest 100 samples to the position of the initial camera. On every run, we sample a random navigable position and random orientation to set for the initial camera state and re-sample if the score of the initial view is too low (more than a standard deviation below the mean of the *entire* scene). This is because a low-score initialization is likely in a region with poor aesthetics, and the model is unlikely to learn useful policies from such regions. Initialization re-sampling is not done during evaluation. We train the model for 1.5 million steps, using a batch size of 8, and change the associated scene of each element in the batch after every 250 episodes to minimize the overhead of switching between simulated scenes.

TABLE I: **Aesthetics Model Performance.** Compared to [5], our model performs similarly well in assessing aesthetics of image crops (CPC). However, our model performs better on cross-view ranking (AVA), ranks under-/over-exposed images below well-exposed images more often, and assigns more similar scores to nearly identical images. The latter properties are important for assessing aesthetics of different viewpoints under realistic conditions.

Task	Aesthetics (CPC)	Aesthetics (AVA)	Exposure	Minimal Cropping (MSE)
VEN [5]	**75.8**	62.2	81.0	0.29
Ours	72.2	**84.4**	**99.7**	**0.03**

VII. Results

In the video, we include additional visualizations of initial and final view pairs, and clips of agent behaviors.

In this section, we evaluate both our aesthetics model and the AutoPhoto system that is trained with our aesthetics model. In Section VII-A, we compare the performance of our aesthetics model to an existing model [5]. In Section VII-B, we describe the baseline policies that we compare AutoPhoto against. In Sections VII-C and VII-D, we evaluate the behavior of AutoPhoto on unseen environments in simulation. In Section VII-D, we evaluate AutoPhoto in real life with human judgements. Finally, we include ablation studies to verify AutoPhoto design decisions in Section VII-E.

A. Aesthetic Model Evaluation

We evaluate the accuracy of rankings crops from CPC, as well as accuracy of rankings photos of different views (but the same genre) from AVA in Table I. We also include translation robustness results as measured by MSE of scores between images and their minimally-cropped counterparts, and evaluate ranking accuracy with respect to under- or over-exposure. For reference, we compare our model to a state-of-the-art aesthetics model [5]. With respect to ranking crops from the CPC dataset, our model performs a bit lower than [5]. However, the tradeoff is that our model performs better on cross-view ranking on AVA, and is far more robust to minimal translation and unflattering camera exposure.

B. Baseline Policies for Photo Capture

We describe several baseline policies inspired by existing work. As discussed, existing work in autonomous composition during the capture phase cannot be directly applied as they focus primarily on video, and assume predefined objects of interest. For a fair comparison, we limit the total number of steps to 16 steps for the Key Frame Selection and Greedy policies to match the median number of steps taken by our method, as an unlimited number of steps would trivially achieve high accuracy. The policies are:

a) Random: uniformly samples actions.

b) Rule of Thirds: aligns an object of interest on the lower-left or lower-right third of the image. Since we do not have a predefined object, at each time step we compute the salient objects of the scene (similar to [14], [16]) using saliency detection network BASNet [36]. If a salient object satisfies the rule of thirds, then CAPTURE is selected. Otherwise the agent takes a small turn, or moves to adjust the salient object position towards the lower-left or lower-right third of the frame. If no salient object is found, the agent takes a large turn to explore the environment as it is likely that the camera is facing a wall or featureless scene.

c) Imitation Learning: learns actions directly from demonstrations [26]. We adapt [4] to our setting, whose model is trained to predict camera actions for cinematography. We generate demonstrations by sampling paths that lead to a local aesthetic maxima. Only demonstrations where the captured image exceeds the threshold for the current region are considered. Note that memory is important for modelling actions conditioned on paths. Since our proposed model already contains LSTM memory, we utilize the same architecture for this policy (except no critic branch). The model is trained on 17K demonstrations from Gibson with Adam, initial learning rate 1e-4, and exponential learning rate decay γ=0.95. We found 50 epochs to be sufficient, with 500+ epochs yielding no further improvements.

d) Key Frame Selection: explores the scene, and backtracks to the most aesthetic view seen. This is reminiscent of key frame selection for video summarization [37], [38]. Since we do not have a predefined object of interest to track and scenes may be static, we cannot apply cinematography work [2]–[4], [11] to generate trajectories. The agent instead explores the environment uniformly.

e) Greedy: selects an action at each position by executing every possible movement action, undoing that action, and finally selecting the action that improves the aesthetic score the most. If all movement actions would reduce the aesthetic score, CAPTURE is selected. Due to the large number of actions that the policy needs for a single effective step, this policy is very inefficient, but can eventually reach a local maxima if enough steps are executed.

TABLE II: **Accuracy in Achieving the Aesthetic Threshold.** We show the percentage of photos selected by each policy that achieve the aesthetic threshold. Refer to main text for descriptions of the baseline policies. ($\pm\sigma_{stderr}$)

Metric	$\phi(s_T) > \tau_{aes}$ (%)	
Dataset	Gibson	Replica
Random	13.3 ± 0.8	14.3 ± 0.8
Rule of Thirds [14], [16]	19.1 ± 0.9	17.3 ± 0.9
Greedy	56.6 ± 1.1	56.2 ± 1.2
Key Frame Sel. [37], [38]	56.8 ± 1.1	56.6 ± 1.2
Imitation Learning [4]	57.8 ± 1.2	51.9 ± 1.1
Our Method	**81.7 ± 0.9**	**77.8 ± 1.0**

C. Evaluation in Simulation

To ensure that the model can generalize to unseen scenes, we evaluate performance on 20 unseen scenes from Gibson [17]. We also evaluate on 18 scenes from Replica [31] to measure generalization to a different dataset. We show the quantitative evaluation results on Gibson and Replica in Table II. Our model performs significantly better than the baselines. The low performance of the Rule of Thirds policy suggests that careful selection of objects of interest by human experts instead of automatic selection through saliency is important. Further, the rule of thirds does not fully model human aesthetics preferences (which are better captured by the aesthetics model). While both Key Frame Selection and Greedy can achieve strong performance given unlimited time, their performance is limited by inefficient exploration of the environment. The accuracy achieved by these methods is below that of our method when the number of steps taken is set to be similar as our method is more efficient in finding aesthetic views. Note that the performance of our method on Replica is close to the performance on Gibson, indicating that our model generalizes well to views from a different dataset. In Fig. 3 we show some qualitative results against the Key Frame Selection and Greedy baselines on the Replica dataset. AutoPhoto tends to take well-composed photos of furniture and other appropriate objects in the photo.

Fig. 3: **Simulation Visualizations.** We visualize photos captured by our method AutoPhoto against the strongest baselines on the Replica dataset. Note that AutoPhoto tends to better frame furniture compared to other methods.

D. Deployment in Real Life

We deployed our system in real-world settings on a Clearpath Jackal UGV (shown in Fig. 1), and used it to collect 64 photos of indoors environments. We attached a webcam to the Jackal at 1.5m above ground to approximately match the camera angle that would be typically used by human photographers. The input image is fed to the RL model to decide on which action to take, and then the command is sent to the robot through ROS. If the suggested action would lead to a collision, then the RL model is queried again until the proposed action is valid. Since the model contains a memory module, it is able to learn that repeated

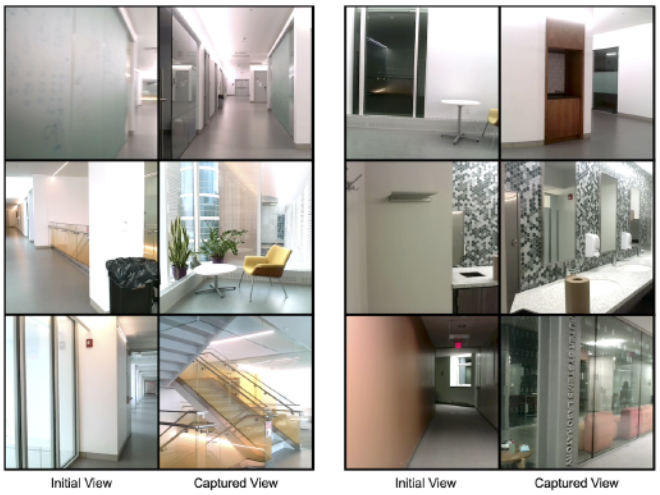

Fig. 4: **Real World Visualizations.** AutoPhoto transfers from simulation to real life to capture well-composed photos.

actions with no change in state means another action should be selected. In Fig. 4 we show sample photos that illustrates the initial views for the robot and the photos it captured.

Since it is not feasible to densely sample photos in real-life to estimate an aesthetic threshold, we conducted a user study on Amazon Mechanical Turk (AMT) to measure how often humans prefer the photos taken by AutoPhoto over the initial view of the environment. Users are shown pairs of images, where each pair consists of an initial view and the final view captured by AutoPhoto. To reduce bias, both the order of pairs as well as images within each pair are shuffled. Users are asked to select which image is more aesthetic, or to select a "tie" option if both images are equally preferred. We select high quality workers through AMT's qualification system by only releasing the study to workers with a Master Qualification, over 95% approval rate, and over 1000 previously approved tasks. For additional quality control, we included three sentinel pairs where the images within each pair are identical to each other. We only keep results from workers who correctly select "tie" for these sentinel pairs, and spend more than 1 second (median) per judgement across all pairs.

We compute a preference score as: 1 if the user preferred the AutoPhoto image, 0 if the user preferred the initial image, and 0.5 if both images are equally preferred. Aggregating scores across 1792 judgements (28 valid user surveys), the mean preference score is 0.63 ± 0.01 (standard error). A t-test with null hypothesis being 0.5 rejects the null hypothesis with $p < 0.05$, indicating that humans prefer the images taken by AutoPhoto.

E. Ablation Studies

We ran ablation studies for (a) the reward function and (b) the model architecture and (c) model actions. For the reward function, the non-terminal terms are ablated (Eq. 3). Specifically, the effect of the exploration term $\Gamma(\zeta)$ and the aesthetic score differences term $(\phi(s_{t+1}) - \phi(s_t))$ are measured. For the architecture, we verify the importance of including LSTM memory in the model, and compare the performance multilayer input features for the agent versus only features from the last layer of the aesthetic model. For the model actions, we verify our decision to include fine and large rotation actions. In Table III we show the results of the ablations on Gibson and Replica. We observe that the exploration and score difference terms are equally important for performance, and removing them both causes the performance to degrade further. Regarding the architecture, we note that removing either memory or multilayer feature extraction lowers performance. Without an LSTM layer, actions cannot depend on the past, making scene exploration to assess the aesthetics of the local region impossible. With regards to input features, using multilayer features provides the agent with both high level and low level information encoded by the aesthetic model, leading to a boost in performance compared to only using the features extracted from the last layer. Finally, the addition of fine rotations (10 °) and large rotations (90°) allows the agent to better adjust view composition and explore the environment.

TABLE III: **Ablation Results**. Our reward function terms, architecture design, and additional rotation actions are important for good performance. ($\pm\sigma_{stderr}$)

	Metric	$\phi(s_T) > \tau_{aes}$ (%)	
	Dataset	Gibson	Replica
Reward	w/o Score Diff. , Explor.	69.3 ± 1.0	58.4 ± 1.2
	w/o Score Diff.	74.4 ± 1.0	66.7 ± 1.1
	w/o Explor.	72.1 ± 1.0	67.7 ± 1.1
Architec.	w/o LSTM	62.7 ± 1.1	55.4 ± 1.2
	w/o Multilayer Feats.	62.3 ± 1.1	62.1 ± 1.1
Actions	w/o 10°, 90° Rotation	73.7 ± 1.0	72.8 ± 1.0
	Our Full Method	**81.7 ± 0.9**	**77.8 ± 1.0**

VIII. Conclusions

In this work, we formulate the problem of photography as a POMDP and train an RL model to automatically capture aesthetically pleasing photos. We demonstrate our AutoPhoto system captures aesthetic photos in unseen scenes across simulation and real life. While our approach can generalize to domains beyond indoor scenes, it depends on having high quality simulated environments which could be harder to get for some domains than others. It is also important to select the aesthetic function appropriately as it influences the system behavior significantly. To extend our work, a possible future direction is to expand the action space to a full 6 degrees of freedom on an aerial drone. Another interesting future direction is to include a user in the loop to ensure that photos taken capture the user intent.

ACKNOWLEDGMENT

We thank Mark Campbell, Yutao Han, Jacopo Banfi, and Vikram Shree for assistance with the Jackal; Utkarsh Mall for providing an implementation of the user study framework; and Aaron Gokaslan for helpful discussions. This work was funded in part by NSF (CHS-1617861 and CHS-1513967) and NSERC (PGS-D 516803 2018).

REFERENCES

[1] J. Chen, H. M. Le, P. Carr, Y. Yue, and J. J. Little, "Learning online smooth predictors for realtime camera planning using recurrent decision trees," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4688–4696. 1, 2

[2] R. Bonatti, W. Wang, C. Ho, A. Ahuja, M. Gschwindt, E. Camci, E. Kayacan, S. Choudhury, and S. Scherer, "Autonomous aerial cinematography in unstructured environments with learned artistic decision-making," *Journal of Field Robotics*, vol. 37, no. 4, pp. 606–641, 2020. 1, 2, 6

[3] H. Jiang, B. Wang, X. Wang, M. Christie, and B. Chen, "Example-driven virtual cinematography by learning camera behaviors," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 45–1, 2020. 1, 2, 6

[4] C. Huang, Z. Yang, Y. Kong, P. Chen, X. Yang, and K.-T. T. Cheng, "Learning to capture a film-look video with a camera drone," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 1871–1877. 1, 2, 6

[5] Z. Wei, J. Zhang, X. Shen, Z. Lin, R. Mech, M. Hoai, and D. Samaras, "Good view hunting: Learning photo composition from dense view pairs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5437–5446. 1, 2, 4, 5

[6] Y.-L. Chen, J. Klopp, M. Sun, S.-Y. Chien, and K.-L. Ma, "Learning to compose with professional photographs on the web," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 37–45. 1, 2, 4

[7] J. Ren, X. Shen, Z. Lin, R. Mech, and D. J. Foran, "Personalized image aesthetics," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 638–647. 1, 2

[8] J.-T. Lee and C.-S. Kim, "Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 2

[9] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis." 1, 2, 4

[10] Y.-L. Chen, T.-W. Huang, K.-H. Chang, Y.-C. Tsai, H.-T. Chen, and B.-Y. Chen, "Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 226–234. 1, 2

[11] C. Huang, C.-E. Lin, Z. Yang, Y. Kong, P. Chen, X. Yang, and K.-T. Cheng, "Learning to film from professional human motion videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4244–4253. 1, 2, 6

[12] J. V. Mascelli, *The five C's of cinematography*. Grafic Publications, 1965. 1

[13] M. L. Gleicher and F. Liu, "Re-cinematography: Improving the camerawork of casual video," *ACM transactions on multimedia computing, communications, and applications (TOMM)*, vol. 5, no. 1, pp. 1–28, 2008. 1

[14] X. Xiong, J. Feng, and B. Zhou, "Automatic view finding for drone photography based on image aesthetic evaluation," in *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 1: GRAPP, (VISIGRAPP 2017)*, INSTICC. SciTePress, 2017, pp. 282–289. 1, 2, 6

[15] Myung-Jin Kim, T. Song, S. Jin, S. Jung, Gi-Hoon Go, K. Kwon, and J. Jeon, "Automatically available photographer robot for controlling composition and taking pictures," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 6010–6015. 1, 2

[16] R. Gallea, E. Ardizzone, and R. Pirrone, "Automatic aesthetic photo composition," in *International Conference on Image Analysis and Processing*. Springer, 2013, pp. 21–30. 1, 2, 6

[17] F. Xia, A. R. Zamir, Z.-Y. He, A. Sax, J. Malik, and S. Savarese, "Gibson env: real-world perception for embodied agents," in *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018, pp. 9068–9079. 2, 4, 5, 6

[18] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019, pp. 9339–9347. 2, 3, 4, 5

[19] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 80–106, 2017. 2

[20] H. Fang and M. Zhang, "Creatism: A deep-learning photographer capable of creating professional work," 2017. 2

[21] D. Li, H. Wu, J. Zhang, and K. Huang, "A2-rl: Aesthetics aware reinforcement learning for image cropping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8193–8201. 2

[22] M. Gschwindt, E. Camci, R. Bonatti, W. Wang, E. Kayacan, and S. Scherer, "Can a robot become a movie director? learning artistic principles for aerial cinematography," *arXiv preprint arXiv:1904.02579*, 2019. 2

[23] C. Gebhardt, S. Stevsic, and O. Hilliges, "Optimizing for aesthetically pleasing quadrotor camera motion," *CoRR*, vol. abs/1906.11686, 2019. [Online]. Available: http://arxiv.org/abs/1906.11686 2

[24] C. Gebhardt, B. Hepp, T. Naegeli, S. Stevsic, and O. Hilliges, "Airways: Optimization-based planning of quadrotor trajectories according to high-level user goals," *CoRR*, vol. abs/1906.11669, 2019. [Online]. Available: http://arxiv.org/abs/1906.11669 2

[25] N. Joubert, D. B. Goldman, F. Berthouzoz, M. Roberts, J. A. Landay, P. Hanrahan, *et al.*, "Towards a drone cinematographer: Guiding quadrotor cameras using visual composition principles," *arXiv preprint arXiv:1610.01691*, 2016. 2

[26] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: A survey of learning methods," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–35, 2017. 2, 6

[27] R. Bonatti, A. Bucker, S. Scherer, M. Mukadam, and J. Hodgins, "Batteries, camera, action! learning a semantic control space for expressive robot cinematography," *arXiv preprint arXiv:2011.10118*, 2020. 2

[28] D. Wierstra, A. Förster, J. Peters, and J. Schmidhuber, "Recurrent policy gradients," *Logic Journal of the IGPL*, vol. 18, no. 5, pp. 620–634, 2010. 3

[29] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *ICML*, 2017. 4

[30] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics*, 2017. [Online]. Available: https://arxiv.org/abs/1705.05065 4

[31] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. D. Nardi, M. Goesele, S. Lovegrove, and R. Newcombe, "The Replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019. 4, 6

[32] R. Zhang, "Making convolutional networks shift-invariant again," *CoRR*, vol. abs/1904.11486, 2019. [Online]. Available: http://arxiv.org/abs/1904.11486 4

[33] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Advances in neural information processing systems*, 2000, pp. 1008–1014. 4

[34] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017. [Online]. Available: http://arxiv.org/abs/1707.06347 5

[35] A. Hill, A. Raffin, M. Ernestus, A. Gleave, A. Kanervisto, R. Traore, P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, and Y. Wu, "Stable baselines," https://github.com/hill-a/stable-baselines, 2018. 5

[36] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7479–7489. 6

[37] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *Journal of visual communication and image representation*, vol. 19, no. 2, pp. 121–143, 2008. 6

[38] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video summarization using deep neural networks: A survey," *arXiv preprint arXiv:2101.06072*, 2021. 6