

---

# Modeling Affect-based Intrinsic Rewards for Exploration and Learning

---

**Dean Zadok\***

Technion  
Haifa, Israel  
deanzadok@  
cs.technion.ac.il

**Daniel McDuff**

Microsoft Research  
Redmond, WA, USA  
damcduff@microsoft.com

**Ashish Kapoor**

Microsoft Research  
Redmond, WA, USA  
akapoor@microsoft.com

## Abstract

Positive affect has been linked to increased interest, curiosity and satisfaction in human learning. In reinforcement learning, extrinsic rewards are often sparse and difficult to define, intrinsically motivated learning can help address these challenges. We argue that positive affect is an important intrinsic reward that effectively helps drive exploration that is useful in gathering experiences. We present a novel approach leveraging a task-independent reward function trained on spontaneous smile behavior that reflects the intrinsic reward of positive affect. To evaluate our approach we trained several downstream computer vision tasks on data collected with our policy and several baseline methods. We show that the policy based on our affective rewards successfully increases the duration of episodes, the area explored and reduces collisions. The impact is the increased speed of learning for several downstream computer vision tasks.

## 1 Introduction

Reinforcement learning (RL) is most commonly achieved via policy specific rewards that are designed for a predefined task or goal. Such *extrinsic* rewards can be sparse and difficult to define and/or only apply to the task at hand. We are interested in exploring the hypothesis that RL frameworks can be designed in a task agnostic fashion and that this will enable us to efficiently learn general representations that are useful in solving several tasks related to perception in robotics. In particular, we consider *intrinsic* rewards that are akin to affect mechanisms in humans and encourage efficient and safe explorations. These rewards are task-independent; thus, the experiences they gather are not specific to any particular activity and can be harnessed to build general representations. Furthermore, intrinsically motivated learning can have advantages over extrinsic rewards as they can reduce the sample complexity by producing rewards signals that indicate success or failure before the episode ends [1].

A key question we seek to answer is how to define such an intrinsic policy. We propose a framework that comprises mechanisms motivated by human affect, that mimic affect experienced by real human subjects. The core insight is that learning agents motivated by drives such as delight, fear, curiosity, hunger, etc. can garner rich experiences that are useful in solving multiple types of tasks. For instance, in reinforcement learning contexts there is a need for an agent to adequately explore its environment [2]. This can be performed by randomly selecting actions, or employing a more intelligent strategy, directed exploration, that incentivizes exploration of unexplored regions [3].

---

\*This paper is the product of work during an internship at Microsoft Research.  
Our implementation is on <https://github.com/microsoft/affectbased>.

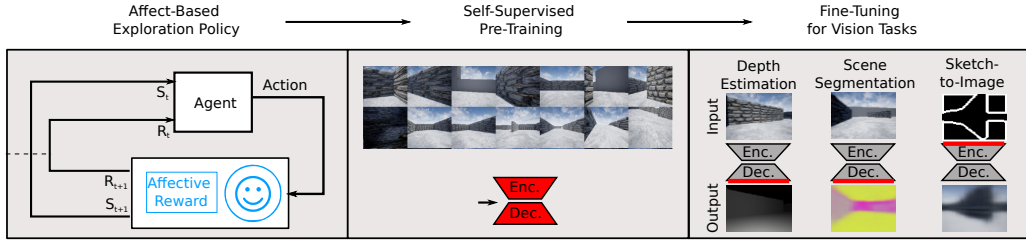


Figure 1: We present a novel approach leveraging a reward that mimics positive affect-based intrinsic rewards to motivate exploration. We use this policy to collect data for self-supervised pre-training and then use the learned representations for multiple downstream computer vision tasks. The red regions highlight the parts of the architecture trained at each stage.

Curiosity is often defined by using the prediction error as the reward signal [4, 5]. As such, the uncertainty, or mistakes, made by the system are assumed to represent what the system should want to learn more about. However, this is a simplistic view as it fails to take into account that new stimuli are not always very informative or useful [6]. Savinov *et al.* uses the analogy of becoming glued to a TV channel surfing when there is the rest of the world outside the window. Their work proposed a new novelty bonus that features episodic memory. McDuff and Kapoor [1] took another approach focusing on safe exploration, proposing intrinsic rewards mimicking responses to that of a human’s sympathetic nervous system (SNS) to avoid catastrophic mistakes during the learning phase, but not necessarily promoting exploration.

In this paper, we specifically focus on the role of positive emotions and study how such intrinsic motivations can enable learning agents to explore efficiently and learn useful representations. In research on education, positive affect has been shown to be related to increased interest, involvement and arousal in learning contexts [7]. Kort *et al.*’s [8] model of emotions in learning posits that the states of curiosity and satisfaction are associated with positive affect in the context of a constructive learning experience. Human physiology is informative about underlying affective states. Smile behavior [9] and physiological signals [1] have been effectively used as feedback in learning systems but not in the context of intrinsic motivation or curiosity. We leverage facial expressions as an unobtrusive measure of expressed positive affect. The key challenges here entail both designing a system that can first model the intrinsic reward appropriately, and then building a learning framework that can efficiently use the data to solve multiple downstream tasks related to perception in robotics.

The core contributions of this paper (summarized in Fig. 1) are to (1) present a novel learning framework in which reward mechanisms motivated by positive affect-mechanisms in humans are used to carry out explorations while being agnostic to any specific tasks, (2) show how the data collected in such an *unsupervised* manner can be used to build general representations useful for solving downstream tasks with minimal task-specific fine-tuning, and (3) report the results on experiments showing that the framework improves exploration as well as enabling efficient learning for solving multiple data-driven tasks. In summary, we argue that such an intrinsically motivated learning framework inspired by affective mechanisms can be effective in increasing the coverage during exploration, decreasing the number catastrophic failures and that the garnered experiences can help us learn general representations for solving tasks including depth estimation, scene segmentation, and sketch-to-image translation.

## 2 Related Work

Our work is inspired by intrinsically motivated learning [10, 11, 12]. One key property of intrinsic rewards is that they are non-sparse [6], this helps aid learning even if the signal is weak. Much of the work in this domain uses a combination of intrinsic and extrinsic rewards in learning. Curiosity is one example of an intrinsic reward that grants a bonus when an agent discovers something new and is vital for discovering successful behavioral strategies. For example, [4, 5] models curiosity via the prediction error as a surrogate and shows that such intrinsic reward mechanism performed similarly to hand-designed extrinsic rewards in many environments. Similarly, Savinov *et al.* [6] defined a different curiosity metric based on how many steps it takes to reach the current observation from those in memory, thus capturing environment dynamics. Their motivation was that previous approaches

were too simplistic in assuming that all changes in the environment should be considered equal. McDuff and Kapoor [1] provided an example of how an intrinsic reward mechanism could motivate safer learning, utilizing human physiological responses to shorten the training time and avoid critical states. Our work is inspired by the prior art, but with the key distinction that we specifically aim to build intrinsic reward mechanisms that are visceral and trained on signals correlated with human affective responses.

Imitation learning (IL) is a popular method for deriving policies. In IL, the model is trained on previously generated examples to imitate the recorded behavior. It has been successfully implemented using data collected from many domains [13, 14, 15, 16, 17, 18]. Simulated environments have been successfully used for training and evaluating IL systems [19, 20]. We use IL as a baseline in our work and perform experiments to show how a combination of IL and positive affect-based rewards can lead to greater and safer exploration.

One of our goals is to explore whether our intrinsically motivated policy can help us learn general representations. Fortunately, the rise of unsupervised generative models, such as generative adversarial networks (GANs) [21] and variational auto-encoders (VAE) [22], has led to progress across many interesting challenges in computer vision that involve image-to-image translation [23, 24]. In our work, we use three tasks in our data evaluation process: scene segmentation [25], depth estimation [26], and sketch-to-image translation [27]. The first two tasks are common in driving scenarios and augmented reality [28, 29, 30], while the third is known for helping people render visual content [31] or synthesize imaginary images [32]. In this paper, we show that by pre-training a VAE in a self-supervised way first using our exploration policy we can obtain better results on all three tasks.

### 3 Our Framework

Fig. 1 describes the overall framework. The core idea behind the proposed methodology is that we model intrinsic motivations in the agents that lead to extensive exploration. Consequently, an agent on its own is able to gather experiences and data in a wide variety of conditions. Note that unlike traditional machine intelligence approaches, the agent is not fixated on a given task - all it is encouraged to do is explore as extensively as possible without getting into perilous situations. The rich data that is being gathered then needs to be harnessed into building representations that will eventually be useful in solving many perceptions tasks. Thus, the framework consists of three core components: (1) A positive-affect based exploration policy, (2) a self-supervised representation learning component and (3) mechanisms that utilize these representations *efficiently* to solve various tasks.

#### 3.1 Affect-Based Exploration Policy

The approach here is to create a model that encourages the agent to explore the environment. We want a reward mechanism that positively reinforces behaviors that mimics a human’s affective responses and lead to discovery and joy.

**Positive Affect-based Reward Mechanism:** In our work, we use a Convolutional Neural Network (CNN) to model the affective responses of a human, as if we were in the same scenario as the agent. We train the CNN model to predict human smile responses as the exploration evolves. Based on the fact that positive affect plays a central role in curiosity and learning [8], we chose to measure smiles as an approximate measure of positive affect. Smiles are consistently linked with positive emotional valence [33, 34] and have a long history of study [35] using electromyography [33] and automated facial coding [36]. We must emphasize that in this work we were not attempting to model the psychological processes that cause people to smile explicitly. We are only using smiles as an outward indicator of situations that are correlated with positive affect as people explore new environments. In particular, the NN was trained to infer the reward  $h(x_i, a)$  directly given that an action  $a$  was taken when at state  $x_i$ . We defer the details of the NN architecture, the process of data collection and the training procedure to Section 4.

**Choosing Actions with Intrinsic Rewards:** Given the reward mechanism, we can use any off-the-shelf sequential decision-making framework, such as RL [1], to learn a policy. It is also feasible to modify an existing policy that is trained to explore or collect data. While the former approach is a

desirable one theoretically, it requires a very large number of training episodes to return a useful policy [37, 38]. We focus on the later, where we assume there exists a function  $f(x_i)$  which can predict a vector of actions probabilities  $p(a)$  when the agent observes state  $x_i$ . Formally, given an observation  $x_i$  and a model  $f$ , the next action,  $a_{i+1}$ , is gets selected as:  $a_{i+1} = \arg \max_a f(x_i)$ . Such a function can be trained on human demonstrations while they explore the environment.

We then use the positive affect model to change the action selection such that it biases the actions that promise to provide better intrinsic rewards. Intuitively, instead of simply using the output of the pre-trained policy  $f$  to decide on the next action, we consider the impact of the intrinsic motivation for every possible action consideration. Formally, given the positive affect model  $h$ , a pre-trained exploration policy  $f$ , observation  $x_i$ , the next action,  $a_{i+1}$ , being selected becomes:

$$a_{i+1} = \arg \max_a \{f(x_i) + \gamma h(x_i, a)\} \quad (1)$$

The above equation adds a weighted intrinsic motivation component to the action probabilities from the original model  $f$ . The weighting parameter  $\gamma$  defines the trade-off between the original policy and the effect of the affect-based reward.

### 3.2 Self-Supervised Learning

Given the exploration policy, the agent has the ability to explore and collect rich data. The next component aims to use this data to build rich representations that eventually could be used for various visual recognition and understanding tasks.

The challenge here is that since the collected data was task agnostic there are no clear labels that could be used for supervised learning. We consequently, use the task of jointly learning an autoencoder and decoder, through a low-dimensional latent representation. Formally, we use a variational autoencoder (VAE) to build such representations. For example, a VAE can be trained to restore just the input image, with the loss constructed as the combination of negative log-likelihood and KL divergence, as follows:

$$L(\theta, \phi) = -\mathbb{E}_{z \sim q_\theta(z|x)} [\log p_\phi(x|z)] + \mathbb{KL}(q_\theta(z|x) || p_\phi(z)) \quad (2)$$

Where the encoder is denoted by  $q_\theta$ , the decoder is denoted by  $p_\phi$  and  $z$  denotes the low-dimensional projection of the input  $x$ . The key intuition here is that if the VAE can successfully encode and decode frames then implicitly it is considering aspects such as depth, segmentation, and textures that are critical to making successful predictions. Thus, it should be possible to tweak and fine-tune these VAE networks to solve a host of visual recognition and perception tasks with minimal effort.

### 3.3 Fine-tuning for Vision Tasks

Given the VAE representation, our goal now is to reuse the learned weights to solve standard machine perception tasks. Formally, given some labeled data corresponding to a visual task, similar to supervised learning, we optimize the negative log-likelihood:

$$L(\phi) = -\mathbb{E}_{z \sim q_\theta(z|x)} [\log p_\phi(x|z)] \quad (3)$$

Note that the goal is to minimally modify the network. In our experiments, we show how we can solve depth map estimation and scene segmentation by only tweaking the weights for the first or last few layers just before the decoder output. We also show how we can use those weights for sketch-to-image translation, even with a small amount of annotated samples.

## 4 Experiments

We conducted experiments to analyze (1) the potential advantages of affect-based exploration policy over other heuristics, (2) the ability to learn general representations and (3) if it was possible to solve diverse computer vision tasks by building upon and minimally modifying the model.

We used a high fidelity simulation environment [39] for autonomous systems, contains a customized 3D maze (dimensions: 2490 meters by 1500 meters), a top-down view can be seen in Fig. 3. The maze is composed of walls and ramps, frames from the environment can be seen in Fig. 2. The agent we used was a vehicle capable of maneuvering comfortably within the maze. To generate random starting points that allowed us to deploy the agent into the maze, we constructed a navigable area, according to the vehicle dimensions and surroundings (the green region in Fig. 3).

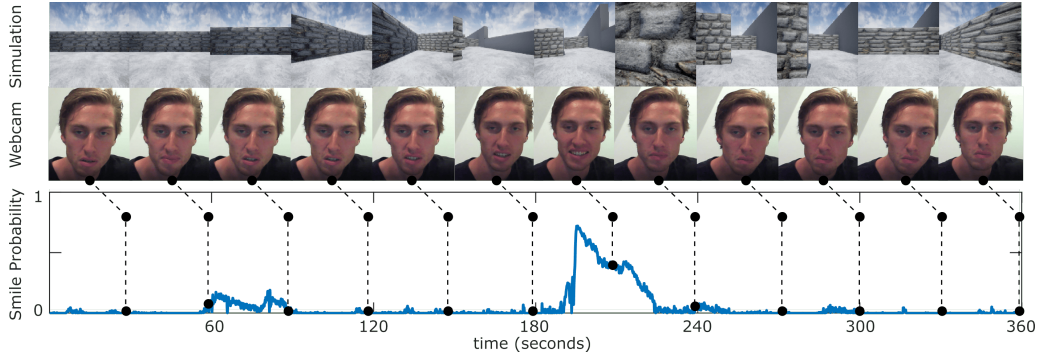


Figure 2: An example of the smile response for a six-minute (360s) period during one of the driving sessions. Frames from the environment and from the webcam video are shown as a reference.

#	Method	Duration (s)	Coverage ( $m^2$ )	Coverage/sec ( $m^2/s$ )	Collisions
(1)	Random	7.57	107.79	<b>14.23</b>	230
(2)	Straight	8.32	115.33	13.86	206
(3)	IL	52.87	727.46	13.75	38
(4)	IL + [1]	<b>87.63</b>	952.82	10.87	<b>23</b>
(5)	Affect-based	79.76	<b>1059.29</b>	13.28	27

Table 1: Evaluation of the driving policies. Given a random starting point, duration is the average time the car drove before a collision and coverage is the average area the car covered.

#### 4.1 Data and Model Training

**Positive Affect-based Reward:** We collected data from five subjects (four males, one female; ages: 23 - 40 years) exploring in the simulated environment. All participants were qualified drivers with multiple years of driving experience. Simultaneously we collected synchronized videos of their face. The participants drove for an average of 11 minutes each, providing a total of approximately 64000 frames. The protocol was approved by our institutional review board. The participants were told to explore the environment but were given no additional instruction about other objectives. We used a well-validated and open-source algorithm to calculate the smile response of the drivers from the webcam videos [40]. When evaluated on a large set of videos of naturalistic facial expressions (very similar to ours) the smile detection had a 0.85 correlation with human coding. An example of a smile response from one subject can be seen in Fig. 2. Using these data we trained our affect-based intrinsic motivation model. The image frames from the camera sensor in the environment served as input to the network and the smile probability in the corresponding webcam frame served as output. The input frames were downsampled to  $84 \times 84$  pixels and normalized to be in the range  $[0, 1]$ . The model architecture is described in the supplemental material.

**Exploration Policy:** We first train the base policy  $f$  by imitation learning, where data was recorded by a human driver that was trained to drive in the simulation using the vehicle. The data set has 50000 images, which were normalized to  $[0, 1]$  and corresponding human actions. The model is a CNN trained to classify the desired steering angle. The input space of the model contains four consecutive images, down-sampled to  $84 \times 84$ , similar to many DQN applications [41]. The action space is discrete and composed of five possible different steering angles:  $40$ ,  $20$ ,  $0$ ,  $-20$  and  $-40^\circ$ . The model architecture is described in the supplemental material. To increase the variations in the collected data and cope better with sharp turns, we shifted sampled frames and post-processed the steering angle accordingly [20]. The final exploration policy embeds this  $f$  as described in Section 3.1, and considers the affective-rewards. Specifically, the reward mechanism was computed for each one of the steering angles, so the positive intrinsic values represent the values inferred from looking directly towards the respective driving directions. We set  $\gamma = 6$  which was determined via cross-validation, as described in the supplemental material.

**Representation Learning:** As the vehicle explores the environment the data it sees is used to train the VAE. Each episode was initiated by placing the vehicle at a random starting point and letting it

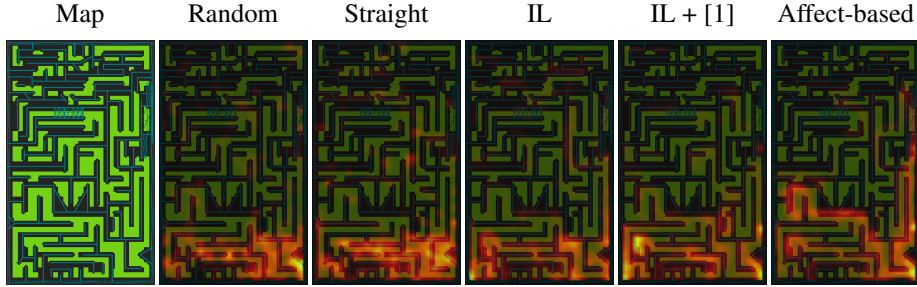


Figure 3: Visualization of the experiment from Table 1 using heat maps. From this visualization we can observe that the better the policy, the longer the paths that are recorded during the trials.

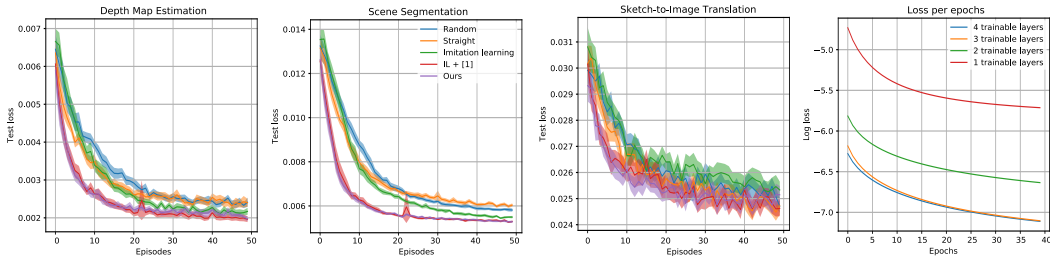


Figure 4: (Left) Test loss as a function of a number of episodes for depth map estimation, scene segmentation, and sketch-to-image translation. Results are averaged over 30 trials. The error bars reflect the standard error. (Right) Test loss as we vary the number of layers tuned for scene segmentation. Fine-tuning just a couple of layers is enough for this task.

drive until the collision. Here we use the task of frame restoration to train the VAE model to restore down-sampled  $64 \times 64$  images. The model architecture is described in the supplemental material. For evaluating performance on depth map estimation and scene segmentation, we collected 2000 images with ground truth, captured by placing simulated cameras randomly in the environment. For sketch-to-image translation, we used the same method except that the sketches were computed by finding the image contours.

## 4.2 Results

**How good is affect-driven exploration?** We compare the proposed method to four additional methods - random, straight, IL and IL + [1]. For the random policy, we simply draw a random action for each timestamp according to a uniform distribution, regardless of the input space. For the straight policy, the model drives straight, without changing course. The IL policy is simply the base policy  $f$ , but without the intrinsically motivated reward component. For [1], to compare their method with ours, we combined their intrinsic reward function with the base policy, the same as we did in ours. When implementing such baselines using the RL approach, we achieved significantly lower results for driving duration.

For this experiment, we select a starting point randomly from the navigable area and let the policy drive until a collision is detected. Then, we reset the vehicle to a random starting position. We continue this process for 2000 seconds, with the vehicle driving at a speed of  $2.5m/s$ . We then consider the mean values of the duration and of the total area covered during exploration per episode. Longer episodes reflect that the policy is able to reason about free spaces and consider that with the vehicle dynamics, whereas higher coverage suggests that the policy indeed encourages novel experiences and that the vehicle is not simply going in circles (or stationary). Coverage is defined as a union of circles, with a 3-meter radius centered around the car.

Fig. 3 shows the map of the environment and how different policies explore the space. The heat signature (yellow is more), indicates the amount of time a vehicle spends at that location. We observe that our policy driven by positive affect-based reward is able to go further and cover significantly bigger space in the allocated 2000 seconds. We present the numerical results in Table 1, which shows

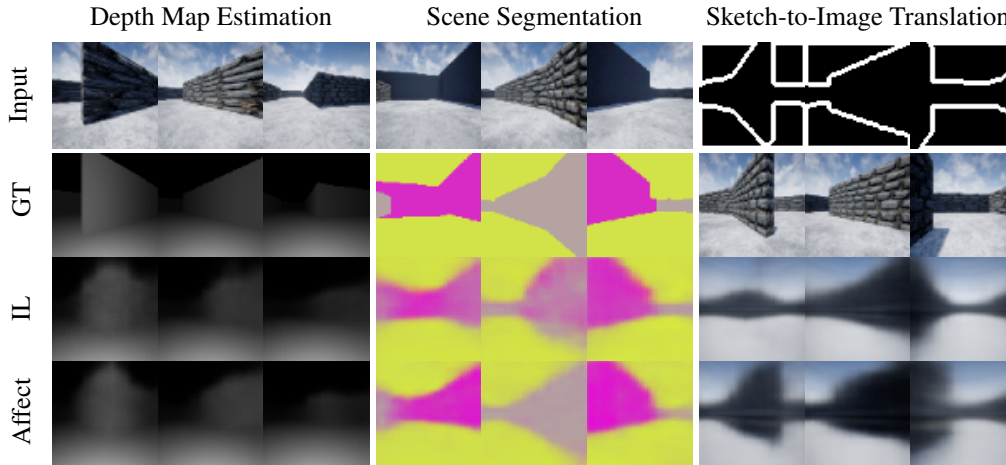


Figure 5: Samples generated using VAEs trained on each of the three tasks. Notice that there are fewer distortions in the depth map estimations, better classification of structures in the scene segmentation, and better generation of the images from sketches when using our proposed policy.

Method	Frame Rest.	Sketch-to-Img.
Random	276.1	273.6
Straight	260.9	271.4
IL	275.4	276.3
IL + [1]	248.7	260.7
Affect-based	<b>242.7</b>	<b>257.8</b>

Table 2: FID scores calculated for the image generation tasks. The FID was calculated on 2000 test images. We compute the metric between the reconstruction and the ground truth. The results are averaged over 30 runs.

that despite being [1] better at holding onto the track, it is less good in exploring the field and cover less area per seconds in average.

**How well can we solve other tasks?** In this experiment, we explore how well the VAE trained on the task-agnostic data collected via the exploration (as described in Section 3.2) can help solve depth-map estimation, scene segmentation, and sketch-to-image translation. First, we explore how much do we really need to perturb the original VAE model to get reasonable performance on these tasks, with only retraining the top few layers of the decoder. In Fig. 4 (right) we plot the log loss, where loss is the task-specific reconstruction L2 loss, as the training evolves over several epochs. The figure shows curves when we choose different numbers of layers to retrain. What we find is that the biggest gains in performance happen with just the top-2 layers. Note that since the loss is in log-scale the difference in performance between training the top-2 layers and retraining a higher number of layers is small. Note that for the sketch-to-image translation it was necessary to retrain the encoder as the input sketches differ from images.

Next, we also study the effect of exploration policies on the three tasks. For these experiments, as the vehicle explores the environment, we in-parallel train the VAE model on the data gathered so far and measure the performance on the three-tasks. We report our results by averaging over 30 trials. Fig. 4 (left) shows the mean test L2 loss as the number of episodes evolves for various exploration policies. While for most of the policies we observe that the error goes down with the number of episodes, for affect-based policy achieves lower errors with fewer episodes. For example, on the task of scene-segmentation, the number of episodes required to achieve a loss of 0.006, is approximately half the number of episodes when using the proposed method than when using the IL policy. We can also see that despite IL + [1] have longer episodes on average, affect-based reaches convergence faster.

Finally, besides the L2 loss, we also examine the realism of the output using the Frechet Inception Distance (FID) [42], a metric that is frequently used to evaluate the realism of images generated using

GANs [21]. The results are presented in Table 2 and show that better FID scores are obtained using the proposed framework.

**How efficient is the framework?** Given that our reward mechanism requires additional computation, we also consider the performance penalty we might be paying. We conducted timed runs for each policy and logged the average frame rate. Our code is implemented using TensorFlow 2.0, with CUDA 10.0 and CUDNN 7.6. For our framework, we observed an average of 16.3 fps using a mobile GTX1060 and 28.9 fps using an RTX2080Ti, which was only slightly worse than 19.1 fps and 41.6 fps respectively, using the IL policy. The results were averaged over 100 seconds of driving.

## 5 Discussion and Conclusions

This paper explores how using positive affect to model an intrinsic motivation for an agent can spur exploration. Greater exploration by the agent can lead to a better representation of the environment and this, in turn, leads to improved performance in a range of downstream tasks. We argued that positive affect is an important drive that spurs more curious behavior. Modeling positive affect as an intrinsically motivated reward led to an exploration policy with 51% longer duration, 46% greater coverage and 29% fewer collisions in comparison to IL. Fig. 3 shows this as a heat map.

Central to our argument is that affective responses to stimuli are intrinsic sources of feedback that lead to exploration and discovery of examples that generalize across contexts. We used our general representations to perform experiments across multiple perception tasks. Comparing performance with and without our affective reward we found a large benefit in using the policy with intrinsic motivation based on the positive affect signal. Qualitative examples (see Fig. 5) show that this led to a better reconstruction of the respective outputs across the tasks. Furthermore, exploration was larger using positive affect as a reward signal than when using a physiological fight-or-flight response [1] which fits with our hypothesis that positive affect is one drive for curiosity.

Here we were not attempting to mimic affective processes. But rather to show that functions trained on affect like signals can lead to improved performance. Smiles are complex nonverbal behaviors that are both common and nuanced; while smiles are interpreted as expressing positive emotion they communicate a variety of interpersonal states [43, 44]. This work establishes the potential for affect like mechanisms in robotics. Extension to other physiological signals presents a further opportunity worth exploring.

## APPENDIX

### 5.1 Network Architectures

Table 3 shows the network architecture for the driving policy described in the paper. The architecture contains three convolutional layers and two dense layers. The input shape is  $[84 \times 84 \times 4]$  (four consecutive grayscale images). The output is a vector of probabilities with a shape equal to the number of possible actions.

Table 4 shows the network architecture for the affect-based reward function described in the paper. The architecture contains three convolutional layers and two dense layers. Batch normalization is applied prior to each intermediate layer. The input shape is  $[84 \times 84 \times 3]$  and contains a single RGB image. The output is a vector of two values. The first is the positive affect-based value that is being used by our affect-based policy, and the second is the proposed reward output from [1] which is being examined using our model architecture in the fourth baseline described in the experiments.

Table 5 shows the network architecture for the VAE model described in the paper. The encoder and the decoder composed of five layers each, with batch normalization prior to each intermediate layer. The input shape is  $[64 \times 64 \times 3]$  and contains a single RGB image. The output of the encoder is an 8-dimensional latent space representation. The output shape of the decoder is  $[64 \times 64 \times 3]$  for a single RGB/segmentation image, or  $[64 \times 64 \times 1]$  for a depth estimation map.

For more details about the training procedure and parameters, our code is publicly available on <https://github.com/microsoft/affectbased>.



Layer	Act.	Out. shape	Params	Layer	Act.	Out. shape	Params
Conv2D	ReLU	$16 \times 8 \times 8$	4.1k	Conv2D	ReLU	$32 \times 5 \times 5$	2.4k
Conv2D	ReLU	$32 \times 4 \times 4$	8.2k	Conv2D	ReLU	$48 \times 4 \times 4$	24.6k
Conv2D	ReLU	$32 \times 3 \times 3$	9.2k	Conv2D	ReLU	$64 \times 4 \times 4$	49.2k
Dense	ReLU	256	401k	Dense	ReLU	2048	8.4M
Dense	Softmax	5	1.2k	Dense	Linear	2	4k
Total			424k	Total			8.4M

Table 3: CNN architecture for the navigation policy. Table 4: CNN architecture for the affect-based reward function.

Layer	Act.	Out. shape	Params
Encoding Layers			
Conv2D	ReLU	$64 \times 4 \times 4$	3.1k
Conv2D	ReLU	$128 \times 4 \times 4$	131k
Conv2D	ReLU	$256 \times 4 \times 4$	524k
Dense	ReLU	1024	9.4M
Dense	Linear	16	16.4k
Decoding Layers			
Dense	ReLU	1024	9.2k
Dense	ReLU	6272	6.4M
Conv2D Transpose	ReLU	$128 \times 4 \times 4$	262k
Conv2D Transpose	ReLU	$64 \times 4 \times 4$	131k
Conv2D Transpose	Sigmoid	$3 \times 4 \times 4$	3k
Total			16.9M

Table 5: Architecture for the convolutional VAE model.

## 5.2 Reward Multiplication Factor

Our method relies on adding the reward component on top of the action probabilities such that it will maximize the exploration results. To find the best  $\gamma$ , we performed the experiment shown in the coverage table that is in the paper for a range of potential values, and then we used this  $\gamma$  to perform the rest of the experiments. An example of an experiment to find the  $\gamma$  can be seen in Fig. 6.

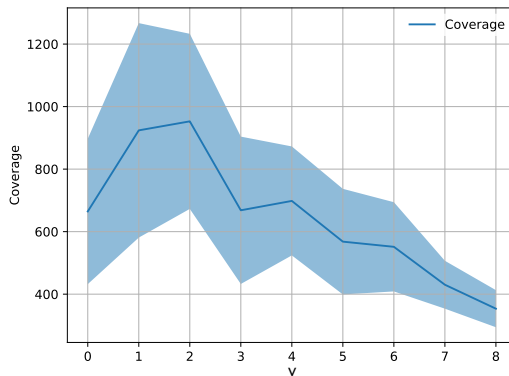


Figure 6: Mean coverage as a function of gamma when searching for the right gamma for IL + [1]. The plot shows that  $\gamma = 2$  resulted in the highest coverage.

## References

- [1] Daniel McDuff and Ashish Kapoor. Visceral machines: Risk-aversion in reinforcement learning with intrinsic physiological rewards. *International Conference on Learning Representations (ICLR)*, 2019.
- [2] Mikhail Frank, Jürgen Leitner, Marijn Stollenga, Alexander Förster, and Jürgen Schmidhuber. Curiosity driven reinforcement learning for motion planning on humanoids. *Frontiers in neurorobotics*, 7:25, 2014.
- [3] Tao Chen, Saurabh Gupta, and Abhinav Gupta. Learning exploration policies for navigation. In *ICLR*, 2019.
- [4] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Large-scale study of curiosity-driven learning. In *ICLR*, 2019.
- [5] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning (ICML)*, 2017.
- [6] Nikolay Savinov, Anton Raichuk, Raphaël Marinier, Damien Vincent, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. *arXiv preprint arXiv:1810.02274*, 2018.
- [7] John C Masters, R Christopher Barden, and Martin E Ford. Affective states, expressive behavior, and learning in children. *Journal of Personality and Social Psychology*, 37(3):380, 1979.
- [8] Barry Kort, Rob Reilly, and Rosalind W Picard. An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In *Proceedings IEEE International Conference on Advanced Learning Technologies*, pages 43–46. IEEE, 2001.
- [9] Natasha Jaques, Jennifer McCleary, Jesse Engel, David Ha, Fred Bertsch, Rosalind Picard, and Douglas Eck. Learning via social awareness: Improving a deep generative sketching model with facial feedback. *arXiv preprint arXiv:1802.04877*, 2018.
- [10] Nuttapon Chentanez, Andrew G Barto, and Satinder P Singh. Intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pages 1281–1288, 2005.
- [11] Nick Haber, Damian Mrowca, Stephanie Wang, Li F Fei-Fei, and Daniel L Yamins. Learning to play with intrinsically-motivated, self-aware agents. In *Advances in Neural Information Processing Systems*, pages 8388–8399, 2018.
- [12] Zeyu Zheng, Junhyuk Oh, and Satinder Singh. On learning intrinsic rewards for policy gradient methods. In *Advances in Neural Information Processing Systems*, pages 4644–4654, 2018.
- [13] Aude Billard and Maja J Matarić. Learning human arm movements by imitation:: Evaluation of a biologically inspired connectionist architecture. *Robotics and Autonomous Systems*, 37(2-3):145–160, 2001.
- [14] Valts Blukis, Nataly Brukhim, Andrew Bennett, Ross A Knepper, and Yoav Artzi. Following high-level navigation instructions on a simulated quadcopter with imitation learning. *arXiv preprint arXiv:1806.00047*, 2018.
- [15] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.
- [16] Luigi Cardamone, Daniele Loiacono, and Pier Luca Lanzi. Learning drivers for torcs through imitation using supervised methods. In *2009 IEEE symposium on computational intelligence and games*, pages 148–155. IEEE, 2009.

- [17] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635, 2011.
- [18] Stéphane Ross, Narek Melik-Barkhudarov, Kumar Shaurya Shankar, Andreas Wendel, Debadeepta Dey, J Andrew Bagnell, and Martial Hebert. Learning monocular reactive uav control in cluttered natural environments. In *2013 IEEE international conference on robotics and automation*, pages 1765–1772. IEEE, 2013.
- [19] Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–9. IEEE, 2018.
- [20] Dean Zadok, Tom Hirshberg, Amir Biran, Kira Radinsky, and Ashish Kapoor. Explorations and lessons learned in building an autonomous formula sae car from simulations. *SIMULTECH*, 2019.
- [21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [24] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.
- [25] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR, 2017 IEEE Transactions on Pattern Analysis and Machine Intelligence*, abs/1606.00915, 2017.
- [26] Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.
- [27] Wengling Chen and James Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9416–9425, 2018.
- [28] Xinlei Pan, Yurong You, Ziyang Wang, and Cewu Lu. Virtual to real reinforcement learning for autonomous driving. *arXiv preprint arXiv:1704.03952*, 2017.
- [29] J Edward Swan, Adam Jones, Eric Kolstad, Mark A Livingston, and Harvey S Smallman. Egocentric depth judgments in optical, see-through augmented reality. *IEEE transactions on visualization and computer graphics*, 13(3):429–442, 2007.
- [30] Dequan Wang, Coline Devin, Qi-Zhi Cai, Philipp Krähenbühl, and Trevor Darrell. Monocular plan view networks for autonomous driving. *arXiv preprint arXiv:1905.06937*, 2019.
- [31] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Trans. Graph.*, 31(4):44–1, 2012.
- [32] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2photo: Internet image montage. In *ACM transactions on graphics (TOG)*, volume 28, page 124. ACM, 2009.
- [33] Serena-Lynn Brown and Gary E Schwartz. Relationships between facial electromyography and subjective experience during affective imagery. *Biological psychology*, 11(1):49–62, 1980.

- [34] Karim Sadik Kassam. *Assessment of emotional experience through facial expression*. Harvard University, 2010.
- [35] Marianne LaFrance. *Lip service: Smiles in life, death, trust, lies, work, memory, sex, and politics*. WW Norton & Company, 2011.
- [36] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*, 2017.
- [37] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.
- [38] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.
- [39] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and service robotics*, pages 621–635. Springer, 2018.
- [40] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.
- [41] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
- [42] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.
- [43] Paul Ekman and Wallace V Friesen. Felt, false, and miserable smiles. *Journal of Nonverbal Behavior*, 6(4):238–252, 1982.
- [44] Magdalena Rychlowska, Rachael E Jack, Oliver G B Garrod, Philippe G Schyns, Jared D Martin, and Paula M Niedenthal. Functional smiles: Tools for love, sympathy, and war. *Psychological Science*, 28(9):1259–1270, 2017.