Atari-HEAD: Atari Human Eye-Tracking and Demonstration Dataset

Ruohan Zhang,[1*] Calen Walshe,[2] Zhuode Liu,[1] Lin Guan,[1] Karl S. Muller,[2]
Jake A. Whritner,[2] Luxin Zhang,[3] Mary M. Hayhoe,[2] Dana H. Ballard[1,2]
[1]Department of Computer Science, University of Texas at Austin
[2]Center for Perceptual Systems, University of Texas at Austin
[3]The Robotics Institute, Carnegie Mellon University
[*]zharu@utexas.edu

Abstract

Large-scale public datasets have been shown to benefit research in multiple areas of modern artificial intelligence. For decision-making research that requires human data, high-quality datasets serve as important benchmarks to facilitate the development of new methods by providing a common reproducible standard. Many human decision-making tasks require visual attention to obtain high levels of performance. Therefore, measuring eye movements can provide a rich source of information about the strategies that humans use to solve decision-making tasks. Here, we provide a large-scale, high-quality dataset of human actions with simultaneously recorded eye movements while humans play Atari video games. The dataset consists of 117 hours of gameplay data from a diverse set of 20 games, with 8 million action demonstrations and 328 million gaze samples. We introduce a novel form of gameplay, in which the human plays in a semi-frame-by-frame manner. This leads to near-optimal game decisions and game scores that are comparable or better than known human records. We demonstrate the usefulness of the dataset through two simple applications: predicting human gaze and imitating human demonstrated actions. The quality of the data leads to promising results in both tasks. Moreover, using a learned human gaze model to inform imitation learning leads to an 115% increase in game performance. We interpret these results as highlighting the importance of incorporating human visual attention in models of decision making and demonstrating the value of the current dataset to the research community. We hope that the scale and quality of this dataset can provide more opportunities to researchers in the areas of visual attention, imitation learning, and reinforcement learning.

Introduction

In modern machine learning, large-scale datasets such as ImageNet (Deng et al. 2009) have played an important role in driving research progress. These datasets provide standardized benchmarks that ensure a fair comparison between algorithms. Recently, imitation learning (IL) and reinforcement learning (RL) have achieved great success in training learning agents to solve sequential decision tasks. The goal for the learning agents is to learn a *policy*–a mapping from states to actions–that maximizes long-term cumulative reward. The policy can be learned through trial and error (RL) or from an expert's demonstration (IL). A major issue of RL is its sample inefficiency and human demonstration has been shown to speed up learning (Silver et al. 2016; Hester et al. 2018; de la Cruz, Du, and Taylor 2018; Zhang et al. 2019).

However, IL results are difficult to reproduce since researchers often collect their own human demonstration data. During this process, many factors are uncontrolled – such as individual expertise, experimental setup, data collection tools, dataset size, and experimenter bias. A publicly available dataset would greatly reduce data collection efforts and allow algorithms to be compared with the same standard. Another concern with IL is the quality of the demonstration data. For supervised IL approaches like behavior cloning, learning from sub-optimal demonstration can result in poor performance. Therefore the quality of human demonstration must be ensured.

Visual perception is a key challenge in modern RL and IL research due to a high-dimensional state space (e.g., raw images) as input. Humans face the same problem when performing all kinds of visuomotor tasks in daily life. One intelligent mechanism that has evolved in humans, but has not yet been fully developed for machines, is visual attention – the ability to identify, process, and respond to a reduced set of important features of visual input. This powerful feature extraction mechanism, if learned by AIs, could make learning more efficient.

Human overt attention is revealed by eye movements (gaze). In complex tasks, human eye movements are used by the visual system to a) identify structures in the environment that are critical for solving the task and b) exploit those structures by moving the high resolution part of the visual field (fovea) to those locations via eye movements. Considerable evidence has shown that human gaze can be considered as an overt behavioral signal that encodes a wealth of information about both the motivation behind an action and the anticipated reward of an action (Hayhoe and Ballard 2005; 2014; Johnson et al. 2014). Recent work has also proposed learning visual attention models from human gaze as an intermediate step towards learning

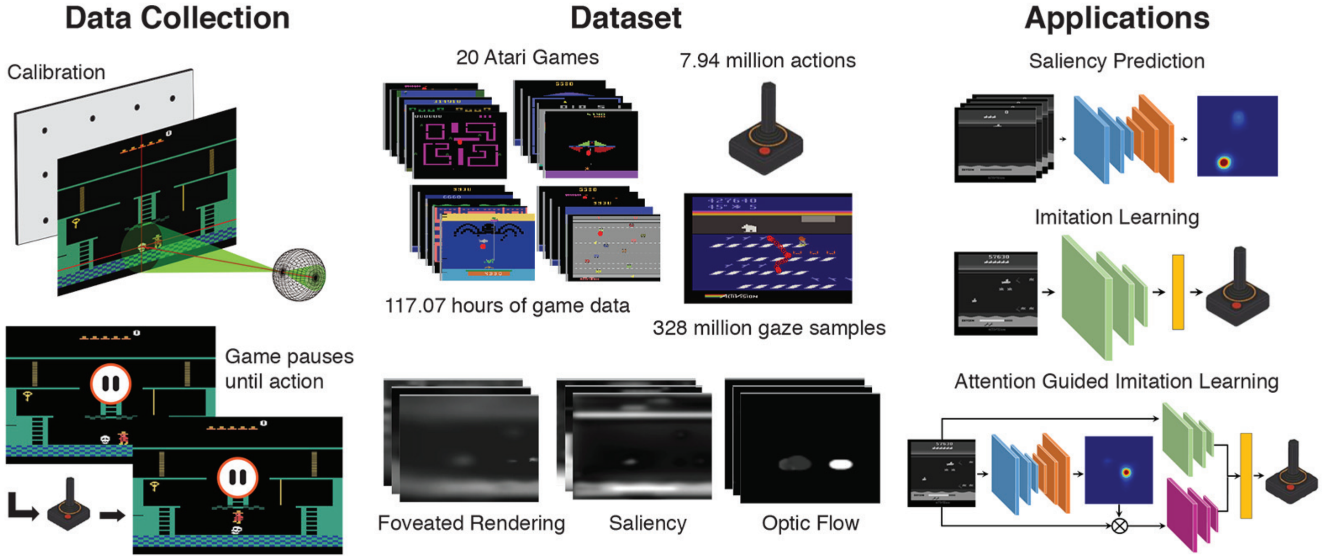


Figure 1: Project schematic for the Atari-HEAD dataset.

the decision policy, and this intermediate signal has been shown to improve policy learning (Li, Liu, and Rehg 2018; Zhang et al. 2018b; Xia et al. 2019; Chen et al. 2019; Liu et al. 2019; Deng et al. 2019)

Addressing the demands and challenges described above, we collected a large-scale dataset of humans playing Atari video games – one of the most widely used task domain in RL and IL research. The dataset is named Atari-HEAD (**Atari** **H**uman **E**ye-Tracking **A**nd **D**emonstration)[1]. An overview of this project can be found in Fig. 1. In collecting Atari-HEAD, we strictly follow standard data collection protocols for human studies and designed a special method to ensure the quality of demonstration policies. The result of these efforts is a dataset with expert-level task performance and minimal recording error. Making this dataset publicly available saves the effort of data collection and provides a benchmark for researchers who use Atari games as their task domain. Having both action and gaze data enables research that aims at bridging attention and control.

Related Work

In imitation learning research, the Atari Grand Challenge dataset pioneered the effort of collecting a large-scale public dataset of Atari games (Kurin et al. 2017). The human demonstration was collected through online crowdsourcing with players of diverse skill levels. Recently, researchers have spent significant effort in building large-scale datasets of human demonstrations in various tasks, including driving (Yu et al. 2018), playing Minecraft (Guss et al. 2019), and manipulating simulated robots (Mandlekar et al. 2018). Our dataset joins their effort in providing a standard dataset for the RL and IL research community.

Gaze prediction was formalized as a visual saliency prediction problem in computer vision research (Itti, Koch, and Niebur 1998). Large-scale datasets have enabled deep learning approaches to make tremendous advances in this area. Examples include MIT saliency benchmark (Bylinskii et al.), CAT2000 (Borji and Itti 2015), and SALICON (Jiang et al. 2015). However, the traditional saliency prediction task does not involve tasks nor human decisions. The humans look at static images or videos in a free-viewing manner without performing any particular task, and only the eye movements are recorded and modeled. How humans distribute their visual attention for dynamic, reward-seeking visuomotor tasks has received less attention in research on saliency. Recently, eye-tracking video datasets of subjects cooking (Li, Liu, and Rehg 2018) and driving (Alletto et al. 2016) in naturalistic environment have been published; this allows researchers to study the relation between attention and decision. We hope that the Atari-HEAD dataset can serve a similar purpose for visual saliency and visuomotor behavior research.

Atari-Head: Design and Data Collection

Our data was collected using the Arcade Learning Environment (ALE) (Bellemare et al. 2012). These games capture many interesting aspects of the natural visuomotor tasks while allowing better experimental control than real-world tasks. ALE is deterministic given the same game seed. While collecting human data, the seed is randomly generated to introduce stochasticity for gameplay. We pick 20 games that span a variety of dynamics, visual features, reward mechanisms, and difficulty levels (for both human and AI). For every game image frame i, we recorded its corresponding image frame I_i, human keystroke action a_i, human decision time t_i, gaze positions $g_{i1}...g_{in}$, and the immediate reward r_i returned by the environment. The gaze data was recorded using an EyeLink 1000 eye tracker at 1000Hz. The game screen was 64.6 × 40.0cm (or 1280×840 in pixels), and the distance to the subjects' eyes was 78.7cm. The visual angle

[1] Available at: https://zenodo.org/record/3451402

of an object is a measure of the size of the object's image on the retina. The visual angle of the screen was 44.6×28.5 visual degrees.

The human subjects were amateur players who were familiar with the games. The human research was approved by the University of Texas at Austin Institutional Review Board with approval number 2006-06-0085. We collected data from 4 subjects playing 20 games. The total collected game time is 117.07 hours, with 7,937,159 action demonstrations and 328,870,044 usable gaze samples.

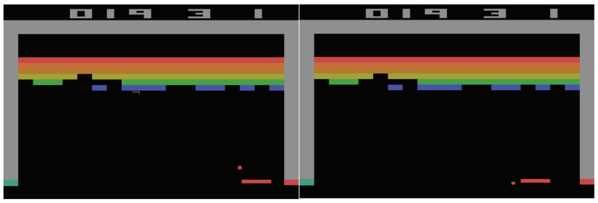
The subjects were only allowed to play for 15 minutes, and were required to rest for at least 15 minutes before the next trial. We mainly collected human data from the first 15 minutes of game play, since for most games AIs have not reached human performance at a 15-minute cutoff. Therefore we reset the game to start from the beginning for every trial. However, it is also interesting to know the human performance limit, hence for each game we let one human player play until the game terminated, or a 2 hour maximum time limit has been reached.

Eye-tracking accuracy The Eyelink 1000 tracker was calibrated using a 16-point calibration procedure at the beginning of each trial, and the same 16 points were used at the end of trial to estimate the gaze positional error. The average end-of-trial gaze positional error across 471 trials was 0.40 visual degrees (or 2.94pixels/0.56cm), less than 1% of the stimulus size. Such high tracking accuracy is critical for Atari games, since many task-relevant objects are small.

Semi-frame-by-frame game mode In the default ALE setting, the game runs continuously at 60Hz, a speed that is very challenging even for expert human players. Previous studies have collected human data, or evaluated human performance at this speed (Mnih et al. 2015; Wang et al. 2016; Kurin et al. 2017; Hester et al. 2018). However, we argue that in order to build a dataset useful for algorithms such as IL, a slower speed should be used. An innovative feature of our setup is that the game pauses at every frame, until a keyboard action is taken by the human player. If desired, the subjects can hold down a key and the game will run continuously at 20Hz, a speed that is reported to be comfortable for most players. The reasons for such a setup are as follows:

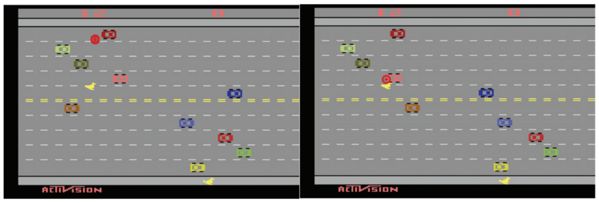
Resolving state-action mismatch Closed-loop human visuomotor reaction time Δt is around 250-300 milliseconds. Therefore, during continuous gameplay, s_t and a_t that are simultaneously recorded at time step t could be mismatched. Action a_t could be intended for a state $s_{t-\Delta t}$ 250-300ms ago. An example that illustrates this point is shown in Fig. 2. This effect causes a serious issue for supervised learning algorithms, since label a_t and input s_t are no longer matched. Frame-by-frame game play ensures that s_t and a_t are matched at every timestep.

Maximizing human performance Frame-by-frame mode makes gameplay more relaxing and reduces fatigue, which could normally result in blinking and would corrupt eye-tracking data. More importantly, this design reduces sub-optimal decisions caused by inattentive blindness. See Fig. 3



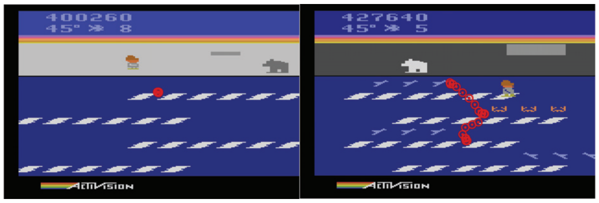
(a) No-op (b) Left

Figure 2: State-action mismatch (game Breakout, game speed at 60Hz). The state s_0 at time t_0 is shown in (a), the correct action would be to move the paddle left to catch the ball. However, due to the human player's delayed reaction, that action is executed 287ms (17 frames) later, as shown in (b). This delay leads to two undesirable consequences: 1) The player loses a life in the game; 2) Action "Left" is paired with state s_{17}, instead of s_0, which posits a serious issue for algorithms that attempts to learn the state-action mapping.



(a) (b)

Figure 3: Inattentive blindness (game Freeway, game speed at 60Hz). (a) The player's gaze (red dot) was on the red car. (b) The pink car hits the chicken controlled by the player 205ms later. Due to the fast pace of the game, the human player was not able to make an eye movement to attend and respond to the pink car.



(a) 217ms (b) 1158ms

Figure 4: Scanpath and reaction time (game Frostbite, frame-by-frame mode). (a) A simple game state that only takes one fixation and 217ms to make a decision. (b) A complicated state which requires a sequence of eye movements and 1158ms to plan the next action. Our game mode allows enough time for the human player to process visual information and find the optimal action.

for an example.

Highlighting critical states that require multiple eye movements Human decision time and all eye movements were recorded at every frame. Hypothetically, the states that could lead to a large reward or penalty, or the ones that require sophisticated planning, will take longer and require multiple eye movements for the player to make a decision. Fig. 4 shows an example. Stopping gameplay means that the observer can use eye-movements to resolve complex situations like (b). This is important because if the algorithm is going to learn from eye-movements it must contain all "relevant" eye-movements.

Dataset statistics The experimental designs result in a high-quality human demonstration dataset. The optimality of demonstrated actions can be intuitively measured by final game scores (when the players lose their last life). In Table 1, we compare our human scores with ones reported in previous literature, along with Atari game world records, as well as one of the best RL agent's performance (Hessel et al. 2018). We reported the average and the best game scores in 15-minute trials, as well as the highest score reached in the 2-hour game play mode. The immediate observation is that our design leads to better human performance compared to those previously reported. The community world record is from Twin Galaxies[2], an official supplier of verified world records by Guinness World Records. For 8 games, our human players have obtained comparable or better scores than world records. For 6 other games, the 2-hour time limit was reached but the human players could surpass the world record if they continued to play.

In recent years, the gap between human and machine performance in many tasks has substantially narrowed (Mnih et al. 2015). AI agents such as DQN play the game in the frame-by-frame manner (although reaction time is not a big issue for RL agents), but in previous literature humans played the game continuously at 60Hz. In our case, allowing human players to have enough decision time sets a stronger human performance baseline for RL agents. Our human score statistics indicate that humans retain advantages in these games, especially ones that require multitasking and divided attention. For difficult games recognized by the RL research community, such as Montezuma's Revenge, human performance is still much higher than that of AI.

Applications of Atari-HEAD

Next, we will demonstrate two main modeling tasks that can be accomplished with this dataset: saliency prediction and learning action from humans. We will define both tasks, discuss inputs and outputs of the models, propose evaluation metrics, show baseline modeling results, and mention potential future directions for both tasks.

Saliency prediction

Task definition The first learning task could be training an agent to imitate human's gaze behaviors, i.e., learning to attend to important regions of a given image. The problem is formalized as a visual saliency prediction problem in computer vision research. The problem can be formulated as:

> Given a state s_t, learn to predict human gaze positions g_t, i.e., learn $P(g|s)$.

Inputs and outputs In the above formulation, note that g_t could be a set of positions in our dataset. s_t could be a single image I_t, or it could include a stack of images $I_{t-n} \dots I_t$ to take into account more history. This includes information such as motion that can make states Markovian (Mnih et al. 2015). The images are published in RGB format, but it is common to convert them to be grayscale (Mnih et al. 2015). Note that for this dataset, two adjacent images, actions, or gaze locations are highly correlated. We suggest that users split data first then shuffle, instead of shuffle first then split, so one can avoid putting one frame in the training set and its neighboring frame in the testing set.

In saliency prediction, additional image statistics are shown to be correlated with visual attention and useful for gaze prediction (Palazzi et al. 2018; Li, Liu, and Rehg 2018; Zhang et al. 2018b). We also provide tools to extract optical flow (Farnebäck 2003) and hand-crafted bottom-up saliency features (orientation and intensity) (Itti, Koch, and Niebur 1998). Examples of these can be seen in the third and fourth columns of Fig. 5. They can be directly used as reasonable guesses for visual saliency (Itti and Koch 2001; Rudoy et al. 2013).

The gaze prediction model should output $P(g_t|s_t)$. In standard practice, discrete gaze positions are converted into a continuous distribution (Bylinskii et al. 2018) by blurring each fixation location using a Gaussian with σ equals to one visual degree (Le Meur and Baccino 2013). Hence the gaze prediction model will learn to predict this continuous probability distribution over the given image, which will be referred as a saliency map.

Evaluation metrics Once the conversion is done, at least eight well-known metrics can be applied to measure prediction accuracy (Bylinskii et al. 2018). Let P denote the predicted saliency map, Q denote the ground truth, and i denote the ith pixel. We discuss four selected metrics here:

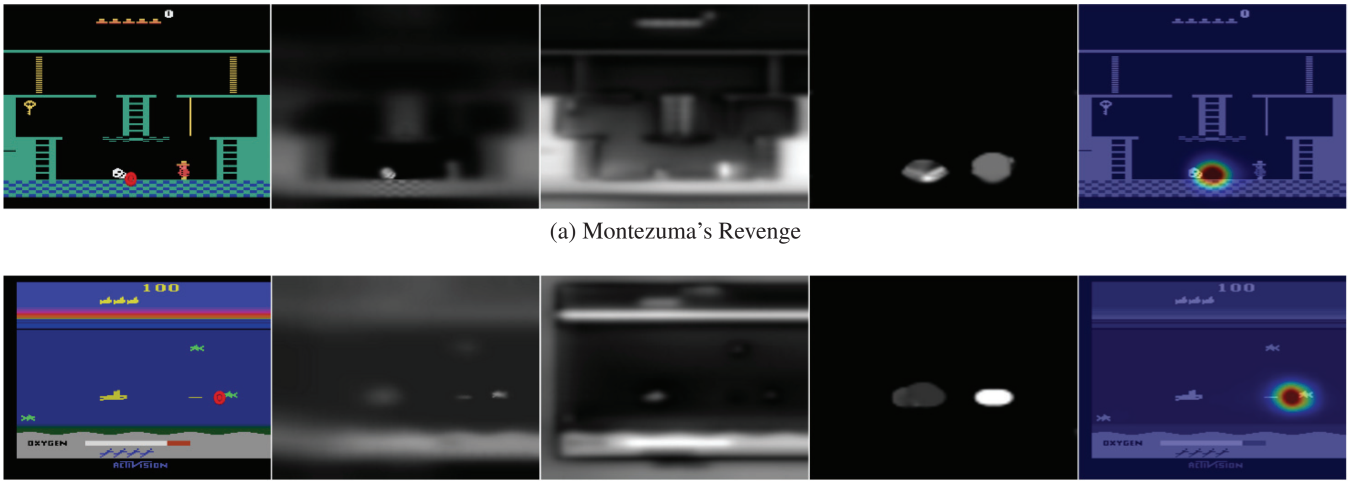
- **Area Under ROC Curve (AUC):** between 0 and 1. One can treat predicted saliency map as a binary classifier to indicate whether a pixel is fixated or not. Hence AUC, one of the most widely used metric in signal detection and classification problems can be applied here.
- **Normalized Scanpath Saliency (NSS):** This metric measures the normalized saliency at gaze positions by subtracting the mean predicted saliency value. It is sensitive to false positives and differences in saliency across predicted saliency map, but is invariant to linear transformations like contrast offsets:

$$NSS(P,Q) = \frac{1}{\sum_i Q_i} \sum_i \Big(\frac{P_i - \mu(P)}{\sigma(P)} \times Q_i \Big) \qquad (1)$$

[2] https://www.twingalaxies.com/games

	Mnih	Wang	Hester	Kurin	de la Cruz	AtariHEAD 15-min avg.	AtariHEAD 15-min best	AtariHEAD 2-hour	Community Record	RL
alien	6,875	7,127.7	29,160	-	-	27,923	34,980	**107,140**†	103,583	9,491.7
asterix	8,503	8,503.3	18,100	-	14,300	110,133.3	135,000	**1,000,000**‡	**1,000,000**	428,200.3
bank_heist	734.4	753.1	7,465	-	-	5,631.3	6,503	**66,531**†	47,047	1,611.9
berzerk	-	2,630.4	-	-	-	6,799	7,950	55,220⋆	**171,770**	2,545.6
breakout	31.8	30.5	79	-	59	439.7	554	**864**‡	**864**	612.5
centipede	11,963	12,017	-	-	-	45,064	55,932	415,160⋆	**668,438**	9,015.5
demon_attack	3,401	3,442.8	6,190	-	-	7,097.3	10,460	107,045⋆	108,075	**111,185.2**
enduro	309.6	860.5	803	-	-	336.4	392	**4,886**⋆	-	2,259.3
freeway	29.6	29.6	32	-	-	31.1	33	33†	**34**	**34.0**
frostbite	4,335	4,334.7	-	-	-	31,731.5	50,630	**453,880**⋆	418,340	9,590.5
hero	25,763	30,826.4	99,320	-	-	59,999.8	77,185	541,640⋆	**1,000,000**	55,887.4
montezuma	4,367	4,753.3	34,900	27,900	-	38,715	46,000	270,400⋆	**400,000**	384.0
ms_pacman	15,693	15,375.0	55,021	29,311	18,241	28,031	36,061	93,721†	**123,200**	6,283.5
name_this_game	4,076	8,049.0	19,380	-	4,840	7,661.5	8,870	**21,850**†	21,210	13,439.4
phoenix	-	7,242.6	-	-	-	30,800.5	40,780	**485,660**⋆	373,690	108,528.6
riverraid	13,513	17,118	39,710	-	-	20,048	22,590	59,420†	**86,520**	-
road_runner	7,845	7,845	20,200	-	-	78,655	99,400	99,400†	**210,200**	69,524.0
seaquest	20,182	42,054.7	101,120	-	-	52,774	64,710	**585,570**⋆	294,940	50,254.2
space_invaders	1,652	1,668.7	-	3,355	1,840	3,527	5,130	49,340⋆	**110,000**	18,789.0
venture	1,188	1,187.5	-	-	-	8,335	11,800	**28,600**†	-	1,107.0

Table 1: A comparison of human scores for 20 Atari games across datasets. The scores reported for (Hester et al. 2018; Kurin et al. 2017; de la Cruz, Du, and Taylor 2018) are the best human scores of each game. Mnih et al. and Wang et al. are average scores. The community world record is from Twin Galaxies, an official supplier of verified world records by Guinness World Records. Note that the display and game difficulty may vary slightly across platforms, here we try to find the game version that matches our setting to the best of our knowledge. For Atari-HEAD 2-hour performance, †: game terminated. ⋆: Two-hour experiment time limit has been reached before the game terminated. If the human players continue to play, they could potentially achieve higher scores. ‡: Maximum score allowed by the game reached. Disclaimer: Our human data is recorded in the semi-frame-by-frame mode discussed above and is intended to be used for research purposes, hence should not be submitted to the gaming community for competition.



(a) Montezuma's Revenge

(b) Seaquest

Figure 5: Example gaze prediction results. First column: game screenshots with red dots indicating the human gaze positions. Second column: biologically plausible retinal image, generated by foveated rendering algorithm (Perry and Geisler 2002). Third column: image saliency calculated by the classic Itti-Koch saliency model (Itti, Koch, and Niebur 1998). Fourth column: Farnebeck optical flow, calculated using the frame in the first column and its previous frame (Farnebäck 2003). Fifth column: predicted gaze distribution by convolution-deconvolution network, overlayed on top of the original image.

- **Kullback-Leibler Divergence (KL):** This metric is widely used to measure the difference between two probability distributions. It is also differentiable hence can be used as the loss function to train neural networks:

$$KL(P,Q) = \sum_i Q_i \log\left(\epsilon + \frac{Q_i}{\epsilon + P_i}\right) \qquad (2)$$

  ϵ is a small regularization constant and determines how much zero-valued predictions are penalized. KL is asymmetric and very sensitive to zero-valued predictions.

- **Pearson's Correlation Coefficient (CC):** between 0 and 1. It measures the linear relationship between two distributions.

$$CC(P,Q) = \frac{\sigma(P,Q)}{\sigma(P) \times \sigma(Q)} \qquad (3)$$

  where $\sigma(P,Q)$ denotes the covariance. CC is symmetric and penalizes false positives and negatives equally.

Note that KL and CC are distribution-based metrics, therefore the aforementioned process of converting discrete gaze positions to distributions is mandatory. However, for location-based metrics (AUC and NSS) the conversion is optional. Other usable metrics include Information Gain, Histogram Intersection, Shuffled AUC, Earth Mover's Distance. For a comprehensive survey about their properties, please see Bylinskii et al. (2018).

Baseline model and results We trained a convolution-deconvolution gaze network (Palazzi et al. 2018; Zhang et al. 2018b; 2018a; Deng et al. 2019) with KL divergence ($\epsilon = 1e - 10$) as loss function to predict human gaze positions. A separate network is trained for each game. We use 80% data for training and 20% for testing. The details of the network design can be found in Appendix Fig. 1(a).

Aggregated modeling results can be seen in Fig. 6. As expected, the learning-based neural network model outperforms optical flow and bottom-up saliency models by a large margin in all metrics. Note that we are only comparing to the basic version of these models. The prediction accuracy overall is high (average AUC of 0.971), although varies across games (min AUC: 0.945-Ms.Pacman, max: 0.988-Enduro). The predicted saliency maps can be visualized in Fig. 5.

The saliency prediction results using the dataset are considered highly accurate in saliency prediction research. One reason is the large amount of training data available provided by the dataset. Another reason is that the chosen tasks are reward-seeking and demanding, therefore human gaze is mostly directed towards image features that are strongly associated with reward and hence become highly predictable (Hayhoe and Ballard 2005; 2014).

Potential future work To further improve the prediction accuracy, researchers can optionally use additional inputs (motion, bottom-up saliency, or image semantics) along with the original images to predict human gaze positions. Several previous works have shown these signals are helpful for gaze prediction in visuomotor tasks (Zhang et al. 2018b; Palazzi et al. 2018; Li, Liu, and Rehg 2018). Another option would be using a recurrent neural network to represent information from past frames as memory, instead of stacking multiple images (Mnih et al. 2014; Xia et al. 2019; Zelinsky et al. 2019). Recurrent network models also allow one to model eye movements scanpaths as a sequence prediction problem.

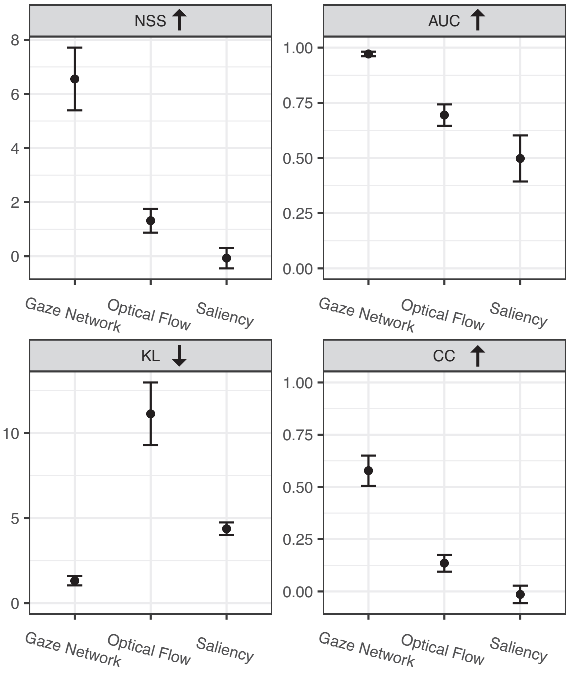


Figure 6: Gaze prediction results measured using four standard metrics, averaged across 20 games. As expected, a convolution-deconvolution network (gaze network) is able to predict human gaze much more accurately than motion-based and image saliency-based models. Error bars indicate standard deviation across games (N=20).

Imitation learning

Task definition The next task is to learn from human demonstrated actions. This standard IL problem is formulated as follows:

> Given a state s_t, learn to predict human action a_t, i.e,. learn $P(a|s)$, or equivalently, policy $\pi(s,a)$.

With human gaze data, we propose to use attention information to improve policy learning. This attention-guided learning problem is formulated as follows:

> Given a state s_t and human gaze positions g_t, learn to predict human action a_t, i.e., learn $P(a|s,g)$.

Another potential formulation is a joint learning problem:

> Given a state s_t, learn to jointly predict human action a_t and gaze positions g_t, i.e., learn $P(a,g|s)$.

Evaluation metrics

- **Behavior matching accuracy:** It measures the accuracy in predicting human actions. Since Atari games have a discrete action space (18 actions), one can treat the prediction task as a 18-way classification problem with standard log likelihood loss:

$$J = -\sum_{t}^{T}\sum_{a=0}^{17} \mathbb{1}_{a_t=a} \log P(a_t = a|s_t) \qquad (4)$$

  This supervised learning approach for imitation learning is commonly referred to as behavior cloning.

- **Game score:** The model that predicts human actions is effectively a gaming AI. Its performance can be directly measured by the final game score.

Note that the results on these two metrics may not necessarily be correlated, as we will show later.

Baseline model and results For standard IL, we trained a convolutional network using the classification loss above. To incorporate gaze information into IL, we use a two-channel policy network in Attention-Guided Imitation Learning (AGIL) (Zhang et al. 2018b). The policy network uses the saliency map predicted by the gaze network to mask the input image. This mask can be applied to the image to generate a "foveated" representation of the image that highlights the attended visual features (Li, Liu, and Rehg 2018; Zhang et al. 2017; 2018b; Xia et al. 2019; Chen et al. 2019). The design of both networks can be found in Appendix Fig. 1(b-c).

The performance measured by behavior matching accuracy can be seen in Fig. 7. The main result is that incorporating gaze information improves accuracy on all games with an average improvement of 7%. However, the magnitude of improvement varies across games. The games with most improvements are Name This Game (19%), Alien (19%), Seaquest (16%), Ms.Pacman (12%), Asterix (12%), and Frostbite (12%). These are games where many task-relevant objects appear on the screen simultaneously. As a result, the current behavioral target is often ambiguous without gaze information, therefore incorporating gaze leads to better prediction.

Next we look at game scores obtained by different models using different datasets, shown in Fig. 8 (additional statistics can be found in Appendix Table 1). We include IL (behavior cloning) results from two previous datasets (Hester et al. 2018; Kurin et al. 2017). The key observation is that the scale and quality of our data results in better performance compared to these datasets. More importantly, this second metric confirms again that gaze information is useful for IL. The AGIL model improves game performance on 19 games, with an average improvement of 115.26%.

Intuitively, knowing where humans look provides useful information on what action they take. Standard IL can only capture *what* the human teacher did, without knowing *why* the decision was made. Gaze is a good indicator of why a particular decision was made. Incorporating such information leads to better performance in both metrics.

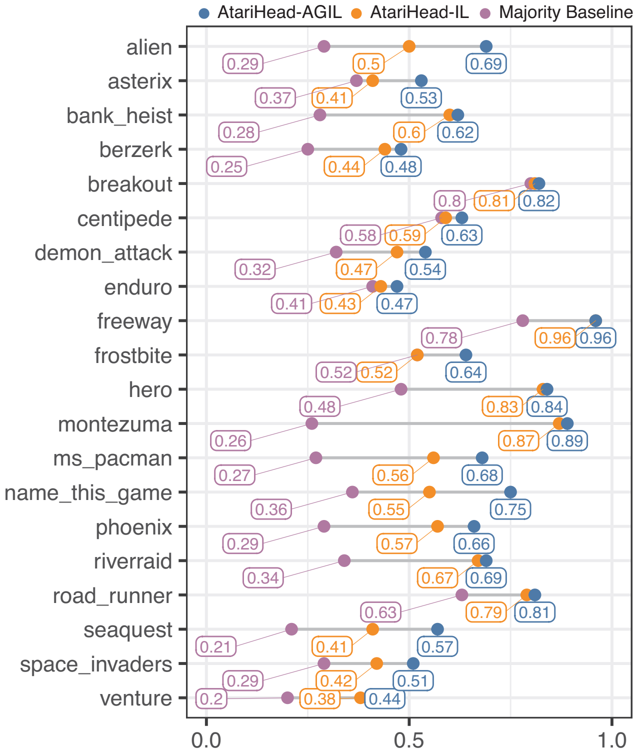


Figure 7: Behavior matching accuracy. Majority baseline simply predicts the majority class in that game (the most frequent action). IL: Standard behavior cloning. AGIL: policy network that includes saliency map predicted by the gaze network. Random guess: 0.06.

Potential future work As mentioned before, the results on the accuracy and score metrics may not necessarily be strongly correlated. For instance a 1% increase in accuracy leads to a 1138% improvement in scores for the game Breakout, while for Space Invaders, a 9% increase leads to minor improvement (0.45%) in game scores. In addition, it was found in experiments that adding dropout to the network improves human action prediction accuracy but hurts game performance, an issue worth further investigation.

There is still a large performance gap between the learning agent and the human scores reported in Table 1. Part of this is due to the inherent problems with behavior cloning such as covariate shift in state distribution (Ross, Gordon, and Bagnell 2011). Using IL and saliency prediction as auxiliary tasks (Hester et al. 2018; Jaderberg et al. 2016) or a pre-training step for RL (de la Cruz, Du, and Taylor 2018) are promising ways to improve game performance.

Discussion and Future Work

We introduce Atari-HEAD, a large-scale dataset of human demonstration playing Atari videos games. The novel features of this dataset include human gaze data, and a semi-frame-by-frame gameplay mode. The latter ensures that states and actions are matched, and allow enough decision

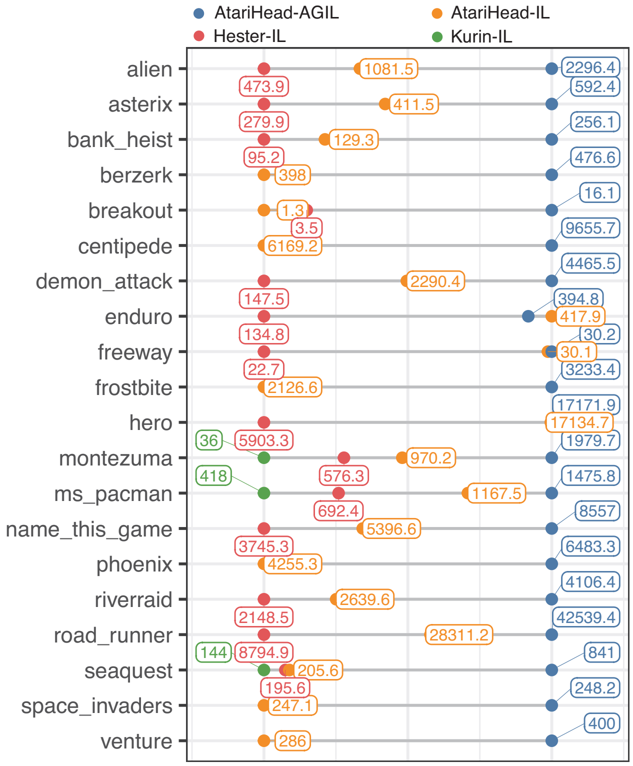


Figure 8: Mean scores of game agents. With this dataset, an IL agent is able to perform better than similar agents reported in previous datasets (Hester et al. 2018; Kurin et al. 2017). Additionally, incorporating the attention model learned from human gaze improves IL agent's performance by 115.26% on average.

time for human players. Software tools to extract image features such as optical flow are published along with the dataset, including a customized video player to visualize data.

This dataset addresses two major issues in IL and IR research; the first being reproducibility and the second being the need to bridge attention and control. We do this by providing human data that allows researchers to study how humans use visual attention to solve visuomotor tasks. We have shown promising results in saliency prediction and IL using Atari-HEAD. The most exciting result is that the human attention model improves the performance of the IL algorithm.

There are a number of promising future directions for research that arise from current progress. In this dataset, decision time is also recorded for every action. Such information could help identify difficult states for human players. Presumably these states should be weighted more during learning process, i.e., through state importance measurements (Li, Bulitko, and Greiner 2007). Another cue about state importance may come from the immediate reward we recorded for every decision.

A byproduct of the semi-frame-by-frame gameplay mode is human *option* (Sutton, Precup, and Singh 1999). We notice that human players often hold a key down until a subgoal is reached, then release the key and plan for the next sequence of actions. This naturally segments the decision trajectories into temporally extended actions, or options. It has yet to be explored whether a learning agent can learn from this type of human demonstrated options, but results from hierarchical imitation learning (Le et al. 2018) indicate that this may indeed be possible.

Although deep RL agents have achieved great performance on many tasks, researchers have attempted to understand these agents' behaviors through visualizing feature saliency maps learned by the deep network (Mousavi, Borji, and Mozayani 2016; Nikulin et al. 2019). It would be desirable to know whether the agents and humans pay attention to the same visual features. This would be a first step in a general understanding of what information humans use that AIs do not have access to. As progress is made in this direction those principles of attention can form the basis for further improvements in attention guided imitation learning.

Acknowledgement

We would like to thank Wilson Geisler's lab at UT-Austin for providing data collection tools. This work is supported by NIH grant EY05729. RZ is supported by The University of Texas at Austin Graduate Continuing Fellowship.

References

Alletto, S.; Palazzi, A.; Solera, F.; Calderara, S.; and Cucchiara, R. 2016. Dr (eye) ve: a dataset for attention-based tasks with applications to autonomous and assisted driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 54–60.

Bellemare, M. G.; Naddaf, Y.; Veness, J.; and Bowling, M. 2012. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research.*

Borji, A., and Itti, L. 2015. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581.*

Bylinskii, Z.; Judd, T.; Borji, A.; Itti, L.; Durand, F.; Oliva, A.; and Torralba, A. Mit saliency benchmark.

Bylinskii, Z.; Judd, T.; Oliva, A.; Torralba, A.; and Durand, F. 2018. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence* 41(3):740–757.

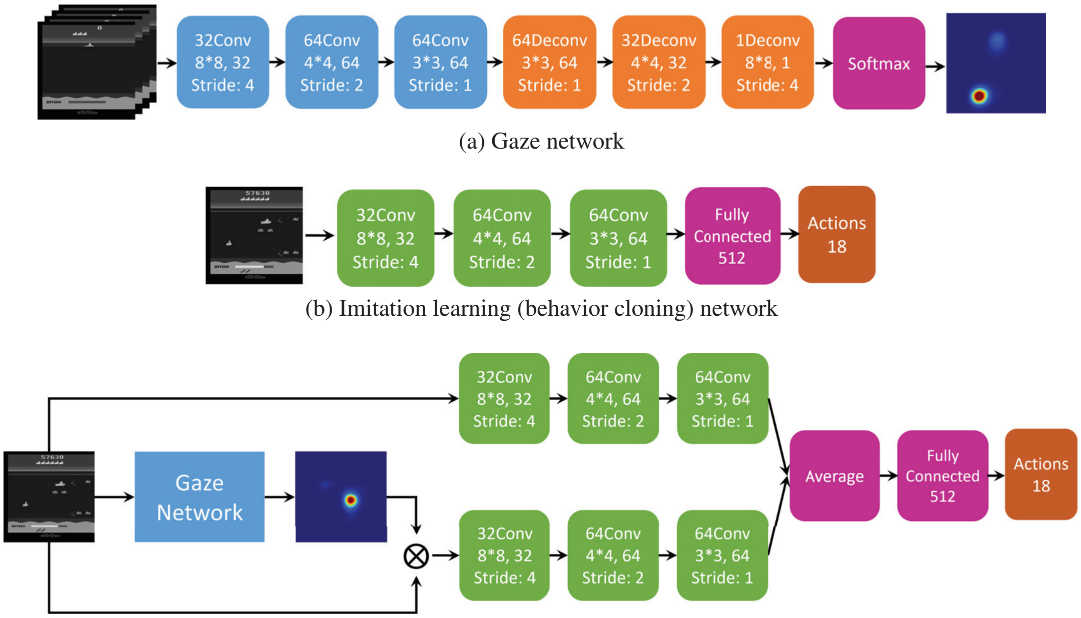
Chen, Y.; Liu, C.; Tai, L.; Liu, M.; and Shi, B. E. 2019. Gaze training by modulated dropout improves imitation learning. *arXiv preprint arXiv:1904.08377.*

de la Cruz, G. V.; Du, Y.; and Taylor, M. E. 2018. Pre-training with non-expert human demonstration for deep reinforcement learning. *arXiv preprint arXiv:1812.08904.*

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deng, T.; Yan, H.; Qin, L.; Ngo, T.; and Manjunath, B. 2019. How do drivers allocate their potential attention? driving fixation prediction via convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems.*

Farnebäck, G. 2003. Two-frame motion estimation based on polynomial expansion. *Image analysis* 363–370.



(a) Gaze network

(b) Imitation learning (behavior cloning) network

(c) Attention-guided imitation learning (AGIL) network

Figure 1: Appendix (a) The gaze prediction network. The network takes in a stack of 4 consecutive game images in grayscale, passes the inputs to 3 convolutional layers followed by 3 deconvolutional layers. The final output is a gaze saliency map that indicates the predicted probability distribution of the gaze. (b) An imitation learning (behavior cloning) network to predict human actions. The network takes in a single grayscale game image as input, and outputs a vector that gives the probability of each action. (c) The policy network architecture for imitating human actions. The top channel takes in the current image frame and the bottom channel takes in the masked image which is an element-wise product of the original image and predicted gaze saliency map by the gaze network. We then average the output of the two channels.

Games	Kurin-IL	Hester-IL	AtariHead-IL	AtariHead-AGIL	Improvement
alien	-	473.9	1081.5 ± 741.8	2296.4 ± 1105.7	+112.33%
asterix	-	279.9	411.5 ± 192.6	592.4 ± 290.5	+43.96%
bank_heist	-	95.2	129.3 ± 75.8	256.1 ± 116.8	+98.07%
berzerk	-	-	398.0 ± 189.4	476.6 ± 197.4	+19.75%
breakout	-	3.5	1.3 ± 1.4	16.1 ± 22.5	+1138.46%
centipede	-	-	6169.2 ± 3856.1	9655.7 ± 5782.8	+56.51%
demon_attack	-	147.5	2290.4 ± 1806.7	4465.5 ± 2603.6	+94.97%
enduro	-	134.8	417.9 ± 91.4	394.8 ± 71.2	-5.53%
freeway	-	22.7	30.1 ± 1.2	30.2 ± 1.0	+0.33%
frostbite	-	-	2126.6 ± 1444.3	3233.4 ± 1857.5	+52.05%
hero	-	5903.3	17134.7 ± 6454.5	17171.9 ± 8939.8	+0.22%
montezuma	36 ± 8.0	576.3	970.2 ± 896.2	1979.7 ± 1291.7	+104.05%
ms_pacman	418 ± 20.0	692.4	1167.5 ± 686.9	1475.8 ± 858.5	+26.41%
name_this_game	-	-	5396.6 ± 1757.0	8557.0 ± 2015.6	+58.56%
phoenix	-	3745.3	4255.3 ± 1967.8	6483.3 ± 3051.5	+52.36%
riverraid	-	2148.5	2639.6 ± 669.3	4106.4 ± 1457.1	+55.57%
road_runner	-	8794.9	28311.2 ± 7261.8	42539.4 ± 11177.2	+50.26%
seaquest	144 ± 12.4	195.6	205.6 ± 103.7	841.0 ± 842.1	+309.05%
space_invaders	-	-	247.1 ± 149.2	248.2 ± 147.1	+0.45%
venture	-	-	286.0 ± 146.8	400.0 ± 175.4	+39.86%

Table 1: Appendix Game scores (mean ± standard deviation) of game agents. Kurin-IL and Hester-IL are imitation learning results reported in (Kurin et al. 2017) and (Hester et al. 2018). Applying IL and AGIL (Zhang et al. 2018b) to our dataset, the mean scores are averaged over 500 episodes per game, with each episode initialized with a randomly generated seed. The game is cutoff after 108K frames. The agent chooses an action a probabilistically using a softmax function with Gibbs (Boltzmann) distribution according to policy network's prediction $P(a)$: $\pi(a) = \frac{\exp(\eta P(a))}{\sum_{a' \in \mathcal{A}} \exp(\eta P(a'))}$ where $\mathcal{A}$ denotes the set of all possible actions, $\exp(.)$ denotes the exponential function, and the temperature parameter η is set to 1. The scale and quality of our data leads to better performance, when comparing to AtariHEAD-IL to Kurin-IL and Hester-IL. The AtariHead-AGIL agent first learns to predict human gaze and uses the learned gaze model to guide the process of learning human decisions. Incorporating attention leads to an average improvement of 115.26% over a standard IL algorithm using our dataset.

Guss, W. H.; Houghton, B.; Topin, N.; Wang, P.; Codel, C.; Veloso, M.; and Salakhutdinov, R. 2019. Minerl: A large-scale dataset of minecraft demonstrations. *arXiv preprint arXiv:1907.13440*.

Hayhoe, M., and Ballard, D. 2005. Eye movements in natural behavior. *Trends in cognitive sciences* 9(4):188–194.

Hayhoe, M., and Ballard, D. 2014. Modeling task control of eye movements. *Current Biology* 24(13):R622–R628.

Hessel, M.; Modayil, J.; Van Hasselt, H.; Schaul, T.; Ostrovski, G.; Dabney, W.; Horgan, D.; Piot, B.; Azar, M.; and Silver, D. 2018. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Hester, T.; Vecerik, M.; Pietquin, O.; Lanctot, M.; Schaul, T.; Piot, B.; Horgan, D.; Quan, J.; Sendonaris, A.; Osband, I.; et al. 2018. Deep q-learning from demonstrations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Itti, L., and Koch, C. 2001. Computational modelling of visual attention. *Nature reviews neuroscience* 2(3):194.

Itti, L.; Koch, C.; and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence* 20(11):1254–1259.

Jaderberg, M.; Mnih, V.; Czarnecki, W. M.; Schaul, T.; Leibo, J. Z.; Silver, D.; and Kavukcuoglu, K. 2016. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*.

Jiang, M.; Huang, S.; Duan, J.; and Zhao, Q. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1072–1080.

Johnson, L.; Sullivan, B.; Hayhoe, M.; and Ballard, D. 2014. Predicting human visuomotor behaviour in a driving task. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 369(1636):20130044.

Kurin, V.; Nowozin, S.; Hofmann, K.; Beyer, L.; and Leibe, B. 2017. The atari grand challenge dataset. *arXiv preprint arXiv:1705.10998*.

Le, H.; Jiang, N.; Agarwal, A.; Dudik, M.; Yue, Y.; and Daumé, H. 2018. Hierarchical imitation and reinforcement learning. In *International Conference on Machine Learning*, 2923–2932.

Le Meur, O., and Baccino, T. 2013. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods* 45(1):251–266.

Li, L.; Bulitko, V.; and Greiner, R. 2007. Focus of attention in reinforcement learning. *Journal of Universal Computer Science* 13(9):1246–1269.

Li, Y.; Liu, M.; and Rehg, J. M. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 619–635.

Liu, C.; Chen, Y.; Tai, L.; Ye, H.; Liu, M.; and Shi, B. E. 2019. A gaze model improves autonomous driving. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 33. ACM.

Mandlekar, A.; Zhu, Y.; Garg, A.; Booher, J.; Spero, M.; Tung, A.; Gao, J.; Emmons, J.; Gupta, A.; Orbay, E.; et al. 2018. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. *arXiv preprint arXiv:1811.02790*.

Mnih, V.; Heess, N.; Graves, A.; et al. 2014. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, 2204–2212.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.

Mousavi, S.; Borji, A.; and Mozayani, N. 2016. Learning to predict where to look in interactive environments using deep recurrent q-learning. *arXiv preprint arXiv:1612.05753*.

Nikulin, D.; Ianina, A.; Aliev, V.; and Nikolenko, S. 2019. Free-lunch saliency via attention in atari agents. *arXiv preprint arXiv:1908.02511*.

Palazzi, A.; Abati, D.; Calderara, S.; Solera, F.; and Cucchiara, R. 2018. Predicting the driver's focus of attention: the dr (eye) ve project. *IEEE transactions on pattern analysis and machine intelligence*.

Perry, J. S., and Geisler, W. S. 2002. Gaze-contingent real-time simulation of arbitrary visual fields. In *Electronic Imaging 2002*, 57–69. International Society for Optics and Photonics.

Ross, S.; Gordon, G. J.; and Bagnell, D. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, 627–635.

Rudoy, D.; Goldman, D. B.; Shechtman, E.; and Zelnik-Manor, L. 2013. Learning video saliency from human gaze using candidate selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1147–1154.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of go with deep neural networks and tree search. *Nature* 529(7587):484–489.

Sutton, R. S.; Precup, D.; and Singh, S. 1999. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112(1):181–211.

Wang, Z.; Schaul, T.; Hessel, M.; Hasselt, H.; Lanctot, M.; and Freitas, N. 2016. Dueling network architectures for deep reinforcement learning. In *International Conference on Machine Learning*, 1995–2003.

Xia, Y.; Kim, J.; Canny, J.; Zipser, K.; and Whitney, D. 2019. Periphery-fovea multi-resolution driving model guided by human attention. *arXiv preprint arXiv:1903.09950*.

Yu, F.; Xian, W.; Chen, Y.; Liu, F.; Liao, M.; Madhavan, V.; and Darrell, T. 2018. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*.

Zelinsky, G.; Yang, Z.; Huang, L.; Chen, Y.; Ahn, S.; Wei, Z.; Adeli, H.; Samaras, D.; and Hoai, M. 2019. Benchmarking gaze prediction for categorical visual search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Zhang, R.; Liu, Z.; Hayhoe, M. M.; and Ballard, D. H. 2017. Attention guided deep imitation learning. In *Cognitive Computational Neuroscience (CCN)*.

Zhang, L.; Zhang, R.; Liu, Z.; Hayhoe, M. M.; and Ballard, D. H. 2018a. Learning attention model from human for visuomotor tasks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhang, R.; Liu, Z.; Zhang, L.; Whritner, J. A.; Muller, K. S.; Hayhoe, M. M.; and Ballard, D. H. 2018b. Agil: Learning attention from human for visuomotor tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 663–679.

Zhang, R.; Torabi, F.; Guan, L.; Ballard, D. H.; and Stone, P. 2019. Leveraging human guidance for deep reinforcement learning tasks. In *Twenty-Eighth International Joint Conference on Artificial Intelligence*.