

Human Gaze Assisted Artificial Intelligence: A Review

Ruohan Zhang*, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo,
Scott Niekum, Dana Ballard and Mary Hayhoe

The University of Texas at Austin

{zharu,yifeng.zhu,sguo19,danab,hayhoe}@utexas.edu, {asaran, bliu,sniekum}@cs.utexas.edu

Abstract

Human gaze reveals a wealth of information about internal cognitive state. Thus, gaze-related research has significantly increased in computer vision, natural language processing, decision learning, and robotics in recent years. We provide a high-level overview of the research efforts in these fields, including collecting human gaze data sets, modeling gaze behaviors, and utilizing gaze information in various applications, with the goal of enhancing communication between these research areas. We discuss future challenges and potential applications that work towards a common goal of human-centered artificial intelligence.

1 Introduction

Humans are surrounded by a complex world full of information. How do humans survive without being overwhelmed? There are often hundreds to thousands of objects and other kinds of information within view, but our sensory and cognitive capacities are limited. Fortunately, not all objects or information matters for our current agenda or long-term goal of survival. Through evolution and learning, humans have gradually developed strategies for selecting information. This is referred to as *selective attention*. As artificial intelligence (AI) migrates from a simple digital world to the complex real world, the same challenge awaits AI agents: How do they select important information from a world full of information? A given computational model, either biological or digital, has limited capacity. Therefore attentional selection is necessary to ensure that resources are devoted to the key components.

Because humans actively seek the information they need, gaze can reveal the underlying attentional patterns [Posner and Petersen, 1990]. Humans have high acuity foveal vision in the central 1-2 visual degrees of the visual field (i.e., covering the width of a finger at arm’s length), with resolution decreasing in the periphery. They have learned to move their foveae to the correct place at the right time to process important task-relevant visual stimuli [Borji and Itti, 2014; Hayhoe, 2017]. This type of selective attentional mechanism

developed through evolution and is refined in a lifelong learning process. Given the amount of training data required during this process, it may be easier for AI agents to learn attention directly from human gaze data. Fortunately, human gaze is one of the most cost-efficient types of physiological data that can be collected in large quantities, as a result of progress in eye-tracking hardware and software. The vision science research community has a long history of studying human gaze behaviors; hence, such behaviors are relatively well understood. Because of these reasons, training AI agents using human gaze has become a viable approach.

Another concern of using human gaze in AI research arises as artificial agents and robots become more prevalent in human society—the importance of making AI agents understand human intentions and goals cannot be overestimated. In many scenarios, AI agents need to gather information about their human fellows to facilitate mutual understanding and coordination. Primates’ social gaze conveys information about their dispositions, intentions, beliefs, emotions, and other cognitive and emotional states [Emery, 2000]. The ability to perceive gaze is critical in learning and social interactions [Emery, 2000].

We have briefly discussed two reasons to include human gaze in AI research, including (1) AI must develop an attention mechanism to cope with the information-rich world and this mechanism can be learned from human gaze data, and (2) AI agents need to perceive and understand human gaze to better interact with humans. Motivated by these reasons, multiple fields of AI, including computer vision, natural language processing, imitation and reinforcement learning, as well as robotics, have started the effort of building human gaze-assisted AI agents. Many state-of-the-art results can only be achieved with human gaze information, especially in realistic complex task domains. In this survey, we review relevant studies in these four research areas that work towards this common goal in the past five years. We further provide a brief overview of modern eye-tracking software technologies that allow for more accurate and accessible tracking results.

2 Gaze in Computer Vision

In order to understand how attention is controlled when viewing natural scenes, vision scientists first explored what image properties or features capture the human gaze. Similarly, computer vision engineers have tried to extract impor-

*Contact Author

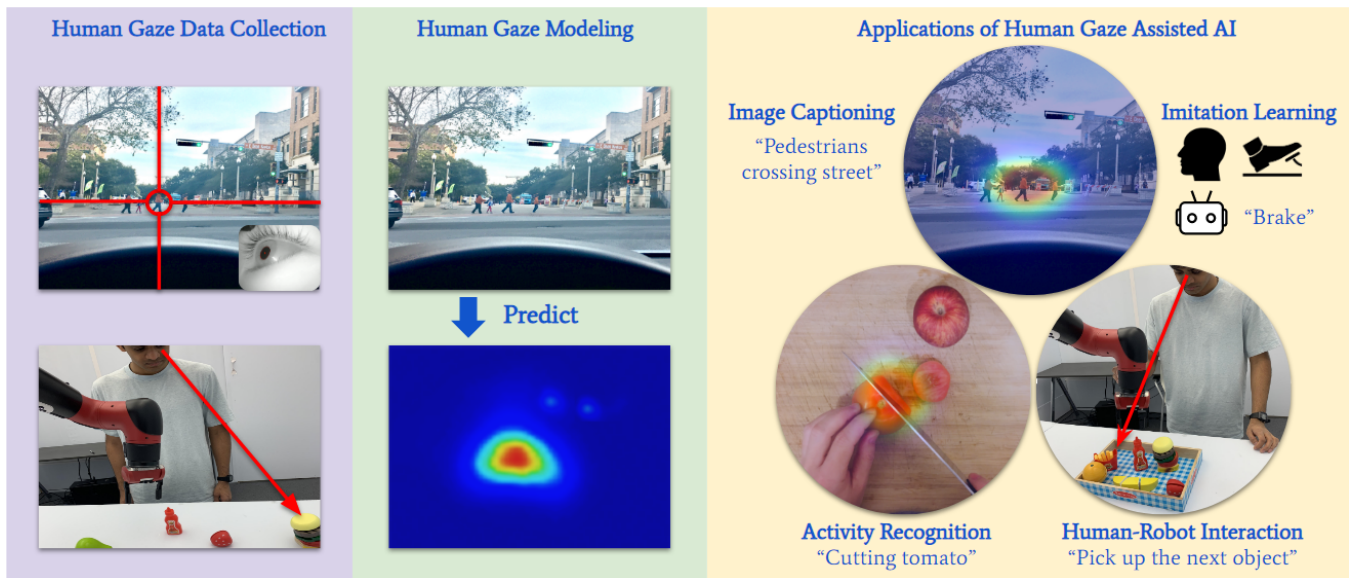


Figure 1: Process of a typical human gaze assisted AI research. The process starts with gathering human gaze data using eye trackers then building models to predict human attention distribution. The human gaze data and models can benefit various AI research fields.

tant (“salient”) visual features from images. Their combined interests led to a large body of research concerning visual saliency. A common goal here has been to develop computational models that can predict the human gaze given visual images, where gaze is often treated as the ground truth to indicate salient features.

Two different approaches are commonly taken to build saliency models: hypothesis-driven and data-driven. Vision scientists and early computer vision researchers mainly used the former approach [Itti *et al.*, 1998]. For instance, the classic work of Itti *et al.*[1998] hypothesized that features derived from image statistics, such as color, intensity, and orientation, capture human gaze. It follows that human gaze data can be used to validate these hypotheses.

In recent years, data-driven approaches became more popular as large-scale eye-tracking datasets became available for images [Papadopoulos *et al.*, 2014; Li *et al.*, 2014; Xu *et al.*, 2014; Bylinskii *et al.*, 2015b; Bylinskii *et al.*, 2015a; Krafcik *et al.*, 2016], videos [Mathe and Sminchisescu, 2014; Wang *et al.*, 2018], and 360-degree videos [Zhang *et al.*, 2018b; Xu *et al.*, 2018]. Eye-tracking devices and software are expensive. Hence researchers have often used alternative methods as surrogates for gaze data, such as mouse tracking [Jiang *et al.*, 2015]. When combined with deep neural networks, data-driven saliency approaches have achieved tremendous progress. Typical saliency networks are convolutional neural networks [Jetley *et al.*, 2016; Kümmerer *et al.*, 2016; Kruthiventi *et al.*, 2017] or convolutional long short-term memory (LSTM) networks [Cornia *et al.*, 2018b]. In practice, one can try to directly predict discrete human gaze positions. Alternatively, it is common to convert discrete gaze positions into a continuous distribution to account for the uncertainty in tracking and modeling. The model should learn to predict the discrete positions or converted probability distribution given the image. This can be done using supervised learning where

several distance metrics can be used as the loss function for training [Bylinskii *et al.*, 2019].

Visual saliency is a well-developed field compared to other emerging ones we are about to discuss. We direct interested readers to recent review papers on the topics of saliency evaluation metrics [Bylinskii *et al.*, 2019], saliency model performance analyses [Bylinskii *et al.*, 2016; He *et al.*, 2019a] and a closely related field called salient object detection [Borji *et al.*, 2015]. These saliency research studies typically model the gaze of an observer *looking at images or videos*. Alternatively, a related line of research named gaze following models the gaze of people *inside images or videos* [Recasens *et al.*, 2015; Recasens *et al.*, 2017].

Saliency models have a wide range of applications in computer vision, graphics, and multimedia. Most of these applications are human-centered. In computer vision, ground truth labels of recognition and detection tasks are often provided by humans. As an example indicating how informative human gaze can be, Karessli *et al.*[2017] showed that human gaze patterns are class discriminative, so that gaze features can be directly used for image classification. In computer graphics and multimedia, many image manipulations such as rendering and compression, must address the need of human users. Saliency-driven manipulations address human demands and, at the same time, reduce the computation burden by selecting only a few image regions to process. For a recent survey on these applications, please see Nguyen *et al.*[2018].

Traditional saliency prediction does not involve active tasks. Datasets have typically been collected by asking human participants simply to look at static images or videos, called free-viewing, and only the gaze data is recorded and modeled. This approach was thought to capture so-called *bottom-up* attention which assumes that attention is driven by visual stimulus. More recent work by Henderson *et al.*[2018], however, suggests that viewers are trying to extract scene

meaning. Salient stimulus features such as contrast correlate highly with meaning, and meaning can explain more of the variance after taking this correlation into account. Thus the free-viewing task might reflect this basic visual process of extracting scene meaning. On the other hand, it is well known that human attention is strongly modulated by *top-down* signals especially when engaged in an explicit task. Progress in bottom-up, stimulus-driven visual saliency research has laid the ground for further research in top-down, task-driven research. Researchers have moved from gaze data collected while passively viewing images and videos to those collected while actively performing a wide range of daily tasks, such as conversation, driving, gaming, social interaction, etc. We will now discuss these types of gaze data.

3 Gaze in Language Tasks

We now consider language learning tasks that involve visual stimuli. The association between human gaze and language has been established since infants have learned the name of an object for the very first time from their caregivers (known as the word-referent association [Yu and Smith, 2011]). Indeed, artificial language learners face a similar challenge as infants do in vision-language learning tasks. Given a complex visual scene and a verbal description, it is unclear which language element refers to which visual entity without prior knowledge. This issue is particularly challenging for modern end-to-end, data-driven learning approaches. Human infants solve the referent problem by following their teachers' gaze and such a gaze-following strategy was shown to be strongly correlated with language learning scores [Brooks and Meltzoff, 2005]. If AI agents are provided the human teacher's gaze that makes the word-referent association clear, the learning task could be simplified.

Consequently, vision and natural language processing (NLP) researchers have recently utilized human gaze data as part of language learning tasks. Multiple datasets of images [Yun *et al.*, 2013; van Miltenburg *et al.*, 2018; Vaidyanathan *et al.*, 2018; Balajee Vasudevan *et al.*, 2018; He *et al.*, 2019b] and videos [Yu *et al.*, 2017] with paired gaze and verbal description data have been made publicly available. As expected, incorporating human gaze information leads to significant improvements in identifying the referred object from all proposals (object referring) [Balajee Vasudevan *et al.*, 2018], generating descriptive captions for images [Sugano and Bulling, 2016; Tavakoli *et al.*, 2017; Cornia *et al.*, 2018a; Chen and Zhao, 2018; He *et al.*, 2019b] and videos [Yu *et al.*, 2017], as well as visual question answering [Qiao *et al.*, 2018].

It was found that the attention maps of neural network models trained without human gaze on these language tasks are different from human attention [Das *et al.*, 2017; Tavakoli *et al.*, 2017; He *et al.*, 2019b]. Understanding and quantifying such differences may provide insights on the performance, especially failure, of current vision-language models. For these models, the ground truth label-verbal annotations are provided by humans, so it is indeed necessary to infer underlying human cognition, such as object referrals, through human gaze. The usefulness of gaze information should be even

more evident when AI agents meant for NLP are deployed to interact with humans in daily conversations—a hypothesis that could be tested in robotic dialogue systems, for example.

4 Gaze in Decision-Making Tasks

In addition to language, another common type of task humans perform on a daily basis is visuomotor decision making, ranging from simple behaviors like walking to sophisticated behaviors such as cooking and driving. One goal of AI research is to develop autonomous machines that can perform these tasks. A common approach to achieve this goal is to make machines act like humans, by training machines to recognize and then imitate human teachers' actions—an approach known as learning from demonstrations (LfD) or imitation learning [Argall *et al.*, 2009].

In the LfD paradigm, human data is typically presented in the form of state-action pairs, where a *state* encodes relevant information for decision making from the environment. The goal of the learning agent is to learn the state-action mapping so it can recognize human activity or perform the task on its own. Learning such mapping is made difficult by the fact that the state-action pairs do not provide enough information and leave ambiguity about the demonstrator's policy or intent. For example, in training an autonomous driving agent to imitate human driving behaviors, it needs to know that the human driver slows down because a pedestrian has appeared. Here, human gaze reveals *why* a particular decision is made. In this sense, changes in gaze positions may also imply task switching or current behavioral target changing. It was proposed that human gaze can be used as an auxiliary guiding signal in the imitation learning paradigm [Zhang *et al.*, 2019a].

We have discussed how one may formalize a task-independent gaze prediction problem as saliency prediction in Section 2. The question remains whether one can model gaze in visuomotor tasks using a similar approach. In recent years, researchers have collected human gaze and action data in meal preparation [Li *et al.*, 2018], human-to-human (non-verbal) interactions [Zuo *et al.*, 2018], driving [Palazzi *et al.*, 2018], and video game playing [Zhang *et al.*, 2019b]. Convolutional neural networks remain the most popular tool for predicting human attention [Li *et al.*, 2018; Zhang *et al.*, 2018a; Palazzi *et al.*, 2018; Deng *et al.*, 2019]. Since the chosen tasks are reward-seeking and cognitively demanding, human gaze is mostly directed towards visual areas that are strongly associated with reward and hence become highly predictable. Not surprisingly, motion features play a more important role in a task-driven case than traditional image features [Zhang *et al.*, 2018a]. A notable challenge here is *egocentric* gaze prediction in which the spatial distribution of the gaze is highly biased towards the image center, a problem further addressed by [Palazzi *et al.*, 2018; Tavakoli *et al.*, 2019].

Being able to model human gaze allows researchers to further investigate whether the gaze information can indeed help agents better learn from human demonstrated actions. To incorporate human attention into action learning, one can treat the predicted gaze distribution of an image as a filter or a mask. This mask can be applied to the image to generate a

representation of the image that highlights the attended visual features. Experimental results have shown that including gaze information leads to higher accuracy in recognizing or predicting human actions, in reaching [Ravichandar *et al.*, 2018], human-to-human interaction [Zuo *et al.*, 2018], driving [Xia *et al.*, 2018; Liu *et al.*, 2019], meal preparation [Li *et al.*, 2018; Shen *et al.*, 2018; Sudhakaran *et al.*, 2019], and video game playing [Zhang *et al.*, 2018a].

An AI agent that has learned both the attention and decision models from humans can perform the task on its own. It has been shown that incorporating a learned gaze model leads to a large performance increase in video games [Zhang *et al.*, 2018a]. For real-world tasks like autonomous driving, it is reasonable to expect a similar improvement when incorporating human attention models. Due to physical constraints and safety reasons, this is yet to be explored but preliminary tests in simulated environments are possible.

The gaze and action datasets in visuomotor decision tasks also provide an opportunity for seeking a deeper understanding of why humans make certain decisions. The gaze is a necessary component in closing the perception-cognition-action loop. For instance, an approach called inverse reinforcement learning (IRL) infers a human’s internal reward function which explains their actions. Since human gaze is closely associated with the task reward [Hayhoe, 2017], a good reward function should also be able to explain human gaze behaviors. In this case, it is desirable to model gaze (perception), reward (cognition), and action in a joint model.

5 Gaze in Robotics

As robots, especially assistive robots, become more prevalent in our daily life, interaction and communication between robots and humans certainly have increased. Human-robot interaction (HRI) research aims to enhance such interaction and communication, and shows that they can be facilitated by the sensitivity to human physiological signals, such as human gaze. We will review recent progress in robotics that utilize human or robot gaze in HRI settings. For earlier work on this topic, we direct interested readers to two previous survey papers [Ruhland *et al.*, 2015; Thomaz *et al.*, 2016].

Unlike vision, language, and decision learning tasks where gaze data is collected in advance, HRI requires robots to acquire human gaze during the interaction. In an ideal setting, a robot and its human partner are both equipped with egocentric cameras, and the human is further equipped with an eye tracker. The robot has direct access to human camera and gaze data, from which it calculates the human’s gaze vector in the robot’s coordinate system [Penkov *et al.*, 2017]. Perhaps a more common but more challenging setting is that humans do not wear a camera nor an eye tracker, and the robot needs to estimate the human gaze vector by looking at their faces [Amos *et al.*, 2016; Saran *et al.*, 2018]. A rough estimate can be computed from the human body and head orientation but this was shown to be much less informative than direct gaze measuring [Palinko *et al.*, 2016].

Once human gaze information is obtained, the next challenge is to interpret the meaning of the gaze. Social gaze

between humans is relatively well studied, and a similar effort has been made for understanding human gaze when interacting with robots [Rich *et al.*, 2010]. The interpretation of human gaze and its benefits are highly context-dependent. Humans and robots engage in various forms of interaction tasks. Similar to decision learning tasks discussed in the previous section, human gaze can facilitate robot learning during teaching [Penkov *et al.*, 2017; Saran *et al.*, 2019]. In a reversed setting, intelligent tutoring systems can monitor a human student’s gaze to infer her mental or emotional state to encourage better engagement [Jaques *et al.*, 2014; Hutt *et al.*, 2016]. Intention-revealing gaze enhances collaboration in object referring [Fang *et al.*, 2015], teleoperation [Yu *et al.*, 2014], shared autonomy [Aronson *et al.*, 2018], collaborative manipulation [Huang and Mutlu, 2016], and assisted reaching and grasping [Shafti *et al.*, 2019]. Human gaze can also help a robot infer the recipient of human verbal communication in a multi-party scenario [Richter *et al.*, 2016].

In the effort of humanizing robots, anthropomorphic humanoid robots can use their own “gaze” to enhance communication with humans [Admoni and Scassellati, 2017]. Robot gaze can resolve object referring [Admoni *et al.*, 2016], communicate intended actions to make interactions more fluent [Moon *et al.*, 2014], effectively manage the conversational floor with humans [Andrist *et al.*, 2014], encourage humans to be more compliant [Admoni *et al.*, 2014], and improve a human teacher’s estimate of the robot learner’s understanding and the human’s teaching strategy [Huang *et al.*, 2019]. But designing robot gaze itself is challenging, at least one study suggested that robotic gaze cues alone have no significant impact on humans in certain scenarios [Fiore *et al.*, 2013].

However, reproducibility is a challenge in HRI studies. Unlike in vision, language, and decision-making tasks, robotics tasks are in general difficult to standardize and benchmark, especially when humans are involved, due to the variations in physical robots and human participants. Another challenge is to make human-robot gaze communication bidirectional [Andrist *et al.*, 2017] and make robot gaze behaviors adaptive to different task settings and users.

6 New Tracking Algorithms

Finally, we briefly review modern eye-tracking technologies that are the foundations for many of the research works discussed above. Modern eye trackers range from desktop trackers that have high spatio-temporal resolution used for psychophysics studies, to wearable trackers that can be mounted on glasses, or even webcams. They differ in tracking accuracy, portability, and cost. Therefore, a wide variety of eye-tracking hardware is made for different applications.

We have discussed how eye-tracking technology can benefit artificial intelligence research. The reverse is also true. Recent progress in computer vision has improved eye tracker accuracy and portability by a significant margin. Appearance-based algorithms using convolutional neural networks have been shown to have better tracking accuracy and are more robust to visual appearance variations [Zhang *et al.*, 2015; Wood *et al.*, 2015; Krafka *et al.*, 2016; Shrivastava *et al.*, 2017; Zhang *et al.*, 2017; Park *et al.*, 2018], compared to

more traditional approaches like hand-crafted feature-based or model-based algorithms. Advanced tracking software has allowed real-time eye tracking on low-cost devices such as webcams [Papoutsaki *et al.*, 2016] and mobile tablets and phones [Huang *et al.*, 2017; Krafka *et al.*, 2016]. Due to this progress, collecting human gaze data along with other forms of human data is now feasible. This is a main reason for the emerging research applications we have discussed.

7 Discussion

We have seen that human gaze can benefit vision, language, decision-making, and robotics research. The main reason for the successes in these fields is the effort of collecting and publishing large-scale high-quality eye-tracking datasets. These datasets are fundamental for modern data-driven research. Another driving force is the progress in machine learning research, especially deep neural networks.

We have also seen how gaze reveals different information in various contexts. In vision tasks, gaze indicates visual features that are generally attractive for humans. In language, gaze helps resolve the word-referent problem. In decision-making tasks, gaze bridges perception and decision-making by indicating the current behavioral target. In robotics, social cues revealed by human or robot gaze facilitate communication and enhance collaboration.

Human gaze information is commonly used in three ways: as an additional channel of information, as a mask on the input to filter out unimportant information, or as a secondary optimization objective. For example, in training a neural network, the above methods correspond to concatenating a gaze map with the input image, masking the input image with the gaze map, and adding gaze prediction as an auxiliary loss term in the objective function, respectively.

We now discuss a few important future research directions.

Human vs. AI attention. AI agents can learn to develop their own attention mechanism which is the key component of many state-of-the-art models [Mnih *et al.*, 2014; Vaswani *et al.*, 2017]. Such a mechanism is often a byproduct of the main learning objective. We can ask at least three questions. First, given the same task and learning objective, does machine learn an attention that is different from humans? Second, if they do differ, which one is more preferred under different conditions? In word-referent association learning tasks human attention is preferred, but what about decision-making tasks especially in which AI agents outperform humans? Could human attention be biased and fail to capture the correct information? Third, if human attention is preferred, how should we incorporate human gaze information into the learning procedure of these machines? Answering these questions can help us better understand the differences between human and machine attention.

Individual differences. A frequently overlooked issue in many studies is individual differences in human gaze behaviors: Given the same visual stimulus, humans may pay attention to different visual entities. Sometimes it is necessary to consider the variability in collected human data. Researchers need to carefully trace the roots of such variability and consider whether to build models to account for this variability.

For example, two distinct gaze distributions may indicate that the two humans are engaged in different tasks and pursuing different behavioral goals. In HRI settings, individual differences require robots to adapt to gaze behaviors on-line for different users.

Assisting humans. Most works we have discussed utilize human gaze to assist AI agents. It is possible in the future that attentive AI systems could assist humans in cognitively demanding tasks. One prototype application is advanced driver-assistance systems (ADAS) that monitors the driver's gaze that is mainly used for fatigue or distraction detection nowadays. We may foresee that ADAS one day could build a gaze model of focused expert drivers, and it could monitor and alert its current driver if abnormal gaze behaviors are detected. Furthermore, for humans with motor or language impairments, their gaze is one of the most important remaining communication channels. AI agents that are built with the ability to perceive and understand their gaze behaviors could better infer their needs, and be able to better assist them in performing daily tasks. However, research in this direction is still limited [Betke, 2010].

Acknowledgements

Part of this work is supported by NIH grant EY 05729. Part of this work has taken place in the Personal Autonomous Robotics Lab (PeARL) at UT Austin. PeARL research is supported in part by the NSF (IIS1724157, IIS-1638107, IIS-1617639, IIS-1749204) and ONR (N00014-18-2243).

References

- [Admoni and Scassellati, 2017] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017.
- [Admoni *et al.*, 2014] Henny Admoni, Anca Dragan, Siddhartha S Srinivasa, and Brian Scassellati. Deliberate delays during robot-to-human handovers improve compliance with gaze communication. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 49–56, 2014.
- [Admoni *et al.*, 2016] Henny Admoni, Thomas Weng, and Brian Scassellati. Modeling communicative behaviors for object references in human-robot interaction. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3352–3359. IEEE, 2016.
- [Amos *et al.*, 2016] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. 2016.
- [Andrist *et al.*, 2014] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. Conversational gaze aversion for human-like robots. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 25–32. IEEE, 2014.
- [Andrist *et al.*, 2017] Sean Andrist, Michael Gleicher, and Bilge Mutlu. Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 2571–2582, 2017.
- [Argall *et al.*, 2009] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

- [Aronson *et al.*, 2018] Reuben M Aronson, Thiago Santini, Thomas C Kübler, Enkelejda Kasneci, Siddhartha Srinivasa, and Henny Admoni. Eye-hand behavior in human-robot shared manipulation. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 4–13, 2018.
- [Balajee Vasudevan *et al.*, 2018] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Object referring in videos with language and human gaze. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4129–4138, 2018.
- [Betke, 2010] Margrit Betke. Intelligent interfaces to empower people with disabilities. In *Handbook of Ambient Intelligence and Smart Environments*, pages 409–432. Springer, 2010.
- [Borji and Itti, 2014] Ali Borji and Laurent Itti. Defending yarbus: Eye movements reveal observers’ task. *Journal of vision*, 14(3):29–29, 2014.
- [Borji *et al.*, 2015] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE transactions on image processing*, 24(12):5706–5722, 2015.
- [Brooks and Meltzoff, 2005] Rechele Brooks and Andrew N Meltzoff. The development of gaze following and its relation to language. *Developmental science*, 8(6):535–543, 2005.
- [Bylinskii *et al.*, 2015a] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. Intrinsic and extrinsic effects on image memorability. *Vision research*, 116:165–178, 2015.
- [Bylinskii *et al.*, 2015b] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark, 2015.
- [Bylinskii *et al.*, 2016] Zoya Bylinskii, Adrià Recasens, Ali Borji, Aude Oliva, Antonio Torralba, and Frédo Durand. Where should saliency models look next? In *European Conference on Computer Vision*, pages 809–824. Springer, 2016.
- [Bylinskii *et al.*, 2019] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2019.
- [Chen and Zhao, 2018] Shi Chen and Qi Zhao. Boosted attention: Leveraging human attention for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–84, 2018.
- [Cornia *et al.*, 2018a] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(2):1–21, 2018.
- [Cornia *et al.*, 2018b] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018.
- [Das *et al.*, 2017] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.
- [Deng *et al.*, 2019] Tao Deng, Hongmei Yan, Long Qin, Thuyen Ngo, and BS Manjunath. How do drivers allocate their potential attention? driving fixation prediction via convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems*, 2019.
- [Emery, 2000] Nathan J Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews*, 24(6):581–604, 2000.
- [Fang *et al.*, 2015] Rui Fang, Malcolm Doering, and Joyce Y Chai. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 271–278, 2015.
- [Fiore *et al.*, 2013] Stephen M Fiore, Travis J Wiltshire, Emilio JC Lobato, Florian G Jentsch, Wesley H Huang, and Benjamin Axelrod. Toward understanding social cues and signals in human-robot interaction: effects of robot gaze and proxemic behavior. *Frontiers in psychology*, 4:859, 2013.
- [Hayhoe, 2017] Mary M Hayhoe. Vision and action. *Annual review of vision science*, 3:389–413, 2017.
- [He *et al.*, 2019a] Sen He, Hamed R Tavakoli, Ali Borji, Yang Mi, and Nicolas Pugeault. Understanding and visualizing deep visual saliency models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10206–10215, 2019.
- [He *et al.*, 2019b] Sen He, Hamed R Tavakoli, Ali Borji, and Nicolas Pugeault. Human attention in image captioning: Dataset and analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8529–8538, 2019.
- [Henderson *et al.*, 2018] John M Henderson, Taylor R Hayes, Gwendolyn Rehrig, and Fernanda Ferreira. Meaning guides attention during real-world scene description. *Scientific reports*, 8(1):1–9, 2018.
- [Huang and Mutlu, 2016] Chien-Ming Huang and Bilge Mutlu. Anticipatory robot control for efficient human-robot collaboration. In *2016 11th ACM/IEEE international conference on human-robot interaction*, pages 83–90. IEEE, 2016.
- [Huang *et al.*, 2017] Qiong Huang, Ashok Veeraraghavan, and Ashutosh Sabharwal. Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications*, 28(5-6):445–461, 2017.
- [Huang *et al.*, 2019] Sandy H Huang, Isabella Huang, Ravi Pandya, and Anca D Dragan. Nonverbal robot feedback for human teachers. *Conference on Robot Learning*, 2019.
- [Hutt *et al.*, 2016] Stephen Hutt, Caitlin Mills, Shelby White, Patrick J Donnelly, and Sidney K D’Mello. The eyes have it: Gaze-based detection of mind wandering during learning with an intelligent tutoring system. *International Educational Data Mining Society*, 2016.
- [Itti *et al.*, 1998] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259, 1998.
- [Jaques *et al.*, 2014] Natasha Jaques, Cristina Conati, Jason M Harley, and Roger Azevedo. Predicting affect from gaze data during interaction with an intelligent tutoring system. In *International Conference on Intelligent Tutoring Systems*, pages 29–38. Springer, 2014.
- [Jetley *et al.*, 2016] Saumya Jetley, Naila Murray, and Eleonora Vig. End-to-end saliency mapping via probability distribution prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5753–5761, 2016.
- [Jiang *et al.*, 2015] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015.

- [Karessli *et al.*, 2017] Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. Gaze embeddings for zero-shot image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4525–4534, 2017.
- [Krafka *et al.*, 2016] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016.
- [Kruthiventi *et al.*, 2017] Srinivas SS Kruthiventi, Kumar Ayush, and Radhakrishnan Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 2017.
- [Kümmerer *et al.*, 2016] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016.
- [Li *et al.*, 2014] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–287, 2014.
- [Li *et al.*, 2018] Yin Li, Miao Liu, and James M Rehg. 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)*, pages 619–635, 2018.
- [Liu *et al.*, 2019] Congcong Liu, Yuying Chen, Lei Tai, Haoyang Ye, Ming Liu, and Bertram E Shi. A gaze model improves autonomous driving. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, page 33. ACM, 2019.
- [Mathe and Sminchisescu, 2014] Stefan Mathe and Cristian Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(7):1408–1424, 2014.
- [Mnih *et al.*, 2014] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.
- [Moon *et al.*, 2014] AJung Moon, Daniel M Troniak, Brian Gleeson, Matthew KXJ Pan, Minhua Zheng, Benjamin A Blumer, Karon MacLean, and Elizabeth A Croft. Meet me where i’m gazing: how shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 334–341, 2014.
- [Nguyen *et al.*, 2018] Tam V Nguyen, Qi Zhao, and Shuicheng Yan. Attentive systems: A survey. *International Journal of Computer Vision*, 126(1):86–110, 2018.
- [Palazzi *et al.*, 2018] Andrea Palazzi, Davide Abati, Simone Calderara, Francesco Solera, and Rita Cucchiara. Predicting the driver’s focus of attention: the dr (eye) ve project. *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [Palinko *et al.*, 2016] Oskar Palinko, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5048–5054. IEEE, 2016.
- [Papadopoulos *et al.*, 2014] Dim P Papadopoulos, Alasdair DF Clarke, Frank Keller, and Vittorio Ferrari. Training object class detectors from eye tracking data. In *European conference on computer vision*, pages 361–376. Springer, 2014.
- [Papoutsaki *et al.*, 2016] Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediya Daskalova, Jeff Huang, and James Hays. Webgazer: Scalable webcam eye tracking using user interactions. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence-IJCAI 2016*, 2016.
- [Park *et al.*, 2018] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 721–738, 2018.
- [Penkov *et al.*, 2017] Svetlin Penkov, Alejandro Boddallo, and Subramanian Ramamoorthy. Physical symbol grounding and instance learning through demonstration and eye tracking. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5921–5928. IEEE, 2017.
- [Posner and Petersen, 1990] Michael I Posner and Steven E Petersen. The attention system of the human brain. *Annual review of neuroscience*, 13(1):25–42, 1990.
- [Qiao *et al.*, 2018] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. Exploring human-like attention supervision in visual question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [Ravichandar *et al.*, 2018] Harish Chaandar Ravichandar, Avnish Kumar, and Ashwin Dani. Gaze and motion information fusion for human intention inference. *International Journal of Intelligent Robotics and Applications*, 2(2):136–148, 2018.
- [Recasens *et al.*, 2015] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, pages 199–207, 2015.
- [Recasens *et al.*, 2017] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1435–1443, 2017.
- [Rich *et al.*, 2010] Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L Sidner. Recognizing engagement in human-robot interaction. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 375–382. IEEE, 2010.
- [Richter *et al.*, 2016] Viktor Richter, Birte Carlmeyer, Florian Lier, Sebastian Meyer zu Borgsen, David Schlangen, Franz Kummert, Sven Wachsmuth, and Britta Wrede. Are you talking to me? improving the robustness of dialogue systems in a multi party hri scenario by incorporating gaze direction and lip movement of attendees. In *Proceedings of the Fourth International Conference on Human Agent Interaction*, pages 43–50, 2016.
- [Ruhland *et al.*, 2015] Kerstin Ruhland, Christopher E Peters, Sean Andrist, Jeremy B Badler, Norman I Badler, Michael Gleicher, Bilge Mutlu, and Rachel McDonnell. A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception. In *Computer graphics forum*, volume 34, pages 299–326. Wiley Online Library, 2015.
- [Saran *et al.*, 2018] Akanksha Saran, Srinjoy Majumdar, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. Human gaze following for human-robot interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8615–8621. IEEE, 2018.
- [Saran *et al.*, 2019] Akanksha Saran, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. Understanding teacher gaze patterns for robot learning. *Conference on Robot Learning*, 2019.
- [Shafti *et al.*, 2019] Ali Shafti, Pavel Orlov, and A Aldo Faisal. Gaze-based, context-aware robotic system for assisted reaching

- and grasping. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 863–869. IEEE, 2019.
- [Shen *et al.*, 2018] Yang Shen, Bingbing Ni, Zefan Li, and Ning Zhuang. Egocentric activity prediction via event modulated attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 197–212, 2018.
- [Shrivastava *et al.*, 2017] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017.
- [Sudhakaran *et al.*, 2019] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9954–9963, 2019.
- [Sugano and Bulling, 2016] Yusuke Sugano and Andreas Bulling. Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203*, 2016.
- [Tavakoli *et al.*, 2017] Hamed R Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. Paying attention to descriptions generated by image captioning models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2487–2496, 2017.
- [Tavakoli *et al.*, 2019] Hamed Rezazadegan Tavakoli, Esa Rahtu, Juho Kannala, and Ali Borji. Digging deeper into egocentric gaze prediction. In *2019 IEEE Winter Conference on Applications of Computer Vision*, pages 273–282. IEEE, 2019.
- [Thomaz *et al.*, 2016] Andrea Thomaz, Guy Hoffman, Maya Cakmak, et al. Computational human-robot interaction. *Foundations and Trends® in Robotics*, 4(2-3):105–223, 2016.
- [Vaidyanathan *et al.*, 2018] Preethi Vaidyanathan, Emily Prud’hommeaux, Jeff B Pelz, and Cecilia Ovesdotter Alm. Snag: Spoken narratives and gaze dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 132–137, 2018.
- [van Miltenburg *et al.*, 2018] Emiel van Miltenburg, Akos Kádár, Ruud Koolen, and Emiel Kraemer. Didec: The dutch image description and eye-tracking corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3658–3669, 2018.
- [Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [Wang *et al.*, 2018] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4894–4903, 2018.
- [Wood *et al.*, 2015] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3756–3764, 2015.
- [Xia *et al.*, 2018] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. Predicting driver attention in critical situations. In *Asian conference on computer vision*, pages 658–674. Springer, 2018.
- [Xu *et al.*, 2014] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014.
- [Xu *et al.*, 2018] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. Gaze prediction in dynamic 360 immersive videos. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5333–5342, 2018.
- [Yu and Smith, 2011] Chen Yu and Linda B Smith. What you learn is what you see: using eye movements to study infant cross-situational word learning. *Developmental science*, 14(2):165–180, 2011.
- [Yu *et al.*, 2014] Mingxin Yu, Yingzi Lin, David Schmidt, Xi-angzhou Wang, and Yu Wang. Human-robot interaction based on gaze gestures for the drone teleoperation. *Journal of Eye Movement Research*, 7(4):1–14, 2014.
- [Yu *et al.*, 2017] Youngjae Yu, Jongwook Choi, Yeonhwa Kim, Kyung Yoo, Sang-Hun Lee, and Gunhee Kim. Supervising neural attention models for video captioning by human gaze data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 490–498, 2017.
- [Yun *et al.*, 2013] Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J Zelinsky, and Tamara L Berg. Studying relationships between human gaze, description, and computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 739–746, 2013.
- [Zhang *et al.*, 2015] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015.
- [Zhang *et al.*, 2017] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017.
- [Zhang *et al.*, 2018a] Ruohan Zhang, Zhuode Liu, Luxin Zhang, Jake A Whritner, Karl S Muller, Mary M Hayhoe, and Dana H Ballard. Agil: Learning attention from human for visuomotor tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 663–679, 2018.
- [Zhang *et al.*, 2018b] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. Saliency detection in 360 videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 488–503, 2018.
- [Zhang *et al.*, 2019a] Ruohan Zhang, Faraz Torabi, Lin Guan, Dana H Ballard, and Peter Stone. Leveraging human guidance for deep reinforcement learning tasks. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6339–6346. AAAI Press, 2019.
- [Zhang *et al.*, 2019b] Ruohan Zhang, Calen Walshe, Zhuode Liu, Lin Guan, Karl S Muller, Jake A Whritner, Luxin Zhang, Mary M Hayhoe, and Dana H Ballard. Atari-head: Atari human eye-tracking and demonstration dataset. *arXiv preprint arXiv:1903.06754*, 2019.
- [Zuo *et al.*, 2018] Zheming Zuo, Longzhi Yang, Yonghong Peng, Fei Chao, and Yanpeng Qu. Gaze-informed egocentric action recognition for memory aid systems. *IEEE Access*, 6:12894–12904, 2018.