A Gaze Model Improves Autonomous Driving

CONGCONG LIU*, Hong Kong University of Science and Technology
YUYING CHEN*, Hong Kong University of Science and Technology
LEI TAI†, Hong Kong University of Science and Technology
HAOYANG YE, Hong Kong University of Science and Technology
MING LIU, Hong Kong University of Science and Technology
BERTRAM E. SHI, Hong Kong University of Science and Technology

End-to-end behavioral cloning trained by human demonstration is now a popular approach for vision-based autonomous driving. A deep neural network maps drive-view images directly to steering commands. However, the images contain much task-irrelevant data. Humans attend to behaviorally relevant information using saccades that direct gaze towards important areas. We demonstrate that behavioral cloning also benefits from active control of gaze. We trained a conditional generative adversarial network (GAN) that accurately predicts human gaze maps while driving in both familiar and unseen environments. We incorporated the predicted gaze maps into end-to-end networks for two behaviors: following and overtaking. Incorporating gaze information significantly improves generalization to unseen environments. We hypothesize that incorporating gaze information enables the network to focus on task critical objects, which vary little between environments, and ignore irrelevant elements in the background, which vary greatly.

CCS Concepts: • **Computing methodologies → Vision for robotics**.

Additional Key Words and Phrases: Eye Tracking, Imitation Learning, Autonomous Driving

ACM Reference Format:
Congcong Liu, Yuying Chen, Lei Tai, Haoyang Ye, Ming Liu, and Bertram E. Shi. 2019. A Gaze Model Improves Autonomous Driving. 1, 1 (July 2019), 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

End-to-end models for vision-based autonomous driving systems have attracted much interest because they simultaneous optimize for all processing stages and eliminate tedious feature engineering procedures [2]. Algorithms such as reinforcement learning (RL) and imitation learning have achieved much success [2, 5, 8, 22]. The major limitations for RL are the need for a predefined reward function, which can be difficult in more complex driving environments, and its sample efficiency. Previous work has shown that using human demonstration in Atari games can speed up RL [7, 9].

*These two authors contributed equally.
†Lei Tai was also with City University of Hong Kong, Shenzhen Research Institute.

Authors' addresses: Congcong Liu, Hong Kong University of Science and Technology, Hong Kong, cliubh@connect.ust.hk; Yuying Chen, Hong Kong University of Science and Technology, Hong Kong, ychenco@connect.ust.hk; Lei Tai, Hong Kong University of Science and Technology, Hong Kong, ltai@connect.ust.hk; Haoyang Ye, Hong Kong University of Science and Technology, Hong Kong, hyeab@connect.ust.hk; Ming Liu, Hong Kong University of Science and Technology, Hong Kong, eelium@ust.hk; Bertram E. Shi, Hong Kong University of Science and Technology, Hong Kong, eebert@ust.hk.

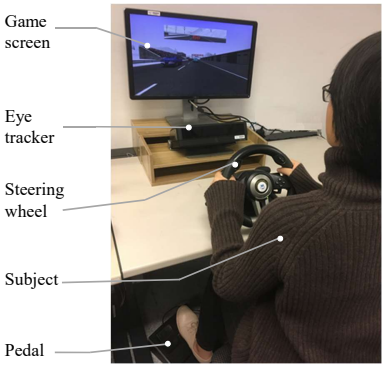

Fig. 1. Experimental Setup. A Tobii Pro X60 remote eye tracker with 60 Hz sampling frequency measures eye gaze. A PXN V3II 180 degree steering wheel and pedal provides steering and accelerator commands.The game screen is a Dell inc. 23 inch LCD monitor.

Imitation learning trains an agent which can mimic the behavior of human demonstrator. One popular approach is behavioral cloning (BC). Given human demonstrations, BC can train a supervised model directly from visual input to control commands like steering angle. It has been applied in tasks such as road following [2].

Human gaze behavior provides a wealth of cues related to driver intent and decision making. However, it has not yet been well investigated and utilized in the context of autonomous driving. Rather, eye gaze has been mostly used for driver-assistance systems like fatigue monitoring [13] and workload classification [20]. Whether and how human gaze can benefit autonomous driving is still an open problem [1]. One recent work uses a multi-branch deep neural network to estimate human gaze in driving scenes[15]. However, their main focus was to analyze the gaze distribution over the objects in road scenes and how car speed influences human gaze. They did not incorporate human gaze to improve autonomous driving.

Here, we investigate the possibility of incorporating gaze behavior into self-driving systems. To do this, we train a conditional GAN (Pix2Pix) to estimate the distribution of human gaze in the visual scene while driving [11]. The GAN minimizes an adversarial loss function, which tries to make the estimated gaze map indistinguishable from the real human gaze map by a trained discriminator network. We incorporate this gaze map into an end-to-end network trained by imitation learning.

There are several key contributions of our work. First, we train a gaze network that accurately estimates human gaze maps both in training scenes and unseen scenes. Second, we demonstrate how gaze maps can be incorporated into an end-to-end driving network to significantly improve the action estimation accuracy. Third, we show that the incorporation of human gaze information improves generalization to unseen environments. To our knowledge, this is the first attempt to incorporate a model of human gaze into an autonomous driving system.

2 EXPERIMENTAL SETUP

Fig. 1 shows the experimental setup and the main devices we used in our experiment. A Tobii Pro X60 remote eye tracker records eye gaze data while subjects are driving in a car simulator. The size of the computer screen is 51 cm $\times$ 28.7 cm. The distance between the screen and driver is 80 cm. The horizontal and vertical visual angles subtended by the screen are 35.5° and 20.0° respectively. The monitor resolution is 1920 by 1280 pixels. In our experiment, one visual degree corresponds to about 54 pixels. The standard nine point calibration scheme is conducted before the experiment. Right

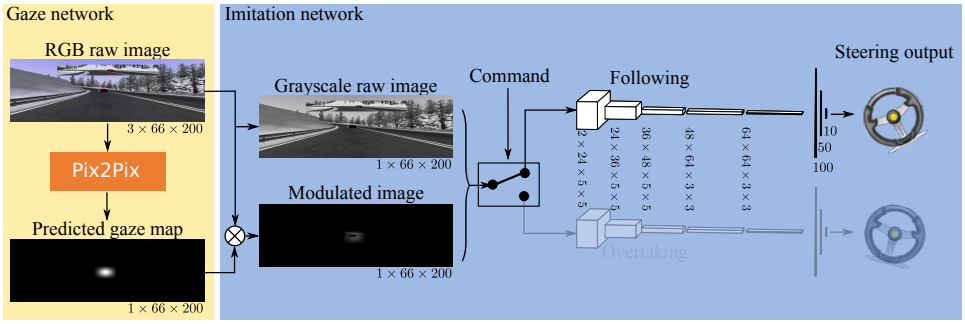

Fig. 2. The system for gaze aided autonomous driving. The gaze network (Pix2Pix) is trained with pairs of driver-view images and ground truth gaze maps. To incorporate the gaze information into the end-to-end network, we pixel-wise multiply the original driver-view image with the estimated gaze map. The gray-scale modulated and original images are stacked as input to the imitation network. The command switches the input between networks for path following or overtaking. The blocks represent the output of each convolutional layer and are labelled by the corresponding filter sizes. After flattening, the output of the last convolutional layer is fed into four fully connected (FC) layers. The three vertical lines show the output of the first three FC layers. The final layer has a single scalar unit encoding the steering command. More details are contained in the supplementary material.

after the calibration, we collected gaze data from the nine points again to calculate the calibration error. The average error across all trials is about 0.6°.

We conducted our driving experiments in TORCS [21], an open source driving simulator for highway driving, which has been used for vision-based self-driving system in recent years[4, 23]. The TORCS simulator runs on a PC with the Ubuntu 16.04 operating system. We chose TORCS for several reasons. First, TORCS is a light-weight simulator without strict hardware requirements. Second, it supports highway driving simulation with multiple lanes which is suitable for our task. Third, it is an open-source software, which makes it convenient for us to modify to suit our needs. To collect training data, we used five different tracks and collected data for four trial runs on each track. During each three minute trial, the subject drives along the race track at a constant speed and overtakes cars if needed. During the time the subject is overtaking another car, the subject presses a button to mark overtaking behavior. To make it more realistic, subjects use a steering wheel to control the direction of the car.

Among the twenty trials collected, we choose two trials each from the first two tracks (S1-S2) as the training dataset (about 40000 action-image pairs). The remaining two trials each from S1 and S2 are for testing and serve as references for evaluations in unseen tracks (S3-S5). After balancing the dataset (down-sampling the dataset since the gazes located in the center area most of the time), we obtained 3843 image-gaze map pairs for training. The remaining twelve trials of testing tracks (S3-S5) are the testing dataset.

During testing, the steering angle output by the system described in the next section was sent to the server to control the car. The car is set to run at a constant speed, making the steering angle the only control variable.

3 METHODOLOGY

Saccadic gaze behavior is often modelled using visual saliency. There are two strategies to model saliency: bottom-up and top-down. Bottom-up saliency models are generally task-independent, driven by image characteristics, such as color, orientation and intensity [12]. They usually model eye gaze during free-viewing. Top-down saliency models are task-specific and emphasize task-relevant object characteristics [16, 19]. Task-irrelevant objects will not be considered

as salient, even they would be in a bottom-up model. Some models combine the bottom-up and top-down strategies using deep networks [15, 24].

Our architecture is shown in Fig. 2. It contains two parts: the gaze network and the imitation network. The structure of these networks is described in more detail below. For details about the training, please refer to the supplementary materials.

3.1 Gaze Network

The gaze network followed the Pix2Pix architecture. It was trained with pairs of driver-view images and ground truth gaze maps to produce estimated gaze maps, analogous to saliency maps.

For each frame, we compute the ground truth gaze map based on the gaze measured in a ten-frame sliding time window centered on the current frame. We created a 2D probability distribution by placing a Gaussian with σ = one degree of visual angle at each gaze location. We chose the value σ equal to one degree of visual angle, because this is the size of the fovea [3, 14].

Most previous deep saliency models attempt to estimate gaze maps from images through convolutional neural networks. Commonly used loss functions are the L2 distance [6] or saliency related loss functions like the KL-divergence [10]. In our case, most pixels of the ground truth gaze map have values close to zero, with only a few significant non-zero values. In our experiments, using the L2 distance results in estimated gaze maps where all pixels had near-zero values. In this paper, we adopt a conditional GAN, the Pix2Pix network [11], to map images to gaze maps. The GAN learns both a generator network and a discriminator network that attempts to distinguish generated and real gaze maps. The loss function penalizes generated gaze maps that are easily distinguished from real gaze maps. This resulted in much better gaze maps, since all zero gaze maps are unrealistic.

3.2 Imitation Learning with Gaze

The imitation network followed a deep CNN architecture to produce steering commands. The input is obtained by stacking the original gray-scale driver-view image and the gray-scale driver-view image modulated by the estimated gaze map. We converted RGB images to gray-scale using the rgb2gray function from scikit-image in Python [17]. We also trained the network with RGB images, but found the performance to be worse.

We trained two convolutional networks, each with the same structure but different parameters: one for following behavior and one for overtaking behavior. We manually selected the driving maneuver (path following or overtaking) and connected the input to the corresponding imitation network to generate the steering command.

As the baseline, we first trained a vanilla end-to-end imitation learning network, the PilotNet proposed by [2], to map driver-view images directly to steering commands. The gaze modulated map was not used. We refer to this network in the results as **Without gaze information**).

We implemented three methods to evaluate the benefits of gaze information.

With real gaze map: The modulated driver-view image was obtained by pixel-wise multiplication by the ground truth gaze map.

With estimated gaze map: The modulated driver-view image was obtained by pixel-wise multiplication with the estimated gaze map from the gaze network.

With center Gaussian blob: The modulated driver-view image was obtained by pixel-wise multiplication with a mask containing a single Gaussian at the center of the image. Based on our observation of the collected gaze trajectories,

Table 1. Similarity of gaze estimates. KL denotes Kullback-Leibler divergence and CC denotes Correlation Coefficient.

		S1	S2	S3	S4	S5
KL	estimated gaze map	2.97	2.99	4.10	4.03	4.18
	Center Gaussian blob	5.22	5.34	4.96	4.11	4.65
CC	estimated gaze map	0.62	0.60	0.54	0.55	0.51
	Center Gaussian blob	0.32	0.30	0.36	0.43	0.39

we find the subject mostly looks at the center area of the scene. This network tested the effect of simply emphasizing the center region.

4 RESULTS

4.1 Gaze Network Evaluation

To evaluate the gaze network quantitatively, we used two standard metrics from the visual saliency literature [3, 18]: the Kullback-Leibler divergence (KL) and the Correlation Coefficient (CC). Better similarity is indicated by a smaller KL divergence and larger CC.

The estimated gaze map closely matches the ground truth gaze. As shown in Table 1, the KL divergence between the estimated and ground truth gaze maps is significantly smaller than the KL divergence between the center Gaussian blob mask and the ground truth gaze map, especially for the two tracks used in training (S1, S2). The KL divergence for the estimated gaze map was 43% smaller for S1 and 44% smaller for S2 than the center Gaussian blob. Similarly, the CC between the estimated and the ground truth gaze maps is larger than the CC between the center Gaussian blob and the ground truth gaze map. For the training tracks, the CC was 94% larger CC for S1 and 100% larger CC for S2.

The gaze model we trained can generalize well to unseen scenes. For the three untrained tracks (S3-S5), we observe better similarity between the estimated and ground truth maps than between the center Gaussian blob and ground truth gaze map. However, the improvement is smaller. The KL divergence was 10% smaller on average. The CC was 36% larger on average.

Fig. 3 shows examples of the estimated gaze map for all five tracks (both training and testing). These also illustrate the good performance of the gaze network. It not only generates realistic gaze maps while following the road, but also reproduces the driver's behavior of gazing at the rear-view mirror before overtaking. The supplementary video shows estimation performance of the gaze network during driving. In addition to fixations on the mirror, the gaze also departs from the center to track the car in front. This suggests gaze helps the network focus on task-critical objects.

4.2 Imitation Learning with Gaze

Incorporating gaze information significantly improves the task performance, as measured by the mean absolute estimation error between the steering actions generated by the network and by humans. As shown in Table 2, for the two training tracks (S1, S2), the imitation network with the estimated gaze map outperforms the baseline vanilla network by 4.5% and the imitation network with the center Gaussian blob by 9.6%. Interestingly, the network with the

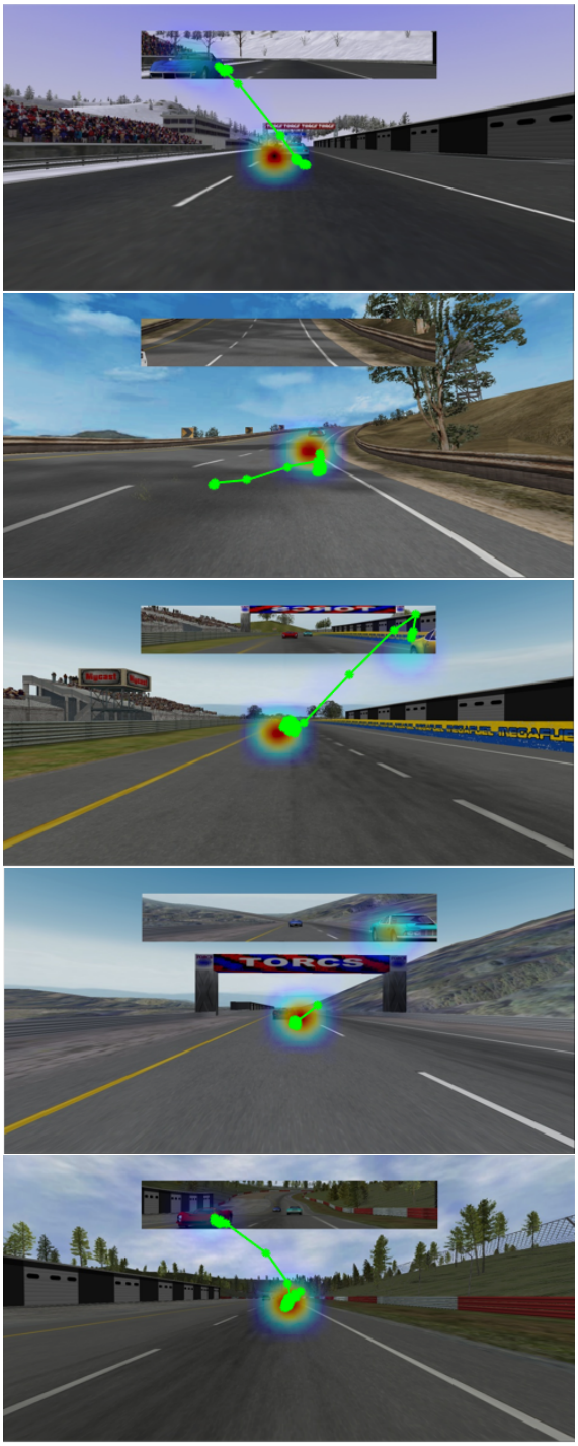

Fig. 3. Estimated gaze heat maps and ground truth gaze trajectories (green line) superimposed on the corresponding scenes of the five different tracks. The first row corresponds to the two training tracks (S1, S2), and the second row corresponds to the three testing tracks (S3, S4, S5). Red area indicates more fixations.

Table 2. Human action estimation results

mean absolute error/deg	S1	S2	S3	S4	S5
Without gaze information	2.45	2.86	6.11	6.39	5.62
With real gaze map	2.57	2.86	5.09	**3.62**	**4.50**
With estimated gaze map	**2.35**	**2.72**	**5.04**	4.19	5.28
With center Gaussian blob	2.56	3.06	5.88	4.97	6.26

estimated gaze maps outperforms that with the real gaze map. We hypothesize that this may be because the real gaze map exhibits more frequent saccades to the background.

The improvement generalizes to unseen scenes. For the untrained three tracks (S3-S5), the network with the estimated gaze map performs on average 19.3% better on average than the vanilla imitation network without gaze. The network with the ground truth gaze map performs even better, achieving a 23.6% decrease on average in the action estimation error. Both networks also perform better than the network with the center Gaussian blob. On average, the network with estimated gaze map achieves a 15.2% decrease while the network with ground truth gaze map achieves a 22.9% decrease in the action estimation error.

To see a performance comparison between imitation network with gaze and vanilla imitation network running in the actual simulator, please refer to our video demo at https://goo.gl/13551M. The networks shown here were trained on tracks S1-S4 and test in track S5 on the simulator.

5 CONCLUSIONS

We demonstrate that a conditional GAN can generate high quality estimates of human gaze maps, which generalize well to unseen environments. Furthermore, we show that incorporating eye gaze information into an end-to-end imitation network significantly improves the accuracy of human action estimation on a driving task. Both estimated and ground truth gaze maps result in comparable performance.

Our work can be extended in several directions in the future. First, the current gaze network does not consider the spatio-temporal dynamics of human gaze. Adding spatio-temporal modules to the gaze network, e.g. a recurrent neural network, may yield better performance. Second, the since eye gaze data holds cues to human intent, the estimated gaze maps may also be used to select between actions, e.g. following and overtaking, which is being done manually in the work presented here.

ACKNOWLEDGMENTS

This work was supported in part by the Hong Kong Research Grants Council under grant 16213617 and in part by the Shenzhen Science, Technology and Innovation Commission (SZSTI) under grant JCYJ20160428154842603.

REFERENCES

[1] Stefano Alletto, Andrea Palazzi, Francesco Solera, Simone Calderara, and Rita Cucchiara. 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.

[2] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316* (2016).

[3] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. 2018. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence* (2018).

[4] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. 2015. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*. 2722–2730.

[5] Felipe Codevilla, Matthias Miiller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. 2018. End-to-end driving via conditional imitation learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1–9.

[6] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2016. A deep multi-level network for saliency prediction. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 3488–3493.

[7] Gabriel V Cruz Jr, Yunshu Du, and Matthew E Taylor. 2017. Pre-training Neural Networks with Human Demonstrations for Deep Reinforcement Learning. *arXiv preprint arXiv:1709.04083* (2017).

[8] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. *arXiv preprint arXiv:1711.03938* (2017).

[9] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Gabriel Dulac-Arnold, et al. 2017. Deep Q-learning from Demonstrations. *arXiv preprint arXiv:1704.03732* (2017).

[10] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. 2015. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 262–270.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 5967–5976.

[12] Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence* 20, 11 (1998), 1254–1259.

[13] Qiang Ji, Zhiwei Zhu, and Peilin Lan. 2004. Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE transactions on vehicular technology* 53, 4 (2004), 1052–1068.

[14] Olivier Le Meur and Thierry Baccino. 2013. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods* 45, 1 (2013), 251–266.

[15] Andrea Palazzi, Davide Abati, Simone Calderara, Francesco Solera, and Rita Cucchiara. 2017. Predicting the Driver's Focus of Attention: the DR (eye) VE Project. *arXiv preprint arXiv:1705.03854* (2017).

[16] Robert J Peters and Laurent Itti. 2007. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*. IEEE, 1–8.

[17] Charles Poynton. 1998. Frequently asked questions about gamma. *Rapport Technique, janvier* 152 (1998).

[18] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. 2013. Saliency and human fixations: state-of-the-art and study of comparison metrics. In *Proceedings of the IEEE international conference on computer vision*. 1153–1160.

[19] Constantin A Rothkopf, Dana H Ballard, and Mary M Hayhoe. 2007. Task and context determine where you look. *Journal of vision* 7, 14 (2007), 16–16.

[20] Ruohan Wang, Pierluigi V Amadori, and Yiannis Demiris. 2018. Real-Time Workload Classification during Driving using HyperNetworks. *arXiv preprint arXiv:1810.03145* (2018).

[21] Bernhard Wymann, Eric Espié, Christophe Guionneau, Christos Dimitrakakis, Rémi Coulom, and Andrew Sumner. 2000. Torcs, the open racing car simulator. *Software available at http://torcs. sourceforge. net* 4 (2000), 6.

[22] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. 2017. End-to-end learning of driving models from large-scale video datasets. *arXiv preprint* (2017).

[23] Jiakai Zhang and Kyunghyun Cho. 2016. Query-efficient imitation learning for end-to-end autonomous driving. *arXiv preprint arXiv:1605.06450* (2016).

[24] Ruohan Zhang, Zhuode Liu, Luxin Zhang, Jake A Whritner, Karl S Muller, Mary M Hayhoe, and Dana H Ballard. 2018. AGIL: Learning Attention from Human for Visuomotor Tasks. *arXiv preprint arXiv:1806.03960* (2018).