Using Eye Gaze to Enhance Generalization of Imitation Networks to Unseen Environments

Congcong Liu, *Graduate Student Member, IEEE*, Yuying Chen, *Student Member, IEEE*, Ming Liu, *Senior Member, IEEE*, and Bertram E. Shi, *Fellow, IEEE*

***Abstract*— Vision-based autonomous driving through imitation learning mimics the behavior of human drivers by mapping driver view images to driving actions. This article shows that performance can be enhanced via the use of eye gaze. Previous research has shown that observing an expert's gaze patterns can be beneficial for novice human learners. We show here that neural networks can also benefit. We trained a conditional generative adversarial network to estimate human gaze maps accurately from driver-view images. We describe two approaches to integrating gaze information into imitation networks: eye gaze as an additional input and gaze modulated dropout. Both significantly enhance generalization to unseen environments in comparison with a baseline vanilla network without gaze, but gaze-modulated dropout performs better. We evaluated performance quantitatively on both single images and in closed-loop tests, showing that gaze modulated dropout yields the lowest prediction error, the highest success rate in overtaking cars, the longest distance between infractions, lowest epistemic uncertainty, and improved data efficiency. Using Grad-CAM, we show that gaze modulated dropout enables the network to concentrate on task-relevant areas of the image.**

***Index Terms*— Autonomous driving, eye gaze, generalization, human attention, imitation learning (IL).**

I. Introduction

END-TO-END deep learning has attracted great interest and has been successfully applied to numerous autonomous control tasks. Many attempts have been made in end-to-end vision-based driving through imitation learning (IL) [1], [2] and reinforcement learning (RL) [3]. The IL can be adapted to complex driving scenarios more efficiently than RL, which requires careful selection of an appropriate reward function.

One typical solution to IL is behavioral cloning (BC), which has been used successfully in many autonomous systems, including off-road driving [4] and lane following [1]. It is simple and efficient. It follows a student–teacher paradigm, where teachers give demonstrations for students to learn from. Previous works in BC for autonomous driving systems largely concentrated on establishing a mapping from the sensor data to the control commands explicitly, without consideration of other cues from the teachers.

While executing tasks, humans pay attention to an important area in the visual environment using saccades. A driver's gaze contains rich information related to his/her intent and decision making. Previous research has found that exposure to experts' gaze trajectories can be beneficial for novice human learners. For example, Vine *et al.* [5] showed that exposing gaze strategies of experts can improve the laparoscopic skills of novice learners. Yamini *et al.* [6] found that novice drivers can improve their hazard anticipation ability by viewing a video of expert gaze patterns.

Thus, it is quite promising to explore whether BC for autonomous driving might also benefit from observation of expert gaze patterns.

There has been little work studying how the human gaze can help autonomous driving. Palazzi *et al.* [7] analyzed gaze data in different driving conditions. They trained a network to predict eye gaze and demonstrated a strong relationship between gaze patterns and driving conditions. However, they did not apply their results to the autonomous driving system.

In this article, we applied a conditional generative adversarial network (GAN) to anticipate human gaze maps accurately while running in both seen and unseen environments. We incorporated the estimated gaze maps into deep driving networks through two different methods.

In the first approach, we use the gaze map as an additional input to the network [8]. While this is a fairly straightforward approach to incorporating additional information, it has the disadvantage of increased network complexity. Since most units of the gaze map are zero value, this additional complexity is utilized inefficiently.

In the second approach, the gaze map is applied to spatially modulate the dropout probability. Areas close to the estimated gaze position possess a smaller dropout probability than areas distant. Since human saccadic eye movements direct high-resolution processing and attention to different areas of the scene, we hypothesized that incorporating a human gaze model as a modulatory effect, rather than as an additional input, may be more effective. This helps the network concentrate

Manuscript received August 26, 2019; revised February 24, 2020; accepted May 14, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant U1713211, in part by the Shenzhen Science, Technology and Innovation Commission under Grant JCYJ20160428154842603. The work of Ming Liu and Bertram E. Shi was supported by the Research Grant Council of Hong Kong SAR Government, China, under Grant 11210017, Grant 21202816, and Grant 16211015. *(Congcong Liu and Yuying Chen contributed equally to this work.) (Corresponding authors: Ming Liu; Bertram E. Shi.)*

The authors are with the Department of Electrical and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong (e-mail: cliubh@connect.ust.hk; ychenco@connect.ust.hk; eelium@ust.hk; eebert@ust.hk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.2996386

on task-relevant areas and neglect task-irrelevant areas, such as the background, which should be useful when the network is presented with new and unfamiliar inputs.

To quantify the generalization ability, we estimated the *epistemic* (model) uncertainty of the trained neural networks. The epistemic uncertainty quantifies the similarity between new inputs and previous observations and is inversely related to model confidence [9]. In autonomous driving, epistemic uncertainty can help quantify the capability that a trained model generalizes to unseen environments.

The main contributions of our work are as follows.

1) We train a gaze model to estimate gaze maps accurately, given the driver-view images.
2) We propose two approaches to incorporate gaze information into deep imitation networks that both improve generalization to unseen environments.
3) We demonstrate that deep networks can benefit from a complete understanding of human behavior, by showing how auxiliary cues not directly related to control commands can improve IL.
4) We validate the effectiveness of gaze modulated dropout using a quantitative measure of performance and the epistemic uncertainty.
5) Through a CNN visualization technique, we find support for an intuitive explanation for the effectiveness of incorporating eye gaze information.

II. Related Work

In this section, we first review work in end-to-end vision-based autonomous driving systems. We then describe prior work using gaze in driver assistance and in characterizing the human gaze during driving. Finally, we discuss different uncertainty measures and their application in deep learning.

A. End-to-End Autonomous Driving

End-to-end learning has become increasingly popular for vision-based autonomous driving because it does not require separation into several steps and can be trained to optimize overall system performance [1].

The RL learns a policy through trial and error without the need for human demonstrations. You *et al.* [10] used a virtual-to-real style translation network to transfer a driving policy trained by RL in the *TORCS* simulator to real-world driving. Liang *et al.* [11] used an imitative RL network to learn the policy. First, they trained the imitation network from human demonstration. Then, they used this as the initial policy for further refinement by RL. The main drawbacks of RL are its low sample efficiency and the need to shape an appropriate reward function, which becomes intractable for complex driving scenarios.

The IL learns policy that mimics human behavior from human demonstrations. Bojarski *et al.* [1] trained a feed-forward five-layer CNN followed by three fully connected (FC) layers (*PilotNet*) to generate steering output for lane following given images of the road ahead. Chen and Huang [12] used a similar architecture for a lane keeping task. Muller *et al.* proposed a similar framework for an off-road obstacle avoidance task [4]. Codevilla *et al.* [2] trained a branched deep imitation network for policy learning. Given high-level command from a human or a global planning module, the imitation network maps camera images and car speed to action. The major limitation of IL systems is that they do not generalize well, i.e., their performance degrades in unfamiliar environments [11]. For instance, despite the many data augmentation methods applied in [2], the policy network trained in Town1 of the Carla simulator had clearly degraded driving performance while tested in Town2.

B. Human Gaze in Driving Systems

During driving, rich cues about a driver's intent, mental state, and decision making are conveyed by his/her eye gaze. In assisted driving systems, eye gaze information has been used to evaluate driver mental state, such as tiredness detection [13] and mental load classification [14]. However, eye gaze has yet to be well investigated for autonomous driving research [15].

To the best of our knowledge, the most related article to ours is by Palazzi *et al.* [7], who introduced a deep neural network with multiple branches to predict human gaze in urban driving scenarios. They investigated the distribution of human gaze over several semantic classes in visual scenes, how driving speed correlated with the driver's attention, and how these measures varied over different scenarios. However, they did not consider the application of an estimated eye gaze to autonomous driving.

C. Types of Uncertainty in Deep Learning

The measurement of uncertainty can quantify a model's confidence and indicate what the model does not know [16]. Uncertainty can be divided into two types: *epistemic* and *aleatoric*.

Aleatoric uncertainty captures noise in the observations caused by sensor noise. Kendall [16] proposed to evaluate aleatoric uncertainty by increasing an extra output. Feng *et al.* [17] used a similar approach to catch the uncertainty in 3-D object detection tasks. Wang *et al.* [18] quantified the aleatoric uncertainty of image segmentation and used it to investigate the effect of multiple image transformations on segmentation.

Epistemic (model) uncertainty quantifies "familiarity," by measuring the similarity of a new input to previously seen observations [9]. Kendall [16] quantified epistemic uncertainty by multiple sampling over the distribution of model weights through dropout at the testing stage. This method has been adopted in many tasks, such as semantic segmentation [19] and depth regression [20]. Objects that are rare in training data sets will lead to larger epistemic uncertainties. The major drawback of the dropout method is the need for expensive sampling, which makes it unsuitable for real-time applications. However, it can be applied offline to measure model confidence.

III. Methodology

In this section, we first introduce our experimental setup to collect training data and to test model performance. After that, we introduce the network for estimating the gaze map.

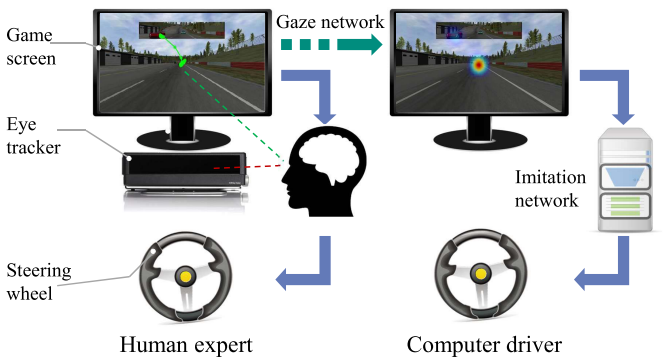


Fig. 1. Illustration of the general idea behind the IL augmented with gaze maps. Typically only the image-action pairs are collected as human expert demonstrations. Other informative cues, such as human gazes, are ignored. It is expected that the network can better imitate expert behaviors with extra gaze information.

We then describe the two ways we incorporated the estimated gaze map into the imitation network. Finally, we present the method used to evaluate the model uncertainty.

A. Experimental Setup

Fig. 1 shows the setup for data collection and testing. During data collection, the human expert controlled the car using the steering wheel while viewing the driving scenes on the computer monitor. An eye tracker recorded gaze location simultaneously for every frame. We stored data as image-gaze-steering action tuples. The eye tracker used in this experiment is Tobii Pro X60 commercial eye tracker, whose error is 0.6° after nine-point calibration. For our experiment, one degree of visual angle corresponds to about 54 pixels.

The experiments were conducted in TORCS simulator [21], which simulates highway driving and supports different tracks designs and settings. We collected data on five tracks with diverse routes and scenarios. For each track, the subject drove the car for four trials, each lasting about three minutes. We set the car to run at a constant speed. Drivers were asked to follow the road and overtake other cars to avoid collisions. The driver pressed a button doing each overtaking maneuver so that we could segment data into lane following versus overtaking. Four subjects participated in data collection. We have eighty trials collected in total. For each trial, we collected an average of 2500 action-image-gaze map tuples. Thus, our entire data set consists of 200 000 tuples.

About 40 000 image-gaze-action tuples from the first two trials of two tracks (16 trials in total) were selected as the training data set. As the eye gaze was located at the center of the visual scene most of the time, we created a balanced data set containing about 4 000 tuples for gaze network training by downsampling the training data set randomly. The three remaining tracks were used for open-loop testing. We also used the *TORCS* simulator for closed-loop driving tests, where the trained networks generated steering commands input to the simulator. During these closed-loop tests, we measured the percentage of cars successfully overtaken, and the average distance traveled between infractions. Infractions are defined by collisions or driving outside the lane.

B. Gaze Map Synthesis

The gaze network was trained to synthesize gaze maps on pairs of driver-view images and real gaze maps similar to the way deep networks have been trained to generate saliency maps [22].

For each frame, the ground-truth gaze map was generated from the gaze data collected in a sliding window of 10 frames centered on the current frame. We generated a 2-D probability distribution by placing a 2-D circularly symmetric Gaussian at each gaze point. The standard deviation of the Gaussian, $\sigma = 2.6$ degrees of visual angle, was chosen by a grid search to yield the best performance on a validation set.

Most previous deep saliency models have used loss functions, such as L2 distance [22] or KL-divergence [23]. In our case, most units of the real gaze map have values close to zero. Only a few pixels have significant nonzero values. In our experiments, we found that using the L2 distance as a loss function resulted in estimated gaze maps where all pixels had close to zero values.

To avoid this problem, we estimated gaze maps using a conditional GAN following the *Pix2Pix* architecture [24]. A GAN trains both a generator network and a discriminator network that tries to differentiate between the generated and ground truth gaze maps. Generated gaze maps that are easily distinguished from ground truth gaze maps are penalized. This led to much better gaze map estimations since gaze maps with all zero values are not realistic.

The generator was a U-Net encoder–decoder architecture [25]. The discriminator network consists of

$$CB_{64} - CB_{128} - CB_{256} - CB_{512} - CB_{512} - CB_{512}$$

where CB_k denotes a 4×4 Convolution-BatchNorm-ReLU layer with k filters. The GAN was trained for 200 epochs (7 h on the collected data set with an NVIDIA GTX 970). We initialized the weights randomly from a Gaussian distribution with zeros mean and a standard deviation of 0.02. The batch size was set to 16.

C. Imitation Network

Fig. 2 shows the two proposed frameworks for incorporating gaze information. They shared the same network for gaze map synthesis but differed in the way the gaze map is used in the imitation network.

The imitation network followed a deep CNN architecture (PilotNet [1]) to estimate steering actions. It comprises five convolutional (Conv) layers and four FC layers. The activation function for the networks was the ReLU. Two CNN networks with the same structure but different parameters were trained: one for following and one for overtaking. The driving maneuver (overtaking or lane following) to follow at each point was based on the driver input.

Weights were initialized in the same way as the gaze network. The batch size was 128. The size of the input image was 66×200. The steering output was a continuous number between -1 and 1.

We implemented two different methods to incorporate gaze information into the driving network, as shown in Fig. 2.

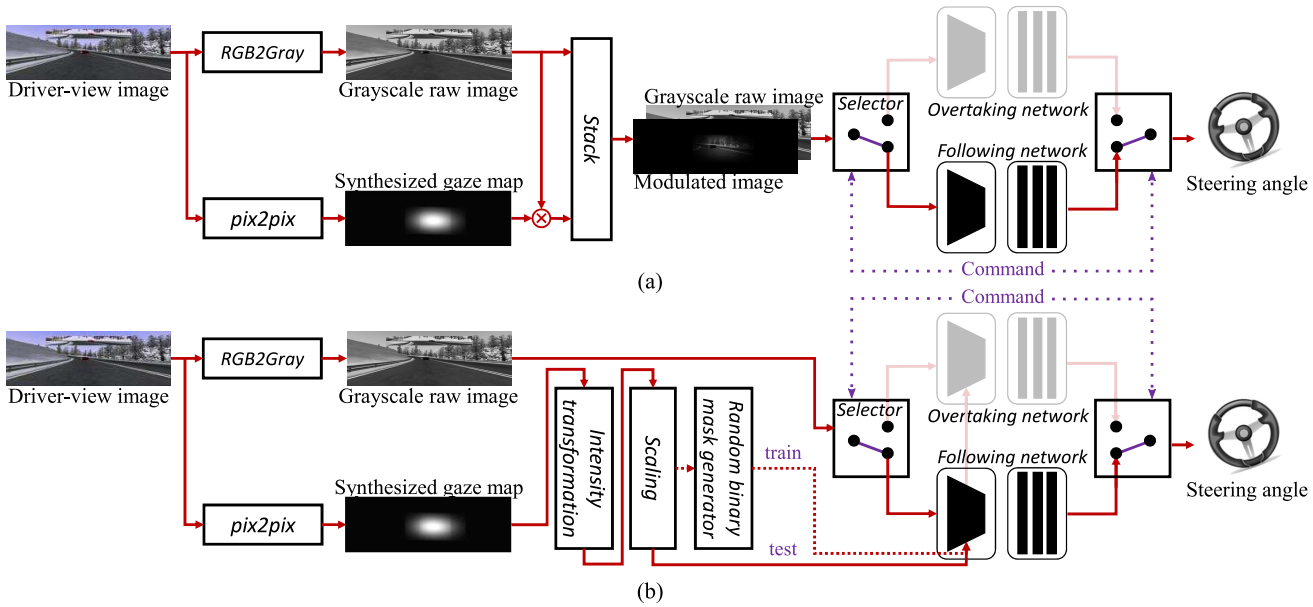


Fig. 2. Two approaches for incorporating gaze information into autonomous driving. (a) For the model that uses the gaze map as an additional input, the gaze map modulates the input image by pixelwise multiplication. The grayscale image and the modulated image are stacked and input to the network. For the model that uses the gaze map to modulate the dropout probability, the gaze map was first used to calculate the keep probability map through (1). During the training (dashed line), the keep probability map is scaled to the size of the feature map to generate the dropout binary mask. During the testing (solid line), we directly modulate the features by the scaled keep probability map. The command switches the input to the steering angle between the networks for overtaking or lane following. (a) With gaze map as input. (b) With gaze modulated dropout.

1) With Gaze Map as Input: We used the synthesized gaze map from the gaze network to create an additional input to the network. As shown in Fig. 2(a), we multiplied the grayscale original driver-view image pixelwise with the estimated gaze map. The modulated and driver-view images were stacked as input to the imitation network.

2) With Gaze Modulated Dropout: We used the synthesized gaze map for a keep probability map to modulate the dropout in the network. As shown in Fig. 2(b), gaze modulated dropout is applied to the convolutional layers of the imitation networks.

Based on the idea that saccades focus attention on important areas and de-emphasize task-irrelevant areas, we utilized the gaze to spatially modulate the keep probability (the complement of the dropout probability) to be higher at gaze points and lower elsewhere, as shown in Fig. 3. We first applied gaze modulated dropout to each convolutional layer individually. We found that dropout in lower layers results in lower error than dropout in higher layers on the validation set (Conv1: 2.84 deg, Conv2: 2.85 deg, Conv3: 2.87 deg, Conv4: 3.06 deg, Conv5: 3.18 deg). Based on this finding, we applied gaze modulated dropout cumulatively to successive convolutional layers starting from the lowest layer and found that applying gaze modulated dropout to the first two layers resulted in the best performance. This setting is used in all results reported in the following.

For uniform dropout, we generated a random array with the same size as the feature map independently at each location according to a standard uniform distribution on [0, 1]. This was converted to a binary mask by setting a pixel to one if the corresponding position in the random array was smaller than the keep probability KP and to zero otherwise. The feature map activation was modulated by the binary mask.

For gaze modulated dropout, the procedure was similar, but KP varied from pixel to pixel according to

$$\mathrm{KP} = (1 - dp) + dp\frac{G - G_{\min}}{1 - G_{\min}} \tag{1}$$

where G is the gaze map scaled to the range $[G_{\min}, 1]$, $G_{\min}$ is the minimum value of the gaze map over all pixels and $dp \in [0, 1]$ is the maximum drop probability.

At test time, we did not apply dropout. Instead, following [26], we approximated the effect of averaging by modulating the feature map by the keep probability KP.

D. Validation of the Gaze Effect

To determine whether gaze-modulated dropout helped to improve the model's generalization capability, we measured the epistemic uncertainty of several models that shared the same structure but were trained with different dropout methods. A model with better generalization should have lower epistemic uncertainty for unseen inputs.

As introduced in Section II-C, stochastic dropout is a practical and effective approach to modeling epistemic uncertainty. With x as the input, we measure the variance of the output of the model $y = f(x)$ by using multiple forward passes using the same input, but different initializations of the weights for each forward pass. Denoting $f_n(\cdot)$ to be the mapping from input to output for the dropout masked weight for pass n, we evaluated

$$\overline{y} = \frac{1}{N}\sum_{n=1}^{N} f_n(x) \tag{2}$$

$$\sigma^2(y) = \frac{1}{N}\sum_{n=1}^{N}(f_n(x) - \overline{y})^2. \tag{3}$$

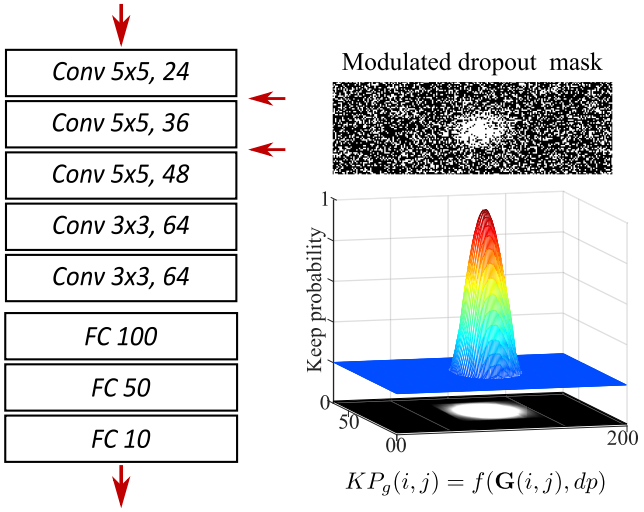


Fig. 3. Details of the implementation of gaze modulated dropout. The blocks on the left show the structure of the driving network. We applied the gaze modulated dropout after the first and second convolutional (Conv) layers. The keep probability for gaze-modulated dropout is shown on the right. At the training stage, keep probabilities are utilized to control whether units are kept or dropped. A typical gaze modulated dropout mask is shown on the top left. At the testing state, similar to dropout, the gaze modulated dropout multiplies the feature maps pixelwise with the keep probability maps.

TABLE I
KL DIVERGENCE AND CC BETWEEN REAL AND ESTIMATED GAZE MAPS

		Seen tracks	Unseen tracks
KL	Estimated gaze	**0.69**	**0.88**
	Central blob	2.83	2.32
CC	Estimated gaze	**0.86**	**0.83**
	Central blob	0.70	0.76

IV. RESULTS

A. Gaze Network Evaluation

Fig. 4 shows examples of estimated gaze maps superimposed with ground truth gaze trajectories. There is a strong overlap between the estimated gaze map and the actual trajectories.

To evaluate the gaze network quantitatively, we compute the two standard metrics for similarity evaluation: the Kullback–Leibler divergence (KL) and the correlation coefficient (CC). Smaller KL and larger CC denote better similarity.

From the observation of the ground truth gaze trajectories, we found that the subject mostly looks at the center region of the image. Therefore, we considered a static gaze map consisting of a single Gaussian at the image center as a reference.

The estimated gaze map closely matches the real-gaze map (Table I). Compared with the baseline (central Gaussian blob), the average KL divergence between the estimated gaze map and the real gaze map is markedly smaller (75.6% for seen tracks and 62.1% for unseen tracks). The CC between the estimated gaze map and the real-gaze map is significantly larger (22.9% for seen tracks and 9.2% for unseen tracks).

Please refer to our video (https://sites.google.com/view/gazedriving) to see gaze map predictions for both trained and unseen tracks. From the video, we can see that, besides mirror fixations, the human gaze data also deviates from the center of the image to track the car in front. This suggests that gaze information helps the network concentrate on task-relevant areas. Consistent with that, we found that 60.7% of the gaze is located within a circle around the center with a radius equal to one standard deviation of the center Gaussian blob, 5.7% is located in the mirror, and 33.6% is located in other regions.

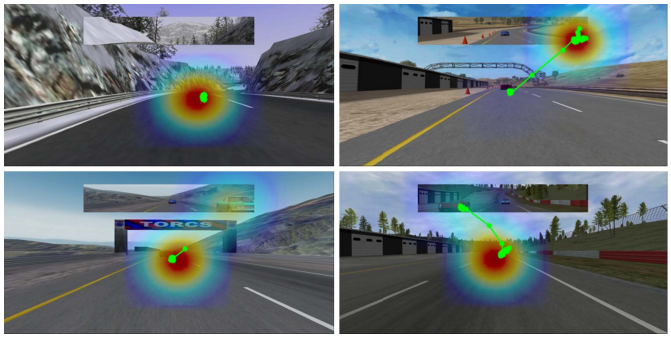


Fig. 4. Estimated gaze map and ground truth gaze trajectories visualized as heatmaps and green lines superimposed on the driver-view images. The first row is from environments seen during training. The second row is from two environments unseen during training. On the heatmaps, red areas indicate areas with more fixations.

B. Imitation Network Evaluation

An imitation network with only image input was trained utilizing uniform dropout as the baseline. We refer to it as **No gaze** in the results. We compared the performance of five different cases.

Real Gaze as Input: The ground truth gaze map was used to generate the gaze-modulated driver-view image provided as an additional input to the imitation network.

Estimated Gaze as Input: The estimated gaze map was used to generate the gaze-modulated driver-view image provided as an additional input to the imitation network.

Real Gaze Dropout: The ground truth gaze map was used to modulate the dropout probability.

Estimated Gaze Dropout: The estimated gaze map was used to modulate the dropout probability.

Center Blob Dropout: We observed that the subjects' gaze map was concentrated primarily at the center of the scene. To rule out the possibility that the effects observed are primarily due to increasing the weighting in the center part of the scene over the periphery, we also considered a model where the keep probability was modulated by a single-Gaussian blob located at the center of the scene. The variance of the Gaussian blob was chosen by the grid search to yield the best performance on the validation set.

1) Test on Data Set: We chose the dropout probability dp for both uniform and gaze modulated dropout by a brute-force search over the range from 0.1 to 0.8 with search step 0.1. The same dp settings were used for both training and testing. Fig. 5 shows the mean average error (MAE) for the models utilizing uniform and gaze modulated dropout. MAE on the seen tracks remains nearly constant at around 3° overall values

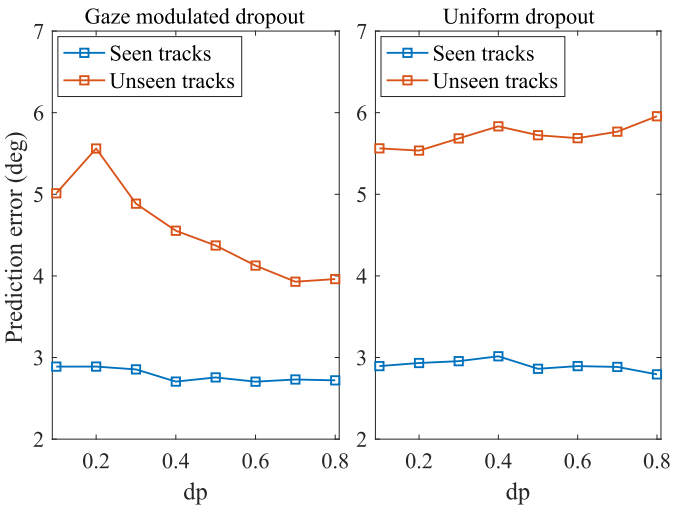


Fig. 5. Results of a brute-force search for the optimal values of dp for models utilizing gaze-modulated dropout and uniform dropout.

TABLE II

MEAN ABSOLUTE ERROR BETWEEN COMMANDS GENERATED BY THE SYSTEM AND THE HUMAN DRIVER ON THE TESTING SET

	Seen tracks (deg)	Unseen tracks (deg)
No gaze	2.90	5.58
Real gaze as input	2.86	4.29
Estimated gaze as input	2.85	4.63
Real gaze dropout	**2.82**	**4.00**
Estimated gaze dropout	2.84	4.27
Central blob dropout	2.84	4.67

of dp for the two dropout methods. However, the MAE for gaze modulated dropout on unseen tracks drops remarkably with dp increasing from 0.2 to 0.8 for unseen tracks. This is in stark contrast with the slight increase of the MAE for uniform dropout.

Based on the results of this experiment, we set dp to 0.1 for uniform dropout and to 0.7 for gaze modulated dropout for all following experiments, unless specified otherwise.

We evaluated the MAE between the commands generated by the various models and the human driver. The testing results shown in Table II demonstrate that both approaches decrease the action estimation error, especially in unseen environments. Networks with (real and estimated) gaze input outperform the baseline (No gaze) by 20.1% and outperform the central blob dropout network by 4.5% for unseen tracks on average. Networks with (real and estimated) gaze dropout give even greater improvements. They outperform the baseline (No gaze) by 25.9% and the central blob dropout network by 11.5% for unseen tracks on average.

Gaze modulated dropout performs better than gaze as input for both real gaze and estimated gaze. Using the real gaze, dropout outperforms input by 6.8% on unseen tracks and 1.4% on seen tracks. Using the estimated gaze, the improvements are 7.8% and 0.35%.

Gaze data used in our experiments and training was collected when subjects were driving alone without distractions.

TABLE III

QUANTITATIVE PERFORMANCE RUNNING ON AN UNSEEN TRACK IN THE SIMULATOR

		w/cars	w/o cars
Success rate of cars overtaking (%)	No gaze	67.6	N/A
	Gaze as input	85.2	N/A
	Gaze dropout	**88.9**	N/A
Ave. dist. traveled between two infractions (km)	No gaze	0.40	0.48
	Gaze as input	0.54	0.67
	Gaze dropout	**0.61**	**0.79**

In real environments, gaze might be noisier, e.g., if the driver is distracted by talking to a passenger. Thus, using gaze data collected in real-unconstrained environments may result in poorer performance. However, in autonomous driving applications, we expect that estimated, rather than actual, the gaze will be used.

2) Closed-Loop Performance: We evaluated the closed-loop performance of the models utilizing estimated gaze in the simulator for an unseen track. Gaze maps are obtained from the gaze network running in real time. The network with uniform dropout (No gaze), the network with estimated gaze map as input (Gaze as input) and the network utilizing estimated gaze-modulated dropout (Gaze dropout) are compared. For each episode, the starting position of the car was selected randomly. The models controlled the car to follow the lane and to overtake cars if required. The choice of whether to follow the lane or to overtake was made by human drivers observing the simulator. Human drivers also intervened to bring the car back to the middle of the lane when infractions (collisions or lane departures) occurred. We tested the models in two different settings: without other cars running on the road (w/o cars) and with cars (w/cars). For the setting w/o cars, the agent only performs lane following. We evaluated the performance by the success rate in overtaking cars (w/cars only), and the average distance traveled between infractions. Higher numbers are better for both measurements.

As shown in Table III, both approaches improve closed-loop performance. Consistent with our previous result, gaze dropout outperforms gaze as input in both measures. The success rate of overtaking cars for the network with estimated gaze dropout is 31.5% better than the baseline and 4.3% better than the network with an estimated gaze as input. The average distance traveled between two infractions for the network with estimated gaze dropout is 58.5% better than the baseline and 15.4% better than the network with an estimated gaze as input on average in the two environments.

Examples of the closes-loop behavior of the models with uniform and gaze modulated dropout in an unseen environment are shown in our video (https://sites.google.com/view/gazedriving).

C. Dropout Versus Inverted Dropout

As discussed in the original dropout article [26], it is too time-consuming to average exponentially many predictions during testing, especially for real-time applications. There are two ways typically proposed to deal with this. In standard

TABLE IV
Epistemic (MODEL) UNCERTAINTY

	Seen tracks (deg^2)	Unseen tracks (deg^2)
No gaze (dp=0.7)	1.23	4.37
Center blob dropout (dp=0.66)	1.29	3.88
Real gaze dropout (dp=0.7)	1.35	**2.23**
Estimate gaze dropout (dp=0.7)	**1.08**	2.60

dropout, during testing, the output of each unit is scaled by the keep probability. In inverted dropout, during training, the output of each unit is multiplied not only by the binary dropout mask but also by the multiplicative inverse of the keep probability. Then at test time, no scaling is applied. The advantage of inverted dropout is that the network during deployment is much simpler. In our case, there would be no need to estimate the gaze map. Prior work has suggested that standard uniform dropout and inverted uniform dropout perform similarly. Unfortunately, in our experiments, inverted gaze modulated dropout performed much worse than standard gaze modulated dropout. For example, on seen tracks using standard real gaze modulated dropout, the MAE was around 3°. However, the MAE for inverted real gaze modulated dropout can reach 6°degrees. Thus, it is still important to estimate the gaze map during testing.

We also compared the performance of scaling the output activations by the keep probability during testing with averaging the outputs of 50 networks with different binary masks sampled independently according to the keep probability. We found that the testing errors to be quite similar to those shown in Table II, differing by only 0.08° on average. Thus, simply scaling the output activations by the keep probability is a good choice during deployment, since it is simpler, more computationally efficient, and results in a similar performance.

D. Discussion on Model Uncertainty

Table IV shows epistemic (model) uncertainty of models using uniform (No gaze), gaze modulated, and center Gaussian blob modulated dropout as computed using (3). For a fair comparison, we unified the average drop probability of these models by setting the parameter dp accordingly. To be specific, dp was set to 0.66 for uniform dropout, 0.7 for gaze and center Gaussian blob modulated dropout. Note that the definition of drop probability is different for modulated dropout and uniform dropout.

Model uncertainties for unseen tracks are much higher than for seen tracks (Table IV). This is consistent with our interpretation of model uncertainty as reflecting the familiarity of the model with the input. In addition, there is little difference between the epistemic uncertainties of the different models for seen tracks.

To compare the generalization capability of different models, we concentrated on the results for unseen tracks. Lower model uncertainties were achieved by gaze modulated dropout, with real-gaze dropout achieving the lowest uncertainty. The uncertainty for estimated gaze dropout was 14.2% higher, but still much lower than for center blob dropout (33.0% higher) and no gaze (40.5% higher), which was the worst among the four.

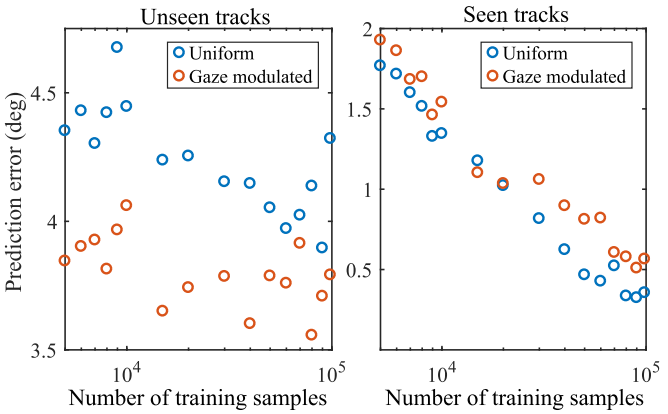


Fig. 6. Prediction error of models trained with different size of training data set.

Combining with the results from Section IV-B1, we find that models with lower epistemic uncertainty also have smaller prediction errors. This is consistent with the finding of [27], which showed epistemic uncertainty and positional error were positively correlated in a camera relocalization task.

E. Data Efficiency

We conducted experiments varying the number of training samples to compare the data efficiency of gaze-modulated and uniform dropout. The data set used in this section contains about 100 000 image-action pairs from the first four tracks. We obtained smaller training data sets by randomly sampling from the larger data set. Data from the fifth track were reserved for testing.

Fig. 6 compares the prediction errors of the two dropout methods in unseen and seen environments as the number of training samples varies. For seen tracks, the prediction error is comparable for the two methods. The error reduces as the number of training samples increases.

For unseen tracks, the error for uniform dropout decreases much more slowly than for gaze modulated dropout. Gaze modulated dropout achieves similar performance as a uniform dropout with much fewer training samples.

Thus, the network with gaze modulated dropout has better data efficiency.

F. CNN Visualization

To illustrate how incorporating the gaze map helps the imitation network, we utilize Grad-CAM [28], a technique for visual explanations of deep neural networks.

Fig. 7 shows the four examples of Grad-CAM visualizations for the vanilla imitation network (without gaze information) and the imitation network with gaze map. These visualizations highlight the regions that contribute most to the output of the networks. For the vanilla imitation network, a significant proportion of the contributing areas (shown in red) is distributed in the background. For the imitation network with gaze map, the contributing areas are mainly associated with the road and car in front and less on the background.

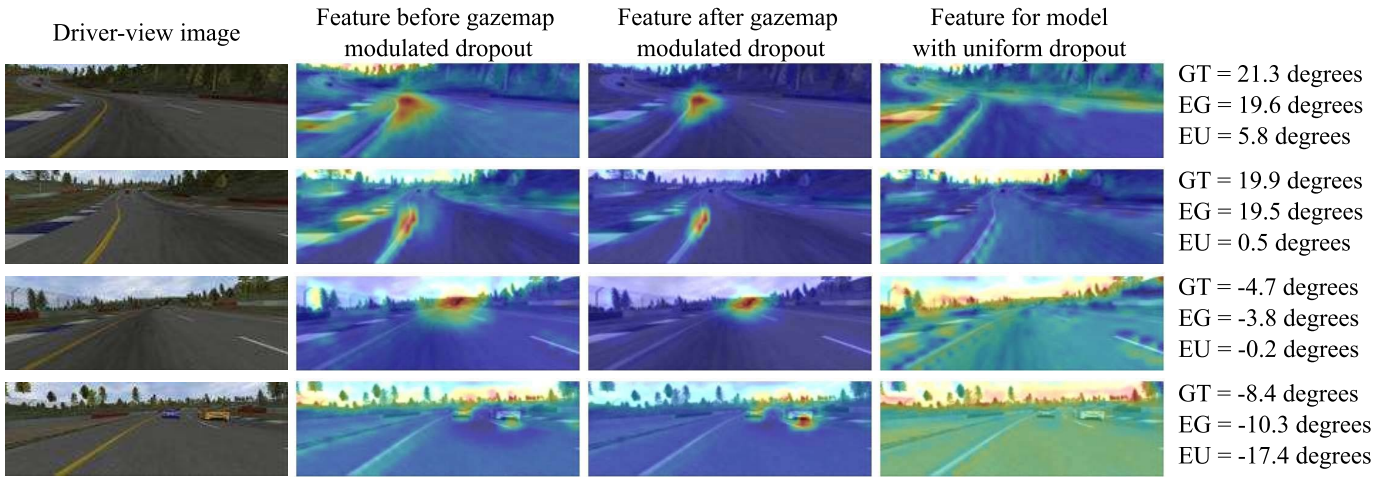


Fig. 7. Visualization of the features by Grad-Cam [28]. For better visualization, we enlarged the road area and overlapped the features with the grayscale image. Red regions correspond to high importance to the steering output. The first column shows the scaled driver-view images. The following columns show the features of models after the second convolutional layer. The second and third columns show the features of models trained with gaze modulated dropout, while the third column displays the features of models trained with uniform dropout. The corresponding steering angles of the scenes are given with GT marking the human demonstrated action. EU and EG mark the estimated steering angle given by the model trained with uniform and gaze modulated dropout, respectively.

These results suggest that the incorporation of the gaze map helps the imitation network concentrate on task-related areas and ignore irrelevant areas, such as the scene background. This helps to explain why gaze modulated dropout leads to much lower MAE and better closed-loop performance on unseen scenes, where the main differences with those observed during training are in the background.

V. Conclusion

This article proposed the use of gaze information contained in expert demonstrations to improve the generalization of IL networks for autonomous driving to unseen environments. We show that a conditional GAN can estimate human gaze maps accurately during driving. We studied two ways to incorporate gaze information. Both significantly improve human action estimation accuracy. Better performance is obtained with the gaze-modulated dropout.

This article demonstrates for the first time that it is possible to incorporate human information about gaze behavior into deep driving networks so that they receive similar benefits novice human drivers do when viewing expert gaze patterns. By exploiting expert behavior implicitly, this article makes an effort to raise IL to the next level. Furthermore, we found that imitation networks with gaze information have lower model uncertainty and have better data efficiency. We also show that integrating gaze enables the network to focus more on task-relevant information.

There are several potential directions to extend this article. First, the current gaze network and policy network do not take spatiotemporal aspects of eye gaze into account. Incorporating a spatiotemporal unit, such as recurrent module, may benefit the gaze prediction and action prediction. Second, human gaze data involves rich information regarding human intent. Estimated gaze maps may also be useful for the selection of the driving maneuvers, e.g., selecting automatically between lane following and car overtaking, which is currently manually selected in our experiments. Finally, other visuomotor tasks, such as vision-based robot manipulation and robot navigation, may also benefit from the incorporation of the proposed gaze-modulated dropout method.

References

[1] M. Bojarski *et al.*, "End to end learning for self-driving cars," 2016, *arXiv:1604.07316*. [Online]. Available: http://arxiv.org/abs/1604.07316

[2] F. Codevilla, M. Muller, A. Lopez, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1–9.

[3] J. Zhang *et al.*, "VR-goggles for robots: Real-to-sim domain adaptation for visual control," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1148–1155, Apr. 2019.

[4] U. Muller, J. Ben, E. Cosatto, B. Flepp, and Y. L. Cun, "Off-road obstacle avoidance through end-to-end learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 739–746.

[5] S. J. Vine, R. S. W. Masters, J. S. McGrath, E. Bright, and M. R. Wilson, "Cheating experience: Guiding novices to adopt the gaze strategies of experts expedites the learning of technical laparoscopic skills," *Surgery*, vol. 152, no. 1, pp. 32–40, Jul. 2012.

[6] Y. Yamani, P. Bıçaksız, D. B. Palmer, J. M. Cronauer, and S. Samuel, "Following expert's eyes: Evaluation of the effectiveness of a gaze-based training intervention on young drivers' latent hazard anticipation skills," in *Proc. 9th Int. Driving Symp. Hum. Factors Driver Assessment, Training, Vehicle Design*, 2017, pp. 1–8.

[7] A. Palazzi, D. Abati, S. Calderara, F. Solera, and R. Cucchiara, "Predicting the driver's focus of attention: The DR(eye)VE project," 2017, *arXiv:1705.03854*. [Online]. Available: http://arxiv.org/abs/1705.03854

[8] C. Liu, Y. Chen, L. Tai, H. Ye, M. Liu, and B. E. Shi, "A gaze model improves autonomous driving," in *Proc. 11th ACM Symp. Eye Tracking Res. Appl.* New York, NY, USA: ACM, Jun. 2019, pp. 1–5.

[9] R. McAllister *et al.*, "Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 4745–4753, doi: 10.24963/ijcai.2017/661.

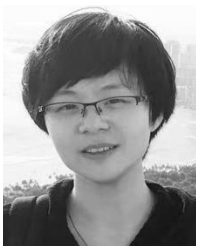
[10] X. Pan, Y. You, Z. Wang, and C. Lu, "Virtual to real reinforcement learning for autonomous driving," 2017, *arXiv:1704.03952*. [Online]. Available: http://arxiv.org/abs/1704.03952

[11] X. Liang, T. Wang, L. Yang, and E. Xing, "Cirl: Controllable imitative reinforcement learning for vision-based self-driving," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 584–599.

[12] Z. Chen and X. Huang, "End-to-end learning for lane keeping of self-driving cars," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 1856–1860.

[13] Q. Ji, Z. Zhu, and P. Lan, "Real-time nonintrusive monitoring and prediction of driver fatigue," *IEEE Trans. Veh. Technol.*, vol. 53, no. 4, pp. 1052–1068, Jul. 2004.

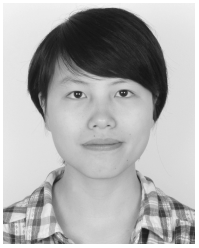
[14] R. Wang, P. V. Amadori, and Y. Demiris, "Real-time workload classification during driving using hypernetworks," 2018, *arXiv:1810.03145*. [Online]. Available: https://arxiv.org/abs/1810.03145
[15] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara, "DR(eye)VE: A dataset for attention-based tasks with applications to autonomous and assisted driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 54–60.
[16] A. G. Kendall, "Geometry and uncertainty in deep learning for computer vision," Ph.D. dissertation, Dept. Eng., Univ. Cambridge, Cambridge, U.K., 2019.
[17] D. Feng, L. Rosenbaum, and K. Dietmayer, "Towards safe autonomous driving: Capture uncertainty in the deep neural network for Lidar 3D vehicle detection," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3266–3273.
[18] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, Apr. 2019.
[19] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," 2015, *arXiv:1511.02680*. [Online]. Available: http://arxiv.org/abs/1511.02680
[20] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5574–5584.
[21] B. Wymann, E. Espié, C. Guionneau, C. Dimitrakakis, R. Coulom, and A. Sumner, "Torcs, the open racing car simulator," *Softw. Sourceforge. net*, vol. 4, no. 6, p. 2, 2000. [Online]. Available: http://torcs
[22] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3488–3493.
[23] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 262–270.
[24] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.
[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[27] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 4762–4769.
[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.



Congcong Liu (Graduate Student Member, IEEE) received the B.Eng. degree in information engineering from Zhejiang University, Hangzhou, China, in 2015. She is currently pursuing the Ph.D. degree in electronic and computer engineering with The Hong Kong University of Science and Technology, Hong Kong.

Her research interests include gaze tracking, visual attention, and computer vision.

Ms. Liu was a recipient of the Hong Kong Ph.D. Fellowship.



Yuying Chen (Student Member, IEEE) received the B.S. degree in information engineering from Zhejiang University, Hangzhou, China, in 2015. She is currently pursuing the Ph.D. degree with the Department of Electronics and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong.

Her research interests include mobile robot navigation, trajectory planning, and autonomous driving.

Ms. Chen was a recipient of the Hong Kong Ph.D. Fellowship.

Ming Liu (Senior Member, IEEE) received the B.A. degree in automation from Tongji University, Shanghai, China, in 2005, and the Ph.D. degree from the Department of Mechanical and Process Engineering, ETH Zürich, Zürich, Switzerland, in 2013, under the supervision of Prof. R. Siegwart.

He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, where he is also an Assistant Professor with the Department of Computer Science and Engineering. He is the Principal Investigator of over 20 projects, funded by the Research Grants Council (RGC), the National Natural Science Foundation of China (NSFC), the Innovation and Technology Commission (ITC), and the Shenzhen Science, Technology and Innovation Commission (SZSTI). His research interests include dynamic environment modeling, deep-learning for robotics, 3-D mapping, machine learning, and visual control.

Prof. Liu was a recipient of the Best RoboCup Paper Award at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2013. He received the innovation contest Chunhui Cup Winning Award in 2012 and 2013, respectively. He was also a recipient of the 2018 IEEE IROS Toshio Fukuda Young Professional Award.

Bertram E. Shi (Fellow, IEEE) is currently a Professor and the Head of the Department of Electronics and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong. His research interests are in bio-inspired signal processing and robotics, neuromorphic engineering, computational neuroscience, developmental robotics, machine vision, image processing, and machine learning.

Prof. Shi has served as a Distinguished Lecturer for the IEEE Circuits and Systems Society. He has also served on the editorial boards of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, the IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS and *Frontiers in Neuromorphic Engineering*. He served as the Chair for the IEEE Circuits and Systems Society Technical Committee on Cellular Neural Networks and Array Computing and as the general chair and the technical program chair for conferences in that area.