UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Driver Eye Movements and the Application in Autonomous Driving

Permalink
https://escholarship.org/uc/item/5xk929x8

Author
Xia, Ye

Publication Date
2019

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Driver Eye Movements and the Application in Autonomous Driving

by

Ye Xia

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor David Whitney, Chair
Professor Ken Nakayama
Professor John Canny
Professor Richard Ivry

Summer 2019

Driver Eye Movements and the Application in Autonomous Driving

Copyright 2019
by
Ye Xia

Abstract

Driver Eye Movements and the Application in Autonomous Driving

by

Ye Xia

Doctor of Philosophy in Psychology

University of California, Berkeley

Professor David Whitney, Chair

Despite the exciting progress in computer vision in the field of autonomous driving, understanding efficiently which cues or objects are the most crucial ones in a crowded traffic scene is still a big challenge. Human drivers can quickly identify the important visual cues or objects in their blurry periphery vision and then make eye movements to direct their more accurate foveal vision to the important regions. Therefore, driver eye movements may be what computer vision can borrow from human vision to make autonomous driving systems better at locating and understanding the important regions of crowded traffic scenes. Meanwhile, the large-scale datasets and advanced object recognition algorithms that emerged in the field of autonomous driving make it possible to study classical human vision science problems in natural driving situations.

Here, we used driver eye movements to improve autonomous driving models and studied visual crowding—the bottleneck of human object recognition—in realistic driving situations through driver eye movements. First, we developed a new protocol that collects driver eye movements in an offline manner for large-scale driving video datasets. We built a deep neural network that predicts human driver gaze from dash camera videos for various driving scenarios. Our model outperformed the current state-of-the-art model. Furthermore, we incorporated the driver gaze prediction model into an autonomous driving model to make a new periphery-fovea multi-resolution driving model that predicts vehicle speed from dash camera videos. This model combines low-resolution input of the whole video frames and high-resolution input from predicted gaze locations to predict vehicle speed. We show that the added human gaze significantly improves the driving accuracy and that our periphery-fovea multi-resolution model outperforms a uni-resolution periphery-only model that has the same amount of floating-point operations. Finally, we studied visual crowding in driving situations. We show that crowding occurs in natural driving scenes and that the degree of crowding correlates with altered saccade localization in realistic driving-like situations. Together, these studies demonstrate the application of driver eye movements in making safer and more efficient autonomous driving models and show strong evidence of visual crowding in driving situations via the analysis of driver eye movements. These studies also present examples of

combining human vision and computer vision to get mutual benefits from both fields.

Dedication

This doctoral dissertation is dedicated to my parents, Fan Xia and Yonghong Li, for always supporting me to pursue whatever I love to do.

Contents

Acknowledgments

This work would not have been possible without the support and assistance of the members of the Whitney Lab over the last five years: Wesley Chaney, Zhimin Chen, Aneesa Conine-Nakano, Victor Hiltner, Anna Kosovicheva, Alina Liberman, Allison Yamanashi Leib, Mauro Manassi, Gerrit Maus, Yuki Murai, William Schloetel, Zixuan Wang, Benjamin Wolfe, Yuan Yuan, and Kathy Zhang. Sincere thanks to my collaborators of this research: Jinkyu Kim and Danqing Zhang. Special thanks to David Whitney, Ken Nakayama, Bill Prinzmetal, Karl Zipser and John Canny for their advice, ideas, and support.

This research was supported by funding from Berkeley DeepDrive.

Chapter 1

Introduction

In this era of deep learning, exciting progress has been made in computer vision approaching autonomous driving. However, how to quickly identify important visual cues and understand risks in crowded traffic scenes remains a major obstacle in achieving safe autonomous driving. Human vision can provide critical insights into this problem. Human drivers are able to quickly identify and locate hazards or important visual cues that occur in their blurry peripheral visual fields and make saccadic eye movements to those regions to further process the input with their foveal vision. Can driver eye movements be used to teach computer vision models to understand the important cues and potential risks in crowded traffic scenes? Can the periphery-fovea structure of human vision be applied to autonomous driving models and improve their performances?

Similarly, the recent advances in computer vision can also help better understand human vision science, especially in the study of driver eye movements. Driver eye movements are often behavioral consequences of the recognition of objects in the periphery. The large-scale driving video datasets and advanced object detection algorithms emerged from the field of autonomous driving provide us ideal experiment material for studying human object recognition in realistic cluttered driving scenes through driver eye movements.

Here, we use interdisciplinary approaches to draw insights from driver eye movements for questions in both computer vision and human vision. The first necessary step of this research is to collect driver eye movements for various driving situations including the critical ones that contain immediate risks. This challenge is addressed in Chapter 2. We introduce a new in-lab driver eye movement collection protocol and the dataset collected with this new protocol. The protocol efficiently collects driver eye movements in an offline manner using diverse driving videos containing critical events and yields corresponding driver gaze/attention maps. We further present a driver gaze/attention prediction model that highlights important visual cues and potential risks in dynamic driving scenes.

In Chapter 3, we continue to discuss whether the predicted driver gaze/attention can guide autonomous driving models and improve their performances. We introduce a novel autonomous driving model that mimics the multi-resolution periphery-fovea design of human vision. The model initially processes the video frames in low resolution and predicts where human drivers would gaze. It then extracts high-resolution input from the predicted gaze locations. Finally, the model

combines the global low-resolution information and the local high-resolution information to predict the speed control of the vehicle. We show that the guidance of predicted human gaze improves the driving performance; most importantly, the performance gain is even more significant for critical situations involving pedestrians than for other non-critical cases. We also demonstrate that our multi-resolution periphery-fovea driving model outperforms a uni-resolution model that has the same amount of floating-point operations.

In Chapter 4, we use driver eye movements collected through our new protocol as a tool to study the major bottleneck of human object recognition in cluttered scenes, visual crowding. Visual crowding is generally defined as the deleterious influence of nearby visual inputs on object recognition. We demonstrate that visual crowding occurs in natural driving scenes and has behavioral consequences in driving-like situations (i.e., altered saccadic localization). This finding provides not only new knowledge about visual crowding but also important implications for driving safety.

Overall, the research discussed in this dissertation shows how driver eye movements can be influenced by the fundamental bottleneck of our visual object recognition and how driver eye movements can be used to improve autonomous driving models. This research also highlights the benefits of utilizing interdisciplinary approaches to address challenges in both human and computer vision.

Chapter 2

Predicting Driver Attention in Critical Situations

Human visual attention enables drivers to quickly identify and locate potential risks or important visual cues across the visual field, such as a darting-out pedestrian, an incursion of a nearby cyclist or a changing traffic light. Drivers' gaze behavior has been studied as a proxy for their attention. Recently, a large driver attention dataset of routine driving (Alletto, Palazzi, Solera, Calderara, & Cucchiara, 2016) has been introduced and neural networks (Palazzi, Solera, Calderara, Alletto, & Cucchiara, 2017; Tawari & Kang, 2017) have been trained end-to-end to estimate driver attention, mostly in lane-following and car-following situations. Nonetheless, datasets and prediction models for driver attention in rare and critical situations are still needed.

However, it is nearly impossible to collect enough driver attention data for crucial events with the conventional in-car data collection protocol, *i.e.*, collecting eye movements from drivers during driving. This is because the vast majority of routine driving situations consist of simple lane-following and car-following. In addition, collecting driver attention in-car has two other major drawbacks. (i) Single focus: at each moment the eye-tracker can only record one location that the driver is looking at, while the driver may be attending to multiple important objects in the scene with their covert attention, *i.e.*, the ability to fixate one's eyes on one object while attending to another object (Cavanagh & Alvarez, 2005). (ii) False positive gazes: human drivers also show eye movements to driving-irrelevant regions, such as sky, trees, and buildings (Palazzi et al., 2017). It is challenging to separate these false positives from gazes that are dedicated to driving.

An alternative that could potentially address these concerns is showing selected driving videos to drivers in the lab and collecting their eye movements with repeated measurements while they perform a proper simulated driving task. Although this third-person driver attention collected in the lab is inevitably different from the first-person driver attention in the car, it can still potentially reveal the regions a driver should look at in that particular driving situation from a third-person perspective. These data are greatly valuable for identifying risks and driving-relevant visual cues from driving scenes. To date, a proper data collection protocol of this kind is still missing and needs to be formally introduced and tested.

Another challenge for driver attention prediction, as well as for other driving-related machine

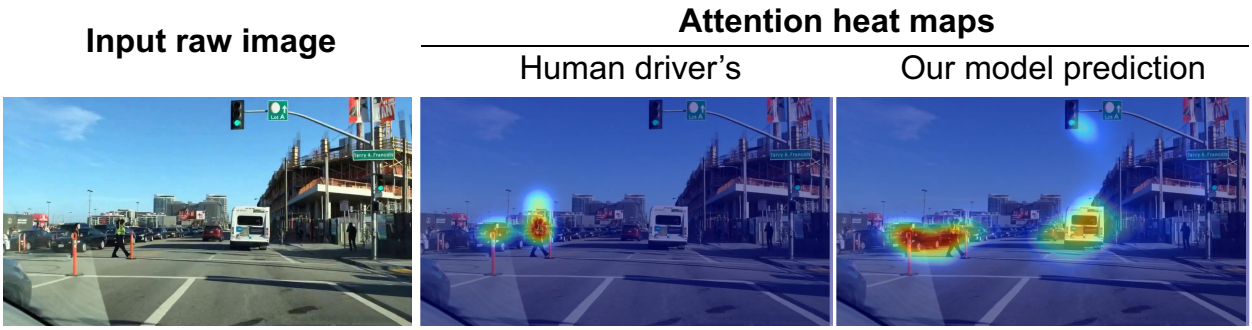

Figure 2.1: An example of input raw images (*left*), ground-truth human attention maps collected by us (*middle*), and the attention maps predicted by our model (*right*). The driver had to sharply stop at the green light to avoid hitting two pedestrians running the red light. The collected human attention map accurately shows the multiple regions that simultaneously demand the driver's attention. Our model correctly attends to the crossing pedestrians and does not give false alarms to other irrelevant pedestrians

learning problems, is that the actual cost of making a particular prediction error is unknown. Attentional lapses while driving on an empty road does not cost the same as attentional lapses when a pedestrian darts out. Since current machine learning algorithms commonly rely on minimizing average prediction error, the critical moments, where the cost of making an error is high, need to be properly identified and weighted.

Here, our paper offers the following novel contributions. First, in order to overcome the drawbacks of the conventional in-car driver attention collection protocol, we introduce a new protocol that uses crowd-sourced driving videos containing interesting events and makes multi-focus driver attention maps by averaging gazes collected from multiple human observers in lab with great accuracy (Fig. 2.1). We will refer to this protocol as the in-lab driver attention collection protocol. We show that data collected with our protocol reliably reveal where a experienced driver should look and can serve as a substitute for data collected with the in-car protocol. We use our protocol to collect a large driver attention dataset of braking events, which is, to the best of our knowledge, the richest to-date in terms of the number of interactions with other road agents. We call this dataset Berkeley DeepDrive Attention (BDD-A) dataset and will make it publicly available. Second, we introduce Human Weighted Sampling (HWS), which uses human driver eye movements to identify which frames in the dataset are more crucial driving moments and weights the frames according to their importance levels during model training. We show that HWS improve model performance on both the entire testing set and the subset of crucial frames. Third, we propose a new driver attention prediction model trained on our dataset with HWS. The model shows sophisticated behaviors such as picking out pedestrians suddenly crossing the road without being distracted by the pedestrians safely walking in the same direction as the car (Fig. 2.1). The model prediction is nearly indistinguishable from ground-truth based on human judges, and it also matches the state-of-the-art performance level when tested on an existing in-car driver attention dataset collected during driving.

Related works

Image / Video Saliency Prediction

A large variety of the previous saliency studies explored different bottom-up feature-based models (N. Bruce & Tsotsos, 2006; Valenti, Sebe, & Gevers, 2009; Erdem & Erdem, 2013; Murray, Vanrell, Otazu, & Parraga, 2011; Jianming Zhang & Sclaroff, 2013; N. D. Bruce & Tsotsos, 2009) combining low-level features like contrast, rarity, symmetry, color, intensity and orientation, or topological structure from a scene (Jianming Zhang & Sclaroff, 2013; Harel, Koch, & Perona, 2007; Wei, Wen, Zhu, & Sun, 2012). Recent advances in deep learning have achieved a considerable improvement for both image saliency prediction (Kümmerer, Theis, & Bethge, 2015; Xun Huang, Shen, Boix, & Zhao, 2015; N. Liu, Han, Zhang, Wen, & Liu, 2015; Kümmerer, Wallis, & Bethge, 2016) and video saliency prediction (Bazzani, Larochelle, & Torresani, 2016; Cornia, Baraldi, Serra, & Cucchiara, 2016; Y. Liu, Zhang, Xu, & He, 2017). These models have achieved start-of-the-art performance on visual saliency benchmarks collected mainly when human subjects were doing a free-viewing task, but models that are specifically trained for predicting the attention of drivers are still needed.

Driver Attention Datasets

DR(eye)VE (Alletto et al., 2016) is the largest and richest existing driver attention dataset. It contains 6 hours of driving data, but the data was collected from only 74 rides, which limits the diversity of the dataset. In addition, the dataset was collected in-car and has the drawbacks we introduced earlier, including missing covert attention, false positive gaze, and limited diversity. The driver's eye movements were aggregated over a small temporal window to generate an attention map for a frame, so that multiple important regions of one scene might be annotated. But there was a trade-off between aggregation window length and gaze location accuracy, since the same object may appear in different locations in different frames. Reference (Fridman, Langhans, Lee, & Reimer, 2016) is another large driver attention dataset, but only six coarse gaze regions were annotated and the exterior scene was not recorded. References (Simon, Tarel, & Brémond, 2009) and (Underwood, Humphrey, & Van Loon, 2011) contain accurate driver attention maps made by averaging eye movements collected from human observers in-lab with simulated driving tasks. But the stimuli were static driving scene images and the sizes of their datasets are small (40 frames and 120 frames, respectively).

Driver Attention Prediction

Self-driving vehicle control has made notable progress in the last several years. One of major approaches is a mediated perception-based approach – a controller depends on recognizing human-designated features, such as lane markings, pedestrians, or vehicles. Human driver's attention provides important visual cues for driving, and thus efforts to mimic human driver's attention have increasingly been introduced. Recently, several deep neural models have been utilized to predict where human drivers should pay attention (Palazzi et al., 2017; Tawari & Kang, 2017). Most of

Table 2.1: Comparison between driver attention datasets

Dataset	# Rides	Durations (hours)	# Drivers	# Gaze providers	# Cars (per frame)	# Pedestrians (per frame)	# Braking events
DR(eye)VE	74	6	8	8	1.0	0.04	464
BDD-A (ours)	1,232	3.5	1,232	45	4.4	0.25	1,427

existing models were trained and tested on the DR(eye)VE dataset (Alletto et al., 2016). While this dataset is an important contribution, it contains sparse driving activities and limited interactions with other road users. Thus it is restricted in its ability to capture diverse human attention behaviors. Models trained with this dataset tend to become vanishing point detectors, which is undesirable for modeling human attention in urban driving environment, where drivers encounter traffic lights, pedestrians, and a variety of other potential cues and obstacles. In this paper, we provide our human attention dataset as a contribution collected from a publicly available large-scale crowd-sourced driving video dataset (H. Xu, Gao, Yu, & Darrell, 2017a), which contains diverse driving activities and environments, including lane following, turning, switching lanes, and braking in cluttered scenes.

Berkeley DeepDrive Attention (BDD-A) Dataset

Dataset Statistics

The statistics of our dataset are summarized and compared with the largest existing dataset (DR(eye)VE) (Alletto et al., 2016) in Table 2.1. Our dataset was collected using videos selected from a publicly available, large-scale, crowd-sourced driving video dataset, BDD100k (H. Xu et al., 2017a; Yu et al., 2018). BDD100K contains human-demonstrated dashboard videos and time-stamped sensor measurements collected during urban driving in various weather and lighting conditions. To efficiently collect attention data for critical driving situations, we specifically selected video clips that both included braking events and took place in busy areas. We then trimmed videos to include 6.5 seconds prior to and 3.5 seconds after each braking event. It turned out that other driving actions, *e.g.*, turning, lane switching and accelerating, were also included. 1,232 videos (=3.5 hours) in total were collected following these procedures. Some example images from our dataset are shown in Fig. 2.6. Our selected videos contain a large number of different road users. We detected the objects in our videos using YOLO (Redmon & Farhadi, 2017).On average, each video frame contained 4.4 cars and 0.3 pedestrians, multiple times more than the DR(eye)VE dataset (Table 2.1).

Data Collection Procedure

For our eye-tracking experiment, we recruited 45 participants who each had more than one year of driving experience. The participants watched the selected driving videos in the lab while performing a driving instructor task: participants were asked to imagine that they were driving instructors sitting in the copilot seat and needed to press the space key whenever they felt it necessary to correct or warn the student driver of potential dangers. Their eye movements during the task were recorded at 1000 Hz with an EyeLink 1000 desktop-mounted infrared eye tracker, used in conjunction with the Eyelink Toolbox scripts (Cornelissen, Peters, & Palmer, 2002) for MATLAB. Each participant completed the task for 200 driving videos. Each driving video was viewed by at least 4 participants. The gaze patterns made by these independent participants were aggregated and smoothed to make an attention map for each frame of the stimulus video (see Fig. 2.6).

Psychological studies (Mannan, Ruddock, & Wooding, 1997; Groner, Walder, & Groner, 1984) have shown that when humans look through multiple visual cues that simultaneously demand attention, the order in which humans look at those cues is highly subjective. Therefore, by aggregating gazes of independent observers, we could record multiple important visual cues in one frame. In addition, it has been shown that human drivers look at buildings, trees, flowerbeds, and other unimportant objects non-negligibly frequently (Alletto et al., 2016). Presumably, these eye movements should be regarded as noise for driving-related machine learning purposes. By averaging the eye movements of independent observers, we were able to effectively wash out those sources of noise (see Fig. 2.2B).

Comparison with In-Car Attention Data

We collected in-lab driver attention data using videos from the DR(eye)VE dataset. This allowed us to compare in-lab and in-car attention maps of each video. The DR(eye)VE videos we used were 200 randomly selected 10-second video clips, half of them containing braking events and half without braking events.

We tested how well in-car and in-lab attention maps highlighted driving-relevant objects. We used YOLO (Redmon & Farhadi, 2017) to detect the objects in the videos of our dataset. We identified three object categories that are important for driving and that had sufficient instances in the videos (car, pedestrian and cyclist). We calculated the proportion of attended objects out of total detected instances for each category for both in-lab and in-car attention maps. The results showed that in-car attention maps highlighted significantly less driving-relevant objects than in-lab attention maps (see Fig. 2.2A).

The difference in the number of attended objects between the in-car and in-lab attention maps can be due to the fact that eye movements collected from a single driver do not completely indicate all the objects that demand attention in the particular driving situation. One individual's eye movements are only an approximation of their attention (Rizzolatti, Riggio, Dascola, & Umiltá, 1987), and humans can also track objects with covert attention without looking at them (Cavanagh & Alvarez, 2005). The difference in the number of attended objects may also reflect the difference between first-person driver attention and third-person driver attention. It may be that the human

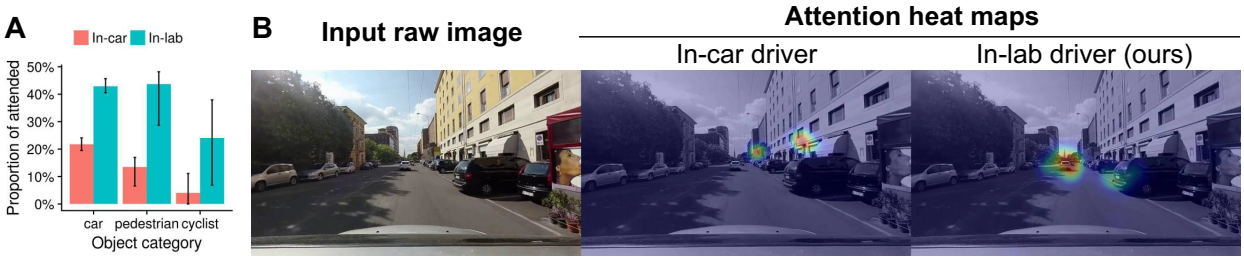

Figure 2.2: Comparison between in-car and in-lab driver attention maps. (A) Proportions of attended objects of different categories for in-car and in-lab driver attention maps. In-car attention maps tend to highlight significantly fewer driving-relevant objects than in-lab attention maps. (B) An example of in-car driver attention maps showing irrelevant regions. The in-lab attention map highlights the car in front and a car that suddenly backed up, while the in-car attention map highlights some regions of the building

observers in our in-lab eye-tracking experiment also looked at objects that were not relevant for driving. We ran a human evaluation experiment to address this concern.

Human Evaluation

To verify that our in-lab driver attention maps highlight regions that should indeed demand drivers' attention, we conducted an online study to let humans compare in-lab and in-car driver attention maps. In each trial of the online study, participants watched one driving video clip three times: the first time with no edit, and then two more times in random order with overlaid in-lab and in-car attention maps, respectively. The participant was then asked to choose which heatmap-coded video was more similar to where a good driver would look. In total, we collected 736 trials from 32 online participants. We found that our in-lab attention maps were more often preferred by the participants than the in-car attention maps (71% versus 29% of all trials, statistically significant as $p = 1 \times 10^{-29}$, see Table 2.2). Although this result cannot suggest that in-lab driver attention maps are superior to in-car attention maps in general, it does show that the driver attention maps collected with our protocol represent where a good driver should look from a third-person perspective.

In addition, we will show in the Experiments section that in-lab attention data collected using our protocol can be used to train a model to effectively predict actual, in-car driver attention. This result proves that our dataset can also serve as a substitute for in-car driver attention data, especially in crucial situations where in-car data collection is not practical.

To summarize, compared with driver attention data collected in-car, our dataset has three clear advantages: multi-focus, little driving-irrelevant noise, and efficiently tailored to crucial driving situations.

Table 2.2: Two human evaluation studies were conducted to compare in-lab human driver attention maps with in-car human driver attention maps and attention maps predicted by our HWS model, respectively. In-car human driver attention maps were preferred in significantly less trials than the in-lab human driver attention maps. The attention maps predicted by our HWS model were not preferred in as many trials as the in-lab human driver attention maps, but they achieved significantly higher preference rate than the in-car human driver attention maps

	# trials	Attention maps	Preference rate
Study 1	736	in-car human driver	29%
		in-lab human driver	71%
Study 2	462	HWS model predicted	41%
		in-lab human driver	59%

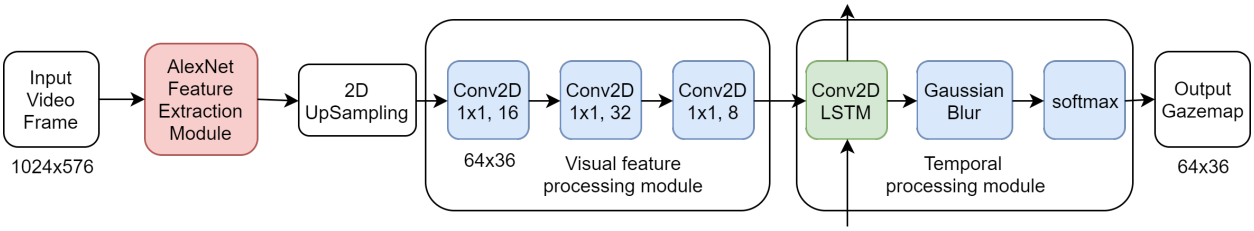

Figure 2.3: An overview of our proposed model that predicts human driver's attention from input video frame. We use AlexNet pre-trained on ImageNet as a visual feature extractor. We also use three fully convolutional layers (Conv2D) followed by a convolutional LSTM network (Conv2D LSTM)

Attention Prediction Model

Network Configuration

Our goal is to predict the driver attention map for a video frame given the current and previous video frames. Our model structure can be divided into a visual feature extraction module, a visual feature processing module, and a temporal processing module (Fig. 2.3).

The visual feature extraction module is a pre-trained dilated fully convolutional neural network, and its weights are fixed during training. We used ImageNet pre-trained AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) as our visual feature extraction module. We chose to use the features from the conv5 layer. In our experiment, the size of the input was set to 1024×576 pixels, and the feature map by AlexNet was upsampled to 64×36 pixels and then fed to the following visual feature processing module.

The visual feature processing module is a fully convolutional neural network. It consists of

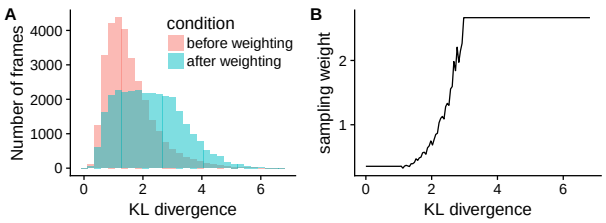

Figure 2.4: Human Weighted Sampling: (A) For each video frame, we measure the KL divergence between the collected driver attention maps and the mean attention map for that entire video clip (≈10s). We use this computed KL divergence as a weight value to sample image frames during training phase, *i.e.*, training a model more often with uncommon attention maps. Histograms show that more uncommon attention maps were selected for training the model, *e.g.*, seeing pedestrians or traffic lights is weighted more than just seeing the vanishing point of roads. (B) Normalized sampling weights as a function of KL divergence values. A normalized sampling weight value of 1 indicates that the video frame is sampled once on average during a single epoch

three convolutional layers with 1×1 kernels and a dropout layer after each convolutional layer. It further processes the visual features from the previous extraction module and reduces the dimensionality of the visual features from 256 to 8. In our experiments, we observed that without the dropout layers, the model easily got stuck in a suboptimal solution which simply predicted a central bias map, i.e. an attention map concentrated in a small area around the center of the frame.

The temporal processor is a convolutional LSTM network with a kernel size of 3×3 followed by a Gaussian smooth layer (σ set to 1.5) and a softmax layer. It receives the visual features of successive video frames in sequence from the visual feature processing module and predicts an attention map for every new time step. Dropout is used for both the linear transformation of the inputs and the linear transformation of the recurrent states. We had also experimented with using an LSTM network for this module and observed that the model tended to incorrectly attend to only the central region of the video frames. The final output of this model is a probability distribution over 64×36 grids predicting how likely each region of the video frame is to be looked at by human drivers. Cross-entropy is chosen as the loss function to match the predicted probability distribution to the ground-truth.

Human Weighted Sampling (HWS)

Human driver attention datasets, as well as many other driving related datasets, share a common bias: the vast majority of the datasets consist of simple driving situations such as lane-following or car-following. The remaining small proportion of driving situations, such as pedestrians darting out, traffic lights changing, etc., are usually more crucial, in the sense that making errors in these moments would lead to greater cost. Therefore, ignoring this bias and simply using mean prediction error to train and test models can be misleading. In order to tackle this problem, we developed a new method that uses human gaze data to determine the importance of different frames of a driving dataset and samples the frames with higher importance more frequently during training.

In simple driving situations human drivers only need to look at the center of the road or the car in front, which can be shown by averaging the attention maps of all the frames of one driving video. When the attention map of one frame deviates greatly from the average default attention map, it is usually an important driving situation where the driver has to make eye movements to important visual cues. Therefore, the more an attention map varies from the average attention map of the video, the more important the corresponding training frame is. We used the KL divergence to measure the difference between the attention map of a particular frame and the average attention map of the video. The KL divergence determined the sampling weight of this video frame during training.

The histogram of the KL divergence of all the training video frames of our dataset is shown in Fig. 2.4. As we expected, the histogram was strongly skewed to the left side. Our goal was to boost up the proportion of the frames of high KL divergence values by weighted sampling. The sampling weight was determined as a function of KL divergence (D_{KL}) illustrated in Fig. 2.4B. The middle part of this function ($D_{KL} \in [1,3]$) was set to be proportional to the inverse of the histogram so that after weighted sampling the histogram of KL divergence would become flat on this range. The left part of the function ($D_{KL} < 1$) was set to a low constant value so that those frames would be sampled occasionally but not completely excluded. The right part of the function was set to a saturated constant value instead of monotonically increasing values in order to avoid over-fitting the model to this small proportion of data. Besides, the attention maps collected in the beginning and the end of each video clip can deviates from the average default attention map merely because the participants were distracted by the breaks between video clips. We therefore restricted the sampling weights of the first second and the last 0.5 seconds of each video to be less or equal to once per epoch. The histogram of KL divergence after weighted sampling is shown in Fig. 2.4A. In our experiment, we needed to sample the training frames in continuous sequences of 6 frames. For a particular sequence, its sampling weight was equal to the sum of the sampling weights of its member frames. These sequences were sampled at probabilities proportional to the sequence sampling weights.

Results and Discussion

Here, we first provide our training and evaluation details, then we summarize the quantitative and qualitative performance comparison with existing gaze prediction models and variants of our model. To test how natural and reasonable our model prediction look to humans, we conduct a human evaluation study and summarize the results. We further test whether our model trained on in-lab driver attention data can also predict driver attention maps collected in-car.

Training and Evaluation Details

We made two variants of our model. One was trained with a regular regime, *i.e.*, equal sampling during training, and the other was trained with Human Weighted Sampling (HWS). Except for the sampling method during training, our default model and HWS model shared the same following

training settings. We used 926 videos from our BDD-A dataset as the training set and 306 videos as the testing set. We downsampled the videos to 1024×576 pixels and 3Hz. After this preprocessing, we had about 30k frames in our training set and 10k frames in our testing set. We used cross-entropy between predicted attention maps and human attention maps as the training loss, along with Adam optimizer (learning rate $= 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$). Each training batch contained 10 sequences and each sequence had 6 frames. The training was done for 10,000 iterations. The two models showed stabilized testing errors by iteration 10,000.

To our knowledge, (Palazzi et al., 2017) and (Tawari & Kang, 2017) are the two deep neural models that use dash camera videos alone to predict human driver's gaze. They demonstrated similar results and were shown to surpass other deep learning models or traditional models that predict human gaze in non-driving-specific contexts. We chose to replicate (Palazzi et al., 2017) to compare with our work because their prediction code is public. The model designed by (Palazzi et al., 2017) was trained on the DR(eye)VE dataset (Alletto et al., 2016). We will refer to (Palazzi et al., 2017)'s model as DR(eye)VE model in the following. The training code of (Palazzi et al., 2017) is not available. We implemented code to fine-tune their model on our dataset, but the fine-tuning did not converge to any reasonable solution, potentially due to some training parameter choices that were not reported. We then tested their pre-trained model directly on our testing dataset without any training on our training dataset. Since the goal of the comparison was to test the effectiveness of the combination of model structure, training data and training paradigm as a whole, we think it is reasonable to test how well DR(eye)VE model performs on our dataset without further training. For further comparison, we fine-tuned a publicly available state-of-the-art image gaze prediction model, SALICON (Xun Huang et al., 2015) on our dataset. We used the open source implementation (Thomas, 2016). We also tested our models against a baseline model that always predicts the averaged human attention map of training videos.

Kullback-Leibler divergence (KL divergence, D_{KL}), Pearson's Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS) and Area under ROC Curve (AUC) are four commonly used metrics for attention map prediction (Palazzi et al., 2017; Tawari & Kang, 2017; Bylinskii, Judd, Oliva, Torralba, & Durand, 2018). We calculated the mean prediction errors in these four metrics on the testing set to compare the different models. In order to test how well the models perform at important moments where drivers need to watch out, we further calculated the mean prediction errors on the subset of testing frames where the attention maps deviate significantly from the average attention maps of the corresponding videos (defined as KL divergence greater than 2.0). We will refer to these frames as non-trivial frames. Our models output predicted attention maps in the size of 64×36 pixels, but the DR(eye)VE model and the SALICON outputs in bigger sizes. For a fair comparison, we scaled the DR(eye)VE model and the SALICON model's predicted attention maps into 64×36 pixels before calculating the prediction errors.

Another important evaluation criterion of driver attention models is how successfully they can attend to the objects that demand human driver's attention, e.g. the cars in front, the pedestrians that may enter the roadway, etc. Therefore, we applied the same attended object analysis described in the Berkeley DeepDrive Attention Dataset section. We used YOLO (Redmon & Farhadi, 2017) to detect the objects in the videos of our dataset. We selected object categories that are important for driving and that have enough instances in both our dataset and the DR(eye)VE dataset for

Table 2.3: Performance comparison of human attention prediction. Mean and 95% bootstrapped confidence interval are reported

	Entire testing set				Testing subset where D_{KL}(GT, Mean) > 2			
	KL divergence		Correlation coefficient		KL divergence		Correlation coefficient	
	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
Baseline	1.50	(1.45, 1.54)	0.46	(0.44, 0.48)	1.87	(1.80, 1.94)	0.36	(0.34, 0.37)
SALICON	1.41	(1.39, 1.44)	0.53	(0.51, 0.54)	1.76	(1.72, 1.80)	0.39	(0.37, 0.41)
DR(eye)VE	1.95	(1.87, 2.04)	0.50	(0.48, 0.52)	2.63	(2.51, 2.77)	0.35	(0.33, 0.37)
Ours (default)	1.24	(1.21, 1.28)	0.58	(0.56, 0.59)	1.71	(1.65, 1.79)	0.41	(0.40, 0.43)
Ours (HWS)	**1.24**	**(1.21, 1.27)**	**0.59**	**(0.57, 0.60)**	**1.67**	**(1.61, 1.73)**	**0.44**	**(0.42, 0.45)**

comparison (car, pedestrian and cyclist). We calculated the proportions of all the detected instances of those categories that were actually attended to by humans versus the models. The technical criterion of determining attended objects was the same as described in the Berkeley DeepDrive Attention Dataset section.

Evaluating Attention Predictor

Quantitative Analysis of Attention Prediction

The mean prediction errors of different models are summarized in Table 2.3. Both of our models significantly outperformed the DR(eye)VE model, the SALICON model and the baseline model in all metrics on both the entire testing set and the subset of non-trivial frames. Our model trained with HWS was essentially trained on a dataset whose distribution was altered from the distribution of the testing set. However, our HWS model showed better results than our default model even when being tested on the whole testing set. When being tested on the subset of non-trivial frames, our HWS model outperformed our default model even more significantly. These results suggest that HWS has the power to overcoming the dataset bias and better leveraging the knowledge hidden in crucial driving moments.

The results of the attended object analysis are summarized in Fig. 2.5A. Cars turned out to be easy to identify for all models. This is consistent with the fact that a central bias of human attention is easy to learn and cars are very likely to appear in the center of the road. However, for pedestrians and cyclists, the DR(eye)VE model, SALICON model and baseline model all missed a large proportion of them compared with human attention ground-truth. Both of our models performed significantly better than all the other competing models in the categories of pedestrians and cyclists, and our HWS model matched the human attention performances the best.

Importantly, our HWS model did not simply select objects according to their categories like an object detection algorithm. Considering the category that has the highest safety priority, pedestrian, our models selectively attended to the pedestrians that were also attended to by humans.

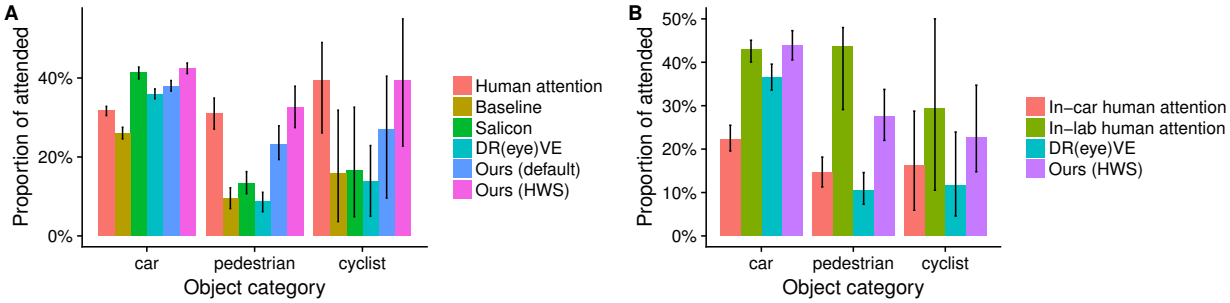

Figure 2.5: Analysis of attended objects for human attention and different models tested on our dataset (A) and the DR(eye)VE dataset (B). Error bars show 95% bootstrapped confidence intervals

Let us refer to the pedestrians that were actually attended to by humans as the important pedestrians and the rest of them as non-important pedestrians. Among all the pedestrians detected by the object detection algorithm, the proportion of important pedestrians was 33%. If our HWS model were simply detecting pedestrians at a certain level and could not distinguish between important pedestrians and non-important pedestrians, the proportion of important pedestrians among the pedestrians attended to by our model should also be 33%. However, the actual proportion of important pedestrians that our HWS model attended to was 48% with a bootstrapped 95% confidence interval of $[42\%, 55\%]$. Thus, our HWS model predicts which of the pedestrians are the ones most relevant to human drivers.

Qualitative Analysis of Attention Prediction

Some concrete examples are shown in Figure 2.6. These examples demonstrates some important driving scenarios: pedestrian crossing, cyclist getting very close to the vehicle and turning at a busy crossing. It can be seen from these examples that the SALICON model and the DR(eye)VE model mostly only predicted to look at the center of the road and ignored the crucial pedestrians or cyclists. In the examples of row 1, 2 and 3, both our default model and HWS model successfully attended to the important pedestrian/cyclist, and did not give false alarm for other pedestrians who were not important for the driving decision. In the challenging example shown in row 4, the driver was making a right turn and needed to yield to the crossing pedestrian. Only our HWS model successfully overcame the central bias and attended to the pedestrian appearing in a quite peripheral area in the video frame.

Human Evaluation

To further test how natural and reasonable our HWS model's predicted attention maps look to humans, we conducted an online Turing Test. In each trial, a participant watched one driving video clip three times: the first time with no edit, and then two times in random order with the ground-truth human driver attention map and our HWS model's predicted attention map overlaid

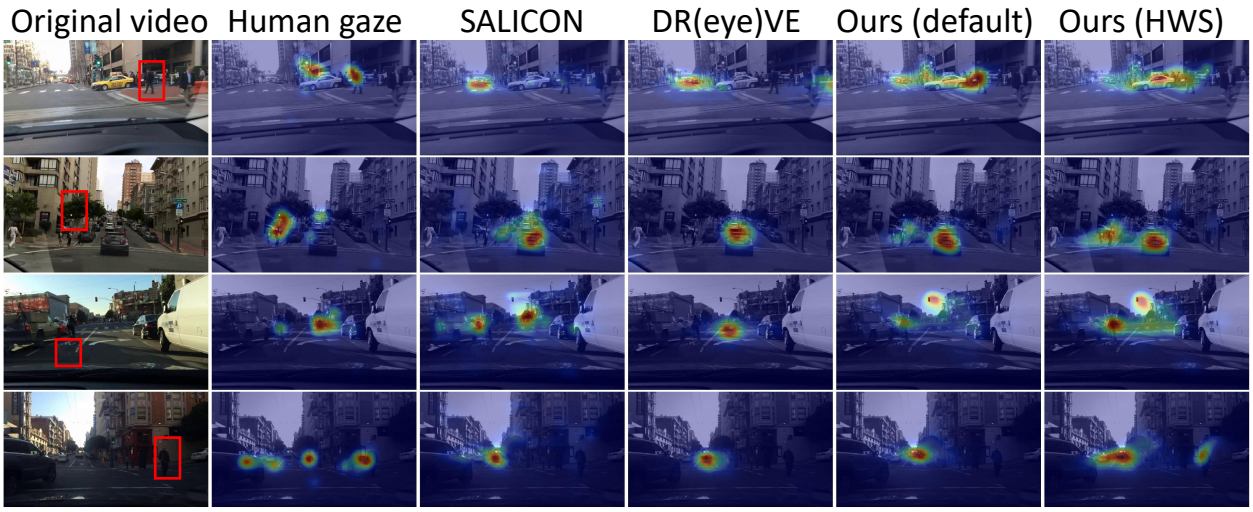

Figure 2.6: Examples of the videos in our dataset, ground-truth human attention maps and the prediction of different models. The red rectangles in the original video column highlight the pedestrians that pose a potential hazard. Row 1: the driver had the green light, but a pedestrian was about to cross the road while speaking on a phone without looking at the driver. Another pedestrian was present in the scene, but not relevant to the driving decision. Row 2: the driver had a yellow light and some pedestrians were about to enter the roadway. Another pedestrian was walking in the same direction as the car and therefore not relevant to the driving decision. Row 3: a cyclist was very close to the car. Row 4: the driver was making a right turn and needed to yield to the crossing pedestrian. Other pedestrians were also present in the scene but not relevant to the driving decision

on top, respectively. The participant was then asked to choose whether the first or the second attention map video was more similar to where a good driver would look.

Note that the experiment settings and instructions were the same as the online study described in the dataset section, except that one compares model prediction against the in-lab driver attention maps, and the other compares the in-car driver attention maps against the in-lab driver attention maps. Therefore, the result of this Turing Test can be compared with the result of the previous online study. In total, we collected 462 trials from 20 participants. If our HWS model's predicted attention maps were perfect and indistinguishable from the ground-truth human driver attention maps, the participants would had to make random choices, and therefore we would expect them to choose our model prediction in about 50% of the trials. If our HWS model's prediction was always wrong and unreasonable, we would expect a nearly zero chosen rate for our model prediction. Our results showed that in 41% of all trials the participants chose our HWS model's predicted attention maps as even better than the in-lab human attention maps (see Table 2.2). In the previous online study, the in-car attention maps of DR(eye)VE only achieved a chosen rate of 29%. This result suggests that our HWS model's predicted attention maps were even more similar to where a good driver should look than the human driver attention maps collected in-car (permutation test $p = 4 \times 10^{-5}$).

Table 2.4: Test results obtained on the DR(eye)VE dataset by the state-of-the-art model (DR(eye)VE) and our finetuned model. Mean and 95% bootstrapped confidence interval are reported

	KL divergence		Correlation coefficient	
	Mean	95% CI	Mean	95%CI
DR(eye)VE	1.76	(1.65, 1.87)	0.54	(0.51, 0.56)
Ours (finetuned)	1.72	(1.66, 1.81)	0.51	(0.48, 0.53)

Predicting In-Car Driver Attention Data:

To further demonstrate that our model has good generalizability and that our driver attention data collected in-lab is realistic, we conducted a challenging test: we trained our model using only our in-lab driver attention data, but tested it on the DR(eye)VE dataset, an in-car driver attention dataset. Note that the DR(eye)VE dataset covers freeway driving, which is not included in our dataset due to the small density of road user interactions on freeway. The high driving speed on freeway introduces strong motion blur which is not present in our dataset videos. Furthermore, drivers need to look further ahead in high speed situations, so the main focus of driver gaze pattern shifts up as the driving speed increases. In order to adapt our model to these changes, we selected 200 ten-second-long video clips from the training set of the DR(eye)VE dataset and collected in-lab driver attention maps for those video clips (already described in the Berkeley DeepDrive Attention Dataset section). We fine-tuned our HWS model with these video clips (30 minutes in total only) and the corresponding in-lab driver attention maps, and then tested the model on the testing set of the DR(eye)VE dataset (with in-car attention maps). The mean testing errors were calculated in D_{KL} and CC because the calculation of NSS and AUC requires the original fixation pixels instead of smoothed gaze maps and the original fixation pixels of the DR(EYE)VE dataset were not released. Our fine-tuned model showed a better mean value in KL Divergence and a worse mean value in CC than the DR(eye)VE model (see Table 2.4). But the 95% bootstrapped confidence intervals for the two models in both metrics overlapped with each other. So overall we concluded that our fine-tuned model matched the performance of the DR(eye)VE model. Note that the DR(eye)VE model was trained using the DR(eye)VE dataset and represents the state-of-the-art performance on this dataset.

We also calculated proportions of attended objects of important categories for our fine-tuned model and the DR(eye)VE model (Fig. 2.5B). Our fine-tuned model showed significantly higher proportions of attended objects in the car, pedestrian and cyclist categories and was more similar to the in-lab driver attention than the DR(eye)VE model. Note that we have shown in the Berkeley DeepDrive Attention Dataset section that humans rated the in-lab attention maps as more similar to where a good driver should look from a third-person perspective than the in-car attention maps.

Conclusions

In this paper, we introduce a new in-lab driver attention data collection protocol that overcomes drawbacks of in-car collection protocol. We contribute a human driver attention dataset which is to-date the richest and will be made public. We propose Human Weighted Sampling which can overcome common driving dataset bias and improve model performance in both the entire dataset and the subset of crucial moments. With our dataset and sampling method we contribute a novel human driver attention prediction model that can predict both in-lab and in-car driver attention data. The model demonstrates sophisticated behaviors and show prediction results that are nearly indistinguishable from ground-truth to humans.

Chapter 3

Periphery-Fovea Multi-Resolution Driving Model guided by Human Attention

Vision-based deep autonomous driving models have shown promising results recently (Kim & Canny, 2017; Kim, Rohrbach, Darrell, Canny, & Akata, 2018; Bojarski et al., 2016; H. Xu et al., 2017a). However, their performance is still far behind humans. An important aspect of human vision that distinguishes it from existing autonomous driving models is its multi-resolution property, with distinct foveal and peripheral structures that carry high-resolution and low-resolution information, respectively. The human fovea covers approximately two degrees of the central visual field; the rest of our visual field, *i.e.*, , the periphery, is blurry. Eye movements, guided by visual attention, are therefore necessary to gather high resolution foveal information from different parts of the visual field. One advantage of this design is its efficiency: resources are saved for particularly salient or important regions in what are otherwise redundant visual scenes. Driving scenes seem to be highly redundant, as well, considering the large portions of uniform areas such as the sky, buildings, and roads. Inspired by the human vision, we propose a new periphery-fovea multi-resolution driving model and show that it achieves higher driving accuracy and better efficiency.

The first challenge in designing this model is to effectively combine the global low-resolution peripheral vision and the local high-resolution foveal vision that dynamically scans across the frame. We propose two ways to merge the two visions by either using a combined peripheral-foveal planner or two independent visual planners. We will compare their performances and discuss the differences.

The second challenge is how to dynamically guide foveal vision to the critical locations. The foveal location selection is a non-differentiable process. A potential solution is to use reinforcement learning, but it could take a great deal of data and training. We choose a different approach: guiding the foveal vision to where human drivers would gaze. Recently proposed large driver gaze datasets (Xia et al., 2018; Alletto et al., 2016) and driver gaze prediction models (Xia et al., 2018; Palazzi, Abati, Calderara, Solera, & Cucchiara, 2018; Palazzi et al., 2017) allow us to predict human gaze for our videos. However, it has not been tested whether predicted human gaze or even ground-truth human gaze can benefit autonomous driving models. Note that in order to be highly

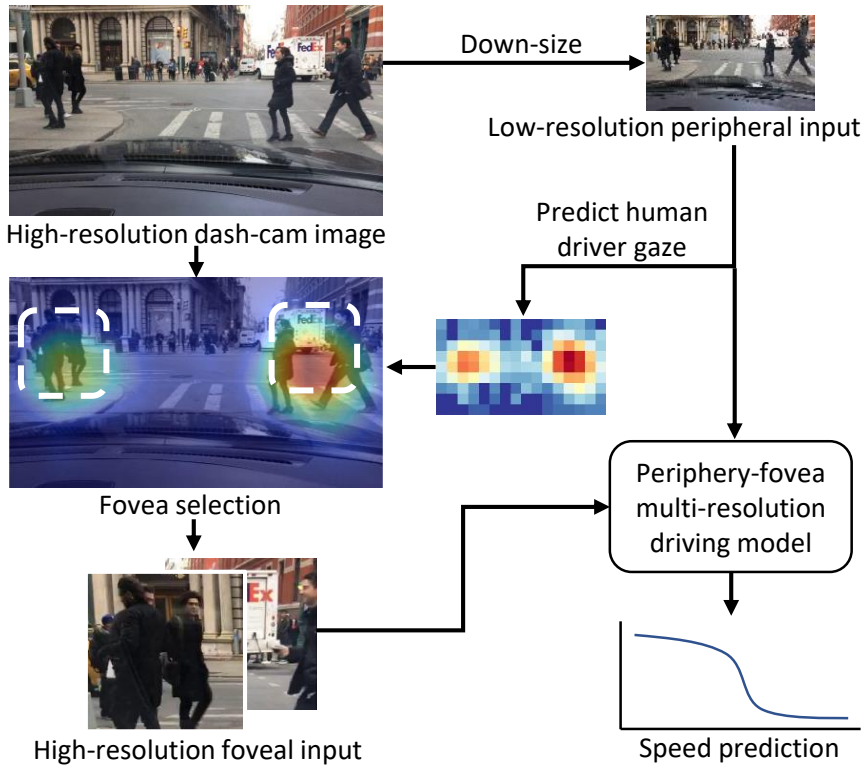

Figure 3.1: Our model uses the low-resolution full video frame as the peripheral visual input to predict human driver gaze and gets high-resolution image patches from the predicted gaze locations. It then combines the peripheral input and foveal input to predict the vehicle speed at high accuracy and high efficiency.

efficient, the human gaze can only be predicted using low-resolution input images, which makes the question even more complex.

A unique property of human gaze is that it reveals the relative urgency of locations and objects of potential interest. Different moments during driving and different road agents are not equally urgent. Human drivers look at the most critical regions when emergencies arise. Incorporating human gaze into a driving model may not only increase its average performance but also bring even higher performance gain at critical moments. We use a driving video dataset that has human-annotated explanations about the driver's actions. We demonstrate that our driving model guided by human gaze shows even higher performance gain in the cases where reactions to pedestrians are necessary than in other presumably less critical cases.

Related work

End-to-End Learning for Self-driving Vehicles. Recent successes (Bojarski et al., 2016; H. Xu et al., 2017a) suggest that a driving policy can be successfully learned by neural networks with the supervision of observation (*i.e.*, raw images)-action (*i.e.*, steering) pairs collected from human demonstration. Bojarski *et al.* (Bojarski et al., 2016) trained a deep neural network to map a dashcam image to steering controls, while Xu *et al.* (H. Xu et al., 2017a) utilized a dilated deep neural network to predict a vehicle's discretized future motions. Hecker *et al.* (Hecker, Dai, & Van Gool, 2018) explored an end-to-end driving model that consists of a surround-view multi-camera system, a route planner, and a CAN bus reader. Explainability of deep neural networks has been increasingly explored. Kim *et al.* (Kim & Canny, 2017; Kim et al., 2018) explored an interpretable end-to-end driving model that explains the rationale behind the vehicle controller by visualizing attention heat maps and generating textual explanation. Recently, Wang *et al.* (Wang, Devin, Cai, Yu, & Darrell, 2019) introduced an instance-level attention model that finds objects (*i.e.*, , cars, and pedestrians) that the network needs to pay attention to.

Incorporating human visual attention. Attention mechanisms have shown promising results in various computer vision tasks, *e.g.*, , image caption generation (K. Xu et al., 2015), visual question answering (VQA) (Zhu, Groth, Bernstein, & Fei-Fei, 2016), and image generation (Gregor, Danihelka, Graves, Rezende, & Wierstra, 2015). Most of these models do not supervise the generated attention by human attention. Recently, Das *et al.* (Das, Agrawal, Zitnick, Parikh, & Batra, 2017) has shown that explicitly supervising the attention of VQA models by human attention improves the models' VQA performance. Zhang *et al.* (R. Zhang et al., 2018) has trained a network that predicts human attention for Altari games and shown that incorporating the predicted human attention into the policy network significantly improves the action prediction accuracy. However, incorporating human visual attention in driving tasks has not yet been explored. Besides, the previously mentioned attention models use high-resolution images to generate attention. Predicting attention using low-resolution input and combining global low-resolution input and attended local high-resolution input has not been explored.

Predicting driver attention. Recently, deep driver attention prediction models (Xia et al., 2018; Palazzi et al., 2018; Palazzi et al., 2017) have been proposed. The input of these models is video recorded by cameras mounted on the car. The output is an attention map indicating the driver's gaze probability distribution over the camera frame. These models are trained using large-scale driver attention datasets (Xia et al., 2018; Alletto et al., 2016) collected with eye trackers, and they use high-resolution input images (576×1024 or higher) to achieve optimal accuracy. How reliable the prediction would be using low-resolution input images have not been explored.

Periphery-Foveal Multi-Resolution Model

Here, we propose a novel driving model that mimics the key aspect of the human vision system: the peripheral and the foveal systems. Our model mainly uses the peripheral vision to predict a

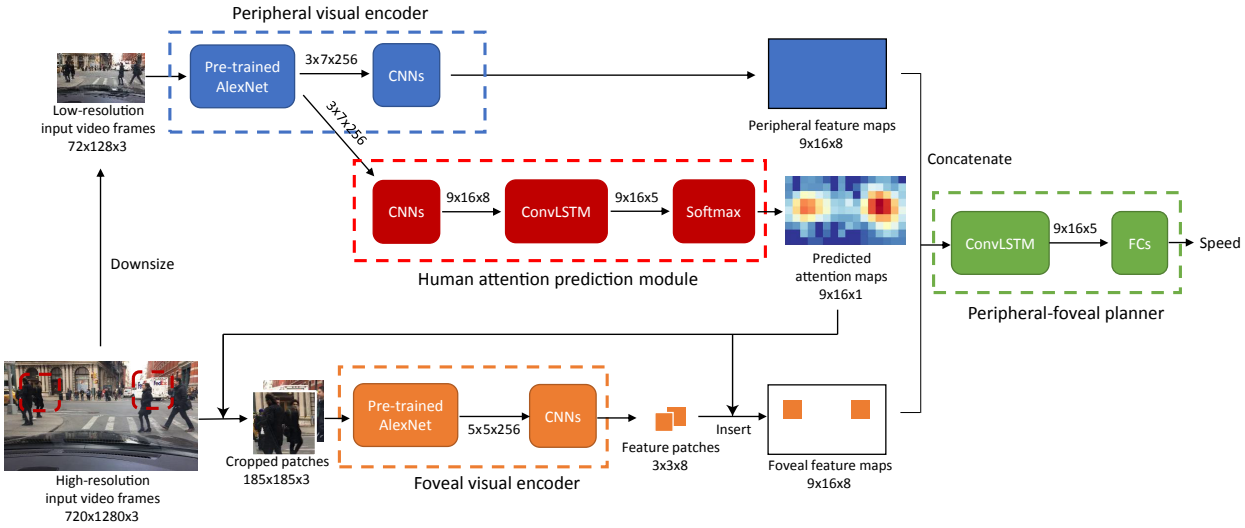

Figure 3.2: Our model consists of four parts: (1) the peripheral visual encoder, which extracts high-level convolutional visual features (CNN here); (2) the human attention prediction module, which learns the behavior of human attention as a supervised learner over image-gaze pairs collected from humans; (3) the foveal visual encoder, which selects fovea locations, crops the high-resolution fovea image patches and encodes them into visual features; (4) the peripheral-foveal planner, which combines the peripheral and foveal visual features and predicts a low-level control command, *i.e.*, , a vehicle's speed.

control command (*i.e.*, , speed) in an end-to-end manner, but we add the foveal vision to improve the model's perceptual primitives. While the peripheral vision sees the whole but blurry image, the foveal vision fixates on parts of the images with a higher resolution. To this end, our model needs three main capabilities: (1) the ability to extract perceptual primitives to manipulate the vehicle's behavior, (2) the ability to find out image regions where the model needs to attend with a high resolution (*i.e.*, , pedestrians, traffic lights, construction cones, etc), (3) the ability to augment the peripheral vision system with the foveal vision.

As we summarized in Figure 3.2, our model consists of four parts: (1) the *peripheral visual encoder*, which extracts high-level convolutional visual features (CNN here); (2) the *human attention prediction module*, which learns the behavior of human attention as a supervised learner over image-gaze pairs collected from humans; (3) the *foveal visual encoder*, which selects fovea locations, crops the high-resolution fovea image patches and extracts visual features from the high-resolution image patches; (4) the *peripheral-foveal planner*, which combines the peripheral and foveal visual features and predicts a low-level control command, *i.e.*, a vehicle's speed.

Peripheral Visual Encoder

We sample the video frames at 10 Hz. The original frame images have a resolution of 720×1280 pixels. We downsample them to 72×128 pixels as the input for the peripheral vision input of our model. The raw pixel values are subtracted by [123.68, 116.79, 103.939] as (Krizhevsky et al., 2012).

The low-resolution frame images are first passed to the peripheral feature encoder. This feature encoder consists of an ImageNet pre-trained AlexNet and three additional convolutional layers. The weights of the pre-trained AlexNet are fixed and not further trained during the training of our driving model. Each of the additional convolutional layers is followed by Batch Normalization and Dropout. The output feature maps of this feature encoder have a size of 3×7 pixels and 8 channels. These feature maps are then upsampled to 9×16 pixels for the next steps.

Human Attention Prediction Module

The low-resolution frame images are also passed to a human attention prediction module to determine where human drivers would gaze. We used the model described in (Xia et al., 2018) as our human attention prediction module. This model consists of a fixed ImageNet pre-trained AlexNet, three additional convolutional layers, and a Convolutional Long Short-Term Memory (ConvLSTM) module. Since both the peripheral feature encoder and the human attention prediction module start with passing the low-resolution through the same fixed AlexNet, this passway is shared by both modules. The human attention prediction module is separately trained using a human driver attention dataset and is fixed during the training of the driving model. The predicted human attention maps have a resolution of 9×16 pixels.

Foveal Visual Encoder

The foveal visual encoder chooses two independent fovea locations for each input frame. In the following experiments, the fovea locations can be chosen in four different ways: random selection over the frame, always selected from the frame center, a top-k method and a sampling method. The top-k method selects the two pixels that have the highest attention intensities in each predicted 9×16-pixel human attention map. The sampling method samples two fovea locations following the predicted attention probability distribution modulated by a temperature factor described by the following formula:

$$p_i = \frac{\exp\left(\log q_i / T\right)}{\sum_j \exp\left(\log q_j / T\right)} \tag{3.1}$$

where p_i is the probability of the i-th pixel being selected as the fovea location, q_i is the predicted human attention probability at the i-th pixel, and T is the temperature factor. A temperature factor of 1 means sampling faithfully following the predicted human attention distribution. A higher temperature factor means sampling more uniformly. A lower temperature factor means sampling more from the pixel that has the highest human attention intensity.

An image patch of 240×240 pixels centered at each selected fovea location is cropped out from the 720×1280-pixel high-resolution frame image. The images patches are then downsized to 185×185 pixels to fit the receptive fields and strides of the following encoder network. The raw pixel values are subtracted by [123.68, 116.79, 103.939] as (Krizhevsky et al., 2012) before being passed to the encoder network. The foveal visual encoder has the same structure as the peripheral visual encoder except for the kernel sizes and strides of the additional convolutional layers.

Peripheral-Foveal Planner

The peripheral-foveal planner further processes the peripheral and foveal features to predict speed for the future. It first creates a foveal feature map that has the same size as the peripheral feature map (9×16 pixels, eight semantic channels). The foveal feature map is initialized with zeros. Each foveal image patch is encoded into a $3 \times 3 \times 8$ feature patch by the foveal feature encoder. These foveal feature patches ($\mathbf{y}_{i,j}$) are inserted into the foveal feature map ($\mathbf{x}^f_{i,j}$) at locations corresponding to the foveal locations:

$$\mathbf{x}^f_{i+h,j+w} = \mathbf{y}_{i,j} \tag{3.2}$$

where h and w are the height and width coordinates of the top-left corner of the fovea patch.

In the cases where the feature patches of two foveae overlap, the maximum of each pair of overlapping feature values is kept. Then the peripheral feature maps ($\mathbf{x}^p_{i,j}$) and foveal feature maps ($\mathbf{x}^f_{i,j}$) are concatenated along the semantic dimension to form the combined feature maps ($\mathbf{x}^c_{i,j}$).

$$\mathbf{x}^c_{i,j} = \begin{pmatrix} \mathbf{x}^p_{i,j} \\ \mathbf{x}^f_{i,j} \end{pmatrix} \tag{3.3}$$

The combined feature maps are then processed by a ConvLSTM layer and four fully-connected layers to predict a continuous value for the vehicle speed.

Experiments

In this section, we first present the datasets we used and our training and evaluation details. Then, we make quantitative and qualitative analyses of our proposed periphery-fovea multi-resolution driving model.

Datasets

We used the Berkeley DeepDrive eXplanation (BDD-X) dataset (Kim et al., 2018) to train and evaluate the driving models. This dataset contains human-demonstrated dashboard videos of urban driving scenes in various weather and lighting conditions. The dataset also provides a set of time-stamped sensor measurements, *e.g.*, , vehicle's velocity and course, and time-stamped human annotations for vehicle action descriptions and justifications. The training set contains 5,588 videos and the validation and testing sets contain 698 videos. Most videos are 40 seconds long.

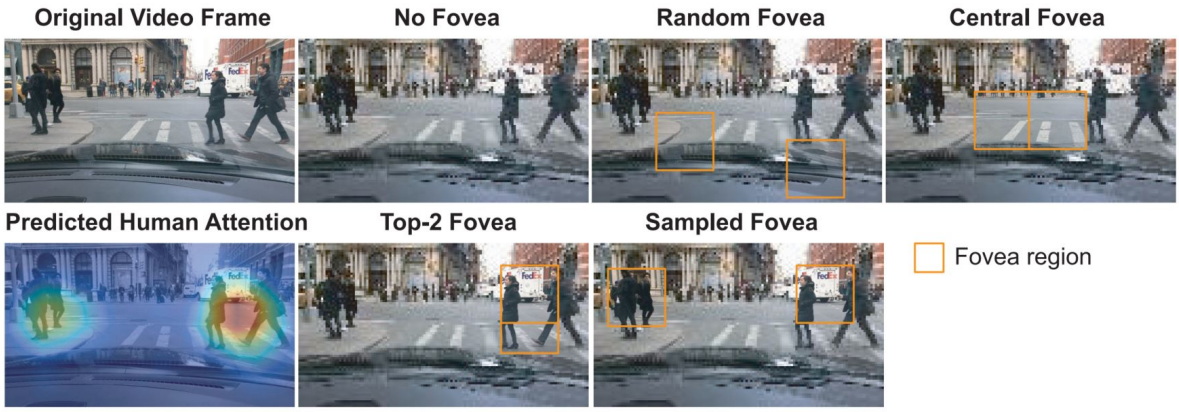

Figure 3.3: Examples of different approaches of foveal region selection. We present the original input video frame and the predicted human attention heat map at the left column. Our baseline model only uses peripheral vision (without fovea). We studied four different types of foveal vision selection: random, central, top-2, and sampling. Top-2 and sampled foveae are chosen according to the predicted human attention. For better visualization, we present orange boxes to indicate the foveal regions.

We used the Berkeley DeepDrive Attention (BDD-A) dataset (Xia et al., 2018) to train the human attention prediction module. The BDD-A dataset contains driving videos collected in the same way as the BDD-X dataset. (But the two datasets do not share the same videos.) The BDD-A dataset also provides human attention map annotations. The human attention maps were collected by averaging multiple drivers' eye movements while they were watching the videos and performing a driver instructor task (Xia et al., 2018). The attention maps highlight where human drivers need to gaze when making driving decisions in the particular situations. The BDD-A dataset contains 926, 200 and 303 videos in the training, validation and testing sets, respectively. Each video is approximately 10-second-long.

Training and Evaluation Details

The AlexNet modules in the driving models were pre-trained on ImageNet and frozen afterwards. The human attention prediction module was trained following (Xia et al., 2018) except that the input image resolution was 72×128 pixels. Other parts of the driving models were trained end-to-end from scratch. We used the Adam optimization algorithm (Kingma & Ba, 2014), dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) at a drop rate of 0.2, and the Xavier initialization (Glorot & Bengio, 2010). The training of our model took approximately one day on one NVIDIA GeForce GTX 1080 GPU. Our implementation is based on Tensorflow and our code will be publicly available upon publication. The models were set to predict the vehicle speed one second in the future. We used three metrics, *i.e.*, , the mean absolute error (MAE), the root-mean-square error (RMSE), and the correlation coefficient (Corr), to compare the prediction against the

Table 3.1: We compared the vehicle control (*i.e.*, speed) prediction performance of four different types of vision systems. We evaluated their performance in terms of the mean absolute error (MAE), the root-mean-square error (RMSE), and the correlation coefficient (Corr).

Model	Speed (km/h)		
	MAE	RMSE	Corr
Peripheral vision only (no fovea, baseline)	9.6	14.4	.594
w/ Random fovea	11.2	15.4	.520
w/ Central fovea	9.4	13.9	.592
w/ Human-guided fovea (ours)	**9.1**	**13.4**	**.596**

ground-truth speed signals to evaluate the performances of the driving models. At inference time, the longest single video duration that our GPU memory could process was 30 seconds. Therefore, during training, unless otherwise stated, the original testing videos that were longer than 30 seconds were divided into 30-second-long segments and the remaining segments.

Effect of the foveal vision guided by human attention

To test the effect of the foveal vision guided by human attention, we compared our peripheral-foveal multi-resolution driving model against three baseline models (Figure 3.3). The first baseline model (no fovea) uses only low-resolution full video frames as input and has only the peripheral branch of the driving model we introduced. The second baseline model (random fovea) select fovea locations randomly over the video frame. The third baseline model (central fovea) always assigns its two foveae to the central 240×480 region of the frame. The central-fovea model is a strong baseline because the central regions mostly cover the area the vehicle is driving into and human drivers mostly localize their attention around the center of the road. We compared these baseline models with our peripheral-foveal multi-resolution driving model guided by human attention (human-guided fovea). The fovea locations were selected using the top-2 method. The mean testing errors of these models are summarized in Table 3.1. Our driving model outperformed all of the baseline models. This result suggests that the foveal vision guided by predicted human attention can effectively improve the model's accuracy. Note that the random-fovea model performed worse than the no-fovea model. This suggests that adding high-resolution foveal input would not necessarily improve the model. If fovea locations are not selected in a proper way, it may add distracting information to the driving model.

Sampling according to multi-focus human attention

Human attention can be multi-focus (Cavanagh & Alvarez, 2005), especially during driving when the driver needs to react to multiple road agents or objects. A concern about using the top-2 method

Table 3.2: Mean testing errors of our driving model using different fovea selection methods.

Fovea selection	Temperature	Likelihood	Overlap	MAE	RMSE	Corr
Top-2 fovea	-	0.48	92%	9.1	13.4	.596
Sampled fovea	0.5	0.46	55%	8.6	12.7	.622
Sampled fovea	1	0.37	32%	**8.5**	**12.4**	**.626**
Sampled fovea	2	0.18	11%	8.7	12.9	.621

Table 3.3: Mean testing errors of our driving models using either combined or dual peripheral-foveal planner.

Model	MAE	RMSE	Corr
Ours w/ Dual Peripheral-foveal Planner	9.4	13.2	.602
Ours w/ Combined Peripheral-foveal Planner	**8.5**	**12.4**	**.626**

to select fovea locations is that it may select adjacent locations around a single focus in one frame and also select locations from the same focus in the next frames. To address this concern, we brought a sampling method to select fovea locations (described in the Model section). It samples fovea locations according to the predicted human attention probability distribution and modulated by a temperature factor (Figure 3.3). We tested our driving model using both the top-2 method and the sampling method and experimented with three different temperature factor values for the sampling method. To quantify to how much extend the fovea selection followed the predicted human attention, we calculated the likelihood of the selected foveae. To quantify the redundancy in fovea location selection, we calculated the overlap ratio between the fovea patches of adjacent frames. The results are summarized in Table 3.2. The results showed the trend that a balance between high likelihood and low overlap would result in the optimal performance. In our experiments, sampling completely following the predicted human attention distribution (*i.e.*, , temperature factor $T = 1$) showed the best prediction accuracy.

Comparison between combined and dual peripheral-foveal planner

The previously presented design of our peripheral-foveal planner combines peripheral and foveal features to process with one ConvLSTM network. We call this design the combined peripheral-foveal planner design. In this design, the peripheral and foveal feature maps need to have the same resolution in order to be concatenated along the semantic dimension (9×16 in our case). This constraint determines that the feature patch corresponding to one foveal input image patch cannot be bigger than 3×3 pixels.

To break this constraint, we experimented with a different design, *i.e.*, , the dual peripheral-foveal planner structure. It bypasses the uni-resolution constraint by processing the peripheral and foveal features with separate ConvLSTM networks. It generates a feature patch of 14×14 pixels for each foveal input image patch. In stead of inserting the foveal feature patch into a bigger grid

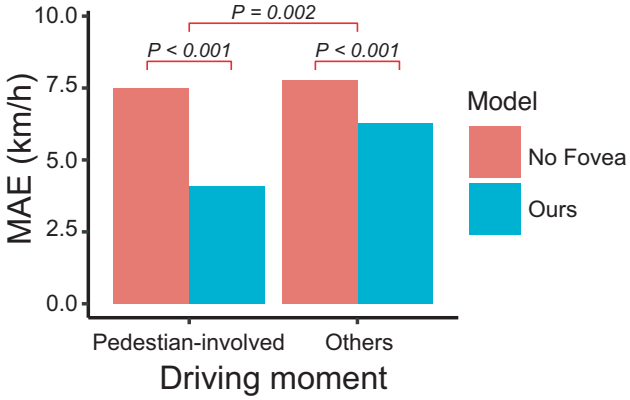

Figure 3.4: Testing errors of the no-fovea baseline model and our model at pedestrian-involved moments and other moments when the vehicle speed is under $10\,\mathrm{m/s}$ ($36\,\mathrm{km/h}$). Statistical significance levels given by permutation tests are noted in the graph.

that corresponded to the full video frame, it adds the positional encoding (Vaswani et al., 2017) of the fovea location into the fovea features to preserve the fovea location information.

We tested the dual planner and compared it against the combined planner. The dual planner did not show higher accuracy than the combined planner (Table 3.3). We think this is because the combined planner also have its own unique advantages. In the combined planner design, the fovea location is clearly indicated by the location of the features in the feature map. Besides, the foveal features and peripheral features that are calculated from the same frame region are aligned into one vector in the combined feature maps. So the kernel of the upcoming ConvLSTM network can process the peripheral and foveal features of the same region jointly.

Larger performance gain in pedestrian-involved critical situations

The textual annotations of the BDD-X dataset allowed us to identify the critical situations where the driver had to react to pedestrians. These pedestrian-involved situations were defined as the video segments where the justification annotations contained the word "pedestrian", "person" or "people". We tested whether our model showed a stronger performance gain in the pedestrian-involved situations than in the remaining situations which should be on average less critical.

We calculated the mean prediction errors of our model and the no-fovea model separately for the pedestrian-involved video segments and the remaining segments in the test set. Note that the prediction error correlates with the vehicle speed and the pedestrian-involved segments only covered a speed range up to $10\,\mathrm{m/s}$ ($36\,\mathrm{km/h}$). For a fair comparison, we excluded the frames in which the vehicle speed was higher than $10\,\mathrm{m/s}$ from this analysis. In order to determine the statistical significance levels, we ran permutation tests that could address the concern that the frames of a video are not independent.

The results are summarized in Figure 3.4. Our model showed significant performance gains in

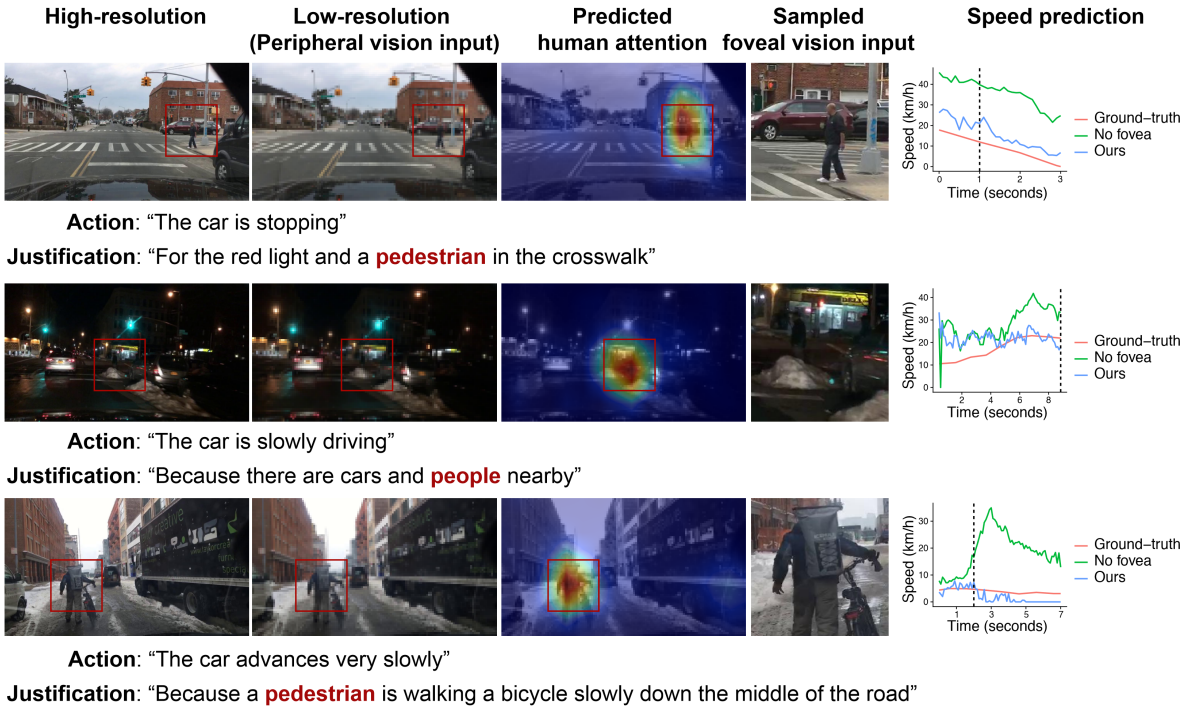

Figure 3.5: Examples showing how our model and the no-fovea model react in pedestrian-involved situations. From left to right: original high-resolution frame images, low-resolution frame images used as peripheral vision input, predicted human attention maps, selected high-resolution image patches as foveal vision input, and ground-truth and predicted speed curves. The vertical dashed lines in the speed curve graphs indicate the moments depicted by the frame images. The textual action and justification human annotations are displayed below the images of each example.

both the pedestrian-involved situations and the remaining situations (P value < 0.001). More importantly, the gain achieved in the pedestrian-involved situations was significantly bigger than the gain in the remaining situations (P value $= 0.002$). Some examples are demonstrated in Figure 3.5.

Multi-resolution vs. Uni-resolution

We further compared the performance of our periphery-fovea multi-resolution model with an uni-resolution periphery-only design, *i.e.*, , allocating all the resources to increase the resolution of the periphery vision without adding foveal vision. The number of floating-point operations (FLOPs) of our multi-resolution model for processing every video frame at inference is 3.4 billion. A medium-resolution periphery-only model that matches the same amount of FLOPs has a periphery input resolution size of 209×371 pixels. The structure of this model was the same as the periphery branch of our model except one change due to the enlarged input resolution. The periphery encoder of our model output feature maps of 3 pixels and then upsampled them to 9×16 pixels. The

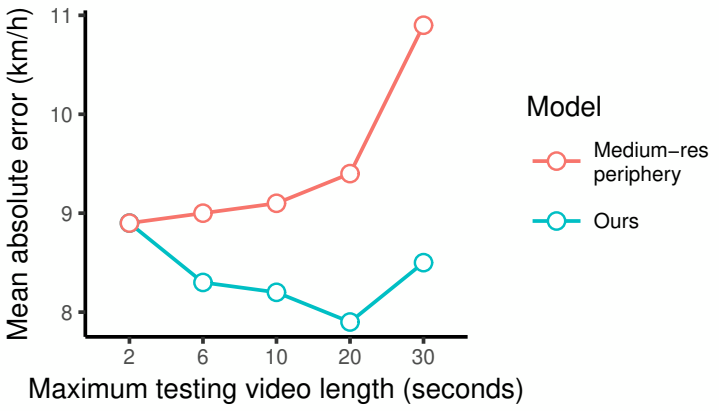

Figure 3.6: Testing errors of the medium-resolution periphery-only model and our model calculated using different lengths of testing videos. The two models have the same amount of FLOPs at inference time, but our model consistently showed greater driving accuracy than the competing model.

periphery encoder of the medium-resolution model output feature maps of 12×22 pixels and then downsampled them to 9×16 pixels. We tested this medium-resolution periphery-only (medium-res periphery) model against our periphery-fovea multi-resolution model. For a thorough analysis, we did the comparison for multiple rounds. In each round we cut the test videos into segments no longer than a certain length and tested the models using those segments. We tried segment lengths from two seconds up to 30 seconds (the longest single segment that we could process with our GPU memory). The prediction errors of the two models measured in MAE are summarized in Figure 3.6. The prediction error of the medium-res periphery model kept increasing with increasing video length, while the prediction error of our model stayed more stable. Our model showed smaller prediction errors than the medium-res periphery model with all video lengths except with 2 seconds the two models showed the same error. Over all, the result suggested that the periphery-fovea multi-resolution design would achieve better driving accuracy than a uni-resolution periphery-only design given the same amount of computation.

Conclusion

We have proposed a new periphery-fovea multi-resolution driving model that combines global low-resolution visual input and local high-resolution visual input. We have shown that guiding the foveal vision module by predicted human gaze significantly improves driving accuracy with high efficiency. The performance gain is even more significant in pedestrian-involved critical situations than other average driving situations. Our approach has demonstrated a promising avenue to incorporate human attention into autonomous driving models to handle crucial situations and to enhance the interpretability of the model's decisions.

Chapter 4

Visual Crowding in Driving

We live in a constantly cluttered visual world: from letters in text to products on the shelves of supermarkets, and to the crowds of cars and pedestrians in busy crossings. The natural crowdedness of our visual input confronts us with a fundamental limitation on object recognition, which is known as visual crowding: objects that can be easily identified in isolation seem jumbled and indistinct in clutter (Dennis M. Levi, 2008; Denis G Pelli and Tillman, 2008; Whitney and Levi, 2011). Visual crowding operates over a wide part of our visual field, in particular in peripheral vision, and it is considered as the major bottleneck on recognizing objects in clutter (Dennis M. Levi, 2008; Manassi and Whitney, 2018; Denis G Pelli and Tillman, 2008; Strasburger, Rentschler, and Juttner, 2011; Whitney and Levi, 2011). Given its ubiquity and significance, the impact of crowding on object recognition has been studied for decades (Flom, Heath, and Takahashi, 1963; Westheimer and Hauske, 1975; D. M. Levi, Klein, and Hariharan, 2002; D. M. Levi, Hariharan, and Klein, 2002; D. G. Pelli, Palomares, and Majaj, 2004; Strasburger et al., 2011). However, the vast majority of studies in the field has been restricted to experiments in psychophysics laboratory. The stimuli used were almost exclusively simple, static, and artificial, such as oriented gratings, shapes, letters, symbols, and even faces. Only few experiments have studied the impact of crowding in static natural scenes or of natural textures (Wallis and Bex, 2012; Gong, Xuan, Smart, and Olzak, 2018). Furthermore, most crowding studies rely on participants' explicit responses made during psychophysical tasks, e.g., pressing buttons or clicking the mouse to explicitly report how they recognize the stimuli. These unnatural response methods, along with laboratory setting and artificial stimuli, raise concerns on to how much crowding can be generalized to daily life. Therefore, it is unclear how crowding operates when viewing natural dynamic scenes in real-life situations.

Driving is a frequent and potentially fatal real-life situation where crowding may play a critical role. Crowding may occur in driving because hazards, obstacles, and pedestrians frequently appear in clutter in the periphery (see Figure 4.1A for an example). However, on the other hand, many types of information, e.g., object configurations, facial expressions, textures, and even scene gist, are known to get through the bottleneck of crowding (for a review see Manassi and Whitney, 2018). Some of this information could guide behavior, at least in principle. Therefore, whether crowding limits behavioral performance in the context of driving is a crucial and open question.

A major challenge that one would face in the study of crowding in driving (and natural scenes

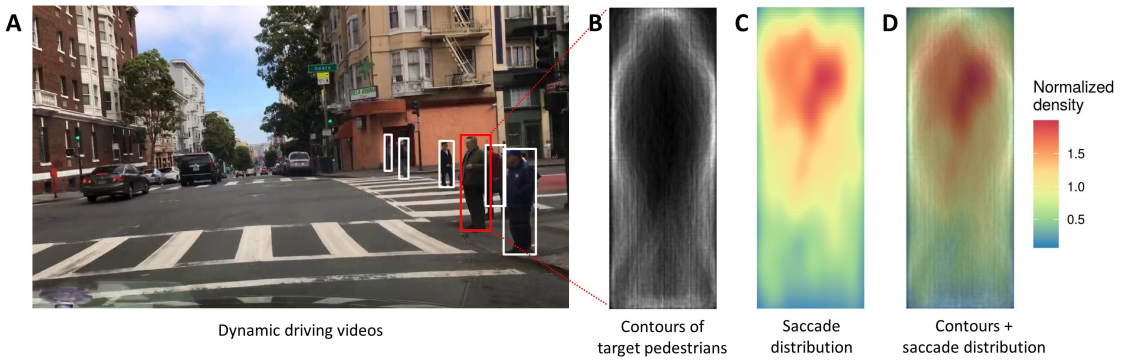

Figure 4.1: Participants watched crowd-sourced driving videos in a simulated driving environment while we recorded eye movements. Pedestrians were detected by a state-of-the-art object detection algorithm, Mask R-CNN (He, Gkioxari, Dollar, and Girshick, 2017). The saccades that landed on pedestrians were then identified. (A) One example frame of the crowd-sourced driving videos and the bounding boxes of the detected pedestrian given by Mask R-CNN. The red bounding box highlights the pedestrian that was targeted by one of the participants' saccades. The white bounding boxes show the other detected pedestrians. (B) Overlaid contours of the pedestrians on which participants' saccades landed, such as the one highlighted in red in Panel A. (C) Distribution of the landing points of the pedestrian-targeted saccades within the bounding box of a pedestrian. (D) The overlay of Panel A and B.

in general) is the availability and analysis of realistic stimuli. Most recently, however, in the field of computer vision and autonomous driving, diverse large-scale driving video datasets (Cordts et al., 2016; Maddern, Pascoe, Linegar, and Newman, 2017; Xinyu Huang et al., 2018; H. Xu, Gao, Yu, and Darrell, 2017b; Yu et al., 2018) have been created and powerful object detection algorithms using Deep Learning have been proposed (He, Gkioxari, Dollar, and Girshick, 2017; S. Liu, Qi, Qin, Shi, and Jia, 2018; Chen et al., 2019). We adopt driving videos from one of the new datasets and use a state-of-the-art object detection algorithm to analyze the videos at object-level for our experimental purposes.

In the present study, we first investigated the impact of the visual clutter in dynamic driving scenes on a fundamental kind of behavior in driving, i.e., eye movements (altered saccadic localization). Second, we further tested whether visual crowding occurs in the recognition of peripheral flanked objects (pedestrians), using static video frames and a conventional psychophysical paradigm. To foreshadow our results, we found that both saccadic localization and pedestrians recognition were impacted in manners that were consistent with the diagnostic criteria of crowding: Bouma's rule-of-thumb (Bouma, 1970), target-flanker similarity tuning (Kooi, Toet, Tripathy, and Levi, 1994; see Dennis M. Levi, 2008 for review), and radial-tangential anisotropy (Toet and Levi, 1992). Importantly, the altered saccadic localization was associated with the degree of crowding of the saccade targets. These results provide strong evidence that crowding strongly impacts both recognition and goal-directed actions in natural driving situations.

Experiment 1

In Experiment 1, we recorded eye movements while observers watched natural driving videos and we analyzed the saccades that landed on pedestrians (pedestrian-targeted saccades). We tested whether corrections in saccadic localization occurred in manners that were consistent with the diagnostic criteria of crowding (Bouma's rule of thumb, flanker similarity tuning, radial-tangential anisotropy, etc.). A positive result would suggest that crowding of pedestrians could potentially exist in driving and be associated with altered saccadic localization.

Methods

Participants

Eight naive participants participated in the experiment for course credits. All of the participants had driven for more than one year and had normal or corrected normal vision.

Stimuli and display setup

We used 519 videos from Berkeley DeepDrive Attention dataset (BDD-A, Xia et al., 2018). BDD-A contains crowd-sourced driving videos recorded by vehicle-mounted dashboard cameras in cities under various weather and lighting conditions (Xia et al., 2018). The videos are mostly 10-seconds long and contain diverse driving activities, e.g., lane following, turning, switching lanes, and braking.

Stimuli were displayed on a CRT monitor (display area size 34 cm $\times$ 23 cm). Display resolution was set to 1024 $\times$ 768 and the refresh rate to 60 Hz. Participants viewed the stimuli binocularly in a darkened experimental booth and head position was stabilized with a chinrest at a viewing distance of 57 cm. At this distance, 30 pixels subtended approximately 1° of visual angle.

Eye tracking

Eye movements were recorded at 1000 Hz monocularly with an EyeLink 1000 desktop mounted infrared eye tracker (SR Research Ltd., Mississauga, Ontario, Canada) used in conjunction with the Eyelink Toolbox scripts for Matlab. Participants were calibrated with a standard 9-point calibration procedure before completing each run (average error < 0.5°). The saccades were parsed out by the EyeLink online parser with the default high-sensitivity configuration (velocity threshold = 22°/s, acceleration threshold = $4000\,^\circ/\mathrm{s}^2$, and motion threshold = 0°).

Procedure

Participants performed a driver instructor task that was adopted from Xia et al., 2018. They watched the driving videos after they were informed that they were driving instructors sitting in the copilot seat. They were asked to press the space key whenever they felt it necessary to correct or warn the student driver of potential dangers. Their eye movements during the task were recorded.

It has been shown that the gaze maps collected by this method are considered as reasonable driver attention maps by independent human viewers (Xia et al., 2018) and can be used to improve autonomous driving models (see Chapter 3). Therefore, we think the driving instructor task can simulate an engaging driving environment for the participants while allowing them to make natural eye movements throughout the scene. The conclusions drawn from these eye movements can be presumably generalized to drivers' eye movements during actual driving.

Each participant watched 200 driving videos in a random order and performed the driving instructor task. Before each driving video, a yellow bullseye was displayed at the center of the screen on top of a uniform gray background. The participant was asked to gaze at the yellow bullseye and press the enter key to start the next video.

Results and discussion

Identification of saccades that landed on pedestrians

To identify the saccades that landed on pedestrians (pedestrian-targeted saccades), we extracted the video frame at the ending point of each saccade. We applied an object detection model to those extracted video frames using a state-of-the-art deep learning object detection algorithm, Mask R-CNN (He et al., 2017), and acquired bounding boxes around the detected objects. We identified the pedestrian-targeted saccades by looking for saccades with landing points within the bounding box of a detected pedestrian. The saccades with landing points beyond 15° away from the starting points were excluded as outliers (1.8% of total). In total, 2,067 pedestrian-targeted saccades were found and kept.

We extracted the contours of the target pedestrians (the ones on which saccades landed), scaled them to the median height-to-width ratio of all the target pedestrians (the median height-to-width ratio = 2.8), and overlaid them together. The overlaid contours showed the shape of a standing/walking pedestrian (Figure 4.1A). We also calculated the distribution of the landing points of the pedestrian-targeted saccades within the bounding box of the standardized pedestrian. The distribution showed that the participants directed their saccades mostly to the upper part of the pedestrian, presumably because the participants wanted to look at the pedestrians' faces (Figure 4.1B; Boucart et al., 2016). This result suggested that the pedestrian-targeted saccades were directed to specific locations of the pedestrians. If the localization of one saccade was inaccurate, i.e., the saccade was not going toward the desired location, a correction might be made in the landing stage of the saccade, which marks a corrected/altered saccade.

Identification of altered saccades

To identify the altered saccades, i.e. pedestrian-targeted saccades that contained correction in the landing stage, we analyzed the speed-time curves of the saccades. For most of the pedestrian-targeted saccades, the speed increased monotonically to the peak velocity and then monotonically decreased. One example is shown in Figure 4.2A. For some saccades, after reaching the peak values, speed decreased to a low value below the speed threshold for saccade detection but increased

again to above the threshold and then finally ended under the threshold (one example shown in Figure 4.2B). The intermediate low-speed stages typically occurred 13 ms prior to the ends of the saccades (Figure 4.2D) and were usually accompanied with a direction change in the saccade trajectory. We used the presence of the intermediate low-speed stage as a mark of saccade landing correction, i.e., altered saccadic localization. We used $30°$/s as a speed threshold (which is also equal to the conservative speed threshold for saccade parsing suggested by the Eyelink online parser) and defined the intermediate low-speed stages as the stages where at least three consecutive speed measurements (i.e., longer than 3 ms) were below the speed threshold prior to the landing stage, i.e, the last consecutive speed measurements that were below the speed threshold. The saccades that contained intermediate low-speed stages were defined as altered saccades. 396 of the 2,067 pedestrian-targeted saccades were defined as altered saccades (more examples shown in Figure 4.2C), and the rest were defined as direct saccades.

More altered saccades when the target pedestrians were flanked

We studied whether the proportion of altered saccades differed for flanked and unflanked pedestrians. Before that, we first tested whether the visual size of the target pedestrian influenced the proportion of altered saccades to see whether we can rule out this confounder from the following analysis. The correlation coefficient between the target visual size and the proportion of altered saccades was -0.001 and the p-value was 0.96. Therefore, we ruled out the target visual size as a confounder and did not include it in the following analyses.

We divided the target pedestrians (i.e., the pedestrians on which saccades landed) into flanked and unflanked targets according to whether there were other pedestrians (flanking pedestrians) in the $2.5°$ vicinity around them (the circular areas with a radius of $2.5°$ viewing angle centered at the centers of the target pedestrians). Notice that only pedestrians are considered as flankers in this analysis and we will consider other potential flanking objects in the next analysis. There were 1,123 flanked target pedestrians and 944 unflanked ones. For both flanked and unflanked targets, the data showed that the saccades became less accurate when the targets were more peripheral: the proportion of altered saccades (PAS) increased with increasing target eccentricity (the angular distance between the starting point of the saccade and the landing point of the saccade, Figure 4.2). If crowding occurred and led to more saccade inaccuracy, we would expect higher PAS for flanked targets than for unflanked targets given the same target eccentricity. According to Bouma's rule, crowding happens when the target-flanker spacing is below approximately one half of the target eccentricity (Bouma, 1970). Therefore, more specifically, we expected that flanked and unflanked targets would show similar PAS on the eccentricity range between $0°$ and $5°$; beyond $5°$, PAS for both flanked and unflanked targets would increase with increasing target eccentricity, but the increase for flanked targets should be significantly faster. To test our hypothesis, we fit the following logistic regression model between PAS and target eccentricity:

$$\text{Model 1}: \quad \log(\frac{p}{1-p}) = \alpha + \beta \cdot (eccen - 2.5°) \tag{4.1}$$

where p is the PAS and α and β are fitted parameters. α indicates the fitted PAS at an eccentricity

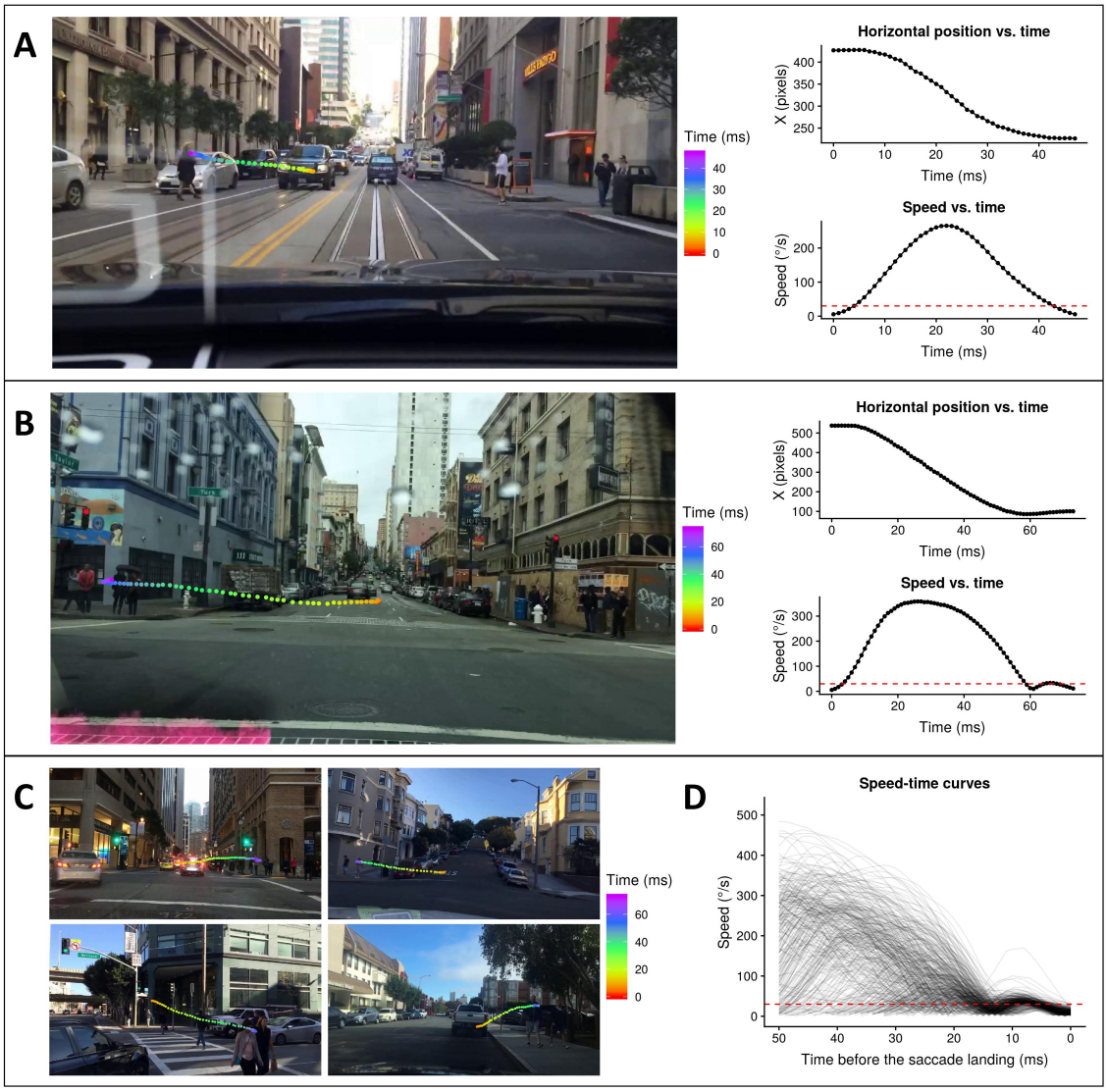

Figure 4.2: Direct and altered pedestrian-targeted saccades. The trajectories, horizontal pixel-time curves, and speed-time curves of one example direct saccade (A) and one example altered saccade (B). (C) The trajectories of various examples of altered pedestrian-targeted saccades. (D) The speed-time curves of all the altered pedestrian-targeted saccades over the last 50 ms prior to the saccade landing. The red dashed lines show the speed threshold of 30°/s.

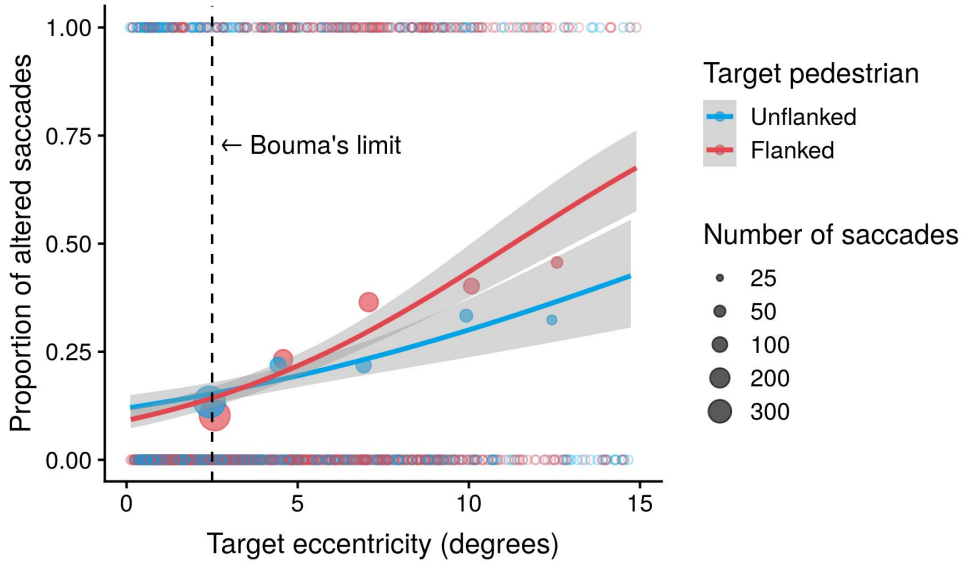

Figure 4.3: The proportion of altered saccades (PAS) versus target eccentricity for flanked and unflanked targets. Red represents the flanked targets, i.e., the target pedestrians with other flanking pedestrians in their $2.5°$ vicinity. Blue represents the unflanked targets, i.e., the target pedestrians with no other flanking pedestrian in their $2.5°$ vicinity. The hollow circles show the data of individual saccades. The solid circles show the mean PAS of the eccentricity bins, and the circle size indicates the number of saccades in the bin. The solid curves show the logistic regression fitting, and the gray ribbons represent the 95% confidence intervals. The vertical dashed line shows $2.5°$ eccentricity around which the PAS for flanked and unflanked targets were expected to be similar if the data followed Bouma's rule.

of $2.5°$ and β quantifies how fast the PAS increases with increasing eccentricity. We compared the parameters fitted for flanked targets and unflanked targets and the results followed our expectation ($\alpha_f - \alpha_u = -0.08$, permutation test $p = 0.56$; $\beta_f - \beta_u = 0.088$, permutation test $p = 0.006$; Figure 4.3).

In addition to the logistic regression, we also calculated the mean PAS for the eccentricity range within $5°$ and the eccentricity range beyond $5°$. The results (Figure 4.4) showed similar PAS for flanked and unflanked targets for eccentricities within $5°$ (permutation test $p = 0.36$) and a significantly higher mean PAS for flanked targets than for unflanked targets for target eccentricity larger than $5°$ (permutation test $p = 0.01$).

More altered saccades for pedestrian flankers than for car flankers

Crowding literature has previously shown that flankers similar to the target crowd more than dissimilar ones (Reuther and Chakravarthi, 2014; for a review see Manassi and Whitney, 2018). Therefore, the flanking effect on pedestrian-targeted saccades discussed in the previous section

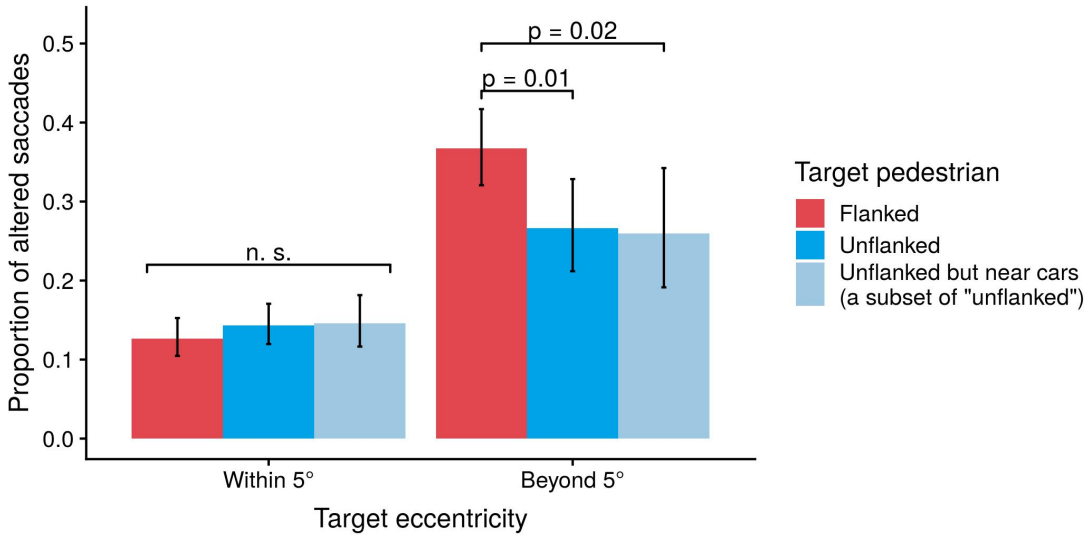

Figure 4.4: Mean proportion of altered saccades (PAS) for different target eccentricity ranges and different kinds of target pedestrians. Red represents the target pedestrians with other flanking pedestrians in their $2.5°$ vicinity (flanked). Dark blue represents the target pedestrians with no other flanking pedestrian in their $2.5°$ vicinity (unflanked). Light blue represents the target pedestrians with cars but no other pedestrians in their $2.5°$ vicinity, which are a subset of the unflanked target pedestrians. The error bars represent 95% confidence intervals.

should be specific to pedestrian flankers. The unflanked targets in the previous section were defined as target pedestrians with no other pedestrians in their $2.5°$ vicinity, so they might very well be flanked by other types of objects commonly seen in driving scenes. To more specifically address this concern, we identified the subset of unflanked target pedestrians that had cars but no other pedestrians within a $2.5°$ vicinity. We calculated mean PAS for two groups: target pedestrians within $5°$ that were flanked by cars, and target pedestrians beyond $5°$ eccentricity that were flanked by cars. We obtained values similar to the whole set of unflanked targets (i.e., flanked by no pedestrian). Comparison against the targets flanked by pedestrians showed that, beyond $5°$ eccentricity, the mean PAS associated with car flankers was significantly lower than that associated with pedestrian flankers (permutation test $p = 0.02$, Figure 4.4). Within $5°$ eccentricity, there was no significant difference between car flankers and pedestrian flankers (permutation test $p = 0.37$, Figure 4.4). The results suggest that similar flankers were more effective at crowding the target, consistent with the well-known similarity tuning of crowding (Andriessen and Bouma, 1976; Kooi et al., 1994; Chung, Levi, and Legge, 2001; Dennis M. Levi, 2008).

Consistency with Bouma's Rule

A signature of crowding is that regardless of the target and flanker size, critical spacing (i.e., the largest target-flanker spacing at which the recognition of the target is affected) is roughly one

half of the target eccentricity (Bouma's rule-of-thumb, Bouma, 1970). Bouma's rule-of-thumb was proven to be consistent across a wide range of stimuli (e.g., oriented gratings, shapes, letters, faces, etc.), although the exact value can be strongly affected depending on the task, stimulus, attentional demands, etc. (Strasburger et al., 2011; Whitney and Levi, 2011). Nevertheless, the half-eccentricity rule of thumb is a reasonable estimate of the average critical spacing at which crowding would often be expected to occur. To test whether the influence of flanking pedestrians on saccade landing accuracy follows this rule, we collected all the valid saccades that landed on pedestrians for the following analysis regardless of the flanker size, target size, and target eccentricity. We plotted the PAS against the spacing-to-eccentricity ratio (the ratio between the target-flanker spacing and the target eccentricity, Figure 4.5A). We performed a clipped line fitting to the data, i.e., we fit two straight lines for the spacing-to-eccentricity ratio range below 0.5 and the ratio range above 0.5 respectively with a shared intercept at the ratio equal to 0.5. In other words, the fitting included three free parameters: the two slopes of the two lines and the shared intercept at the ratio equal to 0.5. The data points over the ratio range above 5 were sparse and were therefore excluded from the fitting for robustness. The fitting result (Figure 4.5A) showed a steep negative slope over the ratio range below 0.5 ($\beta_{<0.5} = -0.33$) and a relatively flat slope over the ratio range above 0.5 ($\beta_{\geqslant 0.5} = 0.03$). The difference between the two slopes was consistent with Bouma's rule-of-thumb. The target eccentricity could be a confounder here because larger target eccentricity leads to higher PAS (as seen in Figure 3) and that the saccades of different spacing-to-eccentricity ratios have different mean target eccentricities. To determine the significance level of the difference in slopes ($\beta_{<0.5} - \beta_{\geqslant 0.5}$) after accounting for target eccentricity, we added target eccentricity as a regressor into the clipped line fitting model and conducted a permutation test where we shuffled the spacing-to-eccentricity ratio values of the saccades. The results showed a significant slope change ($\beta_{<0.5} - \beta_{\geqslant 0.5} = -0.36$, permutation-test $p < 0.001$, Figure 4.5B).

Radial-tangential anisotropy

Another signature of crowding is radial-tangential anisotropy: flankers aligned along the radial direction cause stronger crowding than flankers aligned along the tangential direction (Toet and Levi, 1992). To test whether the influence of flanking pedestrians on saccade landing accuracy follows the radial-tangential anisotropy, for each flanked target, we calculated the angle between the line connecting the target and the flanker and the line connecting the starting and landing points of the saccade (α, $\alpha \in [0°, 90°]$). If the angle α was smaller than $30°$, the flanker was identified as a radial flanker. If the angle α was greater than $60°$, the flanker was identified as a tangential flanker.

First, among the saccades with a spacing-to-eccentricity ratio below 1, where crowding might happen, we calculated the mean PAS separately for the saccades with radial flankers and the ones with tangential flankers. The mean PAS for radial flankers was higher than the mean PAS for tangential flankers (Figure 4.6). This positive difference in mean PAS between radial and tangential flankers is consistent with the radial-tangential anisotropy of crowding. However, target eccentricity might contribute to this difference as a confounder. To determine the baseline PAS values under the null hypothesis that PAS depends on target eccentricity but not flanker alignment, we fit the

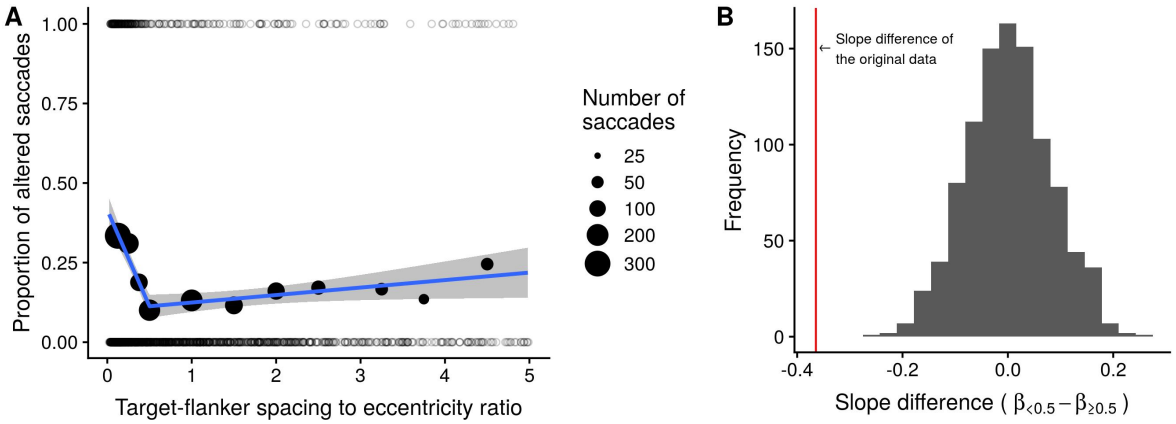

Figure 4.5: (A) Proportion of altered saccades (PAS) versus spacing-to-eccentricity ratio. The hollow circles show the data of individual saccades. The solid circles show the mean PAS of the ratio bins, and the circle size indicates the number of saccades in the bin. The solid curves show the clipped line fit, and the gray ribbons represent the 95% confidence intervals. (B) A permutation test was conducted to test the significance level of the difference between the two slopes of the clipped line fit ($\beta_{<0.5} - \beta_{\geqslant 0.5}$). The spacing-to-eccentricity ratio values of the saccades were shuffled 1000 times. The histogram summarizes the slope differences fitted to the shuffled data. The red line shows the slope difference of the original data, which was significantly negative ($\beta_{<0.5} - \beta_{\geqslant 0.5} = -0.36$, permutation-test $p < 0.001$).

following logistic regression model that only has an intercept and target eccentricity as regressors:

$$\text{Model 2}: \quad \log(\frac{p}{1-p}) = \alpha + \gamma \cdot eccen \tag{4.2}$$

Model 2 allowed us to simulate how the data would distribute under the null hypothesis. We simulated whether a saccade would be altered under the null hypothesis based on a binomial distribution with the probability of being altered equal to the probability predicted by Model 2. Then we calculated the mean PAS of the simulated data separately for the radial and tangential flankers. These mean PAS values are the baseline PAS values under the null hypothesis (plotted as dashed lines in Figure 4.6). To determine the significance level of the difference in mean PAS between radial and tangential flankers, we first fit the following model:

$$\text{Model 3}: \quad \log(\frac{p}{1-p}) = \alpha + \beta \cdot X_{radial} + \gamma \cdot eccen \tag{4.3}$$

where X_{radial} is a dummy variable that is equal to 1 when the flanker alignment of the saccade is radial and 0 when the flanker alignment is tangential. β quantifies the independent influence of radial versus tangential flanker alignment on PAS after accounting for the influence of target eccentricity. The fitting showed that $\beta = 0.75$, permutation test $p = 0.01$. The effect was significant even after accounting for target eccentricity.

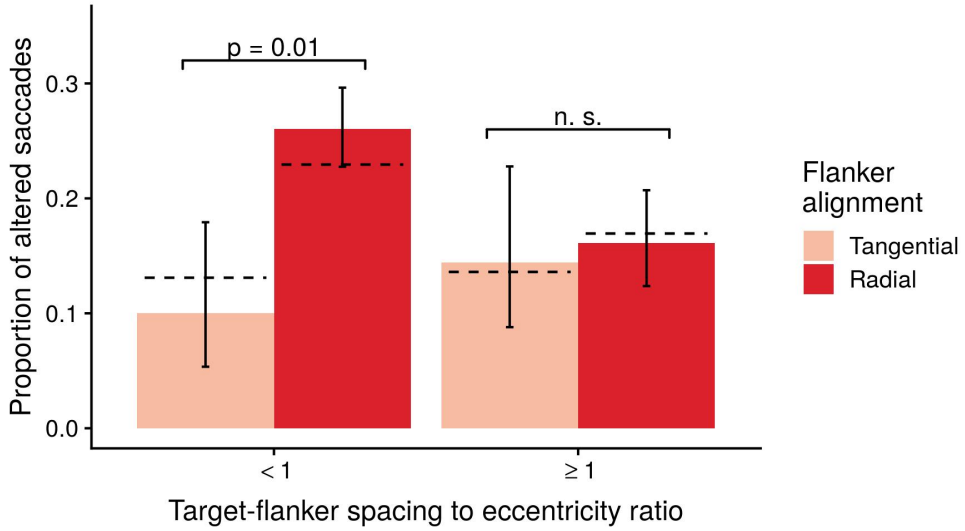

Figure 4.6: Mean proportion of altered saccades for tangential and radial flankers. Pink bars and red bars represent the tangential and radial flankers, respectively. The error bars represent 95% confidence intervals. The dashed lines show the baseline values under the null hypothesis that mean PAS only depends on the target eccentricity but not the flanker alignment.

We applied the same calculations to the saccades with a spacing-to-eccentricity ratio above 1, where crowding was unlikely to occur. The results showed that there was no significant difference in mean PAS between the saccades with radial and tangential flankers after accounting for target eccentricity (Figure 4.6, permutation test $p = 0.80$). Overall, the result suggested that the influence of flanking pedestrians on saccade landing accuracy is consistent with the radial-tangential anisotropy.

To summarize the analyses of Experiment 1, we found that visual clutters around the target pedestrians are associated with altered saccadic localization in manners that are consistent with diagnostic criteria of crowding, i.e., Bouma's rule of thumb, flanker similarity tuning and radial-tangential anisotropy.

Experiment 2

In order to further confirm that the saccade landing inaccuracy discussed above is a behavioral consequence of crowding, we conducted a more conventional crowding experiment with a pedestrian gender discrimination task. We identified the pedestrian-targeted saccades from Experiment 1 and used the video frames at the ending points of those saccades as static stimulus images for Experiment 2. Participants were asked to fixate at the starting point of the saccades and to identify the gender of the pedestrian on which the saccade landed. In Experiment 2, we tested whether crowding impaired gender recognition. We also explored the correlation between the saccade landing

accuracy in Experiment 1 and the gender discrimination accuracy in Experiment 2.

Methods

Participants

Ten participants (five males and five females) participated in this experiment. They had normal or corrected normal vision.

Stimuli and display setup

602 pedestrian-targeted saccades identified in Experiment 1 were used to make the stimuli of Experiment 2. The video frames at the ending points of those saccades were extracted as static stimulus images. The display setup was kept the same as Experiment 1. The stimuli were displayed on a CRT monitor (display area size 34 cm $\times$ 23 cm) The display resolution was set to 1024 $\times$ 768 and the refresh rate to 60 Hz. Participants viewed the stimuli binocularly in a darkened experimental booth and head position was stabilized with a chinrest at a viewing distance of 57 cm.

Procedure

There were three runs with self-paced pauses in between and 140 trials in each run. For each participant, 140 stimulus images were repeatedly used in the three runs but displayed in different random orders. The 140 stimulus images viewed by each participant were extracted from all different videos so that the target pedestrians for each participant were all different people. The participants viewed the stimulus images with required fixation in the first two runs and freely in the third run, and reported the perceived gender of the target pedestrian by key press. Their responses made in the third run were used as the subjective ground-truth of the gender of the target pedestrians to determine whether their response made in the first two runs were correct. Using other participants' responses made in the third run as ground-truth did not change the results qualitatively.

In the first two runs, participants' eye movements were tracked by an Eyelink 1000 at 1000 Hz for fixation monitoring. In each trial, first a pre-cue image was displayed. The pre-cue image consisted of a gray background, a white fixation cross showing the required fixation point (the starting point of the pedestrian-targeted saccade in Experiment 1) and a red bounding box showing the location of the target pedestrian (the pedestrian on which the pedestrian-targeted saccade in Experiment 1 landed). The pre-cue image was displayed for at least one second and until the participant fixated on the fixation cross. Then the stimulus image was displayed with the fixation cross superimposed and two red bars right above and below the target pedestrian. The two red bars were to help the participant appreciate which pedestrian was the target pedestrian. The fixation was monitored in real-time, and the target was presented gaze-contingently. Once the participant's gaze was more than 50 pixels ($1.5^\circ \sim 1.7^\circ$ in viewing angle) away from the required fixation point, the stimulus image was masked by the pre-cue image. The stimulus was displayed again once the fixation was restored. This process continued until the response was made, but no longer than two

seconds after the initial onset of the stimulus image. The participant reported the perceived gender by key press and then the next trial started.

In the third run, there was no pre-cue image. In each trial, the stimulus image was displayed with the fixation cross and the red bounding box around the target pedestrian on top. The participant viewed the stimulus image freely with unlimited time until they made a response by key press. The next trial started right after the response was made.

Results and discussion

Lower accuracy when the target pedestrians were flanked

Following the selection criterion in Experiment 1, we divided the target pedestrians into flanked and unflanked targets according to whether there were other flanking pedestrians in the 2.5° vicinity of the target pedestrians. For both flanked and unflanked targets, the data showed that gender discrimination accuracy dropped with increasing target eccentricity (Figure 4.7). If crowding occurred, we would expect lower accuracy for flanked targets than for unflanked targets given the same target eccentricity. According to Bouma's rule, we expected that flanked and unflanked targets would show similar accuracy on the eccentricity range between 0° and 5°; beyond 5°, the accuracy for both flanked and unflanked targets would decrease with increasing target eccentricity, but the decrease for flanked targets should be significantly faster. Therefore, to test our hypothesis, we fit the following logistic regression model between gender discrimination accuracy and target eccentricity:

$$\text{Model 4}: \quad \log(\frac{accur - 0.5}{1 - accur}) = \alpha + \beta \cdot (eccen - 2.5^\circ) \tag{4.4}$$

where $accur$ is the gender discrimination accuracy, $eccen$ is the eccentricity of the target and α and β are fitted parameters. α indicates the fitted accuracy at an eccentricity of 2.5° and β quantifies how fast the accuracy decreases with increasing eccentricity. We compared the parameters fitted for flanked targets and unflanked targets and the results followed our expectation ($\alpha_f - \alpha_u = 0.016$, permutation test $p = 0.92$; $\beta_f - \beta_u = -0.22$, permutation test $p < 0.001$; Figure 4.7).

However, the visual size of the target pedestrian might be a confounder here because the data showed that flanked pedestrians on average had smaller visual sizes and that low gender discrimination accuracy was correlated with small target visual size. To determine the baseline accuracy-eccentricity curves for flanked and unflanked targets under the null hypothesis that the accuracy was influenced by both the eccentricity and target visual size but not whether the target was flanked or not, we added target visual size as a regressor into Model 4 and fit the following new model to all the data (i.e., the union of flanked and unflanked targets):

$$\text{Model 5}: \quad \log(\frac{accur - 0.5}{1 - accur}) = \alpha + \beta \cdot (eccen - 2.5^\circ) + \mu \cdot size \tag{4.5}$$

where $size$ is the visual size of the target pedestrian. This model captured how the accuracy varied based on target eccentricity and target visual size regardless whether the target was flanked

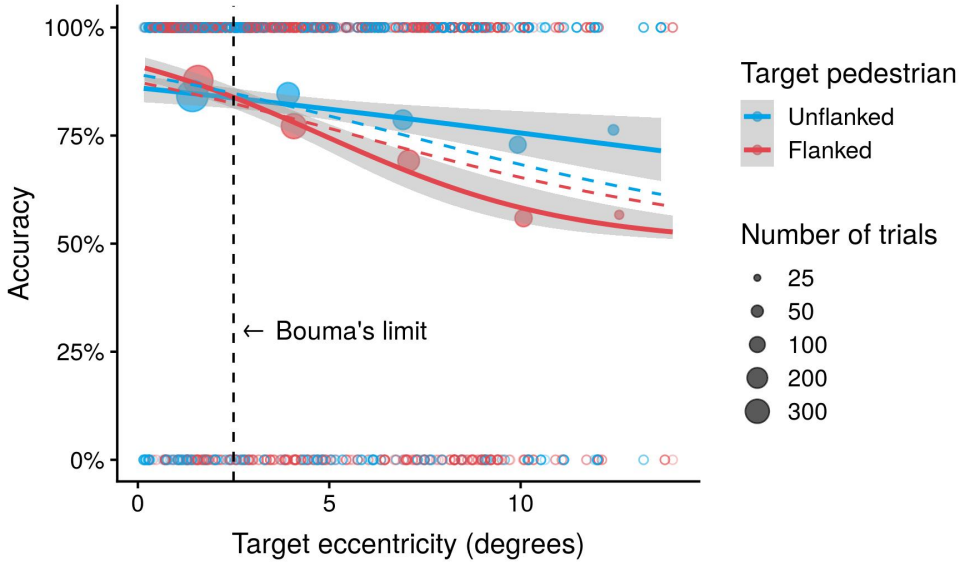

Figure 4.7: Gender discrimination accuracy versus target eccentricity for flanked and unflanked targets. The hollow circles show individual trial data. The solid circles show the mean accuracies of the eccentricity bins, and the circle size indicates the number of trials in the bin. The solid curves show the logistic regression fit, and the gray ribbons represent the 95% confidence intervals. The dashed lines are showing the baseline accuracy-eccentricity curves for flanked and unflanked targets under the null hypothesis that the accuracy depends on target eccentricity and target visual size but not whether the target pedestrian is flanked or not. The vertical dashed line shows $2.5°$ eccentricity around which the accuracy for flanked and unflanked targets are expected to be similar if the data follows Bouma's rule.

or not. Then for each trial, we simulated the outcome of the trial (i.e., correct or wrong) based on a binomial distribution with the probability of correctness equal to the accuracy predicted by Model 5. The simulated data showed how the data would distribute under the null hypothesis. Then we fit Model 4 to the simulated data separately for flanked targets and unflanked targets to obtain the baseline accuracy-eccentricity curves under the null hypothesis. The baseline curves are plotted as dashed curves in Figure 4.7. To calculate the significance level of the difference in how fast accuracy decreased with eccentricity between flanked and unflanked targets after discounting the effect of target visual size, we fit Model 5 separately to the flanked data and unflanked data. This time $\beta_f - \beta_u = -0.24$. We did a permutation test where we shuffled the flanked/unflanked label of each trial to get a null distribution of $\beta_f - \beta_u$. The permutation test showed that $p < 0.001$. So we confirmed that even after discounting the effect of target visual size the gender discrimination accuracy still dropped with increasing eccentricity significantly faster for flanked targets than for unflanked targets. Similarly, we confirmed that after discounting the effect of target visual size, there was no significant difference in the accuracy at $2.5°$eccentricity between flanked

and unflanked targets ($\alpha_f - \alpha_u = 0.23$, permutation test $p = 0.49$).

Besides the logistic regression, we also calculated the mean gender discrimination accuracy for the eccentricity range below 5° and the eccentricity range beyond 5°. The results (Figure 4.8) showed similar accuracies for flanked and unflanked targets for target eccentricities less than 5° ($accur_f - accur_u = -0.05\%$) and a lower mean accuracy for flanked targets than for unflanked targets for target eccentricities beyond 5° ($accur_f - accur_u = -13.4\%$).

Again, in order to account for the influence of target visual size, we calculated the baseline accuracies under the null hypothesis that either below 5° eccentricity or above 5° eccentricity the mean accuracy is only influenced by target visual size but not whether the target is flanked or not. We first fit the following logistic regression model to the data below 5° eccentricity:

$$\text{Model 6}: \quad \log(\frac{accur}{1 - accur}) = \alpha + \mu \cdot size \tag{4.6}$$

Then for each trial below 5° eccentricity, we simulated the outcome of the trial (i.e., correct or wrong) based on a binomial distribution with the probability of correctness equal to the accuracy predicted by Model 6. The simulated data showed how the data would distribute under the null hypothesis. Then we calculated the mean accuracies of the flanked and unflanked trials of the simulated data, which are the baseline accuracies under the null hypothesis (plotted as dashed lines in Figure 4.8). To determine the significance level of the difference between flanked and unflanked data in the original data, we first fit the following model to the data below 5° eccentricity:

$$\text{Model 7}: \quad \log(\frac{accur}{1 - accur}) = \alpha + \beta \cdot X_{flanked} + \mu \cdot size \tag{4.7}$$

where $X_{flanked}$ is a dummy variable that is equal to 1 when the target is flanked and 0 otherwise and β quantifies the independent influence of being flanked versus unflanked on mean accuracy after accounting for the influence of target visual size. The fitting showed that $\beta = 0.18$, permutation test $p = 0.20$. So no significant independent effect from being flanked versus unflanked for the data below 5° eccentricity. We applied the same calculations to the data above 5° eccentricities to get the baseline mean accuracies and the significance level of the independent effect from being flanked versus unflanked. The baseline accuracies are shown as dashed lines in Figure 4.8 and $\beta = -0.54$, permutation test $p < 0.001$, which showed that being flanked versus unflanked significantly decreased the mean accuracy even after accounting for the influence of target visual size.

Lower accuracy for pedestrian flankers than for car flankers

Similar to Experiment 1, we identified the subset of unflanked target pedestrians that had cars but no other pedestrian in their 2.5° vicinity. We calculated the mean gender discrimination accuracy for the target pedestrians flanked by cars for within 5° eccentricity and beyond 5° eccentricity and obtained values similar to the whole set of unflanked targets (i.e., flanked by no pedestrian; Figure 4.8). We then compared these mean accuracies calculated with these trials with car flankers to the mean accuracies calculated with the trials with pedestrian flankers. We applied the same calculations to account for the influence of target visual size as a confounder. The baseline mean

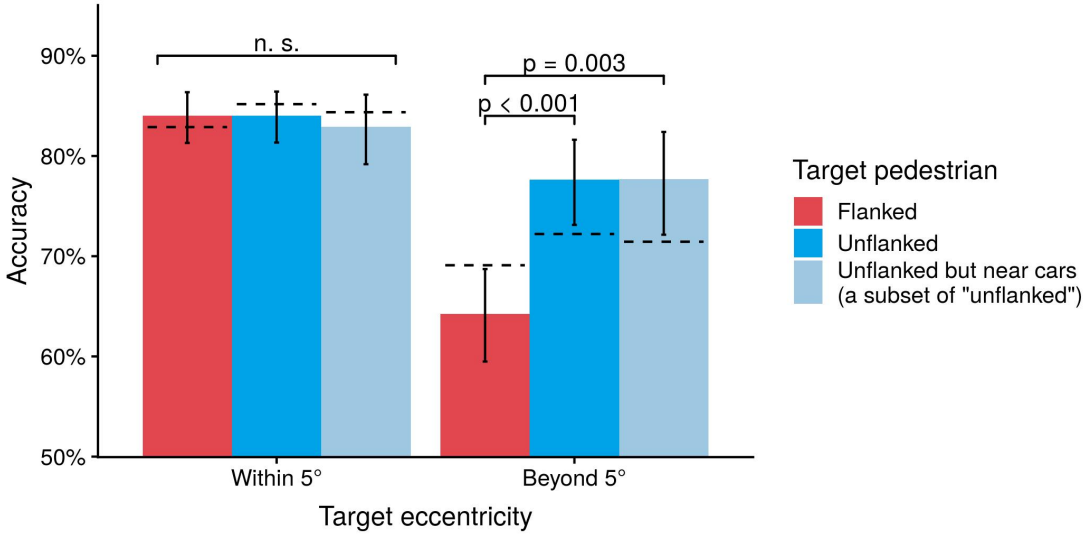

Figure 4.8: Mean gender discrimination accuracy for different target eccentricity ranges and different kinds of target pedestrians. The dashed lines show the baseline values under the null hypothesis that within $5°$ eccentricity or beyond $5°$ eccentricity the gender discrimination accuracy only depends on target visual size but not the types of the target pedestrians. The error bars represent 95% confidence intervals.

accuracies are shown as dashed lines in Figure 4.8. The result showed that, beyond $5°$ eccentricity, having pedestrian flankers versus car flankers significantly decreased the mean accuracy ($\beta = -0.53$, permutation test $p = 0.003$). Within $5°$ eccentricity, there was no significant independent effect from having pedestrian flankers versus pedestrian flankers ($\beta = 0.17$, permutation test $p = 0.30$). These results, again, show the target-flanker similarity tuning of crowding.

Consistency with Bouma's Rule

To test Bouma's rule, we collected all the trials for the following analysis regardless of flanker size, target size, and target eccentricity. We plot gender discrimination accuracy against spacing-to-eccentricity ratio (the ratio between the target-flanker spacing and the target eccentricity, Figure 4.9). We performed the same clipped line fitting as described in Experiment 1 (Figure 4.5A). The data points over the ratio range above 5 were sparse and were therefore excluded from the fitting for robustness. The fitting result (Figure 4.9A) showed a steep positive slope over the ratio range below 0.5 ($\beta_{<0.5} = 0.52$) and a relatively flat slope over the ratio range above 0.5 ($\beta_{\geqslant 0.5} = 0.01$). The difference between the two slopes was consistent with Bouma's rule-of-thumb. To determine the significance level of the slope difference ($\beta_{<0.5} - \beta_{\geqslant 0.5}$) after accounting for target eccentricity and target visual size, we added target eccentricity and target visual size as additional regressors into the clipped line fitting model and conducted a permutation test where we shuffled

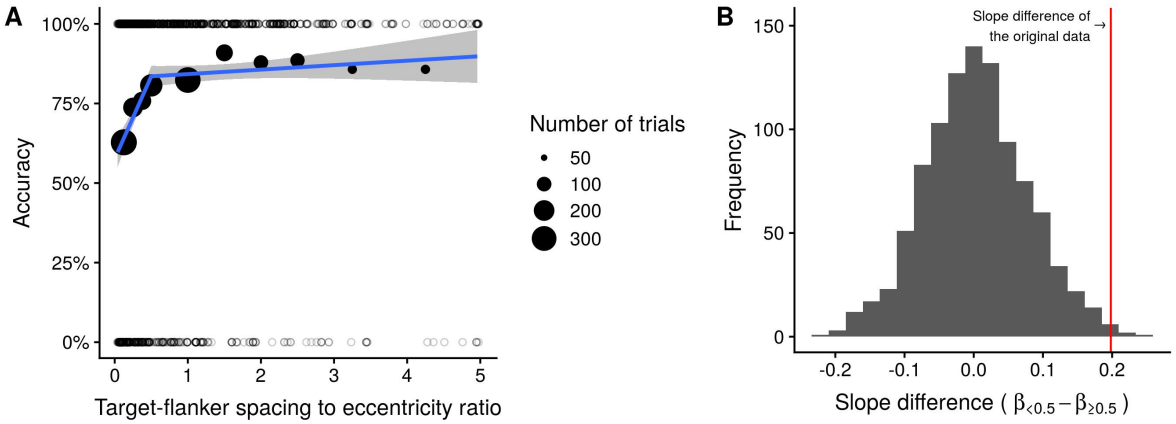

Figure 4.9: (A) Gender discrimination accuracy versus spacing-to-eccentricity ratio. The hollow circles show individual trial data. The solid circles show the mean accuracy of the ratio bins, and the circle size indicates the number of trials in the bin. The solid curves show the clipped line fit, and the gray ribbons represent the 95% confidence intervals. The dashed lines show the baseline fitting under the null hypothesis that the accuracy depends on visual target size but not the spacing-to-eccentricity ratio. (B) A permutation test was conducted to test the significance level of the difference between the two slopes of the clipped line fit ($\beta_{<0.5} - \beta_{\geqslant 0.5}$). The spacing-to-eccentricity ratio values of the trials were shuffled 1000 times. The histogram summarizes the slope differences fitted to the shuffled data. The red line shows the slope difference of the original data, which was significantly positive ($\beta_{<0.5} - \beta_{\geqslant 0.5} = 0.20$, permutation-test $p = 0.01$).

the spacing-to-eccentricity ratio values of the trials. The results showed a significant slope change ($\beta_{<0.5} - \beta_{\geqslant 0.5} = 0.20$, permutation-test $p = 0.01$, Figure 4.9B).

Radial-tangential anisotropy

Among the saccades with a spacing-to-eccentricity ratio below 1, where crowding might happen, we calculated mean gender discrimination accuracies separately for the saccades with radial flankers and the ones with tangential flankers. To account for target eccentricity and target visual size as potential confounders, we fit the following logistic regression model:

$$\text{Model 8}: \quad \log(\frac{accur}{1-accur}) = \alpha + \gamma \cdot eccen + \mu \cdot size \tag{4.8}$$

Model 8 allowed us to simulate how the data would distribute under the null hypothesis that gender discrimination accuracy is only influenced by target eccentricity and target visual size but not flanker alignment. We simulated the outcome of each trial (i.e., correct or wrong) based on a binomial distribution with the probability of correctness equal to the accuracy predicted by Model 8. Then we calculated the mean accuracies of the flanked and unflanked trials of the simulated data, which are the baseline accuracies under the null hypothesis (plotted as dashed lines in Figure

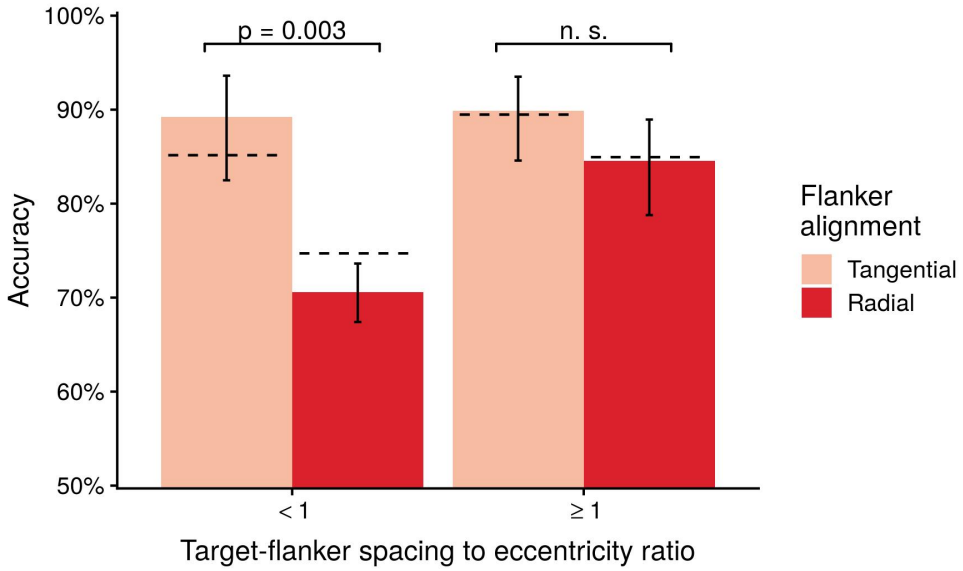

Figure 4.10: Mean gender discrimination accuracy for tangential and radial flankers with low spacing-to-eccentricity ratios (< 1) and high spacing-to-eccentricity ratios (≥ 1). The dashed lines show the baseline mean accuracies under the null hypothesis that either when spacing-to-eccentricity ratio is low or high the mean accuracy depends on target eccentricity and target visual size but not the flanker alignment. The error bars represent 95% confidence intervals.

4.10). To determine the significance level of the difference between flanked and unflanked data in the original data, we first fit the following model:

$$\text{Model 9}: \quad \log(\frac{accur}{1-accur}) = \alpha + \beta \cdot X_{radial} + \gamma \cdot eccen + \mu \cdot size \tag{4.9}$$

where X_{radial} is a dummy variable that is equal to 1 when the flanker alignment is radial and 0 when the flanker alignment is tangential and β quantifies the independent influence of being radial versus tangential on mean accuracy after accounting for the influence of target eccentricity and target visual size. The result showed that radial versus tangential flanker alignment significantly decreased the mean accuracy after accounting for target eccentricity and visual size ($\beta = -0.75$, permutation test $p = 0.003$). We applied the same calculations to the data with spacing-to-eccentricity ratios above 1 where crowding was unlikely to occur. The results showed that there was no significant independent effect of radial versus tangential flanker alignment after accounting for target eccentricity and visual size ($\beta = -0.10$, permutation test $p = 0.76$). Overall, the result confirmed the presence of a radial-tangential anisotropy in Experiment 2, consistent with crowding.

Correlation between saccade landing accuracy and gender discrimination accuracy

Since the stimuli of Experiment 2 were made based on the pedestrian-targeted saccades collected in Experiment 1, we could correlate the gender discrimination accuracy of Experiment 2 with the saccade landing accuracy of Experiment 1. If the altered saccadic localization discussed in Experiment 1 is a behavioral consequence of crowding observed in Experiment 2, the trials in Experiment 2 that corresponded to the altered saccades in Experiment 1 would show a lower mean gender discrimination accuracy than the trials that corresponded to the direct saccades. The result confirmed our hypothesis ($accur_{altered} - accur_{direct} = -8.6\%$). To account for target eccentricity and target visual size, we fit Model 8 to the data and then simulated the outcome of each trial (i.e., correct or wrong) based on a binomial distribution with the probability of correctness equal to the accuracy predicted by Model 8. Then we calculated the mean gender discrimination accuracies associated with altered and direct saccades, which are the baseline mean accuracies under the null hypothesis (plotted as dashed lines in Figure 4.11). To determine the significance level of the difference between altered and direct saccades in the original data, we fit the following model to the data:

$$\text{Model 10}: \quad \log(\frac{accur}{1-accur}) = \alpha + \beta \cdot X_{altered} + \gamma \cdot eccen + \mu \cdot size \tag{4.10}$$

where $X_{altered}$ is a dummy variable that is equal to 1 when the trial corresponded to an altered saccade in Experiment 1 and 0 otherwise and β quantifies the independent influence of corresponding to an altered saccade versus to a direct saccade on the mean gender discrimination accuracy after accounting for the confounding influence of target eccentricity and target visual size. The result showed that altered versus direct saccades significantly decreased the mean gender discrimination accuracy after accounting for target eccentricity and target visual size ($\beta = -0.31$, permutation test $p = 0.02$).

General discussion and conclusion

In Experiment 1, we used crowd-sourced natural driving videos as stimuli and recorded participants' eye movements during a simulated driving task. We identified the saccades that landed on pedestrians by using a Deep Learning object detection algorithm. We found that visual clutters around the target pedestrians are associated with altered saccadic localization in manners that are consistent with the diagnostic criteria of crowding. In Experiment 2, we used a pedestrian gender discrimination task with a conventional psychophysical paradigm to confirm that visual crowding, defined consistently with previous studies, indeed occurs in the recognition of the pedestrians targeted by the saccades recorded in Experiment 1. Importantly, we have shown that the altered saccadic localization observed in Experiment 1 is associated with the degree of crowding of the saccade targets measured in Experiment 2. Taken together, the results of Experiment 1 and 2 show strong evidence that visual crowding occurs in natural driving scenes and has behavioral consequences in driving-like situations (i.e., altered saccadic localization).

Figure 4.11: Mean gender discrimination accuracy of the trials using stimuli from altered saccades and the trials using stimuli from direct saccades. The dashed lines show the baseline mean accuracies under the null hypothesis that the mean accuracy depends on target eccentricity and target visual size but not whether the trial corresponds to an altered or direct saccade in Experiment 1.

To the best of our knowledge, our work demonstrates for the first time that crowding occurs in dynamic natural scenes of real life. Crowding has been studied for decades from low level features, such as oriented gratings, shapes, letters and symbols, to high level features, such as faces and biological motions. One important motivation of these studies is that crowding supposedly influences how we recognize cluttered objects in real life. However, the stimuli used in these studies are artificial, unnatural and often static (except for biological motions). Studies of crowding with dynamic natural stimuli/scenes are still needed for studying the impact of crowding in object recognition in real life. Wallis and Bex (Wallis and Bex, 2012) conducted an experiment where participants identified synthetic "dead leave" patches in natural scenes and they showed that the threshold size of "dead leave" patches scaled with eccentricity in a manner consistent with crowding. However, the "dead leave" patches were added artificially and sometimes inconsistent with the perspective and depth of the natural scene. Gong et al. (Gong et al., 2018) studied crowding in the recognition of the gist of natural scenes but they did not study crowding in object recognition in natural scenes. Previous studies also showed crowding of moving targets but limited to oriented gratings and simple shapes (Peter J. Bex, Dakin, and Simmers, 2003; Peter J. Bex and Dakin, 2005).

Importantly, our study uses saccades as a tool to study crowding in driving to avoid forcing participants to give unnatural and explicit responses. The link between saccades and crowding is supported by previous studies. Recent studies suggest important links between crowding and saccades (Harrison, Mattingley, and Remington, 2013; Wolfe and Whitney, 2014; Yildirim, Meyer, and Cornelissen, 2015; Greenwood, Szinte, Sayim, and Cavanagh, 2017). Specifically, Greenwood et al. (Greenwood et al., 2017) found that saccade precision and the size of crowding zone vary across the visual field with a strong correlation. Yildirim et al. (Yildirim et al., 2015) demonstrated that saccadic target localization is tuned to target-flanker similarity.

The crowding of the recognition of pedestrian that we found is consistent with previous studies on lower level features. It has been shown that crowding occurs in the recognition of faces (Louie, Bressler, and Whitney, 2007; Farzin, Rivera, and Whitney, 2009) and both local features and global configuration of flanking faces contribute to crowding (Sun and Balas, 2015). In addition, Ikeda et al. demonstrated crowding of biological motions using moving dots with configurations of walkers (Ikeda, Watanabe, and Cavanagh, 2013), providing further notion to the idea that crowding can occur between dynamic representations, similar to moving pedestrians.

Our study focuses on driving because driving is presumably the most important real-life situation that may involve crowding given its frequency and potentially fatal risks. Sanocki et al. (Sanocki, Islam, Doyon, and Lee, 2015) conducted an experiment where observers looked for pedestrians in briefly presented traffic scenes, and demonstrated higher miss rates associated with cluttered scenes. However, the diagnostic criteria of crowding were not tested, so it is not clear whether the effect they found was due to crowding or other phenomena such as visual masking and surround suppression. We believe that our study is the first one that demonstrates the behavioral consequences of crowding in dynamic driving-like situations, i.e., altered saccadic localization. This finding has significant implications for public safety. On one hand, it raises safety concerns about visual clutter in traffic scenes; on the other hand, it suggests that the knowledge that we have gained from decades of studies of crowding can be used in traffic designs to address these concerns. For example, road signs may need to be placed with enough spacing in between, or construction workers may need to wear safety vests in different colors.

Chapter 5

Conclusion

In the research discussed in this dissertation, we have seen the application of driver eye movements in improving autonomous driving models and studying the bottleneck of our object recognition in cluttered scenes. Chapter 2 lays the foundation of this research by introducing a new in-lab driver eye movement collection protocol. The new protocol allows accurately and efficiently collecting driver eye movements in an offline manner using videos of specific driving situations. We also developed a driver gaze/attention prediction model that highlights the important regions in dynamic driving scenes. We extend this work in Chapter 3 by incorporating the driver gaze prediction model into an autonomous driving model. The new autonomous driving model processes the whole video frames in low resolution and the predicted gaze regions in high resolution and then predicts the speed control of the vehicle. We show that the guidance from human attention significantly improves the driving model's prediction accuracy and makes it especially more robust in critical situations involving pedestrians. The periphery-fovea multi-resolution design mimicking human vision is also shown to be more efficient than a conventional uni-resolution design. In Chapter 4, we use driver eye movements collected through our new protocol as a tool to study visual crowding, which is the major bottleneck of human visual object recognition in cluttered scenes. We show that visual crowding also occurs in realistic driving scenes and has behavioral consequences in driving-like situations. The degree of crowding correlates with altered saccadic localization. Besides its implication in human vision science, this result also suggests that our knowledge gained from decades of studies on visual crowding can be used to guide driving safety design.

In closing, the research discussed in this dissertation uses driver eye movements to develop better computer vision models for autonomous driving and to study the bottleneck of human visual object recognition in realistic driving contexts. The success of the interdisciplinary approaches used in this research suggests that the combination of human vision and computer vision can lead to fruitful progress for both fields.

Bibliography

Alletto, S., Palazzi, A., Solera, F., Calderara, S., & Cucchiara, R. (2016). Dr (eye) ve: A dataset for attention-based tasks with applications to autonomous and assisted driving. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 54–60).

Andriessen, J. & Bouma, H. (1976). Eccentric vision: Adverse interactions between line segments. *Vision Research*, *16*(1), 71–78.

Bazzani, L., Larochelle, H., & Torresani, L. (2016). Recurrent mixture density network for spatiotemporal visual attention. *arXiv preprint arXiv:1603.08199*.

Bex, P. J. [Peter J.] & Dakin, S. C. (2005). Spatial interference among moving targets. *Vision Research*, *45*(11), 1385–1398.

Bex, P. J. [Peter J.], Dakin, S. C., & Simmers, A. J. (2003). The shape and size of crowding for moving targets. *Vision Research*, *43*(27), 2895–2904.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., ... Zhang, J. [Jiakai], et al. (2016). End to end learning for self-driving cars. *CoRR abs/1604.07316*.

Boucart, M., Lenoble, Q., Quettelart, J., Szaffarczyk, S., Despretz, P., & Thorpe, S. J. (2016). Finding faces, animals, and vehicles in far peripheral vision. *Journal of Vision*, *16*(2), 10.

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature*, *226*(5241), 177–178.

Bruce, N. D. & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, *9*(3), 5–5.

Bruce, N. & Tsotsos, J. (2006). Saliency based on information maximization. In *Advances in neural information processing systems* (pp. 155–162).

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

Cavanagh, P. & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends in cognitive sciences*, *9*(7), 349–354.

Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., ... Lin, D. (2019). Hybrid Task Cascade for Instance Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4974–4983.

Chung, S. T., Levi, D. M., & Legge, G. E. (2001). Spatial-frequency and contrast properties of crowding. *Vision Research*, *41*(14), 1833–1850.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 3213–3223.

Cornelissen, F. W., Peters, E. M., & Palmer, J. (2002). The eyelink toolbox: Eye tracking with matlab and the psychophysics toolbox. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 613–617.

Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R. (2016). Predicting human eye fixations via an lstm-based saliency attentive model. *arXiv preprint arXiv:1611.09571*.

Das, A., Agrawal, H., Zitnick, L., Parikh, D., & Batra, D. (2017). Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, *163*, 90–100.

Erdem, E. & Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of vision*, *13*(4), 11–11.

Farzin, F., Rivera, S. M., & Whitney, D. (2009). Holistic crowding of Mooney faces. *Journal of Vision*, *9*(6), 18–18.

Flom, M. C., Heath, G. G., & Takahashi, E. (1963). Contour interaction and visual resolution: Contralateral effects. *Science*, *142*(3594), 979–980. doi:10.1126/science.142.3594.979

Fridman, L., Langhans, P., Lee, J., & Reimer, B. (2016). Driver gaze region estimation without use of eye movement. *IEEE Intelligent Systems*, *31*(3), 49–56.

Glorot, X. & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249–256).

Gong, M., Xuan, Y., Smart, L. J., & Olzak, L. A. (2018). The extraction of natural scene gist in visual crowding. *Scientific Reports*, *8*(1), 14073. doi:10.1038/s41598-018-32455-6

Greenwood, J. A., Szinte, M., Sayim, B., & Cavanagh, P. (2017). Variations in crowding, saccadic precision, and spatial localization reveal the shared topology of spatial vision. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(17), E3573–E3582.

Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., & Wierstra, D. (2015). Draw: A recurrent neural network for image generation. In *Proceedings of the 32nd international conference on machine learning (icml-15)* (pp. 1462–1471).

Groner, R., Walder, F., & Groner, M. (1984). Looking at faces: Local and global aspects of scanpaths. In *Advances in psychology* (Vol. 22, pp. 523–533). Elsevier.

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. In *Advances in neural information processing systems* (pp. 545–552).

Harrison, W. J., Mattingley, J. B., & Remington, R. W. (2013). Eye movement targets are released from visual crowding. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *33*(7), 2927–33.

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Hecker, S., Dai, D., & Van Gool, L. (2018). End-to-end learning of driving models with surround-view cameras and route planners. In *Proceedings of the european conference on computer vision (eccv)* (pp. 435–453).

Huang, X. [Xinyu], Wang, P., Cheng, X., Zhou, D., Geng, Q., & Yang, R. (2018). The ApolloScape Open Dataset for Autonomous Driving and its Application. *arXiv: 1803.06184*.

Huang, X. [Xun], Shen, C., Boix, X., & Zhao, Q. (2015). Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the ieee international conference on computer vision* (pp. 262–270).

Ikeda, H., Watanabe, K., & Cavanagh, P. (2013). Crowding of biological motion stimuli. *Journal of Vision*, *13*(4), 20–20.

Kim, J. & Canny, J. (2017). Interpretable learning for self-driving cars by visualizing causal attention. *ICCV*.

Kim, J., Rohrbach, A., Darrell, T., Canny, J., & Akata, Z. (2018). Textual explanations for self-driving vehicles. In *Proceedings of the european conference on computer vision (eccv)* (pp. 563–578).

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kooi, F. L., Toet, A., Tripathy, S. P., & Levi, D. M. (1994). The effect of similarity and duration on spatial interaction in peripheral vision. *Spatial vision*, *8*(2), 255–79.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kümmerer, M., Theis, L., & Bethge, M. (2015). Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *International Conference on Learning Representations (ICLR 2015)*.

Kümmerer, M., Wallis, T. S., & Bethge, M. (2016). Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*.

Levi, D. M. [D. M.], Hariharan, S., & Klein, S. A. (2002). Suppressive and facilitatory spatial interactions in peripheral vision: Peripheral crowding is neither size invariant nor simple contrast masking. *Journal of Vision*, *2*(2), 3–3.

Levi, D. M. [D. M.], Klein, S. A., & Hariharan, S. (2002). Suppressive and facilitatory spatial interactions in foveal vision: Foveal crowding is simple contrast masking. *Journal of Vision*, *2*(2), 2–2.

Levi, D. M. [Dennis M.]. (2008). Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research*, *48*(5), 635–654.

Liu, N., Han, J., Zhang, D., Wen, S., & Liu, T. (2015). Predicting eye fixations using convolutional neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 362–370).

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path Aggregation Network for Instance Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8759–8768.

Liu, Y., Zhang, S., Xu, M., & He, X. (2017). Predicting salient face in multiple-face videos. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4420–4428).

Louie, E. G., Bressler, D. W., & Whitney, D. (2007). Holistic crowding: Selective interference between configural representations of faces in crowded scenes. *Journal of Vision*, *7*(2), 24.

Maddern, W., Pascoe, G., Linegar, C., & Newman, P. (2017). 1 year, 1000 km: The Oxford RobotCar dataset. *The International Journal of Robotics Research*, *36*(1), 3–15.

Manassi, M. & Whitney, D. (2018). Multi-level Crowding and the Paradox of Object Recognition in Clutter. *Current Biology*, *28*(3), R127–R133.

Mannan, S., Ruddock, K., & Wooding, D. (1997). Fixation sequences made during visual examination of briefly presented 2d images. *Spatial vision*, *11*(2), 157–178.

Murray, N., Vanrell, M., Otazu, X., & Parraga, C. A. (2011). Saliency estimation using a non-parametric low-level vision model. In *Computer vision and pattern recognition (cvpr), 2011 ieee conference on* (pp. 433–440). IEEE.

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

Palazzi, A., Solera, F., Calderara, S., Alletto, S., & Cucchiara, R. (2017). Learning where to attend like a human driver. In *Intelligent vehicles symposium (iv), 2017 ieee* (pp. 920–925). IEEE.

Pelli, D. G. [D. G.], Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision*, *4*(12), 12.

Pelli, D. G. [Denis G] & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience*, *11*(10), 1129–1135.

Redmon, J. & Farhadi, A. (2017). Yolo9000: Better, faster, stronger. In *Computer vision and pattern recognition (cvpr), 2017 ieee conference on* (pp. 6517–6525). IEEE.

Reuther, J. & Chakravarthi, R. (2014). Categorical membership modulates crowding: Evidence from characters. *Journal of Vision*, *14*(6), 5.

Rizzolatti, G., Riggio, L., Dascola, I., & Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, *25*(1), 31–40.

Sanocki, T., Islam, M., Doyon, J. K., & Lee, C. (2015). Rapid scene perception with tragic consequences: observers miss perceiving vulnerable road users, especially in crowded traffic scenes. *Attention, Perception, & Psychophysics*, *77*(4), 1252–1262.

Simon, L., Tarel, J.-P., & Brémond, R. (2009). Alerting the drivers about road signs with poor visual saliency. In *Intelligent vehicles symposium, 2009 ieee* (pp. 48–53). IEEE.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Strasburger, H., Rentschler, I., & Juttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of Vision*, *11*(5), 13–13.

Sun, H.-M. & Balas, B. (2015). Face features and face configurations both contribute to visual crowding. *Attention, Perception, & Psychophysics*, *77*(2), 508–519.

Tawari, A. & Kang, B. (2017). A computational framework for driver's visual attention using a fully convolutional architecture. In *Intelligent vehicles symposium (iv), 2017 ieee* (pp. 887–894). IEEE.

Thomas, C. L. (2016). *Opensalicon: An open source implementation of the salicon saliency model* (tech. rep. No. TR-2016-02). University of Pittsburgh.

Toet, A. & Levi, D. M. [Dennis M.]. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research*, *32*(7), 1349–1357.

Underwood, G., Humphrey, K., & Van Loon, E. (2011). Decisions about objects in real-world scenes are influenced by visual saliency before and during their inspection. *Vision research*, *51*(18), 2031–2038.

Valenti, R., Sebe, N., & Gevers, T. (2009). Image saliency by isocentric curvedness and color. In *Computer vision, 2009 ieee 12th international conference on* (pp. 2185–2192). IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Nips*.

Wallis, T. S. A. & Bex, P. J. [P. J.]. (2012). Image correlates of crowding in natural scenes. *Journal of Vision*, *12*(7), 6–6.

Wang, D., Devin, C., Cai, Q.-Z., Yu, F., & Darrell, T. (2019). Deep object centric policies for autonomous driving. *ICRA*.

Wei, Y., Wen, F., Zhu, W., & Sun, J. (2012). Geodesic saliency using background priors. In *European conference on computer vision* (pp. 29–42). Springer.

Westheimer, G. & Hauske, G. (1975). Temporal and spatial interference with vernier acuity. *Vision Research*, *15*(10), 1137–1141.

Whitney, D. & Levi, D. M. [Dennis M.]. (2011). Visual crowding: a fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, *15*(4), 160–168.

Wolfe, B. A. & Whitney, D. (2014). Facilitating recognition of crowded faces with presaccadic attention. *Frontiers in Human Neuroscience*, *8*, 103.

Xia, Y., Zhang, D., Kim, J., Nakayama, K., Zipser, K., & Whitney, D. (2018). Predicting Driver Attention in Critical Situations. In *Asian conference on computer vision* (pp. 658–674). Springer, Cham.

Xu, H., Gao, Y., Yu, F., & Darrell, T. (2017a). End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2174–2182).

Xu, H., Gao, Y., Yu, F., & Darrell, T. (2017b). End-to-end Learning of Driving Models from Large-scale Video Datasets. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2174–2182.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., … Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048–2057).

Yildirim, F., Meyer, V., & Cornelissen, F. W. (2015). Eyes on crowding: Crowding is preserved when responding by eye and similarly affects identity and position accuracy. *Journal of Vision*, *15*(2), 21–21.

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

Zhang, J. [Jianming] & Sclaroff, S. (2013). Saliency detection: A boolean map approach. In *Computer vision (iccv), 2013 ieee international conference on* (pp. 153–160). IEEE.

Zhang, R., Liu, Z., Zhang, L., Whritner, J. A., Muller, K. S., Hayhoe, M. M., & Ballard, D. H. (2018). Agil: Learning attention from human for visuomotor tasks. In *Proceedings of the european conference on computer vision (eccv)* (pp. 663–679).

Zhu, Y., Groth, O., Bernstein, M., & Fei-Fei, L. (2016). Visual7w: Grounded question answering in images. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4995–5004).