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

Ye Xia Jinkyu Kim John Canny Karl Zipser Teresa Canas-Bajo David Whitney
University of California, Berkeley
{yexia, jinkyu.kim, canny, karlzipser, teresa_canasbajo, dwhitney}@berkeley.edu

Abstract

Inspired by human vision, we propose a new periphery-fovea multi-resolution driving model that predicts vehicle speed from dash camera videos. The peripheral vision module of the model processes the full video frames in low resolution with large receptive fields. Its foveal vision module selects sub-regions and uses high-resolution input from those regions to improve its driving performance. We train the fovea selection module with supervision from driver gaze. We show that adding high-resolution input from predicted human driver gaze locations significantly improves the driving accuracy of the model. Our periphery-fovea multi-resolution model outperforms a uni-resolution periphery-only model that has the same amount of floating-point operations. More importantly, we demonstrate that our driving model achieves a significantly higher performance gain in pedestrian-involved critical situations than in other non-critical situations. Our code is publicly available at `https://github.com/pascalxia/periphery_fovea_driving`.

1. Introduction

Vision-based deep autonomous driving models have shown promising results recently [13, 14, 4, 26]. 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. Human visual system has distinct foveal and peripheral structures that carry localized high-resolution information and long-range low-resolution information, respectively. The human fovea covers approximately two degrees of the central visual field and our vision within this area is featured by high spatial acuity and small receptive fields of visual neurons (*e.g.*, ganglian cells, LGN cells, visual cortical cells); the rest of our visual field, *i.e.*, the periphery, has low spatial acuity and large neural receptive fields. While peripheral vision constantly covers a large space, eye movements, guided by visual attention, direct the fovea to gather high-resolution information from salient/important regions. One advantage of this design is its efficiency: the peripheral vision efficiently processes long-range structures in low resolution with large receptive fields and the localized foveal vision dedicates resources to particularly salient or important regions. Driving scenes typically contain both important long-range structures (*e.g.*, curving roads and crossing) and localized regions with critical details (*e.g.*, pedestrians' gaze and traffic sign contents). Therefore, 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.

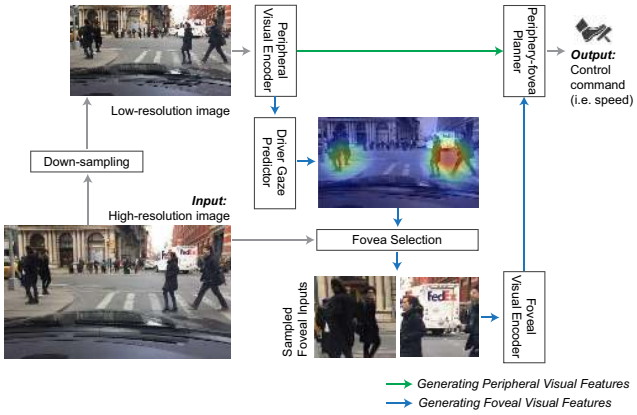

Figure 1: Our model uses the low-resolution full video frame as the peripheral visual input to predict human driver gaze and gets high-resolution image patches from the predicted gaze locations. It then combines the peripheral input and foveal input to predict the vehicle speed at high accuracy and high efficiency.

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 [25, 2] and driver gaze prediction models [25, 18, 19] 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 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 [12]. 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.

In summary, we designed a novel periphery-fovea multi-resolution driving model that efficiently processes both long-range visual structures as well as localized and detailed visual information. We also demonstrated the usefulness of guiding this driving model by human driver attention. Our model showed efficiency superiority, good interpretability, and strong robustness to critical driving situations involving pedestrians.

2. Related work

End-to-End Learning for Self-driving Vehicles. Recent successes [4, 26, 13] suggest that a driving policy can be successfully learned by neural networks with the supervision of observation (*e.g*. raw images)-action (*e.g*. steering) pairs collected from human demonstration. Bojarski *et al*. [4] trained a deep neural network to map a dashcam image to steering controls, while Xu *et al*. [26] utilized a dilated deep neural network to predict a vehicle's discretized future motions. Hecker *et al*. [11] explored an end-to-end driving model that consists of a surround-view multi-camera system, a route planner, and a CAN bus reader. Codevilla *et al*. [6] proposed a goal-conditioned end-to-end driving model to make the system responsive to high-level navigational commands (*i.e*. turn left/right, pass the intersection, follow the road). To further reduce the complexity, there is also a growing interest in end-to-mid [28] and mid-to-mid approaches [3].

Despite the encouraging successes in some specific driving scenarios, learning how to drive in general in urban areas is still challenging. We suggest two major factors that limit the performance of end-to-end driving controllers. (i) Low-resolution inputs: these neural controllers usually take in visual data in a low spatial resolution to reduce computational burdens. However, high-resolution inputs are necessary for recognizing objects at far distances and interpreting complex visual cues such as pedestrians' facial expressions. (ii) Uni-resolution and the lack of a separate decoder dedicated to salient regions: driving scenes are often highly redundant, and human drivers use low-resolution peripheral vision to determine the salient areas for high-resolution foveal vision to process. However, the neural controllers process the entire input frame in one resolution with a single decoder. This design leads to a waste of computational resources in non-salient regions and/or suboptimal performance in salient areas.

In this work, we propose an end-to-end driving model that uses periphery-foveal multi-resolution inputs. A periphery-fovea design has been shown to be effective for predicting human gazes [24]. As shown in Figure 1, our model follows the landmark model by Bojarski *et al*. [4] – a deep model that predicts control command only from (low-resolution) raw images, but, in addition, our model jointly encodes high-resolution image patches that are extracted under the supervision of our driver gaze prediction module.

Incorporating Human Visual Attention. Attention mechanisms have shown promising results in various computer vision tasks, *e.g*., image caption generation [27], visual question answering (VQA) [30], and image generation [9]. Most of these models do not supervise the generated attention by human attention. Recently, Das *et al*. [7] has shown that explicitly supervising the attention of VQA models by human attention improves the models' VQA performance. Zhang *et al*. [29] 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. Li *et al*. [17] proposed a deep model that jointly recognizing what a person is doing and estimating their gaze in First Person Vision.

However, to the best of our knowledge, incorporating human visual attention in driving tasks has not yet been explored. Human-demonstrated vehicle control, which serves as the supervision signal of the conventional end-to-end driving models, is very sparse. The sparsity of this supervision signal limits the models' semantic understanding of

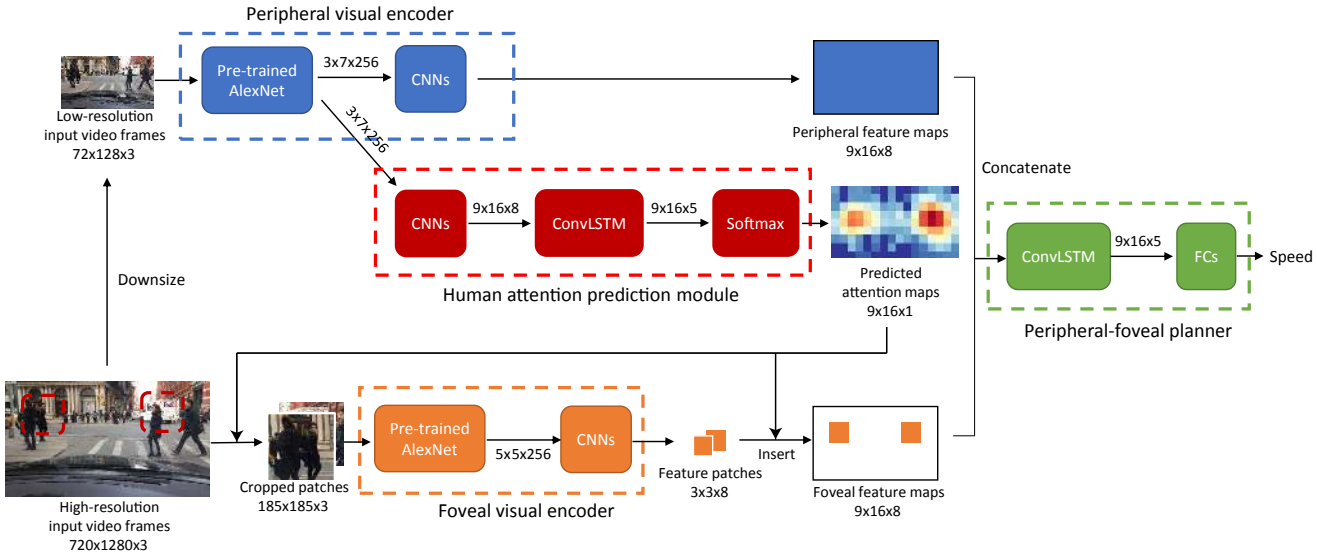

Figure 2: Our model consists of four parts: (1) the peripheral visual encoder, which extracts high-level convolutional visual features; (2) the human attention prediction module, which learns the behavior of human attention as a supervised learner over image-gaze pairs collected from humans; (3) the 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.

the scene. The visual cues that the models can appropriately react to are often limited to the simple ones that frequently lead to consistent control actions (*e.g.*, stop signs and intersections). The models often misinterpret complex and vital cues such as pedestrians, which require deep semantic understanding. In this work, we explore whether human visual attention as additional supervision signals can potentially improve the models' semantic understanding of the scenes (and provide better control prediction performance).

Predicting Driver Attention. Human driver attention has been successfully modeled with probabilistic modeling techniques [12, 10] and has been demonstrated to be critical for the cognitive and social challenges in driving (*e.g.* communicating with pedestrians) [20, 21]. Recently, deep driver attention prediction models [25, 18, 19] 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 [25, 2] collected with eye trackers. In this work, we follow the model by Xia *et al*. [25] to predict the human driver gaze.

3. Periphery-Foveal Multi-Resolution Model

Similar to conventional end-to-end driving models, our model takes dash-camera video frames as input and outputs a control signal. We chose to predict the continuous value of the speed in one second in the future. The unique design of our model is that it mimics the key aspect of the human vision system: the peripheral and the foveal systems. The peripheral vision system of our model processes the whole video frames in low resolution to efficiently captures the long-range scene structures. The foveal vision system processes salient regions in high resolution. Visual features from both peripheral and foveal systems are then combined to predict the vehicle speed.

Therefore, as we summarized in Figure 2, our model consists of four parts: (1) the *peripheral visual encoder*, which extracts high-level convolutional visual features; (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.

3.1. Peripheral Visual Encoder

The low-resolution image inputs $\mathcal{I}_L$ are fed to a convolutional network, which outputs features $\mathbf{X}_t$ that capture a high-level visual representation at each timestep t. This feature $\mathbf{X}_t \in \mathcal{R}^{w \times h \times d}$ contains a set of d-dimensional latent vectors over the spatial dimension, i.e. $\mathbf{X}_t =$

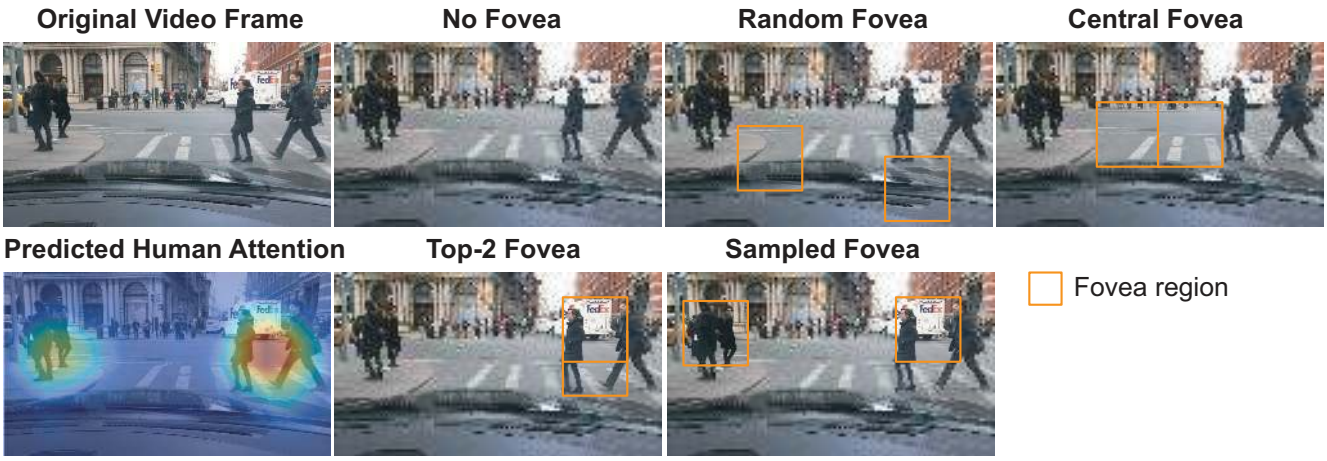

Figure 3: Examples of different approaches of foveal region selection. We present the original input video frame and the predicted human attention heat map in the left column. Our baseline model only uses peripheral vision (without fovea). We studied four different types of foveal vision selection: random, 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.

$\{\mathbf{x}_{t,1}, \mathbf{x}_{t,2}, \ldots, \mathbf{x}_{t,l}\}$, where $l = w \times h$ and $\mathbf{x}_{t,i} \in \mathcal{R}^d$. This feature $\mathbf{X}_t$ is then consumed both by the human attention prediction module (Section 3.2) and by the peripheral-foveal planner (Section 3.4).

3.2. Human Attention Prediction

Our model is trained to predict multi-focal human driver attention $\mathbf{A}_t \in \mathcal{R}^{w \times h}$ conditioned on visual features $\mathbf{X}_t$, i.e. $\alpha_t = f_{\text{attn}}(\mathbf{X}_t)$. To implement the function f_{attn}, we follow the model by Xia *et al*. [25]. We use three additional convolutional layers with 1×1 kernels and a single layer of convolutional LSTM (called *AttnLSTM*). Note that we use a dropout layer after each convolutional layers to avoid over-fitting. Our *AttnLSTM* has a kernel size of 3×3 followed by a Gaussian smoothing layer (σ is set to 1.5). A softmax layer is used to yield normalized $\{\alpha_{t,i}\}$ for $i = \{1, 2, \ldots, l\}$, where $0 \leq \alpha_{t,i} \leq 1$ and $\sum_i \alpha_{t,i} = 1$. We use the cross-entropy loss to train our human driver attention prediction model as follows:

$$\mathcal{L}_{\text{att}} = H(\alpha_t, \alpha_t^{\text{gt}}) \tag{1}$$

where H is the cross entropy function and the superscript gt indicates ground-truth.

3.3. Foveal Visual Encoder

Our model chooses k independent fovea locations for each input image. Given the predicted human attention α_t as a prior probability distribution, we sampled k fovea locations as follows:

$$p_{t,i} = \frac{\exp\left(\log \alpha_{t,i}/T\right)}{\sum_j \exp\left(\log \alpha_{t,j}/T\right)} \tag{2}$$

where $p_{t,i}$ for $i = \{1, 2, \ldots, l\}$ is the probability of the i-th spatial location being selected as the fovea location and T is the temperature factor. A temperature factor of 1 means sampling faithfully following the predicted probabilistic distribution of multi-focal gaze. A higher temperature means sampling more uniformly, while a lower temperature means sampling more from the pixel that has the highest human attention intensity.

As shown in Figure 3, we also explore three different ways to select these fovea locations as follows. (a) Random Fovea where we randomly choose k fovea locations. (b) Central Fovea where we always select k fovea locations from the center of the frame. (c) Top-k Fovea where we choose locations that have the highest attention intensities in each predicted $w \times h$-pixel human attention map.

Encoding. An image patch centered at each selected fovea location is cropped out from the 720×1280-pixel high-resolution frame image. The size of this image patch, i.e., fovea size, can potentially be dynamically determined based on clustering of the predicted attention map. However, in order to keep the efficiency of the model and to focus on the proof of the concept, we chose a fixed fovea size of 240×240 pixels. The foveal image 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 a global mean as [16] 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.

3.4. 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 $\mathbf{Y}_t = \{\mathbf{y}_{t,1}, \mathbf{y}_{t,2}, \ldots, \mathbf{y}_{t,l}\}$, where $l = w \times h$ and $\mathbf{y}_{t,i} \in \mathcal{R}^d$. This foveal feature map is initialized with zeros. Each foveal image patch is encoded into a $3 \times 3 \times d$ feature patch by the foveal feature encoder. These foveal feature patches are inserted into the foveal feature map at locations corresponding to the foveal locations. 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 ($\mathbf{X}_t$) and foveal feature maps ($\mathbf{Y}_t$) are concatenated along the semantic dimension to form the combined feature maps ($\mathbf{Z}_t = [\mathbf{X}_t; \mathbf{Y}_t]$). The combined feature maps are then processed by an additional ConvLSTM layer (called *ControlLSTM*) and four fully-connected layers to predict a continuous value for the vehicle speed.

Loss Function. We minimize the difference between human-demonstrated and predicted control (*i.e.* speed) to train our peripheral-foveal planner. Concretely, we minimize the following loss function $\mathcal{L}$ to train our model:

$$\mathcal{L} = \sum_t |v(t) - \hat{v}(t)| \tag{3}$$

where $v(t)$ and $\hat{v}(t)$ are the predicted speed and the human demonstrated speed, respectively.

4. 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.

4.1. Datasets

BDD-X. We used the Berkeley DeepDrive eXplanation (BDD-X) dataset [14] 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.*, vehicles 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.

DR(eye)VE. We also evaluated our driving models on the DR(eye)VE dataset [2] to test the generalizability of the models. The DR(eye)VE dataset contains human-demonstrated dashboard videos of driving collected in a balanced way across various urban areas (downtown, countryside and highway), different times of the day (daytime and night), and multiple weather conditions (sunny, cloudy and rainy).

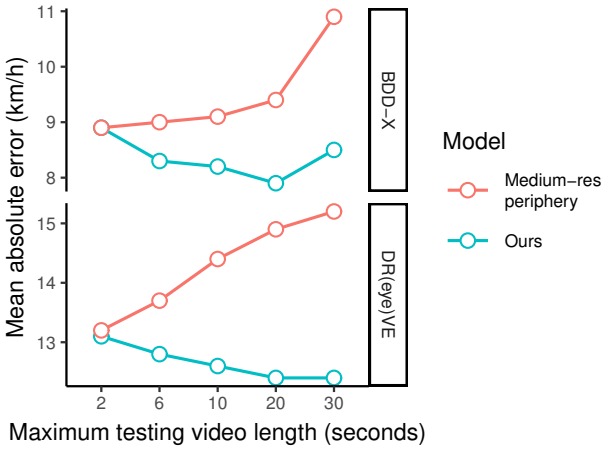

Figure 4: Testing errors of the medium-resolution periphery-only model and our model calculated using different lengths of testing videos of the BDD-X and DR(eye)VE datasets. The two models have the same amount of FLOPs at inference time, but our model consistently showed greater driving accuracy than the competing model.

BDD-A. We used the Berkeley DeepDrive Attention (BDD-A) dataset [25] 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 [25]. 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.

4.2. Training and Evaluation Details

To obtain the low-resolution image input, we sample the video frames at 10Hz and downsize them from 720×1280 pixels to 72×128 pixels by applying bilinear interpolation. Each image is then normalized by subtracting the global mean from the raw pixels [16].

Following [25], we use the identical ConvNet base (*i.e.* AlexNet) where each of the additional convolutional layers is followed by Batch Normalization and Dropout. This provides a convolutional feature map of size $3 \times 7 \times 256$. The AlexNet modules in the driving models were pre-trained

Table 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	BDD-X (km/h)			DR(eye)VE (km/h)		
	MAE ↓	RMSE ↓	Corr ↑	MAE ↓	RMSE ↓	Corr ↑
Peripheral vision only (no fovea, baseline)	9.6	14.4	.594	14.3	18.8	.752
w/ Random fovea	11.2	15.4	.520	13.8	17.8	.789
w/ Central fovea	9.4	13.9	.592	13.1	16.8	.816
w/ Human-guided fovea (ours)	**9.1**	**13.4**	**.596**	**12.6**	**16.4**	**.823**

on ImageNet and frozen afterwards. The human attention prediction module was trained following [25]. Other parts of the driving models were trained end-to-end from scratch. We used the Adam optimization algorithm [15], dropout [22] at a drop rate of 0.2, and the Xavier initialization [8]. The training of our model took approximately one day on one NVIDIA GeForce GTX 1080 GPU. Our implementation is based on Tensorflow [1].

Evaluation. To quantitatively evaluate the speed prediction performance, we used three metrics: (i) the mean absolute error (MAE), (ii) the root-mean-square error (RMSE), and (iii) the correlation coefficient (Corr). These metrics compare the prediction against the ground-truth speed commands 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 testing, unless otherwise stated, the original testing videos that were longer than 30 seconds were divided into 30-second-long segments and the remaining segments. Moreover, since human attention prediction is not the main goal of our model and one of our datasets does not contain human attention ground-truth, the evaluation of the predicted human attention maps is out of the scope of the current work.

4.3. Multi-resolution vs. Uni-resolution

We first compare 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 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 4. 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.

4.4. Effect of Incorporating Human Attention-guided Foveal Vision

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). 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 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.

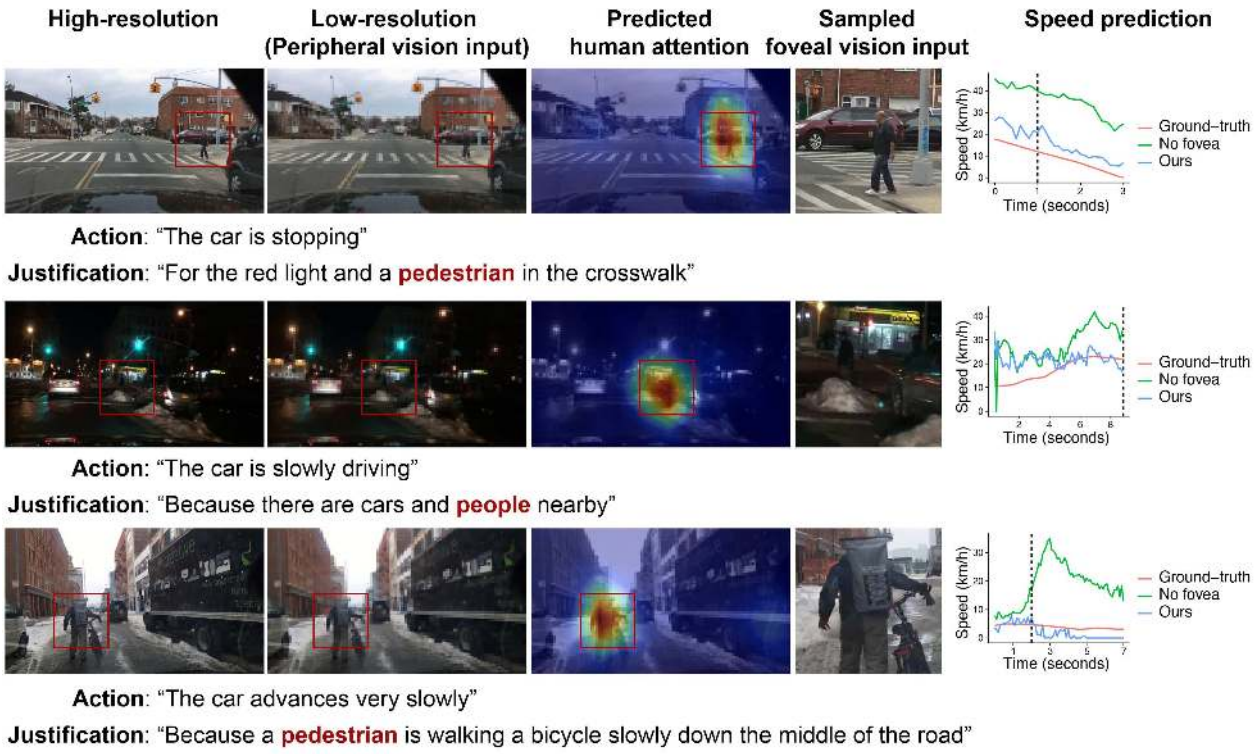

Figure 5: Examples showing how our model and the no-fovea model react in pedestrian-involved situations. From left to right: original high-resolution frame images, low-resolution frame images used as peripheral vision input, predicted human attention maps, selected high-resolution image patches as foveal vision input, and ground-truth and predicted speed curves of the pedestrian-involved events. The vertical dashed lines in the speed curve graphs indicate the moments depicted by the frame images. The ground-truth and predicted speed curves of the entire videos of these examples are in the supplementary materials. The textual action and justification human annotations are displayed below the images of each example.

Table 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	2.0	0.18	11%	8.7	12.9	.621
Sampled fovea	0.5	0.46	55%	8.6	12.7	.622
Sampled fovea	1.0	0.37	32%	**8.5**	**12.4**	**.626**

4.5. Top-2 vs. Sampled Fovea

Human attention can be multi-focus [5], especially during driving when the driver needs to react to multiple road agents or objects. A concern about using the top-2 method 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 (see Figure 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 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.

4.6. Combined vs. 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 com-

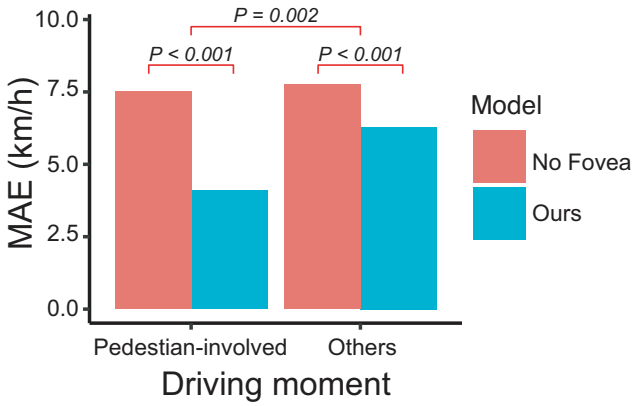

Figure 6: Testing errors of the no-fovea baseline model and our model at pedestrian-involved moments and other moments when the vehicle speed is under $10\,\mathrm{m/s}$ ($36\,\mathrm{km/h}$). Statistical significance levels given by permutation tests are noted in the graph.

Table 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**

bined 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 that corresponded to the full video frame, it adds the positional encoding [23] 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). 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.

4.7. 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 6. Our model showed significant performance gains in 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 5. The videos of those examples are in the supplementary materials.

5. 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.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th*

USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), pages 265–283, 2016.

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

[3] M. Bansal, A. Krizhevsky, and A. Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst.

[4] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, et al. End to end learning for self-driving cars. *CoRR abs/1604.07316*, 2016.

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

[6] F. Codevilla, M. Miiller, A. López, V. Koltun, and A. Dosovitskiy. End-to-end driving via conditional imitation learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–9. IEEE, 2018.

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

[8] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256, 2010.

[9] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1462–1471, 2015.

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

[11] S. Hecker, D. Dai, and L. Van Gool. End-to-end learning of driving models with surround-view cameras and route planners. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–453, 2018.

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

[13] J. Kim and J. Canny. Interpretable learning for self-driving cars by visualizing causal attention. *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[14] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–578, 2018.

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012.

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

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

[19] A. Palazzi, F. Solera, S. Calderara, S. Alletto, and R. Cucchiara. Learning where to attend like a human driver. In *2017 IEEE Intelligent Vehicles*, pages 920–925. IEEE, 2017.

[20] A. Rasouli, I. Kotseruba, and J. K. Tsotsos. Agreeing to cross: How drivers and pedestrians communicate. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 264–269. IEEE, 2017.

[21] A. Rasouli, I. Kotseruba, and J. K. Tsotsos. Understanding pedestrian behavior in complex traffic scenes. *IEEE Transactions on Intelligent Vehicles*, 3(1):61–70, 2017.

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

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[24] C. Wloka, I. Kotseruba, and J. K. Tsotsos. Active fixation control to predict saccade sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3184–3193, 2018.

[25] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney. Predicting driver attention in critical situations. In *Proceedings of the 14th Asian Conference on Computer Vision*. Springer, 2018.

[26] H. Xu, Y. Gao, F. Yu, and T. Darrell. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2174–2182, 2017.

[27] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 2048–2057, 2015.

[28] W. Zeng, W. Luo, S. Suo, A. Sadat, B. Yang, S. Casas, and R. Urtasun. End-to-end interpretable neural motion planner. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8660–8669, 2019.

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

[30] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004, 2016.