

# PERIPHERY-FOVEA MULTI-RESOLUTION DRIVING MODEL GUIDED BY HUMAN ATTENTION

PREPRINT, COMPILED MARCH 26, 2019

Ye Xia, Jinkyu Kim, John Canny, Karl Zipser, and David Whitney

University of California, Berkeley

## 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. 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.

## 1 INTRODUCTION

Vision-based deep autonomous driving models have shown promising results recently [1, 2, 3, 4]. 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 [5, 6] and driver gaze prediction models [5, 7, 8] 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

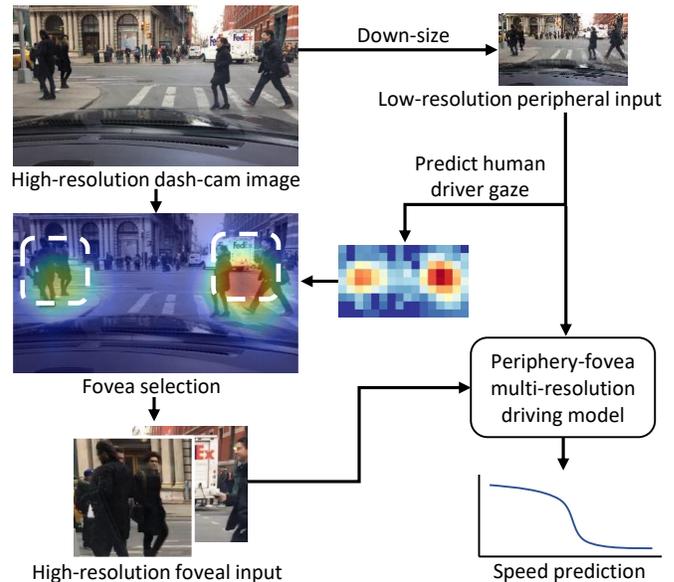


Figure 1: Our model uses the low-resolution full video frame as the peripheral visual input to predict human driver gaze and gets high-resolution image patches from the predicted gaze locations. It then combines the peripheral input and foveal input to predict the vehicle speed at high accuracy and high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

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.

## 2 RELATED WORK

**End-to-End Learning for Self-driving Vehicles.** Recent successes [3, 4] 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.* [3] trained a deep neural network to map a dashcam image to steering controls, while Xu *et al.* [4] utilized a dilated deep neural network to predict a vehicle’s discretized future motions. Hecker *et al.* [9] 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.* [1, 2] 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.* [10] 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 [11], visual question answering (VQA) [12], and image generation [13]. Most of these models do not supervise the generated attention by human attention. Recently, Das *et al.* [14] has shown that explicitly supervising the attention of VQA models by human attention improves the models’ VQA performance. Zhang *et al.* [15] has trained a network that predicts human attention for Atari 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 [5, 7, 8] 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 [5, 6] collected with eye trackers, and they use high-resolution input images ( $576 \times 1024$  or higher) to achieve optimal accuracy. How reliable the prediction would be using low-resolution input images have not been explored.

## 3 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 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 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.

### 3.1 Peripheral Visual Encoder

We sample the video frames at 10 Hz. The original frame images have a resolution of  $720 \times 1280$  pixels. We downsample them to  $72 \times 128$  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 [16].

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  $3 \times 7$  pixels and 8 channels. These feature maps are then upsampled to  $9 \times 16$  pixels for the next steps.

### 3.2 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 [5] 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  $9 \times 16$  pixels.

### 3.3 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

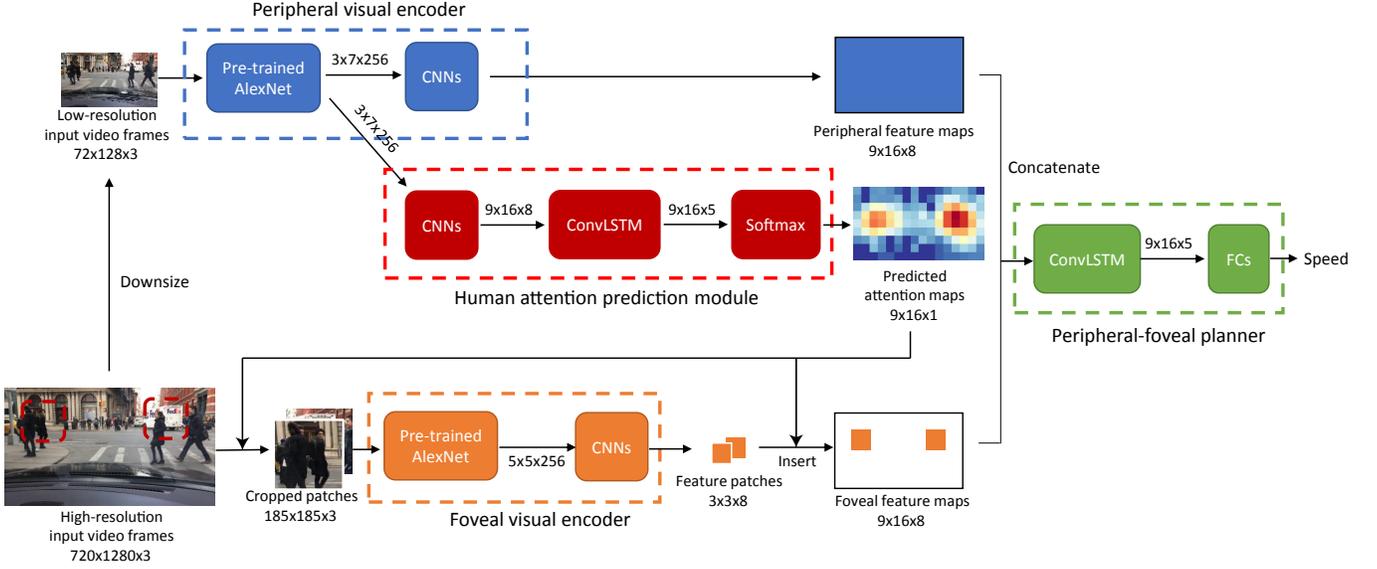


Figure 2: Our model consists of four parts: (1) the peripheral visual encoder, which extracts high-level convolutional visual features (CNN here); (2) the human attention prediction module, which learns the behavior of human attention as a supervised learner over image-gaze pairs collected from humans;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

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  $9 \times 16$ -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:

$$p_i = \frac{\exp(\log q_i/T)}{\sum_j \exp(\log q_j/T)} \quad (1)$$

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  $240 \times 240$  pixels centered at each selected fovea location is cropped out from the  $720 \times 1280$ -pixel high-resolution frame image. The images patches are then downsized to  $185 \times 185$  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 [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 that has the same size as the peripheral feature map ( $9 \times 16$  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 ( $\mathbf{x}_{i,j}^f$ ) at locations corresponding to the foveal locations:

$$\mathbf{x}_{i+h,j+w}^f = \mathbf{y}_{i,j} \quad (2)$$

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 ( $\mathbf{x}_{i,j}^p$ ) and foveal feature maps ( $\mathbf{x}_{i,j}^f$ ) are concatenated along the semantic dimension to form the combined feature maps ( $\mathbf{x}_{i,j}^c$ ).

$$\mathbf{x}_{i,j}^c = \begin{pmatrix} \mathbf{x}_{i,j}^p \\ \mathbf{x}_{i,j}^f \end{pmatrix} \quad (3)$$

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.

## 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.

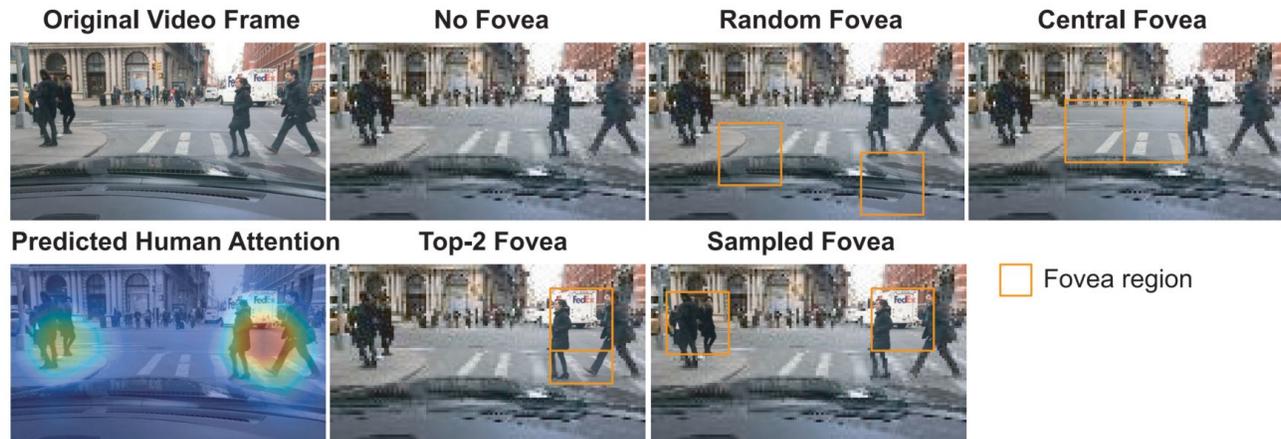


Figure 3: Examples of different approaches of foveal region selection. We present the original input video frame and the predicted human attention heat map at the left column. Our baseline model only uses peripheral vision (without fovea). We studied four different types of foveal vision selection: random,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

#### 4.1 Datasets

We used the Berkeley DeepDrive eXplanation (BDD-X) dataset [2] 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.

We used the Berkeley DeepDrive Attention (BDD-A) dataset [5] 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 [5]. 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

The AlexNet modules in the driving models were pre-trained on ImageNet and frozen afterwards. The human attention prediction module was trained following [5] except that the input image resolution was  $72 \times 128$  pixels. Other parts of the driving models were trained end-to-end from scratch. We used the Adam optimization algorithm [17], dropout [18] at a drop rate of 0.2, and the Xavier initialization [19]. The training of our model took approximately one day on one NVIDIA GeForce GTX 1080 GPU. Our implementation is based on Tensorflow [20] 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 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.

#### 4.3 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). 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  $240 \times 480$  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.

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	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)	<b>9.1</b>	<b>13.4</b>	<b>.596</b>

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	0.5	0.46	55%	8.6	12.7	.622
Sampled fovea	1	0.37	32%	<b>8.5</b>	<b>12.4</b>	<b>.626</b>
Sampled fovea	2	0.18	11%	8.7	12.9	.621

#### 4.4 Sampling according to multi-focus human attention

Human attention can be multi-focus [21], 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 (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.

#### 4.5 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 ( $9 \times 16$  in our case). This constraint determines that the feature patch

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	<b>8.5</b>	<b>12.4</b>	<b>.626</b>

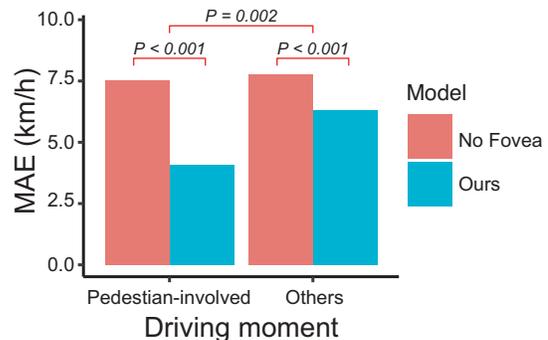


Figure 4: Testing errors of the no-fovea baseline model and our model at pedestrian-involved moments and other moments when the vehicle speed is under 10 m/s (36 km/h). Statistical significance levels given by permutation tests are noted in the graph.

corresponding to one foveal input image patch cannot be bigger than  $3 \times 3$  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  $14 \times 14$  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 [22] 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.

#### 4.6 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

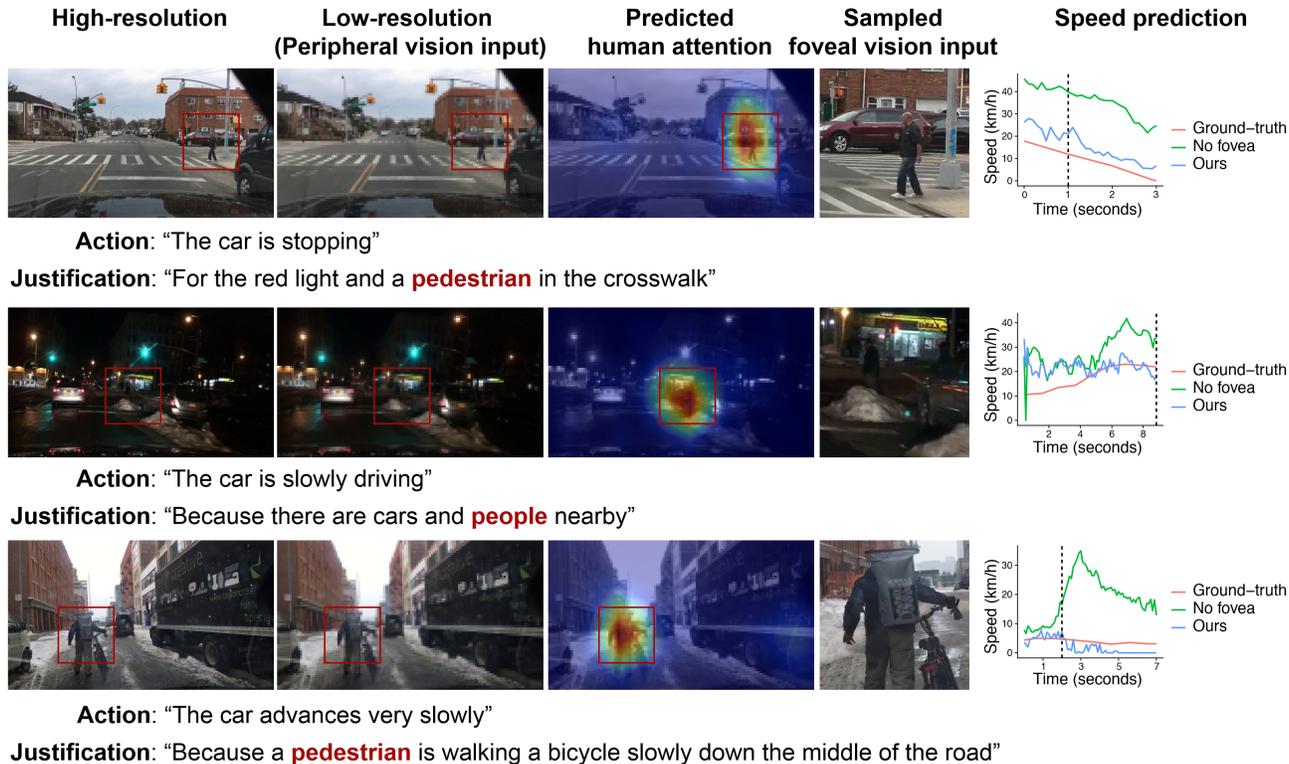


Figure 5: Examples showing how our model and the no-fovea model react in pedestrian-involved situations. From left to right: original high-resolution frame images, low-resolution frame images used as peripheral vision input, predicted human attention maps, selected high-resolution image patches as foveal vision input, and ground-truth and predicted speed curves. 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.

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 m/s (36 km/h). For a fair comparison, we excluded the frames in which the vehicle speed was higher than 10 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 4. 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.

#### 4.7 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

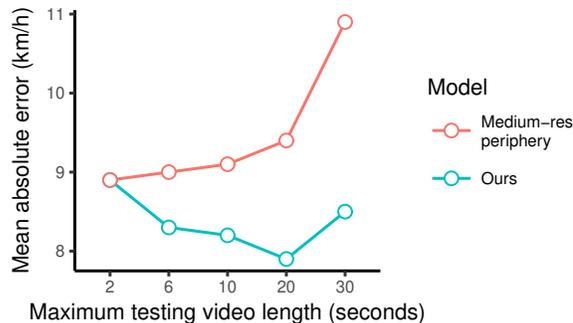


Figure 6: Testing errors of the medium-resolution periphery-only model and our model calculated using different lengths of testing videos. The two models have the same amount of FLOPs at inference time, but our model consistently showed greater driving accuracy than the competing model.

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  $209 \times 371$  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  $3 \times 3$  pixels and then upsampled them to  $9 \times 16$  pixels. The periphery encoder of the medium-resolution model output feature maps of  $12 \times 22$  pixels and then downsampled them to  $9 \times 16$  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 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.

## 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] Jinkyu Kim and John Canny. Interpretable learning for self-driving cars by visualizing causal attention. *ICCV*, 2017.
- [2] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–578, 2018.
- [3] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Praseoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *CoRR abs/1604.07316*, 2016.
- [4] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor 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*, pages 2174–2182, 2017.
- [5] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. Predicting driver attention in critical situations. In *Asian Conference on Computer Vision*. Springer, 2018.
- [6] Stefano Alletto, Andrea Palazzi, Francesco Solera, Simone Calderara, and Rita 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.
- [7] Andrea Palazzi, Davide Abati, Simone Calderara, Francesco Solera, and Rita Cucchiara. Predicting the driver’s focus of attention: the dr (eye) ve project. *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [8] Andrea Palazzi, Francesco Solera, Simone Calderara, Stefano Alletto, and Rita Cucchiara. Learning where to attend like a human driver. In *Intelligent Vehicles Symposium (IV), 2017 IEEE*, pages 920–925. IEEE, 2017.
- [9] Simon Hecker, Dengxin Dai, and Luc 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.
- [10] Dequan Wang, Coline Devin, Qi-Zhi Cai, Fisher Yu, and Trevor Darrell. Deep object centric policies for autonomous driving. *ICRA*, 2019.
- [11] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.
- [12] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016.
- [13] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1462–1471, 2015.
- [14] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163: 90–100, 2017.
- [15] Ruohan Zhang, Zhuode Liu, Luxin Zhang, Jake A Whritner, Karl S Muller, Mary M Hayhoe, and Dana H Ballard. Agil: Learning attention from human for visuomotor tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 663–679, 2018.
- [16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

- [18] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [19] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.
- [20] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael 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.
- [21] Patrick Cavanagh and George A Alvarez. Tracking multiple targets with multifocal attention. *Trends in cognitive sciences*, 9(7):349–354, 2005.
- [22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.