How Do Drivers Allocate Their Potential Attention? Driving Fixation Prediction via Convolutional Neural Networks

Tao Deng, Hongmei Yan, Long Qin, Thuyen Ngo, and B. S. Manjunath, *Fellow, IEEE*

***Abstract*—The traffic driving environment is a complex and dynamic changing scene in which drivers have to pay close attention to salient and important targets or regions for safe driving. Modeling drivers' eye movements and attention allocation in traffic driving can also help guiding unmanned intelligent vehicles. However, until now, few studies have modeled drivers' true fixations and allocations while driving. To this end, we collect an eye tracking dataset from a total of 28 experienced drivers viewing 16 traffic driving videos. Based on the multiple drivers' attention allocation dataset, we propose a convolutional-deconvolutional neural network (CDNN) to predict the drivers' eye fixations. The experimental results indicate that the proposed CDNN outperforms the state-of-the-art saliency models and predicts drivers' attentional locations more accurately. The proposed CDNN can predict the major fixation location and shows excellent detection of secondary important information or regions that cannot be ignored during driving if they exist. Compared with the present object detection models in autonomous and assisted driving systems, our human-like driving model does not detect all of the objects appearing in the driving scenes, but it provides the most relevant regions or targets, which can largely reduce the interference of irrelevant scene information.**

***Index Terms*—Fixation prediction, visual attention, eye tracking, convolutional neural networks, traffic driving.**

I. Introduction

HUMAN-CENTRIC advanced driver assistance systems (ADAS), such as collision avoidance systems, blind spot control, and lane change assistance, have significantly improved the safety and comfort of driving. Among the ADAS solutions, the most ambitious example is the monitoring system [1]–[3]. It is expected to parse the driver's attentional behaviors as well as the road scene to predict the potential unsafe maneuvers and then have the car react to avoid danger either by signaling the driver or by braking.

In fact, the traffic driving environment is a complex and dynamic changing scene in which many objective and subjective factors fuse together and govern the driver's gaze and attention automatically. These factors can be bottom-up sensory stimulus, such as a posted speed limit sign or traffic lights, and they can also be top-down aims or experiences, such as looking for a gas station or recalling a nearby restaurant. During traffic driving, drivers usually allocate their attention to the most important and salient region or target at the current second. Sometimes, there may be more than one salient region or target that drivers should focus on. For example, drivers must notice the traffic light and the roadside pedestrians when crossing a busy crossroad. **Understanding how drivers allocate their potential attention and where/what drivers mainly look at are important and challenging problems for driving assistance systems.**

Traffic saliency detection, which computes the important and salient regions or objects that drivers should care about in a given driving environment, is a hot topic in intelligent vehicle systems. Many algorithms and models have been proposed to predict the traffic saliency or drivers' attention [4]. Some researchers utilized driver monitoring systems to estimate drivers' gaze direction or fixation region from head pose and eye location cues [5]–[7]. Bremond et al. [8] presented a visual saliency model based on a nonlinear support vector machine (SVM) classifier for the detection of traffic signs. Pugeault et al. analyzed drivers' pre-attention at T junctions [9]. The authors studied the looked-but-failed-to-see effect by analyzing the object saliency. There are some other studies focusing on drivers' head orientations by detecting facial landmarks [5], [10]–[12].

However, these studies lack the prediction of the drivers' true fixation during the driving task. Our previous studies [13], [14] analyzed the eye tracking data of 20 experienced drivers when viewing traffic images and then proposed a bottom-up and top-down combined saliency detection model via the random forest learning method to predict drivers' direct attentional area [15]. However, the work was based on static images, which was not suited for the prediction of a complex dynamic traffic video stream. In the field of computer vision, there are some natural saliency image/video datasets and saliency models, such as the MIT benchmark [16], the SLICON dataset [17], and Action in the Eye [18], but they do not aim at specific driving scenes. Recently, Alletto et al. [19] recorded one driver's eye tracking video during actual driving and built a publicly available video dataset (DR(eye)VE). The distribution of eye tracking data depended on the characteristics of the driver (e.g., driving proficiency level or culture [20]). Palazzi et al. [21] proposed an attention prediction model based on the DR(eye)VE dataset using the deep learning method. Tawari and Kang [22] proposed a Bayesian framework to model the visual attention of a human driver and developed a fully convolutional neural network to detect the salient region based on the DR(eye)VE dataset.

Although DR(eye)VE is a good public dataset that consists of 74 videos and eight drivers' eye tracking data while real driving, the data collection scheme determines only one driver's attention to be recorded on each video. Therefore, the models based on DR(eye)VE can predict only one salient region, and they cannot predict drivers' endogenous attentional allocation when two or more salient targets or regions must be focused on, as mentioned above. In order to solve the problem, we did the following works:

- We built a traffic driving video dataset based on an eye movement experiment that recorded 28 experienced drivers' eye tracking data.

Manuscript received October 18, 2018; revised January 28, 2019 and March 21, 2019; accepted April 30, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61773094, Grant 61573080, and Grant 61503059, in part by the Fundamental Research Funds for the Central Universities under Grant 2682019CX59, in part by the Key program for International S&T Cooperation of Sichuan Province under Grant 2019YFH0097, and in part by the Science and Technology Project of Sichuan Province under Grant 2018GZ0008. The Associate Editor for this paper was J. M. Alvarez. *(Corresponding author: Hongmei Yan.)*

T. Deng is with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China (e-mail: tinydao@163.com).

H. Yan and L. Qin are with the MOE Key Lab for Neuroinformation, University of Electronic Science and Technology of China, Chengdu 610054, China (e-mail: hmyan@uestc.edu.cn; qin476290209@gmail.com).

T. Ngo and B. S. Manjunath are with the Department of Electrical and Computer Engineering, University of California Santa Barbara, Santa Barbara, CA 93106-9560 USA (e-mail: thuyen@ece.ucsb.edu; manj@ece.ucsb.edu).

Digital Object Identifier 10.1109/TITS.2019.2915540

- Based on the dataset, we proposed a traffic video saliency detection model with compact convolutional-deconvolutional neural networks (CDNN) to predict the drivers' fixation location. The CDNN network was trained by multiple drivers' eye tracking data and contained bottom-up and top-down information related to traffic driving.
- Finally, we compared our model with other methods. The experimental results demonstrated that our model can predict the drivers' fixational areas more accurately.

Moreover, by taking advantage of multiple drivers' attention experiences, our model can predict the drivers' potential attention allocation, including the main target or region and also the secondary one if it exists. Compared with the state-of-the-art object detection models in autonomous and assisted driving systems, such as the Faster Region-CNN (RCNN), Mask RCNN and YOLO, our model did not detect all of the objects appearing in the driving scenes. Rather, it provided the most relevant regions or targets, which can largely reduce the interference of irrelevant scene information. We made the dataset and source code of our method publicly available.[1]

Fig. 1. Video samples recorded by driving recorders. Each video is approximately 52 to 181 seconds, and its resolution ratio is 1280×720 pixels.

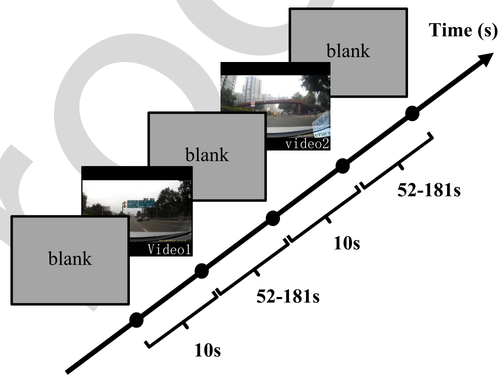


Fig. 2. Flow chart of the eye movement experiment.

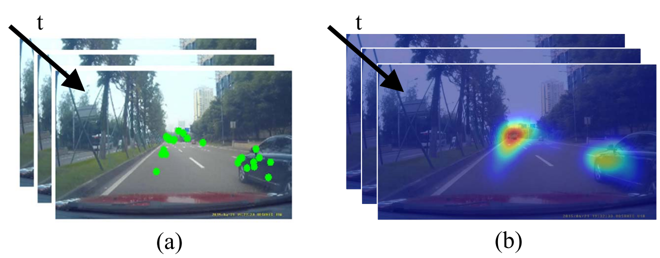


Fig. 3. Example of eye tracking data and the corresponding fixation saliency map. (a) Fixation points when 28 drivers view the videos. (b) eye tracking data placed with a 2-D Gaussian distribution.

II. Eye Tracking Data

In this section, an eye movement experiment was designed to collect drivers' eye tracking data while viewing the driving videos.

A. Participants

Twenty-eight participants took part in the eye movement experiment, including 12 females and 16 males that ranged from 23 to 43 years old (M=32.0; SD=6.4). The participants were required to be drivers who had at least 2 years driving experience and drove a car frequently. As a result, their driving experience ranges from 2 to 16 years (M=5.7; SD=3.8). All participants had normal or corrected-to-normal vision and were provided with written informed consent prior to participation. The experimental paradigms were approved by the Ethics and Human Participants in Research Committee at the University of Electronic Sciences and Technology of China in Chengdu, China.

B. Stimuli and Apparatus

The visual material consisted of 16 traffic driving videos, as illustrated as Fig. 1. Each traffic video was collected by a driving recorder while the cars were running on an urban road. The videos lasted from 52 to 181 seconds (M=161.4; SD=38.0), had a resolution of 1280×720 pixels (34.2×19.2 squared degrees of visual angle), and had a frame rate of 30 frames per second. Participants were seated 57 cm away from a 21-inch CRT monitor with a spatial resolution of 1280×1024 pixels and a refresh rate of 75 Hz. The head was stabilized with a chin and forehead rest. A steering wheel is placed in front of the participants who were asked to view the videos by assuming that they were driving a car. Eye movements were recorded using an eye-tracker (Eyelink 2000, SR Research, Eyelink, Ottawa, Canada) with a sampling rate of 1000 Hz and a nominal spatial resolution of 0.01 degree of visual angle.

C. Procedure

Before each participant viewed the stimuli videos, a calibration was run to ensure the accuracy of the eye tracking data. The calibration was repeated if the quality of eye tracking was not satisfactory. Each participant was asked to 'task-view' the 16 different traffic driving videos. The 'task-view' denoted that participants should view these stimuli videos under a hypothetical driving attentional condition. Each participant performed 8 blocks, and each block consisted of 2 trials. Each block cost approximately 6 minutes (calibration excluded) with a 2 minute break between blocks. Overall, it took approximately 1 hour for a participant to complete the whole experiment. The video sequences were shown to each subject in a random order, as illustrated in Fig. 2.

D. Eye-Movement Analysis

The subjects' eye fixations were recorded to construct the human saliency map. In the eye tracking dataset, there were 28 drivers' fixation points that were recorded per video frame (Fig. 3(a)). The drivers' eye tracking data fitted with the 2-D Gaussian distribution (Fig. 3(b)) were used as the ground truth in our work. By taking

[1]https://github.com/taodeng/CDNN-traffic-saliency

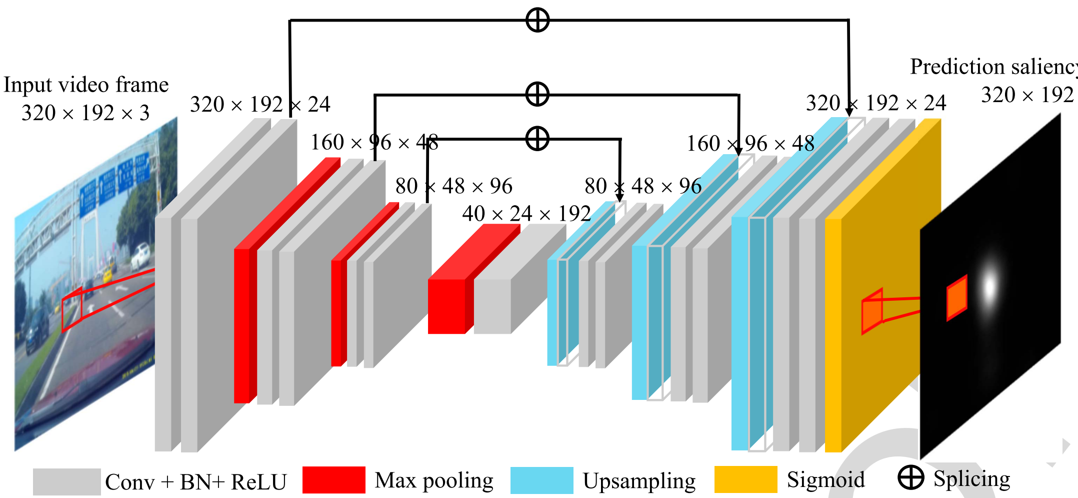


Fig. 4. CDNN architecture in our work.

advantage of multiple drivers' attention experiences, the dataset included both the primary salient region and the secondary one if it existed. Figure 3 illustrates an example of two salient regions that drivers may pay attention to under certain situations.

III. Fixation Prediction Based on a Convolutional Neural Network

A. Convolutional-Deconvolutional Neural Network

The choice of the architecture is very important when utilizing a neural network framework. In this paper, we propose a convolutional-deconvolutional neural network (CDNN) inspired by U-Net [23] to predict the drivers' fixation locations in traffic scenes. The CDNN architecture is shown in Fig. 4.

The CDNN consists of a contracting path (convolution) and an expansive path (deconvolution). The contracting path follows the typical architecture of a convolutional network. This path consists of the repeated application of two 3×3 convolutions, each followed by a rectified linear unit (ReLU), batch normalization (BN) and a 2×2 max pooling operation with a stride of 2 for downsampling. Every step in the expansive path consists of an upsampling of the feature map followed by a 2×2 convolution (deconvolution) that halves the number of feature channels, a concatenation with the corresponding feature map from the contracting path, and two 3×3 convolutions, each followed by a ReLU and BN. Table I shows more details of the convolutional-deconvolutional network.

Although the architecture of CDNN is similar with U-Net, there are some differences between them. The aim of the proposed CDNN is to predict the drivers' fixation and attention allocation, so calculation complexity and speed are important considerations for potential application. A shallow convolutional network layer is chosen in our work, so the parameters of CDNN are less than U-Net. We set the padding parameter as 0 at the maxpooling layer so that the model can make full use of the edge information for saliency detection. Besides, each convolution and deconvolution layer include a batch normalization operation that allows each layer to learn independently by itself and reduce the overfitting.

B. Loss Function

We choose the binary cross entropy during the training phase. The loss function $L(S, \hat{S})$ is defined between the predicted saliency map $\hat{S}$ and its corresponding ground truth fixation saliency map S.

$$L_{BCE}(S, \hat{S}) = -\frac{1}{N}\sum_{i=1}^{N} S_i \log(\hat{S}_i) + (1 - S_i)\log(1 - \hat{S}_i) \qquad (1)$$

where S_i denotes the i^{th} pixel of the fixation saliency map S, $\hat{S}_i$ is the i^{th} pixel of predicted saliency map $\hat{S}$, N is the total number of pixels.

TABLE I
The Detailed Parameters of the Convolutional-Deconvolutional Neural Network

Layer	Depth	Kernel	Stride	Pad	Activation
Conv2d 1_1	24	3×3	1	1	ReLU
Conv2d 1_2	24	3×3	1	1	ReLU
Max pooling		2×2	2	0	
Conv2d 2_1	48	3×3	1	1	ReLU
Conv2d 2_2	48	3×3	1	1	ReLU
Max pooling		2×2	2	0	
Conv2d 3_1	96	3×3	1	1	ReLU
Conv2d 3_2	96	3×3	1	1	ReLU
Max pooling		2×2	2	0	
Conv2d 4_1	192	3×3	1	1	ReLU
Conv2d 4_2	192	3×3	1	1	ReLU
Upsampling		2×2	2	0	
Conv2du 3_1	96	3×3	1	1	ReLU
Conv2du 3_2	96	3×3	1	1	ReLU
Upsampling		2×2	2	0	
Conv2du 2_1	48	3×3	1	1	ReLU
Conv2du 2_2	48	3×3	1	1	ReLU
Upsampling		2×2	2	0	
Conv2du 1_1	24	3×3	1	1	ReLU
Conv2du 1_2	24	3×3	1	1	ReLU
Conv2du 0	1	3×3	1	1	
FC					Sigmoid

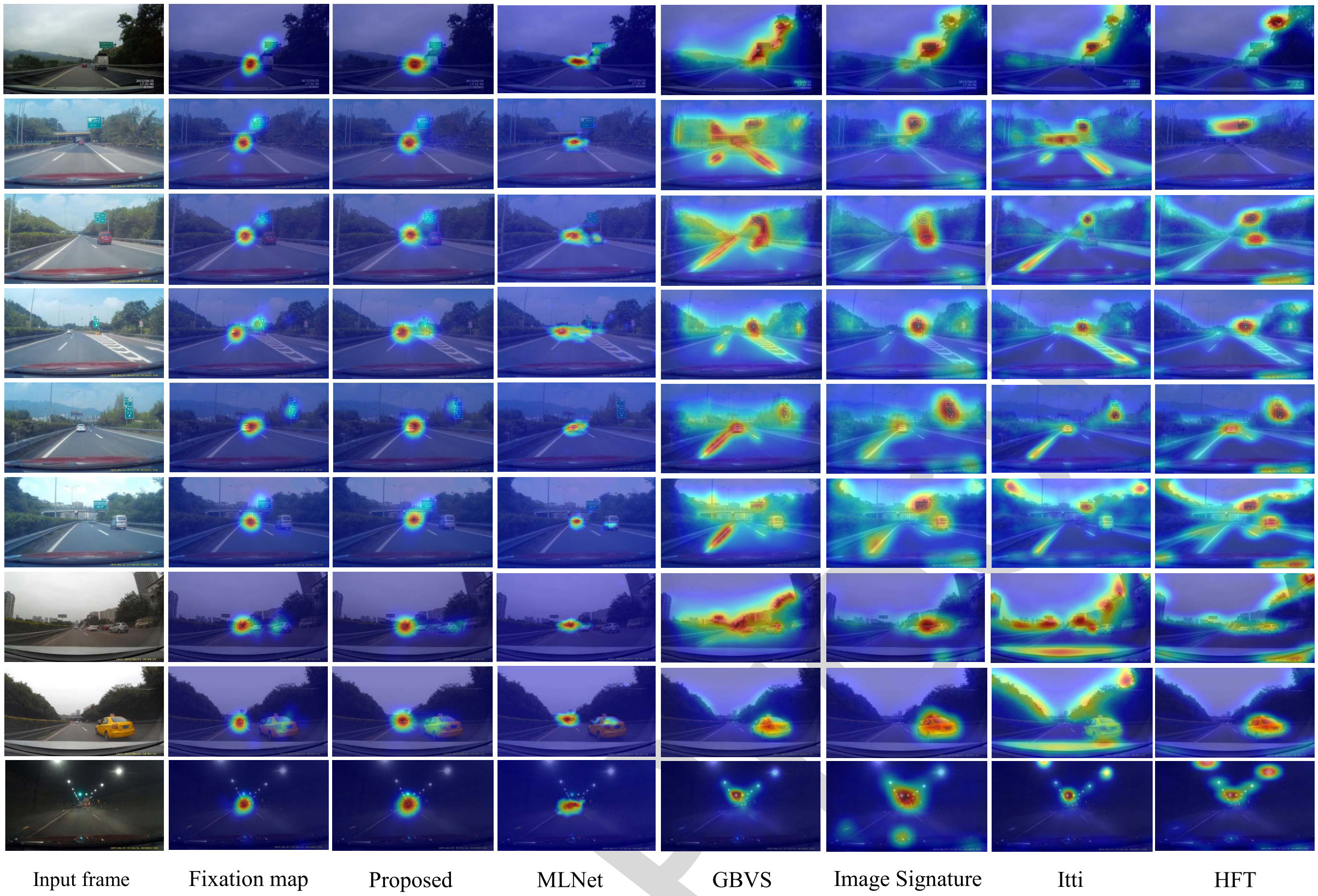


Fig. 5. Qualitative assessment of our proposed model and the classical state-of-the-art methods. From left to right: the input frames, the ground truth fixation maps, our predicted saliency maps, and the predictions of MLNet [25], GBVS [26], Image Signature [27], Itti [28] and HFT [29].

IV. Results

In this section, we first describe the preparation of the drivers' eye tracking dataset. The training and testing datasets are composed of these video frames and the corresponding saliency maps. Then, we train the proposed CDNN with the training set and evaluated the performance of the model with the testing set both qualitatively and quantitatively.

A. Dataset

In our experiments, the dataset is divided into three subsets. Ten videos are used as the training set, 2 videos are used as the validating set and 4 videos were used as the testing set. All of these videos are untrimmed videos, but the first five frames and last five frames are deleted to ensure the accuracy of the eye tracking recording. There are 49035 frames in the training phase and 6655 frames in the validating phase. A total of 19135 frames are used to test the performance of the prediction model.

All of these training frames are randomly input into the model during the training phase. The Adam optimizer with the parameters as suggested in the original paper [24] is applied in this work. The learning rate is set to 10^{-3}, with the momentum and weight decay valued as 0.9 and 10^{-4}, respectively. To reduce the training time, the video frames are resized to 320×192. The model is trained using a GPU server consisting of four NVIDIA TITAN Xp 12 GB GPUs and two Intel Xeon E5-2673 v3 CPUs. The CDNN implementation is based on PyTorch.

TABLE II

Indicators of Similarity and Dissimilarity Between the Prediction and the Ground Truth

Metrics	Location-based	Distribution-based
Similarity	AUC-Borji, AUC-Judd, NSS, IG	CC, SIM
Dissimilarity	-	EMD, KL-Div

B. Qualitative Evaluation

Figure 5 presents a visual comparison of our model and some state-of-the-art saliency models, i.e., Multi-Level Net (MLNet) [25], Graph-based Visual Saliency (GBVS) [26], Image Signature [27], Itti [28], and Hypercomplex Fourier Transform (HFT) [29]). The predicted saliency maps are overlaid with original traffic images for better viewing. The results show that our prediction model can predict the drivers' fixational areas more accurately than the classical saliency models can. In Fig. 5, we can see that the state-of-the-art saliency models show excellent prediction of traffic lights, traffic signs, cars and some road lanes in traffic scenes. However, the models cannot detect the most important top-down information about driving. They match poorly against human eye tracking data. That is, the models cannot predict drivers' attention allocation precisely. By contrast, our model can detect both the driving related bottom-up information (e.g., traffic signs and nearby cars) and the important top-down information (i.e., the right front of the driving road). Namely, our model can predict the drivers' potential attention allocation accurately for both the main target or region and also the secondary one if it exists, which is consistent with the drivers' driving experience. Please note that the last row in Fig. 5 is the result tested with tunnel scene. Our model shows a robust performance in a dark and faint tunnel environment, which indicates that our model can predict drivers' fixation areas, even in severe scenes such as tunnels and night.

Especially, a deep learning-based saliency model MLNet is also compared in this section. The MLNet is re-trained on our dataset. The fourth column of Fig. 5 shows some prediction results by MLNet. We can see that MLNet outperforms all the other bottom-up saliency models. However, it still cannot precisely predict all the drivers' fixation regions, for example, it does not detect the location of traffic sign in the third, fifth and sixth rows.

C. Quantitative Evaluation Metrics

To quantitatively compare the performance of our model with state-of-the-art saliency models, we employ two categories of saliency evaluation metrics: location-based and distribution-based [16], [30], [31]. Location-based metrics include the area under the ROC curve (AUC-Borji [32] and AUC-Judd [33], [34]), the normalized scanpath saliency (NSS [35]) and Information Gain (IG [36]), which indicate the similarity between the prediction and the ground truth. Distribution-based metrics include Pearson's correlation coefficient (CC [37]), Kullback-Leibler divergence (KL-Div [16]), the Earth mover's distance (EMD [38]), and Similarity (SIM [34]). CC and SIM are indicators of similarity, while EMD and KL-Div are indicators of dissimilarity between the prediction and the ground truth (Tab. II). Different metrics use different formats of the ground truth for evaluating saliency models. Location-based metrics consider the saliency map values at discrete fixation locations, while the distribution-based metrics treat the ground truth as continuous distributions. In other words, location-based metrics use the fixation point map (Fig. 3(a)) as the ground truth and distribution-based use the fixation saliency map (Fig. 3(b)) as the ground truth. In the following, the saliency evaluation metrics are introduced briefly.

1) Area Under ROC Curve (AUC): Evaluating Saliency as a Classifier of Fixations: In [32], Borji et al. proposed a variant of the AUC called the AUC-Borji. It uses a uniform random sample of image pixels as negatives and defines the saliency map values of pixels that are above a threshold as false positives. Judd et al. [33], [34] proposed a variant of the AUC called the AUC-Judd consisting of the true positive rate (TP rate) and the false positive rate (FP rate).

2) Normalized Scanpath Saliency (NSS): Measuring the Normalized Saliency at Fixations: The NSS metric quantifies the saliency map values at the eye fixation locations and computes the average normalized saliency at all fixations as follows:

$$NSS = \frac{1}{N} * \sum_{i=1} \frac{\hat{S}(x_i, y_i) - \mu_{\hat{S}}}{\sigma_{\hat{S}}} \tag{2}$$

where (x_i, y_i) is the location of one fixation point, $\mu_{\hat{S}}$ and $\sigma_{\hat{S}}$ are the mean and standard deviation of the prediction saliency map $\hat{S}$, respectively. $NSS = 1$ indicates that the subject's eye position falls within a region where the predicted density is one standard deviation above the average, while $NSS = 0$ means that the model performs at a chance level [13], [39].

3) Information Gain (IG): Evaluating Information Gain Over a Baseline: Information gain metric [36] is an information theoretic method that measures saliency model performance beyond systematic bias. Given a binary map of fixations S_B, a saliency map $\hat{S}$, and a baseline map B, information gain is computed as:

$$IG(\hat{S}, S_B) = \frac{1}{N} \sum_{i}^{N} S_B[\log_2(\varepsilon + \hat{S}_i) - \log_2(\varepsilon + B_i)] \tag{3}$$

where i indexes the i^{th} pixel, N is the total number of fixated pixels, ε is for regularization constant ($\varepsilon = 2.2204\text{e-}16$ in MATLAB), and information gain is measured in bits per fixation. In this work, the center bias saliency map is regarded as baseline map B. This metric measures the average information gain of the saliency map over the center prior baseline at fixated locations (i.e., where $S_B = 1$). A score above zero indicates the saliency map predicts the fixated locations better than the center prior baseline.

4) Pearson's Correlation Coefficient (CC): Evaluating the Linear Relationship Between Distributions: Pearson's correlation coefficient, also called the linear correlation coefficient, is a statistical method used generally for measuring how correlative or dependent two variables are. The linear CC output ranges is between -1 and 1 and is calculated as follows:

$$CC = \frac{\operatorname{cov}(\hat{S}, S)}{\sigma_{\hat{S}} * \sigma_S} \tag{4}$$

where S is the fixation saliency map, and σ is the standard deviation. It means that the maps are correlated when the correlation value is close to -1 and 1. A score of 0 indicates that the maps are completely uncorrelated.

5) Similarity (SIM): Measuring the Intersection Between Distributions: The similarity metric [34] also uses the normalized probability distributions of the predicted saliency map $\hat{S}$ and human fixation saliency map S as follows:

$$SIM = \sum_{i=1}^{N} \min(S(i), \hat{S}(i)) \tag{5}$$

where

$$\sum_{i}^{N} S(i) = \sum_{i}^{N} \hat{S}(i) = 1 \tag{6}$$

$SIM = 1$ indicates the distributions are the same, while $SIM = 0$ indicates no overlap.

6) Kullback-Leibler Divergence (KL-Div): Evaluating Saliency With a Probabilistic Interpretation: The Kullback-Leibler divergence is a general information theoretic measure of the difference between two probability distributions. It is calculated as follows:

$$KL_{div} = \sum_{i=1}^{N} S(i) * \log\left(\frac{S(i)}{\hat{S}(i) + \varepsilon} + \varepsilon\right) \tag{7}$$

where N is the number of pixels and ε is a regularization constant ($\varepsilon = 2.2204\text{e-}16$ in MATLAB) that is used to avoid the log and division by zero. The S and $\hat{S}$ distributions are both normalized as follows:

$$Norm(i) = \frac{Norm(i)}{\sum_{i=1}^{N} Norm(i) + \varepsilon}, Norm = \{S, \hat{S}\} \tag{8}$$

The $KL_{div} = 0$ indicates that the two maps are strictly equal [30].

7) Earth Mover's Distance (EMD): Incorporating Spatial Distance Into Evaluation: The Earth mover's distance metric is a measure of the distance between two probability distributions over a region.

$$EMD = \min_{\{f_{ij}\}} \sum_{i,j} f_{ij} d_{ij} + \left| \sum_{i} S_i - \sum_{j} \hat{S}_j \right| \max_{i,j} d_{ij} \tag{9}$$

$$s.t. f_{ij} \geq 0, \sum_{j} f_{ij} \leq S_i, \sum_{i} f_{ij} \leq \hat{S}_j, \tag{10}$$

TABLE III

PERFORMANCE COMPARISON OF OUR MODEL WITH THE STATE-OF-THE-ART SALIENCY MODELS USING MULTIPLE EVALUATION METRICS. DIFFERENT TYPE OF GROUND TRUTH IS USED FOR VARIOUS METRICS

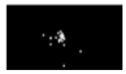
Ground truth	Fixation point map 				Fixation saliency map 			
Models	*AUC-Borji* ↑	*AUC-Judd* ↑	*NSS* ↑	*IG* ↑	*CC* ↑	*SIM* ↑	*KLD* ↓	*EMD* ↓
Human	0.9578	0.9863	6.4827	2.1544	1.0	1.0	0	0
ITTI	0.7023	0.7256	0.8627	-2.0573	0.1668	0.1736	2.1418	2.2353
Image Signature	0.8298	0.8526	1.6486	-1.5032	0.3148	0.2102	2.0430	2.2408
GBVS	0.8942	0.9076	1.8363	-1.1009	0.3665	0.5223	1.7484	1.7200
HFT	0.7015	0.7329	0.9729	-2.2359	0.1750	0.1687	2.5579	2.3961
MLNet	0.8734	0.8957	5.6942	0.1869	0.8666	0.4516	0.8709	2.9803
Proposed	**0.9261**	**0.9745**	**5.8288**	**1.4945**	**0.9451**	**0.7779**	**0.2897**	**0.3416**

TABLE IV

PERFORMANCE COMPARISON OF THE RESIDUAL AND THE DEFORMED RESIDUAL UNIT WITH OUR PROPOSED ORIGINAL CDNN MODEL

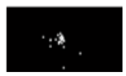
Ground truth	Fixation point map 				Fixation saliency map 			
Models	*AUC-Borji* ↑	*AUC-Judd* ↑	*NSS* ↑	*IG* ↑	*CC* ↑	*SIM* ↑	*KLD* ↓	*EMD* ↓
Human	0.9578	0.9863	6.4827	2.1544	1.0	1.0	0	0
CDNN	0.9261	0.9745	5.8288	1.4945	**0.9451**	**0.7779**	**0.2897**	**0.3416**
Basic Residual	**0.9274**	0.9747	5.9374	1.4589	0.9344	0.7150	0.3577	1.1323
Deformed Residual	0.9253	**0.9751**	**5.9927**	**1.5070**	0.9358	0.7550	0.3151	0.9676

TABLE V

TRAINING COST COMPARISON OF THE DIFFERENT NETWORK STRUCTURES. CDNN IS MUCH FASTER THAN OTHERS

	Train cost
CDNN	**19h**
Basic Residual	66h
Deformed Residual	78h

and

$$\sum_{i,j} f_{ij} = \min(\sum_i S_i - \sum_j \hat{S}_j) \tag{11}$$

where each f_{ij} represents the amount transported from the i^{th} supply to the j^{th} demand. d^{th} is the ground distance between the i^{th} and j^{th} points in the distribution. Starting from zero, a larger EMD indicates a larger overall difference between the two distributions.

Table III shows the quantitative performance of our proposed model compared with other state-of-the-art saliency models [25]–[29] using the aforementioned evaluation metrics. The first row Human represents the fixation saliency map of drivers (Fig. 3(b)). As expected, our proposed model (last row of Table III) shows the highest similarity and lowest dissimilarity with the ground truth. We can draw a conclusion that the proposed CDNN architecture can predict human's fixation area more precisely than other models can.

D. Performance Comparison With Residual and Deformed Residual Networks

He et al. [40] proposed a deep residual learning named ResNet (Fig. 6(a)). Moreover, they improved the ResNet (Fig. 6(b)) [41]. The

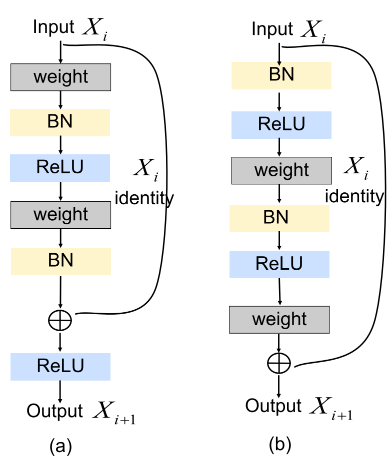


Fig. 6. (a) Basic residual unit. (b) Deformed residual unit.

authors analyzed the propagation formulations behind the residual building blocks in the revised ResNet, which suggested that the forward and backward signals could be directly propagated from one block to any other block. Here, their methods are also applied in our CDNN model at each convolutional phase.

We compare these results with those of our original model in Table IV. The results show that the prediction of the deformed ResNet is slightly better than that of the basic ResNet, which is consistent with the authors' results [41]. The AUC, NSS and IG evaluation metrics of our CDNN model are slightly smaller than those of ResNet. However, the CC, KL-Div, EMD and SIM of our CDNN model are the best. More importantly, Table V shows that the original CDNN costs only 19 hours to train the frames, which is much faster than ResNet is.

Fig. 7. Our trained model tests on the DR(eye)VE dataset. (a) The input frames that the camera recorded. (b) The driver's fixation maps in DR(eye)VE. (c) The salient regions predicted by our model. The red circles label the location of a biker.

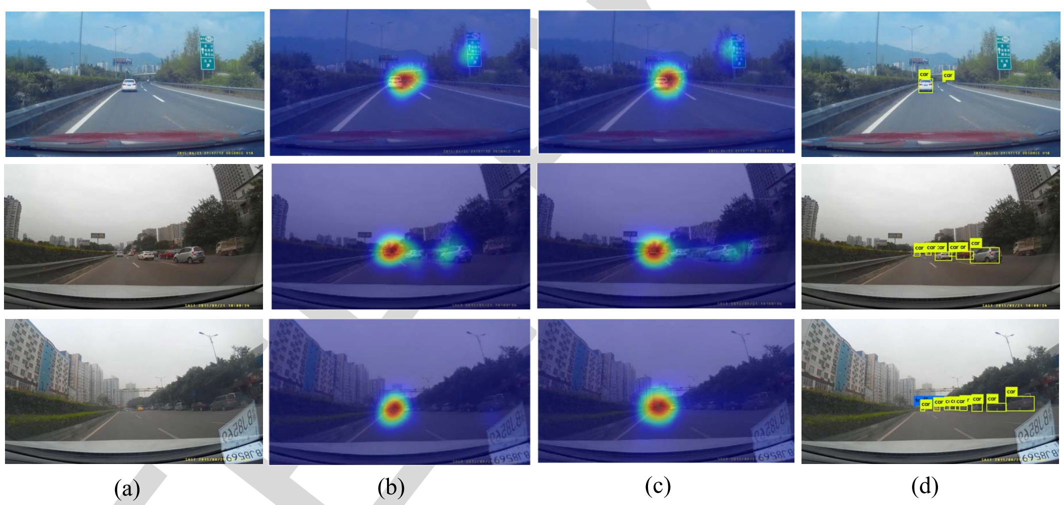


Fig. 8. Comparison with YOLOv3 object detection model. (a) The input frames that the camera recorded in our dataset. (b) The driver's fixation maps. (c) The salient regions predicted by our model. (d) The objects detected by the YOLOv3 model. The yellow rectangles mark the location of the detected cars and label the object category.

V. Discussion

Palazzi et al. [19], [21] recently built the DR(eye)VE dataset that consists of 74 videos, and each video lasts 5 minutes. The dataset provides videos both from a roof-mounted camera and a head mounted camera. The dataset comes from eight drivers' eye tracking data while driving. However, there is only one driver's eye tracking data on each video.

We use our trained model to test the DR(eye)VE videos. Fig. 7 shows some results of our prediction using the DR(eye)VE dataset. Fig. 7(a) shows the input frames that the camera recorded. Fig. 7(b) shows the driver's fixation maps of DR(eye)VE, and Fig. 7(c) gives the predicted salient regions using our model. Most of our predictions are consistent with the driver's fixation region (the first row of Fig. 7), but some are not. For example, in the second and third rows of Fig. 7, we notice that there is a biker who is preparing to go across the street (red circle shown in the figures). We think that the biker is also an important factor that the driver should consider for driving safety. Since only one driver's eye tracking data are recorded, DR(eye)VE can give only one salient region, which is the most important area that the driver is currently focusing on (right in front of the road), as shown as Fig. 7(b). The location of the biker is ignored in the DR(eye)VE eye tracking dataset. However, we find that our model can predict both the most important driving information (right in front of the road) and also the second most important information (the biker), as shown in Fig. 7(c). This is because our model is trained with multiple drivers' eye tracking data, and thus, our model can detect more driving-related information, including bottom-up and top-down attention.

Although the DR(eye)VE dataset is composed by a real driving eye-movement experiment, the data collection scheme determines only single driver's attention to be recorded on each video, so it cannot indicate multiple salient regions for the traffic driving scenes. By comparison, our dataset is constructed of 28 drivers' attention, which may cover more key information related with driving safety. By taking advantage of the multiple drivers' attention dataset, our model can predict the drivers' potential attention allocation for both the main target or region and also the secondary/tertiary ones if they exist (the second and third rows of Fig. 7(c)).

Currently, there are some state-of-the-art object detection models such as the Faster RCNN [42], Mask RCNN [43], and YOLO [44]–[46] that can detect all objects that appear in traffic scenes precisely and in real time. The detected objects include cars, bikers, traffic signs/lights, roads, pedestrians, and the sky. Some image segmentation methods have been used in commercial intelligent driving vehicles. All of the objects and areas that appear in the environment can be detected and recognized. However, we think that not all objects are critical and helpful for driving, for example static cars parking on the wayside, distant cars, pedestrians walking on the sidewalk, some irrelevant advertising signs and trees. We consider that these static or irrelevant objects could be the redundant information for driving. The objects may even interfere with the judgment and control of safe driving if an assistant system provides too many redundant objects.

In Fig. 8, we compare our model with YOLOv3 [46], which is a state-of-the-art real-time object detection model trained on the COCO dataset, using our driving videos. Because YOLOv3 is one deep learning method that is dependent on the dataset, the first row of Fig. 8 shows that YOLOv3 cannot detect the traffic sign on the right roadside where drivers look in our dataset. In the second and third rows of Fig. 8, we can see that YOLOv3 can accurately detect all of the cars appearing in the traffic scenes. Actually, drivers do not gaze at all of the cars, but they allocate some attention to some key objects, such as crossing cars and related traffic signs. However, although the state-of-the-art object detection models can precisely discover and recognize all of the objects in driving scenes, some irrelevant objects are redundant information for drivers. These redundant detection results may interfere with the control of the intelligent driving system. Furthermore, the object detection results cannot indicate the drivers' attentional area, nor the drivers' attention allocation. On the contrary, our model can not only detect the locations of safe driving related objects (selective attention) but also show the attention allocation when driving. The deep red in the saliency map illustrates the most important area that the drivers should consider for driving safety, and the yellow or light blue shows the secondary/tertiary important areas. Therefore, we hope that the human-like driving method based on visual attention would be taken into account in future intelligent driving systems.

VI. Conclusion

In conclusion, in this paper, we provide a traffic driving video dataset with multiple drivers' eye tracking data that includes bottom-up and top-down attention on traffic driving. We further propose a convolutional-deconvolutional neural network (CDNN) for predicting drivers' eye fixations on traffic driving videos. The proposed network is trained by multiple drivers' eye tracking data, and it also includes bottom-up and top-down visual attentional information on traffic driving. The experimental results indicate that the proposed CDNN outperforms the state-of-the-art saliency models and predicts drivers' fixation locations more accurately. The proposed CDNN can predict the major fixation locations, and it also shows excellent detection of secondary important regions that cannot be ignored during driving if it exists. Compared with the present object detection models in autonomous and assisted driving system, such as YOLOv3, our human-like driving model does not detect all of the objects appearing in the driving scenes, but it provides the most important and relative regions or targets, which can largely reduce the interference of irrelevant scene information.

However, there are some limitations in our current work. For example, the temporal information that is critically important for video fixation prediction is not considered in our model. Some temporal-spatial networks such as long short-term memory (LSTM) network or optical flow model can be considered in the further work. Besides, more driving video samples including various weather and traffic environment can be trained and tested in the future study.

Acknowledgment

The authors would like to thank each eye movement participant and China Scholarship Council.

References

[1] B. Fröhlich, M. Enzweiler, and U. Franke, "Will this car change the lane?—Turn signal recognition in the frequency domain," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2014, pp. 37–42.

[2] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena, "Car that knows before you do: Anticipating maneuvers via learning temporal driving models," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3182–3190.

[3] P. Kumar, M. Perrollaz, S. Lefèvre, and C. Laugier, "Learning-based approach for online lane change intention prediction," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2013, pp. 797–802.

[4] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[5] L. Fridman, P. Langhans, J. Lee, and B. Reimer, "Driver gaze region estimation without use of eye movement," *IEEE Intell. Syst.*, vol. 31, no. 3, pp. 49–56, May/Jun. 2016.

[6] A. Tawari, S. Martin, and M. M. Trivedi, "Continuous head movement estimator for driver assistance: Issues, algorithms, and on-road evaluations," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 818–830, Apr. 2014.

[7] A. Tawari, K. H. Chen, and M. M. Trivedi, "Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation," in *Proc. IEEE Conf. Intell. Transp. Syst.*, Oct. 2014, pp. 988–994.

[8] L. Simon, J.-P. Tarel, and R. Brémond, "Alerting the drivers about road signs with poor visual saliency," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2009, pp. 48–53.

[9] N. Pugeault and R. Bowden, "How much of driving is preattentive?" *IEEE Trans. Veh. Technol.*, vol. 64, no. 12, pp. 5424–5438, Dec. 2015.

[10] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "POSEidon: Face-from-depth for driver pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5494–5503.

[11] A. Tawari and M. M. Trivedi, "Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2014, pp. 344–349.

[12] F. Vicente, Z. Huang, X. Xiong, F. D. L. Torre, W. Zhang, and D. Levi, "Driver gaze tracking and eyes off the road detection system," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2014–2027, Aug. 2015.

[13] T. Deng, K. Yang, Y. Li, and H. Yan, "Where does the driver look? Top-down-based saliency detection in a traffic driving environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 2051–2062, Jul. 2016.

[14] T. Deng, A. Chen, M. Gao, and H. Yan, "Top-down based saliency model in traffic driving environment," in *Proc. IEEE Conf. Intell. Transp. Syst.*, Oct. 2014, pp. 75–80.

[15] T. Deng, H. Yan, and Y.-J. Li, "Learning to boost bottom-up fixation prediction in driving environments via random forest," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 9, pp. 3059–3067, Sep. 2018.

[16] Z. Bylinskii *et al.* (2015). *Mit Saliency Benchmark*. [Online]. Available: http://saliency.mit.edu/

[17] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1072–1080.

[18] S. Mathe and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1408–1424, Jul. 2015.

[19] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara, "DR(eye)VE: A dataset for attention-based tasks with applications to autonomous and assisted driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun./Jul. 2016, pp. 54–60.

[20] T. Amer, K. W. J. Ngo, and L. Hasher, "Cultural differences in visual attention: Implications for distraction processing," *Brit. J. Psychol.*, vol. 108, no. 2, pp. 244–258, 2017.

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

[22] A. Tawari and B. Kang, "A computational framework for driver's visual attention using a fully convolutional architecture," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2017, pp. 887–894.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[24] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980

[25] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in *Proc. Int. Conf. Pattern Recogit. (ICPR)*, Dec. 2016, pp. 3488–3493.

[26] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Neural Inf. Process. Syst.*, 2006, pp. 545–552.

[27] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, Jan. 2012.

[28] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[29] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 996–1010, Apr. 2013.

[30] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1153–1160.

[31] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. (2016). "What do different evaluation metrics tell us about saliency models?" [Online]. Available: https://arxiv.org/abs/1604.03605

[32] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 55–69, Jan. 2013.

[33] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2106–2113.

[34] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," MIT, Cambridge, MA, USA, Tech. Rep., 2012.

[35] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vis. Res.*, vol. 45, no. 8, pp. 2397–2416, 2005.

[36] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Information-theoretic model comparison unifies saliency metrics," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 52, pp. 16054–16059, 2015.

[37] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vis. Res.*, vol. 47, no. 19, pp. 2483–2498, 2007.

[38] O. Pele and M. Werman, "A linear time histogram metric for improved SIFT matching," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2008, pp. 495–508.

[39] A. Borji, D. N. Sihite, and L. Itti, "What/where to look next? Modeling top-down visual attention in complex interactive environments," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 5, pp. 523–538, May 2014.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 630–645.

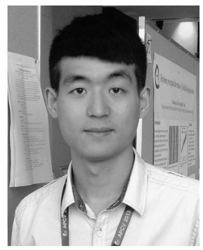
[42] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[43] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.

[44] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

[45] J. Redmon and A. Farhadi. (2017). "Yolo9000: Better, faster, stronger." [Online]. Available: https://arxiv.org/abs/1612.08242

[46] J. Redmon and A. Farhadi. (2018). "YOLOv3: An incremental improvement." [Online]. Available: https://arxiv.org/abs/1804.02767



Tao Deng received the Ph.D. degree from the MOE Key Lab for Neuroinformation, University of Electronic Science and Technology of China, in 2018. He is currently with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China.

His research interests include visual attention, cognition, computer vision, image/video processing, and intelligent transportation.

Hongmei Yan received the Ph.D. degree from Chongqing University in 2003. She is currently a Professor with the MOE Key Lab for Neuroinformation, University of Electronic Science and Technology of China, Chengdu, China.

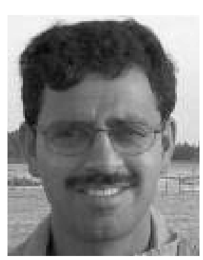
Her research interests include visual cognition, eye movements, visual attention, and saliency detection.

Long Qin is currently pursuing the M.S. degree in biomedical engineering from the MOE Key Lab for Neuroinformation, University of Electronic Science and Technology of China, Chengdu, China.

His research interests include visual attention, cognition, vision computation, saliency detection, and object detection.

Thuyen Ngo is currently pursuing the Ph.D. degree with Department of Electrical and Computer Engineering, University of California, Santa Barbara, Santa Barbara, CA, USA.

His research interests include computer vision, image/video processing, and eye tracking.



B. S. Manjunath (F'05) received the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1991.

Since then, he has been with the University of California, Santa Barbara, CA, USA, where he is currently a Distinguished Professor of Electrical and Computer Engineering and directs the Center for Multimodal Big Data Science and Healthcare. He has authored over 300 peer-reviewed articles and is a coeditor of the book entitled *Introduction to MPEG-7* (Wiley, 2002). His broad research interests include image processing and computer vision.