Improving Driver Gaze Prediction with Reinforced Attention

Kai Lv, Hao Sheng*, *Member, IEEE,* Zhang Xiong, Wei Li, *Member, IEEE,* Liang Zheng, *Member, IEEE*

***Abstract*—We consider the task of driver gaze prediction: estimating where the location of the focus of a driver should be, based on a raw video of the outside environment. In practice, we output a probability map that gives the normalized probability of each point in a given scene being the object of the driver attention. Most existing methods (*i.e.*, *Coarse-to-Fine* and *Multi-branch*) take an image or a video as input and directly output the fixation map. While successful, these methods can often produce highly scattered predictions, rendering them unreliable for real-world usage. Motivated by this observation, we propose the reinforced attention (RA) model as a regulatory mechanism to increase prediction density. Our method is built directly on top of existing methods, making it complementary to current approaches. Specifically, we first use *Multi-branch* to obtain an initial fixation map. Then, RA is trained using deep reinforcement learning to learn a location prediction policy, producing a reinforced attention. Finally, in order to obtain the final gaze prediction result, we combine the fixation map and the reinforced attention by a mask-guided multiplication. Experimental results show that our framework improves the accuracy of gaze prediction, and provides state-of-the-art performance on the DR(eye)VE dataset.**

***Index Terms*—gaze prediction, driver attention, reinforcement learning, video processing, deep learning**

I. Introduction

Autonomous and assisted driving are some of the most active research areas in computer vision. Typically, these works have focused on lane change assistance [1], traffic signs recognition [2], and many more [3]. Recently, several works [4], [5] have advocated for a new assisted driving paradigm - driver gaze prediction. The goal of gaze prediction is to provide useful suggestions to the driver where where they should focus. In this task, the gaze points are gathered from real driving scene, and are defined as the ground truth of the training dataset. In practice, gaze is defined as a probability map where each point in a given scene has a value. This value denotes how much probability this point is the gaze of the driver.

Kai Lv and Hao Sheng are with State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100191, P.R. China, with Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China, and with Beihang Hangzhou Institute for Innovation at Yuhang, Beihang University, Hangzhou, P.R. China, 311121. Email: {lvkai, shenghao}@buaa.edu.cn.

Zhang Xiong and Wei Li are with School of Computer Science and Engineering, Beihang University, Beijing 100191, China. Email: xiongz@buaa.edu.cn, liwei@nlsde.buaa.edu.cn.

Liang Zheng is with Research School of Computer Science, Australian National University, Canberra 2601, Australia. Email: liang.zheng@anu.edu.au.

Hao Sheng is the corresponding author.

(a) RGB image (b) Ground truth

(c) Multi-branch [9] (d) Ours

Fig. 1. An example of gaze prediction while driving. The results of (c) Multi-branch and (d) ours are produced with the input (a) a video clip. Comparing to (c) the result of Multi-branch, (d) the predicted map of ours is more concentrated and more accurate.

Some previous works take saliency detection methods to address the challenge of driver gaze prediction. Some attempt to capture salient objects or events that occur naturally in the environment as driver gaze [6], [7], [8]. Other models combine saliency details with motion cues to handle gaze prediction task [4], [9], [10], [11]. These methods do not directly use ground truth human gaze and instead attempt to approximate the task purely via environment cues. Recent works [5], [9] propose to specifically use supervised gaze prediction to achieve higher performance. For example, Palazzi *et al.* [5] propose a model based on C3D [12] that takes videos of driver outside environment as input, allowing the model to explicitly take into account the temporal dimension.

Multi-branch [9] is another work solving driver gaze prediction in a supervised strategy and is considered state-of-the-art. Multi-branch has the architecture of three branches: RGB image, optical flow and semantic segmentation. Each branch provides complementary details for the overall model to contribute to the final prediction.

However, as shown in Fig. 1(c), the Multi-branch [9] tends to produce scattered gaze prediction maps. The prediction map of Multi-branch does not reflect the distribution of the ground truth gaze, as shown in Fig. 1(b), where gaze tends to appear in tightly clustered areas. In other words, the predicted gaze should mainly be concentrated in one area. The concentrated map reflects the nature of human perception: the focus of driver gaze is localized to the most important part of the environment at any given point in time.

In contrast to most previous approaches, our work aims

to introduce attention as a regulatory mechanism to increase prediction density and accuracy. In this paper, the attention means the location where the driver should mainly focus on. We believe that there is only one attention at a moment. The attention localization can be solved by a regression method (*i.e.*, Recurrent Neural Network (RNN)). Given a sequence of frames, we can directly apply a standard RNN to produce an attention location. In RNN, we need to apply convolution for the entire frames. However, this process is computationally expensive, as the computational cost scales linearly with the number of image pixels. While down-sampling the input frames can reduce the computational burden, many local details (*e.g. lanes and signs*) are lost in the process. Instead, we propose to use a reinforced attention model to estimate the attention, which we term Reinforced Attention (RA). Our method has the following characteristics: 1) The backbone of the RA is a recurrent neural network where the input frames are processed sequentially. 2) RA selectively chooses parts of the images to process to save computational resources. At each frame, the model has a sampler to select the next patch to sample based on the previous internal state. At the final frame, the model makes the decision on where is the attention. The above procedure uses Williams's REINFORCE [13] to address the non-differentiates due to the control problem.

In addition, we also introduce speed and course details into the reinforced model. Previous works [9] only analyse the relationship between the gaze locations and speeds, and do not make use of speed and course details to train their gaze prediction model. In contrast, our method uses these details to improve the gaze prediction accuracy. Specifically, we feed these details as well as the patches obtained by the reinforced model into the RA model.

Overall, we propose a driver gaze prediction method by introducing reinforced attention. The overall framework can be described as follows. First, we use Multi-branch [9] to obtain an initial prediction. Multi-branch integrates three sources of information: raw video, motion and scene semantics. We use the predicted gaze map as our baseline. Then, the RA is employed to estimate the attention location for the input video. Finally, we combine the gaze prediction map and the attention to generate the final result.

To summarize, we make the following contributions.

- We introduce RA to estimate the attention, which helps produce more clustered and accurate predictions.
- We argue that speed and course details are useful cues for attention localization. To the best of our knowledge, we are the first to introduce speed and course into the attention localization task.
- We employ a reinforcement learning strategy, achieving competitive results at a lower computational cost.
- Extensive experiment confirms the consistent effectiveness of RA and the overall framework.

II. Related Work

Appearance-based gaze prediction. There are two kinds of gaze prediction tasks, appearance-based gaze prediction and scene-based gaze prediction. Appearance-based gaze prediction takes images of human face and annotated eye gaze as input to learn a direct image-to-gaze mapping. Feng *et al.* [14] propose a hidden Markov model based gaze prediction system that utilizes the visual saliency of the content being viewed. Davies *et al.* [15] present a multicue gaze prediction framework for open signed video content, and inverstigate which cues are relevant for gaze prediction. Kellnhofer *et al.* [16] present Gaze360, a large-scale gaze-tracking dataset. In [16], the authors propose a 3D gaze model that includes temporal information and output the probability of gaze prediction. Other deep learning methods employed to solve the appearance-based gaze prediction task are described in [17], [18], [19]. These methods need to apply eye or face detection before training. The appearance-based task is primarily studied as a behavioral cue to better understand human thought processes.

Scene-based gaze prediction. In scene-based gaze prediction, images or videos of the outside environment are taken as input. Since Alletto *et al.* [5] propose the DR(eye)VE dataset, the task of driver gaze prediction has been extensively studied. Footage for the DR(eye)VE dataset is recorded as individuals are driving and the driver gaze is also saved. Palazzi *et al.* [4] model the driver gaze by training a coarse-to-fine convolutional network on short video clips from the DR(eye)VE dataset. In [9], the authors propose a complementary model based on a deep multi-branch architecture. This model integrates three sources of information: raw video, motion cues (in terms of optical flow) and scene semantics. This work focuses on the task of scene-based gaze prediction. The input is 16 raw video frames and the output is a gaze prediction map. Compared to appearance-based methods, the scene-based method primarily studies the relationship between the drivers and the scenes they are viewing. In this work, we focus on the scene-based gaze prediction.

Assisted driving. Gaze prediction tasks are primarily studied in the context of assisted driving [20]. Simon *et al.* [21] present an improvement of the advanced driver assistance systems by estimating the saliency of road signs using SVM learning techniques. Bredmond *et al.* [22] review a cluster of visual attention studies and conduct more realistic experiments with a larger set of targets, including pedestrians and bicycles. Pugeault *et al.* [23] propose a novel vision-based method that predicts driver behavior in real-time. They find that the field of view used by the computational model is closely related to driver gaze locations. In addition, Li *et al.* [20] focus on monitoring the driver attention level and propose a driver monitoring system that is able to sense inattentive drivers.

Deep Reinforcement Learning. Our method utilizes deep reinforcement learning (DRL) to locate the regions of attention. DRL is a framework in which decision-making networks interact with an environment and seek to learn a policy to take actions that maximise an environment reward. Similar to deep learning methods [24], [25], [26], DRL has been applied to many computer vision tasks [27], [28], including bounding box location prediction [29], image caption [30], and seeding points for segmentation [31]. Several works utilize DRL to predict the focus of human attention. Minut *et al.* [32] propose a model of selective attention for visual search tasks and introduce a reinforcement learning framework for

sequential decision-making. Mnih *et al.* [33] present a novel recurrent neural network model for classification task. This method can extract information from an image by adaptively selecting a sequence of regions or locations and processing the selected regions only with high resolution. In [34], the authors propose a Dueling Network, which represents two separate estimators: one for the state value function and one for the state-dependent action advantage function. They also illustrate that the area of the network paying attention to is reasonable and important. Xu *et al.* [35] apply A3C [36] to predict head movement positions and focus of attention. Inspired by these reinforced methods, we train a network based on Williams's REINFORCE [13] to learn the policy in continuous action space to produce the driver attention.

III. The Proposed Method

In this section, we review the Multi-branch [9] model, which is applied to predict the driver gaze in Section III-A. We then describe the process of acquiring the reinforced attention in Section III-B. After getting the initial gaze prediction map and the attention, we employ mask-guided multiplication, in Section III-C, to generate the final prediction.

A. Multi-branch Review

In this work, we employ the Multi-branch [9] model to get an initial gaze prediction map. The input of Multi-branch is a video clip with 16 frames. The Multi-branch model is composed of three different branches: RGB image (I branch), optical flow (F branch) and semantic segmentation (S branch). Each branch exploits complementary details which contribute to the final prediction.

The three branches have the same architecture. Each branch is a two-input two-output architecture composed of two streams, the cropped stream and the resized stream. The two inputs are fed into a weight shared C3D [37] model which has been pre-trained. The resized stream differs from the cropped stream due to a set of refine layers following the C3D model. The prediction of the resized input is stacked with the last frame of the video clip and then fed to the refine layers. The refine layers then process and upsample the tensor back to the input spatial resolution.

The model is trained in two steps. The first is single branch training in which each branch is trained separately using the aforementioned method. Following this, the three branches are then simultaneously fine-tuned. Prediction cost is minimized in terms of Kullback-Leibler divergence:

$$D_{KL}(Y||\hat{Y}) = \sum_i Y(i) log(\epsilon + \frac{Y(i)}{\epsilon + \hat{Y}(i)}), \tag{1}$$

where $\hat{Y}$ is the prediction map of Multi-branch and Y is the ground truth. i is the summation index that spans across image pixels and ϵ denotes a small constant.

B. Reinforced Attention

This section presents how the Reinforced Attention (RA) obtains the attention. Fig. 2 shows the overall framework of RA, in which *REINFORCE* [38] is embedded to generate the reinforced attention.

In a typical DRL-based method, there exists an agent taking actions in an environment to maximize the cumulative rewards. The process by which an agent makes decisions is called policy π. π can be a deep network and map states to the action with some probability. The environment can be modelled as a Markov decision process, where the current state and action depend only on previous state and action. Given an observation o_t at each time step t, the agent chooses an action a_t and receives a reward R_t. The reward R_t at each step t is either a future or discounted reward, which can be described as $R_t = \sum_{i=t}^{T} \gamma^{i-t} r(s_i, a_i)$.

We formulate the driver gaze prediction task in an RL framework as shown in Fig. 2. The input of the attention model is video frames $\{F_t\}_{t=1}^{T}$ with frame number t ranging from 1 to T and the output is a predicted location l_p, where the attention is located.

Actions. The parameters of the RA agent are composed of several networks described as: $\{\theta_g, \theta_k, \theta_h, \theta_s, \theta_p\}$. Here $f_h(\theta_h)$ is the core network. As shown in Fig. 2, the core network at time t takes three kinds of inputs: the localization and observation feature g_t, the speed-course feature k_t and the internal representation h_{t-1} at previous time step. $f_s(\theta_s)$ has the same architecture with $f_p(\theta_p)$, which is applied to produce a location. The difference is that $f_s(\theta_s)$ is a sampler deciding where to sample the patch at frame $t+1$, while $f_p(\theta_p)$ determines the final action. This final action is the predicted location of attention for the input video clip. Our objective is to maximize the expectation of accumulated future rewards:

$$J(\theta) = \mathbb{E}_{p(s_{1:T};\theta)}[\sum_{t=1}^{T} r_t] = E_{p(s_{1:T};\theta)}[R], \tag{2}$$

where $s_{1:T}$ is a possible interaction sequences and $p(s_{1:T};\theta)$ is the probability of $s_{1:T}$. Here $p(s_1 : T;\theta)$ is based on current policy θ. During training, We use REINFORCE [38] to calculate the gradient,

$$\begin{aligned} \nabla_\theta J &= \sum_{t=1}^{T} \mathbb{E}_{p(s_{1:T};\theta)}[\nabla_\theta log\pi(a_t|s_{1:t};\theta)R] \\ &\approx \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T} \nabla_\theta log\pi(a_t^i|s_{1:t}^i;\theta)R^i, \end{aligned} \tag{3}$$

where a_t is the action we take in time step t.

Rewards. One of the major tasks in training the RA network is formulating the reward function. A reward R reflects the quality of the action and is given back to the agent. In the case of RA, R is determined by the ground truth attention $l_g = (x_g, y_g)$ and final action $l_p = (x_p, y_p)$, which is produced by $f_p(\theta_p)$. We use Euclidean distance $d(\cdot,\cdot)$ to calculate the distance between two points and the reward function R is defined as:

$$R = 1 - d(l_p, l_g). \tag{4}$$

Training the RA model. In training the RA model, the actor interacts with the environment. The interaction is achieved in our RA model through the following procedure:

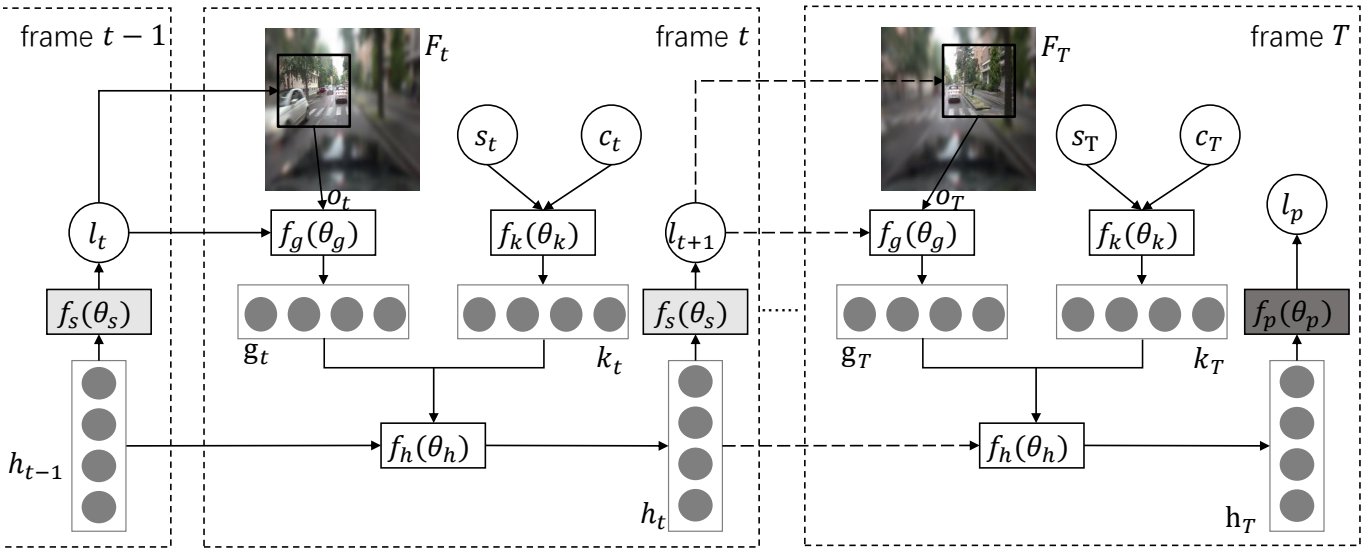

Fig. 2. Architecture of the *reinforced attention* (RA) model. Our model is based on Recurrent Neural Network (RNN) and the RNN iteration is repeated for T steps. *1) At Frame t*, the core network $f_h(\theta_h)$ takes multiple features as input: observation feature g_t, speed-course feature k_t and the internal feature h_{t-1}. Note that h_{t-1} is produced at the previous time step $t-1$. *2) At the last Frame* T, the model produces the final location l_p. Instead of applying $f_s(\theta_s)$ to generate a sample location, $f_p(\theta_p)$ is applied to produce the final attention location l_p of the video clip. Note that $f_p(\theta_p)$ and $f_s(\theta_s)$ have the same architecture but different parameters.

(1) At frame $t \in [1, ..., T-1]$, the sampler obtains the current observation o_t from the input frame F_t, according to the position l_t. Note that l_t is generated by processing the previous frames. More specifically, o_t is a square with a side length of 64 pixels centered at l_t.

(2) Multiple details from the current frame t and the RNN feature h_{t-1} in the last frame are fed into the core network $f_h(\theta_h)$. These details of current frame include the observation o_t, the location l_t, speed s_t and course c_t. As shown in Fig. 2, the RA model contains several deep modules, which are to extract the features described above.

(3) The RA model produces the RNN feature h_t by $f_h(\theta_h)$. Based on h_t, the sample network $f_s(\theta_s)$ produces the next sample location l_{t+1}. Note that h_t and l_{t+1} will be delivered to next frame $t+1$.

(4) When arriving at the final frame $t = T$, RA uses action $f_p(\theta_p)$ to output the final location of attention. $f_p(\theta_p)$ and $f_s(\theta_s)$ have the same network architecture but different parameters. This is primarily because $f_s(\theta_s)$ is a sampler which mines useful details, including the sample location and the corresponding observation. However, $f_s(\theta_s)$ is employed to make final decisions based on the driving state collected from the video clip.

(5) Once the RA model meets the termination condition of giving the final decision l_p, all experiences, together with the rewards R are delivered to the optimizer to upgrade the RA network.

C. Prediction with Reinforced Attention

In this section, we propose mask-guided multiplication to generate the final gaze prediction. After obtaining the reinforced attention, we need to combine it with the fixation map F described in Section III-A. The fixation map is a $W \times H$ distribution matrix and the reinforced attention l_p is a location. Thus, we generate a mask M by using the attention location l_p. Here M is 2-D gaussian distribution with mean (x_p, y_p) and $\sigma = 0.1$. Following this, each fixation map, F, generated by *Multi-branch* is masked with the learned attention mask M. The final gaze prediction map $\hat{F}$ is defined as:

$$\hat{F} = F \odot M, \tag{5}$$

where $\odot$ denotes element-wise multiplication.

Fig. 3. A demonstration for the selective process mechanism. At each time step, the sampler $f_s(\theta_s)$ chooses a patch rather than processing the entire frame.

D. Discussion

The difference between the RA model and the typical RNN model. Attention localization can be regarded as a regression task with video as input and a prediction map as output. Typically this is achieved by feeding all video frames into the RNN model and then processing them with convolution filter

maps. This process is computationally expensive, because all the pixels would be processed by convolutional process.

In contrast to the typical RNN, the proposed RA model adaptively selects a region to process at high resolution. Comparing to RNN, RA process less pixels and thus is more computational efficient. This technique is inspired by the perception mechanisms of human drivers who do not process a scene in full. Experiment in Section IV-E indicates that RA has a competitive accuracy when compared with the typical RNN. The main reason is that continuous frames contain much redundant information, it is unnecessary to process the entire frame at each step.

The mechanism of the sampler. We illustrate the sampling process in Fig. 3. Taking $T = 16$ frames as input, the sampler selects a specific area for each frame to process. As the frames move forward, the sampler is found to select different locations to process. A possible explanation is that the small differences between successive frames lead the sampler to explore other areas in which it can collect new details. This can also be explained by the fact that the reinforced model is fed with appearance and location features so the sampler can perceive which areas have been collected and which are not.

IV. Experiment

A. Datasets and Evaluation Metrics

Datasets. We evaluate our method on DR(eye)VE [5], which is the first publicly available dataset addressing driver gaze prediction. While there exist several dataset for gaze prediction [16], [18], [39], only DR(eye)VE is setup for scene-based driver gaze prediction, the rest being designed for appearance-based gaze prediction tasks.

The DR(eye)VE dataset consists of 74 sequences with 555,000 frames. Each sequence is 5 minutes long and is captured at 1080p/25fps. Additionally, the dataset includes other relevant driving state such as GPS data, accelerometer and gyroscope measurements. In this work, we mainly concern the speed and course details, which are then employed in our reinforced attention model.

Moreover, we evaluate the methods on the complete test set as well as the acting subset. The acting subset is particularly interesting as the deviation of driver gaze from central pattern denotes an intention related to some driving actions (*e.g.* changing lanes and overtaking).

Evaluation metrics. To evaluate the proposed method, we compare our approach with the state-of-the-art methods primarily in two aspects, *i.e.,* the accuracy of gaze prediction and the accuracy of reinforced attention. We evaluate the gaze prediction result following the guidelines in [40], [9]. Specifically, we use Person's Correlation Coefficient (CC), Kullback-Leibler Divergence (D_{KL}) and Information Gain (IG) for evaluation on DR(eye)VE. For CC and IG, higher is better, while for D_{KL} the opposite is true. IG is a measure of the quality of a predicted map P with respect to a ground truth map Y in presence of a strong bias,

$$IG(P, Y, B) = \frac{1}{N} \sum_i Y_i [log_2(\epsilon + P_i) - log_2(\epsilon + B_i)], \quad (6)$$

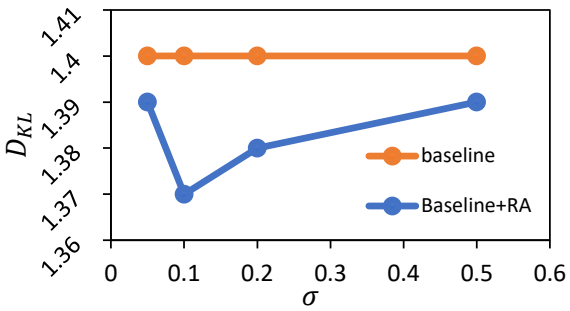

Fig. 4. Evaluation with different σ on the complete dataset of DR(eye)VE.

where i is an index spanning all the N pixels in the image and ϵ is a pre-defined constant to ensure numerical stability. Note that B is the bias map computed by averaging the training fixation map. The ground truth map Y for a frame is built by accumulating the gaze points of the nearby 25 frames [9]. For evaluating the reinforced attention model, we use the Euclidean distance $d(l_p, l_g)$ between two locations to evaluate the distance between the predicted location l_p with ground truth attention l_g. l_g is a location, which is obtained by calculating the gaze prediction ground truth Y. As shown in Fig. 5, for each ground truth Y (column 2), there exist several gathered points. All the points are Gaussian distribution and they have the same maximum value. In this work, the ground truth attention l_g is defined by averaging the coordinates of these maximum points:

$$l_g = \frac{1}{N} \sum_{n=1}^{N} l_n, \quad (7)$$

where N is the number of maximum value points in Y and (x_n, y_n) is the coordinate of n-th point l_n. Note that $(x_n, y_n) \in [-1, 1]$.

B. Experiment Setting

Baseline model for gaze prediction. We use Multi-branch as the baseline of our overall framework. To train the baseline, we follow the training strategy in [9]. Specifically, the Multi-branch model is split into three branches where each branch is fed with 16 frames clips in raw RGB color space, 16 frames clips with optical flow maps and 16 frames clips with semantic segmentation maps, respectively. The training process is divided into two stages: training the three branches independently and then fine-tuning the complete Multi-branch model with a lower learning rate value. More details can be found in [9].

Reinforced attention model. We train the reinforced attention model following Sec. III-B. In our implementation, input frames are resized to 256×256 and the sampled patch has the size of 64. The dimension of the internal state vector h_t, appearance feature g_t and speed-course feature k_t is 256. In order to extract the observation feature o_t, a pre-trained Resnet-50 [41] is deployed and fine-tuned when training the RA model. The model is trained via stochastic gradient descent with batch size 32 and momentum of 0.9. The learning rate is initialised at 0.003.

Method	*CC* ↑	D_{KL} ↓	*IG* ↑
Baseline Gaussian	0.40	2.16	-0.49
Baseline Mean	0.51	1.60	0.00
Mathe *et al.* [10]	0.04	3.30	-2.08
Wang *et al.* [42]	0.04	3.40	-2.21
Wang *et al.* [43]	0.11	3.06	-1.72
MLNet [44]	0.44	2.00	-0.88
RMDN [**?**]	0.41	1.77	-0.06
Palazzi *et al.* [4]	0.55	1.48	-0.21
Multi-branch [9]	0.56	1.40	0.04
Ours	**0.58** ± 0.008	**1.37** ± 0.004	**0.05** ± 0.016

TABLE I
COMPARISON WITH STATE-OF-THE-ART ON THE COMPLETE TEST SET OF DR(EYE)VE.

Method	*CC* ↑	D_{KL} ↓	*IG* ↑
Baseline Gaussian	0.26	2.41	0.03
Baseline Mean	0.22	2.35	0.00
MLNet [44]	0.32	2.35	-0.36
RMDN [**?**]	0.31	2.13	0.31
Palazzi *et al.* [4]	0.37	2.00	0.20
Multi-branch [9]	0.41	1.80	0.51
Ours	**0.42** ± 0.016	**1.79** ± 0.004	**0.51** ± 0.008

TABLE II
COMPARISON WITH STATE-OF-THE-ART ON THE ACTING SUBSEQUENCES OF DR(EYE)VE.

C. Parameters Analysis

An important hyper-parameter of RA is σ described in Section III-C. This parameter is used to generate the weighted mask M. We show its impact by varying its value in Fig. 4. We observe that our method with different σ consistently improves the *Multi-branch* method. The best performance is achieved when $\sigma = 0.1$. Note that we set $\sigma = 0.1$ in the following experiments.

D. Comparison with State-of-the-art Methods

We compare our approach with state-of-the-art methods on the complete test set and the acting subsequences of DR(eye)VE. Note that we use the same model to test on the two sets. Table I and II report the comparison when tested on the complete set and the acting set, respectively. We compare with two baseline methods: *Baseline Gaussian* and *Baseline Mean*, several salience detection based methods: MLNet [44] and RMDN [**?**], and previous gaze prediction methods: Palazzi *et al.* [4] and Multi-branch [9].

We first compare with two baseline methods which do not require training. The *Baseline Gaussian* method is employed by a centered gaussian baseline and the *Baseline Mean* is generated by averaging all training set prediction maps. On the complete test set and the acting subsequences, the proposed method outperforms the *Baseline Gaussian* and the *Baseline Mean*. For example, the D_{KL} value of *Baseline Gaussian* and *Baseline Mean* are 2.16 and 1.60, respectively, both higher than the value of our method. This indicates that our method can deal with task-driven changes in gaze prediction, even if the gaze distribution is often strongly biased to the vanishing point of the road.

Next, we compare with several saliency prediction based methods, which can be employed to solve gaze prediction task.

Method	complete set	acting set
RNN *w/o* speed and course	0.27	0.27
RNN *w/o* speed	0.22	0.25
RNN *w/o* course	0.27	0.27
RNN	0.20	0.24
RA *w/o* speed and course	0.26	0.28
RA *w/o* speed	0.21	0.25
RA *w/o* course	0.24	0.26
RA	**0.20**	**0.23**

TABLE III
COMPARISON OF ATTENTION LOCALIZATION METHODS ON THE DR(EYE)VE DATASET IN TERMS OF L_2 DISTANCE.

As shown in Fig. I, RMDN [**?**] yields a D_{KL} of 1.77 on the test set. It is lower than the Baseline Gaussian but is higher than Palazzi *et al.* [4] by 0.29. This is probably due to that the [4] is a task-orientated method designed for gaze prediction, while RMDN is not.

We also evaluate the methods on the complete set and the acting subset. We observe that the results of *Baseline Gaussian* and the *Baseline Mean* performed on the complete set are not as accurate as those performed on the acting set. For example, the *Baseline Gaussian* achieves $D_{KL} = 2.16$ when tested on complete set, and obtains $D_{KL} = 2.41$ on when tested on the acting set. It indicates that the deviation of driver gaze from the central pattern does occur when undergoing task-specific actions (*e.g. changing lanes and overtaking*). One possible solution is to introduce attention mechanism, which can highlight the valuable ares. Results show that our method provide more accurate gaze maps on both sets.

Comparing to state-of-the-art gaze prediction methods, our approach clearly achieves higher performance on both sets. Specifically, our method achieves $D_{KL} = 1.37$ and $CC = 0.58$ on complete set, and obtains $D_{KL} = 1.79$ and $CC = 0.42$ on acting set. The D_{KL} value is 0.03 and 0.01 lower than the current best results (Multi-branch [9]) when tested on the complete set and the acting set, respectively. These results show that our method is a good complementary to the existing methods. Qualitatively results are shown in Fig. 5. The results illustrate that our method can generate more concentrated and more accurate gaze maps.

E. Evaluation of Reinforced Attention

Compared to RNN, our approach is competitive in accuracy and more computational efficient. In this section, we evaluate the reinforced attention (RA) model. On the DR(eye)VE dataset, we compare RA with typical recurrent neural network (RNN). The RNN takes a set of images as input and can be viewed as a standard regression problem. On the complete test set, the proposed RA yields an L_2 of 0.20, while the L_2 of RNN is also 0.20. On the acting set, RA outperforms RNN by 0.01 in terms of L_2. It clearly indicates that our method is competitive in accuracy. Meanwhile, to process a video clip, RNN need to apply convolution operation for the entire frames, while RA selectively chooses a patch of each frame. The patch is one quarter of the size of the entire image. It indicates that the computational cost of RA is about a quarter of that of RNN.

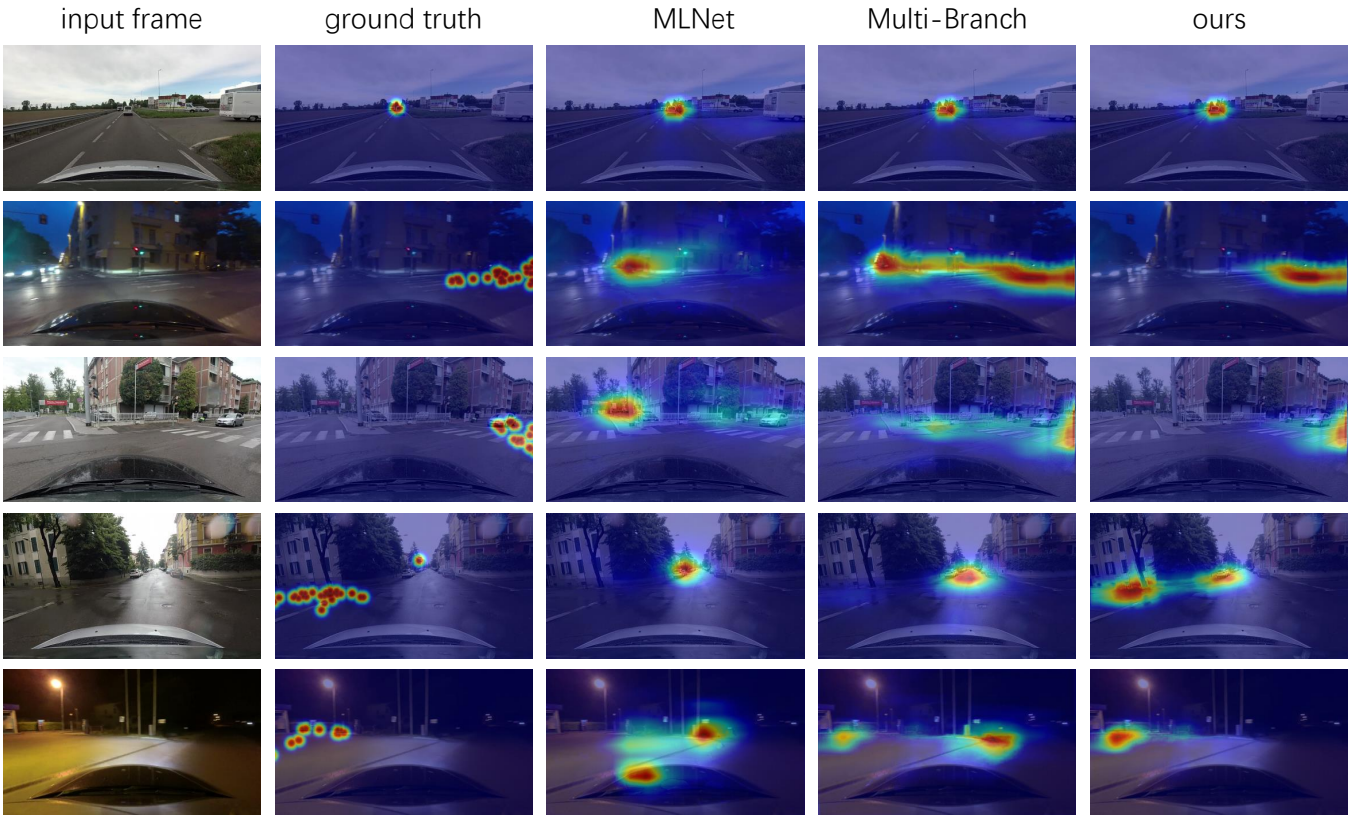

Fig. 5. Qualitative examples of the gaze prediction maps. Each row is a set of examples. Each test sample is represented by five images: input raw image, ground truth map, prediction of MLNet [44], prediction of Multi-branch [9] and our prediction. The results show that the prediction maps produced by our method are more focused (the first three rows) and more accurate (the last two rows).

Speed and course details are useful for attention localization. This section also presents ablation studies of the RA model. Since the driving condition details are involved, *i.e.*, speed and course, we remove them one at a time to evaluate their contribution respectively. Results on DR(eye)VE are shown in Table III.

When removing speed and course from the RA system, L_2 will be 0.06 higher than the full RA model in the complete set. Meanwhile, the L_2 of RNN is 0.20, which is also lower than RNN *w/o* speed and course. Similar observation can be made on the acting set as well. Recall that lower L_2 is better. These results indicate that speed and course details play an important role in estimating the driver attention.

To further evaluate the respective importance of speed and course details, we add another experiment. we remove speed or course details, one at a time from the system. Overall, the models *w/o* speed or *w/o* course achieve lower L_2 than the ones *w/o* speed and course, but achieve higher L_2 than the complete system. For example, RA *w/o* course yields 0.24 in L_2, which is slightly lower than the RA *w/o* speed and course but higher than complete RA. We also find that the course detail is more import than the speed. For example, on acting set, RA *w/o* speed yields a L_2 of 0.25, which is lower than the RA *w/0* course.

Method	*CC* ↑	D_{KL} ↓	*IG* ↑
I	0.55	1.41	-0.01
F	0.51	1.61	-0.13
S	0.47	1.69	-0.11
IFS	0.56	1.40	0.04
I+RA	0.56(+0.01)	1.40(-0.01)	0.01(+0.02)
F+RA	0.52(+0.01)	1.55(-0.06)	-0.12(+0.01)
S+RA	0.50(+0.03)	1.59(-0.10)	-0.10(+0.01)
IFS+RA	**0.58 (+0.02)**	**1.37(-0.03)**	**0.05(+0.01)**

TABLE IV
VARIANT STUDY OF OUR METHOD ON THE COMPLETE TEST SET OF DR(EYE)VE. I, F AND S REPRESENT IMAGE, OPTICAL FLOW AND SEMANTIC SEGMENTATION BRANCHES, RESPECTIVELY. (IFS) EQUALS THE MULTI-BRANCH MODEL. RA MEANS THE REINFORCED ATTENTION.

F. Variant Study.

We further evaluate three different variants of Multi-branch. *i.e.*, Image Branch, Flow Branch and Segmentation Branch. As mentioned above, our method is built directly on top of existing methods. Thus, we use other methods as our baseline. As described in Section III-A, image branch, flow branch and segmentation branch are parts of the Multi-branch model. Table IV details the CC, D_{KL} and IG results of each variant in the DR(eye)VE complete set. *I* refers to the RGB image branch. As the result shows, *I+RA* yields 1.40 in D_{KL}, which is slightly lower than the baseline *I* 1.41. However, compared to *S* (1.69), *S+RA* (1.59) significantly improves the result by 0.1 in terms of D_{KL}. Meanwhile, when using the RA model,

the D_{KL} value of *(I+F+S)+RA* will drop 0.03, compared with *I+F+S*. These results prove that the RA model has the ability to improve the gaze prediction accuracy on top of existing methods.

V. Conclusion

In this paper, we propose to use attention to improve the driver gaze prediction task. Based on existing gaze prediction approaches, we use attention to regulate the initial predicted results to obtain more concentrated and accurate gaze maps. Specifically, we propose a reinforcement learning based model, termed Reinforced Attention (RA), for attention localization. RA is able to produce competitive localization accuracy while only processing a small subset of the video. When training RA, we also feed RA with speed and course details, which are proven to be indispensable. Experiments on the complete test set and the acting set of DR(eye)VE show that our method yields consistent improvement over several baselines and compares favorably with the state-of-the-art approaches.

Acknowledgment

This study is partially supported by the National Key R&D Program of China(No.2019YFB2101600), the National Natural Science Foundation of China(No.61861166002, 61872025, 61635002) , the Science and Technology Development Fund, Macau SAR (File no.0001/2018/AFJ), the Fundamental Research Funds for the Central Universities and the Open Fund of the State Key Laboratory of Software Development Environment(No. SKLSDE2019ZX-04). Thank you for the support from HAWKEYE Group.

References

[1] S. Habenicht, H. Winner, S. Bone, F. Sasse, and P. Korzenietz, "A maneuver-based lane change assistance system," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, 2011, pp. 375–380.

[2] L. Chen, Q. Li, M. Li, and Q. Mao, "Traffic sign detection and recognition for intelligent vehicle," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, 2011, pp. 908–913.

[3] K.-H. Lee and J.-N. Hwang, "On-road pedestrian tracking across multiple driving recorders," *IEEE Transactions on Multimedia*, vol. 17, no. 9, pp. 1429–1438, 2015.

[4] A. Palazzi, F. Solera, S. Calderara, S. Alletto, and R. Cucchiara, "Where should you attend while driving?" *arXiv preprint arXiv:1611.08215*, 2016.

[5] 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*, 2016, pp. 54–60.

[6] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, no. CONF, 2009, pp. 1597–1604.

[7] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2014.

[8] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proceedings of the ACM International Conference on Multimedia*, 2003, pp. 374–381.

[9] A. Palazzi, D. Abati, F. Solera, R. Cucchiara, *et al.*, "Predicting the driver's focus of attention: the dr (eye) ve project," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1720–1733, 2018.

[10] S. Mathe and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 7, pp. 1408–1424, 2014.

[11] S.-h. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren, "Video saliency detection via dynamic consistent spatio-temporal attention modelling," in *Twenty-seventh AAAI Conference on Artificial Intelligence*, 2013.

[12] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[13] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proceedings of the Advances in Neural Information Processing Systems*, 2000, pp. 1057–1063.

[14] Y. Feng, G. Cheung, W.-t. Tan, P. Le Callet, and Y. Ji, "Low-cost eye gaze prediction system for interactive networked video streaming," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1865–1879, 2013.

[15] S. J. Davies, D. Agrafiotis, C. N. Canagarajah, and D. R. Bull, "A multicue bayesian state estimator for gaze prediction in open signed video," *IEEE Transactions on Multimedia*, vol. 11, no. 1, pp. 39–48, 2008.

[16] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6912–6921.

[17] T. Fischer, H. Jin Chang, and Y. Demiris, "Rt-gene: Real-time eye gaze estimation in natural environments," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 334–352.

[18] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2176–2184.

[19] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4511–4520.

[20] N. Li, J. J. Jain, and C. Busso, "Modeling of driver behavior in real world scenarios using multiple noninvasive sensors," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1213–1225, 2013.

[21] L. Simon, J.-P. Tarel, and R. Brémond, "Alerting the drivers about road signs with poor visual saliency," in *Proceedings of the IEEE Intelligent Vehicles Symposium*. IEEE, 2009, pp. 48–53.

[22] R. Bremond, J. M. Auberlet, V. Cavallo, L. Desire, V. Faure, S. Lemonnier, R. Lobjois, and J. P. Tarel, "Where we look when we drive: A multidisciplinary approach," 2014.

[23] N. Pugeault and R. Bowden, "How much of driving is preattentive?" *IEEE Transactions on Vehicular Technology*, vol. 64, no. 12, pp. 5424–5438, 2015.

[24] K. Lv, H. Sheng, Z. Xiong, W. Li, and L. Zheng, "Pose-based view synthesis for vehicles: A perspective aware method," *IEEE Transactions on Image Processing*, vol. 29, pp. 5163–5174, 2020.

[25] H. Sheng, Y. Zheng, W. Ke, D. Yu, X. Cheng, W. Lyu, and Z. Xiong, "Mining hard samples globally and efficiently for person re-identification," *IEEE Internet of Things Journal*, pp. 1–1, 2020.

[26] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528–2536, 2016.

[27] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[28] Y. Zhang, X. Qian, X. Tan, J. Han, and Y. Tang, "Sketch-based image retrieval by salient contour reinforcement," *IEEE Transactions on Multimedia*, vol. 18, no. 8, pp. 1604–1615, 2016.

[29] J. C. Caicedo and S. Lazebnik, "Active object localization with deep reinforcement learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2488–2496.

[30] N. Xu, H. Zhang, A. Liu, W. Nie, Y. Su, J. Nie, and Y. Zhang, "Multi-level policy and reward-based deep reinforcement learning framework for image captioning," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1372–1383, 2020.

[31] G. Song, H. Myeong, and K. Mu Lee, "Seednet: Automatic seed generation with deep reinforcement learning for robust interactive segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1760–1768.

[32] S. Minut and S. Mahadevan, "A reinforcement learning model of selective visual attention," in *Proceedings of the ACM International Conference on Autonomous Agents*, 2001, pp. 457–464.
[33] V. Mnih, N. Heess, A. Graves, *et al.*, "Recurrent models of visual attention," in *Proceedings of the Advances in Neural Information Processing Systems*, 2014, pp. 2204–2212.
[34] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *Proceedings of the International Conference on Machine Learning*, 2016, pp. 1995–2003.
[35] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang, "Predicting head movement in panoramic video: A deep reinforcement learning approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
[36] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proceedings of the International Conference on Machine Learning*, 2016, pp. 1928–1937.
[37] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.
[38] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.
[39] M. X. Huang, J. Li, G. Ngai, and H. V. Leong, "Screenglint: Practical, in-situ gaze estimation on smartphones," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2017, pp. 2546–2557.
[40] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2018.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[42] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3395–3402.
[43] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4185–4196, 2015.
[44] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in *Proceedings of the International Conference on Pattern Recognition*, 2016, pp. 3488–3493.

Kai Lv received the B.S. degree from the School of Computer Science and Technology, Tianjin University of Science and Technology, Tianjin, China, in 2012. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Beihang University, Beijing, China.

Hao Sheng received his B.S. and Ph.D. degrees from the School of Computer Science and Engineering of Beihang University in 2003 and 2009, respectively. Now he is an associate professor in the School of Computer Science and Engineering, Beihang University, China. He is working on computer vision, pattern recognition and machine learning. He is the corresponding author of this paper.

Zhang Xiong received his B.S. degree from Harbin Engineering University in 1982. He received his M.S. degree from Beihang University in 1985. He is a professor and Ph.D. supervisor in the School of Computer Science and Engineering, Beihang University, China. He is working on computer vision, information security and data vitalization.

Wei Li is a computer scientist. He graduated from the Department of Mathematics and Mechanics at Peking University in 1966 and has been teaching at Beihang University (formerly Beijing Institute of Aeronautics) since then. After four years of graduate study at the University of Edinburgh, he obtained his Ph.D. degree in computer science there in 1983. He has been Professor in the School of Computer Science and Engineering at Beihang University since 1986 and served as President of Beihang University from 2002 to 2009. He was elected as a member of the Chinese Academy of Sciences in 1997. Currently, he serves as Director of the State Key Lab of Software Development Environment, Member of the National Educational Advisory Committee.

Liang Zheng is a Lecturer and a Computer Science Futures Fellow in the Research School of Computer Science, Australian National University. He received the Ph.D. degree in Electronic Engineering from Tsinghua University, China, in 2015, and the B.E. degree in Life Science from Tsinghua University, China, in 2010. He was a postdoc researcher in the Centre for Artificial Intelligence, University of Technology Sydney, Australia. His research interests include image retrieval, classification, and person re-identification.