IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 November 2023, accepted 29 November 2023, date of publication 5 December 2023,
date of current version 15 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3339855

RESEARCH ARTICLE

Accident Prediction Model Using Divergence Between Visual Attention and Focus of Expansion in Vehicle-Mounted Camera Images

YUTA MARUYAMA AND GOSUKE OHASHI, (Member, IEEE)
Department of Electrical and Electronic Engineering, Shizuoka University, Hamamatsu, Shizuoka 432-8561, Japan
Corresponding author: Yuta Maruyama (maruyama.yuta.18@shizuoka.ac.jp)

This work was supported by the Suzuki Foundation Science and Technology Research Grant.

ABSTRACT Recently, accident prediction models, which predict the occurrence of traffic accidents through deep learning algorithms have been proposed. The application of these models demands both high precision and visualization of the decision basis applied. Current models use the motion features of objects in the surrounding environment, but they do not predict well when the motion feature of the risk factor is small. Meanwhile, drivers can avoid accidents because they utilize visual attention functions. This study focuses on the divergence between visual attention and focus of expansion (FOE), which are highly correlated in normal driving situations, as the basis for an accident prediction method. The proposed model can visualize decision basis with high accuracy, even when the motion feature of the risk factors is small, by combining it with Dynamic-Spatial-Attention, a deep-learning-based accident prediction method. In this experiment, we classified data from the Dashcam Accident Dataset, a widely used accident dataset, into categories of accidents. Using the Dashcam Accident Dataset, the proposed method achieves higher accident prediction performance in categories for which the motion feature of risk factors tends to be small while maintaining the same accident prediction performance as achieved by the baseline Dynamic-Spatial-Attention method in categories for which the motion feature of risk factors tends to be large. In addition, the proposed method visualizes the risk factors using visual attention and FOE to provide a visual explanation of the decision basis.

INDEX TERMS Deep learning, driver visual attention, focus of expansion, traffic accident prediction.

I. INTRODUCTION

Systems that improve driving safety are essential for drivers and automated driving technology. In automated driving technology, accurately and swiftly predicting the occurrence of future traffic accidents as well as explaining the decision basis to users, contribute to assisting drivers and promoting the practical application of automatic driving. Recently, deep learning has demonstrated great results in various fields such as image recognition. Therefore, accident prediction models that acquire images from widely used in vehicle-mounted cameras and use deep learning to predict the occurrence of accidents have been studied [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27]. Fig. 1 shows an overview of accident prediction tasks using accident prediction models. Here, the occurrence probability of an accident is estimated by an accident prediction model using images from in-vehicle cameras as inputs. Many recent high-precision accident prediction models utilize both object and motion features [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. These methods extract object features limited to pedestrians and vehicles, which are the main candidates for risk factors through object detection. Then, the accident probability is estimated by evaluating the abnormal motion from motion features using optical flow and Long Short Term Memory(LSTM) neural networks [28]. In addition, some models such as Dynamic-Spatial-Attention(DSA) [1], [5], [6], [7], [8], [9], [12], [13], [14] explain the decision basis of the occurrence of

The associate editor coordinating the review of this manuscript and approving it for publication was Thomas Canhao Xu.

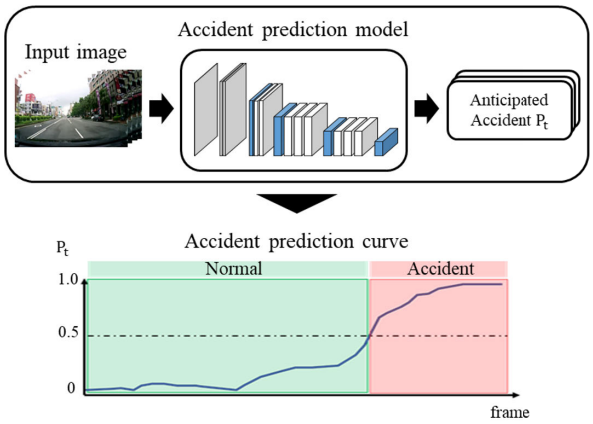

FIGURE 1. Overview of accident prediction task. The accident prediction model takes in vehicle-mounted camera images as input and outputs the accident probability for each frame. The vertical axis of the graph shows the accident probability P_t estimated by the model, and the horizontal axis shows the frame number. The horizontal dotted line shows a threshold value of 0.5. Frames with a probability of accident occurrence higher than the threshold are shown in red.

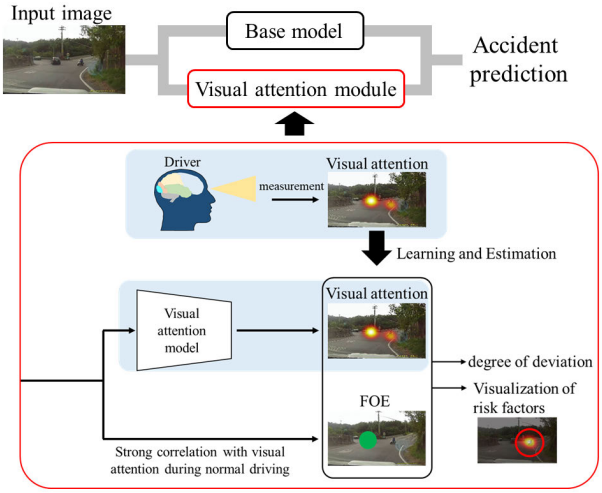

FIGURE 2. Overview of the proposed method. The proposed method incorporates a visual attention module consisting of visual attention and FOE into the base accident prediction model. The driver visual attention can be estimated by a visual attention model trained using eye tracking data measured by an eye tracker.

an accident by calculating the hazard level for each bounding box and visualizing the risk factors. However, these methods are difficult to estimate when the risk factor is an unexpected object that is in an unknown class for object detection, such as a falling object from a vehicle in front. In addition, a motion feature is not valid if the motion feature of the corresponding risk factor is difficult to obtain. For example, in the collision course phenomenon [29], [30], the apparent movement of the risk factor is affected by relative factors between vehicles, such as angle and speed. As similar phenomena can occur in crossing collision and curve scenes, models that evaluate the danger from such motion features cannot predict well [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23]. Therefore, it is not possible to encourage drivers and automatic controls to take actions to prevent extensive damage.

Meanwhile, the visual attention of the driver is focused on the risk factor when the motion feature is small, regardless of whether or not the class of object detection is known [31], [32], [33]. This is because, unlike bottom-up current models based on low-dimensional image features, driver visual attention has top-down properties that are based on driving experiences. This visual function makes avoiding accidents possible, suggesting that driver visual attention is useful for accident prediction tasks. When using visual attention for accident prediction tasks, it is difficult to measure the driver visual attention in real-time with an eye tracker because of the adaptation to automated driving and the cost of installing an eye tracker. Therefore, we use a visual attention model that can estimate the driver visual attention from in-vehicle camera images by using actual driver gaze data for training. Estimated driver visual attention is strongly correlated with the focus of expansion (FOE) during normal driving [34], [35], [36]. However, during abnormal events, visual attention is likely to diverge from FOE because visual attention is directed toward risk factors.

In this study, we propose an accident prediction model that incorporates visual attention and FOE divergence as shown in Fig. 2. Furthermore, even in a highly accurate and open-source DRIVE [27] model, such as that based on the saliency map, it is impossible to explicitly show the risk factors as the predicted saliency map itself is used. However, the proposed model visualizes the risk factors by using visual attention and FOE, enabling the visual explanation of the decision basis of the model. Visualization of the decision basis is an important element for the practical application of accident prediction models because it reduces the black box behavior of deep learning models. The contributions of this study are as follows.

1) By a visual attention model learned with driver gaze data, top-down knowledge of the driver is utilized in an accident prediction model.
2) An accident prediction model is proposed that incorporates the divergence between driver visual attention and FOE and enables the prediction of accidents difficult to estimate by only object and motion features.
3) The proposed accident prediction method achieves high accuracy in accident prediction using the Dashcam Accident Dataset, including accidents such as rear-end collision, head-on collision, turn, and crossing collision.
4) The visualization of risk factors from visual attention and FOE can provide a visual explanation of the decision basis for accident prediction.

II. RELATED WORKS

A. ACCIDENT PREDICTION MODEL WITH DEEP LEARNING

Deep learning models applied to accident prediction can be divided into four categories. The first category represents models based on the motion features of objects [17], [18],

[19], [20], [21], [22], [23]. The second category represents models based on multimodal data consisting of images, sensor data, and voice information [24], [25], [26]. The third category is saliency-based models [27]. The fourth category represents models based on both object and motion features [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16].

1) MODELS BASED ON MOTION FEATURES

This model uses motion features of objects, capturing abnormal motion by extracting motion features from time-series data, using optical flow or LSTM to predict accidents. Anopred [17] uses optical flow and U-Net [37] to predict future frames. Anomalies are detected by comparing the predicted future's frame object motion with the actual motion. Kataoka et al. [18] used semantic flow to separate the background for risk factor recognition. CST-S3D [19] extracts motion features based on 3D CNN [38] that can process spatio-temporal information on training data augmented by an image transformation model. SP [20] uses LSTM to account for longer-term temporal relationships. This mechanism learns the general movements of pedestrians and predicts their future trajectories. ConvLSTMAE [21] also uses LSTM to extract motion features and detect anomalies using reconstruction errors with the input image obtained by an autoencoder. Karim et al. [22] also used a recurrent neural network and proposed a model based on Gated Recurrent Unit (GRU) [39], which is more computationally efficient than LSTM. AdaLEA [23] employed a quasi-recurrent neural network [40], which can learn faster than LSTM by parallelizing the computational process to capture spatiotemporal relationships. The method introduces Adaptive Loss, which adaptively changes the penalty weights during each epoch of training. Although these methods can extract abnormal risk factors from motion features, they cannot estimate accident risk when the motion features of the risk factors are small. In addition, they do not focus on the major risk factors such as vehicles and pedestrians. Therefore, in recent years, object detection has been used to detect vehicles and pedestrians and to recognize risk factors based on the relationships among objects.

2) MODELS BASED ON MULTIMODAL DATA

Models using multimodal data are those that use not only video images but also audio information and data obtained from various sensors. Yamamoto et al. [24] proposed a method that combines sensor data, such as acceleration and speed, with video images. Tanno et al. [25] also proposed a configuration with three streams that extract time-series multimodal data consisting of voice information and sensor data, such as moving images and speed. This study confirms the improvement in accuracy using voice information and shows the effectiveness of this method. Monjuru et al. [26] proposed a model that uses textual data to handle cases in which extracting features from moving images is difficult such as, at night. However, the above-mentioned models have not yet been combined with methods that show rational decision bases, such as the visualization of risk factors.

3) MODELS BASED ON SALIENCY MAP

A saliency-based model is a model that uses saliency maps showing the salient regions in an image. DRIVE [27] is an open source and highly accurate accident prediction model. It uses reinforcement learning to estimate the probability of an accident occurring at each time point, while simultaneously predicting the prominence of the next frame and providing feedback. Depending on the estimated value, the probability of the occurrence of an accident in the next frame is predicted, and the weights utilized in the predicted saliency are changed. However, many datasets that include accident scenes exclude annotated gaze data. In such cases, the entire model cannot be trained properly because part of the reward in reinforcement learning is removed. Therefore, the estimated probability of accidents may be high even in normal operation scenarios in which no accidents occur, which may degrade the accuracy of the prediction. In addition, because a saliency map is used for the visualization of the decision basis, it is the output of the image not only in accident scenes but also in normal driving scenes. Therefore, the visualized maps cannot always indicate the risk factors, making interpretation of the decision basis difficult. Hence, the accident prediction model is required to visualize only the areas judged to be risk factors.

4) MODELS BASED ON MOTION AND OBJECT FEATURES

Models using both object and motion features extract object features, through object detection, and estimate the probability of accidents by acquiring motion features through optical flow and recurrent neural networks. DSA [1], CDAP [3], L-RAI [5], DSTA [6], Yamamoto et al. [7], and FA [16] have proposed models that process object detection results in conjunction with LSTM and other recurrent neural networks to extract location relationships among surrounding objects. Ustring [2] uses graph convolution (GNN) to clearly adapt the distance between objects to the edges in the graph convolution for object-position extraction. Ichiki et al. [12] used features such as dynamic obstacle presence and static road information by combining semantic segmentation and object detection. Object detection is also used in FRPN [15] and Zhou et al. [14], which use changes in the size of the detected bounding box and shifts in the center of gravity. SSC [13] proposed an unsupervised accident prediction method using the predictions of object movement and whole frames. FOL-Ensemble [8], AM-Net [9], MAMTCF [10], and THAT-Net [11] are models that utilize optical flow. AM-Net, MAMTCF, and THAT-Net generate object-level flow images by using the center coordinates of the object's bounding box. FOL-Ensemble uses optical flow in the same manner but predicts the position of the next frame from the optical flow. This position information is used to calculate the probability of an accident occurrence. Some models such

as DSA [1], [5], [6], [7], [8], [9], [12], [13], [14] calculate the risk per bounding box to visualize risk factors by using the bounding box with the highest risk. However, when using object features, if the risk factor is an unexpected object such as a falling object from a vehicle in front of the user, it may be an example of out-of-class data for object detection, making estimation difficult. In addition, when using motion features, there are cases in traffic scenes in which it is difficult to obtain the motion features of risk factors. One example is the collision course phenomenon, in which the apparent movement of a risk factor is reduced by relative factors such as the angle and speed between vehicles. These models, which use motion features to determine hazards, cannot predict accidents well because similar phenomena can occur in crossing collisions, curves, and other situations. Therefore, there is a need for an accident prediction model that can visualize the basis of a decision with high accuracy, even for risk factors with small movement characteristics.

B. VISUAL ATTENTION MODEL WITH DEEP LEARNING

A model for estimating human visual attention from input images is called a visual attention model. A variety of methods have been proposed since the preliminary study on visual attention models by Itti et al. [41] in 1998. In recent years, deep learning based visual attention models have been proposed [31], [34], [42], [43], [44], [45], [46], [47], [48], [49], [50], [52], [53], [54], [55], [56], [57], [59]. These models use human gaze data measured by an eye tracker for training and estimate visual attention based on factors such as depth, context, and flicker in the image. UNISAL [46] uses domain adaptation to predict visual attention with a unified model for different types of datasets of still and moving images. The model also uses a recurrent neural network, which is not capable of simultaneously encoding both spatial and temporal information. Therefore, methods [42], [45], ViNet [47], HD^2S [48], and STSANet [49], which can simultaneously process spatio-temporal information based on 3D CNN, have been implemented. VSFT [50] incorporates Transformer [51] into the model structure to consider longer-term spatio-temporal dependencies compared with 3D CNN.

Moreover, top-down characteristics, such as those based on task and experience, are important in estimating the visual attention of a driver. Therefore, these proposed methods [34], [52], [53], [54], [55], [56], [57], [59] for driving tasks are trained on the driver gaze data. BDD-A [31] proposed Human Weighted Sampling, a method in which the sampling rate is varied according to the degree of separation between the average map of the driver visual attention, which is the correct image during training, and the visual attention in each frame. This allows the model to learn effective visual attention to accident scenes by identifying important frames in the dataset. DR(eye)VE [34] applied semantic segmentation to explicitly extract the respective relationships among people, vehicles, and roads. Visual attention is estimated by adding the output of branches in RGB images, semantic segmentation images, and images showing optical flow. Meanwhile, SCAFNet [54] and Rui et al. [55] proposed a method of concatenating features obtained from segmentation images using Convolution LSTM to take advantage of features obtained from 3D CNN. Watanabe et al. [57] created a predictive model that reproduces human vision on the basis of PredNet [58], which incorporates predictive coding. Using this predictive model, Seki et al. [33] demonstrated the characteristics of gaze regarding potential dangers during driving. ARAGAN [59] proposed a method that combines Conditional Generative Adversarial Network [60] and Multi-Head Attention algorithms to generate a driver visual attention map from input RGB images. In this study, by using a visual attention model learned from driver gaze data, we apply driver top-down knowledge to the accident prediction model.

III. METHOD

The proposed method consists of a base model and visual attention module. Using DSA, a highly accurate open source accident prediction model as a base model, we calculate the divergence between visual attention and FOE in the visual attention module. The outputs are combined to calculate the probability of accident occurrence for the input image. The model structure of the proposed method is shown in Fig. 3.

A. BASE MODEL (DYNAMIC SPATIAL ATTENTION)

The proposed method uses DSA as a base model to predict accidents from object and motion features. We use Faster R-CNN [61], which was pre-trained on the KITTI dataset [62] to extract object features, such as vehicles and pedestrians. Faster R-CNN is a typical end-to-end object detection model. Vehicles and pedestrians in the input image are detected by determining the location and rectangular shape of the object using the region proposal network [61] and classifying its class. Each of the J objects $\hat{x}_t^j$ detected at time t is assigned a weight α_t^j for accident prediction. The α_t^j is computed from the output of the hidden layer at time $t-1$, $\hat{x}_t^j$, and e_t, which consists of several model parameters. The calculation formulas for α_t^j and e_t^j are as follows.

$$\alpha_t^j = \frac{exp(e_t^j)}{\sum_j exp(e_t^j)} \tag{1}$$

$$e_t^j = \omega^T \rho(\omega_e h_{t-1} + U_e \hat{x}_t^j + b_e) \tag{2}$$

ω, ω_e, U_e, b_e are learning parameters. ρ indicates the tanh function, which improves the expressive power of the model by performing nonlinear transformation. h_{t-1} is calculated at the output gate of the RNN in the previous frame and has a size of 10×512. Then, at each time step, the weight α_t^j for each object is embedded into the Holistic Feat X_t, which is a set of $\hat{x}_t^j$ at each time, to obtain $\phi(X_t, \alpha_t)$. The formula for

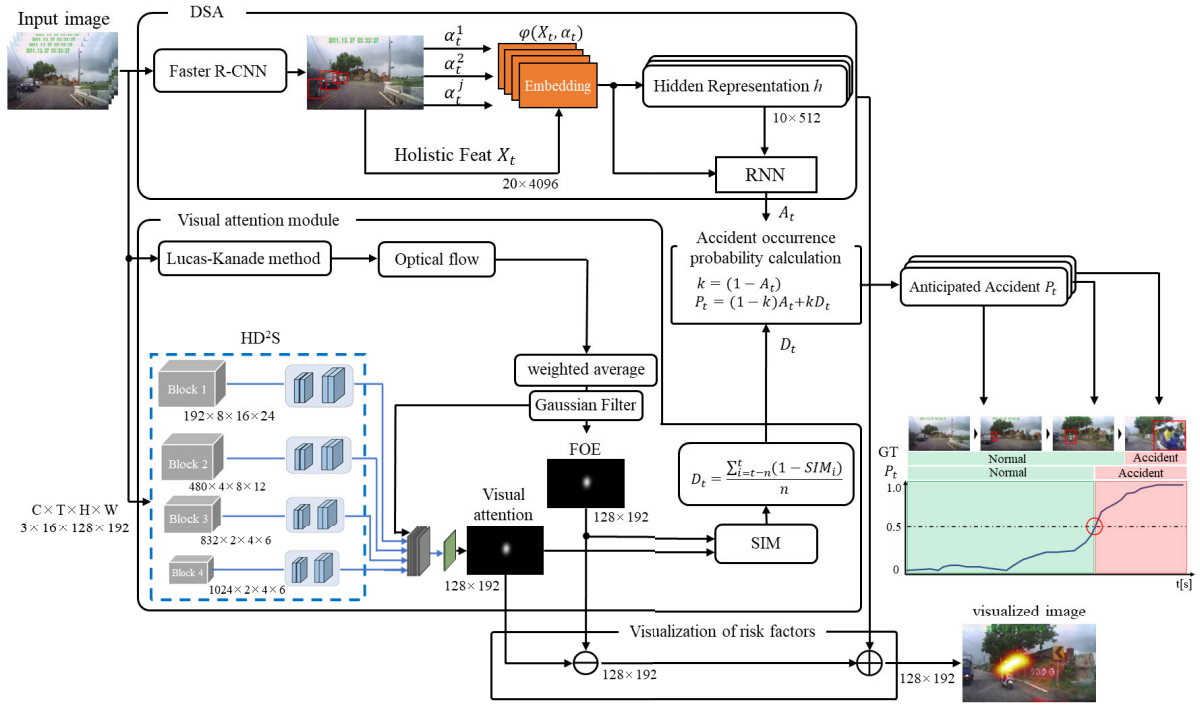

FIGURE 3. Model structure of the proposed method. The proposed method consists of "DSA", "Visual attention module", and "Accident occurrence probability calculation" and outputs the accident occurrence probability P_t. A graph is created using the probability of accident occurrence P_t for each frame. GT in the graph shows the frames after the collision in red.

embedding is as follows.

$$\phi(X_t, \alpha_t) = \sum_{j=1}^{J} \alpha_t^j \hat{x}_t^j \tag{3}$$

$\phi(X_t, \alpha_t)$ uses RNN to calculate A_t at time t based on the weights of detected objects such as vehicles and pedestrians.

B. VISUAL ATTENTION MODULE

1) VISUAL ATTENTION MODEL

The proposed method utilizes a visual attention model to predict the probability of occurrence of accidents for risk factors with small image variation. Highly accurate and open source visual attention models include ViNet [47], HD^2S [48], and STSANet [49]. However, STSANet uses images up to frame $t + 16$ to predict visual attention at frame t. Therefore, it is inappropriate to incorporate it into an accident prediction model. In addition, because HD^2S has a smaller model size than ViNet, we useÂ a visual attention model [36] that composed of HD^2S. HD^2S estimates visual attention by combining the outputs of four streams encoded by 3D convolutional layers for each level of abstraction. For pre-training of HD^2S, we use the BDD-A dataset [31], which includes dangerous traffic scenes among the driver's gaze datasets. Through this, we obtain top-down visual attention according to the traffic accident prediction task. The estimated visual attention is output as a 128 × 192 grayscale image.

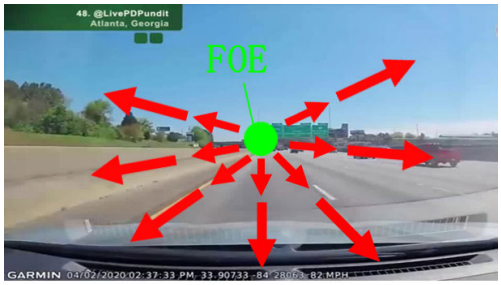

FIGURE 4. FOE and optical flow. This shows FOE and optical flow in the in-vehicle camera images. Optical flow is indicated by the direction and size of the red arrow. Also, FOE is shown as a green circle.

2) FOE (FOCUS OF EXPANSION)

FOE is defined as the origin of the optical flow [63], [64]. Fig. 4 shows FOE and optical flow in an in-vehicle camera image. As shown in Fig. 4, the norm of the optical flow increases with separation from FOE. Therefore, FOE obtained by the weighted average, which increases the weights of the optical flow norm, is extended to a two-dimensional image distribution by the Gaussian filter. The calculation formula for the x-coordinate of FOE is shown below. Also, calculate the y coordinate using the same formula.

$$FOE_x = \frac{\sum \omega_1 x_1 + \sum \omega_2 x_2 + \sum \omega_3 x_3}{\sum \omega_1 + \sum \omega_2 + \sum \omega_3} \tag{4}$$

x_1 and x_2 represent the coordinates of the starting point of the 10% and 20% optical flow with small norms, and x_3 represents the coordinates other than the above. In this paper, we set ω_3=1, ω_2=2, and ω_1=3 because the smaller the norm, the larger the weight. There are two methods for determining optical flow: dense optical flow [65], [66] and sparse optical flow [67]. In the sparse optical flow calculation method, optical flow is calculated only for pixels where feature points in the image can be obtained. On the other hand, the method that calculates dense optical flow calculates it for all pixels. Therefore, it is calculated even in areas where the norm of optical flow is close to 0, such as the sky. In this paper, we calculate FOE using optical flow with a small norm. Hence, optical flow in areas such as the sky may become noise. Here, we use Lucas-Kanade method [67], which is the method for finding sparse optical flow. The estimated FOE is output as a 128 × 192 grayscale image.

3) CALCULATION OF DIVERGENCE

The divergence between the obtained visual attention and FOE is calculated as presented here. SIM, an index for evaluating the overlap of histograms, is used for the divergence. In SIM, the distribution is normalized so that the total value is 1, and the sum of the minimum values of each corresponding pixel i is calculated. The definition formula for SIM is as follows.

$$SIM(mapA, mapB) = \sum_i min(mapA_i, mapB_i) \tag{5}$$

$$where \sum_i mapA_i = \sum_i mapB_i = 1 \tag{6}$$

SIM indicates that the distributions are perfectly matched when it is 1, and 0 indicates that there is no overlap in the distributions. Therefore, the value obtained by subtracting SIM from 1 is used as the divergence. The equation for calculating divergence D_t is shown below. Since this paper uses real images, FOE may fluctuate as the vehicle shakes, and the SIM may be affected. Therefore, to absorb the fluctuation, we set $n = 5$ and take the average of the previous 5 frames.

$$D_t = \frac{\sum_{i=t-n}^{t}(1 - SIM_i)}{n} \tag{7}$$

C. ACCIDENT OCCURRENCE PROBABILITY CALCULATION

Using the resulting A_t from the base model, the final accident probability P_t is calculated using A_t and D_t, as shown below.

$$k = (1 - A_t) \quad P_t = (1 - k)A_t + kD_t \tag{8}$$

When the output from the base model is low, the coefficient k increases the weight of the output from the visual attention module, allowing the treatment of accidents that are difficult to estimate from only object and motion features. For normal driving scenes, the divergence between visual attention and FOE is low, allowing the calculation of accident probability without over-detection. Let P_t be the probability of accident occurrence, and use the learning process of DSA, which is the base model. The loss used for learning in the accident prediction model makes the penalty for failure in prediction in a frame close to the accident larger than that in the case of prediction in a frame far from the accident. In addition, cross-entropy error is used in scenes where no accidents occur. Therefore, set each loss as follows.

$$L_p = \sum_t -e^{-max(0,y-t)} log(P_t) \tag{9}$$

$$L_n = \sum_t -log(1 - P_t) \tag{10}$$

Here, L_p indicates the loss in the positive scene where the accident occurred. Let y be the time of occurrence of the accident. In addition, L_n indicates the loss in a negative scene where no accident occurs. Finally, calculate the sum of each loss.

$$\sum_{j \in P} L_p^j + \sum_{j \in N} L_n^j \tag{11}$$

Here, let P be the set of positive scenes and N be the set of negative scenes.

TABLE 1. Experimental conditions.

Train dataset		DAD
Train data number	Rear end collision	5400 frames
	Head on collision	1900 frames
	Turn	2600 frames
	Crossing collision	5800 frames
	Normal	29800 frames
Test dataset		DAD
Test data number	Rear end collision	900 frames
	Head on collision	600 frames
	Turn	700 frames
	Crossing collision	2600 frames
	Normal	3300 frames
Comparative approach		DSA, DRIVE
Valuation index		F1, TTA, FT
Learning rate		0.0001
Epoch number		30
Batch size		10
Optimizer		Adam
Input image size		128×192
CPU		Intel Core(TM)i9-9900K
CPU memory		32 GB
GPU		NVIDIA GeForce RTX 2080 Ti
GPU memory		11GB

D. VISUALIZATION OF RISK FACTORS

The proposed accident prediction model enables visual explanation through the visualization of only risk factors. Risk factors are visualized using the output from the base-model DSA, visual attention, and FOE. Visual attention and FOE are output as two-dimensional image distributions to compute the difference of the image. In addition, the bounding box coordinates of the highly hazardous object obtained by DSA object detection are extended to the distribution by the Gaussian filter. By visualizing the risk

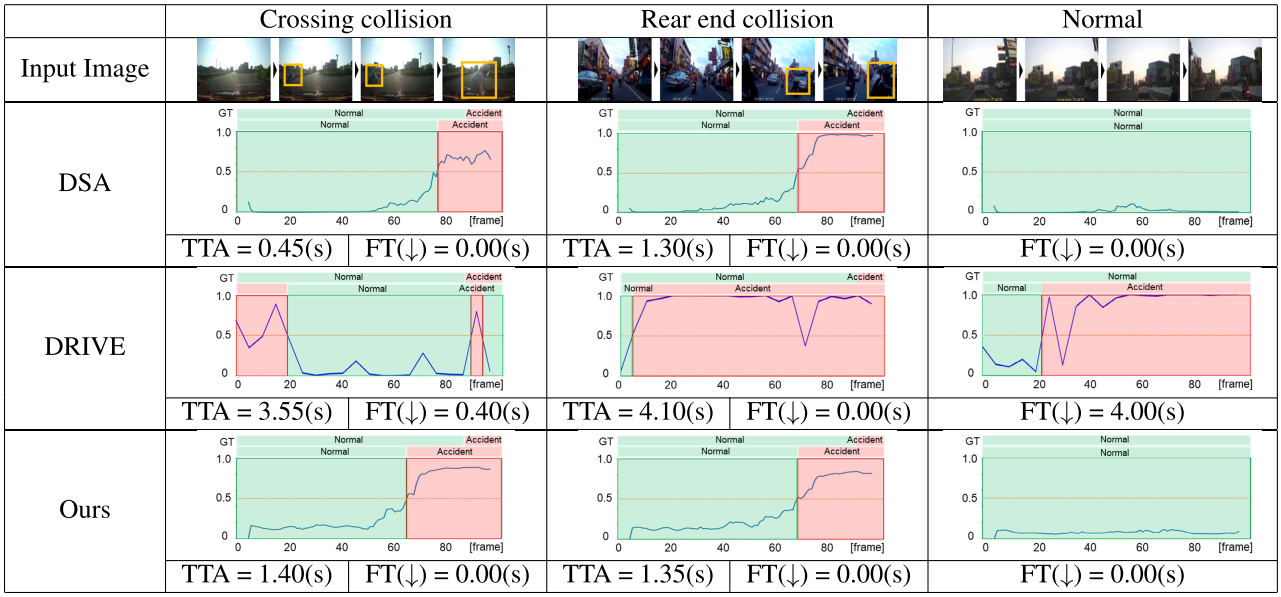

FIGURE 5. Accident prediction curves in DAD. Risk factors in the input image are shown in yellow. The two on the left show accident scenes, and the third from the left is a normal scene. In each graph, the vertical axis shows the estimated accident prediction probability, and the horizontal axis shows the frame number. In addition, FT is the estimation failure time. TTA is the difference between time t_1 when the predicted accident probability exceeds the threshold and time t_2 when the accident occurs.

factors from adding these two images, a heat map is created for the risk factor in the frame where the risk factors are present. There is no heat map created in the typical usual operation scene without the risk factor. Visualization of solely risk factors allows for the rational explanation of the basis for model decisions.

IV. EXPERIMENTS

To verify the effectiveness of the proposed model, categories were generated representing each type of accident, and accident prediction experiments were conducted.

A. DATA SET

For the experiments, a widely used dataset containing accident scenes named DAD [1] is used. The DAD consists of various accident scenes involving cars, pedestrians, and motorcycles, which are filmed with in-vehicle cameras such as drive recorders and published on the website. The resolution of the dataset is 720 × 1280, and the frame rate is fixed at 20 fps. In this study, accident scenes from datasets are classified into four categories. In the categories ''Rear end collision'' and ''Head on collision,'' it is assumed that risk factors exist in the center of images. When the given vehicle moves, the flow of the risk factor becomes relatively large in the direction of motion, because the optical flow of the peripheral background is small. Therefore, these categories are assumed to be accidental scenes in which the image variation of the risk factors tends to increase. In addition, in the cases of ''Turn'' and ''Crossing collision,'' considering that the motion features cancel each other out owing to the relative movement between the vehicles, these categories are assumed to be accident scenes in which the image variation of the risk factors tends to be small. In addition, the dataset includes some scenes in which vehicles are not moving, such as those captured by fixed-point cameras, and such scenes are excluded.

B. EXPERIMENTAL CONDITIONS

Experimental conditions are shown in Table 1. For the experiment, we used DAD [1], which is widely used as a dataset containing accident scenes categorized by accident category. The learning rate of the model is 0.0001, the number of epochs is 30, and the batch size is 10. Use Adam as the optimization function. These conditions are set so that the loss in learning can be sufficiently converged. DSA and DRIVE [27] are used as methods for comparison in the experiment. The CPU configuration is Intel Core i9-9900K CPU @ 3.60 GHz and the GPU configuration is NVIDIA GeForce RTX 2080Ti. Time to accident(TTA) is used as an evaluation index. TTA is the time range of risk perception before an accident occurs and is defined as the difference between the time t_1 when the predicted accident probability exceeds the threshold value and the time t_2 when the accident occurs. False Time(FT) is defined as the average of the time corresponding to false negatives in accident scenes and false positives in normal scenes, with a smaller value serving as an indicator of better results.

C. RESULTS

Estimated accident prediction curves from the Dashcam Accident Dataset are shown in Fig. 5. The horizontal axis is time, and the vertical axis is accident occurrence probability.

TABLE 2. Quantitative Evaluation in DAD. In each category, "Rear end collision" and "Head on collision" are categories that tend to have large motion features, and "Turn" and "Crossing collision" are categories that tend to have small motion features.

		DSA			DRIVE			Ours		
Motion feature	Accident category	F1	TTA	FT	F1	TTA	FT	F1	TTA	FT
Large	Rear end collision	**0.89**	1.03	0.37	0.24	**4.20**	1.55	0.82	0.87	**0.35**
	Head on collision	**0.50**	0.63	0.48	0.24	**4.33**	1.46	**0.50**	0.60	**0.38**
Small	Turn	0.77	0.42	0.40	0.22	**4.19**	1.55	**0.86**	0.51	**0.39**
	Crossing collision	0.77	0.45	0.74	0.21	**4.00**	1.59	**0.85**	0.77	**0.71**
Average		0.72	0.64	0.50	0.23	**4.18**	1.53	**0.77**	0.69	**0.46**

FIGURE 6. Visual description of the model in DAD. Risk factors for the input image are shown in yellow. The three on the left are accident scenes, and the three on the right are normal scenes. In the proposed method, only the areas determined to be risk factors are visualized using a heat map.

The period that exceeds a threshold of 0.5 (50%) in the accident prediction task is indicated in red. The proposed method can recognize the danger even in an accident scene in which the image variation of the risk factor tends to become small and confirm that the probability of an accident is low and predictable under normal conditions. A quantitative assessment is shown in Table 2. The proposed method exhibits the same performance as that of DSA in accident scenes where the image variation of the ''Rear end collision'' and ''Head in collision'' risk factors tend to be large, and improves by 12% in the F1 compared to DSA in accident scenes where the image variation of the ''Turn'' and ''Crossing collision'' risk factors tend to be small. In addition, the TTA is 0.5 [s] better than the TTA of the DSA, and the FT is better than those of both the DSA and DRIVE. Our proposed model performs 5% better than DSA and 49% better than DRIVE in terms of F1 for all test images. These results show that the proposed method can predict the probability of accidents with high speed and accuracy while maintaining predictability in categories for which the image variation of risk factors tends to be large and suppressing over-detection in categories for which the image variation of risk factors tends to be small. Fig. 6 shows the results of the visualization of the decision basis for the model. Risk factors in the input image are indicated by yellow boxes. In each accident scene, DRIVE can confirm risk factors such as cars and motorcycles, but it can also confirm vehicles traveling in front of the scene other than risk factors, such as the scenario in the third scene from the left. This can be considered as the gazing area during normal operation, and similar output can be seen in the three non-accident scenes shown on the right. This constant display of the human gazing area is not appropriate for visualizing the basis for decisions in the accident prediction task because it makes distinguishing risk factors impossible. Meanwhile, the proposed method provides output only for the risk factors that suddenly present themselves, such as the motorcycle in the first and third scenes from the left for accident scenes. Therefore, in the normal driving scene shown on the right, nothing is displayed because there are no risk factors. Although DSA also provides visualization for risk factors, it displays many bounding boxes containing features other than risk factors. These qualitative evaluations confirm that the proposed method adequately captures risk factors and provides reasonable visual explanations. We have discussed the issues with the proposed method. The proposed method has the same performance as the baseline in ''Rear end collision'' and ''Head on collision,'' and no major changes due to the visual attention module can be observed. This could be the case the vehicle, which is a risk factor, is likely to be located in the center of the screen. Therefore, because the visual attention and FOE regions overlap, no discrepancy occurs and each evaluation index has the same value.

V. CONCLUSION

This study proposes an accident prediction model based on DSA that predicts accidents based on object and motion features, combined with the divergence between visual attention and FOE. By applying the visual attention model

learned from the driver gaze data, the driver top-down knowledge is incorporated into the accident prediction model. The proposed method is applied to the DAD dataset and compared with the DSA and DRIVE. The results show the effectiveness in the metrics F1, TTA, and FT by confirming that it is possible to predict accidents with high accuracy for all accident scenes containing risk factors with small motion features. Visualization of the risk factors using differential images of visual attention and FOE demonstrates that a visual explanation of the basis for decision making is possible. These qualitative evaluations confirm that the proposed method adequately captures risk factors and provides reasonable visual explanations.

ACKNOWLEDGMENT

The purpose of the Suzuki Foundation is to fund scientific research on small cars and other machinery for people's daily lives in Japan.

REFERENCES

[1] F. H. Chan, Y. T. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *Proc. Asian Conf. Comput. Vis.*, Nov. 2016, pp. 136–153.
[2] W. Bao, Q. Yu, and Y. Kong, "Uncertainty-based traffic accident anticipation with spatio-temporal relational learning," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2682–2690.
[3] A. P. Shah, J.-B. Lamare, T. Nguyen-Anh, and A. Hauptmann, "CADP: A novel dataset for CCTV traffic camera based accident analysis," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–9.
[4] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 273–280.
[5] K.-H. Zeng, S.-H. Chou, F.-H. Chan, J. C. Niebles, and M. Sun, "Agent-centric risk assessment: Accident anticipation and risky region localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1330–1338.
[6] M. Monjurul Karim, Y. Li, R. Qin, and Z. Yin, "A dynamic spatial–temporal attention network for early anticipation of traffic accidents," 2021, *arXiv:2106.10197*.
[7] S. Yamamoto, T. Kurashima, and H. Toda, "Classifying near-miss traffic incidents through video, sensor, and object features," *IEICE Trans. Inf. Syst.*, vol. 105, no. 2, pp. 377–386, 2022.
[8] Y. Yao, X. Wang, M. Xu, Z. Pu, Y. Wang, E. Atkins, and D. J. Crandall, "DoTA: Unsupervised detection of traffic anomaly in driving videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 444–459, Jan. 2023.
[9] M. M. Karim, Z. Yin, and R. Qin, "An attention-guided multi stream feature fusion network for localization of risky objects in driving videos," *IEEE Trans. Intell. Veh.*, pp. 1–12, May 2023.
[10] R. Liang, Y. Li, Y. Yi, J. Zhou, and X. Li, "A memory-augmented multi-task collaborative framework for unsupervised traffic accident detection in driving videos," 2023, *arXiv:2307.14575*.
[11] W. Liu, T. Zhang, Y. Lu, J. Chen, and L. Wei, "THAT-Net: Two-layer hidden state aggregation based two-stream network for traffic accident prediction," *Inf. Sci.*, vol. 634, pp. 744–760, Jul. 2023.
[12] M. Ichiki, C. Miyajima, A. Carballo, and K. Takeda, "Detection of potentially risky driving scenes and identification of associated risk factors," in *Proc. Int. Symp. Future Act. Saf. Technol. Toward Zero Traffic Accidents (FAST-zero)*, Sep. 2021, pp. 1–5.
[13] J. Fang, J. Qiao, J. Bai, H. Yu, and J. Xue, "Traffic accident detection via self-supervised consistency learning in driving scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9601–9614, Jul. 2022.
[14] Z. Zhou, X. Dong, Z. Li, K. Yu, C. Ding, and Y. Yang, "Spatio-temporal feature encoding for traffic accident detection in VANET environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19772–19781, Oct. 2022.
[15] Y.-F. Zhou, K. Xie, X.-Y. Zhang, C. Wen, and J.-B. He, "Efficient traffic accident warning based on unsupervised prediction framework," *IEEE Access*, vol. 9, pp. 69100–69113, 2021.
[16] M. Fatima, M. U. Karim Khan, and C.-M. Kyung, "Global feature aggregation for accident anticipation," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 2809–2816.
[17] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545.
[18] H. Kataoka, T. Suzuki, S. Oikawa, Y. Matsui, and Y. Satoh, "Drive video analysis for the detection of traffic near-miss incidents," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 3421–3428.
[19] H. Pradana, M.-S. Dao, and K. Zettsu, "Augmenting ego-vehicle for traffic near-miss and accident classification dataset using manipulating conditional style translation," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2022, pp. 1–8.
[20] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 961–971.
[21] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Proc. Adv. Neural Netw. (ISNN)*, Jun. 2017, pp. 189–196.
[22] M. M. Karim, Y. Li, and R. Qin, "Toward explainable artificial intelligence for early anticipation of traffic accidents," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2676, no. 6, pp. 743–755, Jun. 2022.
[23] T. Suzuki, H. Kataoka, Y. Aoki, and Y. Satoh, "Anticipating traffic accidents with adaptive loss and large-scale incident DB," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3521–3529.
[24] S. Yamamoto, E. Yuki, and H. Toda, "Near miss accident detection method from drive recorder using video and sensor signals," *IPSJ Trans. Databases*, vol. 10, no. 4, pp. 26–30, 2017.
[25] R. Tanno, D. Ozawa, and K. Ito, "Multimodal deep learning techniques for hazardous driving scene extraction," in *Proc. Forum Data Eng. Inf. Manage. (DEIM)*, 2019, pp. 1–5.
[26] J. Fang, L.-L. Li, K. Yang, Z. Zheng, J. Xue, and T.-S. Chua, "Cognitive accident prediction in driving scenes: A multimodality benchmark," 2022, *arXiv:2212.09381*.
[27] W. Bao, Q. Yu, and Y. Kong, "DRIVE: Deep reinforced accident anticipation with visual explanation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7599–7608.
[28] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, Oct. 2000.
[29] N. Uchida, D. de Waard, and K. A. Brookhuis, "Countermeasures to prevent detection failure of a vehicle approaching on collision course," *Appl. Ergonom.*, vol. 42, no. 4, pp. 540–547, May 2011.
[30] T. Keiji, N. Iwaki, and S. Oyama, "Research on peripheral vision characteristics at intersections with good visibility," *IEEEJ Trans. Electron., Inf. Syst.*, vol. 131, no. 12, pp. 2148–2153, 2011.
[31] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney, "Predicting driver attention in critical situations," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Dec. 2018, pp. 658–674.
[32] B. Wolfe, B. D. Sawyer, and R. Rosenholtz, "Toward a theory of visual information acquisition in driving," *Hum. Factors, J. Human Factors Ergonom. Soc.*, vol. 64, no. 4, pp. 694–713, Jun. 2022.
[33] T. Seki, M. Amamiya, M. Kato, K. Emura, and E. Watanabe, "Analysis of driver's predictive characteristics using eye tracking and a deep learning model that mimics human vision," *Trans. Soc. Automot. Eng. Jpn.*, vol. 54, no. 3, pp. 628–634, 2023.
[34] A. Palazzi, D. Abati, S. Calderara, F. Solera, and R. Cucchiara, "Predicting the driver's focus of attention: The DR(eye)VE project," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1720–1733, Jul. 2019.
[35] I. R. Johnston, G. R. White, and R. W. Cumming, "The role of optical expansion patterns in locomotor control," *Amer. J. Psychol.*, vol. 86, no. 2, p. 311, Jun. 1973.
[36] Y. Maruyama and G. Ohashi, "Prediction model for gazing at in vehicle camera images focusing on FOE," in *Proc. Symp. Sens. Imag. Inf. (SSII)*, Jun. 2022, pp. 1–2.
[37] 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.*, Oct. 2015, pp. 234–241.
[38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[39] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.
[40] J. Bradbury, S. Merity, C. Xiong, and R. Socher, "Quasi-recurrent neural networks," 2016, *arXiv:1611.01576*.
[41] 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, Mar. 1998.
[42] K. Min and J. J. Corso, "TASED-Net: Temporally-aggregating spatial encoder–decoder network for video saliency detection," 2019, *arXiv:1908.05786*.
[43] C. Bak, A. Kocak, E. Erdem, and A. Erdem, "Spatio-temporal saliency networks for dynamic saliency prediction," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1688–1698, Jul. 2018.
[44] Q. Lai, W. Wang, H. Sun, and J. Shen, "Video saliency prediction using spatiotemporal residual attentive networks," *IEEE Trans. Image Process.*, vol. 29, pp. 1113–1126, 2020.
[45] Q. Chang and S. Zhu, "Temporal–spatial feature pyramid for video saliency detection," 2021, *arXiv:2105.04213*.
[46] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in *Proc. Int. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 419–435.
[47] S. Jain, P. Yarlagadda, S. Jyoti, S. Karthik, R. Subramanian, and V. Gandhi, "ViNet: Pushing the limits of visual modality for audio-visual saliency prediction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 3520–3527.
[48] G. Bellitto, F. Proietto Salanitri, S. Palazzo, F. Rundo, D. Giordano, and C. Spampinato, "Hierarchical domain-adapted feature learning for video saliency prediction," *Int. J. Comput. Vis.*, vol. 129, no. 12, pp. 3216–3232, Dec. 2021.
[49] Z. Wang, Z. Liu, G. Li, Y. Wang, T. Zhang, L. Xu, and J. Wang, "Spatio-temporal self-attention network for video saliency prediction," *IEEE Trans. Multimedia*, vol. 25, pp. 1161–1174, 2023.
[50] C. Ma, H. Sun, Y. Rao, J. Zhou, and J. Lu, "Video saliency forecasting transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6850–6862, Oct. 2022.
[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 5998–6008.
[52] T. Deng, H. Yan, L. Qin, T. Ngo, and B. S. Manjunath, "How do drivers allocate their potential attention? Driving fixation prediction via convolutional neural networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 2146–2154, May 2020.
[53] S. Baee, E. Pakdamanian, I. Kim, L. Feng, V. Ordonez, and L. Barnes, "MEDIRL: Predicting the visual attention of drivers via maximum entropy deep inverse reinforcement learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13158–13168.
[54] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "DADA: Driver attention prediction in driving accident scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4959–4971, Jun. 2022.
[55] R. Fu, T. Huang, M. Li, Q. Sun, and Y. Chen, "A multimodal deep neural network for prediction of the driver's focus of attention based on anthropomorphic attention mechanism and prior knowledge," *Expert Syst. Appl.*, vol. 214, Mar. 2023, Art. no. 119157.
[56] C. Gou, Y. Zhou, and D. Li, "Driver attention prediction based on convolution and transformers," *J. Supercomput.*, vol. 78, no. 6, pp. 8268–8284, Apr. 2022.
[57] E. Watanabe, A. Kitaoka, K. Sakamoto, M. Yasugi, and K. Tanaka, "Illusory motion reproduced by deep neural networks trained for prediction," *Frontiers Psychol.*, vol. 9, pp. 1–12, Mar. 2018.
[58] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," 2016, *arXiv:1605.08104*.
[59] J. Araluce, L. M. Bergasa, M. Ocaña, R. Barea, E. López-Guillén, and P. Revenga, "ARAGAN: A dRiver attention estimation model based on conditional generative adversarial network," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2022, pp. 1066–1072.
[60] F. Lateef, M. Kas, and Y. Ruichek, "Saliency heat-map as visual attention for autonomous driving using generative adversarial network (GAN)," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5360–5373, Jun. 2022.
[61] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[62] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.
[63] W. H. Warren, M. W. Morris, and M. Kalish, "Perception of translational heading from optical flow," *J. Experim. Psychol., Hum. Perception Perform.*, vol. 14, no. 4, pp. 646–660, 1988.
[64] R. Jain, "Direct computation of the focus of expansion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-5, no. 1, pp. 58–64, Jan. 1983.
[65] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow (extended abstract)," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 402–419.
[66] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. 13th Scandin. Conf. SCIA*, Jun. 2003, pp. 363–370.
[67] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. DARPA Imag. Understand. Workshop*, 1981, pp. 121–130.

YUTA MARUYAMA received the B.E. degree from the Department of Electrical and Electronic Engineering, Faculty of Engineering, Shizuoka University, Hamamatsu, Japan, in 2022. He is currently with the Department of Electrical and Electronic Engineering, Graduate School of Integrated Science and Technology, Shizuoka University. His research interests include deep learning, artificial intelligent, and computer vision.

GOSUKE OHASHI (Member, IEEE) received the B.E., M.E., and D.E. degrees from Keio University, Yokohama, Japan, in 1992, 1994, and 1997, respectively. From 2003 to 2004, he was a Visiting Researcher with the University of California, Santa Barbara. Since 1997, he has been an Assistant Professor. He is currently a Professor with the Department of Electrical and Electronic Engineering, Shizuoka University. His research interests include image processing, computational vision, and visual perception.