arXiv:2308.02530v1 [cs.CV] 1 Aug 2023

Gated Driver Attention Predictor

Tianci Zhao¹, Xue Bai², Jianwu Fang¹ and Jianru Xue³

Abstract-Driver attention prediction implies the intention understanding of where the driver intends to go and what object the driver concerned about, which commonly provides a driving task-guided traffic scene understanding. Some recent works explore driver attention prediction in critical or accident scenarios and find a positive role in helping accident prediction, while the promotion ability is constrained by the prediction accuracy of driver attention maps. In this work, we explore the network connection gating mechanism for driver attention prediction (Gate-DAP). Gate-DAP aims to learn the importance of different spatial, temporal, and modality information in driving scenarios with various road types, occasions, and light and weather conditions. The network connection gating in Gate-DAP consists of a spatial encoding network gating, long-shortterm memory network gating, and information type gating modules. Each connection gating operation is plug-and-play and can be flexibly assembled, which makes the architecture of Gate-DAP transparent for evaluating different spatial, temporal, and information types for driver attention prediction. Evaluations on DADA-2000 and BDDA datasets verify the superiority of the proposed method with the comparison with state-of-the-art approaches.

I. INTRODUCTION

The interaction between the driver and the surrounding road environment implies frequent intention prediction. The driver fixation contains the intention of where to intend to go or be interested in safe decision-making. Driver attention is a typical cognition load that reflects the capacity for selecting and perceiving the useful road context [1], and is investigated largely for normal driving situations [2]. With an important expansion, recent researches find that driver attention shows manifested promotion for accident prediction in driving scenes [3], [4]. Driver attention prediction can help to find the crashing (to be involved in an accident) object in advance under many adverse environment conditions [5], [6].

The popular prototype in this topic is to leverage the powerful fitting ability of deep learning architectures. In different driving scenarios, different drivers may focus on different scene regions because of their subjective will, which makes the driver attention prediction with large prediction uncertainty. Consequently, some works begin to adopt multitudinous information, such as *road semantics, scene motion, intended goals, object locations, etc.*, to weaken the prediction uncertainty and find the key elements in accident scenarios for driver attention prediction [7].



Fig. 1: Illustration with a critical scenario in BDDA dataset [8], where we formulate a driver cognition system with the neural links for spatial feature, temporal memory, and information type (e.g., vision, semantics, and motion) selection (i.e., by gating functions).

Although existing works improve the performance of driver attention prediction, most of them are not explainable for which kinds of information or what module play the key role in the improvement. Commonly, fusing all information or integrating all modules is universal for final implementation. In fact, it is natural that different scene information implies different promotion abilities from the aspects of the spatial region, temporal memory, or information types. Claimed by the recent research that appeared in Science [9], [10], no neuron is an island, and the outcomes of the exhibited human behaviors are driven by the connections between the neurons. The connection of the brain neurons stimulates human intelligence. How to leverage this finding, measure, or discover the expressive ability of different information and encoding modules in this topic? This question implies information selection or counterfactual reasoning problems [11]. The work of [2] explores the driver's visual fixation behavior with a model-driven white-box representation, which introduces the driving task-aware representation, such as steering angles, speed, etc. Then, the driving taskdriven representation is fed into a weight learning module with motion and bottom-up saliency maps to select the information. Some works explore the counterfactual role of different input information by masking or changing some types of them [11], [12]. These formulations involve many ad-hoc enumeration tricks, which introduce further questions for different masking strategies.

In this work, we aim to explore the network connection gating mechanism, for finding the flexible and transparent architecture for driver attention prediction (called Gate-DAP). Specifically, we design the network connection gating from the Spatial Feature Gating (SpaG), Long-Short-Term Memory Gating (MemoG), and Information-Type Gating (InfoG) (to be described in Sec. III-A), as illustrated in Fig 1. The connection gating units in Gate-DAP can be flexibly

¹T. Zhao and J. Fang are with the College of Transportation Engineering, Chang'an University, Xi'an, China fangjianwu@chd.edu.cn.

²X. Bai is with the Science and Technology on Complex System Control and Intelligent Agent Cooperative Laboratory, Beijing, China 18829281638@163.com.

³J. Xue is with the College of Artificial Intelligence, Xi'an Jiaotong University, Xi'an, China jrxue@mail.xjtu.edu.cn.

assembled for checking the roles of different information and encoding modules. In this work, we introduce four types of information of RGB video frames, road semantic images, optical flow (motion) images, and the drivable road area. Different types of information are encoded with the Vision Transformer (ViT) to leverage the self-attention mechanism. To conveniently evaluate the role of different information, we introduce an object-centric counterfactual analysis to check the role of certain types of information but maintain the whole model unchanged. The **contributions** are threefold.

- We explore the network connection gating mechanism for achieving a transparent architecture of driver attention prediction (Gate-DAP)¹, which refers to the spatial feature encoding network, temporal memory encoding network, and information fusion network.
- The connection gating units are plug-and-play and can be flexibly assembled. We introduce the object-centric counterfactual analysis for the information importance evaluation, which avoids to re-train the whole model with different configurations.
- We evaluate the performance on two datasets, DADA-2000 [5] and BDDA [8], and superior performance to other state-of-the-art methods is obtained.

II. PRECEDENT WORK

A. Driver Attention Prediction in Driving Scenes

The prediction of driver attention reflects the intention prediction of where intends to go or what is of interest in driving scenes and is investigated in many applications, such as important object detection [13], driver distraction detection [14], driving model, or policy learning [15]. With the emergence of some large-scale benchmarks, such as the ones for normal driving situations (DR(eye)VE [16], TrafficGaze [17], and CoCAtt [18]) and the critical or accident scenarios (Eyecar [7], DADA-2000 [5], and BDDA [8]), the driver attention prediction has fast progress in recent years. Driver attention prediction in driving scenes exhibits three kinds of research prototypes: *data-driven, model-driven*, and *cognitive-conditioned* approaches.

Data-driven formulation aims to leverage the data distribution of driver attention in different scenarios or datasets. The motion, semantic, and RGB frames are commonly used in the data encoding modules [7], [16]. For example, Deng *et al.* [17], [19] collect the driver attention data the targeting highway scenario and the rainy condition, and some Convolution Neural Network (CNN) models are proposed for the video frame encoder and driver attention map decoder.

Model-driven formulation is accompanied by the innovation of many kinds of learning models for learning the driver attention patterns, such as the conditional Generative Adversarial Network (GAN), Inverse Reinforcement Learning (IRL) [7], or some explainable models [2], etc.

Cognitive-conditioned approaches explore the driver status in driver attention prediction, such as the distraction state and the intention, to guide the drive attention prediction. Co-CAtt [18] investigates the driver attention pattern in turning behavior or straight-moving intention. Huang and Fu [14] adopt the predicted driver attention map to detect the state of driver distraction. Analogously, the driver's attention is also estimated by the head pose [20].

The aforementioned works for driver attention prediction show manifested progress, while most of them are not explainable for which kinds of information or what module plays the key role in the improvement.

B. Gating Networks

In recent years, many kinds of Dynamic Neural Networks (DNN) [21] have been proposed for fusing different information with various fusion strategies. Gating networks are the types of typical paradigms. Different from the previous attention-based works [22], the gating network adopts different gating functions to restrain the link of encoding networks in spatial, temporal, or modality aspects. Gating networks usually concentrate on the feature layer skipping [23], feature channel gating [24], network path selection [25], and information type allocation [26]. Commonly, the gating functions are added as a lateral skip residual link on the original feature extraction pathways. The gating function is a plug-in module that can be used in arbitrary locations in different networks.

Spatial gating networks can be divided into pixel-level gating networks, region-level gating networks, and scale-level gating networks. For example, gated convolution [27] is one typical pixel-level gating function by learning a soft mask from the data, which achieves a dynamic spatial feature selection for each feature channel and spatial location.

Temporal gating networks commonly add the gating function in the hidden state or the input of Recurrent Neural Networks (RNN). For example, Wu *et al.* propose an efficient video recognition method, which uses a conditional gating module to decide whether more discriminative information is needed for the current video frame.

Modality gating networks aim to explore the modality selection for final decisions. For example, Dynamic Multimodal Fusion (DynMM) [28] proposes a text-vision-audio fusion method for the final decision, which has a gating network for selecting the expert networks on each modality.

Most gating networks explore spatial, temporal, and modality gating separately, while different information is woven together and each kind of gating function may have an influence on each other. In this work, we explore the spatial, temporal, and information types gating together for verifying different encoding modules for driver attention prediction.

III. GATE-DAP

In this section, we first describe the network connection gating functions, and then present the whole model of driver attention prediction. The network connection gating is inspired by the connection mechanism between the neurons in the brain, where each kind of connection stimulates the intelligent understanding of different information. Meanwhile, the gating fulfills an information selection.

¹The code will be available in https://github.com/JWFangit/Gate-DAP.



Fig. 2: The structure of SpaG, MO-InfoG, MU-InfoG, where ⊙ is the Hadamard product.

A. Network Connection Gating Functions

This work leverages the gating mechanism from the aspects of spatial region selection, temporal memory selection, and information type selection, and fulfills them by the Spatial Region Gating (**SpaG**), Long-Short-Term Memory Gating (**MemoG**), and Information-Type Gating (**InfoG**) (as shown in Fig. 2) to be described in following.

1) SpaG: SpaG aims to select the spatial region feature for subsequent information encoding and driver attention prediction. SpaG is fulfilled by a spatial attention module, which multiplies the generated feature attention tensors with the original feature tensor to achieve spatial gating of each image patch. As shown by the SpaG structure in Fig. 2, for the input feature tensor $\mathbf{S} \in \mathbb{R}^{C \times W \times H}$, where *C*, *H*, and *W* are the number of feature channels, the height, and width of the feature map, respectively. Similar to [29], we first use the global average-pooling (GAP) and global max-pooling (GMP) operations along the channel axes for highlighting the spatial activation regions, and generate two types of 2D feature maps, i.e., $\mathbf{F}_{avg}^{S} \in \mathbb{R}^{1 \times W \times H}$ and $\mathbf{F}_{max}^{S} \in \mathbb{R}^{1 \times W \times H}$. Based on the spatial attention convolution, these 2D feature maps are then concatenated and convolved by a standard convolution layer, producing a 2D spatial attention map $A_s =$ $\sigma(W_s([\mathbf{F}^S_{avg}; \mathbf{F}^S_{max}])) \in \mathbb{R}^{W \times H}$, where [;] is the concatenation, σ is the Sigmoid function and W_s is the weight of a convolution operation. Then the final output S after spatial gated convolution is:

$$\mathbf{S}' = A_s \odot \mathbf{S}.\tag{1}$$

2) InfoG: InfoG concentrates on the selection of different types of information, such as RGB video frames, motion features, semantic features, etc. It is inspired by that different types of information in the same driving scene may have differing importance for drivers. For this purpose, we design two kinds of information-type gating functions: Multiple Information Type Gating (MU-InfoG) and Monocular Information Type Gating (MO-InfoG). MU-InfoG fulfills a cross-attention model among different types of information, and MO-InfoG gates the single type of information.

MO-InfoG: MO-InfoG filters the input feature tensor $\mathbf{M} \in \mathbb{R}^{C \times W \times H}$ by:

$$\mathbf{M}' = \boldsymbol{\phi}(W_f \cdot \mathbf{M}) \odot \boldsymbol{\sigma}(W_g \cdot \mathbf{M}), \tag{2}$$

where \mathbf{M}' is the output after MO-InfoG, ϕ can be any activation function (e.g., ReLU, LeakyReLU, etc.), and the ELU activation function [30] is chosen in this work for relaxing the gradient and making the neuron be active all the time. ELU activation function is defined as $e^x - 1$ for x < 0



Fig. 3: The structure of MemoG, where G_t is the output label of \mathbf{X}'_t and omitted in this work.

and x for $x \ge 0$. W_g and W_f are the gated convolution filters [27] and the original feature convolution filters, respectively. Here, $\sigma(.)$ restrains the output of the gating value to [0, 1].

MU-InfoG: As shown in Fig. 2, for the input feature tensor $\mathbf{M}_i \in \mathbb{R}^{C \times W \times H}$, we first use 1D convolution to reduce their channel dimension and generate $\mathbf{m}_i \in \mathbb{R}^{1 \times W \times H}$. Then, we concatenate \mathbf{m}_i along the information type dimension to obtain a tensor **MU** with the size of $n \times W \times H$, where *n* is the number of information types. Next, we use the *softmax* function to ensure that the sum of the feature values of **MU** at each spatial dimension is 1 to achieve the feature selection. Finally, we rearrange **MU** to the shape of the original feature tensor \mathbf{M}_i and obtained *n* masks $\{\mathbf{mask}_i\}_{i=1}^n$ with size of $H \times W$. The output of MU-InfoG for each kind of information is denoted as \mathbf{M}'_i .

3) *MemoG*: MemoG focuses on the temporal memory gating with a long and short window consideration. MemoG stands at the gate recurrent unit (GRU) and filters out redundant features in historical video frames with short and long-term information gating. We treat the hidden state \mathbf{H}_t as a representation of a long-term memory at time *t* after several times of temporal recurrences. The input \mathbf{X}_t at time *t* is denoted as the representation of short-term memory.

Specially, in short-term memory, we consider the **uncertainty** in driver attention prediction, which is fulfilled by the cross-attention of the input feature tensors of *k* frames, i.e., $[\mathbf{X}_t, \mathbf{X}_{t-1}, ..., \mathbf{X}_{t-k}]$. If we encounter a sudden change in driving scenes, the uncertainty estimation will give a large weight to the frame with sudden change. This consideration is achieved by the MU-InfoG operation on $[\mathbf{X}_t, \mathbf{X}_{t-1}, ..., \mathbf{X}_{t-k}]$ and generates $[\mathbf{X}'_t, \mathbf{X}'_{t-1}, ..., \mathbf{X}'_{t-k}]$.

For long-term memory, we employ the MO-InfoG to gate the hidden state \mathbf{H}_{t-1} obtained in the previous time. Consequently, the MemoG is modeled by:

$$\mathbf{H}'_{t-1} = \text{MO-InfoG}(\mathbf{H}_{t-1}), \mathbf{H}_t = \text{GRU}(\mathbf{H}'_{t-1}, \mathbf{X}'_t), \quad (3)$$

Fig. 3 demonstrates the structure of MemoG. We omit Y_t for the driver attention decoding. We explicitly enforce the gating function to the input hidden state and current



Fig. 4: The pipeline of **Gate-DAP**. Notably, we estimate the uncertainty in the memory gating module, which is useful for feature learning with sudden scene changes.

observation, which aims to purify the temporal information before the GRU unit (with 256 dimensions of hidden state).

Thus, the SpaG, InfoG, and MemoG are described, which can be flexibly adopted in any deep learning model, and are carefully utilized in our driver attention prediction network.

B. Driver Attention Prediction

In driver attention prediction tasks for driving scenarios, the driver's eye movements are usually influenced by multiple factors due to complex and variable road conditions. The whole pipeline of the Gated Driver Attention Predictor (Gate-DAP) model is shown in Fig. 4. In this work, we consider four kinds of information to model the driving scenes, and each input sample clip of Gate-DAP consists of a group of $[I_{1:t}, F_{1:t}, S_{1:t}, D_{1:t}]$, where the t^{th} frame at one clip is denoted as I_t for RGB information, F_t for motion information, S_t for semantic information, and D_t for drivable region information. Notably, motion frame F_t is obtained by computing the motion correlation between I_t and I_{t-1} .

Backbone Model: Each frame in one clip is encoded by a backbone model, respectively. As shown in Fig. 4, different types of information follow a parallel encoding but share the encoder weights. Because the encoding of each type of information is the same, we take the t^{th} RGB frame in one clip as an example. In this work, we take the vision-transformer model (ViT) as the backbone model, which is pre-trained in ImageNet-1K by masked auto-encoder (MAE) [31]. ViT is fulfilled by a multi-head self-attention module on image patches with position embedding. In this work, the number of self-attentive heads of ViT is set as 12, the depth of layers is set as 12, and the patch size is 16.

Connection Gating: Denote the feature embedding of the RGB frame clip is $[\mathbf{Z}_1^I, ..., \mathbf{Z}_t^I]$. The feature embedding of each frame is separately gated by the SpaG to achieve a spatial feature selection. The output of the SpaG at each frame is correlated by the MemoG to fulfill a temporal memory gating over the frames within the frame clip. MU-InfoG is adopted by following the MemoG at the *t*th frame for gating different

types of information. The gating part of Gate-DAP is:

$$\mathbf{Z}_{t}^{I} = \mathrm{ViT}|_{\mathrm{MAE}}(I_{t}), \tag{4}$$

$$\mathbf{Z}_{t}^{\prime I} = \operatorname{SpaG}(\mathbf{Z}_{t}^{I}), \tag{5}$$

$$\mathbf{H}_{t}^{I} = \operatorname{MemoG}(\mathbf{H}_{t-1}^{I}, [\mathbf{Z'}_{1}^{I}, ..., \mathbf{Z'}_{t}^{I}]),$$
(6)

$$\mathbf{I}_{t}^{\prime I}, \mathbf{M}_{t}^{\prime F}, \mathbf{M}_{t}^{\prime S}, \mathbf{M}_{t}^{\prime D}] = \mathrm{MU} \operatorname{InfoG}(\mathbf{H}_{T}^{I}, \mathbf{H}_{T}^{S}, \mathbf{H}_{T}^{F}, \mathbf{H}_{T}^{D}), \quad (7)$$

$$\mathbf{I}'_{t} = \operatorname{Stack}[\mathbf{M}'_{t}, \mathbf{M}'_{t}, \mathbf{M}'_{t}, \mathbf{M}'_{t}].$$
(8)

where $\mathbf{M'}_t$ is the final feature representation at time *t* after stacking all information types and adopted to decode the future attention map at time *t* + 1. Because embedding each type of information is time-consuming by the ViT model, we share the weight for different types of information.

Attention Map Decoding: After the MU-InfoG operation, \mathbf{M}'_t needs to be converted to an attention map with the same size as the input frame. Therefore, \mathbf{M}'_t is decoded as the final driver attention map Y_{t+1} for the $(t+1)^{th}$ frame.

The decoding module consists of three interleaved blocks, each one of which contains a 2D convolution, batch normalization, ReLU function, and upsampling operation, fulfilled by $[conv(3 \times 3, 128) \rightarrow Batch Normalization \rightarrow ReLU \rightarrow upsampling] \times 4 \rightarrow conv(3 \times 3, 1)$, and the final block ends with Sigmoid function for mapping the output value to [0,1]to highlight the focused regions of the $(t+1)^{th}$ frame.

Loss Function: Similar to DADA [5], we also take the joint loss function to train the gated network, which contains the Kullback-Leibler distance (*KLD*), linear correlation coefficient (*CC*), and normalized scan path significance (*NSS*). The specific loss function is denoted as:

$$\mathscr{L}(Y_{t+1}, \hat{Y}_{t+1}) = \sum_{i=1}^{N} Y_{t+1}(i) log(\varepsilon + \frac{Y_{t+1}(i)}{\varepsilon + \hat{Y}_{t+1}(i)}) - \alpha \cdot \frac{cov(Y_{t+1}, Y_{t+1})}{\rho(Y_{t+1})\rho(\hat{Y}_{t+1})} - \beta \cdot \frac{1}{\sum_{i=1} P_{t+1}(i)} \sum_{i=1} \frac{\hat{Y}_{t+1}(i) - \mu(\hat{Y}_{t+1})}{\rho(\hat{Y})} P_{t+1}(i)$$
(0)

where Y_{t+1} and P_{t+1} are ground-truth saliency and fixation point maps, respectively, \hat{Y}_{t+1} is the predicted attention map. *i* indexes the *i*th pixel across the all *N* pixels of the saliency map. α and β are the coefficient that adjusts the weight of *CC* and *NSS*, *N* indicates the number of image pixel points, $cov(Y, \hat{Y})$ indicates the covariance of Y_{t+1} and \hat{Y}_{t+1} ; ρ indicates standard deviation operation, and ε is a very small constant to prevent the operation from errors such as the number of denominators being 0.

IV. EXPERIMENTS

A. Dataset

[N]

In this paper, we evaluate the performance of the proposed Gate-DAP on two challenging datasets with critical or accident scenarios, i.e., BDD-A [8] and DADA-2000 [5].

BDD-A [8] consists of 1,232 sequences (each one owns about 10 seconds). It focuses on critical situations such as occlusions, truncations, and emergency braking. To obtain annotations, 45 drivers are asked to watch videos, and their eye movements are recorded by an eye tracker to generate fixations. We follow its partition and obtain 28k frames for training, 6k frames for validation, and 9k frames for testing. DADA-2000 [5] concentrates on driver attention prediction in accident scenarios. This dataset contains 2000 videos with over 658,746 frames. We follow the work [5] that 1000 videos are used for performance evaluation, which provides 598 training sequences (about 214k frames) and 222 testing sequences (about 70k frames), respectively.

B. Implementation Details

The proposed method is implemented using the PyTorch framework. During the training process, we used the Adam optimizer with a learning rate of 10^{-6} and a weight decay of 0.0001. The entire model is trained in end-to-end mode, and the entire training process takes about 20 hours and 6 hours on one NVIDIA RTX2080Ti GPU with 11GB RAMs for DADA and BDD-A datasets, respectively. In addition, regarding the number of input frames in one clip, based on our previous research [5], [32], we found that for the frame or map prediction problems, more input frames will consume more computing resources with little performance gain. Therefore, in our implementation, due to the limitation of RAM space, each input clip contains 4 consecutive frames. We pre-prepare the semantic images, optical flow images, and drivable area images in advance using the DeeplabV3 [33], FlowNet2.0 [34], and Yolo-P [35], respectively.

Metrics: Following the previous driver attention prediction methods [5], [7], [16], we utilize five metrics to evaluate the performance, which contains three distribution-based metrics, i.e., Kullback-Leibler Divergence (KLD), Pearson Correlation Coefficient (CC), and Similarity (SIM), and two location-based metrics, i.e., Normalized Scanpath Saliency (NSS) and the area under the receiver operating characteristic (ROC) curve (AUC). Here, two variants of AUC were used, namely AUC-Judd (AUC-J) and shuffled AUC (AUC-S).

C. Ablation Study

1) Which information is important? A counterfactual analysis. To evaluate the importance of each type of information, this work introduces a counterfactual analysis strategy. We all know that most participants in driving scenes are pedestrians and vehicles, and these two types of semantics basically attract driver attention in most situations. If we remove these two kinds of semantics in the images, the input images may only contain the background. Specifically, we maintain the whole architecture of Gate-DAP and remove these two kinds of semantics one by one for each type of information (See Fig. 5 for semantic information). For the drivable area image, we remove the binary mask.

This strategy does not need to re-train the model with different information configurations and check the importance of each type of information by the metric value difference with Gate-DAP-Full-Model. Totally, we obtain ten versions after counterfactual analysis, denoted as three RGB versions ("Gate-DAP-I w/o P", "Gate-DAP-I w/o V", and "Gate-DAP-I w/o V-P"), three motion versions ("Gate-DAP-F w/o P", "Gate-DAP-F w/o V", and "Gate-DAP-F w/o V-P"), three semantic versions ("Gate-DAP-S w/o P", "Gate-DAP-S w/o V", and "Gate-DAP-S w/o V-P"), and one drivable mask



Fig. 5: The counterfactual operation on a semantic image (removing pedestrians and vehicles, i.e., "*Gate-DAP-S w/o V-P*") and removing the binary mask in drivable area image (i.e, "*Gate-DAP-D w/o Mask*").



Fig. 6: Performance influence with the counterfactual operation on each kind of information, respectively. This evaluation is conducted on the testing set of the DADA-2000 dataset.

version ("*Gate-DAP-D w/o Mask*"). Here, "*P*", "V", and "*Mask*" denote the indication of pedestrians, vehicles, and road mask regions, respectively. Larger differences mean that the information with our counterfactual operation has more importance. If one kind of information with counterfactual operation hardly affects the result of each metric, it is useless.

Fig. 6 demonstrates the performance influence of each kind of information. We observe that the motion information in this work has the weakest influence on the performance, verified by "*Gate-DAP-F w/o P*", "*Gate-DAP-F w/o V*", and "*Gate-DAP-F w/o V-P* with little difference. We also show some snapshots of predicted driver attention maps in Fig. 7. The visualization results demonstrate that removing the pedestrian and vehicles in the motion information (Fig. 7(f)) has little change with the Full-Model. On the contrary, the drivable area mask (we denote it as an indirect driving task representation) has the largest performance influence (marked by the black lines in Fig. 6). Besides motion information, other kinds of information have an impact on performance to a large extent. Therefore, we think the role of motion information in this work is very little.

2) Temporal uncertainty evaluation in MemoG. As aforementioned, we consider the temporal uncertainty in MemoG by cross-attention model for successive frames. This consideration aims to find the sudden scene change in critical or accident scenarios. Here, we evaluate the role of this consideration by the setting of MemoG with or without the Temporal Uncertainty (i.e., MemoG-w-TU. and MemoG-w/o-TU.). Accordingly, the structure of MemoG for the short-term memory gating is changed, as shown in Fig. 8. Table. I presents the results on the testing set of the DADA-2000 dataset, and we can see that **TU**. shows a slight promotion role for driver attention prediction.

3) How about the role of different gating modules?



Fig. 7: Some predicted driver attention maps in DADA-2000 for checking the importance of different types of information.





Fig. 9: The visualization of some predicted driver attention examples in DADA-2000 for evaluating different gating modules. (a): ground-truth; (b): Gate-DAP-Full-Model; (c): Gate-DAP w/o SpaG; (d): Gate-DAP w/o MemoG; (e): Gate-DAP w/o MU-InfoG.

The primary insight of this work is to introduce the gating modules. To evaluate their roles, we take contrastive experiments, where we close the relative gating modules in the Gate-DAP model. Consequently, we have three versions in this evaluation, i.e., "Gate-DAP w/o SpaG", "Gate-DAP w/o MemoG", and "Gate-DAP w/o MU-InfoG".

"Gate-DAP w/o SpaG" is fulfilled by setting the weight map W_s in Eq. 1 as an identity matrix.

"Gate-DAP w/o MemoG" has two kinds of gating modules: MO-InfoG and MU-InfoG. Eliminating MO-InfoG is achieved by setting the weight matrix W_g and W_f in Eq. 2 as two identity matrixes. Eliminating MU-InfoG is fulfilled by setting all the weight of different information equally.

"*Gate-DAP w/o MU-InfoG*" is obtained by the same setting for multiple kinds of information (with the same weight for each type of information).

The results are shown in Tab II, and we can see that the gating modules are positive for driver attention prediction. Among them, the MU-InfoG module has the greatest contribution. We can see that after adding the MU-InfoG module, all metric values improve. We also demonstrate some snapshots of predicted attention maps by different gating configurations in Fig 9. From the figure, we can clearly see without the SpaG and MU-InfoG, the predicted driver attention is intended on the road region, while the actual fixations concentrate on the vehicle. Gate-DAP with all gating modules can localize the true driver fixations well.

TABLE II: Evaluation on different gating modules.

(Gating M	odules	DADA2000				BDD-A			
SpaG	MemoG	MU-InfoG	$KLD\downarrow$	$\mathbf{CC}\uparrow$	$\text{SIM} \uparrow$	$\text{NSS} \uparrow$	$\mathrm{KLD}\downarrow$	$\mathbf{CC}\uparrow$	$\text{SIM} \uparrow$	
			1.70	0.47	0.35	3.07	1.47	0.52	0.40	
\checkmark			1.69	0.47	0.35	3.11	1.44	0.53	0.42	
	\checkmark		1.69	0.47	0.35	3.09	1.46	0.52	0.40	
		\checkmark	1.67	0.48	0.35	3.13	1.24	0.59	0.45	
\checkmark	\checkmark		1.68	0.47	0.35	3.13	1.43	0.53	0.42	
\checkmark		\checkmark	1.67	0.48	0.36	3.13	1.19	0.59	0.47	
	\checkmark	\checkmark	1.67	0.48	0.36	3.12	1.24	0.59	0.45	
\checkmark	\checkmark	\checkmark	1.65	0.52	0.36	3.14	1.12	0.61	0.49	

D. Comparison with State-of-The-Arts

To validate the superiority of Gate-DAP, seven representative driver attention prediction methods are compared, with five video-based BDDA [8], DR (eye) VE [16], TwoStream [36], SCAFNet [5], TASEDNet [37], Vi-Net [39], Flow-DA [40], and ASIAF-Net [38]. We compare the results reported in their works in this evaluation.

In addition, in the ablation studies, we introduce the counterfactual analysis and *gating closing* to check the information's importance and gating roles, where different configurations are not re-trained. Certainly, to prove the reasonability of the ablation study ways, we re-train two versions, i.e., removing all gating modules (*Gate-DAP w/o Gs*) and removing motion information (*Gate-DAP w/o F*), to check whether our finding is reasonable or not. The evaluation results are shown in Table. III. From the results, we can see that our Gate-DAP shows promising performance in all datasets, especially for the KLD metric and AUC-J metric. Besides, the results of *Gate-DAP w/o Gs* and *Gate-DAP w/o F* indicate that the ablation study ways in this work are promising for checking the role or importance of different model configurations without model re-training.

V. CONCLUSIONS

This work proposes a Gated Driver Attention Predictor (Gate-DAP), which explores spatial feature gating, temporal memory gating, and information type gating to fulfill a transparent architecture of driver attention prediction networks. The gating modules are a plug-and-play that can be used to check the role of different kinds of features, i.e., spatial feature, temporal feature, and information type feature. In addition, we introduce a counterfactual analysis to evaluate the importance of different types of information. Based on the analysis, motion features have the least

TABLE III: Comparison with several state-of-the-art methods.

	DADA2000						BDD-A		
	$\text{KLD}\downarrow$	$\mathbf{CC}\uparrow$	$\text{SIM}\uparrow$	$\text{NSS} \uparrow$	AUC-J \uparrow	AUC-S \uparrow	$\mathrm{KLD} \!\!\downarrow$	$\mathbf{CC}\uparrow$	$\text{SIM}\uparrow$
DR(eye)VE [16]	2.27	0.45	0.32	2.92	0.91	0.64	1.95	0.50	-
BDDA [8]	3.32	0.33	0.25	2.15	0.86	0.63	1.49	0.51	0.35
TwoStream [36]	2.85	0.23	0.14	1.48	0.84	0.64	-	_	_
TASEDNet [37]	1.78	0.46	0.31	3.20	0.92	-	1.24	0.55	0.42
Vi-Net [39]	-	_	_	-	-	-	1.39	0.61	0.45
SCAFNet [5]	2.19	0.50	0.37	3.34	0.92	0.66	1.39	0.54	0.43
Flow-DA [40]	_	-	-	-	-	-	1.39	0.61	0.45
ASIAF-Net [38]	1.66	0.49	0.36	3.39	0.93	0.78	1.24	0.66	-
Gate-DAP w/o Gs	1.72	0.46	0.35	3.03	0.92	0.84	1.36	0.54	0.39
Gate-DAP w/o F	1.66	0.47	0.35	3.12	0.92	0.85	1.18	0.61	0.42
Gate-DAP	1.65	0.52	0.36	3.14	0.93	0.85	1.12	0.61	0.49

influence after removing the pedestrians and vehicles in motion images. On the contrary, drivable area images show manifest performance influence. Through the comparison with other state-of-the-art methods, Gate-DAP generates the best performance on DADA-2000 and BDDA datasets.

References

- A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 900–918, 2020.
- [2] X. Tang, J. Yu, and Y. Su, "Modeling driver's visual fixation behavior using white-box representations," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15434–15449, 2022.
- [3] W. Bao, Q. Yu, and Y. Kong, "DRIVE: deep reinforced accident anticipation with visual explanation," in *ICCV*, 2021, pp. 7599–7608.
- [4] J. Fang, D. Yan, J. Qiao, J. Xue, H. Wang, and S. Li, "DADA-2000: can driving accident be predicted by driver attention? analyzed by A benchmark," in *ITSC*, 2019, pp. 4303–4309.
- [5] 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, 2022.
- [6] S. Gan, X. Pei, Y. Ge, Q. Wang, S. Shang, S. E. Li, and B. Nie, "Multisource adaption for driver attention prediction in arbitrary driving scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 20912–20925, 2022.
- [7] S. Baee *et al.*, "MEDIRL: predicting the visual attention of drivers via maximum entropy deep inverse reinforcement learning," in *ICCV*, 2021, pp. 13 158–13 168.
- [8] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney, "Predicting driver attention in critical situations," in ACCV, vol. 11365, 2018, pp. 658–674.
- [9] P. Stern, "No neuron is an island," Science, vol. 378, no. 6619, pp. 486–487, 2022.
- [10] M. T. D. Scholtten and S. J. Forkel, "The emergent properties of the connected brain," *Science*, vol. 378, no. 6619, pp. 505–510, 2022.
- [11] P. Jacob, É. Zablocki, H. Ben-Younes, M. Chen, P. Pérez, and M. Cord, "STEEX: steering counterfactual explanations with semantics," in *ECCV*, vol. 13672, 2022, pp. 387–403.
- [12] C. Li, S. H. Chan, and Y. Chen, "Who make drivers stop? towards driver-centric risk assessment: Risk object identification via causal inference," in *IROS*, 2020, pp. 10711–10718.
- [13] Y. Rong, N. Kassautzki, W. Fuhl, and E. Kasneci, "Where and what: Driver attention-based object detection," *Proc. ACM Hum. Comput. Interact.*, vol. 6, pp. 1–22, 2022.
- [14] T. Huang and R. Fu, "Driver distraction detection based on the true driver's focus of attention," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19374–19386, 2022.
- [15] A. Morando, T. W. Victor, and M. Dozza, "A reference model for driver attention in automation: Glance behavior changes during lateral and longitudinal assistance," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 2999–3009, 2019.
- [16] 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, 2019.

- [17] 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, 2020.
- [18] Y. Shen, N. Wijayaratne, P. Sriram, A. Hasan, P. Du, and K. D. Campbell, "Cocatt: A cognitive-conditioned driver attention dataset," pp. 32–39, 2022.
- [19] H. Tian, T. Deng, and H. Yan, "Driving as well as on a sunny day? predicting driver's fixation in rainy weather conditions via a dualbranch visual model," *IEEE CAA J. Autom. Sinica*, vol. 9, no. 7, pp. 1335–1338, 2022.
- [20] S. Jha and C. Busso, "Estimation of driver's gaze region from head position and orientation using probabilistic confidence regions," *IEEE Trans. Intell. Veh.*, vol. 8, no. 1, pp. 59–72, 2023.
- [21] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7436–7456, 2022.
- [22] Y. Wang, X. Zhang, X. Hu, B. Zhang, and H. Su, "Dynamic network pruning with interpretable layerwise channel selection," in AAAI, 2020, pp. 6299–6306.
- [23] X. Wang, F. Yu, Z. Dou, T. Darrell, and J. E. Gonzalez, "Skipnet: Learning dynamic routing in convolutional networks," in *ECCV*, vol. 11217, 2018, pp. 420–436.
- [24] C. Li, G. Wang, B. Wang, X. Liang, Z. Li, and X. Chang, "Dynamic slimmable network," in CVPR, 2021, pp. 8607–8617.
- [25] Y. Li, L. Song, Y. Chen, Z. Li, X. Zhang, X. Wang, and J. Sun, "Learning dynamic routing for semantic segmentation," in *CVPR*, 2020, pp. 8550–8559.
- [26] Y. Meng, R. Panda, C. Lin, P. Sattigeri, L. Karlinsky, K. Saenko, A. Oliva, and R. Feris, "Adafuse: Adaptive temporal fusion network for efficient action recognition," in *ICLR*, 2021.
- [27] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *ICCV*, 2019, pp. 4470– 4479.
- [28] Z. Xue and R. Marculescu, "Dynamic multimodal fusion," *CoRR*, vol. abs/2204.00102, 2022.
- [29] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *ECCV*, vol. 11211, pp. 3–19.
- [30] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," in *ICLR*, 2016.
- [31] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022, pp. 16 000– 16 009.
- [32] S. Li, J. Fang, H. Xu, and J. Xue, "Video frame prediction by deep multi-branch mask network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1283–1295, 2021.
- [33] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017.
- [34] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *CVPR*, 2017, pp. 2462–2470.
- [35] D. Wu, M. Liao, W. Zhang, and X. Wang, "Yolop: You only look once for panoptic driving perception," arXiv preprint arXiv:2108.11250, 2022.
- [36] K. Zhang and Z. Chen, "Video saliency prediction based on spatialtemporal two-stream network," *IEEE Trans. Circuits Syst. Video Tech*nol., vol. 29, no. 12, pp. 3544–3557, 2018.
- [37] K. Min and J. J. Corso, "Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection," in *ICCV*, 2019, pp. 2394–2403.
- [38] Q. Li, C. Liu, F. Chang, S. Li, H. Liu, and Z. Liu, "Adaptive shorttemporal induced aware fusion network for predicting attention regions like a driver," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 18695–18706, 2022.
- [39] S. Jain, P. Yarlagadda, S. Jyoti, S. Karthik, R. Subramanian, and V. Gandhi, "Vinet: Pushing the limits of visual modality for audiovisual saliency prediction," in *IROS*, 2021, pp. 3520–3527.
 [40] R. Sultana and G. Ohashi, "Prediction of driver's visual attention in
- [40] R. Sultana and G. Ohashi, "Prediction of driver's visual attention in critical moment using optical flow," *IEICE Trans. Inf. Syst.*, vol. 106, no. 5, pp. 1018–1026, 2023.