ARAGAN: A dRiver Attention estimation model based on conditional Generative Adversarial Network

Javier Araluce[1], Luis M. Bergasa[1], Manuel Ocaña[1], Rafael Barea[1], Elena López-Guillén[1], Pedro Revenga[1]

***Abstract*— Predicting driver's attention in complex driving scenarios is becoming a hot topic due to it helps the design of some autonomous driving tasks, optimizing visual scene understanding and contributing knowledge to the decision making. We introduce ARAGAN, a driver attention estimation model based on a conditional Generative Adversarial Network (cGAN). This architecture uses some of the most challenging and novel deep learning techniques to develop this task. It fuses adversarial learning with Multi-Head Attention mechanisms. To the best of our knowledge, this combination has never been applied to predict driver's attention. Adversarial mechanism learns to map an attention image from an RGB traffic image while mapping the loss function. Attention mechanism contributes to the deep learning paradigm finding the most interesting feature maps inside the tensors of the net. In this work, we have adapted this concept to find the saliency areas in a driving scene.**

An ablation study with different architectures has been carried out, obtained the results in terms of some saliency metrics. Besides, a comparison with other state-of-the-art models has been driven, outperforming results in accuracy and performance, and showing that our proposal is adequate to be used on real-time applications. ARAGAN has been trained in BDDA and tested in BDDA and DADA2000, which are two of the most complex driver attention datasets available for research.

I. INTRODUCTION

Intelligent vehicles have reached multiple advances in recent years with the goal of achieving the fully autonomous driving architecture, defined as the fifth level (L5) according to the J3016 SAE international Standard [1]. L5 architectures are divided into different tasks to fulfill the outstanding goal of autonomous driving. This work is focused on visual scene understanding, which includes tasks as pedestrian/vehicle detection and tracking, signs/traffic lights recognition or semantic segmentation. But, do autonomous vehicles need to detect all the objects on the scene or just only the ones that affect them? Do humans recognize the complete scene when they are driving or just a few objects?

Humans keep their visual attention on the most saliency objects in order to focus their attention on a potential hazard and be aware of the traffic situation to be able to make decisions over it. Human attention is a fusion of two mechanisms, bottom-up (color or intensity) and top-down (goals or intention) [2], [3]. This fusion enables drivers to focus their attention only on the necessary objects to get the sufficient information to make decisions [4], [5]. This concept transferred to the task proposed here, optimizes the visual scene understanding [6], saving computational efforts when the vehicle is driving in a complex scenario [7].

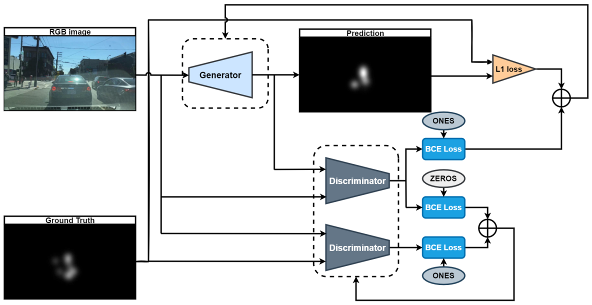


Fig. 1: ARAGAN framework. Generator, Discriminator and the flow of the information generated by these two nets is represented.

Driver gaze focalization has been used to obtain driver's attention. Recently some datasets have been published to explore this field in intelligent vehicles. The first one to come out was DR(eye)VE dataset [8], which recorded and annotated 550,000 frames of driving sequences in different traffic and weather conditions. Every frame provides the driver's gaze obtained through an accurate eye tracker while driving. After that, Berkeley DeepDrive Attention (BDD-A) [9] dataset was launched, claiming that it is nearly impossible to collect enough driver attention data for crucial events with a conventional in-car data collection protocol, like DR(eye)VE project did. This protocol only captures a single focus and false positive gazes, which will confuse the model during training. To overcome this concern, they proposed an in-lab recorded dataset after a number of experienced drivers were looked. With this new protocol, DADA-2000 dataset [10] came out, which enabled annotating critical accidental situations with this mechanism. This collected 658,476 frames from accidents and annotated them using an in-lab protocol with experienced driver's gaze to obtain their attention. From our knowledge, the referenced datasets are the only ones that tackle driver attention and are public nowadays. Many works have been focused on obtaining driver's attention through deep learning models trained on these datasets. More competitive proposals are based on conditional Generative Adversarial Nets (cGANs) [11].

The release of Transformers [12] supposed a change in

*This work has been funded in part from the Spanish MICINN/FEDER through the Techs4AgeCar project (RTI2018-099263-B-C21) and from the RoboCity2030-DIH-CM project (P2018/NMT- 4331), funded by Programas de actividades I+D (CAM) and cofunded by EU Structural Funds.

[1] Javier Araluce, Luis M. Bergasa, Manuel Ocaña, Rafael Barea, Elena López-Guillén and Pedro Revenga are with the Electronics Department, University of Alcalá (UAH), Spain. {javier.araluce, luism.bergasa}@uah.es, mocana@depeca.uah.es, {rafael.barea, elena.lopezg, pedro.revenga}@uah.es

Code is available in https://github.com/javierAraluce/ARAGAN the deep learning paradigm regarding attention mechanism because allowed multi-head attention in an effective way. This technique, firstly used to outperform Natural Language Processing (NLP) models, has recently been implemented in computer vision applications. There are several works that have overcome state-of-the-art results. This is the case of [13], that presents a Transformer architecture tested on different benchmarks including image-level classification [14], region-level object detection [15] and pixel-level semantic segmentation [16].

This work presents a driver attention estimation model based on a conditional Generative Adversarial Net (cGANs) [17], which is a network that learn to map an image of distribution (b) from an image of distribution (a) mapping a loss function during training. cGANs differ from GANs [18] in the fact that a condition is fed to the generator and the discriminator instead of using random noise. In our case, this feeding is an RGB image extracted from traffic scene. Besides, cGAN generator uses the encoder of a Transformer providing our architecture with Multi-Head Attention capacity. To the best of our knowledge this is the first time that these two techniques (cGAN and Transformer) have been used together to predict driver's attention.

The main contribution of this work is our novel and efficient architecture, which could be used in real-time applications, as well as its high performance, surpassing results obtained by other state-of-the-art proposals on open-source datasets.

II. RELATED WORKS

In recent years, prior efforts have been made in the literature to get driver attention models using computer vision techniques. Moreover, some new Deep Learning (DL) techniques have been used in intelligent vehicle applications.

A. *Driver attention models.*

Research has been conducted in the field of fixation models based on computer vision for some years using DL techniques [19], [20] and [21]. These are the first ranked models on the MIT300 [22] benchmark. On the other hand, [23] and [24] are the first ranked models on the CAT2000 [25] benchmark. However, it was not until the release of the DR(eye)VE dataset [8] that these models started to be used in intelligent vehicles scenarios.

With this release, several DL models were trained on this dataset, including the one proposed by the authors in [8], which was based on a multi-branch architecture. Three different branches (RGB, Flow estimation and Semantic Segmentation) composed the model and were fused to obtain the attention prediction. Other trained models were proposed on [2], [26] and [27].

After this dataset, Berkeley DeepDrive Laboratory launched its Berkeley DeepDrive Attention dataset (BDD-A) [9], which not only proposed a new protocol to estimate driver's attention but obtained the peripheral vision, since humans have the ability to fixate their gaze on an object while attending others. To include this functionality authors propose using more than one observer to capture the attention map. They presented their dataset and a model trained on it. The proposed model used a feature extractor based on ImageNet and pre-trained in AlexNet [28]. The input was based on 6 consecutive frames with a temporal processing unit in the form of 2D convolutions and Convolutional LSTM (Conv2D-LSTM).

With this new protocol, another dataset came out, DADA2000 [10], which not only proposed a new dataset based on accidents but also proposed an architecture to predict driver's attention [29]. This model is based on a multi-branch architecture with a semantic-guided attentive fusion to predict the attention map from the RGB image and the semantic segmentation. Like its predecessors, this model used a sequence of images as input with Convolutional LSTM (Conv2D-LSTM).

In recent years, there have been other works that aimed to be different from the state of the art models in this field. Authors of [30] proposed a Maximum Entropy Deep Inverse Reinforcement Learning model to predict visual attention on driving scenarios. They also announce a dataset to be tested by the community, but it is not completely public. They predict the eyes' movement, which is the action of their inverse reinforcement learning model.

B. *Deep learning techniques.*

In order to build the application described earlier some DL techniques have been proposed. Adversarial learning was proposed in [18], which gives the ability to learn to generate a data distribution (Generator) and a discriminative model (Discriminator) that learns to estimate if a sample came from the ground truth or it has been generated by the Generator. This training procedure corresponds to minimax two-net game. Image to image generation has been a topic developed in the DL field for a long time, but it was not until [31] came out that it was investigated with conditional adversarial networks. They proposed an architecture to solve different image processing problems with a generator based on the U-Net [32] and a Markovian discriminator denominated PatchGan.

Driver's attention through adversarial learning has recently been presented in [11], which is based on the architecture described in [31].

Apart from adversarial learning, there is another tide that is changing the DL paradigm, the Multi-Head Attention. Because not all the stimuli are equal, focus attention has enabled the human species to direct attention to objects of interest, such as preys and predators have been doing in the complex visual environment during history. Finding the saliency regions of a traffic scene making the most of Transformers [12], which allow efficient Multi-Head Attention capability, is the key aspect of this work.

Before this architecture, there were some others that estimated attention. The mostly used ones were additive attention [33], and dot-product (multiplicative) attention [34]. There are different approaches that have implemented Transformers for computer vision tasks, despite they were firstly designed to be used in NLP tasks. In [35], they used an attention module based on a Channel Attention ‘what to pay attention’ and a Spatial Attention ‘where to pay attention’. This implementation was based on the work proposed in [36], which implements channel attention in a different way. These works show progress in using attention mechanisms in computer vision tasks. But they did not end up using the Transformer. In [37], Self-Attention is employed to model an adversarial learning that was tested on the ImageNet-1K [14] dataset. Self-Attention is the base of the work proposed on [12] but still needs some modifications to explore the Transformers’ capability. Finally, [13] presented a work where the Transformer was used for computer vision tasks. They tested and got state of the art results on different benchmarks, such as, ImageNet-1K image classification (V1 and V2) [38], [39], COCO object detection [15], ADE20K semantic segmentation [40] and Kinetics-400 video action classification [41].

There is a recent work [42] that has presented a driver attention model using the fusion of Transformers and convolutions. This model has been trained on BDD-A but obtains worse performance results that the obtained with our proposal in the same test dataset.

After this literature revision and to the best of our knowledge, we can claim that the architecture proposed here is the first that fuse the novel DL techniques cGANs and Transformers to predict driver’s attention for real-time intelligent vehicles applications.

III. ARAGAN ARCHITECTURE

Taking novel state-of-the-art DL techniques (cGANs and Transformers), we proposed ARAGAN architecture for driver’s attention estimation. ARAGAN improves state-of-the-art results, opening the possibility to perform real-time applications due to the efficiency of our architecture. Figure 1 shows the framework of our proposal, where the Generator, Discriminator and the information flow generated by these two nets is represented. Then, loss function calculation is included in a graphical way.

Based on the Pix2Pix architecture [31], which presents a cGAN that learns from an image to obtain another image, we propose to use conditional adversarial learning for driver attention prediction. We feed the network with a RGB image of the traffic scene, expecting to obtain a predicted attention map over the image. The Generator predicts “fake” attention maps that cannot be distinguished from the “real” ones by the Discriminator, which is trained adversarially to identify these synthetics attention maps. Both nets apply the gradients at every batch.

Hereafter, we present the learning strategy applied in our system, the different attention modules used in the Generator design, as well as the architectures propose for the Generator and the Discriminator.

A. Adversarial learning procedure

Loss function of a cGAN is shown in Equation 1 where the Generator (G) tries to minimize the function and the Discriminator (D) aims to maximize it,

$$\mathscr{L}_{cGAN} = \mathbb{E}_{x,y}[logD(x,y)] + \mathbb{E}_{x,z}[log(1 - D(G(x,z))] \tag{1}$$

being x the RGB image, y the driver attention estimation and z a random noise. This logic is based on what authors claimed in [31]. We have removed the random noise added to the input image as the generator learns to avoid the noise as exposed in [43]. Namely, we have provided the net with a dropout that provides noise to it, without jeopardizing the input. Besides, we have added to the function a L1 distance, represented in Equation 2. It is based on the work presented in [44], where they claim that adding a traditional loss to the Generator benefits its performance.

$$\mathscr{L}_{L1}(G) = \mathbb{E}_{x,y,z}[y - G(x,z)||_1] \tag{2}$$

With this said, the final loss function for our architecture is shown in Equation 3.

$$G^* = arg\,\min_G \max_D \mathscr{L}_{cGAN} + \lambda \mathscr{L}_{L1}(G) \tag{3}$$

B. Attention modules

In this subsection, we introduce the different attention modules used in the Generator design.

1) Convolutional Block Attention Module (CBAM): This module was proposed in [35]. With an intermediate feature map, this architecture infers attention in two ways, channel (“what to pay attention”) and spatial (“where to pay attention”). Figure 2 shows the architecture of this module. The Channel attention (M_c) is done as shown in Equation 4 and the Spatial attention (M_s) as exposed on the Equation 5.

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{4}$$

In Equation 4, σ denotes the sigmoid function and MLP is a shared multi-layer perceptron of one hidden layer with an activation size set to $\mathbb{R}^{\frac{C}{r}x1x1}$, where r is the reduction ratio.

$$M_s(F) = \sigma(f^{7x7}([AvgPool(F); MaxPool(F)])) \tag{5}$$

In Equation 5, σ denotes the sigmoid function, and f^{7x7} is a convolutional layer with a kernel size of 7x7.

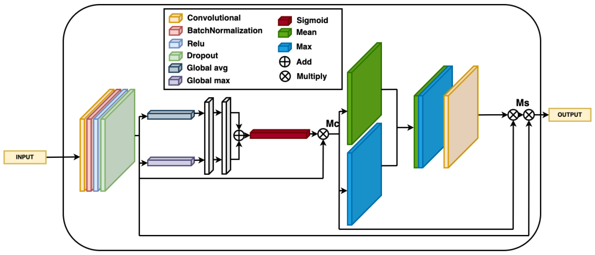


Fig. 2: Convolutional Block Attention Module (CBAM).

2) Self-Attention: After CBAM, we used the Self-Attention module [37], which has been claimed in the literature as one of the best attention procedures when all the information comes from the same input. This module is represented in Figure 3. The input feature map experiences three transformations. The first two are query (q) and key (k), which are used to calculate the attention, where $q(x) = f_q^{1x1}x$ and $k(x) = f_k^{1x1}x$. With these two feature maps, Equation 6 learns to attend to the i^{th} location (pixel) when synthesizing the j^{th} region, building the attention map of the module.

$$\beta_{j,i} = \frac{exp(s_{ij})}{\sum_{i=1}^{N} exp(s_{ij})},\ where\ s_{ij} = q(x_i)^T k(x_j) \tag{6}$$

The output layer of the attention (Equation 7) has the form of $o = (o_1, o_2, ..., o_j, ..., o_N) \in \mathbb{R}^{CxN}$, where C is the number of channels of the input feature map and N is the number of feature locations from the previous hidden layer. This output is calculated after a matricial product of the attention layer ($\beta_{j,i}$) and the last transformation values (v) experienced by the input.

$$o_j = f_h^{1x1}(\sum_{i=1}^{N} \beta_{j,i} v(x_i)),\ where\ v(x) = f_v^{1x1} x \tag{7}$$

Convolution operations, with kernel size of 1, are done to extract the feature maps. These operations are the following ones: f_q^{1x1}, f_k^{1x1} and $f_v^{1x1} \in \mathbb{R}^{C/rxC}$, where r is a decreasing factor to increase efficiency without compromising performance, and $f_h^{1x1} \in \mathbb{R}^{CxC}$.

$$y_i = \gamma o_i + x_i \tag{8}$$

Additionally, we multiply the output by a learnable scalar γ and add the feature input as a residual connection, as can be seen on Equation 8. This scalar is initialized to 0 to provide the network with the ability to firstly rely on the local neighborhood and then assign more weight to the non-local evidence.

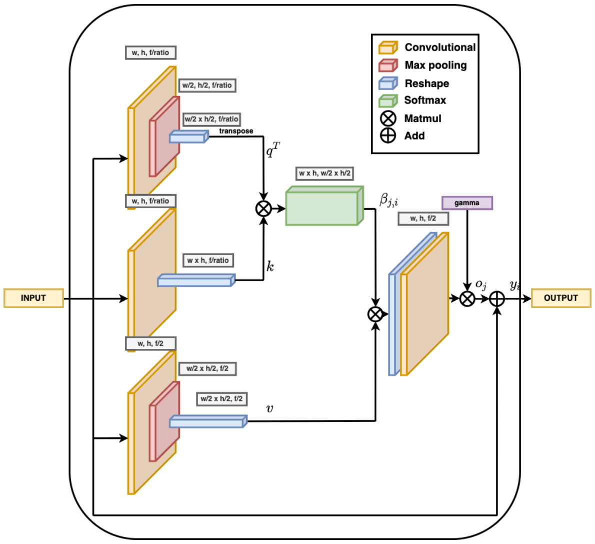


Fig. 3: Self-Attention module

3) Multi-Head Attention module: This module implements a Multi-Head Attention mechanism through a Transformer encoder defined in [12]. Its architecture can be seen in Figure 4. To build the Multi-Head Attention mechanism, M Self-Attention modules, as the above described, are concatenated, which acts as heads in this new architecture. These concatenated feature maps are fused through a convolutional layer f^{1x1} followed by a residual connection and a normalization layer. This mechanism is shown in Equation 9.

$$MultiHead(x_i) = f_o^{1x1}(Concat(head_1, head_2, \dots, head_M))$$
$$where\ head_i = \gamma o_i + x_i \tag{9}$$

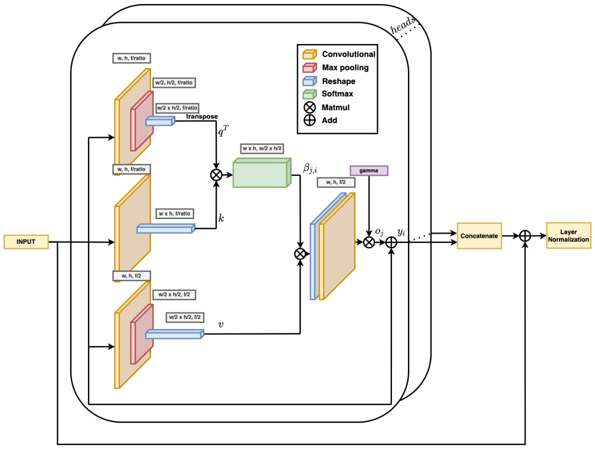


Fig. 4: Multi-Head Attention module.

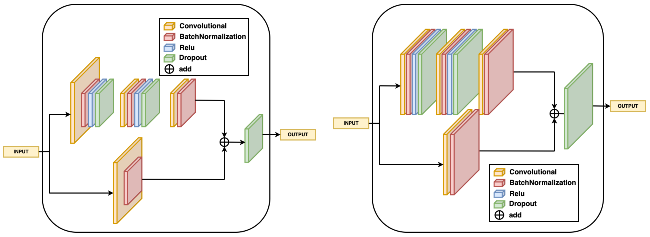
C. Generator architecture.

Using the different attention modules and the learning procedure explained above, as well as a Residual convolutional module with and without stride as the presented in Figure 5, we are ready to tackle the Generator design.

Firstly, a Generator presented in [31] and based on U-Net [32] was implemented. It is an encoder-decoder architecture built by convolutions with skip connections between mirrored layers of the encoder and the decoder. This architecture is a great option for many image to image problems, as information of the input is shared with the output. For example, in semantic segmentation, the output has the same objects shape as the input, so this information could be passed in early steps of the network instead of vanishing it with a deeper network. This information is not so important in our application as the shapes are not maintained. The encoder is built with convolutional modules, which are made up of a Convolutional layer f^{3x3}, a dropout of 0.5 and a Relu activation function. The decoder is made up of upsample modules that are composed of a Convolution Transpose, Batch Normalization, a dropout of 0.5 and a Leaky-Relu activation function. The input is a 256x256 RGB image and the output an 256x256 attention map. These sizes are the same in all implementations. This is the first Generator for our model and will be treated as a baseline and a proof of what we can achieve in the driver attention estimation with an adversarial training.

In our second proposal, we change the convolutional part of the previous Generator for an encoder based on CBAM. The proposed Generator is composed of a first stage module with a convolutional layer (f^{7x7}), Batch Normalization, a Relu activation function and a Max Pooling (f^{3x3}) with stride = 2. After that, sixteen CBAMs modules as the explained before (Figure 2) are used. The odd ones have stride = 1 and pairs ones have stride = 2. Afterwards, a decoder is implemented to upsample the feature maps obtained to get the final output as a 256x256 attention map.

Our third approach is based on Self-Attention mechanism (Figure 3). This generator configuration is a fusion of Residual convolutional downsample modules (Figure 5) with a



(a) Residual convolotutional module (Resblock) with stride = 2 for dimension reduction.

(b) Residual convolotutional module (Resblock) with stride = 1

Fig. 5: Residual convolotutional modules used as image feature extractor.

Self-Attention module. It has the same stride configuration as the CBAM-based Generator. After downsamples and feature extraction are conducted, a Self-Attention module is implemented to extract the most interesting characteristics of these feature maps. This is made to feed the decoder with this key information to obtain the final output as a 256x256 attention map.

Our final proposal takes advance of a Transformer encoder to provide architecture with Multi-Head Attention capability in an efficient way. We have convolutional steps based on Resblocks downsample (Figure 5(a)) and deconvolutional steps that follow the same architecture as the previous tested models. We add some Transformer encoder modules at different steps of the downsample procedure to test the benefits of this architecture.

D. Discriminator architecture

Due to the adversarial training, an additional net is required to complement the Generator. This net has been built following the implementation explained in [31]. In this work, they proposed the use of a convolutional PatchGAN classifier, previously introduced in [45]. This classifier, instead of classifying between "real" or "fake", for the full image, classifies NxN patches, where N is much smaller than the image size. Among its benefits, it gives the possibility for a Discriminator with less parameters, which trains faster and uses larger images.

We concatenate the two feeded images and pass them through some convolutional steps, to end up with the patch classification.

IV. EVALUATION METRICS

To evaluate the proposed models, we have used some evaluation metrics applied in saliency models, which are collected and explained in [46]. From the cited metrics, we will use in our experiments the following: Kullback-Leibler Divergence (KLD), Pearson's Correlation Coefficient (CC) and shuffled Area under the ROC curve (s-AUC). Hereafter, we briefly explain each of them.

Kullback-Leibler Divergence (KLD) shows the difference between two probability distributions. Equation 10 shows this metric where the distributions are P and Q^D, and ε is a regularization constant. In our case, the two comparing distributions correspond to the generated output and the Ground Truth attention map, respectively.

$$KL(P,Q^D) = \sum_i Q_i^D log(\varepsilon + \frac{Q_i^D}{\varepsilon + P_i}) \tag{10}$$

The Pearson's Correlation Coefficient, CC, also called Linear Correlation Coefficient, is a statistical method generally used to measure how correlated or dependent two variables are. CC can be used to consider saliency and fixation maps, P and Q^D, as random variables to measure their linear relationship. Equation 11 shows how it is calculated:

$$CC(P,Q^D)) = \frac{\sigma(P,Q^D)}{\sigma(P) \times \sigma(Q^D)} \tag{11}$$

,where $\sigma(P,Q^D)$ is the covariance of P and Q^D.

Shuffled Area Under the ROC Curve (s-AUC) is a modification of the Area Under ROC Curve (AUC), which has been one of the most used metrics in saliency. The s-AUC assumes that the center bias has not been modeled, and penalizes models where this happened. This metric penalizes when the model only predicts the attention map on the center of the image, due to it is the common area to pay attention.

V. EXPERIMENTS

This section present results obtained for the different proposals. It is divided in two main experiments. Firstly, an ablation study was carried out to test our cGAN architecture with different Generator models. Secondly, the best model has been compared with other state-of-the-art proposals in order to evaluate its real contribution. Architectures were trained using Adam optimizer with a declining learning rate and parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$. Using a Batch size of 8. The GPU used for all the experiments described below is the NVIDIA 2080 Ti with 11GB of VRam. Both experiments have been quantitatively analyzed with the metrics exposed above. Some qualitative results will be shown at the end of this section to give a graphical view of the performance of our work.

A. Ablation study

All architectures were trained for 15 epochs on the BDDA training set dataset. Images and attention maps were resized to 256x256. After the training procedure, a testing step was carried out. This testing was done in two different datasets. Firstly, the BDDA testing set was used, to continue with the DADA-2000 testing set, which is composed of accidental scenarios.

Table I shows some different metrics. It is divided into: accuracy metrics (KLD, CC and s-AUC) and performance metrics (Training time and inference frequency). Performance metrics are not normally evaluated in the literature, but for us, it is a key concept that needs to be evaluated, as it proves the ability of the model to work during inference in real-time. For that, we present training time and inference frequency.

We are gonna evaluate different proposals for the downsample section of the Generator: U-Net (baseline), CBAM, Self-Attention (SA) and Multi-Head Attention (MHA).

Results show that the best model is the combination of the residual modules with the Multi-Head Attention module for the downsample section of the Generator, achieving the best

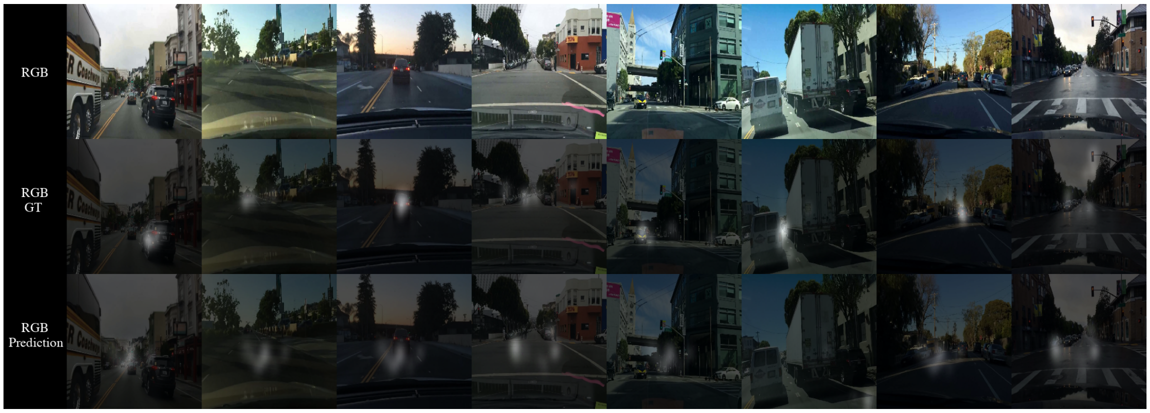


Fig. 6: Qualitative results for the model proposed using the combination of Residual modules + Multi-Head Attention.

TABLE I: Ablation study performed using adversarial learning with different generators configuration. All models have been trained in BDDA and testing in BDDA and DADA2000.

Generator	Performance metrics						Training performance	
	BDDA dataset			DADA dataset				
	KLD ↓	CC ↑	s-AUC ↑	KLD ↓	CC ↑	s-AUC ↑	Training Time	Inference (Hz)
U-Net	0.44	0.91	0.54	0.43	0.94	0.55	10h 45m 25s	7.70
CBAM	0.27	**0.92**	0.68	0.27	0.94	0.69	6h 12m 39s	7.67
Resisual modules + Self-Attention	0.13	0.87	**0.71**	0.22	0.96	**0.70**	**5h 26m 24s**	8.70
Residual modules + Multi-Head Attention	**0.05**	**0.92**	0.66	**0.10**	**0.97**	0.65	6h 30m 52s	**10.53**

performance in both datasets for all metrics except for s-AUC and training time, where using the Self-Attention module gives the best values. Moreover, the use of attention modules has decreased the training time, which is an important fact in terms of research purposes.

B. Comparison with other state-of-the-art models

The best model obtained in the ablation study is compared with some other popular driver attention estimation implementations. Results are shown in Table II. Our proposal obtains better performance results than the referenced implementations in the two challenging datasets for KLD and CC metrics. And it is on par with the best one evaluated in s-AUC. Moreover, unlike most proposals of the literature, our application shows inference times that can be used in real-time applications.

Figure 6 shows qualitative result for our model, the images are displayed in the following order: the first row is composed by the RGB image, the second one by the RGB image with the ground truth (humans fixations) and the third one is the RGB image with the ARAGAN's predicted attention map.

VI. CONCLUSIONS AND FUTURE WORKS

We presented ARAGAN, a novel architecture that combines adversarial learning with Multi-Head Attention, two of the most advanced techniques in DL, to obtain driver attention estimation.

Our proposal outperforms state-of-the-art results in a comparison experiment with other popular models in two different challenging datasets (BDDA and DADA2000). An extensive study of different architectures for the Generator design and the results obtained for each of them in the referenced datasets have been presented. This architecture enables the use of driver's attention in real driving applications due to its low inference time, and may be helpful in the design of autonomous vehicles.

For future works, we plan to explore the use of positional encoding at the beginning of the Transformer module, as well as the implementation of a Transformer decoder, which has not been tackled in this work. Moreover, we plan to test ARAGAN in real environments using our electric autonomous vehicle.

REFERENCES

[1] S. Grubmüller, J. Plihal, and P. Nedoma, "Automated driving from the view of technical standards," in *Automated driving*, pp. 29–40, Springer, 2017.

[2] T. Deng, K. Yang, Y. Li, and H. Yan, "Where does the driver look? top-down-based saliency detection in a traffic driving environment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 2051–2062, 2016.

[3] F. Katsuki and C. Constantinidis, "Bottom-up and top-down attention: different processes and overlapping neural systems," *The Neuroscientist*, vol. 20, no. 5, pp. 509–521, 2014.

[4] A. J.-W. Chen, M. Britton, G. R. Turner, J. Vytlacil, T. W. Thompson, and M. D'Esposito, "Goal-directed attention alters the tuning of object-based representations in extrastriate cortex," *Frontiers in human neuroscience*, vol. 6, p. 187, 2012.

[5] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proceedings of the IEEE international conference on computer vision*, pp. 2722–2730, 2015.

[6] J. S. Perry and W. S. Geisler, "Gaze-contingent real-time simulation of arbitrary visual fields," in *Human vision and electronic imaging VII*, vol. 4662, pp. 57–69, International Society for Optics and Photonics, 2002.

[7] J. Guo, U. Kurup, and M. Shah, "Is it safe to drive? an overview of factors, metrics, and datasets for driveability assessment in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3135–3151, 2019.

[8] A. Palazzi, D. Abati, F. Solera, R. Cucchiara, *et al.*, "Predicting the driver's focus of attention: the dr (eye) ve project," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1720–1733, 2018.

TABLE II: Comparison with other state of the art models trained in BDDA and testing in BDDA and DADA2000

Architecture	Performance metrics					
	BDDA dataset			DADA dataset		
	KLD ↓	CC ↑	s-AUC ↑	KLD ↓	CC ↑	s-AUC ↑
DR(eye)VE [8]	1.95	0.5	-	3.37	0.29	0.63
BDDA [9]	1.24	0.59	-	2.5	0.4	**0.66**
MEDIRL [30]	2.51	0.74	**0.67**	2.93	0.63	0.61
ACT-Net [42]	1.11	0.63	-	-	-	-
Residual modules + Multi-Head Attention (ours)	**0.05**	**0.90**	0.66	**0.10**	**0.97**	0.65

[9] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney, "Predicting driver attention in critical situations," in *Asian conference on computer vision*, pp. 658–674, Springer, 2018.

[10] J. Fang, D. Yan, J. Qiao, J. Xue, H. Wang, and S. Li, "Dada-2000: Can driving accident be predicted by driver attentionƒ analyzed by a benchmark," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 4303–4309, IEEE, 2019.

[11] F. Lateef, M. Kas, and Y. Ruichek, "Saliency heat-map as visual attention for autonomous driving using generative adversarial network (gan)," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[13] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, *et al.*, "Swin transformer v2: Scaling up capacity and resolution," *arXiv preprint arXiv:2111.09883*, 2021.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[16] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

[17] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[19] A. Linardos, M. Kümmerer, O. Press, and M. Bethge, "Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12919–12928, 2021.

[20] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in *European Conference on Computer Vision*, pp. 419–435, Springer, 2020.

[21] G. Ding, N. Imamouglu, A. Caglayan, M. Murakawa, and R. Nakamura, "Salfbnet: Learning pseudo-saliency distribution via feedback convolutional networks," *arXiv preprint arXiv:2112.03731*, 2021.

[22] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark," 2015.

[23] M. Kummerer, T. S. Wallis, L. A. Gatys, and M. Bethge, "Understanding low-and high-level contributions to fixation prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4789–4798, 2017.

[24] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2798–2805, 2014.

[25] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," *CVPR 2015 workshop on "Future of Datasets"*, 2015. arXiv preprint arXiv:1505.03581.

[26] A. Palazzi, F. Solera, S. Calderara, S. Alletto, and R. Cucchiara, "Learning where to attend like a human driver," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, pp. 920–925, IEEE, 2017.

[27] A. Tawari and B. Kang, "A computational framework for driver's visual attention using a fully convolutional architecture," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, pp. 887–894, IEEE, 2017.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[29] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "Dada: Driver attention prediction in driving accident scenarios," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[30] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13178–13188, 2021.

[31] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[32] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[33] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[34] D. Britz, A. Goldie, M.-T. Luong, and Q. Le, "Massive exploration of neural machine translation architectures," *arXiv preprint arXiv:1703.03906*, 2017.

[35] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[37] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*, pp. 7354–7363, PMLR, 2019.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[39] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?," in *International Conference on Machine Learning*, pp. 5389–5400, PMLR, 2019.

[40] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 302–321, 2019.

[41] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[42] C. Gou, Y. Zhou, and D. Li, "Driver attention prediction based on convolution and transformers," *The Journal of Supercomputing*, pp. 1–17, 2022.

[43] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," *arXiv preprint arXiv:1511.05440*, 2015.

[44] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

[45] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *European conference on computer vision*, pp. 702–716, Springer, 2016.

[46] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 3, pp. 740–757, 2018.