Leveraging Driver Attention for an end-to-end Explainable Decision-making from frontal images

Javier Araluce[1], Luis M. Bergasa[1], Manuel Ocaña[1], Ángel Llamazares[1] and Elena López-Guillén[1]

***Abstract*—Explaining the decision made by end-to-end autonomous driving is a difficult task. These approaches take raw sensor data and compute the decision as a black box with large deep learning models. Understanding the output of deep learning is a complex challenge due to the complicated nature of explainability; as data passes through the network, it becomes untraceable, making it difficult to understand. Explainability increases confidence in the decision by making the black box that drives the vehicle transparent to the user inside. Achieving a Level 5 autonomous vehicle necessitates the resolution of that challenging task.**

In this work, we propose a model that leverages the driver's attention to obtain explainable decisions based on an attention map and the scene context. Our novel architecture addresses the task of obtaining a decision and its explanation from a single RGB sequence of the driving scene ahead. We base this architecture on the Transformer architecture with some efficiency tricks in order to use it at a reasonable frame rate. Moreover, we integrate in this proposal our previous ARAGAN model [1], which obtains SOTA attention maps, to improve the performance of the model thanks to understand the sequence as a human does. We train and validate our proposal on the BDD-OIA dataset, achieving on-pair results or even better than other state-of-the-art methods. Additionally, we present a simulation-based proof of concept demonstrating the model's performance as a copilot in a close-loop vehicle to driver interaction.

***Index Terms*—Driver attention, decision-making, explainability, deep learning, self-driving**

I. Introduction

Vehicle intelligence is primarily located within the decision-making layer [2]. This executive layer is responsible for executing the necessary manoeuvres during an event. According to current literature, there exist two approaches to building the decision-making layer: modular and end-to-end. Modular approaches divide the whole AV architecture into a modular pipeline [3], [4]. On the other hand, end-to-end proposals execute the entire driving task via a single neural network ("black box") that takes in the raw sensor data and produces the driving manoeuvre (actions) as output [5], [6]. This second approach is mainly relies on deep learning techniques due to the complexity of the task.

Two concepts summarise the explainability of the previous approaches. The former has the advantage that the different modules of the architecture explain the decision-making. However, these approaches tend to propagate errors to subsequent pipeline modules, which means that the error is cumulative. Key modules or layers are: perception, localisation, mapping, planning, decision-making and control.

On the flip side, end-to-end proposals avoid the issue of error propagation. However, they lack interpretability, making it difficult to explain errors. Explaining end-to-end proposals is a complex task and has recently been tackled through the Explainable AI (XAI) approach. Several techniques such as Shapley values [7] or saliency maps [8], have been employed to calculate the influence of each input feature on the model output. Our approach deviates from these methods because it does not explain the input features.

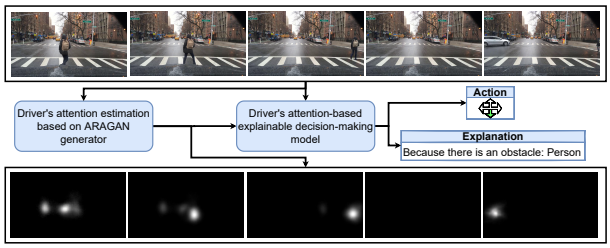

Fig. 1: Driver's attention-based explainable decision-making framework.

In order to achieve this objective, we propose a driver attention-based explainable decision-making model to predict the vehicle's actions and explains them. Our proposed framework is illustrated in Figure 1. Our ARAGAN Generator [1] predicts the driver's attention for the input sequence. Using only the input image sequence and the driver's attention sequence our model is able to predict the ego-vehicle decision and its explanation. Unlike the state-of-the-art (SOTA) methods [9], [10], which rely on a single image (i.e. the last one in the sequence) to make decisions, our approach considers the entire sequence. Figure 1 highlights the significance of this approach since it accounts for sequential events that occur throughout the sequence. For example, in the first three frames a pedestrian is crossing the street, but in the final two frames they are out of view of the camera. The automobile ought to apply the brakes in this scenario. Nevertheless, a comprehensive judgement of the action and explanation behind this behaviour cannot be determined solely based on the final frame due to the lack of temporal analysis.

We propose a deep learning model for vehicle-driver interaction based on driver attention. The model predicts decisions and explanations from a video sequence captured by the vehicle's forward-facing camera. The proposed model has potential applications in the field of autonomous driving. This information can assist the driver in performing manual manoeuvres (acting as a copilot), or guide imitation or reinforcement learning of autonomous driving. The main contributions of

Javier Araluce, Luis M. Bergasa, Manuel Ocaña and Ángel Llamazares ({javier.araluce, luism.bergasa, manuel.ocanna, angel.llamazares, elena.lopezg}@uah.es) are with the Department of Electronics, University of Alcalá (UAH), Spain.

This work has been supported from the Spanish PID2021-126623OB-I00 project, funded by MICIN/AEI and FEDER, TED2021-130131A-I00, PDC2022-133470-I00 projects, funded by MICIN/AEI and the European Union NextGenerationEU/PRTR, and ELLIS Unit Madrid funded by Autonomous Community of Madrid.

our proposal are:

- We leverage driver's attention to obtain an explainable decision. We generate the attention maps with our previous work ARAGAN. We prove that the driver's attention helps to get a better explanation of the decision.
- We propose a novel architecture based on the Transformer Encoder to predict the explainable decision from an RGB urban road sequence. It understands the spatio-temporal features in an efficient way and uses cross-attention to take into account the decision to get the explanation.
- Our proposal is verified on an open source dataset (BDD-OIA [9]) outperforming current standard methods in terms of explainability. This is possible due to the addition of temporal context and the additional information provided by the driver's attention in the sequence.
- We utilize the CARLA simulator in conducting proof-of-concept testing and validating our model as a copilot in close-loop vehicle to driver interaction. This digital environment enables the evaluation of system performance in different scenarios beyond the training dataset achieving on pair results with the obtained in the dataset.

II. Related works

The literature review section is split into three subsections to summarise the explainability of autonomous driving using deep learning. The first part consists of a study of the open-source datasets within this topic to show the available data to accomplish this task. The second section describes approaches trained in the same dataset as this proposal to show their strengths and weaknesses. And we end the literature review with some Transformers architecture in the autonomous vehicles research field.

A. Datasets

Training deep learning models in a supervised way requires a huge amount of data, as models learn to generalize from seen data. For this purpose, many datasets are open-source to reach different tasks. In the context of autonomous vehicles, object detection, semantic segmentation, and driver attention are usual tasks that use deep learning. These AV tasks have their own datasets that allow training supervised deep learning models.

Another task that needs its own dataset is the explainability of the decision-making modules. There are different open-source datasets in the literature to learn models for this task. Table I summarises them with their sizes and the annotation method. BDD-X [11] provides a textual explanation of the decision making with their heatmaps. DoTA [12] collects temporal, spatial and categorical annotations of road anomalies. CTA [13] is composed of traffic accidents with their causes and effects annotated. HDD [14] was annotated with a 4-layer annotation scheme: Goal-oriented action, Stimulus-driven action, Cause and Attention. BDD-A Extended [15] is an extension of BDD-A dataset [16] with first-person explanations.

Among them we have chosen BDD-OIA dataset [9] as it provides a sequence with the actions and the explanations that the driver performed during a complex driving manoeuvre. Moreover, it is a double classification problem instead of a textual task, which makes the task less complex. The prediction of an explanation in text form, as done in some datasets exposed above, is within the Natural Language Processing (NLP) paradigm, which is out of the scope of this work, focused on computer vision techniques.

The selected videos have various weather conditions recorded at different times of the day. Their dataset has 22,924 5-second video clips annotated for 4 actions (Move forward, Stop/Slow down, Turn left and Turn right) and 21 explanations (i.e. Traffic light is green, No lane on the left, On the right-turn lane, Obstacle: rider, etc.). This annotation method allows the definition of a multi-label classification problem, which is easier to train because the outputs are bounded. However, dealing with this dataset is complicated because the same video has more than one possible output.

B. State-of-the-art models trained in BDD-OIA

There are some models in the literature trained on the BDD-OIA dataset. The following section describes these architectures used to solve the explainability problem in autonomous vehicles.

The first proposal [9] introduced the prediction of object-induced actions and their explanations. They used an object detection model and a second step called the Action-Inducting Object Selection module. The object detection module is the Faster R-CNN [17], previously trained in BDD100K [18] to perform this task. After this training, they froze it to continue with the second step of the model. They fed the local features of each object (bbox) to the second module that fuses them with the global characteristics (full image) extracted by the Faster R-CNN to select the best detections to complete the task. The Global module produces a feature map of the entire scene that provides context. This shows where the network focuses. We propose a system that uses attention based on human gaze.

The authors proposed a multi-task loss to train the model as the shown in Equation 1. It is a combination of the Binary Cross Entropy (BCE) (Equation 2) and a hyperparameter λ that controls the influence of the explanation. After conducting an ablation study to determine the optimal choice, they set λ to 1. We will retain the same configuration but modify the binary cross-entropy (BCE) loss function with an alternative option that will be elaborated on subsequently.

$$\mathcal{L} = \mathcal{L}_{BCEaction} + \lambda \cdot \mathcal{L}_{BCEexplanation} \tag{1}$$

$$\mathcal{L}_{BCE}(y, x) = \frac{1}{N} \sum_{i=0}^{N} [y \cdot log(x) + (1 - y) \cdot log(1 - x)] \tag{2}$$

They fed the model with the last frame of the recorded sequence to obtain the action and the explanation for the entire clip. They compared their result with other state-of-the-art works (ResNet-101 [19] and the network of [20]), outperforming them with their proposal. They don't have temporal information, which clarifies many situations that occur when driving.

TABLE I: Driving datasets to develop explanation methods for AVs and the stakeholders that would potentially benefit from such explanations.

Dataset	Size	Annotation & Explanation
BDD-OIA [9]	23K × 5s	Actions and *Why explanation*
BDD-X [11]	7K × 40s	Textual *Why explanation* associated with videos segments with heatmaps
DoTA [12]	4,677 videos (73,193s)	*What explanation* (Temporal and spatial anomaly identification with bounding boxes)
CTA [13]	1,935 × 17.7s	Why explanation for accidents with cause and effects
HDD [14]	374,400s	*What explanations* for driver actions
BDD-A Extended [15]	$1,103 \times 10s$	Human gaze inciting *why and/or what explanation*, explanation necessity score

The second proposal [10] claimed that [9] has two shortcomings: 1) It can "only" have a clear look into those cropped-out regions. These regions correspond to the detected object. 2) The proposed attention mechanism depends on the Faster R-CNN. Therefore, the error will propagate to all other modules.

They proposed a Global Soft Attention (GSA) mechanism to overcome these problems. It allows the model to search for correlations more deeply. To complete this proposal, they proposed a model based on two feature extractors (Resnet50 [19] and Mobile Net [21]), which feeds a Multi-Head Self Attention. Therefore, GSA attends to the "raw" features instead of only a few ones (ROIs). They used the same loss as in the previous proposal based on Equation 1. Nevertheless, they do not take into account the temporal information as they only processed the last frame of the sequence, which does not explain the complete sequence. We propose to understand the temporal sequence in an efficiency way to understand and explain the decision.

C. Transformers in the autonomous vehicles research field

The application of transformer architectures in autonomous vehicles' decision-making processes is a burgeoning area of research. Transformer models are explored for end-to-end driving, trajectory prediction, scene understanding, and semantic segmentation tasks, offering the potential to simplify and enhance autonomous driving systems. Researchers investigated techniques for multi-modal sensor fusion, efficiency improvements, and the integration of Reinforcement Learning with transformers. Subsequent iterations extend the application of the attention mechanism to the realm of driving, encompassing tasks such as motion projection [22]–[24], forecasting driver focus [25], [26], and monitoring object movement [27], [28]. In the context of fully autonomous driving systems, TransFuser [29], [30] leverages multiple transformer components to merge information sourced from the frontal camera and LiDAR sensors. We propose a framework that comprehends the spatio-temporal information supplied by the RGB sequence and attention maps, incorporating effective transformers architectures for viable processing of the data at a sensible frame rate.

III. Theoretical Background

This section presents the theory behind comprehending the attention modules of the current proposal, which is a key architectural contribution of this work.

Attention mechanisms in deep learning are a way to selectively focus on certain parts of the input, allowing the model to make more informed decisions. The main idea behind attention is to weigh the importance of different input sections so that the model can focus on the most relevant information. Attention mechanisms are valuable in tasks with large and complex input. The attention mechanism allows to improve model's performance and to reduce the amount of required computation. Various tasks, including image captioning, machine translation, and language modelling, have incorporated attention mechanisms [31], [32].

1) Self-Attention: Self-Attention computes the relationships inside the input itself, attending to different input parts. We will explore with linear layers and with the scaled-dot product [33]. We define the attention weights a_i in Equation 3 to the input feature vector x_i.

$$a_i = \sigma\left(\frac{Q(x_i)K(x_i)^T}{\sqrt{d_k}}\right) \tag{3}$$

Where σ is the softmax function; Q is the query; K is the key; and d_k is the scaling factor. With that said, We compute the attended output z as:

$$z = a_i V(x_i) \tag{4}$$

It is an attention mechanism where the dot products are scaled down by $\sqrt{d_k}$. If we assume that Q and K are d_k-dimensional vectors whose components are independent random variables with mean 0 and variance 1, then their dot product, $Q \cdot K = \sum_{i=1}^{d_k} u_i v_i$, has mean 0 and variance d_k. Since we would prefer these values to have variance 1, we divide by $\sqrt{d_k}$.

2) Cross-Attention: Cross-Attention combines asymmetrically two separate embedding sequences of the same dimension. Generally, in Cross-Attention, the key (K) and the value (V) are computed from one input and the query (Q) from another. A well-known example of this type of attention is Transformer [33], which implements this technique in the Decoder section. In this case, the attention weight a_i for the input feature vectors x_i and y_i is computed as follows:

$$a_i = \sigma\left(\frac{Q(y_i)K(x_i)^T}{\sqrt{d_k}}\right) \tag{5}$$

In this method, the query is derived from a source distinct from the key and value, while σ represents the softmax function, Q denotes the query, K denotes the key, and d_k represents the scaling factor. We compute the attended output z as:

$$z = a_i V(x_i) \tag{6}$$

3) Multi-Head Attention: Multi-Head Attention allows parallel computing and provides the ability to pay attention to different aspects in the same feed.

It is a robust architecture for understanding features and their relationships. This work explores it in two variants: build with Self-Attention and Cross-Attention. In this case, we compute the attention weight a_i for each input element x_i as follows:

$$\begin{aligned} a_i &= Concat(head_1, head_2, \ldots, head_M) \\ where \quad head_i &= \sigma\left(\frac{Q(x_i)K(x_i)^T}{\sqrt{d_k}}\right)V(x_i) \end{aligned} \tag{7}$$

Where M is the number of heads, and z is computed by a feed-forward neural network with parameters θ_k as:

$$z = f_{\theta_k}(a_i) \tag{8}$$

4) Transformer Encoder: The Transformer Encoder [33] is a neural network architecture that uses attention mechanisms to process sequential input data. The Transformer Encoder was presented as an improvement of Recurrent Neural Networks (RNNs) in natural language processing tasks, including but not limited to machine translation and language modelling.

The Transformer Encoder consists of a stack of layers, each with two sub-layers: a Multi-Head Self-Attention mechanism and a feed-forward neural network. The Multi-Head Self-Attention mechanism allows the model to attend to different parts of the input and to compute a weighted sum of the input elements, where the weights are calculated based on the similarity between the input elements. The feed-forward neural network then processes the output of the Multi-Head Self-Attention mechanism to produce the final attended feature vector.

In addition, the Transformer Encoder incorporates a mechanism known as positional encoding, enabling the model to consider the sequence of input elements. In this way, the model can understand the spatial context of the input without fading this information through its layers.

The Transformer Encoder has shown to be very effective in NLP tasks, outperforming traditional RNN-based models. The main advantage of the Transformer Encoder is its ability to handle long-term dependencies in the input data, which is particularly useful in tasks such as machine translation and language modelling. Additionally, the Transformer Encoder can parallelize the computations, which makes it more efficient than traditional RNN-based models. The problem is that it takes many GPUs to calculate these relationships from raw data.

IV. Driver Attention-based Explainable Decision-Making Model

The following section explains the main contribution of this work, the Driver Attention-based Explainable Decision-Making model built to assist the driver or to help the learning of autonomous vehicles. It predicts the decision-making of the ego-vehicle, giving an explanation based on the driver's attention using our ARAGAN [1].

The input data of this model is a sequence of images extracted from the BDD-OIA dataset [9]. The model predicts an action expressed as a four-dimension vector and an explanation as a 21-dimension vector. This section will expose the model architecture based on the theory previously explained.

Figure 2 shows the architecture and modules of our proposal. We explain these modules individually to understand the idea behind this proposal and the data flow in the model. In contrast to [9], [10], the input model is a video sequence from the dataset instead of the last frame. It makes the model heavier, but with efficient modules, this proposal outperforms other architectures in terms of efficiency.

The model is composed of a Feature extractor or backbone; driver's attention fusion based on the ARAGAN Generator; a Spatio-Temporal Transformer Encoder; and the classification heads.

A. Feature extractor

This technique has been widely used in the literature [34]. The most used dataset for this task is ImageNet, composed of 1,281,167 training images, 50,000 validation images and 100,000 test images distributed among its 1,000 classes. Lately, Google has been using a private dataset called JFT-300M [35]. We employ the MobileNetV2 [21], which is a mobile architecture that does not increase the computational cost of the final model. It is a computer vision backbone composed of convolutional and residual bottleneck layers. The used version has 2,261,632 parameters trained in Imagenet.

Using a backbone as an initial feature extractor makes it possible to work with this kind of dataset that has not got much data.

B. Driver's attention fusion based on ARAGAN

In addition to the RGB data, our proposal utilises the driver's attention as input for our transparent decision-making model. To achieve this, we implement the ARAGAN generator, previously trained and developed on the BDD-A dataset [16]. After training, we assessed the performance of the model in two datasets: BDD-A [16] and DADA 2000 [36], using the following metrics: Kullback-Leibler Divergence (KLD), Pearson's Correlation Coefficient (CC) and shuffled Area under the ROC curve (s-AUC). In BDD-A it obtained KLD = 0.05, CC = 0.92 and s-AUC = 0.66, and in DADA2000 it obtained KLD = 0.1, CC = 0.97 and s-AUC = 0.65. We refer the reader to our previous study [1] for more details about that model.

This work has not got pre-processing, and the feed-forward of the driver attention net is done online. We resize the generated attention map to the output of the feature extractor (7×7). After that, we pass the attention map sequence through convolutional layers, one for each clip image. The kernel size is (7×7) to have the complete resized attention map into account. We multiply the result of this layer with the backbone output and concatenate it. With this method, we use

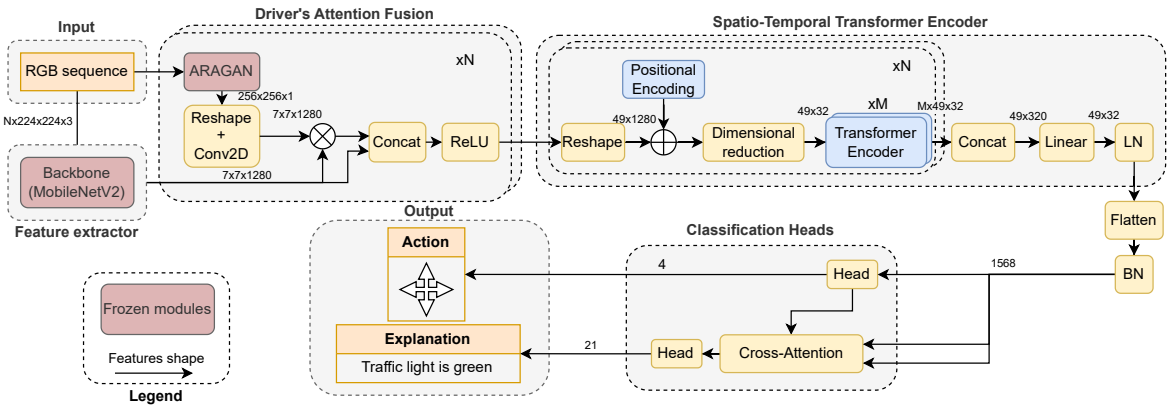

Fig. 2: Driver Attention-based Explainable Decision-Making model diagram. It represents the complete forward pass from the RGB sequences to the final decision and explanation output. The model is composed of 4 modules: 1) Feature extractor, 2) Driver's Attention Fusion, 3) Spatio-Temporal Transformer Encoder and 4) Classification Heads. Note that the shape features are missing from the Batch Size dimension, which has been set to 512.

the driver's attention to understand the output characteristics of the backbone. Moreover, we pass the complete information extracted by the backbone to the following network steps. Finally, we employ a ReLU activation function.

In the complete architecture, we repeat this module N times in parallel, one for each clip image. They do not share weights between different timestamps.

C. Spatio-Temporal Transformer Encoder

After the driver's attention module, the following section shows the Spatio-Temporal Transformer Encoder. It understands the relationships between the features from the previous modules and filters them before passing them to the last network step.

It is composed of N modules that compute in parallel for each sequence frame. Each parallelised module has as input the concatenation of the multiplied attention features and the backbone features, which we flatten using a reshaping method. After that, we employ positional encoding to preserve the spatial information throughout the module, which follows Equation 9. The positional encoding is a vector added to the features to indicate its position in the sequence (feature map). The underlying idea is that it generates a unique coding for each location based on a combination of sine and cosine waves with different frequencies.

$$\vec{p_t} = f(t) := \begin{cases} PE_{(pos,2i)} = sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \\ PE_{(pos,2i+1)} = cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \end{cases} \quad (9)$$

After the positional encoding, an intermediate layer is necessary to reduce computational costs. For this reason, we perform a dimensional reduction. The Transformer encoder has a high memory consumption if the input tensor is too big. The reduction is carried out linearly, with a Linear layer of 32 units.

With the positional encoding and the dimensional reduction explained, a Transformer Encoder is applied, which encodes the Spatio-Temporal features of the sequence. We build it with N Encoders, one for each frame and it has a stack of $M = 8$ identical layers.

It has a Layer Normalization, followed by Multi-Head Self-Attention previously explained, and a Dropout layer. This branch has a residual connection, a shortcut connection that allows the gradient to flow directly from one layer to another, bypassing one or more intermediate layers. After that, we employ Layer Normalization followed by a Multi-Layer Perceptron module (MLP) to end with another residual connection.

We explained the Multi-Head Self-Attention theory earlier, and the only difference here is how the heads are parallelised. We calculate them in a single forward pass instead of one by one. For this purpose, we reshape the outputs of the Linear layers to have Multi-Head computing. After that, the features pass a final MLP composed of a Linear layer, a GELU activation function, a Dropout layer, another Linear layer, and a Dropout.

After that, we concatenate the features and pass them through a Linear layer and Layer Normalization to output the temporal attended features. The Spatio-Temporal Transformer Encoder proposed to encode the spatio-temporal features. It learns the temporal information of the video sequence and the spatial distribution along the image to understand the localisation of the relevant objects and make the best decision. It allows parallel computing efficiently. Moreover, the driver's attention module filters fed features to this module. We compute the image features using a pre-trained backbone, which allows using a low-dimensional encoder.

D. Classification heads

The last modules are the classification heads composed of two branches, one for each output (Action, Explanation). The first branch is in charge of the Action output. It is a standard classification module built of a Linear layer, followed by Batch Normalization, a ReLU activation function, Dropout, another Linear layer and a sigmoid activation function to end the branch. The output is a 4-dimensional vector. The last activation function is sigmoid because the problem to solve is a multi-label classification problem. Unlike other classification methods [37], the softmax function cannot be used as it is unable to predict output for multiple labels.

The second branch, which predicts the Explanation layer, is quite different because the input of this branch is the latent vector that comes from the Spatio-Temporal Transformer Encoder and the action output. Multi-Head Cross-Attention is the first module that models the relationship between the action output and the latent vector. The reason for this procedure is that explanations are correlated with the action, with each action having its own set of explanations.

After the Multi-Head Cross-Attention, we pass the attended latent vector to a classification section that follows the same structure as the Action branch, but in this case with a 21-dimensional vector as output.

V. Experimental Setup

The following section explains the experimental setup of this proposal, with the different loss functions and metrics used to evaluate it.

A. Loss functions

A loss function in deep learning is a function that measures the difference or distance between the predicted output and the actual output. The objective of training a neural network is to minimise the loss function value. Common examples of loss functions include mean squared error and cross-entropy. These loss functions are used to evaluate the model performance during training to optimise the model's parameters.

The task under study is a multi-label classification composed of two outputs. In this work, the aim is to predict multiple labels or classes for each input sample.

Some architectures in the state-of-the-art [9], [10] used the BCE as their loss function. This work proposes to use the F1-score loss to find the optimal threshold that maximises this metric as will be shown in the results section. Moreover, due to the unbalanced dataset, the loss is weighted for the action labels (Turn Left and Turn Right).

B. Metrics

Metrics provide a way to quantitatively assess how well the model makes predictions on unseen data. In other words, metrics measure the model's accuracy, robustness, and generalisation capabilities.

In deep learning, many metrics based on the specific task and the nature of the data can be used. The most commonly used metrics are: accuracy, precision, recall, F1-score, area under the ROC curve (AUC), etc. Each metric has its own strengths and weaknesses, and choosing the appropriate metric for the specific task is significant.

In general, it is crucial to consider both the overall performance of the model as well as its performance on subsets of the data when evaluating the performance of a deep learning model. In this work we use the following.

F1-score (Equation 10). It is the harmonic mean of precision and recall. We use it to balance the trade-off between precision and recall. We calculate it as shown in equation 10.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (10)$$

This proposal evaluates the models with two variations of this metric. Firstly the models are tested using Equation 11, which computes the F1-score for all the predictions. It averages the F1-score over all the forecastings.

$$F1_{all} = \frac{1}{|A|} \sum_{j=1}^{|A|} F1(\hat{A}_j, A_j) \quad (11)$$

The other variation computes the mean F1-score of each label (action or explanation), represented in Equation 12.

$$mF1 = \frac{1}{|A|} \sum_{j=1}^{|A|} \frac{1}{|C|} \sum_{i=1}^{|C|} F1(\hat{A}_j^i, A_j^i) \quad (12)$$

Finally, the mean of both $F1_{all}$ (action and explanation) is computed to show the best model in both metrics. This results follows Equation 13.

$$\overline{F1_{all}} = \frac{F1_{all}^{decision} + F1_{all}^{explanation}}{2} \quad (13)$$

We will use these metrics to compare the proposed modules and the comparison against other proposals from the literature.

VI. Results

In this section, we present the obtained results from different experiments. We evaluate the proposal in two ways. Firstly, we expose a comparison with main state-of-the-art proposals. To continue with the explained proposal module by module as an ablation study to determine the best option and the contribution of each module previously explained is carried out. After this, we present a discussion about the results obtained in the dataset. With that model, we propose a proof of concept to evaluate the proposal as a copilot in a simulation environment using CARLA Simulator [38].

A. Comparison with state-of-the-art proposals

This section compares the driver's attention-based explainable decision-making proposal with the main state-of-the-art models that have published their results. Denote that [9] provides the code and the weights to its model, so we have tested it in the same experimental setup as this proposal. However, [10] do not provide code, and we could not verify the published results.

As we depict in Table II, our proposal gives the best $F1_{all}$ for the explanation. Regarding the action $F1_{all}$, it is only ten tenths behind [9], [19], [20] and two tenths behind [10]. We have added the $\overline{F1_{all}}$ to measure the average score of both results (action and explanations). About this metric, our approach achieves the best outcome for the tested models, and it is one-tenth behind [10].

Finally, we have compared the inference time, and both tested models provide the same values. They work at 5 Hz. However, [9] processes only the last frame, and our driver's attention-based explainable decision-making processes ten sequence frames. We have not tested the other proposals because the code is not available.

TABLE II: Action and explanation prediction performance using the driver's attention-based explainable decision-making. It will be compared with other proposals in the literature. Results in blue colour represent the best of the evaluated models. **Bold** results represent the best over all the proposals.

Models	Action						Explanation		$\overline{\mathbf{F1_{all}}}$	Inf. (ms)
	F	S	L	R	mF1	$\mathbf{F1_{all}}$	mF1	$\mathbf{F1_{all}}$		
ResNet-101 [19]	0.755	0.607	0.098	0.108	0.392	0.601	0.180	0.331	0.466	-
Local selector [20]	0.810	0.762	0.600	0.624	0.699	0.711	0.196	0.406	0.578	-
† FRCNN + BDD-OIA [9]	**0.829**	0.781	**0.630**	**0.634**	0.718	**0.734**	0.208	0.422	0.57	199
GSA (resnet) [10]	-	-	-	-	**0.750**	0.729	**0.644**	0.525	**0.627**	-
GSA (mobilenet) [10]	-	-	-	-	0.746	0.718	0.642	0.531	0.624	-
† Driver attention-based explainable decision-making (ours)	0.634	**0.802**	0.475	0.460	0.593	0.695	0.267	**0.538**	0.617	**199**

† indicates that the proposal has been tested on the same computer and the results have been verified.
F = Forward; S = Stop; L = Left; R = Right

B. Ablation study

We have explained various modules for our architectures, so we carry out an ablation study to known the contribution of each module according to the following items:

- The use of a sequence instead of the last frame.
- The contribution of adding driver attention to the pipeline.
- The advantage of using the Transformer Encoder to understand the Spatio-Temporal relationships.
- The use of Cross-Attention between the action head and the latent vector to obtain the explanation.
- Loss function influence, BCE vs F1-score loss.

We expose the ablation study in Table III. It displays the contributions of individual modules to the final model, allowing for comparison with alternative solutions. It shows the number of parameters for each model (Par. (M)). We employ for the evaluation the metrics explained above. We calculate the F1-score individually for each action and their mean ($mF1$). After that, we compute the F1-score for all the decisions ($F1_{all}$). Finally, we calculate the mean F1-score ($mF1$) for each explanation and the F1-score ($F1_{all}$) for all of these predictions. The last exposed metric is the inference time of each proposal (Inf.).

We have carried out this study with the same experimental setup for every model. Moreover, the seed has been set to a deterministic value to avoid randomness and not reproducibility between experiments.

We have trained every model with a batch size of 512, an Adam optimizer ($b_1 = 0.9$, $b_2 = 0.999$ and $\epsilon = 1e-7$) and have set the learning rate to 0.001 with an Exponential Decay scheduler. The GPU used for training is an NVIDIA A100 with 80 GB of VRAM. Moreover, we have measured the inference time with an NVIDIA 2080 Ti with 11 GB of VRAM, which is more common to have by other researchers in order to compare with their proposals. We have configured the training with an early stopping of 3 epochs to avoid overfitting. Regarding the dataset, we have trained all the models with the training and validation set shuffled, and the exposed results correspond to the testing set. Denote that the dataset has not got the same amount of sequences as the authors claimed in the paper, where the dataset was presented [9]. The dataset is composed of 7,946 videos for training, 1,117 for validation and 2,236 for testing, instead of a training set of 16,082 images, a validation set of 2,270 and a test set of 4,572 indicated in the paper. We have measured the inference time results to the complete forward pass of the model, including the driver attention model. The batch size has been set to 1 to measure this parameter.

The strategy followed in this ablation study consists of using the modules explained above, understanding the contribution of each module to the final proposal and the benefits of using them together. The first model has a feature extractor with classification heads (Id. III.1). We have trained this configuration using only the last frame of the sequence instead of the video clip.

After that, the following approach adds an input sequence to the model (Id. III.2). This configuration computes the temporal relationships along the series and feeds it to the final decision and explanation. The last ten frames of the sequence have been used but skipping pairs, which means they correspond to the two last seconds of the video clip. The problem with this approach is that the high quantity of parameters makes the model overfit in the early steps of the training without being able to generalize. That's why the results are worst, but this information should help the final task because it is crucial to understand the temporal relationships to predict the vehicle decision.

The next model adds the driver attention module based on ARAGAN (Id. III.3). It contributes to the model with the comprehended visual scene. The driver's attention is employed, which is generated by ARAGAN, to achieve this visual understanding. This proposal again has a problem of over-adjustment due to the number of parameters.

We have employed the Spatio-Temporal Encoder to overcome this situation (Id. III.4). It achieves similar performance with 43 times fewer parameters. It contributes with a low-parameter model that understands all the relationships among the temporal and spatial information. Having a model with fewer parameters does not imply that it will infer faster because the architecture is the one that affects this the most. However, the complexity of the architecture proposal makes it slower than the other techniques, and it will require some optimization methods that we have not addressed in this work.

Our final proposal adds the Multi-Head Cross-Attention to the explanation classification head in order to model the relationships between the action and the explanation (Id. III.7). It is the final model of this work.

Moreover, evaluating the cross relationships between modules is relevant. For this reason, we add two more evaluations. The first one assesses the final model without the driver's attention module showing lower performance explaining the benefits of this module (Id. III.5). And the second one shows the final model but was trained with the last frame of the sequence to prove the advantage of understanding the temporal information (Id. III.6).

Finally, we have changed the loss function to the best model (Id. III.7), which uses the F1-score loss, to show its contribution compared with using the BCE loss (Id. III.8). The proposed loss outperforms the other one.

We conducted an experiment to evaluate the effectiveness of weighting loss in our proposal following the same procedure as in [9]. We adjusted the hyperparameter λ in the equation 1. The results are shown in the table IV, where the best scores are obtained when $\lambda = 1$. Nevertheless, training the model with $\lambda = \infty$, i.e. using only the explanation loss, gives a better overall score than when $\lambda = 0$, i.e. using only the action loss. This result is due to the use of cross-attention in the explanation branch.

To quantitatively evaluate the influence of the driver's attention on the proposal, which is one of the main contributions of this work, we conducted two experiments: Gini coefficient calculation and perturbation-based analysis, following the steps proposed in [39]. The Gini coefficient is an unsupervised metric to measure the sparsity of the attention. A high score indicates a high inequality in the distribution of attention. That means the higher the Gini score is, the more sparse the attention matrix is. We obtained a Gini score of 0.909 in the test dataset using ARAGAN. This high score indicates that most of the attention is paid to some few patches in the images. We also analysed the Gini coefficient in BDD-A [16] attention maps generated by humans to compare our result. We obtained an average value of 0.82, revealing that our attention model is sparser and tend to be more concentrated on road scene saliency patches. In conclusion, our model is more flexible and produces more explanatory attention maps than humans in BDD-A dataset.

As second experiment to quantitative assess our ARAGAN model, we performed a perturbation-based analysis [40] to determine the importance of the focused patches in the images. The significant regions of interest (attention higher than 0.995) on the RGB images were deliberately modified to 10 % of their original values, as done in [39]. The results are presented in Table V. The notable difference between the perturbed and normal models underscores the contribution of attention and its ability to emphasize saliency patches in urban road images. This enhancement nearly doubles the explanatory power, providing substantial evidence for the attention mechanism's capability to highlight crucial features.

C. Discussion

Our model outperforms the other proposals in explanations, obtaining the best $F1_{all}$ for this metric. Moreover, it gets the best $\overline{F1_{all}}$ of all the tested models in this experiment and the second-best performance of all the models in the literature. Although our proposal achieves the best explainability results, there are instances where the decision and explanation do not align. This is likely due to insufficient data to extrapolate the complexities of this task, despite Cross-Attention between the decision and explanation heads. In the near future, we will address these errors by imposing constraints on the model to enhance its learning.

The inference time comparison denotes that having more parameters does not imply higher inference time and that the model architecture is the most crucial parameter to evaluate this metric. Nevertheless, the proposed model achieves a reasonable frame rate to make a high-level decision, understanding the previous frames to take the best decision and explaining it based on how drivers look at the scene.

D. Qualitative results

In this section, we show some examples of the evaluated test dataset to show the strengths of our proposal. Table VI explains the qualitative results of Figure 3 to understand their predictions and the misunderstanding of the model. The attention maps are generated by ARAGAN and overlaid on the RGB image to show the attention map fed to the network, which helps the suggestion to perform better, as explained in the ablation study. In addition, Table VI explains the attention maps for each use case in the last column.

In the exposed samples, the model understands the situation obtaining the action to be taken and explains the scene with the explanation and the attention map. The explanation achieves optimal results, with slight differences from the ground truth.

E. Proof of concept using CARLA simulator for close-loop assessment

This experiment aims to assess our driver's attention-based explainable decision-making in a simulation environment for close-loop vehicle to driver interaction in an alternative domain. The CARLA simulator [38] was configured to execute a urban scenario. The driver is required to follow a predetermined route, which is overlaid in the simulator. During the driving, the model acts as a copilot for the driver.

1) Experiment Framework: The experiment is composed of: scenario, vehicle, driver and copilot (our model). The scenario is a urban route defined in the Town01 of CARLA with similar use cases than the found in BDD-OIA dataset. The vehicle provides a view of the road urban scene from the pilot seat and a frontal camera that sends the information to the model. Also includes a steering wheel and two pedals (throttle and brake). The driver is required to drive in a naturalistic way along the route, taking into account the copilot information. He has a view of the vehicle interior similar to a real vehicle using our three-screen simulator. Our model acts as a copilot that warns the pilot about possible decisions to be taken in their manoeuvres and provides explanations about these decisions. Information is overlaid in the frontal screen of the simulator.

The simulation runs at 15 Hz. The model's input is a ten images sequence corresponding to two seconds. This sequence is passed in batches to the ARAGAN generator, which obtains their attention maps in one forward pass. After

(a)

(b)

(c)

(d)

(e)

Fig. 3: Qualitative results for the driver attention explainable decision-making in BDD-OIA test set.

TABLE III: Ablation study of modules carried out to build the driver's attention-based explainable decision-making. Every model has been trained in the same data, and the results exposed are for the testing set of BDD-OIA.

Id.	Modules						$\mathcal{L}$	Par. (M)	Action						Explanation		Inf. (ms)
	MN V2	Seq	Att	ST Enc	CH	CA			F	S	L	R	mF1	$\mathbf{F1_{all}}$	mF1	$\mathbf{F1_{all}}$	
III.1	✓				✓		F1	18.5	0.61	0.77	0.46	0.48	0.58	0.65	0.27	0.48	**5.5**
III.2	✓	✓			✓		F1	165	0.56	0.79	0.37	0.40	0.54	0.68	0.28	0.53	34.7
III.3	✓	✓	✓		✓		F1	328	0.59	0.79	0.37	0.39	0.54	0.67	0.28	0.53	157
III.4	✓	✓	✓	✓	✓		F1	7.6	0.61	0.80	0.45	0.44	0.58	0.67	0.28	0.53	195
III.5	✓	✓		✓		✓	F1	3.8	0.60	0.79	0.43	0.43	0.56	0.68	0.25	0.52	163
III.6	✓		✓	✓		✓	F1	**3.0**	0.60	0.78	0.48	0.48	0.59	0.66	0.26	0.49	21
III.7	✓	✓	✓	✓		✓	F1	4.2	**0.63**	**0.80**	**0.48**	**0.46**	**0.59**	**0.69**	0.27	**0.54**	199
III.8	✓	✓	✓	✓		✓	BCE	4.2	0.52	0.78	0.34	0.29	0.49	0.63	0.14	0.38	199

Notes: Id. = identification; MNV2 = MobiliNet V2; Seq = Use of sequences; Att = Driver attention module; ST Enc = Spatio-Temporal Transformer Encoder; CH = Classification heads; CA = Multi-Head Cross-Attention explanation head; $\mathcal{L}$ = Loss
Par. = Parameters; F = Forward; S = Stop; L = Left; R = Right

TABLE IV: Performance in predicting actions and explanations based on task importance (determined by λ) and its impact on the loss function (1).

λ	Action		Explanation		$\overline{\mathbf{F1_{all}}}$
	mF1	$\mathbf{F1_{all}}$	mF1	$\mathbf{F1_{all}}$	
0	0.571	0.676	-	-	-
0.01	0.558	0.679	0.206	0.485	0.582
1	**0.593**	**0.695**	**0.267**	**0.538**	**0.617**
∞	0.430	0.442	0.265	0.535	0.489

TABLE V: Perturbation-based analysis to ascertain the significance of the attended region.

Models	Action		Explanation		$\overline{\mathbf{F1_{all}}}$
	mF1	$\mathbf{F1_{all}}$	mF1	$\mathbf{F1_{all}}$	
Proposal	0.634	0.695	0.267	0.538	0.617
Perturbed proposal	0.324	0.549	0.107	0.237	0.395

that, the RGB sequence and their attention maps are fed to the explainable decision-making model, to predict the action and the explanation of the sequence. This decision and explanation are sent to the driver to close the loop between the driver and the vehicle.

2) Experiment Results: This section explains the result of our copilot in the scenario designed in CARLA. For this experiment 5 drivers were requested to perform the test. The results are presented in two different ways: firstly, predicted actions and explanations provided by the copilot are compared with the decision and the reason made by the user for each designed use case. A posterior analysis of the route is required for each participant to obtain the explanation ground truth. After that, the predictions are compared with all the possible actions and explanations associated to each use case, not only the ones made by the driver, as the model has not got route information and cannot predict the decision according to that. For this purpose, a supervisor labels the possible action and explanation for the sequence and compares it with the model's output.

Table VII shows both comparisons. The first row shows the comparison against the drivers' decisions and explanations. The results show that the Forward and Stop predictions are perfect while the Left and Right decisions are not fulfilled, with a $F1 = 0.67$. This is because of the unbalanced dataset, which has more sequences of the first two classes. Regarding the global decision results, $mF1$ and $F1_{all}$ are higher than the ones achieved in the test dataset, because the use cases variability for this simulated environment is lower. Table VII second row shows the comparison with all the possible actions and explanations associated to each use case. Decision results show better performance than in the dataset, achieving $F1_{all} = 0.81$, but explanation performance decreased from the obtained in the dataset, achieving $F1_{all} = 0.46$. The cause may be the explanations are more difficult to predict due to the bias between domains and there are 21 possible answers. However, the scenario actions are easier with only 4 possible options in a simulated town with less complexity than the real world.

These results show the close-loop verification of our proposal and the domain adaptation of the model trained in a real dataset in a simulation environment. Fine-tuning in some recorded sequences in the simulator could improve the results. Moreover, the exposed results should be taken in caution, because the users' sample is low, and they are subjective, because the supervisors' bias could influence them. However, the decisions made in almost all use cases are the expected ones.

VII. Conclusions and Future Works

This work proposes a driver's attention-based explainable decision-making validated on BDD-OIA. It presents a novel architecture that understands the temporal data to explain and make a driving decision from a frontal sequence of images. We base the architecture on driver attention obtained with the features provided by our ARAGAN Generator and a Spatio-Temporal Transformer Encoder. The explanation branch has Multi-Head Cross-Attention to understand the relationship between both results.

We have carried out an ablation study to show the best approach using the proposed modules. In addition, we have compared the architecture with other state-of-the-art architectures obtaining on par results for action and the best for explanation. It outperforms the tested architectures, showing the best explanations for all the options in the literature.

The driver's attention helps the driving decision's visual context, resulting in a comprehensive vehicle-to-driver interac-

TABLE VI: Qualitative results explained for Figure 3. Correct predictions Incorrect predictions

Figure	Action		Explanation		Comments	Attention maps comments
	Pre	GT	Pre	GT		
3(a)	Forward Stop Right	Forward Right	Follow traffic Obstacle: car Obstacles on the left lane	Follow traffic Traffic light is green Obstacles on the left lane	The model does not detect the traffic light and predicts to stop due to an obstacle.	It focuses on the car ahead and on obstacles in the left lane.
3(b)	Stop	Stop	Obstacle: car The traffic light Solid line on the left	Obstacle: car Solid line on the left	The model mispredicts the traffic light due to night conditions.	It focuses on the vehicle's brake lights, which provide information about the braking event.
3(c)	Stop	Stop	The traffic light No lane on the left Solid line on the left Obstacles on the right lane	The traffic light No lane on the left Solid line on the left Obstacles on the right lane	Perfect prediction.	It focuses on the current lane because there is no relevant traffic on it.
3(d)	Forward	Forward	Follow traffic No lane on the left Obstacles on the right lane	Follow traffic No lane on the left Obstacles on the right lane	Perfect prediction.	It focuses on the vehicle in front, but not on the right one.
3(e)	Stop	Stop	Obstacle: car	Obstacle: car Obstacle: person/pedestrian	The model understands that the pedestrian on the right is not an obstacle and mispredicts it.	It focuses on the vehicle ahead and the possible new path of the ego-vehicle.

TABLE VII: Comparison between the predicted decisions and explanations in some labelled sequences.

Comparison	Action						Explanation		
	F	S	L	R	mF1	$\mathbf{F1_{all}}$	mF1	$\mathbf{F1_{all}}$	$\overline{\mathbf{F1_{all}}}$
Against the drivers' questionnaire	1	1	0.67	0.67	0.83	0.86	0.16	0.43	0.64
Against all possibilities	0.67	0.80	0.50	0.81	0.69	0.81	0.18	0.46	0.64

tion that can forecast decision-making. This capacity can assist a human driver to take decisions or facilitate the learning of autonomous driving.

We have conducted a proof of concept of the proposed algorithm in a simulation environment where the output model is provided to drivers to assist them during driving in a close-loop, acting as a copilot. The results were compared with the driver's decision and the explanation of that decision obtaining expected results.

In the near future, we intend to carry out a full investigation of the proposed model using an alternative dataset. This will require re-labelling. We also want to expand our simulation experiment with more users and scenarios.

REFERENCES

[1] 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 *2022 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1066–1072, IEEE, 2022.

[2] C. Gómez-Huelamo, L. M. Bergasa, R. Barea, E. López-Guillén, F. Arango, and P. Sánchez, "Simulating use cases for the uah autonomous electric car," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 2305–2311, IEEE, 2019.

[3] C. Gómez-Huélamo, A. Diaz-Diaz, J. Araluce, M. E. Ortiz, R. Gutiérrez, F. Arango, Á. Llamazares, and L. M. Bergasa, "How to build and validate a safe and reliable autonomous driving stack? a ros based software modular architecture baseline," in *2022 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1282–1289, IEEE, 2022.

[4] I. Gog, S. Kalra, P. Schafhalter, M. A. Wright, J. E. Gonzalez, and I. Stoica, "Pylot: A modular platform for exploring latency-accuracy tradeoffs in autonomous vehicles," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8806–8813, IEEE, 2021.

[5] A. Bewley, J. Rigley, Y. Liu, J. Hawke, R. Shen, V.-D. Lam, and A. Kendall, "Learning to drive from simulation without real world labels," in *2019 International conference on robotics and automation (ICRA)*, pp. 4818–4824, IEEE, 2019.

[6] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9329–9338, 2019.

[7] O. Makansi, J. Von Kügelgen, F. Locatello, P. Gehler, D. Janzing, T. Brox, and B. Schölkopf, "You mostly walk alone: Analyzing feature attribution in trajectory prediction," *arXiv preprint arXiv:2110.05304*, 2021.

[8] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proceedings of the IEEE international conference on computer vision*, pp. 2942–2950, 2017.

[9] Y. Xu, X. Yang, L. Gong, H.-C. Lin, T.-Y. Wu, Y. Li, and N. Vasconcelos, "Explainable object-induced action decision for autonomous vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9523–9532, 2020.

[10] J. Dong, S. Chen, S. Zong, T. Chen, and S. Labi, "Image transformer for explainable autonomous driving system," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 2732–2737, IEEE, 2021.

[11] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 563–578, 2018.

[12] Y. Yao, X. Wang, M. Xu, Z. Pu, E. Atkins, and D. Crandall, "When, where, and what? A new dataset for anomaly detection in driving videos," *arXiv preprint arXiv:2004.03044*, 2020.

[13] T. You and B. Han, "Traffic accident benchmark for causality recognition," in *European Conference on Computer Vision*, pp. 540–556, Springer, 2020.

[14] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7699–7707, 2018.

[15] Y. Shen, S. Jiang, Y. Chen, E. Yang, X. Jin, Y. Fan, and K. D. Campbell, "To explain or not to explain: A study on the necessity of explanations for autonomous vehicles," *arXiv preprint arXiv:2006.11684*, 2020.

[16] 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, Cham, 2019.

[17] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[18] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[20] D. Wang, C. Devin, Q.-Z. Cai, F. Yu, and T. Darrell, "Deep object-centric

policies for autonomous driving," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8853–8859, 2019.

[21] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[22] J. Gao, C. Sun, H. Zhao, Y. Shen, D. Anguelov, C. Li, and C. Schmid, "Vectornet: Encoding hd maps and agent dynamics from vectorized representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11525–11533, 2020.

[23] L. L. Li, B. Yang, M. Liang, W. Zeng, M. Ren, S. Segal, and R. Urtasun, "End-to-end contextual perception and prediction with interaction transformer," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5784–5791, IEEE, 2020.

[24] L. Wang, Y. Hu, L. Sun, W. Zhan, M. Tomizuka, and C. Liu, "Hierarchical adaptable and transferable networks (hatn) for driving behavior prediction," *arXiv preprint arXiv:2111.00788*, 2021.

[25] C. Gou, Y. Zhou, and D. Li, "Driver attention prediction based on convolution and transformers," *The Journal of Supercomputing*, vol. 78, no. 6, pp. 8268–8284, 2022.

[26] J. Kim, T. Misu, Y.-T. Chen, A. Tawari, and J. Canny, "Grounding human-to-vehicle advice for self-driving vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10591–10599, 2019.

[27] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "Trackformer: Multi-object tracking with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8844–8854, 2022.

[28] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo, "Transtrack: Multiple object tracking with transformer," *arXiv preprint arXiv:2012.15460*, 2020.

[29] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[30] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7077–7087, 2021.

[31] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3202–3211, 2022.

[32] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International conference on machine learning*, pp. 4055–4064, PMLR, 2018.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[35] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

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

[37] T. Jing, H. Xia, R. Tian, H. Ding, X. Luo, J. Domeyer, R. Sherony, and Z. Ding, "Inaction: Interpretable action decision making for autonomous driving," in *European Conference on Computer Vision*, pp. 370–387, Springer, 2022.

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[39] W. Huang, Y. Zhou, X. He, and C. Lv, "Goal-guided transformer-enabled reinforcement learning for efficient autonomous navigation," *arXiv preprint arXiv:2301.00362*, 2023.

[40] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," *International Journal of Computer Vision*, vol. 130, no. 10, pp. 2425–2452, 2022.

Javier Araluce received his BSc degree in 2018 in Industrial Electronics and Automation Engineering and his MSc in 2020 in Industrial Engineering, focused on Robotics and Perception, from the University of Alcalá. Currently he is a PhD candidate in Robotics and Artificial Intelligence in the RobeSafe research group (Department of Electronics, University of Alcalá), his PhD thesis is focused on "Driver Attention based on Deep Learning for a Smart Driver to Vehicle (V2D) Interaction".

Luis M. Bergasa received the PhD degree in Electrical Engineering in 1999 from the University of Alcalá (UAH), Madrid, Spain. He is Full Professor since 2011 and Director of Digital Transformation since 2022 in this university. From 2000 he had different research and teaching positions at the UAH. He was Head of the Department of Electronics (2004-2010), coordinator of the Doctorate program in Electronics (2005-2010), Director of Knowledge Transfer at the UAH (2014-2018), and Director of the Committee for the Strategic Plan of the UAH (2019-2022). He is author of more than 280 refereed papers in journals and international conferences. His research activity has been awarded/recognized with 28 prizes/recognitions related to Robotics and Automotive fields from 2004 to nowadays. He ranks 65th among Spanish researchers in Computer Science (2022). He was recognized as one of the most productive authors in Intelligent Transportation Systems (ITS) (1996-2014). He was a Distinguished Lecturer of the IEEE Vehicular Technology Society (2019 - 2021). He received the Institutional Lead Award 2019 from the IEEE ITS Society for the longstanding work of his research group. He is Associate Editor of the IEEE Transactions on ITS from 2014 and he has served on Program/Organizing Committees in more than 20 conferences. His research interests include driver behaviors and scene understanding using Computer Vision and Deep Learning for Autonomous Driving.

Manuel Ocaña is a Associate Professor at the Department of Electronics of the University of Alcalá. He received the Tech. Eng. degree in Electrical Engineering in 2000 from the Technical University of Madrid (UPM), Masters' degree in Electrical Engineering in 2002 (with Best Student Award) and the Ph.D. degree in Electrical Engineering (with PhD UAH Award) in 2005 from the University of Alcalá (UAH), Alcalá de Henares, Madrid, Spain. His research interests include localization and navigation in Robotics, eSafety and Intelligent Transportation Systems. He is the author of more than 110 refereed papers in journals and international conferences, and corresponding author of 3 patents.

Ángel Llamazares received the technical degree in industrial electronics, the M.Sc. in computer science engineering and the Ph.D. degree in electrical engineering from the University of Alcalá (UAH), Alcalá de Henares, Spain, in 2007, 2009 and 2017 respectively. He is currently an Assistant Professor in the Electronics Department at University of Alcalá (UAH). His research interests include robotics and intelligent transportation systems. He is the author of more than 34 refereed publications in international journals, book chapters, and conference proceedings and is the co-author of a patent. Dr. Llamazares has been awarded the Accesit for the best Ph.D. thesis by the Spanish Chapter of the IEEE ITS Society and the prize for the best academic performance in his M.Sc.

Elena López-Guillén received the M.S. and Ph.D. degrees in electronics engineering from the University of Alcalá, Alcalá de Henares, Spain, in 1999 and 2004, respectively. From 1996 to 2009 she was an Assistant Professor in the Department of Electronics in the same university, and in 2009 she became an Associate Professor. She is a member of the RobeSafe research group. Her research interests include robotics and control, autonomous driving techniques, multisensorial indoor localization (SLAM techniques), scene understanding, computer vision, probabilistic algorithms, human-behavior analysis and bioengineering. She has authored more than 100 publications in international journals, book chapters and conference proceedings.