Predicting the Driver’s Focus of Attention: the DR(eye)VE Project

Andrea Palazzi*, Davide Abati*, Simone Calderara*, Francesco Solera, and Rita Cucchiara

Abstract—In this work we aim to predict the driver’s focus of attention. The goal is to estimate what a person would pay attention to while driving, and which part of the scene around the vehicle is more critical for the task. To this end we propose a new computer vision model based on a multi-path deep architecture that integrates three sources of information: visual cues, motion and scene semantics. We also introduce DR(eye)VE, the largest dataset of driving scenes, enriched with eye-tracking annotations and other sensors’ measurements. This dataset features more than 500,000 registered frames, matching ego-centric views (from glasses worn by drivers) and car-centric views (from roof-mounted camera). Results highlight that several attentional patterns are shared across drivers and can be reproduced to some extent. This may benefit an ADAS system by providing indication of which elements in the scene are likely to capture the driver’s attention.

Index Terms—video saliency, driver attention prediction

1 INTRODUCTION

According to the J3016 SAE international Standard, which defined the five levels of autonomous driving [25], cars will provide a fully autonomous journey only at the fifth level. At lower levels of autonomy, computer vision and other sensing systems will still support humans in the driving task. Human-centric Advanced Driver Assistance Systems (ADAS) have significantly improved safety and comfort in driving (e.g. collision avoidance systems, blind spot control, lane change assistance etc.). Among ADAS solutions, the most ambitious examples are related to monitoring systems [20], [28], [32], [40]: they parse the attentional behavior of the driver together with the road scene to predict potentially unsafe maneuvers and act on the car in order to avoid them – either by signaling the driver or braking. However, all these approaches suffer from the complexity of capturing the true driver’s attention and rely on a limited set of fixed safety-inspired rules. Oppositely, we advocate for a new ADAS paradigm which suggests to the driver or to the car, with no hard intervention, where most drivers would focus. The problem is thus shifted from a personal level (what the driver is looking at) to a task-driven level (what most drivers would be looking at).

In this work we conform to this new paradigm and propose a computer vision model able to suggest to the driver what others would focus their attention to in the same situation. We achieve this result in two stages: First, we conduct a data-driven study on drivers’ gaze fixations under different circumstances and scenarios. The study concludes that the semantic of the scene, the speed and bottom-up features all influence the driver’s gaze. Second, we advocate for the existence of common gaze patterns that are shared among different drivers. We empirically demonstrate the existence of such patterns by developing a deep learning model that can profitably learn to predict where a driver would be looking at in a specific situation.

To support our claims, we collected DR(eye)VE, a new dataset of on-the-road driving videos. We recorded and annotated more than 500,000 frames (resulting in more than 6 hours) of driving sequences in different traffic and weather conditions. For every frame, we also acquired the driver’s gaze through an accurate eye tracking device, and registered such data to the external scene recorded from an HD roof-mounted camera. The DR(eye)VE data richness enables us to train an end-to-end deep network that predicts salient regions in car-centric driving videos. The network we propose is based on three branches which estimate saliency maps from a) visual information of the scene, b) motion cues (in terms of optical flow) and c) semantic segmentation (Fig. 1). In contrast to the majority of experiments, which are conducted in controlled laboratory settings or employ sequences of unrelated images [7], [29], [58], we train our model on data acquired on the field. Final results demonstrate the ability of the network to generalize across different day times, different weather conditions, different landscapes and different drivers. Eventually, we believe our work can be complementary to the

* The authors are with the Department of Engineering “Enzo Ferrari”, University of Modena and Reggio Emilia, Italy. E-mail: name.surname@unimore.it
The paper is organized as follows. In Sec. 2, related works about computer vision and visual saliency are provided to frame our work in the current state-of-the-art scenario. Sec. 3 describes the $\text{DR(eye)}\text{VE}$ dataset and some insights about several attentional patterns that human drivers exhibit. Sec. 4 illustrates the proposed deep network to replicate such human behavior, and Sec. 5 reports the performed experiments.

2 RELATED WORK

Predicting human attention is strongly related to the concept of visual saliency, which is defined as a dense prediction task determining the probability of each point in a given scene to attract the human attention [15], [16].

Image and video saliency. Image saliency has been addressed exploiting both top-down and bottom-up strategies. In the former approach, visual saliency aims at highlighting objects and cues that could be meaningful in the context of a given task. For this reason, such methods are also known as task-driven. Usually, top-down computer vision models are built to integrate semantic contextual information in the saliency extraction process [56]. This can be achieved by either merging saliency maps at different levels of scale and abstraction [23], or including a-priori cues about relevant objects for the task at hand [14], [21], [63]. Conversely, bottom-up models refer to data-driven saliency and capture salient objects or events naturally popping out in the image, independently of the observer, the undergoing task and other external factors. In this context, computational models focus on spotting visual discontinuities, either by clustering features or considering the rarity of image regions, locally [36], [50] or globally [1], [10], [65]. The recent success of deep networks has smoothed the aforementioned distinction, as models have become more and more powerful by seamlessly combining the benefits of both top-down and bottom-up approaches, achieving state-of-the-art results on public benchmarks [12], [13], [27], [33], [34].

In video, saliency estimation is more complex with respect to still images since motion heavily affects human attention. Some models merge bottom-up saliency with motion maps, either by means of optical flow [67] or feature tracking [66]. Other methods enforce temporal dependencies between bottom-up features in successive frames. Both supervised [52], [67] and unsupervised [39], [61], [62] feature extraction can be employed, and temporal coherence can be achieved either by conditioning the current prediction on information from previous frames [47] or by capturing motion smoothness with optical flow [52], [67]. While deep video saliency models still lack, an interesting work is [3], which relies on a recurrent architecture fed with clip encodings to predict the fixation map by means of a Gaussian Mixture Model (GMM). Nevertheless, most saliency methods limit to bottom-up features accounting for just visual discontinuities in terms of textures or contours. Our proposal, instead, is specifically tailored to the driving task and fuses the bottom-up information with semantics and motion elements that have emerged as attentional factors from the analysis of the $\text{DR(eye)}\text{VE}$ dataset.

Attention and driving. Prior works addressed the task of detecting saliency and attention in the specific context of assisted driving. In such cases, however, gaze and saliency attentive mechanisms have been mainly studied for some driving sub-tasks only, often acquiring gaze maps from on-screen images. Bremond et al. [51] presented a model that exploits visual saliency with a non-linear SVM classifier for the detection of traffic signs. The validation of this study was performed in a laboratory non-realistic setting, emulating an in-car driving session. A more realistic experiment [6] was then conducted with a larger set of targets, e.g. including pedestrians and bicycles.

Driver’s gaze has also been studied in a pre-attention context, by means of intention prediction relying only on fixation maps [45]. The study in [58] inspects the driver’s attention at T junctions, in particular towards pedestrians and motorbikes, and exploits object saliency to avoid the looked-but-failed-to-see effect. In absence of eye tracking systems and reliable gaze data, [4], [19], [55], [59] focus on drivers’ head, detecting facial landmarks to predict head orientation. Such mechanisms are more robust to varying lighting conditions and occlusions, but there is no certainty about the adherence of predictions to the true gaze during the driving task.

Datasets. Many image saliency datasets have been released in the past few years, improving the understanding of the human visual attention and pushing computational models forward. Most of these datasets include no motion information, as saliency ground truth maps are built by aggregating fixations of several users within the same still image. Usually, a gaussian filtering post-processing step is employed on recorded data, in order to smooth such fixations and integrate their spatial locations. Some datasets, such as the MIT saliency benchmark [7], were labeled through an eye tracking system, while others, like the SALICON dataset [29] relied on users clicking on salient image locations. We refer the reader to [5] for a comprehensive list of available datasets. On the contrary, datasets addressing saliency in video still lack. Up to now, Action in the Eye [38] represents the most important contribution, since it consists in the largest video dataset accompanied by gaze and fixation annotations. That information, however, is collected in the context of action recognition, so it is heavily task-driven. A few datasets address directly the study of attentional mechanisms while driving, but are mostly restricted to limited settings and are not publicly available. In some of them [51], [58] fixation and saliency maps are acquired during an in-lab simulated driving experience. In-lab experiments enable several attentional drifts that are influenced by external factors (e.g. monitor distance and others) rather than the primary task of driving [54]. A few in-car datasets exist [6], [45], but were precisely tailored to force the driver to fulfill some tasks, such as looking at people or traffic signs. Coarse gaze information is also available in [18], while the external road scene images are not acquired. We believe that the dataset presented in [45] is, among the others, the closer to our proposal. Yet, video sequences are collected from one driver only and limited to countryside scenarios. Instead, our $\text{DR(eye)}\text{VE}$ dataset includes sequences from several different drivers and presents a high
3 THE DR(eye)VE PROJECT

In this section we present the DR(eye)VE dataset, the protocol adopted for video registration and annotation, the automatic processing of eye-tracker data and the analysis of the driver’s behavior in different conditions.

The dataset. The DR(eye)VE dataset consists of 555,000 frames divided in 74 sequences, each of which is 5 minutes long. Eight different drivers of varying age from 20 to 40, both male and female, took part to the driving experiment, that lasted more than two months. Videos were recorded in different contexts, both in terms of landscape (downtown, countryside, highway) and traffic condition, ranging from traffic-free to highly cluttered scenarios. They were recorded in diverse weather conditions (sunny, rainy, cloudy) and at different hours of the day (both daytime and night). Tab. 1 recap the dataset features and Tab. 2 compares it with other related proposals. DR(eye)VE is now the largest publicly available dataset including gaze and driving behavior in automotive.

The Acquisition System. The driver’s gaze information was captured using the commercial SMI ETG 2w Eye Tracking Glasses (ETG). ETG capture attentional dynamics also in presence of head pose changes, which occur very often during the task of driving. While a frontal camera acquires the scene at 720p/30fps, users pupils are tracked at 60Hz. Gaze information are provided in terms of eye fixations and saccade movements. ETG was manually calibrated before each sequence for every driver.

Simultaneously, videos from the car perspective were acquired using the GARMIN VirbX camera mounted on the car roof. Such sensor captures frames at 1080p/25fps, and includes further information such as GPS data, accelerometer and gyroscope measurements.

Video-gaze registration. The dataset has been processed to move the acquired gaze from the egocentric (ETG) view to the car (GARMIN) view. The latter features a much wider field of view (FoV), and can contain fixations that are out of the egocentric view. For instance, this can occur whenever the driver takes a peek at something at the border of this FoV, but doesn’t move his head. For every sequence, the two videos were manually aligned to cope with the difference in sensors framerate. Videos were then registered frame-by-frame through a homographic transformation that projects fixation points across views. More formally, at each timestep \( t \) the GARMIN frame \( I^G_t \) and the ETG frame \( I^E_t \) are registered by means of a homography matrix \( H_{E\rightarrow G} \), computed by matching SIFT descriptors [35] from one view to the other (see Fig. 3). A further RANSAC [17] procedure ensures robustness to outliers.

Fixation map computation. The pipeline discussed above provides a frame-level annotation of the driver’s fixations. In contrast to image saliency experiments [7], there is no clear and indisputable protocol for obtaining saliency maps from raw fixations when acquired in task-driven real-life scenarios. The main motivation resides in the fact that observer’s subjectivity cannot be removed by averaging different observers’ fixations.

Fig. 2. Examples taken from a random sequence of DR(eye)VE. From left to right: frames from the eye tracking glasses with gaze data, from the roof-mounted camera, temporal aggregated fixation maps (as defined in Sec. 3) and overlays between frames and fixation maps.
TABLE 1
Summary of the DR(eye)VE dataset features.

| # Videos | # Frames | Drivers | Weather conditions | Lighting | Gaze Info | Metadata | Camera POVs |
|----------|----------|---------|--------------------|----------|-----------|----------|-------------|
| 74       | 555,000  | 8       | sunny              | day      | raw fixations | GPS       | driver (720p) |
|          |          |         | cloudy             | evening  | gaze map   | car speed | car (1080p)  |
|          |          |         | rainy              | night    | pupil dilation | car course |             |

TABLE 2
A comparison between DR(eye)VE and other datasets.

| Dataset              | Frames      | Drivers | Scenarios            | Annotations | Real-world | Public |
|----------------------|-------------|---------|----------------------|-------------|------------|--------|
| Pugeault et al. [45] | 158,668     | n.d.    | Countryside, Highway | 9 classes   | Yes        | No     |
|                      |             |         | Downtown             | Environment |            |        |
| Simon et al. [51]    | 40          | 30      | Downtown             | Gaze Maps   | No         | No     |
| Underwood et al. [58]| 120         | 77      | Urban Motorway       | n.d.        | No         | No     |
| Fridman et al. [18]  | 1,860,761   | 50      | Highway              | 6 Gaze Location Classes | Yes | No |
| DR(eye)VE [2]        | 555,000     | 8       | Countryside, Highway | Gaze Maps   | Yes        | Yes    |
|                      |             |         | Downtown             |             |            |        |

This is obviously the case as the scene changes continuously and two drivers cannot be at the same time in the same point of the street. The only chance to average among different observers would be the adoption of a simulation environment, but it has been proved that the cognitive load in controlled experiments is lower than in real test scenarios and it effects the true attentional mechanism of the observer [48]. In our preliminary DR(eye)VE release [2], fixation points were aggregated and smoothed by means of a temporal sliding window. In such a way, temporal filtering discarded momentary glimpses that contain precious information about the driver’s attention. Following the psychological protocol in [37] and [24], this limitation was overcome in the current release where the new fixation maps were computed without temporal smoothing. Both [37] and [24] highlight the high degree of subjectivity of scene scanpaths in short temporal windows (< 1 sec) and suggest to neglect the fixations pop-out order within such windows. This mechanism accounts also for the inhibition of return phenomenon that may prevent interesting objects to be observed twice in short temporal intervals [26], [44].

More formally, the fixation map $F_t$ for a frame at time $t$ is built by accumulating projected gaze points in a temporal sliding window of $k = 25$ frames, centered in $t$. For each time step $t + i$ in the window, where $i \in \{-k, -k + 1, \ldots, k - 1, k\}$, gaze points projections on $F_t$ are estimated through the homography transformation $H_{t+i}$ that projects points from the image plane at frame $t + i$, namely $p_{t+i}$, to the image plane in $F_t$. A continuous fixation maps is obtained from the projected fixations by centering on each of them a gaussian having spatial variance $\sigma_s^2 = 200$ pixels an taking the max value along the time axis:

$$F_t(x, y) = \max_{i \in \{-k, \ldots, k\}} \mathcal{N}((x, y) \mid H_{t+i} \cdot p_{t+i}, \sigma_s^2)$$  (1)

The gaussian variance has been computed by averaging the ETG spatial acquisition errors on 20 observers looking at calibration patterns at different distances from 5 to 15 meters. Visual examples of fixation maps can be appreciated in Fig. 2-4.

**Labeling attentional drifts.** Fixation maps exhibit a very strong central bias. This is common in saliency annotations [53] and even more in the context of driving. For these reasons, there is a strong unbalance between lots of easy-to-predict scenarios and unfrequent but interesting hard-to-predict events.

To enable the evaluation of computational models under such circumstances, the DR(eye)VE dataset has been extended with a set of further annotations. For each video, subsequences whose ground truth poorly correlates with the average ground truth of
that sequence are selected. We employ Pearson’s Correlation Coefficient (CC) and select subsequences with \( CC < 0.3 \). This happens when the attention of the driver focuses far from the vanishing point of the road. Examples of such subsequences are depicted in Fig. 5. Several human annotators inspected the selected frames and manually split them - by majority - into (a) acting, (b) inattentive, (c) errors and (d) uninteresting events:

- errors can happen either due to failures in the measuring tool (e.g. in extreme lighting conditions) or in the successive data processing phase (e.g. SIFT matching);
- inattentive subsequences occur when the driver focuses his gaze on objects unrelated to the driving task (e.g. looking at an advertisement);
- uninteresting subsequences describe situations where external coders did not reach an agreement about the coherence between the fixation points and the actual driving behavior.
- acting subsequences include all the remaining ones.

Acting subsequences are particularly interesting as the deviation of driver’s attention from the common central pattern denotes an intention linked to task-specific actions (e.g. turning, changing lanes, overtaking ...).

Moreover, the gaze location tends to drift from this central attractor when the context changes in terms of car speed and landscape. Indeed, [46] suggests that our brain is able to compensate spatially or temporally dense information by reducing the visual field size, so that every perceived stimulus can be elaborated properly. We also observe this phenomenon in our experiments, as shown in Fig. 6.

DR(e)VE data also highlight that the driver’s gaze is attracted towards specific semantic categories. To reach the above conclusion, the dataset is analysed by means of the semantic segmentation model in [64] and the distribution of semantic classes within the fixation map evaluated. More precisely, several binary versions of the fixation maps (that hold continuous values) are constructed by thresholding their values for each pixel. By linearly moving the threshold towards 1, the area of interest shrinks around the real fixation point. Fig. 7 highlights how different categories respond when the threshold changes. Object classes that are the real focus of the gaze exhibit a positive trend, since the ratio of pixels the fixation map evaluated. More precisely, several binary versions of the fixation maps (that hold continuous values) are constructed by thresholding their values for each pixel. By linearly moving the threshold towards 1, the area of interest shrinks around the real fixation point. Fig. 7 highlights how different categories respond when the threshold changes. Object classes that are the real focus of the gaze exhibit a positive trend, since the ratio of pixels

### 3.1 Dataset analysis

The very first insight is the presence of a strong attraction of driver’s focus towards the vanishing point of the road, that can be appreciated in Fig. 8. It seems indeed that drivers often tend to disregard road signals, cars coming from the opposite direction and pedestrians on sidewalks. This is an effect of human peripheral vision [49], that allows observers to still perceive and interpret stimuli out of - but sufficiently close to - their FoA. A driver can therefore achieve a larger area of attention by focusing on the road’s vanishing point: due to the geometry of the road environment, many of the objects worth of attention are coming from there and have already been perceived when distant.

![Image](https://via.placeholder.com/150)

**Fig. 8.** Mean frame (a) and fixation map (b) averaged across the whole sequence 02, highlighting the link between driver’s focus and the vanishing point of the road.
The DR(eye)VE dataset is a strong starting point for constructing a deep architecture to predict common attentional patterns. Here, we develop a neural network to predict task-driven saliency of the scene, mimicking humans FoA while driving. In the context of high level video analysis (e.g. action recognition and video classification) recent studies demonstrate that a baseline processing single frame inputs can be outclassed by feeding a deep network with an additional input time dimension [30], [57]. Temporal dependencies are usually modeled either by 3D convolutional layers [57], tailored to capture short range correlations, or by recurrent architectures (e.g. LSTM, GRU), that can model long term dependencies. Our model follows the former approach, relying on the assumption that a small time window (e.g. half a second) holds sufficient contextual information for predicting where the driver would focus in that moment. Therefore, we propose a deep network that we feed with samples containing 16 consecutive temporal frames from a video (called clips from now on) and predicts the fixation map for the last frame of the clip.

4.1 Multi-Path Model

The DR(eye)VE dataset analysis (Sec. 3.1) can be summarized in the following observations:

- the drivers’ FoA exhibits recurrent patterns suggesting they can be reproduced by a computational model;
- the drivers’ gaze is affected by a strong prior on objects semantics (e.g. drivers tend to focus on "on the road" items);
- motion (i.e. vehicle speed) is another key factor that influences gaze.

Accordingly, the multi-path model (illustrated in Fig. 9) is composed of three independent saliency branches, presenting identical architectures but disjoint sets of parameters and working on different input domains. The first branch is fed with RGB images and processes raw visual information about the scene. The second path computes saliency from motion, and is fed with optical flow estimates of the clip encoded in RGB (according to the flow field defined in [22]). The last path provides saliency from semantics, and processes the segmentation prediction on the scene given by the model by Yu and Koltun [64], which is trained on the cityscapes dataset [11] and provides 19-channels tensors where each map holds pixel probabilities for a certain class (the class convention is the one adopted in the original benchmark). In the final layer of the network, the three independent maps are summed and normalized to achieve the final probability map.

Single saliency branch. Each branch of the multi-path model is a multiple-input multiple-output architecture composed of two streams, as shown in Fig. 11. This network architecture aims at addressing the strong central bias that occurs in driving gaze data by the adoption of different training policies on each stream. Both streams rely on the same convolutional encoder (i.e. with shared weights) that provides a rough saliency estimate of a given clip. This component is called COARSE module.

The COARSE module represents the backbone of each branch, and provides rough estimates of attentional maps of input clips. It is based on the C3D architecture [57] and maps pixels from a given videoclip into a 512-dimensional feature space. 3D convolutions allow to encode motion by applying a convolutional kernel on the 4D input tensor. As opposed to 2D convolutions that stride along the width and height dimension of the input tensor, a 3D convolution also strides along time. Formally, the \( j \)-th feature...
map in the \(i\)-th layer at position \((x, y)\) at time \(t\) is computed as:

\[
v_{i,j}^{x,y,t} = b_{i,j} + \sum_{m} \sum_{p=0}^{P_m-1} \sum_{q=0}^{Q_m-1} \sum_{r=0}^{R_m-1} w_{i,j,mpq} x+p, y+q, t+r
\]

(2)

where \(m\) indexes different input feature maps, \(w_{i,j,mpq}\) is the value at the position \((p, q)\) at time \(r\) of the kernel connected to the \(m\)-th feature map, and \(P_i\), \(Q_i\) and \(R_i\) are the dimensions of the kernel along width, height and temporal axis respectively; \(b_{i,j}\) is the bias from layer \(i\) to layer \(j\).

From C3D, only the most general-purpose features are retained by removing the top convolutional and fully connected layers which are strongly linked to the original action recognition task. Furthermore, a modified version of the last pooling layer squeezes out the temporal dimension by covering the whole time axis. The clip encoding is now independent of time and is brought back to the input spatial resolution with a bilinear upsampling. Eventually, a \(3 \times 3\) 2D convolution achieves the single channel saliency map by weighting feature channels. The \textit{COARSE} module is depicted in Fig. 10.

During training, the first stream feeds the \textit{COARSE} network with random crops, forcing the model to learn saliency given visual cues rather than prior spatial location. The C3D training process described in [57], employs a \(128 \times 128\) image resize, and then a \(112 \times 112\) random crop. However, the small difference in the two resolutions may limit the variance of random crops and thus the variety in the location of fixation maps, leading to models strongly attracted towards the image center. For this reason, training images are resized to \(256 \times 256\) before the \(112 \times 112\) crops. This crop policy is way more radical, and generates cropped samples that cover less than a quarter of the original image. This strategy allows to reach a sufficient variety in prediction targets, at the cost of a coarser prediction. Indeed, as crops get smaller, the relative area of ground truth maps gets bigger, leading the model to learn better localized yet larger maps.

Conversely, the second stream feeds the same \textit{COARSE} model with images resized to match the crop size. The coarse prediction that is performed on this stream is then upsampled to the original dimension, and concatenated on the 3-channel final clip frame (i.e. the frame with respect to which the prediction is made), that is provided as input as well. Finally, this concatenated tensor is processed through a final block of convolutions, that serves as a refinement network. Indeed, this last module merges coarse saliency maps with high resolution appearance information from the color frame. The output of the refinement block is the actual fixation map prediction (see Fig. 11).

Both the crop and the full frame output are modeled as a probability distribution \(P\), and the cost against ground truth distribution \(Q\) is minimized in terms of Kullback-Leibler divergence:

\[
D_{KL}(P, Q) = \sum_{i} Q_i \log \left( \frac{\epsilon + Q_i}{\epsilon + P_i} \right)
\]

(3)

where the summation index \(i\) spans across image pixels and \(\epsilon\) is a regularization constant.

The proposed cropping policy is particularly beneficial when dealing with spatially biased fixation maps. Indeed, when trained with crops the network captures the bottom-up saliency neglecting its spatial location, thus balancing the central biasing effect. Conversely, when trained with resized clips, the network learns also the prior on fixation map locations. By keeping a balance between these two strategies the model can learn to benefit from both those effects.

5 Experiments

In this section we evaluate the performance of the proposed multi-path model. First, we start by comparing our model against some baselines and other methods in literature. Following the guidelines in [8], for the evaluation phase we rely on Pearson’s
Correlation Coefficient (CC) and Kullback–Leibler Divergence ($D_{KL}$) measures. Moreover, we evaluate the Information Gain ($IG$) measure to assess the quality of a predicted map $P$ with respect to a ground truth map $Q$ in presence of a strong bias, as:

$$IG(P, Q, B) = \frac{1}{N} \sum_i Q_i \left[ \log_2(\epsilon + P_i) - \log_2(\epsilon + B_i) \right]$$

being $i$ an index spanning all the $N$ pixels in the image, $B$ the bias computed as the average training fixation map and $\epsilon$ a regularization constant.

Moreover, we conduct an ablation study to investigate how different branches affect one another and how this mutual influence changes in different scenarios. Eventually, we assess our model qualitatively for an ADAS system.

**Training details.** The three different pathways of the multi-path model (namely saliency from color, from motion and from semantics) have been pre-trained independently using the same cropping policy of Sec. 4.1. Each branch has been respectively fed with:

- 16 frames clips in raw RGB color space;
- 16 frames clips with optical flow maps, encoded as color images through the flow field encoding [22];
- 16 frames clips holding semantic segmentation from [64] encoded as 19 scalar activation maps, one per segmentation class.

During individual branch pre-training clips were randomly mirrored for data augmentation. We employ Adam optimizer with parameters as suggested in the original paper [31], with the exception of the learning rate that we set to $10^{-4}$. Eventually, batch size was fixed to 32 and each branch was trained until convergence without early stopping policies.

Moreover, the complete multi-path architecture was finetuned using the same cropping and data augmentation strategies but by lowering the learning rate value to $10^{-5}$ and batch size to 4 due to GPU memory limitations.

### 5.1 Comparison with state-of-the-art

In Tab. 3 we report results of our proposal against other state-of-the-art saliency models [3], [12], [43], [52], [61] and [62] evaluated both on the complete test set and on acting subsequences only. All the competitors, with the exception of [43] are bottom-up approaches and mainly rely on appearance and motion discontinuities. To test the effectiveness of deep architectures for saliency prediction we compare against the Multi-Level Network (MLNet) [12], which scored favourably in the MIT300 saliency benchmark [7], and the Recurrent Mixture Density Network (RMDN) [3], which represents the only deep model addressing video saliency. While MLNet works on images discarding the temporal information, RMDN encodes short sequences in a similar way to our COARSE module, and then relies on a LSTM architecture to model long term dependencies and estimate the fixation...
map in terms of a GMM. Both models were re-trained on the DR(eye)VE dataset.

Results highlight the superiority of our multi-path architecture on all test sequences. The gap in performance with respect to bottom-up unsupervised approaches [61], [62] is higher, and is motivated by the strong relation between saliency maps and the driving context, that calls for a task-oriented training procedure. Moreover, MLNet low performance testifies for the need of accounting the temporal correlation between consecutive frames that distinguishes the image and video saliency tasks. Indeed, RMDN processes video inputs and outperforms MLNet, yet its performances are still limited. We believe that the recurrent module capturing long term dependencies employed by RMDN is not beneficial, even harmful, for this task. Eventually, our preliminary task-driven model published in [43] scored positively but lower than the final proposal. The gap in performance resides in the awareness of our three path architecture of the aspects that characterize the driving task as emerged from the analysis in Sec. 3.1. The positive performance of our model are confirmed also when evaluated on the acting partition of the dataset. We recall that acting stands for sequences where there is a significant shift of the saliency from the center of the image that resembles a stable change in the gaze direction (Fig. 5). Being able to predict the saliency also on acting sequences means that the model captures the strong centered attention bias but is capable of generalizing when required by the context. Some successful prediction cases are illustrated in Fig. 12.

This is further shown by the comparison against a centered gaussian baseline (BG) and against the average of all training set fixation maps (BM). The former baseline has proven effective on many image saliency detection tasks [7] while the latter represents a more task-driven version of the center bias. Even in this case, the capability of the multi-path model to outperform the centered baselines testifies in favor of the task complexity where the attention is often strongly biased towards the center but the model must be able to deal with sudden task-driven changes in gaze direction.

5.2 Ablation study
In order to assess the design of the multi-path model illustrated in Sec 4.1, we study the individual contributions of the different branches to the final task in Tab. 4, by enabling and disabling one or more pathways of the architecture.

As one might expect, results testify that the RGB branch plays a major role in video saliency prediction. Nevertheless, the motion stream is beneficial and provides a slight improvement, that becomes clearer in the acting subsequences. Indeed, optical flow intrinsically captures a variety of peculiar scenarios that are non-trivial to classify when only color information is provided, e.g. when the car is waiting at a traffic light or is turning. The semantic stream provides a little performance boost as well, although minor with respect to motion. From Tab. 4 we can appreciate that semantics is helpful in increasing the IG measure. This can be intuitively motivated by observing that the semantic of a moving object can help in discerning among distractors and potential attractors far from the road center, such as traffic signs, semaphores and advertising panels.

Eventually, we test our model in different driving contexts (Fig. 13). As expected, the analysis of different landscapes testifies that the human attentional behavior is easier to predict in highways (HGH) rather than downtown (DWT), where the focus can shift towards more distractors. The model seems more reliable in evening scenarios (EVN), rather than morning (MRN) or night (NGH), where we observed better lightning conditions and lack of shadows, over-exposure and so on. Lastly, in rainy conditions (RNY) we notice that human gaze is easier to model, possibly due to the higher level of awareness demanded to the driver.

5.3 Do we capture the attentional dynamics?

The previous sections validate quantitatively the proposed model. Now, we assess its capability to attend like a human driver by comparing its predictions against the analysis performed in Sec. 3.1.

First, we report the average predicted fixation map in several speed ranges in Fig. 14. The conclusions we draw are twofold: i) generally, the model succeeds in modeling the behavior of the driver at different speeds, and ii) as the speed increases fixation maps exhibit lower variance, easing the modeling task, and prediction errors decrease.

We also study how often our model focuses on different semantic categories, in a fashion that recalls the analysis of Sec. 3.1, by employing our predictions rather than ground truth maps as focus of attention. Fig. 15 shows the comparison. Please note that, to highlight differences for low populated categories, values are reported on a logarithmic scale. Gray bars represent average ground truth values across different thresholds (Sec. 3.1), while color bars represent how often the predicted map focuses on a certain category. The plot shows a certain degree of absolute error.
Fig. 14. Model prediction averaged across all test sequences and grouped by driving speed. As the speed increases, the area of the predicted map shrinks, recalling the trend observed in ground truth maps (Fig. 6).

(a) $0 \leq \text{km/h} \leq 10$
(b) $10 \leq \text{km/h} \leq 30$
(c) $30 \leq \text{km/h} \leq 50$
(d) $50 \leq \text{km/h} \leq 70$
(e) $70 \leq \text{km/h}$

Fig. 15. Comparison between ground truth (gray bars) and predicted fixation maps (colored bars) when used to mask semantic segmentation of the scene. Absolute errors exist, but the two bar series agree on the relative importance of different categories.

Fig. 16. In this setting, frames are blurred such that for the observer of the videoclip is provided with the same amount/quality level of information as the driver was. Top: fixation map. Bottom: blurred frame mimicking visual perception.

Fig. 17. Comparison between ground truth (gray bars) and predicted fixation maps (colored bars) when used to mask semantic segmentation of the scene. Absolute errors exist, but the two bar series agree on the relative importance of different categories.

5.4 Qualitative assessment of predicted fixation maps

To validate our model for an Advanced Driver Assisted System (ADAS), we asked 18 people to take part to a qualitative assessment test by employing 200 videoclips (40 seconds long) extracted from DR(eye)VE. First, we replicated the cognitive perception of the scene of the driver by blurring pixels of the videoclips as they get further from fixation points. Such a transformation is inspired by the mechanism of peripheral vision, which captures details right where we focus and only low-frequency contextual information elsewhere. The result of such blurring can be appreciated in Fig. 16.

We let participants watch the videoclips and asked them to answer two questions:

- If you were sitting in the same car of the driver whose attentional behavior you just observed, how safe would you feel? (rate from 1 to 5)
- Would you say the observed attentional behavior comes from a human driver? (yes/no)

The first question aims at measuring the comfort level of the observer during a driving experience when suggested to focus at specific locations in the scene. The underlying assumption is that, if the suggested focus overlaps with a portion of the scene the observer also wishes to stare at, then he is going to feel safe and comfortable. Conversely, if the observer wishes to focus at some specific location but he cannot retrieve details there, he is going to feel unsafe and uncomfortable. The second question measures potential preference towards human common sense. Videos were blurred by randomly selecting either the ground truth fixation map (G videoclips) or the fixation map predicted by our model (P videoclips). To be fair to the ground truth, all the selected videoclips contain less than 0.5% of frames marked as errors. Results are shown in Fig. 17 where it emerges that the center of mass of the distribution of comfort levels collected for P videoclips (Fig. 17 (b)) is significantly higher than the ones collected for G videoclips (Fig. 17 (a)). Specifically, the average comfort score for predicted gaze is 4.16 against 3.26 for ground truth gaze. Interestingly, users found our predicted fixation maps even safer than attentional behaviour recorded from humans.

In Fig. 17 (c) we report the confusion matrix of the second question we asked to participants. We notice that participants were not particularly good at discriminating between human’s gaze and automatically generated maps scoring about the 54% of accuracy, which is comparable to random guessing. Interestingly, most of the higher and lower values of comfort (1 and 5) were attributed while the observer was wrong about the source of the fixation map. This is revealed by Fig. 17 (d), where the first and last bar are mainly composed by FN and FP respectively. This finding highlights that participants were more keen to label as automatically generated
behaviors which were judged unsafe, and vice-versa. Notably, all unsafest behaviors came from humans, and almost all of the behaviors judged as safest came from our model.

6 CONCLUSIONS
This paper presents a study of human attentional dynamics underpinning the driving experience. Our main finding is the emergence of common gaze patterns across drivers and different scenarios. We observed a consistent relation between changes in speed, lightning conditions, weather and landscape, and changes in the driver’s focus. Upon this insight, we developed a multi-path deep network to model and predict the driver’s gaze starting from raw video sequences. Experiments with the proposed architecture and related training strategies yielded state-of-the-art results. To our knowledge, our contribution is the first model able to predict human attention in driving sequences. As our model only input is car-centric videos, it can be seamlessly integrated with already adopted ADAS technologies. A qualitative study revealed that the attentional behavior produced by our system made users feel safer than human behavior.

REFERENCES
[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 1597–1604, June 2009.
[2] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara. Dr(Eye)Ve: A dataset for attention-based tasks with applications to autonomous and assisted driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 54–60, 2016.
[3] L. Bazzani, H. Larochelle, and L. Torresani. Recurrent mixture density network for spatiotemporal visual attention. In International Conference on Learning Representations (ICLR), 2017.
[4] G. Borghi, M. Venturirelli, R. Vezzani, and R. Cucchiara. Poseidon: Face-from-depth for driver pose estimation. In CVPR, 2017.
[5] A. Borji and L. Itti. Cai2000: A large scale fixation dataset for boosting saliency research. CVPR 2015 workshop on “Future of Datasets”, 2015.
[6] R. Brémont, J.-M. Auberlet, V. Cavallo, L. Désiré, V. Faure, S. Lemonnier, R. Lobjois, and J.-P. Tarel. Where we look when we drive: A multidisciplinary approach. In Proceedings of Transport Research Arena (TRA’14), Paris, France, 2014.
[7] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Met saliency benchmark. 2015. 1, 2, 3, 8, 9
[8] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? arXiv preprint arXiv:1604.03605, 2016.
[9] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals for accurate object class detection. In NIPS, 2015.
[10] M. M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S. M. Hu. Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(3):569–582, March 2015.
[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[12] M. Cornia, L. Baralidi, G. Serra, and R. Cucchiara. A Deep Multi-Level Network for Salience Prediction. In International Conference on Computer Vision and Pattern Recognition (ICPR), 2016.
[13] M. Cornia, L. Baralidi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. arXiv preprint arXiv:1611.09571, 2016.
[14] L. Elazary and L. Itti. A bayesian model for efficient visual search and recognition. Vision Research, 50(14):1338 – 1352, 2010.
[15] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin. A video saliency detection model in compressed domain. IEEE transactions on circuits and systems for video technology, 24(1):27–38, 2014.
[16] Y. Fang, Z. Wang, W. Lin, and Z. Fang. Video saliency incorporating spatiotemporal cues and uncertainty weighting. IEEE Transactions on Image Processing, 23(9):3910–3921, 2014.
[17] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6):381–395, 1981.
[18] L. Fridman, P. Langhans, J. Lee, and B. Reimer. Driver gaze estimation without using eye movement. arXiv preprint arXiv:1507.04760, 2015.
[19] L. Fridman, P. Langhans, J. Lee, and B. Reimer. Driver gaze region estimation without use of eye movement. IEEE Intelligent Systems, 31(3):49–56, May 2016.
[20] B. Frohlich, M. Enzweiler, and U. Franke. Will this car change the lane? turn signal recognition in the frequency domain. In 2014 IEEE Intelligent Vehicles Symposium Proceedings, pages 37–42. IEEE, 2014.
[21] D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausibility of the discriminant centersurround hypothesis for visual saliency. Journal of Vision, pages 1–18, 2008.
[22] T. Gkammas and C. Nikou. Guiding optical flow estimation using superpixels. In Digital Signal Processing (DSP), 2011 17th International Conference on, pages 1–6. IEEE, 2011.
[23] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. IEEE Trans. Pattern Anal. Mach. Intell., 34(10):1915–1926, Oct. 2012.
[24] R. Groner, F. Walder, and M. Groner. Looking at faces: Local and global features. In Advances in Psychology, 25(2):253–533, 1984.
[25] S. Grubmüller, J. Pilhal, and P. Nedoma. Automated Driving from the View of Technical Standards, pages 29–40. Springer International Publishing, Cham, 2017.
[26] J. M. Henderson. Human gaze control during real-world scene perception. Trends in cognitive sciences, 7(11):498–504, 2003.
[27] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In 2015 IEEE International Conference on Computer Vision (ICCV), pages 262–270, Dec 2015.
[28] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In Proceedings of the IEEE International Conference on Computer Vision, pages 3182–3190, 2015.
[29] M. Jiang, S. Huang, J. Yuan, and Q. Zhao. Salicon: Saliency in context. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.
[30] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-
Fei. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014.

[31] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR), 2015.

[32] P. Kumar, M. Perrollaz, S. Lefevre, and C. Laugier. Learning-based approach for online lane change intention prediction. In Intelligent Vehicles Symposium (IV), 2013 IEEE, pages 799–802, IEEE, 2013.

[33] M. Kümmerer, L. Theis, and M. Bethge. Deep gaze: boosting saliency prediction with feature maps trained on imagenet. In International Conference on Learning Representations Workshops (ICLRW), 2015.

[34] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu. Predicting eye fixations using convolutional neural networks. In Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, pages 362–370, June 2015.

[35] D. G. Lowe. Object recognition from local scale-invariant features. In Computer vision, 1999. The proceedings of the seventh IEEE international conference on, volume 2, pages 1150–1157. Ieee, 1999.

[36] Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In Proceedings of the Eleventh ACM International Conference on Multimedia, MULTIMEDIA '03, pages 374–381, New York, NY, USA, 2003. ACM.

[37] S. Mannan, K. Rudder, and D. Wooding. Fixation sequences made spatial vision. In Cognitive neuropsychology, 11(2):157–178, 1997.

[38] S. Mathe and C. Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(7):1408–1424, July 2015.

[39] T. Mauthner, H. Possegger, G. Waltner, and H. Bischof. Encoding based saliency detection for videos and images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2494–2502, 2015.

[40] B. Morris, A. Doshi, and M. Trivedi. Lane change intent prediction for driver assistance: On-road design and evaluation. In Intelligent Vehicles Symposium (IV), 2011 IEEE, pages 895–901, 2011.

[41] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka. 3d bounding box estimation using deep learning and geometry. In CVPR, 2017.

[42] D. Nilsson and C. Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. arXiv preprint arXiv:1612.08871, 2016.

[43] A. Palazzi, F. Solera, S. Calderara, S. Alotto, and R. Cucchiara. Where should you attend while driving? In IEEE Intelligent Vehicles Symposium Proceedings, 2017.

[44] M. I. Posner, R. D. Rafal, L. S. Chotard, and J. Vaughan. Inhibition of return: Neural basis and function. Cognitive neuropsychology, 2(3):211–224, 1985.

[45] N. Pugeault and R. Bowden. How much of driving is preattentive? IEEE Transactions on Vehicular Technology, 64(12):5424–5438, Dec 2015.

[46] J. Rogé, T. Pébaye, E. Lambilliotte, F. Spitzerstetten, D. Gisellebret, and A. Muzet. Influence of age, speed and duration of monotonous driving task in traffic on the drivers useful field of view. Vision research, 44(23):2737–2744, 2004.

[47] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor. Learning video saliency from human gaze using candidate selection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1147–1154, 2013.

[48] R. Rukinas, J. Back, P. Curzon, and A. Blandford. Formal modelling of salience and cognitive load. Electronic Notes in Theoretical Computer Science, 208:57 – 75, 2008.

[49] J. Sardegna, S. Shelly, and S. Steidl. The encyclopedia of blindness and vision impairment. Infobase Publishing, 2002.

[50] B. Schlkopf, J. Platt, and T. Hofmann. Graph-Based Visual Saliency, pages 545–552. MIT Press, 2007.

[51] L. Simón, J. P. Tarell, and R. Bremond. Alerting the drivers about road signs with poor visual saliency. In Intelligent Vehicles Symposium, 2009 IEEE, pages 48–53, June 2009.

[52] C. S. Stefan Mathe. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37, 2015.

[53] B. W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. Journal of vision, 7(14):4–4, 2007.

[54] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard. Eye guidance in natural vision: Reinterpreting salience. Journal of Vision, 11(5), May 2011.

[55] A. Tawari and M. M. Trivedi. Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos. In Intelligent Vehicles Symposium Proceedings, 2014 IEEE, pages 344–349, June 2014.
Simone Calderara received a computer engineering masters degree in 2005 and the PhD degree in 2009 from the University of Modena and Reggio Emilia, where he is currently an assistant professor within the Imagelab group. His current research interests include computer vision and machine learning applied to human behavior analysis, visual tracking in crowded scenarios, and time series analysis for forensic applications. He is a member of the IEEE.

Francesco Solera obtained a master degree in computer engineering from the University of Modena and Reggio Emilia in 2013 and a PhD degree in 2017. His research mainly addresses applied machine learning and social computer vision.

Rita Cucchiara received the masters degree in Electronic Engineering and the PhD degree in Computer Engineering from the University of Bologna, Italy, in 1989 and 1992, respectively. Since 2005, she is a full professor at the University of Modena and Reggio Emilia, Italy, where she heads the ImageLab group and is Director of the SOFTECH-ICT research center. She is currently President of the Italian Association of Pattern Recognition, (GIRPR), affiliated with IAPR. She published more than 300 papers on pattern recognition computer vision and multimedia, and in particular in human analysis, HBU and egocentric-vision. The research carried out spans on different application fields, such as video-surveillance, automotive and multimedia big data annotation. Currently she is AE of IEEE Transactions on Multimedia and serves in the Governing Board of IAPR and in the Advisory Board of the CVF.