Where Should You Attend While Driving?

Andrea Palazzi* Francesco Solera*
Simone Calderara Stefano Alletto Rita Cucchiara

University of Modena and Reggio Emilia
name.surname@unimore.it

Abstract

Despite the advent of autonomous cars, it’s likely - at least in the near future - that human attention will still maintain a central role as a guarantee in terms of legal responsibility during the driving task. In this paper we study the dynamics of the driver’s gaze and use it as a proxy to understand related attentional mechanisms. First, we build our analysis upon two questions: where and what the driver is looking at? Second, we model the driver’s gaze by training a coarse-to-fine convolutional network on short sequences extracted from the DR(eye)VE dataset. Experimental comparison against different baselines reveal that the driver’s gaze can indeed be learnt to some extent, despite i) being highly subjective and ii) having only one driver’s gaze available for each sequence due to the irreproducibility of the scene. Eventually, we advocate for a new assisted driving paradigm which suggests to the driver, with no intervention, where she should focus her attention.

1. Introduction

While autonomous driving is quickly reaching maturity, it’s not clear how far in time society will overlook the legal responsibility of the human driver [21]. Conversely, Advanced Driver Assistance Systems (ADAS) are human-centric and already established both in literature and in the market. The aim of assisted driving is to increase the safety of the driver and the road environment at large. This is usually done through collision avoidance systems, blind spot control, lane change assistance, traffic signs recognition and many others.

Some of the most ambitious examples of assisted driving are related to driver monitoring systems [11, 7, 14, 20], where the attentional behavior of the driver is parsed together with the road scene to predict potentially unsafe manoeuvres and act on the car in order to avoid them (either by signaling the driver or braking). However, all these approaches are limited by their ability to capture the true attentional and intentional behavior of the driver, which is still a complex and largely unsolved task today. Conversely, we advocate for a new assisted driving paradigm which suggests to the driver, with no hard intervention, where she should focus her attention. The problem is thus shifted from a personal level (what the driver is looking at) to a task-driven level (what the driver should be looking at). Following the notion that gaze is a primary cue to human visual attention, the contributions of this paper are threefold:

- we investigate the attentional dynamics during the driving task by employing data collected from eye tracking glasses and semantic cues, Fig. 1(c) and (d) respectively (Sec. 3);
- upon these findings, we build a deep network architecture to model human attention while driving, Fig. 1(b), and evaluate the ability of the proposed approach to replicate what we observed on humans (Sec. 4-6);
- to deal with subjectivity of ground truth data (i.e. the driver’s gaze), we extend the DR(eye)VE [1] dataset with new attentive vs inattentive annotation and devise a consistent protocol to remove the bias (Sec. 7).

* Authors contributed equally.
2. Related Works

Our proposal relates to the following topics.

**Head pose and gaze estimation.** Recent methods typically estimate the driver’s gaze direction using both head and eye location cues [30, 33, 2, 19]. The common aim of these works is usually to infer eyes off-the-road and other dangerous driving habits [31, 34, 6]. Other works exploit driver’s attention to predict her next maneuver, and either use a combination of head-pose estimation and GPS/maps [11] or collect precise gaze data through eye tracking devices [15]. On the opposite, [22] does not employ driver’s attention information (neither measured nor estimated) to predict actions. Interestingly, they show a strict correlation between the visual appearance of the scene and driver’s action such as steering, turning or braking. This finding also motivates our work, where we try to model the driver’s attention from the external scene only.

**Video saliency.** In this work we model the driver’s focus through attentional maps, which resemble video saliency maps. The video saliency prediction task is typically tackled by extracting low level (or bottom-up) features from each frame, either in an unsupervised [18, 35, 36] or supervised [38, 29] manner. Temporal dependencies are enforced either by conditioning the prediction on previous frames [24] or by means of optical flow [38, 29], which captures motion smoothness. To the best of our knowledge, this work is the first one to exploit deep learning to tackle the video saliency task.

**Datasets.** There are only three datasets that provide driver’s gaze annotation. In particular, gaze annotations in [26] were collected in lab settings by showing a small amount of still images to participants. Conversely, annotations in [6, 1] are collected during a real driving experience for more than 1,800,000 and 500,000 frames respectively. While [6] records data from many more drivers than [1], the latter presents a higher variety of driving scenarios and among all, it is the only one publicly available. Starting from the DR(eye)VE [1] dataset, we extend it with driver’s attention quality annotations, see Sec. 5.3 for details.

3. Experimental analysis of driver’s attention

In this section we investigate the attentional mechanisms involved in driving. To this end, we rely on the recently proposed DR(eye)VE [1] dataset, which is a collection of 74 sequences 5 minutes long, featuring different landscapes, weather scenarios, lighting conditions and 8 drivers. Every sequence is composed of two videos, capturing both the driver’s and the car point of view. While the driver’s gaze is synchronized with the former video, ground truth attentional maps are provided on the latter through a homographic projection followed by spatial smoothing and time integration. These latter post-processing steps attenuate subjectivity of the drivers’ scan-path and reduce measurement inaccuracies in the ground truth.

3.1. Where do we attend while driving?

The first observation that emerges from the analysis of the dataset is that drivers tend to focus their gaze towards the vanishing point of the road – often uncaring of close-by road signals, pedestrians walking on a sidewalk or cars coming from the opposite direction. This phenomenon can be qualitatively appreciated in Fig. 2. This finding can be explained through the notion of peripheral vision, introduced in physiological studies to indicate the area around the fixation point where objects and events can still be perceived and interpreted [25]. Accordingly, the largest area of attention is covered by focusing on the vanishing point of the road because, in the task of driving, many of the objects worth of attention have already been perceived when distant.

Additionally, we measured that the average location of the gaze deviates from the road vanishing point as a function of speed and landscape. In fact, due to limited resources our brain compensates the increase of information density - either spatial or temporal - with a reduction of the visual field size, so that everything that is perceived can also be elaborated [23]. We have an increase in information spatial density in complex scenes, where both many interesting elements and distractions are present. Similarly, when the driver travels at higher speed, we have an increase in information temporal density, e.g. the amount of information we need to elaborate per unit of time. Contextually to a decreased visual field, the driver needs to move the gaze around the scene more often to capture all the informative elements. Fig. 3 illustrates how the average gaze position changes at different speeds. When the speed increases - from less than 10km/h (a) to more than 70km/h (c) - the average gaze position converges towards the vanishing point of the road, suggesting an increased visual field size able to capture enough of the scene without shifting the gaze. The increased field size is the result of a linear increase of temporal density counterbalanced by a stronger decrease of information spatial density. In fact, low speed frames occur frequently in downtown sequences where high information.

![Figure 2](image-url)
spatial density is expected, while high speed frames mainly belong to highway driving events, where the amount of information is scarcer. The bar plots in Fig. 3 measure the amount of downtown (red), countryside (green) and highway (blue) frames that concurred to generate the average gaze position for a specific speed range.

### 3.2. What do we attend while driving?

To further investigate the attentional mechanism involved in the driving task, in this section we consider what the driver is looking at. To conduct this analysis we leverage on semantic segmentation - the task of assigning a label to each pixel according to what object it belongs to. We rely on the network proposed by Yu and Koltun [37], trained on the cityscapes [5] dataset, to recognize the following 10 categories: Road, sidewalk, buildings, traffic lights / signs, trees, road limits, sky, people, vehicles and cycles. Once all frames are labeled we can answer the question of what the driver looking at by measuring the segmentation labels density inside the attentional map\(^1\), see Fig. 1(d). Since the attentional maps have continuous values in \([0, 1]\), we build several binary versions. By moving the threshold towards 1, we shrink the observed portion of the scene close to the fixation point. Fig. 4 shows how different object categories react when such threshold is changed. In particular, an upward trend reveals that a specific category was indeed the focus of the gaze since, as the observed area decreases, the proportion of pixels labeled in that way increases. Interestingly, this is true for traffic lights, road limits, vehicles and people. Similarly, a downward trend reveals that focus on many categories is often only circumstantial, \(i.e.\) the categories contain objects that happened to be close to what the driver was really looking at. Examples include sidewalks, buildings, trees and sky. The irregular behavior of the road category can be motivated as follows: On one hand, the road is often close to other things we are looking at - so a low threshold should include big portion of the road - on the other hand, the road is often itself the true focus of the driver, hence the upward trend beyond the first threshold. A similar reasoning can explain the behavior of the cycle category. By fixing a threshold, the plot elucidates on the proportion of observed categories across all the sequences. Our focus is dominated by road and vehicles, but we often observe buildings and trees even if they contain little to none information useful to drive.

### 4. Network Models

In the previous section we studied the driver’s focus under a variety of conditions, aiming at providing insights on the mechanisms that govern human attention while driving. This, however, was only a part of our goal: Indeed, paraphrasing the renowned physicist Richard Feynman, we argue that what we cannot reproduce we do not fully understand. We hence tackle the challenge of learning a deep model of driver attention that, given a driving sequence,
Figure 5. **Coarse** prediction architecture. The first part of the network performs the feature encoding. The input videoclip is a tensor of size $3 \times 16 \times 112 \times 112$ that undergoes a sequence of conv3D and pool3D layers that gradually squeeze it to size $512 \times 1 \times 7 \times 7$. All conv3D have kernel size (3,3,3) and ReLU activation units; all pool3D have pool size (2,2,2) except the first one that has pool size (1,2,2). In order to obtain a saliency map with the same spatial size of the input frame, the feature representation is decoded through a series of intertwined layers of conv2D and $\times 2$ upsampling on the spatial dimensions. All conv2D have kernel size (3,3) and are followed by leaky ReLU activations with $\alpha = .001$. As a result, the output of the network is a tensor of size $1 \times 112 \times 112$, i.e. the predicted attentional map.

Figure 6. **Coarse+Fine** prediction architecture. The **Coarse** module (see Figure 5) is applied to both a cropped and a resized version of the input tensor, which is a videoclip of 16 consecutive frames. The cropped input is used during training to augment the data and the variety of ground truth attentional maps. The prediction of the resized input is stacked with the last RGB frame of the videoclip and fed to a series of convolutional layers (**Fine** prediction module) with the aim of refining the prediction. Training is performed end-to-end and weights between **Coarse** modules are shared. At test time, only the refined predictions are used.

would be able to focus where the human driver would.

As various recent works show [32, 12] when dealing with videos, taking explicitly into account the temporal dimension of the input in the network architecture can lead to results that easily outclass the single-frame-input baselines in various high-level video analysis tasks such as video classification and action recognition among others. To this end, we can distinguish at least two main trends that emerged in the recent literature. Those who just want to exploit short-range dependencies in the data structure often make use of 3D convolutional architectures in which data of successive time steps are stacked along an additional dimension of the input tensor. Conversely, if the task requires to capture longer-term interactions recurrent architecture (e.g. LSTM, GRU) are often the clear winners.

Here, we make the assumption that a short video sequence (e.g. half a second) contains enough information to successfully predict where the driver should look in that situation. Indeed, it can be argued that humans take even less time to react to a stimulus. For this reason, our proposed model consists in a deep network that takes as input a fixed-size sequence of 16 consecutive frames from a video (called from now on videoclip) and outputs the gaze map for the last frame of the input clip.

### 4.1. Coarse gaze prediction module

The core of our deep network model is a fully convolutional network whose architecture is represented in Fig. 5. The first half of the network acts as an encoder and maps the input videoclip in feature space. Conversely, the second block decodes the feature representation in an attentional map which has the same width and height of the input videoclip, but singleton temporal dimension. In order to perform the encoding, we employ the C3D network by [32] with pre-trained weights and few minor modifications, such as dropping the last convolutional and the fully connected layer block decodes the feature representation in an attentional map which has the same width and height of the input videoclip, but singleton temporal dimension. In order to perform the encoding, we employ the C3D network by [32] with pre-trained weights and few minor modifications, such as dropping the last convolutional and the fully connected layers in order to maintain spatial information that would be otherwise discarded. See Fig. 5 for further details.

During training we resize the training images to $128 \times 128$ and then randomly crop them to $112 \times 112$, following the training process described in [32]. Due to the strong bias towards the vanishing point, this standard cropping policy turned out to be insufficient for creating a real variety in the location of the attentional maps; thus the network prediction resulted strongly attracted towards the center of the image.
In this section we quantitatively measure the performance of the proposed COARSE and COARSE+FINE models against different baselines. Following the guidelines in [4], for the evaluation phase we rely on Cross-Correlation (CC) and Kullback–Leibler divergence (KL) measures.

**Training details.** The encoding half of the COARSE network is initialized with pre-trained weights [32]. Training sequences are randomly mirrored to augment the data. End-to-end training minimizes the Mean Squared Error for both losses of the COARSE+FINE model; we employ Adam optimizer with parameters suggested in the original paper [13]. 500 central frames from each training sequence compose the validation set and are not used in training.

### 5.1. Keeping it simple: Baselines from saliency

It is widely known that a centered Gaussian, stretched to the aspect ratio of the image, makes for an incredibly effective baseline for the visual saliency task. This static baseline scores better than many methods benchmarked on the MIT300 [3] dataset. Section 3.1 revealed that a similar bias affects the DR (eye)VE dataset (see Fig. 2). Thus, to validate the proposed model, we compare it against both the aforementioned baseline and a more task-driven version of it built as the average of all training set attentional maps, Fig. 7(c) and (d). Results are reported in Tab. 1: The second baseline performs better than both the Gaussian baseline and the COARSE model on the test set, but significantly worse than COARSE+FINE model. Indeed the latter has a relative gain of about +25% over the COARSE model and about +15% over the average training ground truth.
5.2. Comparison with state-of-the-art

In Tab. 2 we report results from two recent video saliency methods [35, 36] on the test set. Both approaches chosen for comparison are unsupervised and mainly rely on appearance and motion discontinuities. As a result, their saliency maps are strongly attracted by moving objects that, due to the aperture problem, are actually still in the scene and typically reside in the peripheral portion of the road. Results shown in Tab. 2 call for supervised methods aware of both the semantic of the scene and the peculiarities of the task. To our knowledge, our proposal is the first deep model for driving attention, and video saliency to a greater extent.

5.3. New annotations to escape the bias

Despite showing good results, the baselines introduced in Sec. 5.1 are of no interest for the driving task as they are not able to generalize when required. There is a strong unbalance between lots of trivial-to-predict scenarios of little interest and few but important hard-to-predict events.

To enable the evaluation of our model under such circumstances, we extend the DR(eye)VE dataset with a set of new annotations. For each sequences of the dataset, we select those sub-sequences whose ground truth poorly correlates with the average ground truth of the whole sequence (CC < 0.3), i.e. the focus of the driver is far from the vanishing point of the road.

Examples of such sequences are depicted in Fig. 8: Three human annotators inspected the selected frames and manually split them - by majority - into (a) attentive vs (b) inattentive behavior, (c) errors and (d) uninteresting events. Errors can happen either due measuring tool failure (e.g. in extreme lighting conditions) or because something went wrong in the successive phase of data processing. Inattentive behavior captures the driver in distracted frames. Uninteresting events occur when the driver’s attention is biased by long-term intentions - as when the driver is committed to taking a turn in 100m and she starts looking in that direction even if she is still driving straight. We label these events uninteresting as they cannot be employed in real time gaze prediction. Eventually, all sub-sequences that are not labeled as above denote attentive behavior.

The last column of Tab. 1 reports the results computed on the attentive subset. When tested on these sequences, the Gaussian baseline outperforms both the COARSE model and the average training ground truth baseline. To interpret this result, consider that when measuring a distance between distributions, a high probability but wrong prediction is severely penalized over a somehow uncertain prediction (i.e. a Gaussian with high variance). Nonetheless, the COARSE+FINE model scores higher than all other methods with a relative gain of about +35% over the second best.

6. Do we capture the attentional dynamics?

In Sec. 5 we quantitatively evaluated the proposed network. Here, we qualitatively investigate the ability of the model to learn both where and what a human driver would focus while driving. The results are then compared against the analysis previously introduced in Sec. 3.

Figure 9 shows the average attentional maps predicted by our model, arranged by speed range. On each plot we also overlay precision errors (green) and recall errors (red), i.e. pixels whose value differs by more than 10% from the analogous ground truth plots reported in Fig 3. We observe that i) the model generally succeeds in capturing the location of the driver gaze at different speed, ii) errors are mostly due to precision (prediction is wider than ground truth) and iii) eventually, errors decrease as the speed increases, as the lower variance of the gaze at high speed makes the modeling task easier.

Figure 10. Comparison between ground truth (gray bars) and predicted attentional 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.
7. Dealing with subjectivity in the data

As it often happens with experiments carried out on-the-field - and this is particularly true for research on autonomous/assisted driving - the irreproducibility of experimental conditions limits the amount, the objectivity and the reliability of collected ground truth. In turn, this questions the significance of the evaluation procedure: Is there any guarantee that the network is not learning to predict undesirable behaviors such as subjective and inattentive attitude, or to replicate measurement errors as in Fig. 8(b) and (c)?

While such undesirable examples are present in the training set, they cannot be safely removed: Deciding which frame represents which behavior (e.g. attentive vs. inattentive) is itself a subjective task. By building on the following proposition, this section we introduce a useful protocol to obtain such guarantee.

**Proposition 1** Test examples classified as (a)ttentive and (u)ninteresting (see Sec. 5.3) are likely to be sampled from the same distribution of the training set; while test examples classified as (i)nattentive and (e)rrors are likely not.

If the above proposition is true to a sufficient extent, then the independent and identically distributed hypothesis that lies at the core of every learning algorithm would not hold for frames with compromised ground truth. As a consequence, inattentive behavior and errors would still be learnt from the training data, but this would be of no practical use in predicting such examples at test time. Eventually, every performance improvement measured on the test data could be safely attributed to conscious and acceptable driving behavior.

In Sec. 7.1 we’ll show i) how to represent both training and testing sets through their respective probability distributions, ii) how these distributions can be estimated by sampling the dataset, and iii) how this is sufficient to reduce the labeling effort to a subset of non-ambiguous frames. In Sec. 7.2, Prop. 1 will be exploited to normalize results of Tab. 1 and make comparison with future methods fairer.

### 7.1 Measuring distance between joint distributions

To empirically validate Prop. 1 we rely on the Joint Distribution Discrepancy (JDD) [16], a way of measuring distance between any two joint distributions. We will take two distance measurements against the whole training set: First with test frames of type (a) and (u), $\text{JDD}_{au}$; second with test frames of type (i) and (e), $\text{JDD}_{ei}$. According to Prop. 1 the former discrepancy $\text{JDD}_{au}$ should be remarkably smaller than $\text{JDD}_{ei}$. To formalize the difference between the two discrepancies, we introduce the following quantity:

$$R = 1 - \frac{\text{JDD}_{au}}{\text{JDD}_{ei}},$$

which reaches the optimal value $R = 1$ when $\text{JDD}_{au} \to 0$ and $\text{JDD}_{ei} \to \infty$. JDD is built on the Maximum Mean Discrepancy (MMD) introduced in Gretton et al. [10] by noticing that a joint distribution can be mapped into a tensor product feature space via kernel embedding. In particular, the biased empirical estimation of JDD can be computed as:

$$\text{JDD} = \left(\frac{1}{n^2} \sum_{i,j=1}^{n} [k_\phi(x_i, x_j)k_\psi(y_i, y_j) + k_\phi(x_i', x_j')k_\psi(y_i', y_j') - 2k_\phi(x_i, x_j')k_\psi(y_i, y_j')] \right)^{\frac{1}{2}}$$

where $\{(x_i, y_i)\}$ and $\{(x_i', y_i')\}$ are sampled from two different joint distributions. We employed a RBF kernel ($k_\phi$).
and \(k(x, y)\) for both variables \(x\) and \(y\), which is known to be characteristic \([9]\), i.e. induces a one-to-one embedding between distributions and points of a RKHS.

Naive computation of Eq. (2) scales quadratically with the number \(n\) of elements sampled from each distribution. Long et al. \([16]\) suggest an efficient linear approximation which trades accuracy for a larger sample set, but the small amount of available annotation for type (a), (i) and (e) discourages this path. Instead, we will exploit the asymptotic behavior of the sampling error for this kind of discrepancies, which is known to decrease at rate \(O(n^{-\frac{1}{2}})\) \([10, 28]\). First, both \(\text{JDD}_{au}\) and \(\text{JDD}_{ei}\) are computed several times with \(n = \{1, 2, \ldots\}\) different samples. Second, a curve of the kind \(\text{JDD}(n) - \text{JDD}^* = \alpha n^{-\frac{1}{2}}\) is fitted on \(\text{JDD}(n)\), obtained by averaging all \(\text{JDD}\) results computed with \(n\) elements. The bias parameter \(\text{JDD}^*\) represents the limit value of the discrepancy as \(n \to \infty\). By least-square fitting we obtained \(\text{JDD}^*_{ei} = 0.0634\) and \(\text{JDD}^*_{au} = 0.0608\). Accordingly, the limit value for Eq. (1) is \(R^* = 1 - \text{JDD}^*_{au}/\text{JDD}^*_{ei} = 0.8921\), whose high value corroborates Prop. 1. Note that due to sampling in the evaluation of \(\text{JDD}\), only a small subset of data required the annotation.

### 7.2. Discounting the subjectivity bias

The computation of kernels in Eq. (2) requires \(x\) and \(y\) to be suitable representation of input and ground truth elements of the learning model. In the following, \(x\) is chosen to be the vector of unrolled activation maps of the last layer of the architecture (see Fig. 6). It is known that at this level of the network feature representation is strongly discriminative. Conversely, \(y\) is simply the vector of the unrolled ground truth attentional map. Features \(x\) and \(y\) are by construction a fair choice for future comparisons, since they can always be obtained independently of the specific architecture of a network. On the downside, the whole analysis and computation of \(R^*\) is now dependant on both the learning model and the training procedure. This had to be expected: The validity of Prop. 1 depends both on the data and on the ability of the learner to model hard input/output relations, \(i.e.\) between training and test examples of type (i) and (e).

\(R^*\) can indeed be interpreted as the hardness of such relation modeling task for the considered learner. A learner for which this task is easy, will have a low \(R^*\) value and will tend to overfit to examples of type (i) and (e) at test time. Conversely, a learner for which this task is hard will have a high \(R^*\) value and will likely not be able to predict type (i) and (e) test examples.

Interestingly, the above interpretation of \(R^*\) opens up a way to remove the bias of inattentive behavior and errors from test results (Tab. 1). \(\text{CC}\) and \(\text{KL}\) values can be discounted as \(\overline{\text{CC}} = Z \cdot \text{CC}\) and \(\overline{\text{KL}} = Z^{-1} \cdot \text{KL}\) through the following multiplicative factor:

\[
Z = \frac{(1 - X) + R^* \cdot X}{1 - R^* \cdot X},
\]

where \(X\) is the percentage of frames of type (i) and (e) in the test set. \(Z\) can be seen as a linear combination of 1 and \(R^*\) weighted by the amount of good and undesirable frames respectively, while the denominator is for normalization. Figure 11(a) depicts the logarithmic values of \(Z\) when \(X\) and \(R^*\) vary. However, the main argument behind this analysis follows from the impossibility of precisely labeling as (a), (i), (u) or (e) all frames of the dataset, mainly due to subjectivity and ambiguity. So the exact percentage \(X\) of frames classified as (i) and (e) is not available. To overcome this limitation, we apply the correction by conjecturing a value \(X\) varying from 0% to 100%. As a consequence, starting from a single value of \(\text{CC}/\text{KL}\) we obtain a set of corrected values \(\overline{\text{CC}}/\overline{\text{KL}}\), as shown in Fig. 11(b). Different methods can be compared through their curves, or their Area Under the Curve if a single value has to be preferred. This comparison guarantees that different methods are evaluated on how well they predict desirable and objective behaviors. To a greater extent, this evaluation policy can be adapted to any method aiming to model task-driven attention and should complete standard experiments for a fairer comparison.

![Figure 11](image_url)

(a) correction factor (log)  
(b) corrected scores  

Figure 11. (a) Correction factor \(Z\) (log values) as a function of \(X\) and \(R^*\). (b) Corrected scores \(\overline{\text{CC}}\) and \(\overline{\text{KL}}\) for \(R^* = 0.8921\).

### 8. Conclusions

In this work we investigated the spatial and semantic attentional dynamics of the human driver and designed a deep network able to replicate attentive behavior during the driving experience. We also discussed the intrinsic subjectivity that resides in the ground truth gaze data and proposed an evaluation protocol to discount the driver’s subjective behavior bias from performance results. These results eventually pave the way for a new assisted driving module, where real-time attentional maps support the driver by both decreasing fatigue and helping in keeping focus. We argue that such attentional maps can be less invasive than other Advanced Driver Assistance Systems that directly act on the car (\(e.g.\) by activating breaks), whereas attentional maps leave full control to the driver.
Acknowledgments

We acknowledge the support of Facebook with the donation of the GPUs used in this work.

References

[1] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara. DrEye/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. 1, 2

[2] T. Bär, J. F. Reuter, and J. M. Zöllner. Driver head pose and gaze estimation based on multi-template icp 3-d point cloud alignment. In 2012 15th International IEEE Conference on Intelligent Transportation Systems, pages 1797–1802. IEEE, 2012. 2

[3] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark. 5

[4] 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. 5

[5] 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. 3

[6] L. Fridman, P. Langhans, J. Lee, and B. Reimer. Driver gaze estimation without using eye movement. arXiv preprint arXiv:1507.04760, 2015. 2

[7] B. Fröhlich, 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. 1

[8] S. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In NIPS, volume 20, pages 489–496, 2007. 14

[9] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In Advances in neural information processing systems, pages 513–520, 2006. 8

[10] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. Journal of Machine Learning Research, 13(Mar):723–773, 2012. 7, 8, 14

[11] 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. 1, 2

[12] 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. 4

[13] D. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 5

[14] 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 797–802. IEEE, 2013. 1

[15] F. Lethaus, M. R. Baumann, F. Köster, and K. Lemmer. Using pattern recognition to predict driver intent. In International Conference on Adaptive and Natural Computing Algorithms, pages 140–149. Springer, 2011. 2

[16] M. Long, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks. arXiv preprint arXiv:1605.06636, 2016. 7, 8, 14

[17] 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. 11

[18] 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. 2

[19] R. O. Mbouna, S. G. Kong, and M.-G. Chun. Visual analysis of eye state and head pose for driver alertness monitoring. IEEE transactions on intelligent transportation systems, 14(3):1462–1469, 2013. 2

[20] 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. IEEE, 2011. 1

[21] U. D. of Transportation. Federal automated vehicles policy. https://www.transportation.gov/sites/dot.gov/files/docs/AV%20Policy%20guidance%20PDF.pdf, September 2016. 1

[22] N. Pugeault and R. Bowden. How much of driving is pre-attentive? IEEE Transactions on Vehicular Technology, 2015. 2

[23] J. Rogé, T. Pébayle, E. Lambilliotte, F. Spitzenstetter, D. Giselbrecht, and A. Muzet. Influence of age, speed and duration of monotonous driving task in traffic on the drivers useful visual field. Vision research, 44(23):2737–2744, 2004. 2

[24] 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. 2

[25] J. Sardegna, S. Shelly, and S. Steidl. The encyclopedia of blindness and vision impairment. Infobase Publishing, 2002. 2

[26] L. Simon, J. P. Tarel, and R. Bremond. Alerting the drivers about road signs with poor visual saliency. In Intelligent Vehicles Symposium, 2009 IEEE, pages 48–53, June 2009. 2

[27] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In International Conference on Algorithmic Learning Theory, pages 13–31. Springer, 2007. 14

[28] F. Solera and A. Palazzi. A statistical test for joint distributions equivalence. arXiv preprint arXiv:1607.07270, 2016. 8

[29] C. S. Stefan Mathe. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition.
[30] A. Tawari, K. H. Chen, and M. M. Trivedi. Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation. In 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), pages 988–994. IEEE, 2014.

[31] A. Tawari and M. M. Trivedi. Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos. In 2014 IEEE Intelligent Vehicles Symposium Proceedings, pages 344–349. IEEE, 2014.

[32] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. arXiv preprint arXiv:1412.0767, 2014.

[33] R. Valenti, N. Sebe, and T. Gevers. Combining head pose and eye location information for gaze estimation. IEEE Transactions on Image Processing, 21(2):802–815, 2012.

[34] F. Vicente, Z. Huang, X. Xiong, F. De la Torre, W. Zhang, and D. Levi. Driver gaze tracking and eyes off the road detection system. IEEE Transactions on Intelligent Transportation Systems, 16(4):2014–2027, 2015.

[35] W. Wang, J. Shen, and F. Porikli. Saliency-aware geodesic video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3395–3402, 2015.

[36] W. Wang, J. Shen, and L. Shao. Consistent video saliency using local gradient flow optimization and global refinement. IEEE Transactions on Image Processing, 24(11):4185–4196, 2015.

[37] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122, 2015.

[38] S.-h. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren. Video saliency detection via dynamic consistent spatio-temporal attention modelling. In AAAI, 2013.
Supplementary materials
Where Should You Attend While Driving?

The following supplementary materials complete the submitted manuscript with additional experimental details and qualitative evaluation. In particular, Sec. A introduces a new comparison against a supervised video saliency method and some visual examples of how competitors perform on the DR(eye)VE dataset; Sec. B presents the results of a visual quality assessment test we conducted to further validate our predictions; Sec. C shows per-sequence results of the proposed approach on all sequences as well as only on (a)-labelled ones (see Sec. 5.3 of the paper for more details); Sec. D shows the convergence trends of the JDD scores as the number of samples increases; and Sec. E is an introduction to the math underlying our subjectivity analysis (Sec. 7 of the paper).

A. Comparison with a supervised video saliency method [17]

To further extend the comparison against other video saliency methods (discussed in Sec. 5.2 of the paper), we report in Tab. 3 the performance of [17] on the DR(eye)VE dataset. Unlike previous methods reported in Sec. 5.2 of the paper, this approach is supervised. [17] detector uses both static (HoG) and motion (MBH) descriptors centered at fixations. On top of these features, extracted at 3 different scale levels, a linear SVM is trained with data from the Action in the Eye dataset, also from the same authors. The detector is then run in a sliding window fashion across the entire video to obtain a saliency map. Unfortunately, the dataset used to train the detector comes from a domain (e.g. action recognition) fairly different than the one of DR(eye)VE and authors do not provide code for re-training. This explains the low results reported on Tab. 3. To the best of our knowledge, no supervised video saliency method releases code for re-training, nor publishes results on DR(eye)VE. Figure 12 shows some qualitative results of the aforementioned approach by Mathe et al. [17] (last row, last column), as well as the unsupervised methods already introduced in Sec. 5.2 of the paper (last row).

| Method          | CC     | KL     |
|-----------------|--------|--------|
| Mathe et al. [17]| 0.04 ± 0.08 | 3.92 ± 0.53 |

Table 3. Comparison with Mathe et al. [17]

B. Qualitative assessment of predicted attentional maps

To validate our new feature 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 the fixation point. 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. 13.

![Figure 13](image)

(a) attentional map  
(b) blurred frame mimicking visual perception

Figure 13. 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.

Second, we let participants watch the videoclips and asked them to answer to 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 wishes to stare too, then he would feel safe and comfortable. Conversely, if the observer wishes to focus at some specific location but he cannot retrieve details there, he would then feel unsafe and uncomfortable. On the other hand, the second question measures potential biases towards humans better judgement. Videos were randomly transformed by using either the ground truth gaze (G) or the attentional map predicted by our model (P). 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. 14.

![Figure 14](image)

(a) (b) (c)

Figure 14. Bar plots in (a) and (b) depict the distributions of comfort level scores as a function of the attentional map source, i.e. groundtruth (a) and our approach prediction (b). Confusion matrix in (c) reports the results of observers’ guesses on the source of attentional maps.

From Fig. 14 we observe that the center of mass of the distribution of comfort levels collected when the videoclip was transformed with the predicted attentional map (P, Fig. 14(b)) is significantly higher than when the ground truth attentional map was used (G, Fig. 14(a)). Specifically, the average comfort score for predicted gaze is 4.16 against 3.26 for ground truth gaze. Interestingly, users found our predicted attentional maps even safer than humans true behavior.

To study the effect of the observers’ bias towards humans better judgement we measured how easy it was for them to understand whether the attentional maps came from a human or were automatically generated. In Fig. 14(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 attentional map. This is revealed by Fig. 15, where the first and last bar are mainly composed by FN and FP respectively.

![Figure 15](image)

**Figure 15.** Composition of scores for different levels of comfort. TP, TN, FP, FN refer to the guess of the observer on whether the attentional map came from a human driver or from our algorithm.

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 system.

### C. Detailed per-sequence results

Figure 16 complements Tab. 1 of the paper, by inspecting our method performance scores for each individual sequence of the test set.

![Figure 16](image)

**Figure 16.** Detailed CC (a) and KL (b) results for all sequences (blue) and attentive subsequences (red) of the test set.

### D. JDD convergence trends

In Sec. 7.1 of the paper, we fitted curves of the kind \( y = \alpha n^{-\frac{1}{2}} + \beta \) on JDD\(_{au}\) and JDD\(_{ei}\) values computed at increasing sample size \( n \), in order to compute their limit values \( \beta \). Below, we depicts the convergence trends and the fitted curves for both discrepancies.

### E. Background on MDD and JDD

The main idea behind MMD and JDD is to measure distance between distributions by comparing their embeddings in a Reproducing Kernel Hilbert Space (RKHS). RKHS \( \mathcal{H} \) is a Hilbert space of functions \( f : \Omega \to \mathbb{R} \) equipped with inner
Figure 17. Convergence trends (green) for the discrepancy computed against inattentive behavior and corrupted groundtruth (a), and attentive behavior and uninteresting events (b). Fitted lines are plotted in blue, while red lines mark their limit values.

products $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norms $\| \cdot \|_{\mathcal{H}}$. In the context of this work, all elements $f \in \mathcal{H}$ of the space are probability distributions that can be evaluated by means of inner products $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$ with $x \in \Omega$, thanks to the reproducing property. $k$ is a kernel function that takes care of the embedding by defining an implicit feature mapping $k(x, \cdot) = \phi(x)$, where $\phi : \Omega \mapsto \mathcal{H}$. As always, $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ can be viewed as a measure of similarity between points $x, x' \in \Omega$. If a characteristic kernel is used, then the embedding is injective and can uniquely preserve all the information about a distribution [8]. According to the seminal work by Smola et al. [27], the kernel embedding of a distribution $p(x)$ in $\mathcal{H}$ is given by $E_x[k(x, \cdot)] = E_x[\phi(x)] = \int_\Omega \phi(x) dP(x)$.

Having all the required tools in place, we can introduce the MMD and the JDD.

**Definition 1 (Maximum Mean Discrepancy (MMD) [10])** Let $\mathcal{F} \subset \mathcal{H}$ be the unit ball in a RKHS. If $x$ and $x'$ are samples from distributions $p$ and $q$ respectively, then the MMD is

$$\text{MMD}(\mathcal{F}, p, q) = \sup_{f \in \mathcal{F}} (E_x[f(x)] - E_{x'}[f(x')])$$

(4)

**Definition 2 (Joint Distribution Discrepancy (JDD) [16])** Let $\mathcal{F} \subset \mathcal{H}$ be the unit ball in a RKHS. If $(x, y)$ and $(x', y')$ are samples from joint distributions $p$ and $q$ respectively, then the JDD is

$$\text{JDD}(\mathcal{F}, p, q) = \sup_{f,g \in \mathcal{F}} \left( E_{x,y}[f(x)g(y)] - E_{x',y'}[f(x')g(y')] \right)$$

$$= \| E_{x,y}[\phi(x) \otimes \psi(y)] - E_{x',y'}[\phi(x') \otimes \psi(y')] \|_{\mathcal{F} \otimes \mathcal{F}},$$

where $\phi$ and $\psi$ are the mappings yielding to kernels $k_\phi$ and $k_\psi$, respectively.

Note that, conversely to Long et al. [16], we don’t square the norm in Eq. (5). A biased empirical estimation of JDD can be obtained by replacing the population expectation with the empirical expectation computed on samples $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \in X \times Y$ from $p$ and samples $\{(x'_1, y'_1), (x'_2, y'_2), \ldots, (x'_n, y'_n)\} \in X' \times Y'$ from $q$:

$$\text{JDD}_b(\mathcal{F}, X, Y, X', Y') = \sup_{f,g \in \mathcal{F}} \left( \frac{1}{m} \sum_{i=1}^{m} f(x_i)g(y_i) - \frac{1}{n} \sum_{i=1}^{n} f(x'_i)g(y'_i) \right)$$

$$= \left\| \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) \otimes \psi(y_i) - \frac{1}{n} \sum_{i=1}^{n} \phi(x'_i) \otimes \psi(y'_i) \right\|_{\mathcal{F} \otimes \mathcal{F}}$$

$$= \left( \frac{1}{m^2} \sum_{i,j=1}^{m} k_\phi(x_i, x_j)k_\psi(y_i, y_j) + \frac{1}{n^2} \sum_{i,j=1}^{n} k_\phi(x'_i, x'_j)k_\psi(y'_i, y'_j) + \ldots 
- \frac{2}{mn} \sum_{i,j=1}^{m} \sum_{j=1}^{n} k_\phi(x_i, x'_j)k_\psi(y_i, y'_j) \right)^{\frac{1}{2}}$$

(6)

Moreover, throughout the paper, we restrict ourselves to the case of $m = n$ and bounded kernels, specifically $0 \leq k(x_i, x_j) \leq K$, for all $i$ and $j$ and for all kernels.