X-AWARE: ConteXt-AWARE Human-Environment Attention Fusion for Driver Gaze Prediction in the Wild

Lukas Stappen
University of Augsburg
Chair of Embedded Intelligence for Health Care and Wellbeing
Augsburg, Germany
stappen@ieee.org

Georgios Rizos
Imperial College London
Group on Language, Audio, & Music
London, United Kingdom
rizos@ieee.org

Björn W. Schuller
Imperial College London
Group on Language, Audio, & Music
London, United Kingdom
schuller@ieee.org

ABSTRACT

Reliable systems for automatic estimation of the driver's gaze are crucial for reducing the number of traffic fatalities and for many emerging research areas aimed at developing intelligent vehicle-passenger systems. Gaze estimation is a challenging task, especially in environments with varying illumination and reflection properties. Furthermore, there is wide diversity with respect to the appearance of drivers' faces, both in terms of occlusions (e. g., vision aids) and cultural/ethnic backgrounds. For this reason, analysing the face along with contextual information – for example, the vehicle cabin environment – adds another, less subjective signal towards the design of robust systems for passenger gaze estimation. In this paper, we present an integrated approach to jointly model different features for this task. In particular, to improve the fusion of the visually captured environment with the driver's face, we have developed a contextual attention mechanism, X-AWARE , attached directly to the output convolutional layers of InceptionResNetV2 networks. In order to showcase the effectiveness of our approach, we use the Driver Gaze in the Wild dataset, recently released as part of the Eighth Emotion Recognition in the Wild Challenge (EmotiW) challenge. Our best model outperforms the baseline by an absolute of 15.03 % in accuracy on the validation set, and improves the previously best reported result by an absolute of 8.72 % on the test set.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Scene understanding**; • **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

gaze detection; attention fusion; context aware; in the wild

ACM ISBN 978-1-4503-7581-8
https://doi.org/10.1145/3382507.3417967

ACM Reference Format:
Lukas Stappen, Georgios Rizos, and Björn W. Schuller. 2020. X-AWARE: ConteXt-AWARE Human-Environment Attention Fusion for Driver Gaze Prediction in the Wild. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3382507.3417967

1 INTRODUCTION

According to the World Health Organization, 1.35 million people worldwide lose their lives in traffic accidents each year, which is currently the leading cause of death among young people aged 5-29 years [25]. Distraction is listed as a major contributing factor, especially for novice drivers. In the interest of developing advanced automatic driving and personal assistance systems, the importance of taking into account the human factor, such as by capturing driver and passenger behavioural indices, is well-established [12, 21, 22, 30]. A popular approach here is to gain some insight into the driver's mental state, for example by means of using head orientation and pupil position, in order to assess driver attention [2, 21, 22, 42].

A non-invasive, non-contact approach towards driver distraction estimation is by the analysis of visual signals [41, 42] recorded by cabin-placed cameras (instead of head-mounted), as it refrains from introducing additional distractions that might deteriorate driving ability, or the integrity of the study. Since there can be a large diversity in the facial characteristics of people in a gaze detection dataset, as well as the background environment, there is an inherent need to address gaze detection and eye tracking studies 'in the wild' [10, 20, 53]. Such a realistic setting is very desirable in the context of driver gaze detection, due to the catastrophic costs of a system failure in driving-related applications, and as such, 'in the wild' studies pose a great challenge of utmost importance.

In this paper, we focus on driver gaze estimation of nine coarse regions in an 'in the wild' setting. We work with the Driver Gaze in the Wild (DGW) dataset [11] released as part of the eighth Emotion Recognition in the Wild Challenge (EmotiW) [1]. It exhibits several modelling challenges, such as diverse ethnic background among the subjects, varying illumination, and potential presence of reflections in the face and environment. Moreover, there is an imbalanced gender representation in the dataset, as well as the fact that several drivers wear glasses, which may occlude parts of the eye. Most importantly, each participant utilises a unique head-eye coordination strategy, in order to focus on areas of interest. Whereas a calibration step is traditionally applied on each specific

user in order to accommodate for the latter issue [10, 20], there is a trend in research for calibration-free methods [9, 53] that eschew the need for collecting personalised training data; something that can be prohibitive when one considers the truly diverse population of drivers.

Whereas head pose, and eye tracking have long been considered as main components of gaze detection [10, 42], an under-explored source of information is the person's environment, in this case the passenger cabin. We hypothesise that instead of treating the environment as unknowable, stochastic noise that can be discarded during preprocessing, we can augment our understanding of the visual scene by providing our classifier with explicit environment context information. Alongside the estimation of the facial gaze as in [9], we assume that continuously or sporadically visible car parts can serve as learnable anchor points and enable the models to automatically learn both, the position of the camera depending on the car parts and the environmental context in a car. This idea represents a first step towards an object- rather than purely human/face-centred calibration approach that could eliminate several problems, such as poor prediction quality in recognition algorithms trained on culturally biased data. Furthermore, detecting gaze in an automotive environment is not only limited to the driver and a fixed camera setting. Other research areas, for example, passenger-vehicle interaction [43] and multimodal sentiment analysis in the wild [36] covering multiple passengers, un- and differently mounted cameras and several perspectives.

Towards this goal, we propose a method to model various confounding environmental context features, potentially present in the visual recording. Inspired by the success of attention mechanism variations [27, 32], e. g., in modelling facial activity [38], we propose to utilise attention on top of convolutional layers for fusion at the core of our method. Thereby, we aim to fulfill two desiderata: a) important regions of the raw image – not necessarily limited to facial features – are identified early during the image processing architecture, and b) separate vision and feature stream representations, including environment context related ones, are fused in an intelligent, data-driven manner. Our best method – **X-AWARE** – outperforms the official challenge baseline, previous work, as well as simpler interaction and fusion blocks we developed.

Since some parts of our method are specifically designed to address and accommodate the input features made available for the eighth EmotiW challenge, we first introduce the dataset and extracted features in Section 3, followed by the proposed approach in Section 4, and our preliminary and fusion experiments in Section 5. Finally, we discuss the results, point out possible improvements of our work, and link gaze estimation to the field of multimodal sentiment analysis in Section 6. The code and the weights of the best model are publicly available on the project repository[1].

2 RELATED WORK

Related fields to gaze zone estimation, such as head pose estimation [24] and gaze tracking [4], have a long history in computer vision. Many apply Bayesian filtering achieving reasonable results [13, 47]. However, most systems designed for controlled environments are not robust enough for the use in the context of human-robot interaction and driving assistance systems. With this goal in mind, [29] developed four Convolutional Neural Network (CNN) architectures and evaluated them on 'in the wild' datasets. The authors of [8] argue that with increasing distance to the person, the face resolution deteriorates. This, combined with inaccurate annotation, worsens prediction accuracy in the real world scenarios, to which end they proposed an invasive annotation method using eye-tracking glasses. To the best of our knowledge, [44] is the only work that focuses on gaze zone estimation and evaluates performance depending on different image crop regions. Some parts of the face, the entire face, and the face with an enlarged bounding box were used separately for training, whereby all of the different CNN architectures tested achieved the best results with first.

One way to improve accuracy, especially in human-machine interaction, is multimodal fusion of face images and poses using deep learning as in [16]. Similarly, the authors of [23] suggested the fusion of close facial images and depth signals. Recently, [50] highlighted the importance of gaze as a weakly supervised signal to attend regions in order to achieve better scene understanding, and introduced a human intention-driven framework fusing gaze, as well as distances between body joints and various object instances. With regard to the driver's gaze estimation, the study performed in [6] reports state-of-the-art results on two gaze datasets in automotive setup by including upper body poses in addition to the face depth images. By merging them into an in-depth learning model as well, the authors of [5] demonstrated the value of extended input in dealing with the 'in the wild' factors present in the automotive context.

In order to make use of eye and gaze tracking technology in real environments, [18] emphasised the crucial importance of algorithms for *implicit calibration*. Dynamic changes of environments and distances usually require intrusive, user-specific calibration methods for precise predictions, which are disliked by users, and are impractical for daily use. The authors of [45] proposed an implicit 3D modelling method based on several eye-tracking principles, e. g., that in most scenarios the fixations are concentrated on the centre of the screen. For their proposed person-free calibration, they used a hard expectation maximisation algorithm with good results. Recently, in the study performed in [19], the detection of probable eye targets for implicit calibration in order to take advantage of static, surrounding objects was introduced. Here instead, we utilised end-to-end deep learning from the raw images that include the cabin environment, in conjunction with Generic, optical Car Part Recognition and Detection (GoCaRD) [37] features, in an attempt to allow for a similar *neurally learnable implicit calibration*.

All state-of-the-art vision architectures we used as a first component for image processing are CNN-based. In the following, we give a brief overview of the most common architecture(s) (building blocks), also evaluated in our preliminary experiments in Section 5.3. In 2014, the VGG16 and VGG19 networks, with 16 and 19 layers with small convolution filters respectively, achieved substantial gains in benchmarks for large-scale image recognition [35], and initiated diverse research involving CNNs in the field of computer vision. The addition of residual connections to the network [15] allowed for training a deep network architecture called ResNet50 with up to

[1] https://github.com/lstappen/xaware

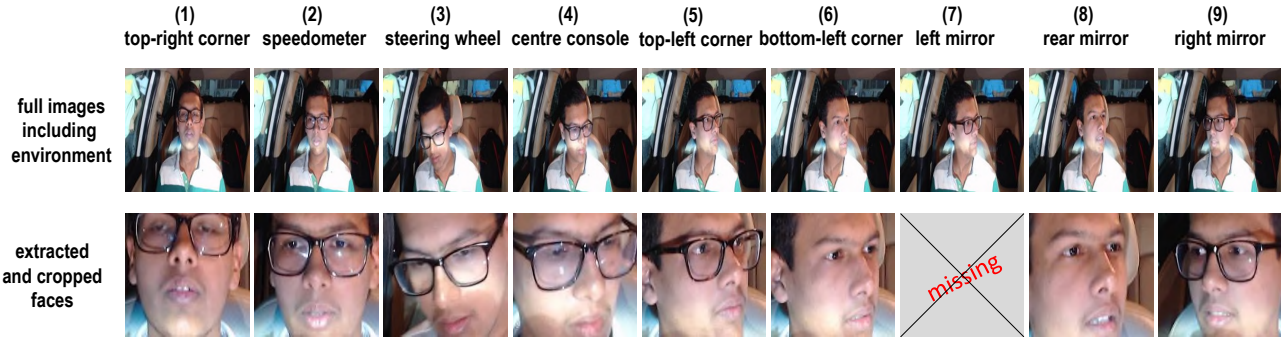

Figure 1: Example of pairs of full and cropped input images for the nine gaze zones for one participant in the Driver Gaze in the Wild (DGW) dataset. Some challenging qualities of the task are noticeable here; the same subject is also in the dataset without spectacles (illustrated for classes 6 and 8) and for class 7 the automatic face extraction failed. In cases such as the latter, we expect that the full image representation stream will still serve as an adequate substitute of purely face related features, especially due to the cross-stream attentional feature learning of X-AWARE. An example motivation for using the full image along the cropped one is also shown here: there is a highly illuminated patch that appears on a different part of the subject's face; however, it is in the same place compared to the background.

152 layers, while keeping the number of parameters from growing prohibitively high. In an attempt to avoid making a hard decision on a single convolution filter shape, the authors of [40] introduced the Inception module, i. e., the parallelisation of several convolution operations of different shapes. To stabilise and accelerate the training process, InceptionV3 was initially updated by adding batch normalisation, auxiliary classifiers and regularisation [40], and then further by the employment of residual connections towards the introduction of the InceptionResNetV2 model [39]. The authors of [7] developed Xception which outperformed InceptionV3 significantly on a dataset consisting of 350 million images and 17 000 classes. This can largely be attributed to the replacement of the Inception modules with depth-wise separable convolutions that receive the input on multiple channels of the previous layer.

3 DATASET AND FEATURES

3.1 Study setting and dataset

The dataset was made available for the Driver Gaze Prediction in the wild sub-challenge, as part of the eighth EmotiW. It consists of 50 thousand frames, partitioned in training, validation, and test set (60-20-20 split) for driver gaze zone estimation. These images are extracted from 586 video recordings of the original DGW dataset [11]. The data collection process can be described as follows: 247 male and 91 female subjects are seated in a car and look at different directions at pre-defined zones as depicted in Figure 1. A Microsoft Lifecam RGB is mounted in position frontally to the test person. The participants indicate the targeted zone themselves by reading aloud the class number on the signs attached to the nine zones. In this purely voice-based labelling approach, relevant frames are identified and automatically labelled by processing the videos using automatic speech recognition. The authors specify that the participants are between 18 and 63 years of age, with most (over 90 %) being between 18 and 35 years old. The participants were not given any guidance on how to look at the zones, and thus, we observe some movement of their head and/or eyes. As this study is considered to be 'in the wild', several factors present in realistic settings are considered. Firstly, 30 % of the participants are filmed with and without wearing their spectacles (cf. Figure 1, class 6). With regard to the environment, there is a diversity of backgrounds (windows scenes) and illumination due to changes in parking positions, weather conditions, and day times (half during the day and the other half in the evening). All our experiments use the officially provided sets.

3.2 Image data preparation

We extract the faces from the images using the computer vision libaries Dlib[2] and OpenCV[3] for an efficient and accurate face extraction. In a two step procedure, we first use a linear classifier on Histogram of Oriented Gradients features. The classifier detects faces on 66 % of all images. We then feed the remaining images into a pretrained convolutional face detection network provided by dlib, leaving us with a total of 0.002 % images without a face detected. An example face detection failure is illustrated in Figure 1, class 7. In order to train a stable model with these missing images, we replace them by a matrix of the same size, initialised with the smallest *epsilon* value available on pytorch. Since the authors [11] stated that an average face is represented by 179 x 179 pixels, we re-scale the identified face areas to 150 x 150 before export. In a random quality performance check, we visually inspected the extracted images and bounding boxes; no false positive recognised faces were spotted, which underlines the high quality of this simple face extraction method for this 'in the wild' dataset. The image pixel values are first scaled to a range between 0 and 1 (dividing by 255) and then standardised using a mean and a standard deviation value of 0.5 on each of the three channels [39].

3.3 Facial features

Over time, researchers have identified positions of expressive facial features, such as, the corners of the mouth and eyes, tip of the nose and pupil movement, which can be extracted from facial images and used to estimate facial posture and direction. Based on the face frameworks mentioned in the previous section, we obtain 138 facial

[2] http://dlib.net/python/index.html
[3] https://opencv.org/

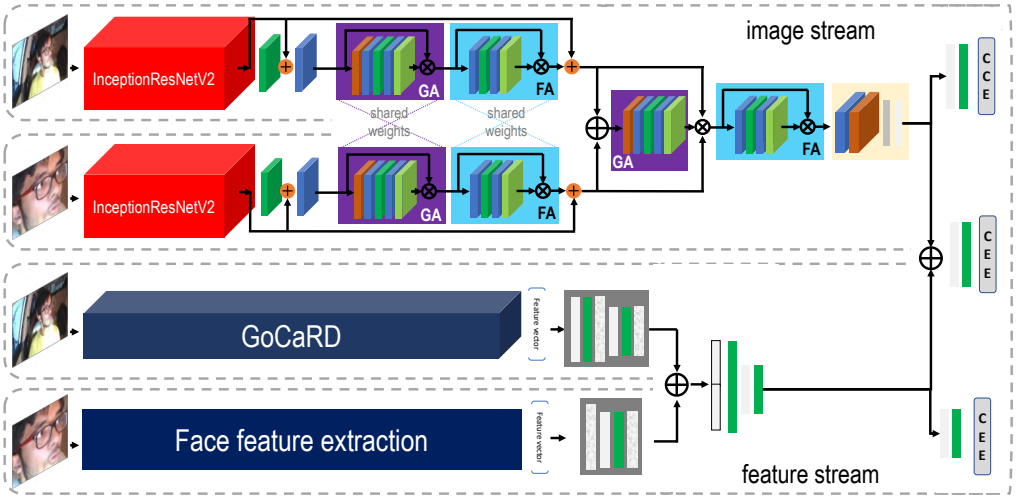

Figure 2: Most complex example configuration using the blocks introduced in Section 4. A full architecture has 4 main parts: one InceptionResNetV2 each for extracting features from the full image and the face image; the GoCaRD feature extracted on the full image; and some manually crafted face features. The first to InceptionResNetV2 outputs are fused e. g., using multiple grid (GA), and feature (FA) attention blocks, residual connections and auxiliary layers. In parallel, the extracted features are enhanced and combined in a simpler fashion. Both fused streams are specifically fine-tuned using (auxiliary) CCE (categorical cross entropy) losses while also combined for the final prediction output.

landmarks. We further extract the coordinates of all corners from the eyes to calculate the center, origin, area, height, width, and a ratio that can indicate whether an eye is closed or not. Given an open eye, the position of the pupil is estimated by detecting the iris. The pupil positions' usage is manifold; firstly, a pupil detection algorithm[4] based on the relative spatial estimation that the iris occupies in relation to the ocular surface is used to calibrate the distance between camera and subject. Secondly, in order to determine the direction a person is looking at, a vertical and horizontal ratio between 0 and 1 is calculated, so that a value of 0.0 reflects the top, 0.5 the middle, and 1.0 the bottom level. A binary value that represents left versus right oriented gaze is also derived by comparing this value with fixed thresholds. We set our horizontal threshold to $\beta = .35$ (<= right, >= $1 - \beta$ left). The final face feature vector has 164 dimensions.

3.4 GoCaRD features

Generic, optical Car Part Recognition and Detection [37] is a visual feature extractor designed specifically for the automotive environment to robustly predict vehicle regions regardless of make and model. It localizes 29 vehicle interior and exterior parts, e. g., ventilation outlets, armrest, steering wheel, roof window, and sun visor. In the challenge "Multimodal Sentiment Analysis in Real-Life Media" (MuSe 2020) [36], which aims at a more profound understanding of vehicle interaction, especially between human emotions (sentiment) and vehicle environment, GoCaRD was successfully applied on 'in the wild' recordings of emotional vehicle reviews (MuSe-CaR). The architecture is based on a Darknet-53 as the backbone [34] and jointly trained over two 'in the wild' datasets. On the datasets introduced with the framework, the underlying model achieves a mean average precision of 41.07 % on 1 124 video frames of recordings which include human interaction and challenging illuminations. Since the number of detected parts varies, we convert the prediction output into a feature vector of fixed size. We obtain a 350-dimensional sparse vector by focusing on the 10 objects with the highest confidence (one-hot-encoding) and their localisation coordinates (x, y, width, and height).

[4]https://github.com/antoinelame/GazeTracking

4 X-AWARE: DEEP, CONTEXT-AWARE GAZE PREDICTION

This section introduces the underlying deep learning architectures and related network blocks. Firstly, the building blocks enhancing the features from Section 3.3 and Section 3.4 are explained in Section 4.1. In Section 4.2, we introduce our chosen state-of-the-art architecture and a fine-tuning classification block. This is followed by a detailed description of X-AWARE individual components and several alternatives illustrated in Figure 3 and Figure 4.

An illustration of one tested architecture configuration is depicted in Figure 2. The **image processing component** is designed to analyse the full, raw image including the environment, as well as the cropped face image in an end-to-end manner. The **extracted feature part** in the lower part utilises the output of the facial (cf. Section 3.3) and the GoCaRD (cf. Section 3.4) feature extraction frameworks. Both parts are combined in a way such that the face and environment information interact via co-processing by the network. In order to create context aware representations, we experiment with several network block choices, while the X-AWARE attention blocks are the most sophisticated in terms of modelling.

4.1 Feature enhancement

Since the GoCaRD features are very sparse after extraction, they are compressed in a simple **compression** block. As shown in Figure 3

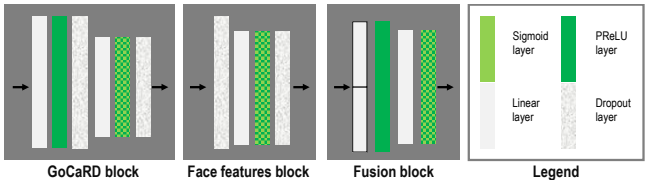

Figure 3: Architectural blocks to enhance, compress and fuse the GoCaRD and facial features.

the compression is made in two steps, each using a combination of linear, Parametric Rectified Linear Unit (PReLU), and dropout layers. The compression block for the face features, face features block, is similarly structured but only compresses one time. Depending on the experiment settings, both can have either a PReLU or sigmoid activation function before fusion.

In addition, two **naive fusion** blocks are used. The first one, called Interaction block, concatenates three representations squashed by a sigmoid function. The concatenated vector is fed into a 3 blocks, each consisting of a dropout, a linear, and a PReLU layer, reducing the neurons towards the end to force interaction. The second one called Fusion block is shown in Figure 3, which follows the same principle but only concatenates two feature sets, followed by a dropout and a linear layer.

4.2 State-of-the-art vision architectures

4.2.1 Core networks. As an initial image processing component, we opted to leverage a top performing computer vision architecture. After performing a preliminary comparative experiment (see Section 5.3) among a selection of architectures [7, 15, 35, 39, 40], we finally selected the InceptionResNetV2 [39] model to be our core vision component before any fusion operations are applied.

4.2.2 Enhancement head. In order to perform the preliminary experiments of Section 5.3, we require all competing architectures to be expressed by a function $y = F(x)$, where $F(x)$ denotes a non-linear function (the neural network), x an input image, and y the classes. The architectures described above, have a pooling operation (e. g., average, max) to reduce the dimensions of the convolution filters before applying the dense prediction layer. A common technique in transfer learning is to remove the top layer, and replace it with a new neural block 'head' that is to be fine-tuned to the class set specific to the new task [46]. Alternative pooling operations are often considered in these heads, that can potentially lead to increased performance. Towards the design of a reasonable baseline, we use such a custom head on top of the last convolution layer. The first layer of the head has 1024 filters, a kernel size of 3×3 and strides of 1, followed by a flattening operation, and the application of dropout with rate equal to 0.5. Finally, two dense layers (1024 and 256 neurons respectively) with sigmoid activation functions further compress the representation for the final prediction. This type of baseline is used in the preliminary experiments in Section 5.3. However, within the context of our full model, the best-performing InceptionResNetV2 model outputs a hidden feature vector, to be further processed by the more elaborate X-AWARE fusion components.

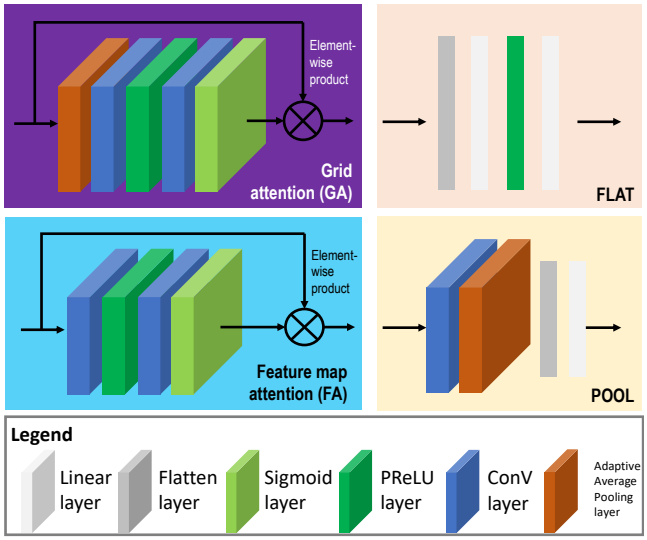

Figure 4: Architectural blocks of X-AWARE .

4.3 X-AWARE attention fusion

In the following, the individual blocks that comprise the **X-AWARE** context information fusion approach are described. These blocks are depicted in Figure 4. In order to facilitate gradient backpropagation, we also utilise residual connections between the input and the output of these blocks.

4.3.1 Grid attention block. Consider a convolutional representation F with D filter channels (depth slices, grids) and a symmetrical $W \times H$ grid, where the slices contain different learnt features. Towards extracting slice-specific importance values (and reduce the amount of parameters), we apply a 2D adaptive average pooling operation and receive a $D \times 1 \times 1$ feature map:

$$m_d = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_d(i, j), \tag{1}$$

where $X_d(i, j)$ represents a value of d-th slice at position (i,j). Next, we transform m_d to receive the attention weights of the slices:

$$\alpha_d = \sigma\left(\text{Conv}\left(\gamma\left(\text{Conv}\left(m_d\right)\right)\right)\right), \tag{2}$$

using convolution layers (Conv) in combination with an outer, bounding sigmoid (σ) and an inner, PReLU (γ) activation function. Similar to the Leaky ReLU, PReLU introduces leakage to allow a non-zero gradient flowing when inactive via a learnable parameter [14]. As usual, these weights α_d are applied to the input representation by element-wise multiplication resulting in F'. Inspired by the success of a squeeze-and-extend mechanism [17] and block attention [48], to enhance the calibration and representational power, especially when modelling interdependencies between depth slice relationships, the first convolution of our block squeezes the filter number by a factor of 16 before extending it again in the second convolutional operation to the original input size.

4.3.2 Feature attention block. Apart from filter-specific attention weights, it may be worth learning element-specific weights for each filter slice, as they might encode features of varying informative

value. With respect to the usual output of a convolutional layer F^* with a shape of $D \times H \times W$, we calculate an attention weight matrix with a shape equal to $1 \times H \times W$ by using two convolutional layers:

$$\alpha_{fe} = \sigma\left(\text{Conv}\left(\gamma\left(\text{Conv}\left(F^*\right)\right)\right)\right), \tag{3}$$

where σ denotes the sigmoid, and γ the PReLU activation functions, respectively. We then multiply F^* with the attention weights α_{fe} in an element-wise manner to yield our output $F^{*'}$ similiar to [32]. Similar to the grid attention block, we also utilise here the squeeze-and-extend mechanism.

4.3.3 Combined grid attention block. Multiple outputs F_m of the same depth size D can also be combined (e. g., F_1 and F_2 for $m \in \{1, 2\}$) as depicted in the later, cross-stream fusion part in Figure 2. Firstly, both are concatenated to a representation shape of $(m \cdot D) \times W \times H$, and passed through the grid attention block yielding the cross-context depth slices shaped $m \cdot D \times 1 \times 1$. The 2D pooling operations introduce some translational invariance to the filters, as well as spatial local correlation. Next, the resulting α_{m*d} feature vector is transformed back into multiple slices of shape $m \times D \times 1 \times 1$, and they are multiplied with the corresponding F_m in an element-wise manner. For improved representation, another feature attention block can be added after the grid fusion block (cf. Figure 2). This can be followed by either a Pool or Flat operation as illustrated in Figure 4. The former flattens the output of the previous 2D block into a 1D representation, and compresses it further by using two consecutive linear layers. The latter reduces the 2D representation first, using an adaptive average pooling layer.

4.4 (Auxiliary) Loss

The softmax categorical cross entropy loss (CCE) used, weights the loss with a precalulated w_c corresponding to the number of the class c occurring in the training set. It can be expressed by:

$$\mathcal{L}_g = w_c\left(-x_c + \log\left(\sum_j \exp(x_j)\right)\right), \tag{4}$$

given x_c is the model linear output for class c. The sample-specific losses are averaged in a weighted manner according to w across observations for each minibatch. Auxiliary losses are well known in computer vision for stabilising downstream layers. The total loss is equal to the sum of the weighted, averaged (auxiliary) losses across all observations for each minibach:

$$\mathcal{L}_{total} = \sum_{g=1}^{G} w_g * \mathcal{L}_g, \tag{5}$$

where G is the number of losses and w_g the scalar weight corresponding to the g-th loss.

5 EXPERIMENTS AND RESULTS

5.1 Experimental settings

The challenge permits participants to submit predictions up to five times for an evaluation on the test set, so in the following, we limit our reporting of results to the development set, and to the test set only for the later stage models we submitted. In addition to the official metric of the challenge – accuracy (ACC), we report the macro-averaged F1 score, as well as the Unweighted Average Recall (UAR) for our preliminary experiments, in order to provide a more comprehensive summary of the results, taking into account the class representation factor. Our experiments are performed using a Tesla V100 graphics processor with 32 GB GPU-RAM. We used Python for data preprocessing and external API integration, and Pytorch [28] for most implementations related to deep learning.

5.2 Baseline

The organisers report an official classification baseline of 56.00 % ACC on the validation set using an Inception V1 network. No further details are given regarding the input data, training procedure, and classification heads/tasks. In the accompanying paper describing the data collection, a set of results is reported on the challenge data partitions. This ACC baseline ranges from 56.25 % (Alexnet) to 61.46 % (Inception-V1 + Illumination Robust Layer) on the validation set. The addition of an attention layer to the latter architecture leads to a further gain of 3 % in ACC to 64.46 %, as well as an F1 score of 52.00, which is also reported as the final result of the proposed approach. On the test set, the ACC is slightly lower at 62.90 %, while the F1 score is identical to the validation set. When mixing training and validation set together, these scores improve to 59.00 and 64.31 % for F1 and ACC, respectively on test.

5.3 Preliminary experiments and results

In a series of preliminary experiments, we evaluate the performance of two static feature sets (facial and GoCaRD), find the most promising network architecture, and derive sensible hyperparmeters.

In order to quantify the effectiveness of the hand-crafted face features and GoCaRD representations, we train Support Vector Machines (SVM), often used for such a task with high-dimensional feature sets [3, 26, 49], using complexity values (C) between 10^{-5} and 1. The best SVM model resulted in an F1 score of 34.29 on the normalised (C=1, UAR: 33.72 %, ACC: 34.29 %), and 42.13 on the unnormalised features (C=10^{-5}, UAR: 43.50 %, ACC: 46.21 %). If *solely* the GoCaRD functions are used to predict the driver's gaze, the result is only slightly above the chance level in all settings, which is reasonable due to the lack of any facial information.

To obtain a first performance overview of the core architectures for image classification, we carry out a naive series of experiments where all networks are fully trainable. Gradient vanishing is a common issue when training very deep networks from scratch. We initialise the networks with the pretrained ImageNet weights for the respective networks in this preliminary stage. We utilise the latter technique to avoid performing hyperparameter optimisation on every architecture in this preliminary stage, and to achieve a reduction of training time. The models are trained using a batch size of 64 and an initial learning rate of 1^{-5}, reducing the learning rate every 2 epochs by a factor of 0.2 when a plateau is reached, for a maximum of 100 epochs (early stopping patience is set to 3).

As summarised in Table 1, models trained on the **cropped facial images** scored higher than those trained on the full ones. Here, InceptionV3 achieved the best results with **68.12 % ACC**, closely followed by InceptionResNetV2 . On the **full images**, the latter outperformed all other models by 2 %, achieving **62.79 % ACC**. VGG16 and VGG19 produce relatively competitive results using the cropped face image, but fail to learn using the full image input.

Table 1: Results of face gaze estimation using a) the full image including the environment and faces (full), and b) exclusively on the cropped faces (face) of several unfrozen computer vision architectures on the development set. We report F1, ACC, and UAR, as well as the number of trainable parameters in millions.

	full			faces			# param.
	F1	UAR [%]	ACC [%]	F1	UAR [%]	ACC [%]	trainable
InceptionResNetV2	**60.58**	**61.03**	**62.79**	66.42	67.14	68.09	72.3
InceptionV3	58.15	58.41	60.10	**67.31**	**68.31**	**68.12**	28.9
ResNet50	55.55	56.35	57.62	64.72	64.65	66.49	94.0
VGG16	2.86	11.11	14.76	66.17	65.90	68.08	71.0
VGG19	2.86	11.11	14.76	66.72	67.08	68.06	76.3
Xception	55.90	56.84	58.02	64.20	64.94	66.37	144.8

However, this could be amended if the training settings were to be specifically finetuned to each model, a process we omitted since, at this stage, we intended a simple comparative study among the different base components. Based on these initial findings, we selected InceptionResNetV2 as our primary architecture. Although such a weight initialisation may come with some drawbacks (details cf. Section 6.2), we believe that it is serviceable in this particular challenge setting: we decided to trade off a potential better final prediction performance for a shorter training durationwhich allowed for a faster experimental throughput, so that we could perform a broader assessment and focus on the fusion of human and environmental information streams.

5.4 Advanced and fusion models

The following models are optimised using a batch size of 32 and an initial learning rate of 0.0001. The results of our more advanced architectures are summarised in Table 2 for four different architecture settings, each with various pooling and input data options. Compared to the preliminary experiments, adding the proposed grid and attention mechanisms to the core architecture improved individual ACC performance for the full (68.00 %) and cropped face (70.81 %) image inputs. The naive concatenation of the full and face *images* and all four stream representations (*all*) combined with the enhancement, and the interaction blocks, decreases the performance compared to our single image baseline experiments in Table 1. The addition of X-AWARE blocks results in improvement in all settings by almost 5 % points compared to the naive fusion approach, and Pool outperforms Flat by at least one percent. By using only the full and face images, we achieve our **overall best result of 72.37 % ACC on devel (71.62 % on test) and an F1 score of 70.26.**

In the last setting, we try to improve the results on all inputs employing auxiliary loss(es) either on the combined image streams only (images) or on both (feature stream and image stream) as illustrated in Figure 2. Here, the Pool operation seems also slightly more effective than Flat . Adding an auxiliary layer after the image fusion improves the result by approximately one percentage point. With two auxiliary layers both after the image fusion, and after the feature fusion, our best performing model, utilising all data, yields an ACC of 71.43 %, and an F1 score of 69.88 on the development set (Pool).

Table 2: Face gaze estimation results comparing the blocks: grid (GA) and feature (FA) attention in sequence on a single input stream (either the full or face images), as well as the interaction, X-AWARE , and X-AWARE plus auxiliary losses (AUX.) for fusion. Latter architectures use either the full combined with face images (images) or these two in combination with the facial and GoCaRD features (all). All experiments utilising the auxiliary layers are conducted using all four inputs, while the auxiliary can either be taken from the fused alone (*all+images*) or from both streams (*all+both*) as depict in Figure 2. We report F1 and ACC on the development, and ACC on the test set.

Configuration		Metrics		
Fusion	Input	F1 (dev.)	ACC [%] (dev.)	ACC [%] (test)
		GA + FA		
–	full image	65.61	68.00	–
–	face image	69.15	70.81	–
		INTERACTION block		
concat	images	62.77	64.91	–
concat	all	63.72	66.28	–
		X-AWARE		
Flat	images	67.30	69.14	68.73
	all	67.84	69.09	70.22
Pool	images	70.26	**72.37**	**71.62**
	all	68.47	70.65	–
	AUX. Layer	**X-AWARE + AUX.**		
Flat	all+images	65.78	68.07	–
	all+both	69.51	71.07	71.15
Pool	all+images	67.14	69.32	–
	all+both	69.88	71.43	71.28

6 DISCUSSION AND FUTURE WORK

6.1 Quantitative results discussion

All our trained models considerably outperformed the reported baselines summarised in Section 5.2. Besides the experimental setup, we attribute this to several network-related factors: a) all trained networks utilise a larger number of parameters (cf. Section 5.3). The smallest network InceptionV3 has 5 times more parameters (the original has 23 millions, including our head this is raised to approximately 29 millions) than the Inception V1; b) our utilised head as well as the other attention heads allow improved "fine-tuning" of the model; and c) the weight initialisation leads to a stable training. Although this last technique can also be exploited for training on large datasets, it often comes at the disadvantage of a reduced maximum performance. Inversely, by training on large amounts of data with randomly initialised weights, more descriptive, problem-specific low-level filters can be learnt, and the risk of rapid overfitting avoided, albeit at increased training time.

In line with other research, the attention modules have led to improved results for models with only one image stream (more than 5% for full images) as well as for the fused image streams. Furthermore, we experimented with completely separated attention modules in the early stage of the image streams as well as with shared weights (cf. Figure 2), whereby the performance of the latter was better.

The experiments showed no clear evidence that the additional facial features (combined with the auxiliary losses) significantly enhances the overall performance. However, the class-specific accuracy seems to be slightly more balanced, especially with regard to the more difficult left side of the vehicle, which led to more confusion within these classes in the purely image-based approaches. We speculate that the head is turned more actively, but also the targeted points are further away, therefore requiring less movement of the pupils. Alternative explanations include the possibility of overfitting, or of learning confounding, biased dataset features.

Regarding the effectiveness of the GoCaRD features as environmental anchors, no clear influences on the performance could be observed. We suspect that the lack of suitable anchor points as prediction classes for the model we used resulted in extremely sparse representation, and prevented the full potential of this idea to be exploited. We expect it becomes more apparent when the cameras are statically mounted in different locations or not at all. We discuss this point more thoroughly in Section 6.3.

6.2 Future work

We can conceive further extensions and orthogonal improvements on our proposed framework regarding the data, training procedure, and modelling to be explored in order to improve performance and increase understanding. For example, we did not perform data augmentation (crop, horizontal or vertical flip, etc.), or denoising of the hazy images, something that might bolster the final prediction. As shown in other studies [23, 33], performance gains can be achieved by sequential modelling of consecutive frames. Therefore, metadata marking successive frames has the potential to significantly improve the model and also allows for studying a more realistic, 'in the wild' driving setting, in which changes in illumination might only be temporary for milliseconds during a drive.

Our training procedure also has room for improvements; for example, the learning rate scheduler was applied at the end of each epoch. Since the dataset is of considerable size, a step-wise reduction during the epoch (e. g., every x00 steps) might reduce overfitting significantly. Furthermore, freezing and unfreezing certain parts of the underlying core network (e. g., the low-level features) might be worth exploring, such that we utilise transferred knowledge from the pretrained weights.

We also plan to further explore the interaction between environment and face gaze using parts of the proposed X-AWARE fusion in different experimental settings. A straightforward approach would be to mask the face area out of the full pictures, and, as such, analyse an image solely containing environment information. Using purely modelling techniques, similar effects can be achieved combining cross-attention with a penalty loss pushing the model to focus on learning different high-level representations of the face and environment. In addition to X-AWARE , we performed preliminary experiments with the self-attention solution of [27] in which we augmented the output of the convolution in later layers before fusion. This approach also showed slight improvements compared to the attention-free block (68.73% ACC and 66.64 F1), however, due to the minimal improvement and time constraints, we decided to not pursue this approach. Nevertheless, we acknowledge the importance of a comparison with (purely) attention-based architectures to reduce the number of parameters, and plan to explore it further in the future.

6.3 Proposing further extensions of the gaze estimation in the wild task

Gaze estimation in the wild has so far almost exclusively been discussed in relation to the development of vehicle assistance systems. In this section, we want to point out the advantages of a deeper integration of the context into the field of gaze estimation. Besides assistance systems, gaze estimation is of immense interest to the emerging field of multimodal sentiment analysis in the wild [31, 51, 52], e. g., in the car video review setting [36], which studies the recognition of emotions and interaction in user-generated video content. Compared to previous work linking human-object interaction and gaze estimation [16], it has to deal with additional 'in the wild' factors, such as different camera equipment, changing of the camera position, and inconsistent distance to the subject. In addition, driver assistance systems would also potentially benefit from deeper research in this direction: classifiers that are trained on such datasets or explicitly model the environment would be more generic and robust, which is of particular interest in the automotive context, where camera systems and distances to the passengers depend strongly on makes and models. *A potential starting point towards this direction might be the use of ground-truth visual anchors.* For example, the upper anchorage of the safety belt might be a good anchor point to estimate the camera position in the driver cabin. These anchors can be extracted by localisation networks; for example, an extend framework of the GoCaRD extractor we used in this study.

7 CONCLUSION

In the presented work, we examined the gaze estimation in the wild task on the new Driver Gaze in the Wild dataset. Starting with a broad evaluation of state-of-the-art vision models, we developed a novel modelling approach to integrate environmental context of the face in the gaze estimation in the wild task. To do so, we developed several X-AWARE fusion components that we evaluated in our experiments. Our multi-stream fusion approach, combining the face with environment images, outperforms the baseline's accuracy by 15.03 % on the development set and achieves 71.62 % on the test set. Furthermore, we investigated the integration of manually derived facial features, such as the eye aspect ratio, as well as a vehicle part location extractor (GoCaRD) that might serve as potential anchor points to more easily calibrate and map the human in relation to the car cabin environment. Our holistic approach fusing all four inputs yields an improved 71.28 % accuracy. Based on these promising results, we also link the two 'in the wild' fields of multimodal sentiment analysis and gaze estimation, and propose to also consider in the future varying camera settings, e. g., positioning, and equipment. We expect that this extension would further improve generalisation; also for related applications, such as human-object interaction. In our work, we merely focused on examining the improvement brought by image input and context fusion – leaving room to improve the usage of data and training settings to further improve the performance to which we provide several suggestions.

REFERENCES

[1] Roland Goecke Abhinav Dhall, Garima Sharma and Tom Gedeon. 2020. EmotiW 2020: Driver Gaze, Group Emotion, Student Engagement and Physiological Signal based Challenges. In *ACM International Conference on Multimodal Interaction*. ACM.

[2] Christer Ahlstrom, Katja Kircher, and Albert Kircher. 2013. A gaze-based driver distraction warning system and its effect on visual behavior. *IEEE Transactions on Intelligent Transportation Systems* 14, 2 (2013), 965–973.

[3] Amer Al-Rahayfeh and Miad Faezipour. 2013. Enhanced eye gaze direction classification using a combination of face detection, CHT and SVM. In *2013 IEEE Signal Processing in Medicine and Biology Symposium*. IEEE, 1–6.

[4] Amer Al-Rahayfeh and Miad Faezipour. 2013. Eye tracking and head movement detection: A state-of-art survey. *Journal of Translational Engineering in Health and Medicine* 1 (2013), 2100212–2100212.

[5] Guido Borghi, Matteo Fabbri, Roberto Vezzani, Rita Cucchiara, et al. 2018. Face-from-depth for head pose estimation on depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

[6] Guido Borghi, Marco Venturelli, Roberto Vezzani, and Rita Cucchiara. 2017. Poseidon: Face-from-depth for driver pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4661–4670.

[7] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1251–1258.

[8] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. 2018. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision*. 334–352.

[9] Lex Fridman, Philipp Langhans, Joonbum Lee, and Bryan Reimer. 2016. Driver gaze region estimation without use of eye movement. *IEEE Intelligent Systems* 31, 3 (2016), 49–56.

[10] Wolfgang Fuhl, Marc Tonsen, Andreas Bulling, and Enkelejda Kasneci. 2016. Pupil detection for head-mounted eye tracking in the wild: an evaluation of the state of the art. *Machine Vision and Applications* 27, 8 (2016), 1275–1288.

[11] Shreya Ghosh, Abhinav Dhall, Garima Sharma, Sarthak Gupta, and Nicu Sebe. 2020. Speak2Label: Using Domain Knowledge for Creating a Large Scale Driver Gaze Zone Estimation Dataset. *arXiv preprint arXiv:2004.05973* (2020).

[12] Feng Guo, Sheila G Klauer, Youjia Fang, Jonathan M Hankey, Jonathan F Antin, Miguel A Perez, Suzanne E Lee, and Thomas A Dingus. 2017. The effects of age on crash risk associated with driver distraction. *International Journal of Epidemiology* 46, 1 (2017), 258–265.

[13] Dan Witzner Hansen and Arthur EC Pece. 2005. Eye tracking in the wild. *Computer Vision and Image Understanding* 98, 1 (2005), 155–181.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1026–1034.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 770–778.

[16] Chaoqun Hong, Jun Yu, Jian Zhang, Xiongnan Jin, and Kyong-Ho Lee. 2018. Multimodal face-pose estimation with multitask manifold deep learning. *IEEE Transactions on Industrial Informatics* 15, 7 (2018), 3952–3961.

[17] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7132–7141.

[18] Pawel Kasprowski and Katarzyna Harezlak. 2018. Comparison of mapping algorithms for implicit calibration using probable fixation targets. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. ACM, 1–8.

[19] Pawel Kasprowski, Katarzyna Harężlak, and Przemysław Skurowski. 2019. Implicit Calibration Using Probable Fixation Targets. *Sensors* 19, 1 (2019), 216.

[20] Rakshit Kothari, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B Pelz, and Gabriel J Diaz. 2020. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific Reports* 10, 1 (2020), 1–18.

[21] Matti Kutila, Maria Jokela, Gustav Markkula, and Maria Romera Rué. 2007. Driver distraction detection with a camera vision system. In *2007 IEEE International Conference on Image Processing*, Vol. 6. IEEE, VI–201.

[22] Yuan Liao, Shengbo Eben Li, Wenjun Wang, Ying Wang, Guofa Li, and Bo Cheng. 2016. Detection of driver cognitive distraction: A comparison study of stop-controlled intersection and speed-limited highway. *IEEE Transactions on Intelligent Transportation Systems* 17, 6 (2016), 1628–1637.

[23] Sankha S Mukherjee and Neil Martin Robertson. 2015. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia* 17, 11 (2015), 2094–2107.

[24] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. 2008. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* 31, 4 (2008), 607–626.

[25] World Health Organization. 2018. *Global status report on road safety 2018*. Geneva: World Health Organization.

[26] Kang Ryoung Park. 2004. Real-time gaze detection via neural network. In *International Conference on Neural Information Processing*. Springer, 673–678.

[27] Niki Parmar, Prajit Ramachandran, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. 2019. Stand-alone self-attention in vision models. In *Advances in Neural Information Processing Systems*. 68–80.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, 8024–8035.

[29] Massimiliano Patacchiola and Angelo Cangelosi. 2017. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition* 71 (2017), 132–143.

[30] Jochen Pohl, Wolfgang Birk, and Lena Westervall. 2007. A driver-distraction-based lane-keeping assistance system. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering* 221, 4 (2007), 541–552.

[31] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174 (2016), 50–59.

[32] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. 2020. FFA-Net: Feature Fusion Attention Network for Single Image Dehazing.. In *AAAI Conference on Artifical Intelligence*. AAAI, 11908–11915.

[33] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. 2017. Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision*. 1435–1443.

[34] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[35] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[36] Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Björn W Schuller, Iulia Lefter, et al. 2020. MuSe 2020–The First International Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop. *arXiv preprint arXiv:2004.14858* (2020).

[37] Lukas Stappen, Xinchen Du, Vincent Karas, Stefan Müller, and Björn W Schuller. 2020. Go-CaRD–Generic, Optical Car Part Recognition and Detection: Collection, Insights, and Applications. *arXiv preprint arXiv:2006.08521* (2020).

[38] Lukas Stappen, Vincent Karas, Nicholas Cummins, Fabien Ringeval, Klaus Scherer, and Björn Schuller. 2019. From speech to facial activity: towards cross-modal sequence-to-sequence attention networks. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing*. IEEE, 1–6.

[39] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artifical Intelligence*. AAAI.

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.

[41] Fabio Tango and Marco Botta. 2013. Real-time detection system of driver distraction using machine learning. *IEEE Transactions on Intelligent Transportation Systems* 14, 2 (2013), 894–905.

[42] Francisco Vicente, Zehua Huang, Xuehan Xiong, Fernando De la Torre, Wende Zhang, and Dan Levi. 2015. Driver gaze tracking and eyes off the road detection system. *IEEE Transactions on Intelligent Transportation Systems* 16, 4 (2015), 2014–2027.

[43] Hans-Jörg Vögel, Christian Süß, Thomas Hubregtsen, Elisabeth André, Björn Schuller, Jérôme Härri, Jörg Conradt, Asaf Adi, Alexander Zadorojniy, Jacques Terken, et al. 2018. Emotion-awareness for intelligent vehicle assistants: a research agenda. In *2018 1st International Workshop on Software Engineering for AI in Autonomous Systems*. IEEE, ACM, 11–15.

[44] Sourabh Vora, Akshay Rangesh, and Mohan Manubhai Trivedi. 2018. Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis. *IEEE Transactions on Intelligent Vehicles* 3, 3 (2018), 254–265.

[45] Kang Wang and Qiang Ji. 2018. 3D gaze estimation without explicit personal calibration. *Pattern Recognition* 79 (2018), 216–227.

[46] Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing* 312 (2018), 135–153.

[47] Yafei Wang, Guoliang Yuan, Zetian Mi, Jinjia Peng, Xueyan Ding, Zheng Liang, and Xianping Fu. 2019. Continuous driver's gaze zone estimation using rgb-d camera. *Sensors* 19, 6 (2019), 1287.

[48] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.

[49] Yi-Leh Wu, Chun-Tsai Yeh, Wei-Chih Hung, and Cheng-Yuan Tang. 2014. Gaze direction estimation using support vector machine with active appearance model. *Multimedia Tools and Applications* 70, 3 (2014), 2037–2062.

[50] Bingjie Xu, Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. 2019. Interact as you intend: Intention-driven human-object interaction detection. *IEEE Transactions on Multimedia* 22, 6 (2019), 1423–1432.

[51] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* (2017).

[52] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31, 6 (2016), 82–88.

[53] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4511–4520.