Deep Pictorial Gaze Estimation

Seonwook Park, Adrian Spurr, and Otmar Hilliges

AIT Lab, Department of Computer Science, ETH Zurich
{firstname.lastname}@inf.ethz.ch

Abstract. Estimating human gaze from natural eye images only is a challenging task. Gaze direction can be defined by the pupil- and the eyeball center where the latter is unobservable in 2D images. Hence, achieving highly accurate gaze estimates is an ill-posed problem. In this paper, we introduce a novel deep neural network architecture specifically designed for the task of gaze estimation from single eye input. Instead of directly regressing two angles for the pitch and yaw of the eyeball, we regress to an intermediate pictorial representation which in turn simplifies the task of 3D gaze direction estimation. Our quantitative and qualitative results show that our approach achieves higher accuracies than the state-of-the-art and is robust to variation in gaze, head pose and image quality.

Keywords: Appearance-based Gaze Estimation, Eye Tracking

1   Introduction

Accurately estimating human gaze direction has many applications in assistive technologies for users with motor disabilities [4], gaze-based human-computer interaction [20], visual attention analysis [17], consumer behavior research [36], AR, VR and more. Traditionally this has been done via specialized hardware, shining infrared illumination into the user's eyes and via specialized cameras, sometimes requiring use of a headrest. Recently deep learning based approaches have made first steps towards fully unconstrained gaze estimation under free head motion, in environments with uncontrolled illumination conditions, and using only a single commodity (and potentially low quality) camera. However, this remains a challenging task due to inter-subject variance in eye appearance, self-occlusions, and head pose and rotation variations. In consequence, current approaches attain accuracies in the order of 6° only and are still far from the requirements of many application scenarios.

While demonstrating the feasibility of purely image based gaze estimation and introducing large datasets, these learning-based approaches [14, 45, 46] have leveraged convolutional neural network (CNN) architectures, originally designed for the task of image classification, with minor modifications. For example, [45, 47] simply append head pose orientation to the first fully connected layer of either LeNet-5 or VGG-16, while [14] proposes to merge multiple input modalities by
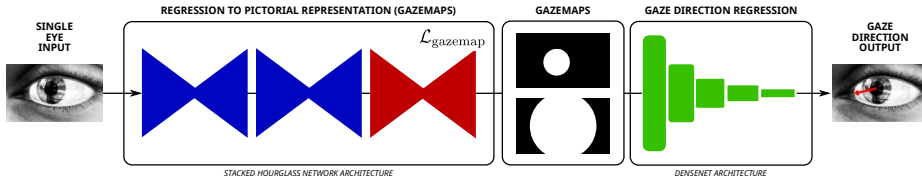
Fig. 1. Our sequential neural network architecture first estimates a novel pictorial representation of 3D gaze direction, then performs gaze estimation from the minimal image representation to yield improved performance on MPIIGaze, Columbia and EYEDIAP.

replicating convolutional layers from AlexNet. In [46] the AlexNet architecture is modified to learn so-called spatial-weights to emphasize important activations by region when full face images are provided as input. Typically, the proposed architectures are only supervised via a mean-squared error loss on the gaze direction output, represented as either a 3-dimensional unit vector or pitch and yaw angles in radians.

In this work we propose a network architecture that has been specifically designed with the task of gaze estimation in mind. An important insight is that regressing first to an abstract but gaze specific representation helps the network to more accurately predict the final output of 3D gaze direction. Furthermore, introducing this gaze representation also allows for intermediate supervision which we experimentally show to further improve accuracy. Our work is loosely inspired by recent progress in the field of human pose estimation. Here, earlier work directly regressed joint coordinates [34]. More recently the need for a more task specific form of supervision has led to the use of confidence maps or heatmaps, where the position of a joint is depicted as a 2-dimensional Gaussian [21, 33, 37]. This representation allows for a simpler mapping between input image and joint position, allows for intermediate supervision, and hence for deeper networks. However, applying this concept of heatmaps to regularize training is not directly applicable to the case of gaze estimation since the crucial eyeball center is not observable in 2D image data. We propose a conceptually similar representation for gaze estimation, called *gazemaps*. Such a gazemap is an abstract, pictorial representation of the eyeball, the iris and the pupil at it's center (see Figure 1).

The simplest depiction of an eyeball's rotation can be made via a circle and an ellipse, the former representing the eyeball, and the latter the iris. The gaze direction is then defined by the vector connecting the larger circle's center and the ellipse. Thus 3D gaze direction can be (pictorially) represented in the form of an image, where a spherical eyeball and circular iris are projected onto the image plane, resulting in a circle and ellipse. Hence, changes in gaze direction result in changes in ellipse positioning (cf. Figure 2a). This pictorial representation can be easily generated from existing training data, given known gaze direction annotations. At inference time recovering gaze direction from such a pictorial representation is a much simpler task than regressing directly from raw pixel values. However, adapting the input image to fit our pictorial representa-

tion is non-trivial. For a given eye image, a circular eyeball and an ellipse must be fitted, then centered and rescaled to be in the expected shape. We experimentally observed that this task can be performed well using a fully convolutional architecture. Furthermore, we show that our approach outperforms prior work on the final task of gaze estimation significantly.

Our main contribution consists of a novel architecture for appearance-based gaze estimation. At the core of the proposed architecture lies the pictorial representation of 3D gaze direction to which the network fits the raw input images and from which additional convolutional layers estimate the final gaze direction. In addition, we perform: (a) an in-depth analysis of the effect of intermediate supervision using our pictorial representation, (b) quantitative evaluation and comparison against state-of-the-art gaze estimation methods on three challenging datasets (MPIIGaze, EYEDIAP, Columbia) in the person independent setting, and a (c) detailed evaluation of the robustness of a model trained using our architecture in terms of gaze direction and head pose as well as image quality. Finally, we show that our method reduces gaze error by 18% compared to the state-of-the-art [47] on MPIIGaze.

2   Related Work

Here we briefly review the most important work in eye gaze estimation and review work touching on relevant aspects in terms of network architecture from adjacent areas such as image classification and human pose estimation.

2.1   Appearance-based Gaze Estimation with CNNs

Traditional approaches to image-based gaze estimation are typically categorized as *feature-based* or *model-based*. Feature-based approaches reduce an eye image down to a set of features based on hand-crafted rules [11, 12, 25, 41] and then feed these features into simple, often linear machine learning models to regress the final gaze estimate. Model-based methods instead attempt to fit a known 3D model to the eye image [30, 35, 39, 42] by minimizing a suitable energy.

Appearance-based methods learn a direct mapping from raw eye images to gaze direction. Learning this direct mapping can be very challenging due to changes in illumination, (partial) occlusions, head motion and eye decorations. Due to these challenges, appearance-based gaze estimation methods required the introduction of large, diverse training datasets and typically leverage some form of convolutional neural network architecture.

Early works in appearance-based methods were restricted to laboratory settings with fixed head pose [1, 32]. These initial constraints have become progressively relaxed, notably by the introduction of new datasets collected in everyday settings [14, 45] or in simulated environments [29, 38, 40]. The increasing scale and complexity of training data has given rise to a wide variety of learning-based methods including variations of linear regression [7, 18, 19], random forests [29], k-nearest neighbours [29, 40], and CNNs [14, 26, 38, 45–47]. CNNs have proven

to be more robust to visual appearance variations, and are capable of person-independent gaze estimation when provided with sufficient scale and diversity of training data. Person-independent gaze estimation can be performed without a user calibration step, and can directly be applied to areas such as visual attention analysis on unmodified devices [22], interaction on public displays [48], and identification of gaze targets [44], albeit at the cost of increased need for training data and computational cost.

Several CNN architectures have been proposed for person-independent gaze estimation in unconstrained settings, mostly differing in terms of possible input data modalities. Zhang *et al.* [45, 46] adapt the LeNet-5 and VGG-16 architectures such that head pose angles (pitch and yaw) are concatenated to the first fully-connected layers. Despite its simplicity this approach yields the current best gaze estimation error of $5.5°$ when evaluating for the within-dataset cross-person case on MPIIGaze with single eye image and head pose input. In [14] separate convolutional streams are used for left/right eye images, a face image, and a 25×25 grid indicating the location and scale of the detected face in the image frame. Their experiments demonstrate that this approach yields improvements compared to [45]. In [46] a single face image is used as input and so-called spatial-weights are learned. These emphasize important features based on the input image, yielding considerable improvements in gaze estimation accuracy.

We introduce a novel pictorial representation of eye gaze and incorporate this into a deep neural network architecture via intermediate supervision. To the best of our knowledge we are the first to apply fully convolutional architecture to the task of appearance-based gaze estimation. We show that together these contribution lead to a significant performance improvement of 18% even when using a single eye image as sole input.

2.2   Deep Learning with Auxiliary Supervision

It has been shown [16, 31] that by applying a loss function on intermediate outputs of a network, better performance can be yielded in different tasks. This technique was introduced to address the vanishing gradients problem during the training of deeper networks. In addition, such intermediate supervision allows for the network to quickly learn an estimate for the final output then learn to refine the predicted features - simplifying the mappings which need to be learned at every layer. Subsequent works have adopted intermediate supervision [21, 37] to good effect for human pose estimation, by replicating the final output loss.

Another technique for improving neural network performance is the use of auxiliary data through multi-task learning. In [24, 49], the architectures are formed of a single shared convolutional stream which is split into separate fully-connected layers or regression functions for the auxiliary tasks of gender classification, face visibility, and head pose. Both works show marked improvements to state-of-the-art results in facial landmarks localization. In these approaches through the introduction of multiple learning objectives, an implicit prior is forced upon the network to learn a representation that is informative to both

tasks. On the contrary, we *explicitly* introduce a gaze-specific prior into the network architecture via gazemaps.

Most similar to our contribution is the work in [9] where facial landmark localization performance is improved by applying an auxiliary emotion classification loss. A key aspect to note is that their network is sequential, that is, the emotion recognition network takes only facial landmarks as input. The detected facial landmarks thus act as a manually defined representation for emotion classification, and creates a bottleneck in the full data flow. It is shown experimentally that applying such an auxiliary loss (for a different task) yields improvement over state-of-the-art results on the AFLW dataset. In our work, we learn to regress an intermediate and minimal representation for gaze direction, forming a bottleneck before the main task of regressing two angle values. Thus, an important distinction to [9] is that while we employ an auxiliary loss term, it directly contributes to the task of gaze direction estimation. Furthermore, the auxiliary loss is applied as an intermediate task. We detail this further in Sec. 3.1.

Recent work in multi-person human pose estimation [3] learns to estimate joint location heatmaps alongside so-called "part affinity fields". When combined, the two outputs then enable the detection of multiple peoples' joints with reduced ambiguity in terms of which person a joint belongs to. In addition, at the end of every image scale, the architecture concatenates feature maps from each separate stream such that information can flow between the "part confidence" and "part affinity" maps. Thus, they operate on the image representation space, taking advantage of the strengths of convolutional neural networks. Our work is similar in spirit in that it introduces a novel image-based representation.

3   Method

A key contribution of our work is a pictorial representation of 3D gaze direction - which we call *gazemaps*. This representation is formed of two boolean maps, which can be regressed by a fully convolutional neural network. In this section, we describe our representation (Sec. 3.1) then explain how we constructed our architecture to use the representation as reference for intermediate supervision during training of the network (Sec. 3.2).

3.1   Pictorial Representation of 3D Gaze

In the task of appearance-based gaze estimation, an input eye image is processed to yield gaze direction in 3D. This direction is often represented as a 3-element unit vector v [6, 26, 46], or as two angles representing eyeball pitch and yaw $g = (\theta, \phi)$ [29, 38, 45, 47]. In this section, we propose an alternative to previous direct mappings to v or g.

If we state the input eye images as x and regard regressing the values g, a conventional gaze estimation model estimates $f : x \rightarrow g$. The mapping f can be complex, as reflected by the improvement in accuracies that have been

(a)

(b) Example gazemaps from UnityEyes
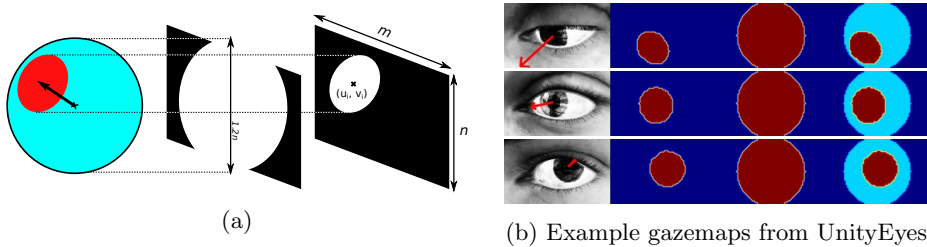
Fig. 2. Our pictorial representation of 3D gaze direction, essentially a projection of simple eyeball and iris models onto binary maps (a). Example-pairs are shown in (b) with (left-to-right) input image, iris map, eyeball map, and a superimposed visualization.

attained by simple adoption of newer CNN architectures ranging from LeNet-5 [26, 45], AlexNet [14, 46], to VGG-16 [47], the current state-of-the-art CNN architecture for appearance-based gaze estimation. We hypothesize that it is possible to learn an intermediate image representation of the eye, m. That is, we define our model as $g = k \circ j(x)$ where $j : x \to m$ and $k : m \to g$. It is conceivable that the complexity of learning j and k should be significantly lower than directly learning f, allowing for neural network architectures with significantly lower model complexity to be applied to the same task of gaze estimation with higher or equivalent performance.

Thus, we propose to estimate so-called *gazemap*s (m) and from that the 3D gaze direction (g). We reformulate the task of gaze estimation into two concrete tasks: (a) reduction of input image to minimal normalized form (gazemaps), and (b) gaze estimation from gazemaps.

The gazemaps for a given input eye image should be visually similar to the input yet distill only the necessary information for gaze estimation to ensure that the mapping $k : m \to g$ is simple. To do this, we consider that an average human eyeball has a diameter of $\approx 24mm$ [2] while an average human iris has a diameter of $\approx 12mm$ [5]. We then assume a simple model of the human eyeball and iris, where the eyeball is a perfect sphere, and the iris is a perfect circle. For an output image dimension of $m \times n$, we assume the projected eyeball diameter $2r = 1.2n$ and calculate the iris centre coordinates (u_i, v_i) to be:

$$u_i = \frac{m}{2} - r' \sin \phi \cos \theta \tag{1}$$

$$v_i = \frac{n}{2} - r' \sin \theta \tag{2}$$

where $r' = r \cos \left(\sin^{-1} \frac{1}{2} \right)$, and gaze direction $g = (\theta, \phi)$. The iris is drawn as an ellipse with major-axis diameter of r and minor-axis diameter of $r \left| \cos \theta \cos \phi \right|$. Examples of our gazemaps are shown in Fig. 2b where two separate boolean maps are produced for one gaze direction g.

Learning how to predict gazemaps only from a single eye image is not a trivial task. Not only do extraneous factors such as image artifacts and partial occlusion need to be accounted for, a simplified eyeball must be fit to the given image

based on iris and eyelid appearance. The detected regions must then be scaled and centered to produce the gazemaps. Thus the mapping $j : \boldsymbol{x} \rightarrow \boldsymbol{m}$ requires a more complex neural network architecture than the mapping $k : \boldsymbol{m} \rightarrow \boldsymbol{g}$.

3.2   Neural Network Architecture

Our neural network consists of two parts: (a) regression from eye image to gazemap, and (b) regression from gazemap to gaze direction $\boldsymbol{g}$. While any CNN architecture can be implemented for (b), regressing (a) requires a fully convolutional architecture such as those used in human pose estimation. We adapt the stacked hourglass architecture from Newell *et al.* [21] for this task. The hourglass architecture has been proven to be effective in tasks such as human pose estimation and facial landmarks detection [43] where complex spatial relations need to be modeled at various scales to estimate the location of occluded joints or key points. The architecture performs repeated multi-scale refinement of feature maps, from which desired output confidence maps can be extracted via 1×1 convolution layers. We exploit this fact to have our network predict gazemaps instead of classical confidence or heatmaps for joint positions. In Sec. 5, we demonstrate that this works well in practice.

In our gazemap-regression network, we use 3 hourglass modules with intermediate supervision applied on the gazemap outputs of the last module only. The minimized intermediate loss is:

$$\mathcal{L}_{\text{gazemap}} = -\alpha \sum_{p \in \mathcal{P}} \boldsymbol{m}(p) \log \hat{\boldsymbol{m}}(p), \tag{3}$$

where we calculate a cross-entropy between predicted $\hat{\boldsymbol{m}}$ and ground-truth gazemap $\boldsymbol{m}$ for pixels p in set of all pixels $\mathcal{P}$. In our evaluations, we set the weight coefficient α to 10^{-5}.

For the regression to $\boldsymbol{g}$, we select DenseNet which has recently been shown to perform well on image classification tasks [10] while using fewer parameters compared to previous architectures such as ResNet [8]. The loss term for gaze direction regression (per input) is:

$$\mathcal{L}_{\text{gaze}} = ||\boldsymbol{g} - \hat{\boldsymbol{g}}||_2^2, \tag{4}$$

where $\tilde{\boldsymbol{g}}$ is the gaze direction predicted by our neural network.

4   Implementation

In this section, we describe the fully convolutional (Hourglass) and regressive (DenseNet) parts of our architecture in more detail.

4.1   Hourglass Network

In our implementation of the Stacked Hourglass Network [21], we provide images of size 150×90 as input, and refine 64 feature maps of size 75×45 throughout the

network. The half-scale feature maps are produced by an initial convolutional layer with filter size 7 and stride 2 as done in the original paper [21]. This is followed by batch normalization, ReLU activation, and two residual modules before being passed as input to the first hourglass module.

There exist 3 hourglass modules in our architecture, as visualized in Figure 1. In human pose estimation, the commonly used outputs are 2-dimensional confidence maps, which are pixel-aligned to the input image. Our task differs, and thus we do not apply intermediate supervision to the output of every hourglass module. This is to allow for the input image to be processed at multiple scales over many layers, with the necessary features becoming aligned to the final output gazemap representation. Instead, we apply 1×1 convolutions to the output of the last hourglass module, and apply the gazemap loss term (Eq. 3).
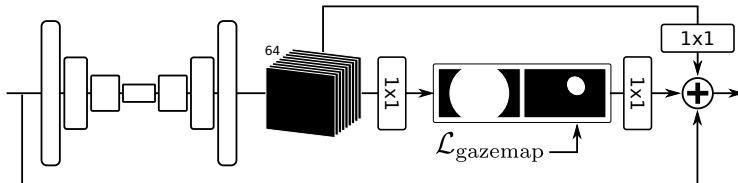
Fig. 3. Intermediate supervision is applied to the output of an hourglass module by performing 1×1 convolutions. The intermediate gazemaps and feature maps from the previous hourglass module are then concatenated back into the network to be passed onto the next hourglass module as is done in the original Hourglass paper [21].

4.2   DenseNet

As described in Section 3.1, our pictorial representation allows for a simpler function to be learnt for the actual task of gaze estimation. To demonstrate this, we employ a very lightweight DenseNet architecture [10]. Our gaze regression network consists of 5 dense blocks (5 layers per block) with a growth-rate of 8, bottleneck layers, and a compression factor of 0.5. This results in just 62 feature maps at the end of the DenseNet, and subsequently 62 features through global average pooling. Finally, a single linear layer maps these features to g. The resulting network is light-weight and consists of just 66k trainable parameters.

4.3   Training Details

We train our neural network with a batch size of 32, learning rate of 0.0002 and L_2 weights regularization coefficient of 10^{-4}. The optimization method used is Adam [13]. Training occurs for 20 epochs on a desktop PC with an Intel Core i7 CPU and Nvidia Titan Xp GPU, taking just over 2 hours for one fold (out of 15) of a leave-one-person-out evaluation on the MPIIGaze dataset.
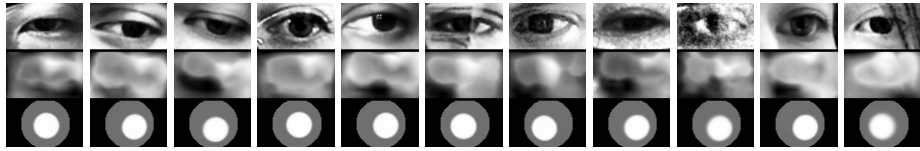
(a) Intermediate representations of training samples without (middle) and with (bottom) intermediate supervision
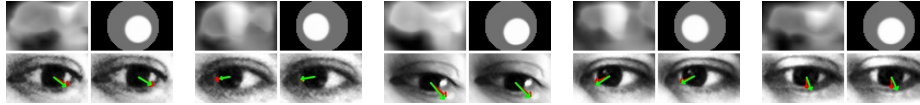
(b) Intermediate representations and predictions from test samples without (left) and with (right) intermediate supervision

Fig. 4. Example of image representations learned by our architecture in the absence or presence of $\mathcal{L}_{\text{gazemap}}$. Note that the pictorial representation is more consistent, and that the hourglass network is able to account for occlusions. Predicted gaze directions are shown in green, with ground-truth in red.

During training, slight data augmentation is applied in terms of image translation and scaling, and learning rate is multiplied by 0.1 after every 5k gradient update steps, to address over-fitting and to stabilize the final error.

5   Evaluations

We perform our evaluations primarily on the MPIIGaze dataset, which consists of images taken of 15 laptop users in everyday settings. The dataset has been used as the standard benchmark dataset for unconstrained appearance-based gaze estimation in recent years [26, 38, 40, 45–47]. Our focus is on cross-person single-eye evaluations where 15 models are trained per configuration or architecture in a leave-one-person-out fashion. That is, a neural network is trained on 14 peoples' data (1500 entries each from left and right eyes), then tested on the test set of the left-out person (1000 entries). The mean over 15 such evaluations is used as the final error metric representing cross-person performance. As MPIIGaze is a dataset which well represents real-world settings, cross-person evaluations on the dataset is indicative of the real-world person-independence of a given model.

To further test the generalization capabilities of our method, we also perform evaluations on two additional datasets in this section: Columbia [28] and EYE-DIAP [7], where we perform 5-fold cross validation. While Columbia displays large diversity between its 55 participants, the images are of high quality, having been taken using a DSLR. EYEDIAP on the other hand suffers from the low resolution of the VGA camera used, as well as large distance between camera and participant. We select screen target (CS/DS) and static head pose sequences (S) from the EYEDIAP dataset, sampling every 15 seconds from its VGA video streams (V). Training on moving head sequences (M) with just single eye input proved infeasible, with all models experiencing diverging test error during train-

ing. Performance improvements on MPIIGaze, Columbia, and EYEDIAP would indicate that our model is robust to cross-person appearance variations and the challenges caused by low eye image resolution and quality.

In this section, we first evaluate the effect of our gazemap loss (Sec. 5.1), then compare the performance (Sec. 5.2) and robustness (Sec. 5.3) of our approach against state-of-the-art architectures.

5.1 Pictorial Representation (*Gazemaps*)

We postulated in Sec. 3.1 that by providing a pictorial representation of 3D gaze direction that is visually similar to the input image, we could achieve improvements in appearance-based gaze estimation. In our experiments we find that applying the gazemaps loss term generally offers performance improvements compared to the case where the loss term is not applied. This improvement is particularly emphasized when DenseNet growth rate is high (eg. $k = 32$), as shown in Table 1.

Table 1. Cross-person gaze estimation errors in the absence and presence of $\mathcal{L}_{gazemap}$, with DenseNet (k=32).

Dataset	$\mathcal{L}_{\text{gazemap}}$	
	no	yes
MPIIGaze	4.67	**4.56**
Columbia	3.78	**3.59**
EYEDIAP	11.28	**10.63**

By observing the output of the last hourglass module and comparing against the input images (Figure 4), we can confirm that even without intermediate supervision, our network learns to isolate the iris region, yielding a similar image representation of gaze direction across participants. Note that this representation is learned only with the final gaze direction loss, $\mathcal{L}_{\text{gaze}}$, and that blobs representing iris locations are not necessarily aligned with actual iris locations on the input images. Without intermediate supervision, the learned minimal image representation may incorporate visual factors such as occlusion due to hair and eyeglasses, as shown in Figure 4a.

This supports our hypothesis that an intermediate representation consisting of an iris and eyeball contains the required information to regress gaze direction. However, due to the nature of learning, the network may also learn irrelevant details such as the edges of the glasses. Yet, by explicitly providing an intermediate representation in the form of gazemaps, we enforce a prior that helps the network learn the desired representation, without incorporating the previously mentioned unhelpful details.

5.2 Cross-Person Gaze Estimation

We compare the cross-person performance of our model by conducting a leave-one-person-out evaluation on MPIIGaze and 5-fold evaluations on Columbia and EYEDIAP. In Section 3.1 we discussed that the mapping k from gazemap to gaze direction should not require a complex architecture to model. Thus, our DenseNet is configured with a low growth rate ($k = 8$). To allow fair comparison, we re-implement 2 architectures for single-eye image inputs (of size 150×90): AlexNet and VGG-16. The AlexNet and VGG-16 architectures have been used in

Table 2. Mean gaze estimation error in degrees for within-dataset cross-person k-fold evaluation. Evaluated on (a) MPIIGaze, (b) Columbia, and (c) EYEDIAP datasets.

(a) MPIIGaze (15-fold)

Model	kNN [47]	RF [47]	[45]	AlexNet	VGG-16	GazeNet [47]	**ours**
# params	0	-	1.8M	86M	158M	90M	0.7M
Inputs	e + h	e + h	e + h	e	e	e + h	e
Error	7.2	6.7	6.3	5.7	5.4	5.5	**4.5**

(b) Columbia (5-fold)

Model	AlexNet	VGG-16	**ours**
Error	4.2	3.9	**3.8**

(c) EYEDIAP (5-fold)

| Model	AlexNet	VGG-16	**ours**
Error	11.5	11.2	**10.3**

where e: single-eye, h: head pose (pitch, yaw)

recent works in appearance-based gaze estimation and are thus suitable baselines [46, 47]. Implementation and training procedure details of these architectures are provided in supplementary materials.

In MPIIGaze evaluations (Table 2a), our proposed approach outperforms the current state-of-the-art approach by a large margin, yielding an improvement of $1.0°$ ($5.5° \rightarrow 4.5° = 18.2\%$). This significant improvement is in spite of the reduced number of trainable parameters used in our architecture (90M vs 0.7M). Our performance compares favorably to that reported in [46] ($4.8°$) where full-face input is used in contrast to our single-eye input. While our results cannot directly be compared with those of [46] due to the different definition of gaze direction (face-centred as opposed to eye centred), the similar performance suggests that eye images may be sufficient as input to the task of gaze direction estimation. Our approach attains comparable performance to models taking face input, and uses considerably less parameters than recently introduced architectures (129x less than GazeNet).

We additionally evaluate our model on the Columbia Gaze and EYEDIAP datasets in Table 2b and Table 2c respectively. While high image quality results in all three methods performing comparably for Columbia Gaze, our approach still prevails with an improvement of $0.4°$ over AlexNet. On EYEDIAP, the mean error is very high due to the low resolution and low quality input. Note that there is no head pose estimation performed, with only single eye input being relied on for gaze estimation. Our gazemap-based architecture shows its strengths in this case, performing $0.9°$ better than VGG-16 - a 8% improvement. Sample gazemap and gaze direction predictions are shown in Figure 5 where it is evident that despite the lack of visual detail, it is possible to fit gazemaps to yield improved gaze estimation error.

By evaluating our architecture on 3 different datasets with different properties in the cross-person setting, we can conclude that our approach provides

(a) Columbia
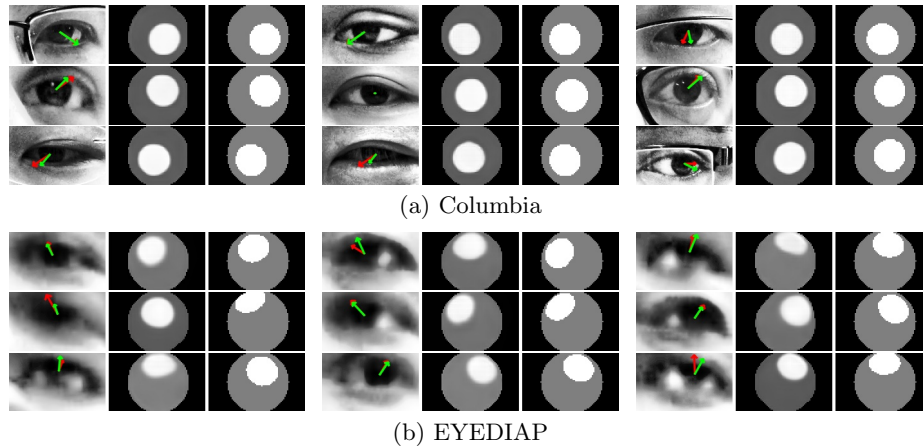
(b) EYEDIAP

Fig. 5. Gazemap predictions (middle) on Columbia and EYEDIAP datasets with ground-truth (red) and predicted (green) gaze directions visualized on input eye images (left). Ground-truth gazemaps are shown on the far-right of each triplet.

significantly higher generalization capabilities compared to previous approaches. Thus, we bring gaze estimation closer to direct real-world applications.

5.3   Robustness Analysis

In order to shed more light onto our models' performance, we perform an additional robustness analysis. More concretely, we aim to analyze how our approach performs under difficult and challenging situations, such as extreme head pose and gaze direction. In order to do so, we evaluate a moving average on the output of our within-MPIIGaze evaluations, where the y-values correspond to the mean angular error and the x-values take *one* of the following factor of variations: head pose (pitch & yaw), gaze direction (pitch & yaw). Additionally, we also consider image quality (contrast & sharpness) as a qualitative factor. In order to isolate each factor of variation from the rest, we evaluate the moving average only on the points whose remaining factors are close to its median value. Intuitively, this corresponds to data points where the person moves only in one specific direction, while staying at rest in all of the remaining directions. This is not the case for image quality analysis, where all data points are used. Figure 6 plots the mean angular error as a function of different movement variations and image qualities. The top row corresponds to variation along the head pose, the middle along gaze direction and the bottom to varying image quality. In order to calculate the image contrast, we used the RMS contrast metric whereas to compute the sharpness, we employ a Laplacian-based formula as outlined in [23]. Both metrics are explained in supplementary materials. The figure shows that we *consistently* outperform competing architectures for extreme head and gaze angles. Notably, we show more consistent performance in particular over large

ranges of head pitch and gaze yaw angles. In addition, we surpass prior works on images of varying quality, as shown in Figures 6e and 6f.

6   Conclusion

Our work is a first attempt at proposing an explicit prior designed for the task of gaze estimation with a neural network architecture. We do so by introducing a novel pictorial representation which we call gazemaps. An accompanying architecture and training scheme using intermediate supervision naturally arises as a consequence, with a fully convolutional architecture being employed for the first time for appearance-based eye gaze estimation. Our gazemaps are anatomically inspired, and are experimentally shown to outperform approaches which consist of significantly more model parameters and at times, more input modalities. We report improvements of up to 18% on MPIIGaze along with improvements on additional two different datasets against competitive baselines. In addition, we demonstrate that our final model is more robust to various factors such as extreme head poses and gaze directions, as well as poor image quality compared to prior work.

Future work can look into alternative pictorial representations for gaze estimation, and an alternative architecture for gazemap prediction. Additionally, there is potential in using synthesized gaze directions (and corresponding gazemaps) for unsupervised training of the gaze regression function, to further improve performance.

Acknowledgements
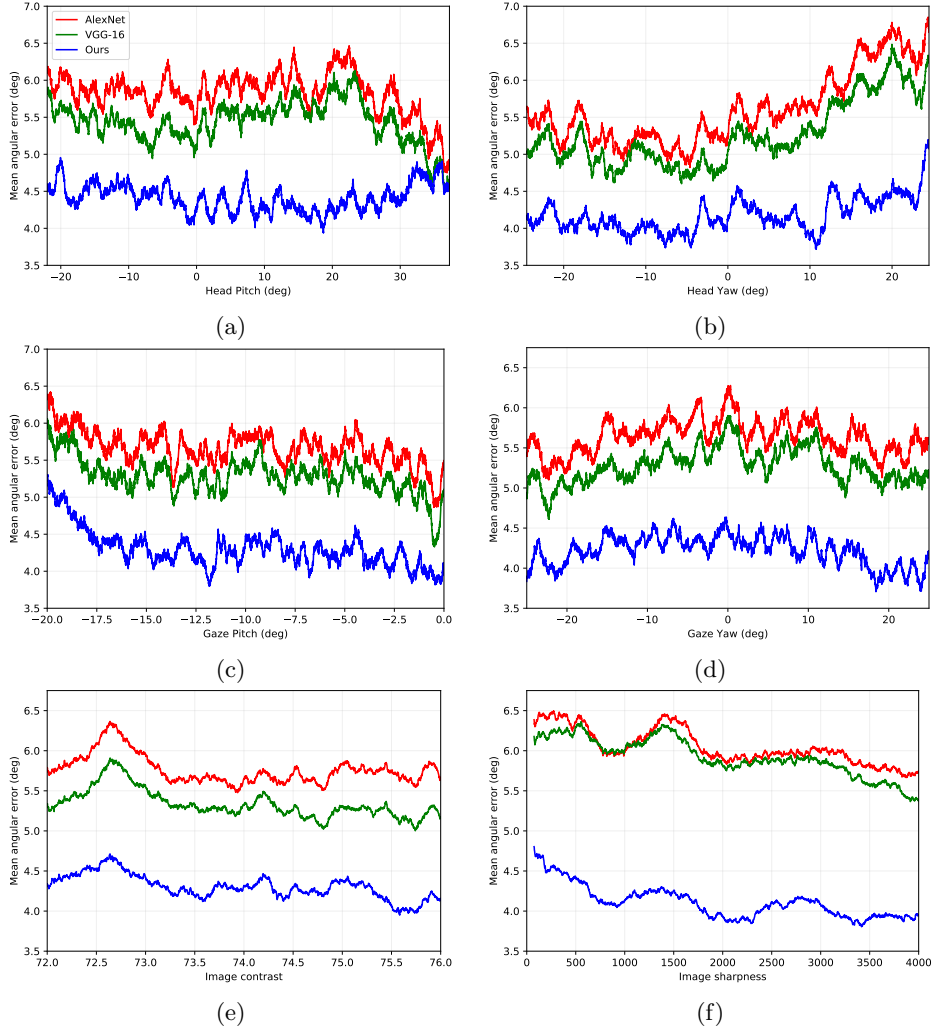
Fig. 6. Robustness of AlexNet (red), VGG-16 (green), and our approach (blue) to different head pose (top), gaze direction (middle), and image quality (bottom). The lines are a moving average.

Bibliography

[1] Baluja, S., Pomerleau, D.: Non-intrusive gaze tracking using artificial neural networks. Tech. rep., Pittsburgh, PA, USA (1994)

[2] Bekerman, I., Gottlieb, P., Vaiman, M.: Variations in eyeball diameters of the healthy adults. Journal of ophthalmology **2014** (2014)

[3] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. vol. 1, p. 7 (2017)

[4] Chin, C.A., Barreto, A., Cremades, J.G., Adjouadi, M.: Integrated electromyogram and eye-gaze tracking cursor control system for computer users with motor disabilities. Journal of rehabilitation research and development **45 1**, 161–74 (2008)

[5] Forrester, J.V., Dick, A.D., McMenamin, P.G., Roberts, F., Pearlman, E.: The Eye E-Book: Basic Sciences in Practice. Elsevier Health Sciences (2015)

[6] Funes-Mora, K.A., Odobez, J.M.: Gaze estimation in the 3d space using rgb-d sensors. International Journal of Computer Vision **118**(2), 194–216 (Jun 2016). https://doi.org/10.1007/s11263-015-0863-4

[7] Funes Mora, K.A., Monay, F., Odobez, J.M.: Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In: Proceedings of the Symposium on Eye Tracking Research and Applications. pp. 255–258. ETRA '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2578153.2578190

[8] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)

[9] Honari, S., Molchanov, P., Tyree, S., Vincent, P., Pal, C., Kautz, J.: Improving landmark localization with semi-supervised learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)

[10] Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)

[11] Huang, M.X., Kwok, T.C., Ngai, G., Leong, H.V., Chan, S.C.: Building a self-learning eye gaze model from user interaction data. In: Proceedings of the 22Nd ACM International Conference on Multimedia. pp. 1017–1020. MM '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2647868.2655031

[12] Huang, Q., Veeraraghavan, A., Sabharwal, A.: Tabletgaze: Dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. Mach. Vision Appl. **28**(5-6), 445–461 (Aug 2017). https://doi.org/10.1007/s00138-017-0852-4

[13] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014)

[14] Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)

[15] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)

[16] Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Artificial Intelligence and Statistics. pp. 562–570 (2015)

[17] Liu, H., Heynderickx, I.: Visual attention in objective image quality assessment: Based on eye-tracking data. IEEE Transactions on Circuits and Systems for Video Technology **21**(7), 971–982 (2011)

[18] Lu, F., Okabe, T., Sugano, Y., Sato, Y.: A head pose-free approach for appearance-based gaze estimation. In: Proceedings of the British Machine Vision Conference. pp. 126.1–126.11. BMVA Press (2011), http://dx.doi.org/10.5244/C.25.126

[19] Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Inferring human gaze from appearance via adaptive linear regression. In: Proceedings of the 2011 International Conference on Computer Vision. pp. 153–160. ICCV '11, IEEE Computer Society, Washington, DC, USA (2011). https://doi.org/10.1109/ICCV.2011.6126237

[20] Majaranta, P., Bulling, A.: Eye Tracking and Eye-Based Human-Computer Interaction, pp. 39–65. Advances in Physiological Computing, Springer (2014)

[21] Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. pp. 483–499. Springer (2016)

[22] Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., Hays, J.: Webgazer: Scalable webcam eye tracking using user interactions. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI). pp. 3839–3845. AAAI (2016)

[23] Pech-Pacheco, J.L., Cristobal, G., Chamorro-Martinez, J., Fernandez-Valdivia, J.: Diatom autofocusing in brightfield microscopy: a comparative study. In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. vol. 3, pp. 314–317 vol.3 (2000). https://doi.org/10.1109/ICPR.2000.903548

[24] Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. arXiv **abs/1603.01249** (2016)

[25] Sesma, L., Villanueva, A., Cabeza, R.: Evaluation of pupil center-eye corner vector for gaze estimation using a web cam. In: Proceedings of the Symposium on Eye Tracking Research and Applications. pp. 217–220. ETRA '12, ACM, New York, NY, USA (2012). https://doi.org/10.1145/2168556.2168598

[26] Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversar-

ial training. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)

[27] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)

[28] Smith, B.A., Yin, Q., Feiner, S.K., Nayar, S.K.: Gaze locking: Passive eye contact detection for human-object interaction. In: Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology. pp. 271–280. UIST '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2501988.2501994

[29] Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based 3d gaze estimation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1821–1828 (June 2014). https://doi.org/10.1109/CVPR.2014.235

[30] Sun, L., Liu, Z., Sun, M.T.: Real time gaze estimation with a consumer depth camera. Inf. Sci. **320**(C), 346–360 (Nov 2015). https://doi.org/10.1016/j.ins.2015.02.004

[31] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)

[32] Tan, K.H., Kriegman, D.J., Ahuja, N.: Appearance-based eye gaze estimation. In: Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision. pp. 191–. WACV '02, IEEE Computer Society, Washington, DC, USA (2002)

[33] Tompson, J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1. pp. 1799–1807. NIPS'14, MIT Press, Cambridge, MA, USA (2014)

[34] Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1653–1660. CVPR '14, IEEE Computer Society, Washington, DC, USA (2014). https://doi.org/10.1109/CVPR.2014.214

[35] Wang, K., Ji, Q.: Real time eye gaze tracking with 3d deformable eye-face model. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). ICCV '17, IEEE Computer Society, Washington, DC, USA (2017)

[36] Wedel, M., Pieters, R.: A review of eye-tracking research in marketing. In: Review of marketing research, pp. 123–147. Emerald Group Publishing Limited (2008)

[37] Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)

[38] Wood, E., Baltruaitis, T., Zhang, X., Sugano, Y., Robinson, P., Bulling, A.: Rendering of eyes for eye-shape registration and gaze estimation. In:

Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). pp. 3756–3764. ICCV '15, IEEE Computer Society, Washington, DC, USA (2015). https://doi.org/10.1109/ICCV.2015.428

[39] Wood, E., Baltrušaitis, T., Morency, L.P., Robinson, P., Bulling, A.: A 3d morphable eye region model for gaze estimation. In: European Conference on Computer Vision. pp. 297–313. Springer (2016)

[40] Wood, E., Baltrušaitis, T., Morency, L.P., Robinson, P., Bulling, A.: Learning an appearance-based gaze estimator from one million synthesised images. In: Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications. pp. 131–138. ETRA '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2857491.2857492

[41] Wood, E., Bulling, A.: Eyetab: Model-based gaze estimation on unmodified tablet computers. In: Proceedings of the Symposium on Eye Tracking Research and Applications. pp. 207–210. ETRA '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2578153.2578185

[42] Xiong, X., Liu, Z., Cai, Q., Zhang, Z.: Eye gaze tracking using an rgbd camera: A comparison with a rgb solution. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication. pp. 1113–1121. UbiComp '14 Adjunct, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2638728.2641694

[43] Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., Shen, J.: The menpo facial landmark localisation challenge: A step towards the solution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (July 2017)

[44] Zhang, X., Sugano, Y., Bulling, A.: Everyday eye contact detection using unsupervised gaze target discovery. In: Proc. of the ACM Symposium on User Interface Software and Technology (UIST). pp. 193–203 (2017). https://doi.org/10.1145/3126594.3126614, best paper honourable mention award

[45] Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4511–4520 (June 2015). https://doi.org/10.1109/CVPR.2015.7299081

[46] Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 2299–2308 (July 2017). https://doi.org/10.1109/CVPRW.2017.284

[47] Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017). https://doi.org/10.1109/TPAMI.2017.2778103

[48] Zhang, Y., Bulling, A., Gellersen, H.: Sideways: A gaze interface for spontaneous interaction with situated displays. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 851–860. CHI '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2470654.2470775

[49] Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI. pp. 94–108. Springer International Publishing, Cham (2014)

A   Baseline Architectures

The state-of-the-art CNN architecture for appearance-based gaze estimation is based on a lightly modified VGG-16 architecture [47], with mean cross-person gaze estimation error of $5.5°$ on the MPIIGaze dataset [45]. We compare against a standard VGG-16 architecture [27] and an AlexNet architecture [15] which has been the standard architecture for gaze estimation in many works [14, 46]. The specific architectures used as baseline are described in Table 3.

Both models are trained with a batch size of 32, learning rate of 5×10^{-5} and L_2 weights regularization coefficient of 10^{-4}, using the Adam optimizer [13]. Learning rate is multiplied by 0.1 every $5,000$ training steps, and slight data augmentation is performed in image translation and scale.

B   Image metrics

In this section we describe the image metrics used for the robustness plots concerning image quality (Figures 6e and 6f in paper).

B.1   Image contrast

The root mean contrast is defined as the standard deviation of the pixel intensities:

$$\text{RMC} = \sqrt{\frac{1}{MN} \sum_{i=1}^{N-1} \sum_{j=1}^{M-1} (I_{ij} - \bar{I})^2}$$

where I_{ij} is the value of the image $I \in M \times N$ at location (i, j) and $\bar{I}$ is the average intensity of all pixel values in the image.

B.2   Image sharpness

In order to have a sharpness-based metric, we calculate the variance of the image I after having convolved it with a Laplacian, similar to [23]. This corresponds to an approximation of the second derivative, which is computed with the help of the following mask:

$$L = \frac{1}{6} \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

After convolving I with L, we compute the standard deviation of the resulting image I_L to get the image sharpness (IS) metric:

$$IS = \sigma(I_L)$$

Table 3. Configuration of CNNs used as baseline for gaze estimation. The style of [27] is followed where possible. s represents stride length, p dropout probability, and conv9-96 represents a convolutional layer with kernel size 9 and 96 output feature maps. maxpool3 represents a max-pooling layer with kernel size 3.

(a) AlexNet
input (150×90 eye image)
conv9-96 ($s = 2$)
local response norm.
maxpool3 ($s = 2$)
conv5-256 ($s = 1$)
local response norm.
maxpool3 ($s = 2$)
conv3-384 ($s = 1$)
conv3-384 ($s = 1$)
conv3-256 ($s = 1$)
maxpool3 ($s = 2$)
FC-4096
dropout ($p = 0.5$)
FC-4096
dropout ($p = 0.5$)
FC-2

(b) VGG-16
input (150×90 eye image)
conv3-64
conv3-64
maxpool2 ($s = 1$)
conv3-128
conv3-128
maxpool2 ($s = 2$)
conv3-256
conv3-256
conv3-256
maxpool2 ($s = 2$)
conv3-512
conv3-512
conv3-512
maxpool2 ($s = 2$)
conv3-512
conv3-512
conv3-512
maxpool2 ($s = 2$)
dropout ($p = 0.5$)
FC-4096
dropout ($p = 0.5$)
FC-4096
FC-2