Gaze estimation via self-attention augmented convolutions

Gabriel Lefundes Vieira
Intelligent Vision Research Lab,
Federal University of Bahia, Bahia, Brazil
Email: gabriel.lefundes@ufba.br

Luciano Oliveira
Intelligent Vision Research Lab,
Federal University of Bahia, Bahia, Brazil
Email: lrebouca@ufba.br

***Abstract*—Although recently deep learning methods have boosted the accuracy of appearance-based gaze estimation, there is still room for improvement in the network architectures for this particular task. Hence we propose here a novel network architecture grounded on self-attention augmented convolutions to improve the quality of the learned features during the training of a shallower residual network. The rationale is that self-attention mechanism can help outperform deeper architectures by learning dependencies between distant regions in full-face images. This mechanism can also create better and more spatially-aware feature representations derived from the face and eye images before gaze regression. We dubbed our framework ARes-gaze, which explores our Attention-augmented ResNet (ARes-14) as twin convolutional backbones. In our experiments, results showed a decrease of the average angular error by 2.38% when compared to state-of-the-art methods on the MPIIFaceGaze data set, while achieving a second-place on the EyeDiap data set. It is noteworthy that our proposed framework was the only one to reach high accuracy simultaneously on both data sets.**

I. Introduction

Gaze estimation is an active area of research within computer vision, and its relevance spans a large array of fields. Gaze pattern analysis during various activities can be, for example, a valuable source of information in behavioral and health research [1]–[3]. In computer systems, gaze can be applied as a mean of interaction in augmented and virtual reality applications [4]–[6], mobile applications [5], [7], and human-computer interaction [8].

Methods of gaze estimation can be categorized as model-based or appearance-based [9]. The former relies on explicitly modeling the subject's eye and using some type of feedback to infer the gaze direction geometrically. This approach can reach accurate results in controlled environments, although suffering from hardware cost and installation overhead. Model-based gaze estimation is usually limited by external factors such as lighting conditions and lower tolerance for subject pose and distance. Appearance-based approaches attempt to directly predict the gaze vector from RGB images of the subject by mapping a regression function that can be ultimately learned from data.

While challenges like lighting conditions and unconstrained subject pose remain, the use of deep learning and in-the-wild large-scale data sets [10]–[12] have greatly improved the accuracy of appearance-based methods, which in general need only monocular cameras as input sensors. Recent publications in the field have focused on exploring different neural network architectures and training conditions to raise the performance of the current state-of-the-art. Notably, many works have remarked that full-face images innately carry relevant information about the subject's pose, and using them as inputs along with the usually extracted eye-patches can improve the prediction accuracy significantly [13]–[15].

A. Contributions

Here we explore the recent trend of attention mechanisms in deep learning [18] as a way to produce higher quality features by improving the spatial awareness of the network. The rationale is to better leverage the relationship between coarse pose information from face images and fine information from eye-patches. We introduce a ResNet-inspired [19] network, dubbed **A**ttention-augmented **Res**Net (ARes-14), conceived upon a self-attention-based mechanism as proposed by [20]. With 14 layers, as the name suggests, ARes-14 was intuitively driven to improve appearance-based gaze estimation, which needs spatial awareness but does not require very deep architectures to be effective.

To provide gaze estimation from a monocular camera, we also propose a framework called ARes-gaze, which is comprised of two ARes-14 networks that act as twin feature extractors, taking as inputs full-face images and isolated eye-patches. We showed that, as reported in [20] for baseline classification tasks, some of the weights of early attention maps can learn to highlight geometric structures from the full-face images, leading us to hypothesize that self-attention augmented convolutions can fulfill a similar role to the spatial importance maps conceptualized in [13]. This ability can help the network better focus on facial regions relevant to gaze estimation. These results further reinforce the intuitive notion that self-attention layers are particularly useful for tasks where the input image has strong spatial correlations with the ground truth, such as appearance-based gaze estimation. As a result, ARes-gaze achieved state-of-the-art performance on two challenging data sets. When compared with similar methods of appearance-based gaze estimation, we found a decrease in the average angular error by 2.38% on the MPIIFaceGaze data set, achieving the second-best result on the EyeDiap data set. Table I summarizes the characteristics of our framework in comparison with other state-of-the-art works.

Method	3D gaze output	Full-face as input	Eye as input	Multimodal inputs	Spatial awareness	Attention augmented
MPIIGaze [10]	✓	–	✓	–	–	–
iTracker [14]	–	✓	✓	✓	–	–
Spatial Weights [13]	✓	✓	–	–	✓	–
RT-Gene [11]	✓	✓	✓	–	–	–
Recurrent CNN [16]	✓	✓	✓	✓	–	–
Dilated Net [15]	✓	✓	✓	–	–	–
FAR-Net [17]	✓	✓	✓	–	–	–
Ours	✓	✓	✓	–	✓	✓

TABLE I: Summary of the state-of-the-art on appearance-based gaze estimation in comparison with our work.

B. *Related works*

Early works in appearance-based gaze estimation used well-established machine-learning algorithms like adaptive linear regression [21], support vector regression [22], and random forests [23] to learn the mapping function from eye images to gaze vectors. Recently, convolutional neural networks (CNNs) have shown great success in gaze estimation, with its first published iteration [10] reporting significant gains over the previous state-of-the-art works.

Subsequent publications have then built upon the notion of using CNNs by proposing different input models for the convolutional networks like images of the entire face and a binary grid to encode head size and position [14]. Other works have proposed taking into account domain knowledge and peculiarities of the gaze estimation task while designing the architecture of the CNN itself. In [17], for example, the asymmetrical nature of left and right eyes is posited to have relevance on the result of gaze estimation, and accuracy gains are reported when encoding and leveraging that asymmetry in a deep neural network. Another example of domain-specific modeling is found in [15] where dilated convolutions are used as a replacement for max-pooling layers to better capture small differences in eye images. In [16], recurrent CNNs are used and shown to improve prediction accuracy significantly on continuous inference. This is so because it is plausible to consider gaze an inherently temporal phenomenon, which is grounded by the notion that where people are looking at, in a particular moment in time, directly depends on where they were looking at, in a previous moment. In [13], a spatial-weights mechanism is proposed to learn spatial importance maps and predict gaze directions using only face images as input. This map serves as a guide to the following layers of the CNN, learning to locate important features on the normalized input image (eyes, nose) while pointing to where the network focus should be.

The Squeeze-and-Excitation (SE) [24] blocks, the Bottleneck Attention Module (BAM) [25] and the Convolutional Block Attention Module (CBAM) [26] are all proposals for drop-in components that should be able to introduce attention capabilities to CNNs. In [20], the principle of multi-headed self-attention from the Transformer network [18] is adapted for 2D inputs, presenting a hybrid layer with attention and convolution operations performed in parallel. Unlike BAM and CBAM, which refine existing convolutional feature maps with attention, self-attention augmented convolutions create new attention maps to be fused with their convolutional counterparts.

II. Gaze estimation with self-attention augmented convolutions

A. *Gaze vector*

3D appearance-based gaze estimation can be comprehended as to find a function capable of mapping an input image, $\boldsymbol{I}$, to a gaze vector, $\hat{\mathbf{g}}$. Given that the gaze direction is usually also dependent on head pose, ($\mathbf{h}$), we include this latter into the formulation, thus generically obtaining:

$$\hat{\mathbf{g}} = f(\mathbf{I}, \mathbf{h}) \,, \tag{1}$$

where $\hat{\mathbf{g}}$ is a 2D unit vector with the origin being in the middle point between the subject's eyes. The components that form $\hat{\mathbf{g}}$ are the pitch ($\hat{\mathbf{g}}_\theta$) and yaw ($\hat{\mathbf{g}}_\phi$) angles. Here, the mapping function is the proposed trained neural network, and $\mathbf{h}$ is implicitly inferred from full-face images. We can then rewrite the generic appearance based formula as $\hat{\mathbf{g}} = f(\mathbf{I}^{eyes}, \mathbf{I}^{face})$.

B. *Attention-augmented convolutional layer*

First proposed as an alternative base layer for classification [20], attention-augmented convolutions (AAConv) extend the multi-head attention concept from the Transformer network [18] by applying self-attention to 2D arrays. In regular convolution layers, inter-pixel correlation is usually spatially constrained by the convolutional kernel. This limits the degree to which is possible to relate distant sections from an image that could have relevant relationships. Similar to what is done in Transformer networks with 1D sequences, attention-augmented convolutions use self-attention to handle pixel matrices. Each pass through an AAConv layer can be split into two main parts: The first one through a regular convolutional layer, while the second through a multi-headed attention layer. The outputs (W_o, H_o) of each individual attention-head are concatenated and projected onto the original spatial dimensions of height and width of the input (W_i, H_i). Additionally, relative positional embeddings [27] are expanded to two dimensions in order to encode spatially-relevant information while maintaining translation equivalence [20]. In the end, the results from both passes of the convolutional and

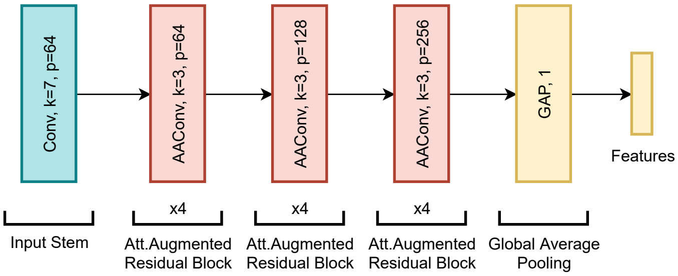


Fig. 1: Self-attention augmented ResNet with 14 layers (ARes-14). Each residual block is comprised of 2 sequential blocks of two convolutional layers each. All convolutions in residual blocks are augmented with self-attention while the input stem remains with conventional convolutions.

the multi-headed attention layers are concatenated, forming spatially-aware convolutional feature maps from the input image. Expanding the neural network principle of long-distant spatial relationships, it is possible to achieve a clear positive effect when applied to straight-forward classification tasks across a range of different architectures [20].

C. ARes-14: A self-attention augmented convolutional backbone

In appearance-based gaze estimation performed by CNNs, shallow networks can be sufficient as long as the task is performed in relatively constrained conditions [11] (across a limited range of head pose and with short distances between subject and camera). These conditions are intrinsic to some available data sets, although they are not a reasonable expectation for in-the-wild applications. These constraints can be simulated in more challenging data by applying preprocessing and normalization procedures (see Section III-B for more details). The use of these strategies allows us to train with more structured data. Also, procedures like those should still perform well in more complex environments by normalizing the input data before sending it through the prediction network during inference time. The use of shallower networks is of particular importance given the significant computational overhead of training with self-attention in convolutional networks (see [20] for a more detailed discussion).

ResNet is a widespread and well understood general-purpose CNN, turning it onto an ideal candidate for a baseline comparison against self-attention augmentation. We started with the shallow version, ResNet-18, and replaced every convolutional layer for an equivalent self-attention augmented convolution with compatible dimensions. The number of parameters was further reduced by removing the last-layer block, essentially shrinking the architecture to 14 layers. Each convolution and AAConv is followed by a batch normalization and activation (ReLU) operation. The ratio between attention channels and output filters (k), as well as, the ratio between the key depth and output filters (v) were both fixed to 0.25 for every self-attention augmented convolution. Unless otherwise specified, the number of attention heads, Nh, is fixed to 8. We called this novel network architecture as **ARes-14**, which is used as the backbone in our proposed framework for gaze estimation. Figure 1 depicts ARes-14 architecture.

D. ARes-gaze: A framework for gaze estimation

To perform gaze estimation, we propose a fairly conventional framework: A two-stemmed network where each branch is an instance of ARes-14, and the extracted features are joined by a shared prediction layer, as shown in Fig. 2. We used a feature vector of 256 elements obtained from each convolutional backbone after the global average pooling layer, resulting in 512 features to be sent through the prediction layers (see Fig. 1).

Many works have used multi-input frameworks in appearance-based gaze estimation [11], [15]–[17], [28] since the gaze direction of a subject relies heavily on more than one factor (eyes, head pose, and location, distance, etc). Here our inputs are RGB-face images, normalized for pose and distance, and grayscale eye images, histogram-normalized.

To extract information from the isolated eye-patches, while some published methods with similar topologies use two networks (one for each eye) [11], [15] or a single network with shared weights (making separate passes for each input) [17], we employed a single-pass, single-network strategy for the eye branch by stacking the left- and right-eye regions, creating a $1 : 1$ ratio square input. We study the practical implications of the use of this method in comparison with the other mentioned works in Section IV-B1. The extracted-feature vectors from the face and the eyes are then joined by concatenation and passed through a prediction block to output the two values of our gaze vector prediction.

III. Materials and Methods

A. Training data

Two challenging and publicly-available data sets were selected to perform our experiments: MPIIFaceGaze [13] and EyeDiap [29].

The MPIIFaceGaze data set [10] was the first to provide unconstrained data for gaze estimation in-the-wild. 15 subjects (9 males, 6 females, and 5 subjects with glasses) were recorded in various sessions during day-to-day use of their laptops, where targets were occasionally displayed at random positions in the screen. The recorded data contains a large number of different conditions of recording locale (inside and outside), illumination, head pose and position, and overall recording quality. Since the original MPIIGaze data set provides cropped-eye regions already, we used its modified version MPIIFaceGaze [13], which provides 3,000 full-face, already normalized images for each subject.

The EyeDiap data set is a collection of 94 videos with 16 different subjects in 3 different modalities: **Discrete screen target** - where a target was displayed in regular intervals on random locations on a screen, **continuous screen target** – in which the target moved along random trajectories

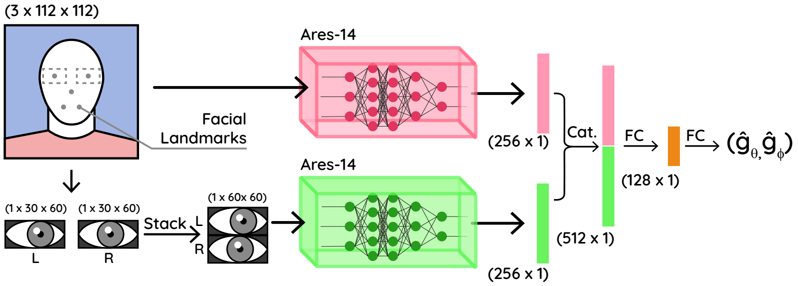


Fig. 2: ARes-gaze framework. Face- and eye-patches are extracted and separately normalized from the source image, subsequently being sent through twin ARes-14 backbones. The resulting features from each backbone are then concatenated and passed through a prediction stage consisting of two fully-connected layers.

in the screen, and **3D floating target** – where a small ball was moved in front of the participant. In our experiments, we used only the modalities where the target was projected onto the screen (continuous and discrete), since, in the floating target-sessions, the small ball would sometimes occlude the subject's face. Two subjects only have video recordings on floating target sessions, so we are left with a total of 14 subjects and 56 videos.

B. Data set normalization

Similar to [30], we applied an affine transformation to rotate the image as to cancel out the roll-axis angle of the head, and to scale it to the desired size (standardizing the distance from the face to the virtual camera). The effect of that transformation is that relevant facial features are always in the same regions on the input, making the network job easier to recognize patterns in important regions. This procedure is only applied on the EyeDiap [29] data set since the MPIIFaceGaze data set [13] is already normalized. For a squared input, we defined that the distance, d, between the left- and right-eye centers should be 40% of the image width. Given that the face should be vertically centered, this gives us left- and right-eye centers to be $(0.7, 0.35)$ and $(0.3, 0.35)$ on a normalized 0 to 1 scale relative to the input dimensions. For the face patch, we used RGB images with an input size of 112×112 pixels. The eye patches are cropped from the normalized face, converted to grayscale, and the histogram normalized before being resized to the input shape of 30×60. These steps are carried out for both data sets.

C. Implementation Details

The code for the models was written with the PyTorch [31] framework[1]. All the models were trained for 120 epochs with a batch size of 48 on an HPC cluster equipped with 8 NVidia V100 GPUs. The high computational overhead of training attention-based methods is prohibitive with regards to the batch size, and this needs to be taken into account during the hyperparameter tuning. We used a stochastic gradient descendant (SGD) [32] solver with a momentum equal to 0.9, and a weight decay of 0.0003 was empirically found to be optimal in preventing over-fitting. The learning rate is linearly warmed-up for 5% of the epochs until reaching the value of 0.128, then gradually decreased by cosine annealing [33]. The loss used for training the whole model was the smooth $L1$ cost function.

IV. Experimental results

To characterize our proposed framework properly, a group of experiments was carried out and divided into two main parts: First, a set of ablation studies were performed to assess the impact of the self-attention augmentation modules on aspects of the ARes-gaze architecture. The main goal of this part is to better understand the optimal conditions to apply AAConvs in our framework. In the second part, some of the external factors that directly impact the performance of gaze estimation were analyzed to explore how our proposed framework can deal with them.

A. Evaluation methodology

To allow reproducibility and reduce the effects of subject's dependence on our evaluations, a leave-one-out cross-validation strategy was used across the subjects from each data set. Considering the characteristics of the data sets used in the experiments, N models were trained, where N is the number of available subjects in a data set. For each model, a different subject is held out and used for testing. The final result is the average of the evaluations of all models. On the EyeDiap data set, for example, the final scores are the average performance of 14 trained models on the held-out subject, each time. Similarly, on the MPIIFaceGaze data set, 15 models were trained and their performance scores were averaged into the final results.

B. Ablation studies

This first evaluation has the goal of comparing our proposed single-branch, single-pass vertical stacking scheme (see Section II-D) with other strategies adopted by similar methods. The second is a study to try to untangle the effect of self-attention augmentation on different inputs (face and eyes)

[1] Link to git repository will be available as paper acceptance

Model type	Average angular error		# Parameters	FLOPs
	MPIIFaceGaze	EyeDiap		
SE	**5.40°**	**7.27°**	**2.810**	**414**
DP	5.54°	7.42°	2.842	422
TB	5.45°	7.36°	5.619	422

TABLE II: Results on different input models of the eye images: Stacked-eyes (SE), double-pass with shared weights (DP) and separate branches for each input (TB). The number of trainable parameters and the approximate floating operations (FLOPs) are in millions.

#Net	**Network type**	**Input**		**Dataset**	
		Eyes	Face	MPIIFaceGaze	EyeDiap
1	Regular	■		5.40°	7.27°
2	Attention	■		5.33°	6.02°
3	Regular		■	4.71°	7.42°
4	Attention		■	4.46°	6.10
5	Regular	■	■	4.46°	6.09°
6	Regular Attention	■ 	 ■	4.42°	5.81°
7	Regular Attention	 ■	■ 	4.52°	5.84°
8	Attention	■	■	**4.17°**	**5.58°**

TABLE III: Results of attention-augmented versus regular convolutional layers on the backbones of ARes-gaze. Best results are shown in bold.

and the network schemes to understand how and where self-attention is effective on the task of gaze estimation. Finally, given the significant computational overhead of using multi-headed attention, we evaluate the effect of choosing different numbers of attention heads for ARes-14.

1) Evaluating different models of the eye images: For the eye-patch branch of our network, the input consists of images of both left and right eyes from the subject. Other published works with similar network topologies either need to perform two forward passes [17] or use a dedicated network branch for each eye [11], [15]. We propose the vertical stacking of eye images to obtain a $1:1$ input image that can be processed in a single pass. We evaluated ARes-gaze against the other mentioned models, considering the parameters in Table Table II. Three models were considered for the eye branch: Stacked-eyes input **(SE)**, Double-pass with shared weights **(DP)**, and three-branch pipeline **(TB)**. As summarized in Table II, although there is arguably only a small difference in the average angular error, the stacked-input model performed better than the other ones on both data sets. Also the stacked-input model presents roughly the same number of trainable parameters of the shared-weights variety and a significantly lower number when compared to the twin-branch network. These results further validate the adoption of the stacked-eye for all subsequent evaluations.

2) ARes-14 evaluation: With the aim of gauging the effect of self-attention augmentation in multiple stages of ARes-gaze, we evaluated and compared multiple models based on the ARes-14 architecture. First, to see how attention affects different types of input, we trained single-branch networks with and without self-attention augmentation in isolated versions with only eye images as inputs, or only face images as inputs. Second, we applied the ARes-gaze and compare models switching between ResNet-14 and ARes-14 backbones for each input branch. The goal is to explore the contrast between fully convolutional features and self-attention augmented features for gaze estimation.

The results for each model are laid out in Table III. The evaluated network variations are: single branch with regular Resnet-14 (networks #1 and #3), single branch attention-augmented (networks #2 and #4), and dual branched with mixed regular/attention backbones (networks #5 through #8). When considering the variation on input modality (eye and face), we obtain a total of 8 different trained models. For the single-branch networks (with either only face or only eyes as inputs), we observe a drop of more than 17% on the average angular error on the EyeDiap data set when using self-attention augmented convolutions. When compared with its regular convolutional form, ARes-gaze reduces the average error by 6.5% on the MPIIFaceGaze data set and by 8.4% on EyeDiap.

3) Determining the number of attention heads: On the evaluation of AAConvs in classification tasks reported in [20], the accuracy gains are on architectures using a fixed number of attention-heads, specifically $Nh = 8$. In this section, we evaluate ARes-gaze considering other values of Nh, but no greater than 8, because of the prohibitive computational cost that was actually observed in practice.

Table IV shows the average angular errors found on the MPIIFaceGaze and EyeDiap data sets. Notably, for the MPIIFaceGaze data set, when using less than 4 attention-heads, the ARes-gaze architecture performs worse than the purely convolutional baseline, with the evaluation error proportionally decreasing with the increase of attention-heads. On the EyeDiap data set, the results follow the same direction with $Nh = 2$ and $Nh = 4$, which are only marginally better than the baseline network. In both data sets, there is a sudden and significant improvement in the results when $Nh = 8$.

C. Comparison of ARes-gaze with other appearance-based methods

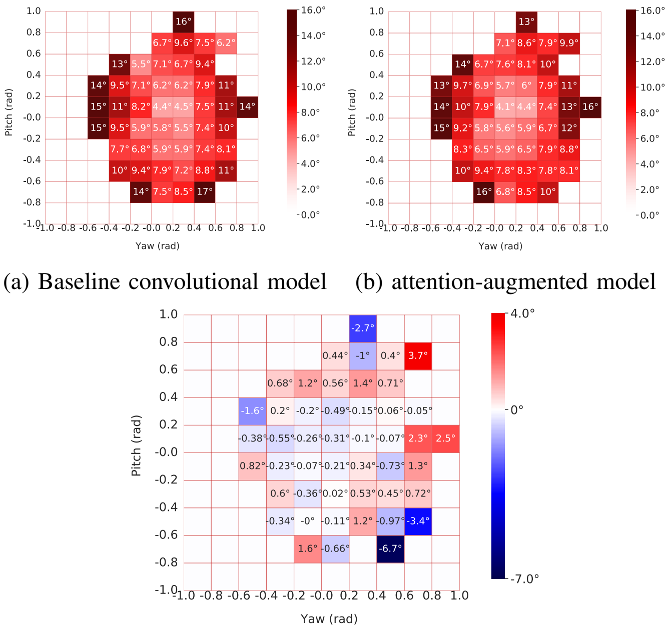
We selected six appearance-based methods that take as input either full-face images or a combination of full-face images and other inputs. All these methods output a single-gaze vector with origin in the center of the face or in the middle-point of the eye. The selected methods were: the iTracker in its original form [14] and with AlexNet backbone [13], the CNN with spatial-weights mechanism [13], RT-Gene (a version of 4

Method	MPIIFaceGaze	EyeDiap
Baseline	4.46°	6.09°
ARes-gaze (Nh=2)	4.93°	5.98
ARes-gaze (Nh=4)	4.36°	5.99
ARes-gaze (Nh=8)	**4.17°**	**5.58°**

TABLE IV: Results of average angular errors on different numbers of attention-heads per attention layer. Best results are highlighted.

Method	MPIIFaceGaze	EyeDiap
iTracker [14]	6.2°	8.3°
iTracker (AlexNet) [13], [14]	5.6°	–
Spatial Weights CNN [13]	4.8°	6.0°
RT-Gene (4 Ensemble) [11]	4.3°	–
Dilated CNN [15]	4.5°	**5.4°**
FAR-Net [17]	4.3°	5.7°
Baseline	4.5°	6.1°
ARes-gaze (Nh = 8)	**4.2°**	5.6°

TABLE V: Results of average angular error compared with other appearance-based methods. Best results are highlighted.



(a) Baseline convolutional model

(b) attention-augmented model

(c) Angle-error difference between attention-augmented and baseline models on the head-pose evaluation on the EyeDiap data set.

Fig. 3: Distribution of mean angular error of baseline (a) and attention-augmented (b) models across head poses in the EyeDiap dataset. (c) summarizes the difference between the previous two plots: Blue boxes mean improvement over the baseline model, while red boxes mean an increase in the average angular error.

ensembles with the best reported results) [11], the CNN with dilated convolutions proposed in [15], and the eye-asymmetry based FAR-Net [17]. These are approaches we consider similar to ours, which were compared over the average 3D-angular error on the chosen data sets. Except for RT-Gene and iTracker (AlexNet), which do not report evaluations on the EyeDiap data set, all compared methods use the same or a similar protocol to extract data from the videos, as described in Section III-A. The results are reported by considering two versions of our architecture: The full ARes-gaze and ARes-gaze without self-attention augmentation. When compared to the other methods, ARes-gaze framework with twin ARes-14 backbones reached state-of-the-art results on the MPIIFaceGaze data set, and the second-best place on the EyeDiap data set, being only 0.2 degrees behind the best result of Dilated CNN [15] (see Table V). It is worth noting that no other method was able to have superior results on both data sets at the same time.

D. Evaluating external factors in gaze estimation

In in-the-wild gaze estimation applications, the subject's head pose and external illumination conditions are to be considered unconstrained. To see how self-attention augmentation in convolutions affects robustness to these factors, we evaluated both our completely attention-augmented model and the traditional convolutional baseline in isolated scenarios. We averaged the angular error obtained to evaluate how the models perform for different angles of the subject's head pose. We used the EyeDiap data set due to its more varied distribution of head pose angles by our model on regular intervals of 0.20 radians for pitch and yaw w.r.t the subject's head pose.

Figure 3 shows the results for both self-attention regular (Fig. 3a) and augmented convolutional (Fig. 3b) based architectures. The plots clearly show that the gains are obtained across most of the pitch and yaw head-pose spectrum (Fig. 3c). The overall decrease in average error appears mostly uniform outside of the most extreme cases. For those, it is noticeable that the larger gains obtained by the ARes-gaze model on the EyeDiap data set were in the regions of extreme pitch angles (negative and positive), and the heavier losses were in the regions of high yaw angles.

1) Illumination conditions: Figure 4 shows an overlapping evaluation of both baseline and ARes-gaze models by light-level intervals. There is a clear inverse relationship between light level and angular error that behaves somewhat linearly. Additionally, the last bin, representing overly lit images, shows a small spike in the averaged angular error. This situation reinforces the intuitive notion that appearance-based gaze estimation models have worse accuracy with both poorly lit and overexposed input images. To quantify the sensibility of each model to lighting conditions, we fit a regression line across the angle error of each bin, and calculated its slope (m). The closer to zero the slope is, the lower is the model sensibility to light. This experiment showed that ARes-gaze had a slightly smaller slope inclination, although the difference was not enough to justify conclusions about its robustness to lighting conditions in comparison with the purely convolutional baseline.

E. Result analysis

1) On the use of self-attention augmented convolutions for gaze estimation: First, we evaluated the difference between using eye images *versus* using the entire face as inputs. Intuitively, the difference between using full-face images and isolated-eye regions as inputs is the scope of the information that the network is able to extract. With full-face images, CNN has the chance to learn not only from the eyes themselves but also extract head-pose information from regions such as the nose and mouth. This comes with the drawback of the subject's eyes having a lower resolution, thus limiting the amount of information present in their regions. In contrast, using isolated

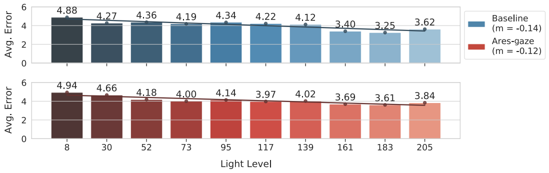


Fig. 4: Model accuracy *versus* lighting conditions of the input images. The MPIIFaceGaze data was split into 10 bins with regard to light levels, with the X-axis showing the average level of each bin. The Y-axis is the average angular error in degrees.

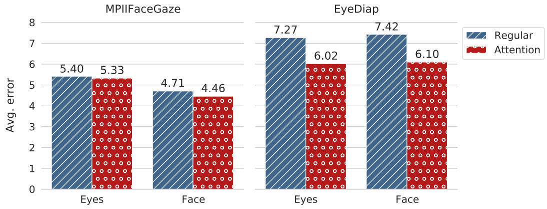


Fig. 5: Average angular error for single-branch gaze estimation networks on the MPIIFaceGaze and EyeDiap data sets. Blue bars represent ResNet-14 as the backbone, while red bars represent ARes-14.

eye-patches should allow the network to extract more detailed information about the pupils' positions, turning the network to be more sensitive to smaller changes in the eye movement. In this case, the drawback is the absence of elements that can inform the network about the subject's head pose, which has relevance to the final prediction.

Figure 5 shows a visual summary of the results of the single-branch networks from Table III. There is a clear and consistent decrease of the average angular error in all instances when using the networks with self-attention augmented convolutions (ARes-14). As to which kind of input benefits the most from attention, on the EyeDiap data set, an error decrease of 17.19% with eyes as input *versus* 17.78% with faces can be observed. On the MPIIFaceGaze data set, the decreases were of 1.28% and 5.31%, respectively. The larger magnitude of gains on the EyeDiap data set can be inferred from the fact that it is a more challenging data set with regards to head pose, which is an issue that made us hypothesize about the self-attention augmentation benefits when applied to gaze estimation. This analysis is further reinforced by the evaluation of angle error grouped by head pose angle presented by the EyeDiap data set, where the more expressive gains are observed in extreme pose angles which could not be found in the MPIIFaceGaze data set.

Figure 6 shows a comparison of the results obtained from multiple inputs across different iterations of our proposed gaze estimation architecture (replacing ResNet backbones by ARes-14 in each branch). It is worth noting that between the networks using ARes-14 as the backbone for only one of the branches, the one with self-attention augmentation on the face branch wins by a slight margin. On the MPIIFaceGaze data set, the one with attention only on the eye branch even had a small but noticeable drop in performance when compared with the regular CNN baseline. When analyzing the results from the single-branch network evaluation, it is possible to note that the face branch benefits slightly more from self-attention augmented convolutions due to having more distant elements that can be correlated by self-attention. This is reinforced by our results on the evaluation of mixed attention and regular convolution networks.

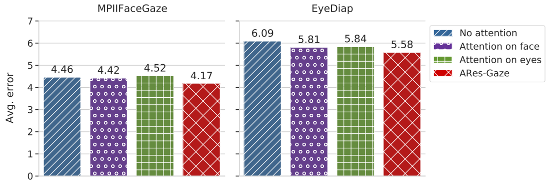


Fig. 6: Average angular error for four different versions of the proposed gaze estimation framework. From left to right: Regular convolutional baseline (blue), a version with ARes-14 on the face branch and ResNet-14 on the eye branch (purple), a version with ResNet-14 on the face branch and ARes-14 on the eye branch (green), and the fully attention-augmented ARes-gaze (red).

Our findings indicate that self-attention augmented convolutions can be used as drop-in replacements to convolutional layers in gaze estimation networks to reduce the angular error in evaluation. Yet, while self-attention augmented convolutions work well with both face and eye-input images, our experiments showed that networks working with the full-face image as input were more prone to improvement when augmented by self-attention.

2) On the number of attention heads per convolutional layer: We obtained the overall best results with the largest number of attention-heads ($Nh = 8$) evaluated in our experiments. It follows that further increasing this parameter could produce even better results, but we were not able to assess this hypothesis due to the significant increase in computational overhead which did not fit our hardware constraints. Notably, we obtained the counter-intuitive results that for a number of attention-heads less than eight, $Nh < 8$, the ARes-gaze framework sometimes actually performed worse than the regular convolutional baseline. It is worth observing that, for the face images, the self-attention augmented layer is capable of highlighting semantically relevant regions of the image. We verified however that when this phenomenon happens, it is only present on the map of the eighth attention head. This leads us to conclude that the attention layer might need a certain depth of attention-heads in order to specialize in very particular tasks.

V. Conclusion

In this paper, we addressed the question "can self-attention augmented convolutions be used to reduce angular error in appearance-based gaze estimation?", and we found that when compared to an equivalent regular convolutional network, the use of our 2D self-attention-based architecture can indeed produce more accurate results. We used ARes-14 twin branches as self-attention augmented CNNs in our experiments, and we guess that further research is merited on the design of optimal architectures for each branch of a multi-input attention-augmented framework such as the proposed ARes-gaze. We showed that the input face images had more to gain from using AAConvs than the input eye images, so incorporating domain knowledge of both attention mechanisms and gaze estimation to refine each branch for its particular input (face and eyes) might produce even better results than those reported. Notably, we highlight the head pose estimation task, and even joint head pose and gaze direction estimation networks, with the possibility of including other types of input images such as facial landmarks and explore their behavior.

Acknowledgment

The authors acknowledge the National Laboratory for Scientific Computing (LNCC/MCTI, Brazil) for providing HPC resources of the SDumont supercomputer. Gabriel was supported by the Fundação de Amparo à Pesquisa do Estado da Bahia [grant number 892/2019], and Luciano was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico [grant number 307550/2018-4].

References

[1] M. A. Hollands, A. E. Patla, and J. N. Vickers, ""look where you're going!": gaze behaviour associated with maintaining and changing the direction of locomotion," *Experimental brain research*, vol. 143, no. 2, pp. 221–230, 2002.

[2] G. Warlop, P. Vansteenkiste, M. Lenoir, J. Van Causenbroeck, and F. J. Deconinck, "Gaze behaviour during walking in young adults with developmental coordination disorder," *Human Movement Science*, vol. 71, p. 102616, 2020.

[3] T. Nakano, K. Tanaka, Y. Endo, Y. Yamane, T. Yamamoto, Y. Nakano, H. Ohta, N. Kato, and S. Kitazawa, "Atypical gaze patterns in children and adults with autism spectrum disorders dissociated from developmental changes in gaze behaviour," *Proceedings of the Royal Society B: Biological Sciences*, vol. 277, no. 1696, pp. 2935–2943, 2010.

[4] S. Nilsson, T. Gustafsson, and P. Carleberg, "Hands free interaction with virtual information in a real environment: Eye gaze as an interaction tool in an augmented reality system." *PsychNology Journal*, vol. 7, no. 2, 2009.

[5] M. Lankes and B. Stiglbauer, "Gazear: Mobile gaze-based interaction in the context of augmented reality games," in *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*. Springer, 2016, pp. 397–406.

[6] J.-Y. Lee, H.-M. Park, S.-H. Lee, T.-E. Kim, and J.-S. Choi, "Design and implementation of an augmented reality system using gaze interaction," in *2011 International Conference on Information Science and Applications*. IEEE, 2011, pp. 1–8.

[7] M. Barz, F. Daiber, D. Sonntag, and A. Bulling, "Error-aware gaze-based interfaces for robust mobile gaze interaction," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 2018, pp. 1–10.

[8] D. Rozado, F. B. Rodriguez, and P. Varona, "Low cost remote gaze gesture recognition in real time," *Applied Soft Computing*, vol. 12, no. 8, pp. 2072–2084, 2012.

[9] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 3, pp. 478–500, 2009.

[10] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4511–4520.

[11] T. Fischer, H. Jin Chang, and Y. Demiris, "Rt-gene: Real-time eye gaze estimation in natural environments," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 334–352.

[12] D. Lian, L. Hu, W. Luo, Y. Xu, L. Duan, J. Yu, and S. Gao, "Multiview multitask gaze estimation with deep convolutional neural networks," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 10, pp. 3010–3023, 2018.

[13] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 51–60.

[14] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2176–2184.

[15] Z. Chen and B. E. Shi, "Geddnet: A network for gaze estimation with dilation and decomposition," *arXiv preprint arXiv:2001.09284*, 2020.

[16] C. Palmero, J. Selva, M. A. Bagheri, and S. Escalera, "Recurrent cnn for 3d gaze estimation using appearance and shape cues," *arXiv preprint arXiv:1805.03064*, 2018.

[17] Y. Cheng, X. Zhang, F. Lu, and Y. Sato, "Gaze estimation by exploring two-eye asymmetry," *IEEE Transactions on Image Processing*, vol. 29, pp. 5259–5272, 2020.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

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

[20] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3286–3295.

[21] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Adaptive linear regression for appearance-based gaze estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 10, pp. 2033–2046, 2014.

[22] T. Schneider, B. Schauerte, and R. Stiefelhagen, "Manifold alignment for person independent appearance-based gaze estimation," in *2014 22nd international conference on pattern recognition*. IEEE, 2014, pp. 1167–1172.

[23] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-synthesis for appearance-based 3d gaze estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1821–1828.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[25] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.

[26] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[27] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[28] W. Zhu and H. Deng, "Monocular free-head 3d gaze tracking with deep learning and geometry constraints," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3143–3152.

[29] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "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*, 2014, pp. 255–258.

[30] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-synthesis for appearance-based 3d gaze estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1821–1828.

[31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.

[32] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[33] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.