IDIAP RESEARCH REPORT

IMPROVING FEW-SHOT USER-SPECIFIC GAZE ADAPTATION VIA GAZE REDIRECTION SYNTHESIS

Yu Yu Gang Liu Jean-Marc Odobez

Idiap-RR-03-2019

APRIL 2019

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Improving Few-Shot User-Specific Gaze Adaptation via Gaze Redirection Synthesis

Yu Yu, Gang Liu, Jean-Marc Odobez
Idiap Research Institute, CH-1920, Martigny, Switzerland
EPFL, CH-1015, Lausanne, Switzerland
{yyu, gang.liu, odobez}@idiap.ch

Abstract

As an indicator of human attention gaze is a subtle behavioral cue which can be exploited in many applications. However, inferring 3D gaze direction is challenging even for deep neural networks given the lack of large amount of data (groundtruthing gaze is expensive and existing datasets use different setups) and the inherent presence of gaze biases due to person-specific difference. In this work, we address the problem of person-specific gaze model adaptation from only a few reference training samples. The main and novel idea is to improve gaze adaptation by generating additional training samples through the synthesis of gaze-redirected eye images from existing reference samples. In doing so, our contributions are threefold: (i) we design our gaze redirection framework from synthetic data, allowing us to benefit from aligned training sample pairs to predict accurate inverse mapping fields; (ii) we proposed a self-supervised approach for domain adaptation; (iii) we exploit the gaze redirection to improve the performance of person-specific gaze estimation. Extensive experiments on two public datasets demonstrate the validity of our gaze retargeting and gaze estimation framework.

1. Introduction

Gaze, as a subtle non-verbal human behaviour, not only indicates the visual content people perceive but also conveys information about the level of attention, mental state or even higher level psychological constructs of human. As a consequence, gaze cues have been exploited in many areas like social interaction analysis [10], stress analysis [8], human robot interaction (HRI) [1, 20], the emerging Virtual Reality industry [21, 24], and they are expected to find a wide range of application in mobile interactions with smart phones [9, 14, 30].

However, gaze extraction from non invasive visual sensors is challenging and has attracted an increased amount of research in recent years. Approaches can be classified in two general categories: geometric based methods (GBM) and appearance based methods (ABM). The former ones rely on a geometrical model of eyes whose parameters can be inferred from localized eye landmarks like iris or eye corners. Although they can be very accurate, they usually require high resolution data to reliably extract eye features. The latter ABM ones directly learn a mapping from the eye images to the gaze directions and have been shown to be more robust against low eye image resolution and other variability factors (illumination, head pose, gaze range,...). Nevertheless, in spite of recent progresses partly due to the use of deep neural networks [2, 14, 16, 23, 44, 45], vision based gaze estimation is still a challenging and open problem due to at least three main factors:

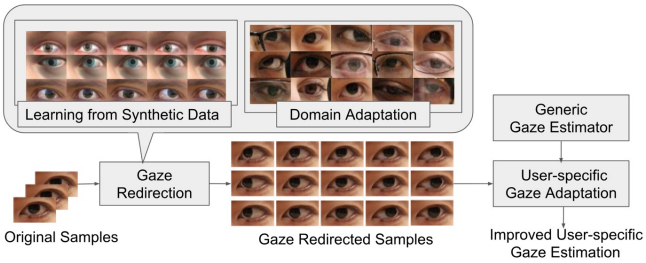


Figure 1. Approach overview. A few reference eye images (with gaze ground truth) from a user are used as input to a gaze redirection synthesis module to generate further training samples. The latter (and reference samples) are used to fine-tune a generic gaze estimator to obtain a user-specific gaze estimator.

- **Lack of data.** The sizes of benchmark gaze datasets [3, 28, 44] are relatively small compared to other vision tasks like image classification, since accurate gaze annotation is complex and expensive. To address the lack of data, domain adaptation methods [27] have proposed to use synthetic images for training, but completely eliminating the domain discrepancies between real and synthetic eye images is hard.
- **Systematic bias.** Existing gaze datasets usually use different gaze coordinate systems and data preprocessing methods, in particular for geometric normalization (rectification) relying on different head pose estimators. This introduces a between-dataset

systematic bias regarding the gaze groundtruth [41].

- **Person-specific bias.** Liu et al. [16] legitimaly argue that gaze can not be fully estimated from the visual appearance since the alignment difference between the optical axis (the line connecting the eyeball center and the pupil center) and the visual axis (the line connecting the fovea and the nodal point [4]) is person specific, and vary within -2 to 2 degrees across the population. Therefore, it is not optimal to train a single generic model for accurate cross-person gaze estimation.

In this paper, we focus on the problem of person-specific gaze adaptation which has not received enough attention compared to cross person gaze estimation. More specifically, the aim is to only rely on few samples since collecting training samples for a new subject is expensive. In this context, a first and interesting result that we show is that a direct and simple fine tuning of a neural network gaze regressor can improve person-specific gaze estimation by a good margin, even if the number of person-specific samples is as small as 9. We then propose to further improve the performance of such gaze adaptation method by using as additional training data gaze-redirected samples synthesized from the given reference samples, as illustrated in Fig. 1. Compared with domain adaptation methods like SimGAN [27], which work by retargeting synthetic images into subject specific eye images, we firmly believe that a gaze redirection framework relying on reference eye images and user defined gaze changes (redirection angles) can generate samples with more realistic appearance (since they are directly derived from real eye images of the subject) and more reliable groundtruth (less systematic and person-specific bias), thus demonstrating better performance when used for person-specific gaze adaptation. By investigating the aboves ideas, we make the following contributions:

- **Gaze redirection network training.** Unlike previous approaches [6, 12], our redirection network is pre-trained with synthetic eye images so that a large amount of well aligned image pairs (the same eye position, eye size, head pose and illumination) can be exploited. As a result, thanks to the large amount of data, the network does not require the eye landmarks as anchoring points. Besides, we also propose to exploit the segmentation map of synthetic samples for regularization during training.
- **Gaze redirection domain adaptation.** Training with synthetic data results in the domain shift problem. However, as we do not have aligned pairs of real images to do domain adaptation, we proposed instead a self-supervised method relying on a cycle consistency loss and a gaze redirection loss.
- **Person-specific gaze adaptation using gaze-redirected samples.** We hypothesize that these samples will provide more diverse visual content and gaze groundtruth compared to the reference samples they originated from, thus improving the person-specific gaze adaptation. To the best of our knowledge, we are the first to propose this idea and a series of experiments to validate its efficacy.

The rest of the paper is organized as follows. We first summarize the related works in Section 2 and then introduce our method in Section 3. Experimental results are reported in Section 4 while a brief discussion is made in Section 5. The final conclusion is given in Section 6.

2. Related Works

Gaze Estimation. As stated in introductionm, non-invasive vision based gaze estimation methods can be divided into geometric ones (GBM) and appearance based ones (ABM) [7]. GBMs build eye models based on some eye features, such as eye corners or iris localization and infer gaze direction using geometric relationship between elements like the line joining the eyeball center to the iris center [4, 31–33, 37, 40]. Usually they do not require much training samples except for a few calibration points, but they suffer from low resolution imaging, noise and variable lighting conditions.

ABMs are more robust to those factors [5, 17, 18, 29, 36], as they learn a regressor from annotated data samples and estimate gaze directly from the images. In particular, recently deep learning approaches [2, 14, 22, 33, 42, 43, 45] have been shown to work well because they train a regression network leveraging large amounts of data. They can capture what are the image features essential for gaze estimation under various conditions, such as various eye shapes, illumination, glasses and head pose.

Gaze Adaptation. However, when testing on unknown person, the different personal eye structures such as eye shapes and visual axis limit the performance of both GBMs and ABMs [16]. Some straightforward solutions to this problem have been proposed, such as to learn person-specific models [15, 29, 43], fine-tune a pre-trained model [19], learn a SVR using a few samples for calibration [14] or learn a differential gaze model [16].

Training a person-specific model or fine-tuning a pre-trained model can achieve very high accuracy for such person, but it usually requires relatively large amount of annotated data from this person, which is not wanted in practice. Calibrating person-specific model with an SVR or relying on differential gaze only require a few reference annotated samples, but those samples do not reflect the global gaze map, and the estimation error will increase when the gaze difference between the test sample and the reference sample becomes large.

Under this circumstance, we propose a gaze redirection method that can alleviate the drawbacks from the aforemen-

tioned methods. More precisely, our algorithm can generate more diverse and realistic images using a few annotated samples from this person. Then these data can be used to fine-tune a pre-trained gaze model.

Gaze Redirection. As far as we know, the computer vision and graphics based gaze redirection for video-conferencing was first studied in [47], in which two components are included to solve this task. The first is tracking the user's head pose and eye ball motion, and the second consists of manipulating the head orientation and eye gaze. Following this work, Weiner *et.al.* [34] evaluated and proved the overall feasibility of gaze redirection in face images via eye synthesis and replacement by integrating the vision and graphical algorithm within a demonstration program. But changes in the eyelid configuration were not considered. Then a simple solution that detects eyes and replaces them with eye images in a front gaze direction was proposed in [26, 35]. Kononenko *et.al.* proposed a pixel-wise replacement method using an *eye flow tree* and could synthesize realistic views with a gaze systematically redirected upwards by 10 to 15 degrees [13]. Then they updated the eye flow tree by a deep warping network trained on pairs of eye images corresponding to eye appearance before and after the redirection [6, 12]. However, these methods require large amount of annotated data for training.

To circumvent this issue, Wood [39] proposed a model based method that does not need any training samples. It first builds and fits a multi-part eye region model using an analysis-by-synthesis method to simultaneously recover the eye region shape, texture, pose, and gaze for a given image. Then, it manipulates the eyes by warping the eyelids and rendering eyeballs in the output image. It achieves better results especially for large redirection angles.

3. Gaze Adaptation approach

Our overall approach for user-specific gaze adaptation is illustrated in Fig. 1. It consists in fine-tuning a generic neural network using labeled training samples. However, rather than only using the very few (less than 10) reference samples, we propose to generate additional samples using a gaze redirection model. As this redirection model is the main component of our approach, we describe it with more details in the sections 3.1 to 3.4. The gaze adaptation part is then described in section 3.5.

3.1. Gaze Redirection Overview

Our framework for gaze redirection is shown in Fig. 2. It is composed of the redirection network itself and a domain adaptation module. The left part of Fig. 2 illustrates the redirection network which takes the eye image, the user defined redirection angle and the head pose as input. It is designed as an encoder-decoder manner where the output of the decoder is an inverse warping field. The gaze-redirected sample is then generated by warping the input eye image with the predicted inverse warping field (via a differentiable sampler). The right part of Fig. 2 is the domain adaptation module which is conducted in a self-supervised way through a cycle consistency loss and a gaze redirection loss.

3.2. Synthetic Data for Gaze Redirection Learning

In principle, the training of a gaze redirection network needs well aligned image pairs where the two images (the input one and the redirection groundtruth for supervision) share the same overall illumination condition, the same person-specific properties (skin color, eye shape, iris color, pupil color) and the same head pose. The only difference should be gaze-related features such as eye ball orientation and eyelid status. This strict requirement make it hard fto collect real data. In this paper, we propose to use synthetic samples instead. Concretely, we use the UnityEyes Engine [38] to produce 3K eye image groups, each containing 10 images generated with the same illumination, the same person-specific parameter, the same head pose, but different gaze parameters, as shown in Fig. 3. A totol of 10*9 image pairs can thus be drawn from each group. In our work, we used 10K image pairs for training.

3.3. Gaze Redirection Network

Architecture. It is illustrated in Fig. 2. The network takes three variables as input, the eye image $\mathbf{I}$, the head pose $\mathbf{h}$ and the user defined redirection angle $\Delta\mathbf{g}$. Among them, $\mathbf{I}$ is processed by an image branch and encoded as a semantic feature, while $\mathbf{h}$ and $\Delta\mathbf{g}$ are processed with another two branches and encoded as features which will guide the gaze related visual changes. Note that the head pose input is a must since it is one of the elements which determine the appearance of eye images. The three output features are then stacked in a bottleneck layer and further decoded into two inverse warping maps $\mathbf{m}_x$ and $\mathbf{m}_y$:

$$\mathbf{m}_{x,y} = \mathbf{R}_\theta(\mathbf{I}, \Delta\mathbf{g}, \mathbf{h}) \tag{1}$$

where $\mathbf{R}$ is the redirection network and θ is the network parameter. Similarly to [12], we then use a differentiable grid sampler s [11] to warp the input image and generate the gaze-redirected image $\mathbf{I}_{\Delta\mathbf{g}}$ whose gaze groundtruth is $\mathbf{g} + \Delta\mathbf{g}$ ($\mathbf{g}$ is the gaze of the original image $\mathbf{I}$) according to:

$$\begin{aligned}\mathbf{I}_{\Delta\mathbf{g}}(x,y) = \sum_i \sum_j \mathbf{I}(i,j) \cdot max(0, 1 - |i - \mathbf{m}_x(x,y)|) \\ \cdot max(0, 1 - |j - \mathbf{m}_y(x,y)|).\end{aligned} \tag{2}$$

For simplicity, we rewrite the above formulas as:

$$\mathbf{I}_{\Delta\mathbf{g}} = \mathbf{I} \circ \mathbf{R}_\theta(\mathbf{I}, \Delta\mathbf{g}, \mathbf{h}) \tag{3}$$

where $\circ$ represents the warping operation. Compared with direct synthesis, this strategy projects the pixels of the input

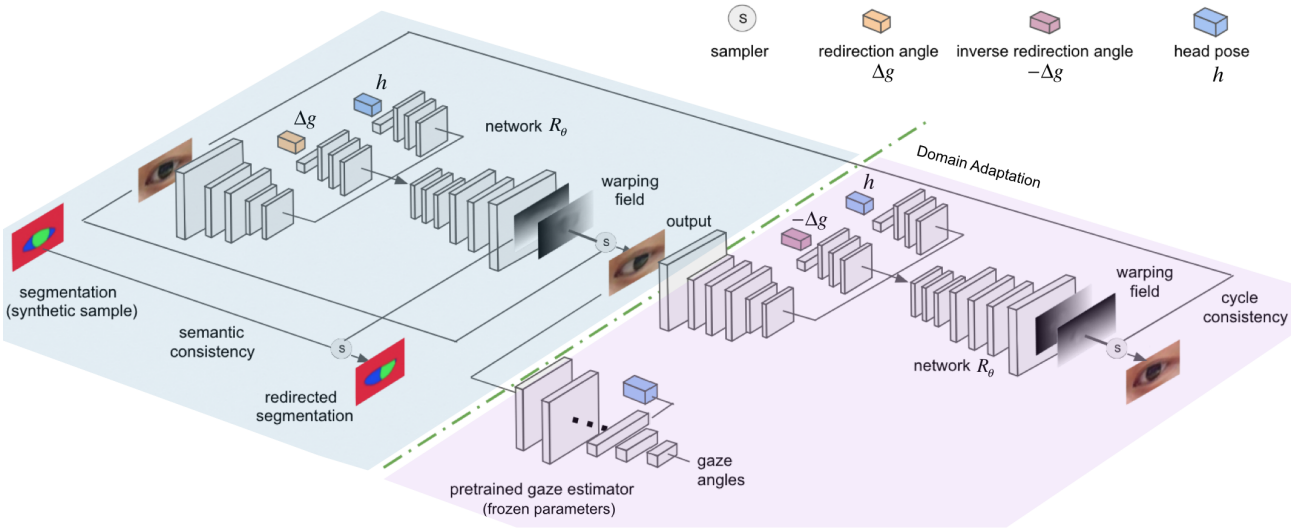


Figure 2. Gaze redirection network (top left), along with learning components (eye segmentation for semantic consistency, cycle consistency, gaze prediction consistency).

Figure 3. Aligned UnityEyes samples (placed in rows)

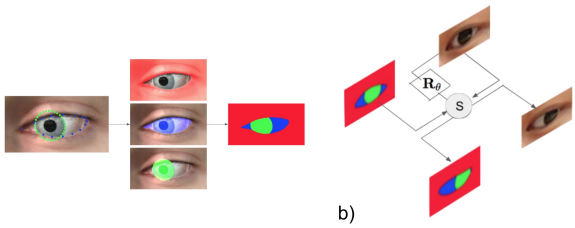


Figure 4. Semantic consistency. (a) Deterministic segmentation of a synthetic sample, red: background, blue: sclera, green: iris + pupil. (b) the gaze redirection of a segmentation map.

to the output, which guarantees that the input and the output will share similar color and illumination distributions.

For training, we use an L1 loss to measure the difference between the redirection output $\mathbf{I}_{\Delta\mathbf{g}}$ and the groundtruth $\mathbf{G_I}$. Therefore, generating the required inverse warping field for redirection is learned in an indirect supervised way.

Semantic consistency. So far, the network can be evaluated by measuring the reconstruction loss between the predicted gaze-redirected eye image $\mathbf{I}$ and the corresponding groundtruth $\mathbf{G_I}$. If the predicted inverse warping field is accurate, then the different semantic parts of the eye (pupil, sclera and background) should also be well redirected. We thus propose to enforce the warping consistency at the semantic level. To do so, for each synthetic image $\mathbf{I}$, we extract the semantic map as follows: we first fit convex shapes to the eyelid landmarks and the iris landmarks (provided by UnityEyes) to get the maps of the iris + pupil region, the sclera region and the background region. We then merge these three maps into a segmentation map $\mathbf{S_I}$, as shown in Fig. 4a. It is important to note that this step is deterministic and is not a part of the network. Then, any segmentation map $\mathbf{S_I}$ can then be redirected with the inverse warping field $\mathbf{R}_\theta(\mathbf{I}, \Delta\mathbf{g}, \mathbf{h})$ (which is predicted from the original image $\mathbf{I}$) and compared with the segmentation map $\mathbf{S}_{\mathbf{G_I}}$ of the target redirected eye $\mathbf{G_I}$.

Overall loss. According to previous paragraphs, our overall redirection loss L_R (for synthetic data) can be defined as the sum of a reconstruction loss and of the semantic loss, using in each case L1 norms. It is thus defined as:

$$L_R = ||\mathbf{I} \circ \mathbf{R}_\theta(\mathbf{I}, \Delta\mathbf{g}, \mathbf{h}) - \mathbf{G_I}||_1 + ||\mathbf{S_I} \circ \mathbf{R}_\theta(\mathbf{I}, \Delta\mathbf{g}, \mathbf{h}) - \mathbf{S}_{\mathbf{G_I}}||_1 \quad (4)$$

Please note that the segmentation map is not processed by the network (looking at Fig. 4b) and will not be required at user gaze adaptation time for generating redirected samples.

3.4. Gaze Redirection Domain Adaptation

Because of the domain difference between synthetic and real data, the performance of the network $\mathbf{R}_\theta$ learned only from synthetic data degrades when it is applied to real data. A straightforward solution to solve this issue would be to fine tune $\mathbf{R}_\theta$ with real image pairs. However, as mentioned above, collecting real image pairs for gaze redirection is difficult. In this section, we introduce a self-supervised domain adaptation method relying on two principles. The first one is gaze redirection cycle consistency, and the second

one is based on the consistency of the estimated gaze from the gaze redirected image.

Cycle consistency loss. It has been used for applications like domain adaptation [46] and identity preserving [25]. The main idea is that when a sample is transferred to a new domain and then converted back to the original domain, the cycle output should be the same as the input. Similarly, in our case, if a gaze redirected sample $\mathbf{I}_{\Delta\mathbf{g}}$ is further redirected with the inverse redirection angle -$\Delta\mathbf{g}$, the cycle output should be close to the original image $\mathbf{I}$.

In this paper, we apply this cycle consistency scheme to the set of real images, and define the cycle loss as:

$$L_{cycle} = ||\mathbf{I}_{\Delta\mathbf{g}} \circ \mathbf{R}_\theta(\mathbf{I}_{\Delta\mathbf{g}}, -\Delta\mathbf{g}, \mathbf{h}) - \mathbf{I}||_1 \quad (5)$$

where $\mathbf{I}_{\Delta\mathbf{g}} = \mathbf{I} \circ \mathbf{R}_\theta(\mathbf{I}, \Delta\mathbf{g}, \mathbf{h})$.

Gaze redirection loss. As a weakness, the cycle loss alone could push the redirection network to collapse to an identity mapping (the output of the redirection network is always equal to the input). To prevent this collapse, we propose to exploit a gaze redirection loss. More concretely, given a set of real data, we first train a generic gaze estimator $\mathbf{E}_\phi$ using them. We then freeze the parameters of $\mathbf{E}_\phi$ and use it to define a loss on the gaze-redirected image, enforcing that the gaze predicted from this image should be close to its target groundtruth (see bottom of Fig. 2). More formally:

$$L_{gaze} = ||\mathbf{E}_\phi(\mathbf{I} \circ \mathbf{R}_\theta(\mathbf{I}, \Delta\mathbf{g}, \mathbf{h})) - (\mathbf{g} + \Delta\mathbf{g})||_2 \quad (6)$$

Besides preventing the collapse, the real data trained gaze estimator $\mathbf{E}_\phi$ can help reducing the systematic bias in the gaze redirection network (arising from intially training the network with only synthetic data) and therefore help the domain adaptation of $\mathbf{R}_\theta$.

Network adaptation optimization. To conduct network adaptation, we do not consider the two losses in the same minibatches, as they are of different nature. In addition, to balance the domain adaptation and the gaze redirection, not all parts of the network need to be adapted simultaneously. In practice, we thus optimize the two losses alternatively according to the following scheme. For the cycle loss L_{cycle}, we only optimize the image encoding branch since i) domain shift usually occurs when encoding an input image into semantic features; ii) the fixed decoder part can further prevent the redirection network from collapsing. For the gaze redirection loss L_{gaze}, only the head pose and gaze branches are updated. The image encoder and decoder remain frozen in this case to prevent an overfitting to L_{gaze}. We use Stochastic Gradient Descent (SGD) to optimize the network.

3.5. Gaze Estimation Adaptation

As stated earlier, the aim of the gaze redirection is to generate more person-specific samples for gaze adaptation. In our work, we first train a generic gaze estimator using the real data from several identities. We then adapt the estimator with the samples of a new person and their gaze-redirected outputs. This adaptation is conducted in a few-shot setting, meaning the number of original samples of this new person is few (less than 10). More concretely, the generic estimation network is fine tuned with the person-specific samples during 10 epochs. In the first 5 ones, we use both the original and the gaze-redirected samples, while in the last 5 ones we only use the original samples to minimize the effects of potentially wrong redirected samples. Since the number of samples is small, we use Batch Gradient Descent instead of Stochastic Gradient Descent. Further details about the generic gaze estimator and its adaptation can be found in the Experiment Section.

4. Experiment

In experiments, our main aim is to evaluate the performance of the person-specific gaze estimators adapted from a generic estimator using few reference samples and their gaze-redirected samples. Nevertheless, we also conduct a subjective test to evaluate to which extent the redirected samples are realistic enough for humans. Note that in this paper, we only target single eye image gaze estmtion (and redirection and adaptation), leaving the full-face case as future work.

4.1. Experimental Setting

Datasets. We use the ColumbiaGaze Dataset [28] and the MPIIGaze Dataset [43] for experiment. The former one contains the gaze samples of 56 persons while the latter contains eye images of 15 persons.

Generic gaze estimator. As our gaze estimator, we use GazeNet [45]. It is is based on a *vgg16* architecture. To train it, we follow the protocols of the ColumbiaGaze and MPIIGaze datasets (i.e. as for cross-subject experiments), using respectively a 5-fold and 15-fold training scheme. The error of our generic gaze estimator on ColumbiaGaze is $3.54°$ ($3.9°$ in [23]) while the error on MPIIGaze is $5.35°$ ($5.5°$ in [45]), showing better performance than the state-of-the-art results. Please note that the generic gaze estimator[1] is also exploited as $\mathbf{E}_\phi$ to define the gaze redirection loss, as defined in section 3.4.

Evaluated models. Starting from the generic gaze estimator, we develop a series of adaptation methods to contrast with our approach. The first two methods are the linear (*LinAdap*, [16]) and the SVR (*SVRAdap*, using the features of the second last layer [14, 16]) gaze adaptation methods which learn additional regressors from the gaze estimator output (*LinAdap*) or features (*SVRAdap*), and thus do not

[1] A generic estimator is trained for each fold. In none of the experiments, data from the test subject is used in either part of the training phase.

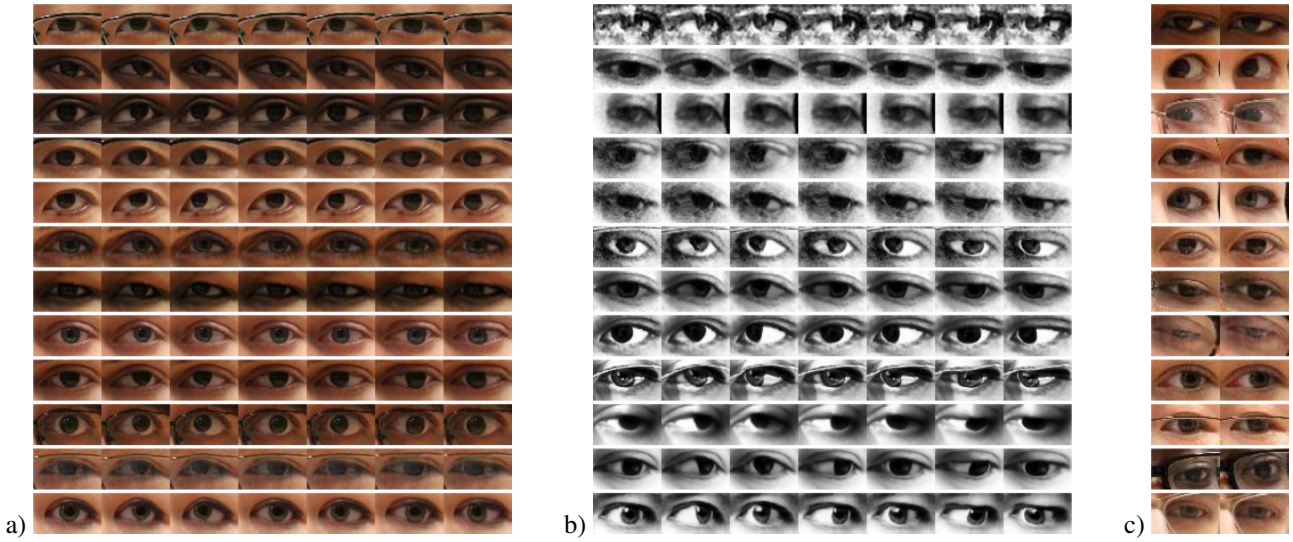


Figure 5. Redirection qualitative results from the ColumbiaGaze (a) and MPIIGaze (b) datasets. In (a) and (b), the first image of each row is an original sample, whereas the remaining images in the row are redirected samples from this original sample. Subfigure (c) displays pairs of images used in the subjective test: in each pair, the left image is an original image from the dataset, while the right one is a redirected sample (obtained from another original sample) which has the same gaze label (i.e. direction) as the left one.

change (or adapt) the generic gaze estimator. In contrast, the third and fourth approaches directly fine tune the generic estimator using either only the reference samples (*FTAdap*, *FT* for fine tuning) or as well the gaze redirected samples (*RedFTAdap*, *Red* for redirection).

In addition, we also implement a differential gaze estimator *DiffNet* [16] for comparison. The *DiffNet* is trained to predict gaze differences, and it exploits the reference samples to predict the gaze of a new eye image. For a fair comparison, we replace the three convolution layers used as feature extractor in [16] with the *vgg16* feature extractor. Please note that the *DiffNet* approach can be regarded as a person-specifc network since person-specific samples (at least one) are required to estimating the gaze of new eye image.

Gaze redirection parameters. For each person, we randomly draw n (n =1, 5 or 9) person-specific samples and generate $t \cdot n$ gaze-redirected samples where the default value of t is 10. For the MPIIGaze dataset in which the gaze groundtruth is continuous, the yaw and pitch components $(\Delta\mathbf{g}_p, \Delta\mathbf{g}_y)$ of the redirection angle $\Delta\mathbf{g}$ are randomly chosen with the range $[-10, 10] \times [-15, 15]$ ($[-10, 10]$ for pitch, and $[-15, 15]$ for yaw). For the ColumbiaGaze dataset, where the annotated gaze is discrete, $\Delta\mathbf{g}$ is chosen from the same range but with discrete values $(\pm 5°, \pm 10°, \pm 15°)$. The impact of t and of the redirection ranges are further studied in the result section.

Performance measurement. We use the angle (in degree) between the predicted gaze vector and the groundtruth gaze vector as the error measurement. Note that gaze vectors are 3D unit vectors constructed from the pitch and yaw angles. To eliminate random factors, we performed 10 rounds of person-specific sample selection, gaze redirection and gaze adaptation, and reported the average estimation error.

Table 1. ColumbiaGaze dataset: gaze adaptation performance

error \ approach / #sample	*Cross Subject*	*LinAdap*	*SVRAdap*	*FTAdap*	*DiffNet*	*RedFTAdap*
1		-	-	5.53	4.64	**3.92**
5	3.54	4.65	7.67	3.11	3.63	**2.88**
9		3.78	5.39	2.79	3.50	**2.60**

Table 2. MPIIGaze dataset: gaze adaptation performance

error \ approach / #sample	*Cross Subject*	*LinAdap*	*SVRAdap*	*FTAdap*	*DiffNet*	*RedFTAdap*
1		-	-	5.28	5.93	**4.97**
5	5.35	5.43	7.68	4.64	4.42	**4.20**
9		4.61	5.79	4.31	4.20	**4.01**

4.2. Results

Gaze redirection qualitative results. We show some qualitative results of the redirection network in Fig. 5(a) and (b). As can be seen, our redirection network does a realistic synthesis for samples with different skin or iris color. Furthermore, we also found that the redirection model is robust when working with noisy eye images, as illustrated in several rows of Fig. 5(b).

Gaze adaptation performance. They are reported in Table. 1 (ColumbiaGaze dataset) and Table. 2 (MPIIGaze dataset). From the tables, we observe that the proposed approach *RedFTAdap* achieves the best results while the *LinAdap* and *SVRAdap* methods obtain the worst results,

Table 3. ColumbiaGaze: Results with different redirection range

error $\Delta\mathbf{g}_p$ \ $\Delta\mathbf{g}_y$	$[-5, 5]$	$[-10, 10]$	$[-15, 15]$
$[-10, 10]$	2.66	2.62	**2.60**

Table 4. MPIIGaze: Results with different redirection range

error $\Delta\mathbf{g}_p$ \ $\Delta\mathbf{g}_y$	$[-5, 5]$	$[-10, 10]$	$[-15, 15]$
$[-5, 5]$	4.15	4.06	4.02
$[-10, 10]$	4.10	4.03	**4.01**

sometimes even degrading the generic gaze estimator. The unsatisfactory performance of the latter models (*LinAdap* and *SVRAdap*) is probably due to the fact that the linear and SVR regressor do not make changes to the generic gaze estimator and thus the capacity of gaze adaptation is limited. We also find that the *DiffNet* is not always superior to the simpler *FTAdap* approach. This is surprising and shows that the ability of direct network fine tuning with small amount of data (less than 10) is often overlooked in the literature and not even unattempted. To the best of our knowledge, we are the first to report this result which can inspire new research on user-specific gaze estimation.

When comparing *RedFTAdap* with the best results of *DiffNet* and *FTAdap*, we note that our approach leads the performance by around $0.2°$. While this may seem a marginal improvement, a more detailed analysis of the results shows that our approach improves the results of **84.2%** of the subjects from the ColumbiaGaze dataset and of **80%** of the subjects from the MPIIGaze dataset (compared with the best results of *both DiffNet* and *FTAdap*), which means that the improvements brought by *RedFTAdap* are stable and rather systematic.

From the two tables, we note that the performances of all the methods improve as the number of reference samples increases. We can also notice that our approach seems to have a larger advantage when the number of reference samples is small, demonstrating that the diversity introduced by our redirected samples is more important when fewer person-specific gaze information is provided.

Finally, while in general adaptation methods improve results, we observe on the ColumbiaGaze dataset that they all perform worse than the generic estimator (cross-subject result) when using only one reference sample. This is most probably due to the large variance of the head pose in this dataset, which makes it difficult to learn (through adaptation) person-specific characteristics from only one sample.

Redirection range. We use different gaze redirection ranges to generate samples for gaze adaptation. The selected redirection ranges are shown in Table. 3 and Table. 4. Note that we only use one redirection range of pitch for the ColumbiaGaze dataset since the gaze groundtruth in this dataset is discrete and there are only three values for the pitch angle, $-10°, 0°, 10°$. It is thus not necessary to produce samples with new groundtruth. From the results, we find that larger redirection ranges do bring an improvement, especially for the MPIIGaze dataset where the performance improves from $4.15°$ to $4.01°$. This result is expected since a larger redirection range will usually bring more gaze diversity, provided that the redirection module produces synthesized samples realistic enough for the given user. Besides, we also find from Table. 4 that a larger redirection range for the yaw angle seems to be more effective than a larger redirection range for the pitch.

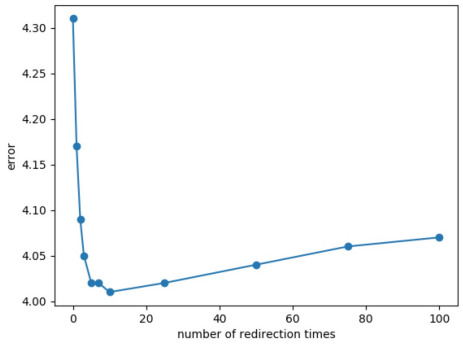


Figure 6. Gaze adaptation performances w.r.t redirection times t.

Table 5. Impact of the gaze redirection network domain adaptation (ColumbiaGaze dataset).

error #sample \ approach	*FTAdap*	*RedFTAdap-noDA*	*RedFTAdap*
1	5.53	4.35	**3.92**
5	3.11	3.01	**2.88**
9	2.79	2.73	**2.60**

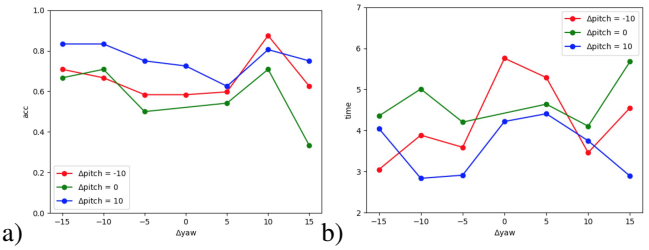
Number t of redirected gaze samples per reference sample. To study the impact of this parameter (the default value was 10 in all other experiments), we randomly selected 9 reference samples for each person and generated $9 \cdot t$ gaze redirected samples, varying t between 0 and 100. We then adapted the generic gaze estimator with these samples as in all other experiments. The corresponding performances are plotted in Fig. 6 for the MPIIGaze dataset (note that we do not use the ColumbiaGaze dataset since its groundtruth and redirection angles are discrete, which limits the number of generated data).

The curve in Fig. 6 starts from $t = 0$ (which means only the initial reference samples are used for adaptation). As can be seen, the error decreases rapidly at first when $t \in [0, 5]$, remains at a relatively stable point within the range $t \in [5, 25]$, and then progressively degrades beyond that. This curve shows that when $t \simeq 10$, the generated samples provide enough diversity to adapt the network, whereas beyond that, the use of too many samples results in an overfit of the network to the generated data which might not reflect the actual distribution of eye gaze appearence of the user.

Domain adaptation. We remove the whole **D**omain

Table 6. Impact of the gaze redirection network domain adaptation (MPIIGaze dataset).

error / approach / #sample	*FTAdap*	*RedFTAdap-noDA*	*RedFTAdap*
1	5.28	4.99	**4.97**
5	4.64	4.22	**4.20**
9	4.31	4.04	**4.01**



a) b)

Figure 7. Subjective test. (a) decision accuracy w.r.t redirection angles. (b) decision time w.r.t redirection angles.

Adaptation step from the redirection network and report the corresponding gaze adaptation results (*RedFTAdap-noDA*) in Table. 5 and Table. 6. On one hand, surprisingly, we note that exploiting the redirection network learned only from synthetic data still helps improving the gaze adaptation process (*FTAdap* vs *RedFTAdap-noDA*). On the other hand, when comparing *RedFTAdap-noDA* and *RedFTAdap*, we find that the domain adaptation further improves the gaze adaptation results. This is particularly the case for the ColumbiaGaze dataset. A possible reason why the domain adaptation is less usefull on the MPIIGaze dataset is that the domain difference between MPIIGaze and the synthetic data (all processed with histogram equalization to match MPIIGaze) is comparatively smaller.

Subjective test. To evaluate whether the gaze redirected samples are realistic, we invited 24 participants for a subjective test. During the test, participants were shown 50 pairs of ColumbiaGaze samples, where one image of the pair did correspond to an actual real data sample, and the second one was a gaze redirected sample. Note that as a result, the eyes in each image pair share the same identity, the same gaze and the same head pose. Some pairs are illustrated in Fig. 5c where the real images are all placed on the left for the purpose of demonstration. In the test, the places of the real and redirected images were selected at random. Participants were asked to choose the sample which they think was real. A software was recording their choices as well as the time they took to make the decisions.

Results are as follows. The average accuracy of making a correct choice is 66%, showing that distingusing genuine samples from redirected ones is difficult. This is further confirmed by the average time to reach a decision, which is around 4 seconds and shows that people have to take some time to make a careful decision.

We also plot the decision precision and the decision time w.r.t redirection angles in Fig. 7. From Fig. 7a, we find a general and expected trend that comparing samples with smaller redirection angles leads to more confusion, i.e. a low accuracy (and although as an artefact, the accuracy declines when $\Delta yaw = 15$). The same trend is observed in Fig. 7b, where a smaller redirection angle corresponds to a longer decision time. Nevertheless, in general, more participants and samples should be used to confirm these results, which we leave as a future work.

5. Discussion

In this section, we discuss techniques we attempted when developing the approach.

More realistic redirected samples. Ganin et al. [6] used a lighness correction refinement module on the gaze image redirected from the inverse warping field to produce a more realistic final redirected image. It indeed removed a lot of artifacts in our case. However, we found out that it was also degrading the performance of gaze adaptation, because the refinement through a set of convolutional layers was altering too much the distribution of color and illumination.

GAN. We also attempted to use GAN (or CycleGAN when combined with the cycle loss) for domain adaptation. However, as our redirected images are already of high quality, the GAN did not further improve the gaze adaptation step.

6. Conclusion

We proposed to improve the adaptation of a generic gaze estimator to a specific person from few shot samples via gaze redirection synthesis. To do so, we first designed a redirection network that was pretrained from large amounts of well aligned synthetic data, making it possible to predict accurate inverse warping fields. We then proposed a self-supervised method to adapt this model to real data. Finally, for the first time to the best of our knowledge, we exploited the gaze redirected samples to improve the performance of a person-specific gaze estimator. Along this way, as a minor contribution, we also showed that the simple fine tuning of a generic gaze estimation network using a very small amount of person-specific samples is very effective.

Notwithstanding the obtained improvements, a limitation of our method is that the redirection synthesis is not good enough for large redirection angles. It hinders further improvements of gaze adaptation because generated samples can not cover the full space of gaze directions and illumination conditions. We leave gaze redirection with larger angles and more illumination variabilities as future work.

Acknowledgement. This work was partly funded by the UBIMPRESSED project of the Sinergia interdisciplinary program of the Swiss National Science Foundation (SNSF), and by the the European Unions Horizon 2020 research and innovation programme under grant agreement no. 688147 (MuMMER, mummer-project.eu).

References

[1] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. Conversational gaze aversion for humanlike robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction (HRI)*, pages 25–32, New York, USA, 2014. 1

[2] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. *European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[3] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc 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 (ETRA)*, pages 255–258, 2014. 1

[4] Kenneth Alberto Funes Mora and Jean-Marc Odobez. Geometric Generative Gaze Estimation (G3E) for Remote RGB-D Cameras. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1773–1780, jun 2014. 2

[5] Kenneth A. Funes-Mora and Jean-Marc Odobez. Gaze Estimation in the 3D Space Using RGB-D Sensors. *International Journal of Computer Vision (IJCV)*, 118(2):194–216, jun 2016. 2

[6] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. DeepWarp: Photorealistic image resynthesis for gaze manipulation. *European Conference on Computer Vision (ECCV)*, pages 311–326, 2016. 2, 3, 8

[7] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32(3):478–500, 2010. 2

[8] Michael Xuelin Huang, Jiajia Li, Grace Ngai, and Hong Va Leong. Stressclick: Sensing stress from gaze-click patterns. In *Proceedings of the ACM on Multimedia Conference (ACMMM)*, pages 1395–1404, 2016. 1

[9] Qiong Huang, Ashok Veeraraghavan, and Ashutosh Sabharwal. Tabletgaze: unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications (MVAP)*, 2017. 1

[10] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. *ACM Transactions on Interactive Intelligent Systems (TIIS)*, 6(1):4:1–4:31, May 2016. 1

[11] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2017–2025. 2015. 3

[12] Daniil Kononenko, Yaroslav Ganin, Diana Sungatullina, and Victor S. Lempitsky. Photorealistic Monocular Gaze Redirection Using Machine Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1–15, 2017. 2, 3

[13] Daniil Kononenko and Victor Lempitsky. Learning to look up: Realtime monocular gaze correction using machine learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4667–4675, 2015. 3

[14] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, and Harini Kannan. Eye Tracking for Everyone. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2176–2184, 2016. 1, 2, 5

[15] Erik Lindén, Jonas Sjöstrand, and Alexandre Proutière. Appearance-based 3d gaze estimation with personal calibration. *CoRR*, abs/1807.00664, 2018. 2

[16] Gang Liu, Yu Yu, Kenneth Alberto Funes-Mora, and Jean-Marc Odobez. A Differential Approach for Gaze Estimation with Calibration. *British Machine Vision Conference (BMVC)*, 2018. 1, 2, 5, 6

[17] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Inferring human gaze from appearance via adaptive linear regression. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 153–160, 2011. 2

[18] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Adaptive Linear Regression for Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(10):2033–2046, 2014. 2

[19] David Masko. Calibration in Eye Tracking Using Transfer Learning. 2017. 2

[20] AJung Moon, Daniel M. Troniak, Brian Gleeson, Matthew K.X.J. Pan, Minhua Zheng, Benjamin A. Blumer, Karon MacLean, and Elizabeth A. Croft. Meet me where i'm gazing: How shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction (HRI)*, pages 334–341, New York, 2014. 1

[21] Nitish Padmanaban, Robert Konrad, Tal Stramer, Emily A Cooper, and Gordon Wetzstein. Optimizing virtual reality for all users through gaze-contingent and adaptive focus displays. *Proceedings of the National Academy of Sciences*, 2017. 1

[22] Cristina Palmero, Javier Selva, Mohammad Ali Bagheri, and Sergio Escalera. Recurrent CNN for 3D Gaze Estimation using Appearance and Shape Cues. *British Machine Vision Conference (BMVC)*, 2018. 2

[23] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep Pictorial Gaze Estimation. In *European Conference on Computer Vision (ECCV)*, pages 741–757. 2018. 1, 5

[24] Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics (TOG)*, 35(6):179, 2016. 1

[25] A. Pumarola, A. Agudo, A.M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 5

[26] Yalun Qin, Kuo-Chin Lien, Matthew Turk, and Tobias Höllerer. Eye Gaze Correction with a Single Webcam Based on Eye-Replacement. In *International Symposium on Visual Computing (ISVC)*, pages 599–609, 2015. 3

[27] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. Learning from simulated and unsupervised images through adversarial training.

In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, page 6, 2017. 1, 2

[28] Brian A Smith, Qi Yin, Steven K Feiner, and Shree K Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology (UIST)*, pages 271–280, 2013. 1, 5

[29] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3D gaze estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1821–1828, 2014. 2

[30] Marc Tonsen, Julian Steil, Yusuke Sugano, and Andreas Bulling. Invisibleeye: Mobile eye tracking using multiple low-resolution cameras and learning-based gaze estimation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 1(3), 2017. 1

[31] Kang Wang and Qiang Ji. Real Time Eye Gaze Tracking with Kinect. *International Conference on Computer Vision (ICCV)*, pages 1003–1011, 2017. 2

[32] Kang Wang and Qiang Ji. 3D gaze estimation without explicit personal calibration. *Pattern Recognition*, 79:216–227, 2018. 2

[33] K. Wang, Y. Wu, and Q. Ji. Head pose estimation on low-quality images. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2018. 2

[34] D. Weiner and N. Kiryati. Gaze redirection in face images. *IEEE Convention of Electrical and Electronics Engineers in Israel*, pages 78–80, 2002. 3

[35] Lior Wolf, Ziv Freund, and Shai Avidan. An eye for an eye: A single camera gaze-replacement method. Technical report, 2010. 3

[36] E Wood, T Baltruaitis, X Zhang, Y Sugano, P Robinson, and A Bulling. Rendering of Eyes for Eye-Shape Registration and Gaze Estimation. *IEEE International Conference on Computer Vision (ICCV)*, pages 3756–3764, 2015. 2

[37] Erroll Wood, Tadas Baltrušaitis, Louis Philippe Morency, Peter Robinson, and Andreas Bulling. A 3D morphable eye region model for gaze estimation. *European Conference on Computer Vision (ECCV)*, pages 297–313, 2016. 2

[38] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA)*, pages 131–138, 2016. 3

[39] Erroll Wood, Tadas Baltrušaitis, Louis Philippe Morency, Peter Robinson, and Andreas Bulling. Gazedirector: Fully articulated eye gaze redirection in video. *Eurographics*, pages 217–225, 2018. 3

[40] Erroll Wood and A Bulling. Eyetab: Model-based gaze estimation on unmodified tablet computers. *ACM Symposium on Eye Tracking Research & Applications (ETRA)*, pages 3–6, 2014. 2

[41] Yu Yu, Gang Liu, and Jean-Marc Odobez. Deep multitask gaze estimation with a constrained landmark-gaze model. *European Conference on Computer Vision Workshop (ECCVW)*, 2018. 2

[42] Xucong Zhang, Michael Xuelin Huang, Yusuke Sugano, and Andreas Bulling. Training Person-Specific Gaze Estimators from User Interactions with Multiple Devices. *ACM Conference on Human Factors in Computing Systems (CHI)*, 2018. 2

[43] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, jun 2015. 2, 5

[44] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016. 1

[45] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1–14, 2017. 1, 2, 5

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *International Conference on Computer Vision (ICCV)*, 2017. 5

[47] C Lawrence Zitnick, Jim Gemmell, and Kentaro Toyama. Manipulation of video eye gaze and head orientation for video teleconferencing. *Microsoft Research MSR-TR-99-46*, 1999. 3