RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments

Fischer, Tobias; Chang, Hyung Jin; Demiris, Yiannis

DOI:
10.1007/978-3-030-01267-0

RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments

Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris

Personal Robotics Laboratory, Department of Electrical and Electronic Engineering,
Imperial College London, UK
{t.fischer, hj.chang, y.demiris}@imperial.ac.uk

Abstract. In this work, we consider the problem of robust gaze estimation in natural environments. Large camera-to-subject distances and high variations in head pose and eye gaze angles are common in such environments. This leads to two main shortfalls in state-of-the-art methods for gaze estimation: hindered ground truth gaze annotation and diminished gaze estimation accuracy as image resolution decreases with distance. We first record a novel dataset of varied gaze and head pose images in a natural environment, addressing the issue of ground truth annotation by measuring head pose using a motion capture system and eye gaze using mobile eyetracking glasses. We apply semantic image inpainting to the area covered by the glasses to bridge the gap between training and testing images by removing the obtrusiveness of the glasses. We also present a new real-time algorithm involving appearance-based deep convolutional neural networks with increased capacity to cope with the diverse images in the new dataset. Experiments with this network architecture are conducted on a number of diverse eye-gaze datasets including our own, and in cross dataset evaluations. We demonstrate state-of-the-art performance in terms of estimation accuracy in all experiments, and the architecture performs well even on lower resolution images.

Keywords: Gaze estimation · Gaze dataset · Convolutional Neural Network · Semantic inpainting · Eyetracking glasses

1 Introduction

Eye gaze is an important functional component in various applications, as it indicates human attentiveness and can thus be used to study their intentions [9] and understand social interactions [41]. For these reasons, accurately estimating gaze is an active research topic in computer vision, with applications in affect analysis [22], saliency detection [42,48,49] and action recognition [31,36], to name a few. Gaze estimation has also been applied in domains other than computer vision, such as navigation for eye gaze controlled wheelchairs [12,46], detection of non-verbal behaviors of drivers [16,47], and inferring the object of interest in human-robot interactions [14].

Deep learning has shown successes in a variety of computer vision tasks, where their effectiveness is dependent on the size and diversity of the image

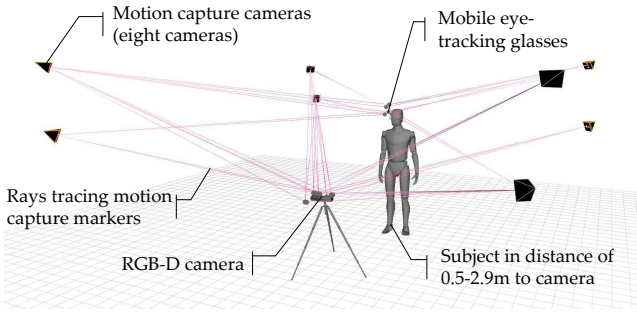

Fig. 1. Proposed setup for recording the gaze dataset. A RGB-D camera records a set of images of a subject wearing Pupil Labs mobile eyetracking glasses [24]. Markers that reflect infrared light are attached to both the camera and the eyetracking glasses, in order to be captured by motion capture cameras. The setup allows accurate head pose and eye gaze annotation in an automated manner.

dataset [29, 51]. However, in deep learning-based gaze estimation, relatively shallow networks are often found to be sufficient as most datasets are recorded in constrained scenarios where the subject is in close proximity to the camera and has a small movement range [15, 20, 28, 60]. In these datasets, ground truth data are typically annotated in an indirect manner by displaying a target on a screen and asking the subject to fixate on this target, with typical recording devices being mobile phones [28], tablets [20, 28], laptops [60], desktop screens [15], or TVs [10]. This is due to the difficulty of annotating gaze in scenarios where the subject is far from the camera and allowed to move freely.

To the best of our knowledge, this work is the first to address gaze estimation in natural settings with larger camera-subject distances and less constrained subject motion. In these settings, gaze was previously approximated only by the head pose [30, 35]. Our novel approach, RT-GENE, involves automatically annotating ground truth datasets by combining a motion capture system for head pose detection, with mobile eye tracking glasses for eye gaze annotation. As shown in Figure 1, this setup directly provides the gaze vector in an automated manner under free-viewing conditions (*i.e.* without specifying an explicit gaze target), which allows rapid recording of the dataset.

While our system provides accurate gaze annotations, the eyetracking glasses introduce the problem of unnatural subject appearance when recorded from an external camera. Since we are interested in estimating the gaze of subjects without the use of eyetracking glasses, it is important that the test images are not affected by an alteration of the subjects' appearance. For this purpose, we show that semantic image inpainting can be applied in a new scenario, namely the inpainting of the area covered by the eyetracking glasses. The images with removed eyetracking glasses are then used to train a new gaze estimation framework, as shown in Figure 2, and our experiments validate that the inpainting improves the gaze estimation accuracy. We show that networks with more depth cope

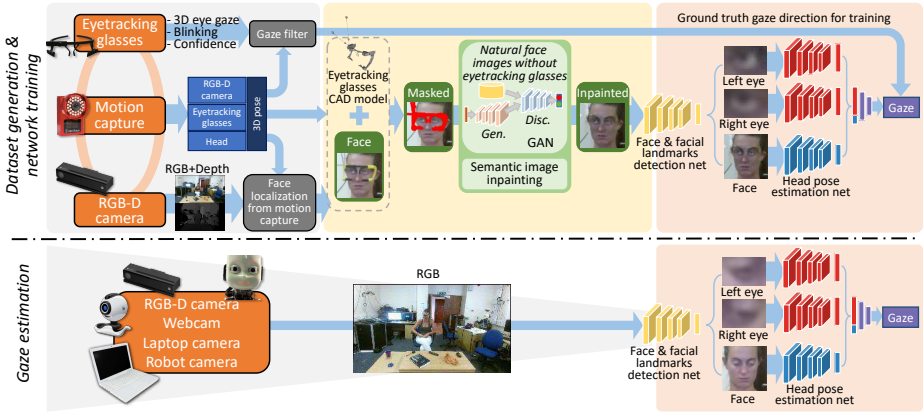

Fig. 2. RT-GENE Architecture overview. During training, a motion capture system is used to find the relative pose between mobile eyetracking glasses and a RGB-D camera (both equipped with motion capture markers), which provides the head pose of the subject. The eyetracking glasses provide labels for the eye gaze vector with respect to the head pose. A face image of the subject is extracted from the camera images, and a semantic image inpainting network is used to remove the eyetracking glasses. We use a landmark detection deep network to extract the positions of five facial landmarks, which are used to generate eye patch images. Finally, our proposed gaze estimation network is trained on the annotated gaze labels.

well with the large variations of appearance within our new dataset, while also outperforming state-of-the-art methods in traditional datasets[1].

2 Related Work

Gaze datasets: In Table 1, we compare a range of datasets commonly used for gaze estimation. In the Columbia Gaze dataset [52], subjects have their head placed on a chin rest and are asked to fixate on a dot displayed on a wall whilst their eye gaze is recorded. This setup leads to severely limited appearances: the camera-subject distance is kept constant and there are only a small number of possible head poses and gaze angles. UT Multi-view [53] contains recordings of subjects with multiple cameras, which makes it possible to synthesize additional training images using virtual cameras and a 3D face model. A similar setup was proposed by Deng and Zhu [10], who captured eye gaze data points at extreme angles by first displaying a head pose target, followed by an eye gaze target.

Recently, several datasets have been collected where subjects are asked to look at pre-defined targets on the screen of a mobile device, with the aim of introducing greater variation in lighting and appearance. Zhang *et al.* [60] presented the MPII Gaze dataset, where 20 target items were displayed on a laptop screen per session. One of the few gaze datasets collected using an RGB-D camera is Eyediap [15].

[1] Dataset and code are available to the public: www.imperial.ac.uk/PersonalRobotics.

Table 1. Comparison of gaze datasets

Dataset	RGB / RGB-D	Image type	Annotation type	#Images	Distance	Head pose annot.	Gaze annot.	Head pose orient.
CMU Multi-Pie [18]	RGB	Camera frame	68 Facial landmarks	755,370	≈300cm	✓	-	All
BIWI [13]	RGB-D	Camera frame	Head pose vector	≈15,500	100cm	✓	-	All
ICT 3D Head pose [2]	RGB-D	Camera frame	Head pose vector	14,000	≈100cm	✓	-	All
Deep Head Pose [38]	RGB-D	Camera frame	Head pose vector	68,000	≈200-800cm	✓	-	All
Vernissage [23]	RGB	(Robot) camera frame	Head pose vector	Unknown	Varying	✓	-	All
Coffeebreak [8]	RGB	Low res. face image	Head pose vector	18,117	Varying	✓	-	All
Eyediap [15]	RGB-D	Face + eye patches	Gaze vector	≈62,500	80-120cm	✓	✓	Frontal
MPII Gaze [60,61]	RGB	Face + eye patches	Gaze vector	213,659	40-60cm	✓	✓	Frontal
Columbia [52]	RGB	High res. camera image	Gaze vector	5,880	200cm	5 orient.	✓	Frontal
SynthesEyes [56]	RGB	Synthesized eye patches	Gaze vector	11,382	Varying	✓	✓	All
UnityEyes [55]	RGB	Synthesized eye patches	Gaze vector	1,000,000	Varying	✓	✓	All
UT Multi-view [53]	RGB	Eye area + eye patches	Gaze vector	1,152,000	60cm	✓	✓	All
Gaze Capture [28]	RGB	Face + eye patches	2D pos on screen	> 2.5M	Close	-	✓	Frontal
Rice TabletGaze [20]	RGB	Tablet camera video	2D pos on screen	≈100,000	30-50cm	-	✓	Frontal
Ours (RT-GENE)	**RGB-D**	**Face + eye patches**	**Gaze vector**	**122,531**	**80-280cm**	✓	✓	**All**

In addition to targets on a computer screen, the dataset contains a 3D floating target which is tracked using color and depth information. GazeCapture [28] is a crowd-sourced dataset of nearly 1500 subjects looking at gaze targets on a tablet screen. For the aforementioned datasets, the head pose is estimated using landmark positions of the subject and a (generic or subject specific) 3D head model. While these datasets are suitable for situations where a subject is directly facing a screen or mobile device, the distance between subject and camera is relatively small and the head pose is biased towards the screen. In comparison, datasets that capture accurate head pose annotations at larger distances typically do not contain eye gaze labels [2, 8, 13, 18, 23, 38].

Another way of obtaining annotated gaze data is to create synthetic image patches [32, 55–57], which allows arbitrary variations in head and eye poses as well as camera-subject distance. For example, Wood *et al.* [55] proposed a method to render photo-realistic images of the eye region in real-time. However, the domain gap between synthetic and real images makes it hard to apply these trained networks on real images. Shrivastana *et al.* [50] proposed to use a Generative Adversarial Network to refine the synthetic patches to resemble more realistic images, while ensuring that the gaze direction is not affected. However, the appearance and gaze diversity of the refined images is then limited to the variations found in the real images.

A dataset employing a motion capture system and eyetracking glasses was presented by McMurrough *et al.* [37]. It only contains the eye images provided by the eyetracking glasses, but does not contain images from an external camera. Furthermore, the gaze angles are limited as a screen is used to display the targets.

Deep learning-based gaze estimation: Several works apply Convolutional Neural Networks (CNN) for gaze estimation, as they have been shown to outperform conventional approaches [60], such as k-Nearest Neighbors or random forests. Zhang *et al.* [60] presented a shallow CNN with six layers that takes an eye image as input and fuses this with the head pose in the last fully connected layer of the network. Krafka *et al.* [28] introduced a CNN which estimates the gaze by combining the left eye, right eye and face images, with a face grid, providing

the network with information about the location and size of the head in the original image. A spatial weights CNN taking the full face image as input, *i.e.* without any eye patches, was presented in [61]. The spatial weights encode the importance of the different facial areas, achieving state-of-the-art performance on multiple datasets. Recently, Deng and Zhu [10] suggested a two-step training policy, where a head CNN and an eye CNN are trained separately and then jointly fine-tuned with a geometrically constrained "gaze transform layer".

3 Gaze Dataset Generation

One of the main challenges in appearance-based gaze estimation is accurately annotating the gaze of subjects with natural appearance while allowing free movements. We propose RT-GENE, a novel approach which allows the automatic annotation of subjects' ground truth gaze and head pose labels under free-viewing conditions and large camera-subject distances (overall setup shown in Figure 1). Our new dataset is collected following this approach. The dataset was constructed using mobile eyetracking glasses and a Kinect v2 RGB-D camera, both equipped with motion capture markers, in order to precisely find their poses relative to each other. The eye gaze of the subject is annotated using the eyetracking glasses, while the Kinect v2 is used as a recording device to provide RGB images at 1920x1080 resolution and depth images at 512x424 resolution. In contrast to the datasets presented in Table 1, our approach allows for accurate annotation of gaze data even when the subject is facing away from the camera.

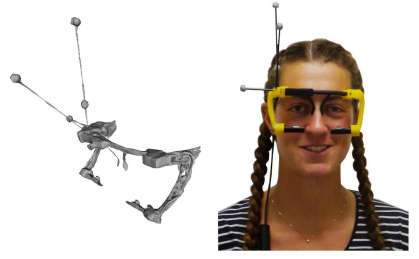

Fig. 3. Left: 3D model of the eyetracking glasses including the motion capture markers. Right: Eyetracking glasses worn by a subject. The 3D printed yellow parts have been designed to hold the eye cameras of the eyetracking glasses in the same place for each subject.

Eye gaze annotation: We use a customized version of the Pupil Labs eyetracking glasses [24], which have a very low average eye gaze error of 0.6 degrees in screen base settings. In our dataset with significantly larger distances, we obtain an angular accuracy of 2.58 ± 0.56 degrees. The headset consists of a frame with a scene camera facing away from the subject and a 3D printed holder for the eye cameras. This removes the need to adjust the eye camera placement for each subject. The customized glasses provide two crucial advantages over the original headset. Firstly, the eye cameras are mounted further from the subject, which leads to fewer occlusions of the eye area. Secondly, the fixed position of the holder allows the generation of a generic (as opposed to subject-specific) 3D model of the glasses, which is needed for the inpainting process, as described

in Section 4. The generic 3D model and glasses worn by a subject are shown in Figure 3.

Head pose annotation: We use a commercial OptiTrack motion capture system [39] to track the eyetracking glasses and the RGB-D camera using four markers attached to each object, with an average position error of 1mm for each marker. This allows to infer the pose of the eyetracking glasses with respect to the RGB-D camera, which is used to annotate the head pose as described below.

Coordinate transforms: The key challenge in our dataset collection setup was to relate the eye gaze $\mathbf{g}$ in the eyetracking reference frame $\mathbf{F}_{\mathrm{E}}$ with the visual frame of the RGB-D camera $\mathbf{F}_{\mathrm{C}}$ as expressed by the transform $\mathbf{T}_{\mathrm{E}\to\mathrm{C}}$. Using this transform, we can also define the head pose $\mathbf{h}$ as it coincides with $\mathbf{T}_{\mathrm{C}\to\mathrm{E}}$. However, we cannot directly use the transform $\mathbf{T}_{\mathrm{E}^*\to\mathrm{C}^*}$ provided by the motion capture system, as the frames perceived by the motion capture system, $\mathbf{F}_{\mathrm{E}^*}$ and $\mathbf{F}_{\mathrm{C}^*}$, do not match the visual frames, $\mathbf{F}_{\mathrm{E}}$ and $\mathbf{F}_{\mathrm{C}}$.

Therefore, we must find the transforms $\mathbf{T}_{\mathrm{C}\to\mathrm{C}^*}$ and $\mathbf{T}_{\mathrm{E}\to\mathrm{E}^*}$. To find $\mathbf{T}_{\mathrm{C}\to\mathrm{C}^*}$ we use the property of RGB-D cameras which allows to obtain 3D point coordinates of an object in the visual frame $\mathbf{F}_{\mathrm{C}}$. If we equip this object with markers tracked by the motion capture system, we can find the corresponding coordinates in the motion capture frame $\mathbf{F}_{\mathrm{C}^*}$. By collecting a sufficiently large number of samples, the Nelder-Mead method [40] can be used to find $\mathbf{T}_{\mathrm{C}\to\mathrm{C}^*}$. As we have a 3D model of the eyetracking glasses, we use the accelerated iterative closest point algorithm [6] to find the transform $\mathbf{T}_{\mathrm{E}\to\mathrm{E}^*}$ between the coordinates of the markers within the model and those found using the motion capture system.

Using the transforms $\mathbf{T}_{\mathrm{E}^*\to\mathrm{C}^*}$, $\mathbf{T}_{\mathrm{C}\to\mathrm{C}^*}$ and $\mathbf{T}_{\mathrm{E}\to\mathrm{E}^*}$ it is now possible to convert between any two coordinate frames. Most importantly, we can map the gaze vector $\mathbf{g}$ to the frame of the RGB-D camera using $\mathbf{T}_{\mathrm{E}\to\mathrm{C}}$.

Data collection procedure: At the beginning of the recording procedure, we calibrate the eyetracking glasses using a printed calibration marker, which is shown to the subject in multiple positions covering the subject's field of view while keeping the head fixed. Subsequently, in the first session, subjects are recorded for 10 minutes while wearing the eyetracking glasses. We instructed the subjects to behave naturally while varying their head poses and eye gazes as much as possible and moving within the motion capture area. In the second session, we record unlabeled images of the same subjects without the eyetracking glasses for another 10 minutes. These images are used for our proposed inpainting method as described in Section 4. To increase the variability of appearances for each subject, we change the 3D location of the RGB-D camera, the viewing angle towards the subject and the initial subject-camera distance.

Post-processing: We synchronize the recorded images of the RGB-D camera with the gaze data $\mathbf{g}$ of the eyetracking glasses in a post-processing step. We also filter the training data to only contain head poses $\mathbf{h}$ between ± 37.5 degrees horizontally and ± 30 degrees vertically, which allows accurate extraction of the images of both eyes. Furthermore, we filter out blinks and images where the pupil was not detected properly with a confidence threshold of 0.98 (see [24] for details).

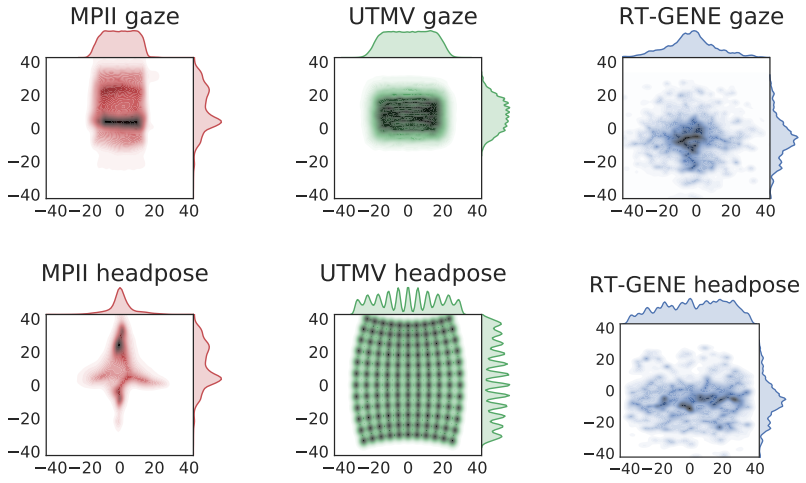

Fig. 4. Top row: Gaze distribution of the MPII Gaze dataset [60] (left), the UT Multi-view dataset [53] (middle) and our proposed RT-GENE dataset (right). Bottom row: Head pose distributions, as above. Our RT-GENE dataset covers a much wider range of gaze angles and head poses, which makes it more suitable for natural scenarios.

Dataset statistics: The proposed RT-GENE dataset contains recordings of 15 participants (9 male, 6 female, 2 participants recorded twice), with a total of 122,531 labeled training images and 154,755 unlabeled images of the same subjects where the eyetracking glasses are not worn. Figure 4 shows the head pose and gaze angle distribution across all subjects in comparison to other datasets. Compared to [53, 60], a much higher variation is demonstrated in the gaze angle distribution, primarily due to the novelty of the presented setup. The free-viewing task leads to a wider spread and resembles natural eye behavior, rather than that associated with mobile device interaction or screen viewing as in [15, 20, 28, 60]. Due to the synthesized images, the UT Multi-view dataset [53] also covers a wide range of head pose angles, however they are not continuous due to the fixed placing of the virtual cameras which are used to render the synthesized images.

The camera-subject distances range between $0.5m$ and $2.9m$, with a mean distance of $1.82m$ as shown in Figure 5. This compares to a fixed distance of $0.6m$ for the UT Multi-view dataset [53], and a very narrow distribution of $0.5m \pm 0.1m$ for the MPII Gaze dataset [60]. Furthermore, the area covered by the subjects' faces is much lower in our dataset (mean: 100×100 px) compared to other datasets (MPII Gaze dataset mean: 485×485 px). Thus compared to many other datasets, which focus on close distance scenarios [15, 20, 28, 53, 60], our dataset captures a more natural real-world setup. Our RT-GENE dataset is the first to provide accurate ground truth gaze annotations in these settings in addition to head pose estimates. This allows application in new scenarios, such as social interactions between multiple humans or humans and robots.

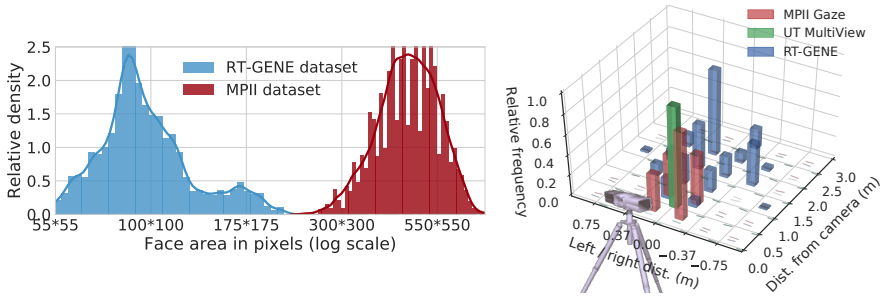

Fig. 5. Left: Face area distribution in the MPII [60] and our proposed RT-GENE datasets. The resolution of the face areas in our dataset is much lower (mean 100×100px) than that of the MPII dataset (mean 485×485px). This is mainly due to the larger camera-subject distance. Right: Distribution of camera-subject distances for various datasets [53,60]. RT-GENE covers significantly more varied camera-to-subject distances than the others, with distances being in the range between $0.5m$ and $2.9m$.

4 Removing Eyetracking Glasses

A disadvantage of using the eyetracking glasses is that they change the subject's appearance. However, when the gaze estimation framework is used in a natural setting, the subject will not be wearing the eyetracking glasses. We propose to semantically inpaint the regions covered by the eyetracking glasses, to remove any discrepancy between training and testing data.

Image inpainting is the process of filling target regions in images by considering the image semantics. Early approaches included diffusion-based texture synthesis methods [1,5,7], where the target area is filled by extending the surrounding textures in a coarse to fine manner. For larger regions, patch-based methods [4, 11,19,54] that take a semantic image patch from either the input image or an image database are more successful.

Recently, semantic inpainting has vastly improved in performance through the utilization of Generative Adversarial Network (GAN) architectures [21,44,58]. In this paper, we adopt this GAN-based image inpainting approach by considering both the textural similarity to the closely surrounding area and the image semantics. To the best of our knowledge, this is the first work using semantic inpainting to improve gaze estimation accuracy.

Masking eyetracking glasses region: The CAD model of the eyetracking glasses is made up of a set of $N = 2662$ vertices $\{\mathbf{v}_n\}_{n=1}^{N}$, with $\mathbf{v}_n \in \mathbb{R}^3$. To find the target region to be inpainted, we use $\mathbf{T}_{\mathrm{E}\to\mathrm{C}}$ to derive the 3D position of each vertex in the RGB-D camera frame. For extreme head poses, certain parts of the eyetracking glasses may be obscured by the subject's head, thus masking all pixels would result in part of the image being inpainted unnecessarily. To overcome this problem, we design an indicator function $\mathbf{1}_{\mathbf{M}}(\mathbf{p}_n, \mathbf{v}_n) = \{0 \text{ if } \|\mathbf{p}_n - \mathbf{v}_n\| < \tau, \text{else } 1\}$ which selects vertices $\mathbf{v}_n$ of the CAD model if they are within a tolerance τ of their corresponding point $\mathbf{p}_n$ in the depth field. Each selected vertex is mapped

Fig. 6. Image pairs showing the original images of the subject wearing the eyetracking glasses (left) and the corresponding inpainted images (right). The inpainted images look very similar to the subjects' appearance at testing time and are thus suited to train an appearance-based gazed estimator. Figure best viewed in color.

using the camera projection matrix of the RGB-D camera into a 2D image mask $\mathbf{M} = \{m_{i,j}\}$, where each entry $m_{i,j} \in \{0, 1\}$ shows whether the pixel at location (i, j) needs to be inpainted.

Semantic inpainting: To fill the masked regions of the eyetracking glasses, we use a GAN-based image generation approach, similar to that of Yeh *et al.* [58]. There are two conditions to fulfill [58]: the inpainted result should look realistic (perceptual loss $\mathcal{L}_{\text{perception}}$) and the inpainted pixels should be well-aligned with the surrounding pixels (contextual loss $\mathcal{L}_{\text{context}}$). As shown in Figure 5, the resolution of the face area is larger than the 64×64px supported in [58]. Our proposed architecture allows the inpainting of images with resolution 224×224px. This is a crucial feature as reducing the face image resolution for inpainting purposes could impact the gaze estimation accuracy.

We trained a separate inpainting network for each subject i. Let D_i denote a discriminator that takes as input an image $\mathbf{x}_i \in \mathbf{R}^d$ ($d = 224 \times 224 \times 3$) of subject i from the dataset where the eyetracking glasses are not worn, and outputs a scalar representing the probability of input $\mathbf{x}_i$ being a real sample. Let G_i denote the generator that takes as input a latent random variable $\mathbf{z}_i \in \mathbf{R}^z$ ($z = 100$) sampled from a uniform noise distribution $p_{\text{noise}} = \mathcal{U}(-1, 1)$ and outputs a synthesized image $G_i(\mathbf{z}_i) \in \mathbf{R}^d$. Ideally, $D_i(\mathbf{x}_i) = 1$ when $\mathbf{x}_i$ is from a real dataset p_i of subject i and $D_i(\mathbf{x}_i) = 0$ when $\mathbf{x}_i$ is generated from G_i. For the rest of the section, we omit subscript i for clarity.

We use a least squares loss [34], which has been shown to be more stable and better performing, while having less chance of mode collapsing [34, 62]. The training objective of the GAN is $\min_D \mathcal{L}_{GAN}(D) = \mathbf{E}_{\mathbf{x}\sim p}[(D(\mathbf{x}) - 1)^2] + \mathbf{E}_{\mathbf{z}\sim p_{\text{noise}}}[(D(G(\mathbf{z})))^2]$ and $\min_G \mathcal{L}_{GAN}(G) = \mathbf{E}_{\mathbf{z}\sim p_{\text{noise}}}[(D(G(\mathbf{z})) - 1)^2]$. In particular, $\mathcal{L}_{GAN}(G)$ measures the realism of images generated by G, which we consider as perceptual loss:

$$\mathcal{L}_{\text{perception}}(\mathbf{z}) = \big[D\big(G(\mathbf{z})\big) - 1\big]^2. \tag{1}$$

The contextual loss is measured based on the difference between the real image $\mathbf{x}$ and the generated image $G(\mathbf{z})$ of non-masked regions as follows:

$$\mathcal{L}_{\text{context}}(\mathbf{z}|\mathbf{M}, \mathbf{x}) = |\mathbf{M}' \odot \mathbf{x} - \mathbf{M}' \odot G(\mathbf{z})|, \tag{2}$$

where $\odot$ is the element-wise product and $\mathbf{M}'$ is the complement of $\mathbf{M}$ (*i.e.* to define the region that should not be inpainted).

The latent random variable $\mathbf{z}$ controls the images produced by $G(\mathbf{z})$. Thus, generating the best image for inpainting is equivalent to finding the best $\hat{\mathbf{z}}$ value which minimizes a combination of the perceptual and contextual losses:

$$\hat{\mathbf{z}} = \arg\min_{\mathbf{z}} \big(\lambda\, \mathcal{L}_{\text{perception}}(\mathbf{z}) + \mathcal{L}_{\text{context}}(\mathbf{z}|\mathbf{M}, \mathbf{x})\big) \tag{3}$$

where λ is a weighting parameter. After finding $\hat{\mathbf{z}}$, the inpainted image can be generated by:

$$\mathbf{x}_{\text{inpainted}} = \mathbf{M}' \odot \mathbf{x} + \mathbf{M} \odot G(\hat{\mathbf{z}}). \tag{4}$$

Poisson blending [45] is then applied to $\mathbf{x}_{\text{inpainted}}$ in order to generate the final inpainted images with seamless boundaries between inpainted and not inpainted regions. In Figure 6 we show the application of inpainting in our scenario.

Network architecture: We performed hyperparameter tuning to generate high resolution images of high quality. We set the generator with the architecture $\mathbf{z}$-dense(25088)-(256)5d2s-(128)5d2s-(64)5d2s-(32)5d2s-(3)5d2s-$\mathbf{x}$, where "(128)5c2s/(128)5d2s" denotes a convolution /deconvolution layer with 128 output feature maps and kernel size 5 with stride 2. All internal activations use SeLU [27] while the output layer uses tanh activation function. The discriminator architecture is $\mathbf{x}$-(16)5c2s-(32)5c2s-(64)5c2s-(128)5c2s-(256)5c2s-(512)5c2s-dense(1). We use LeakyReLU [33] with $\alpha = 0.2$ for all internal activations and a sigmoid activation for the output layer. We use the same architecture for all subjects.

Training hyperparameter details: To train G and D, we use the Adam optimizer [26] with learning rate 0.00005, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and batch size 128 for 100 epochs. We use the Xavier weight initialization [17] for all layers. To find $\hat{\mathbf{z}}$, we constrain all values in $\mathbf{z}$ to be within $[-1, 1]$, as suggested in [58], and we train for 1000 iterations. The weighting parameter λ is set to 0.1.

5 Gaze Estimation Networks

Overview: As shown in Figure 2, the gaze estimation is performed using several networks. Firstly, we use Multi-Task Cascaded Convolutional Networks (MTCNN) [59] to detect the face along with the landmark points of the eyes, nose and mouth corners. Using the extracted landmarks, we rotate and scale the face patch so that we minimize the distance between the aligned landmarks and predefined average face point positions to obtain a normalized face image using the accelerated iterative closest point algorithm [6]. We then extract the eye patches from the

normalized face images as fixed-size rectangles centered around the landmark points of the eyes. Secondly, we find the head pose of the subject by adopting the state-of-the-art method presented by Patacciola *et al.* [43].

Proposed eye gaze estimation: We then estimate the eye gaze vector using our proposed network. The eye patches are fed separately to VGG-16 networks [51] which perform feature extraction. Each VGG-16 network is followed by a fully connected (FC) layer of size 512 after the last max-pooling layer, followed by batch normalization and ReLU activation. We then concatenate these layers, resulting in a FC layer of size 1024. This layer is followed by another FC layer of size 512. We append the head pose vector to this FC layer, which is followed by two more FC layers of size 256 and 2 respectively[2]. The outputs of the last layer are the yaw and pitch eye gaze angles. For increased robustness, we use an ensemble scheme [29] where the mean of the predictions of the individual networks represents the overall prediction.

Image augmentation: To increase the robustness of the gaze estimator, we augment the training images in four ways. Firstly, to be robust against slightly off-centered eye patches due to imperfections in the landmark extraction, we perform 10 augmentations by cropping the image on the sides and subsequently resizing it back to its original size. Each side is cropped by a pixel value drawn independently from a uniform distribution $\mathcal{U}(0, 5)$. Secondly, for robustness against camera blur, we reduce the image resolution to 1/2 and 1/4 of its original resolution, followed by a bilinear interpolation to retrieve two augmented images of the original image size. Thirdly, to cover various lighting conditions, we employ histogram equalization. Finally, we convert color images to gray-scale images so that gray-scale images can be used as input as well.

Training details: As loss function, we use the sum of the individual l_2 losses between the predicted and ground truth gaze vectors. The weights for the network estimating the head pose are fixed and taken from a pre-trained model [43]. The weights of the VGG-16 models are initialized using a pre-trained model on ImageNet [51]. As we found that weight sharing results in decreased performance, we do not make use of it. The weights of the FC layers are initialized using the Xavier initialization [17]. We use the Adam optimizer [26] with learning rate 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.95$ and a batch size of 256.

6 Experiments

Dataset inpainting validation: We first conduct experiments to validate the effectiveness of our proposed inpainting algorithm. The average pixel error of five facial landmark points (eyes, nose and mouth corners) was compared to manually collected ground truth labels on a set of 100 images per subject before and after inpainting. The results reported in Table 2 confirm that all landmark estimation algorithms benefit from the inpainting, both in increased face detection rate and in lower pixel error ($p < .01$). The performance of our proposed inpainting

[2] All layer sizes were determined experimentally.

Table 2. Comparison of various landmark detectors [3,25] on the original images (with eyetracking glasses), images where the eyetracking glasses are filled with a uniform color (the mean color of the image), and inpainted images as proposed in our method. Both the face detection rate and the landmark error improve significantly when inpainted images are provided as input. The performance of MTCNN [59] is not reported, as it would be a biased comparison (MTCNN was used to extract the face patches).

.5Landmark detection method	Face detection rate (%)			Landmark error (pixel)		
	Original	Uniformly filled	Inpainted	Original	Uniformly filled	Inpainted
CLNF [3]	54.6±24.7	75.4±20.9	**87.7±15.6**	6.0±2.4	5.6±2.3	**5.3±1.8**
CLNF in-the-wild [3]	54.6±24.7	75.4±20.9	**87.7±15.6**	5.8±2.3	5.3±1.8	**5.2±1.6**
ERT [25]	36.7±25.3	59.7±23.0	**84.1±17.9**	6.6±2.3	5.8±1.7	**5.1±1.3**

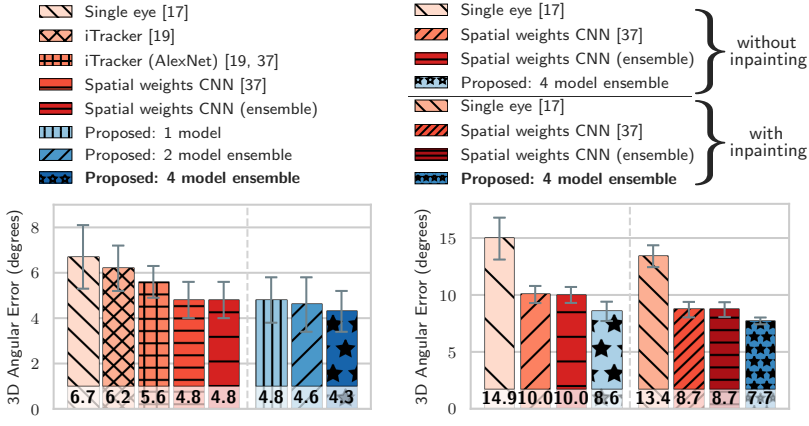

Fig. 7. Left: 3D gaze error on the MPII Gaze dataset. Right: 3D gaze error on our proposed gaze dataset. The inpainting improves the gaze estimation accuracy for all algorithms. Our proposed method performs best with an accuracy of 7.7 degrees.

method is also significantly higher than a method that naively fills the area of the eyetracking glasses uniformly with the mean color ($p < .01$). Importantly however, we found no statistical difference between the inpainted images and images where no eyetracking glasses are worn ($p = .16$).

Gaze estimation performance comparison: We evaluated our method on two de facto standard datasets, MPII Gaze [60] and UT Multi-view [53][3], as well as our newly proposed RT-GENE dataset.

First, we evaluate the performance of our proposed gaze estimation network on the MPII dataset [60]. The MPII dataset uses an evaluation set containing 1500 images of the left and right eye respectively. As our method employs both eyes as input, we directly use the 3000 images without taking the target eye

[3] We do not compare our method on the Eyediap dataset [15] and the dataset of Deng and Zhu [10] due to licensing restrictions of these datasets.

into consideration. The previous state-of-the-art achieves an error of 4.8 ± 0.7 degrees [61] in a leave-one-out setting. We achieve an increased performance of 4.3 ± 0.9 degrees using our method (10.4% improvement), as shown in Figure 7.

In evaluations on the UT Multi-view dataset [53], we achieve a mean error of 5.1 ± 0.2 degrees, outperforming the method of Zhang *et al.* [60] by 13.6% (5.9 degree error). This demonstrates that our proposed method achieves state-of-the-art performance on two existing datasets.

In a third set of experiments, we evaluate the performance on our newly proposed RT-GENE dataset using 3-fold cross validation as shown in Figure 7. All methods perform worse on our dataset compared to the MPII Gaze and UT Multi-view datasets, which is due to the natural setting with larger appearance variations and lower resolution images due to higher camera-subject distances. We confirm that using inpainted images at training time results in higher accuracy compared to using the original images without inpainting for all algorithms including our own (10.5% performance increase). For the inpainted images, our proposed gaze estimation network achieves the best performance with an error of 7.7 ± 0.3 degrees, which compares to [60] with an error of 13.4 ± 1.0 degrees (42.5% improvement) and the previous state-of-the-art network [61] with 8.7 ± 0.7 degrees error (11.5% improvement). These results demonstrate that features obtained using our deeper network architecture are more suitable for this dataset compared to the previous state-of-the-art.

Furthermore, ensemble schemes were found to be particularly effective in our architecture. For a fair comparison, we also applied the ensemble scheme to the state-of-the-art method [61]. However, we did not observe any performance improvement over the single model (see Figure 7). We assume that this is due to the spatial weights scheme that leads to similar weights in the intermediate layers of the different models. This results in similar gaze predictions of the individual models, and therefore an ensemble does not improve the accuracy for [61].

Cross-dataset evaluation: To further validate whether our dataset can be applied in a variety of settings, we trained our proposed ensemble network on

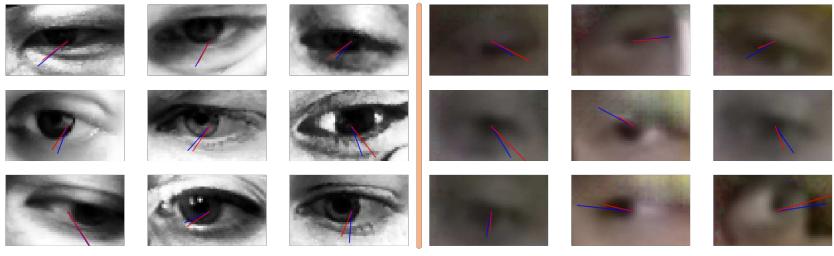

Fig. 8. Sample estimates (red) and ground truth annotations (blue) using our proposed method on the MPII Gaze dataset [60] (left) and our proposed dataset (right). Our dataset is more challenging, as images in our dataset are blurrier due to the higher subject-camera distance and show a higher variation in head pose and gaze angles. Figure best viewed in color.

samples from our RT-GENE dataset (all subjects included) and tested it on the MPII Gaze dataset [60]. This is challenging, as the face appearance and image resolution is very different as shown in Figures 5 and 8. We obtained an error of 7.7 degrees, which outperforms the current best performing method in a similar cross-dataset evaluation [55] (9.9 degrees error, 22.4% improvement). We also conduct an experiment where we train our ensemble network on UT Multi-view instead of RT-GENE as above, and again test the model on MPII Gaze. In this setting, we obtain an angular error of 8.9 degrees, which demonstrates the importance of our new dataset. We also outperform the method of [50] (7.9 degrees error), which uses unlabeled images of the MPII Gaze dataset at training time, while our method uses none.

Qualitative results: Some qualitative results of our proposed method applied to MPII Gaze and RT-GENE are displayed in Figure 8. Our framework can be used for real-time gaze estimation using any RGB or RGB-D camera such as Kinect, webcam and laptop camera, running at 25.3 fps with a latency of $0.12s$. This is demonstrated in the supplementary video. All comparisons are performed on an Intel i7-6900K with a Nvidia 1070 and 64GB RAM.

7 Conclusion and Future Work

Our approach introduces gaze estimation in natural scenarios where gaze was previously approximated by the head pose of the subject. We proposed RT-GENE, a novel approach for ground truth gaze estimation in these natural settings, and we collected a new challenging dataset using this approach. We demonstrated that the dataset covers a wider range of camera-subject distances, head poses and gazes compared to previous in-the-wild datasets. We have shown that semantic inpainting using GAN can be used to overcome the appearance alteration caused by the eyetracking glasses during training. The proposed method could be applied to bridge the gap between training and testing in settings where wearable sensors are attached to a human (*e.g.* EEG/EMG/IMU sensors). Our proposed deep convolutional network achieved state-of-the-art gaze estimation performance on the MPII Gaze dataset (10.4% improvement), UT Multi-view (13.6% improvement), our proposed dataset (11.5% improvement), and in cross dataset evaluation (22.4% improvement).

In future work, we will investigate gaze estimation in situations where the eyes of the participant cannot be seen by the camera, *e.g.* for extreme head poses or when the subject is facing away from the camera. As our dataset allows annotation of gaze even in these diverse conditions, it would be interesting to explore algorithms which can handle these challenging situations. We hypothesize that saliency information of the scene could prove useful in this context.

Acknowledgment: This work was supported in part by the Samsung Global Research Outreach program, and in part by the EU Horizon 2020 Project PAL (643783-RIA). We would like to thank Caterina Buizza, Antoine Cully, Joshua Elsdon and Mark Zolotas for their help with this work, and all subjects who volunteered for the dataset collection.

References

1. Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., Verdera, J.: Filling-in by joint interpolation of vector fields and gray levels. IEEE Transactions on Image Processing **10**(8), 1200–1211 (2001). https://doi.org/10.1109/83.935036
2. Baltrusaitis, T., Robinson, P., Morency, L.P.: 3D Constrained Local Model for Rigid and Non-Rigid Facial Tracking. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2610–2617 (2012). https://doi.org/10.1109/CVPR.2012.6247980
3. Baltrusaitis, T., Robinson, P., Morency, L.P.: Constrained local neural fields for robust facial landmark detection in the wild. In: IEEE International Conference on Computer Vision Workshops. pp. 354–361 (2013). https://doi.org/10.1109/ICCVW.2013.54
4. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Transactions on Graphics **28**(3), 24:1–24:11 (2009). https://doi.org/10.1145/1531326.1531330
5. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Annual Conference on Computer Graphics and Interactive Techniques. pp. 417–424. SIGGRAPH (2000). https://doi.org/10.1145/344779.344972
6. Besl, P.J., McKay, N.D.: A method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **14**(2), 239–256 (1992). https://doi.org/10.1109/34.121791
7. Chan, T.F., Shen, J.: Mathematical models for local nontexture inpaintings. SIAM Journal on Applied Mathematics **62**, 1019–1043 (2002). https://doi.org/10.1137/S0036139900368844
8. Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Bue, A.D., Menegaz, G., Murino, V.: Social interaction discovery by statistical analysis of F-formations. In: British Machine Vision Conference. pp. 23.1–23.12 (2011). https://doi.org/10.5244/C.25.23
9. Demiris, Y.: Prediction of intent in robotics and multi-agent systems. Cognitive Processing **8**(3), 151–158 (2007). https://doi.org/10.1007/s10339-007-0168-9
10. Deng, H., Zhu, W.: Monocular Free-Head 3D Gaze Tracking With Deep Learning and Geometry Constraints. In: IEEE International Conference on Computer Vision. pp. 3143–3152 (2017). https://doi.org/10.1109/ICCV.2017.341
11. Efros, A., Leung, T.: Texture synthesis by non-parametric sampling. In: International Conference on Computer Vision. pp. 1033–1038 (1999). https://doi.org/10.1109/ICCV.1999.790383
12. Eid, M.A., Giakoumidis, N., El-Saddik, A.: A Novel Eye-Gaze-Controlled Wheelchair System for Navigating Unknown Environments: Case Study With a Person With ALS. IEEE Access **4**, 558–573 (2016). https://doi.org/10.1109/ACCESS.2016.2520093
13. Fanelli, G., Weise, T., Gall, J., Gool, L.V.: Real Time Head Pose Estimation from Consumer Depth Cameras. In: Annual Symposium of the German Association for Pattern Recognition. pp. 101–110 (2011). https://doi.org/10.1007/978-3-642-23123-0
14. Fischer, T., Demiris, Y.: Markerless Perspective Taking for Humanoid Robots in Unconstrained Environments. In: IEEE International Conference on Robotics and Automation. pp. 3309–3316 (2016). https://doi.org/10.1109/ICRA.2016.7487504
15. Funes Mora, K.A., Monay, F., Odobez, J.M.: EYEDIAP: A Database for the Development and Evaluation of Gaze Estimation Algorithms from RGB and RGB-

D Cameras. In: ACM Symposium on Eye Tracking Research and Applications. pp. 255–258 (2014). https://doi.org/10.1145/2578153.2578190
16. Georgiou, T., Demiris, Y.: Adaptive user modelling in car racing games using behavioural and physiological data. User Modeling and User-Adapted Interaction **27**(2), 267–311 (2017). https://doi.org/10.1007/s11257-017-9192-3
17. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: International Conference on Artificial Intelligence and Statistics. pp. 249–256 (2010), http://proceedings.mlr.press/v9/glorot10a.html
18. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing **28**(5), 807–813 (2010). https://doi.org/10.1109/AFGR.2008.4813399
19. Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Transactions on Graphics **26**(3), 4:1–4:7 (2007). https://doi.org/10.1145/1276377.1276382
20. Huang, Q., Veeraraghavan, A., Sabharwal, A.: TabletGaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. Machine Vision and Applications **28**(5-6), 445–461 (2017). https://doi.org/10.1007/s00138-017-0852-4
21. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Transactions on Graphics **36**(4), 107:1–107:14 (2017). https://doi.org/10.1145/3072959.3073659
22. Jaques, N., Conati, C., Harley, J.M., Azevedo, R.: Predicting affect from gaze data during interaction with an intelligent tutoring system. In: International Conference on Intelligent Tutoring Systems. pp. 29–38 (2014). https://doi.org/10.1007/978-3-319-07221-0_4
23. Jayagopi, D.B., Sheiki, S., Klotz, D., Wienke, J., Odobez, J.M., Wrede, S., Khalidov, V., Nyugen, L., Wrede, B., Gatica-Perez, D.: The Vernissage Corpus: A Conversational Human-Robot-Interaction Dataset. In: ACM/IEEE International Conference on Human-Robot Interaction. pp. 149–150 (2013). https://doi.org/10.1109/HRI.2013.6483545
24. Kassner, M., Patera, W., Bulling, A.: Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction. In: ACM International Joint Conference on Pervasive and Ubiquitous Computing. pp. 1151–1160 (2014). https://doi.org/10.1145/2638728.2641695
25. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1867–1874 (2014). https://doi.org/10.1109/CVPR.2014.241
26. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: International Conference on Learning Representations (2015), https://arxiv.org/abs/1412.6980
27. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. In: Advances in Neural Information Processing Systems (2017), https://arxiv.org/abs/1706.02515
28. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye Tracking for Everyone. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2176–2184 (2016). https://doi.org/10.1109/CVPR.2016.239
29. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105 (2012). https://doi.org/10.1145/3065386
30. Lemaignan, S., Garcia, F., Jacq, A., Dillenbourg, P.: From real-time attention assessment to with-me-ness in human-robot interaction. In: ACM/IEEE

International Conference on Human Robot Interaction. pp. 157–164 (2016). https://doi.org/10.1109/HRI.2016.7451747
31. Liu, Y., Wu, Q., Tang, L., Shi, H.: Gaze-assisted multi-stream deep neural network for action recognition. IEEE Access **5**, 19432–19441 (2017). https://doi.org/10.1109/ACCESS.2017.2753830
32. Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Gaze Estimation From Eye Appearance: A Head Pose-Free Method via Eye Image Synthesis. IEEE Transactions on Image Processing **24**(11), 3680–3693 (2015). https://doi.org/10.1109/TIP.2015.2445295
33. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: International Conference on Machine Learning (2013), https://sites.google.com/site/deeplearningicml2013/relu_hybrid_icml2013_final.pdf
34. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: IEEE International Conference on Computer Vision. pp. 2794–2802 (2017). https://doi.org/10.1109/ICCV.2017.304
35. Massé, B., Ba, S., Horaud, R.: Tracking gaze and visual focus of attention of people involved in social interaction. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017). https://doi.org/10.1109/TPAMI.2017.2782819, to appear
36. Mathe, S., Sminchisescu, C.: Actions in the eye: dynamic gaze datasets and learnt saliency models for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(7), 1408–1424 (2015). https://doi.org/10.1109/TPAMI.2014.2366154
37. McMurrough, C.D., Metsis, V., Kosmopoulos, D., Maglogiannis, I., Makedon, F.: A dataset for point of gaze detection using head poses and eye images. Journal on Multimodal User Interfaces **7**(3), 207–215 (2013). https://doi.org/10.1007/s12193-013-0121-4
38. Mukherjee, S.S., Robertson, N.M.: Deep Head Pose: Gaze-Direction Estimation in Multimodal Video. IEEE Transactions on Multimedia **17**(11), 2094–2107 (2015). https://doi.org/10.1109/TMM.2015.2482819
39. NaturalPoint: OptiTrack Flex 3 http://optitrack.com/products/flex-3/, http://optitrack.com/products/flex-3/
40. Nelder, J.A., Mead, R.: A simplex method for function minimization. The Computer Journal **7**(4), 308–313 (1965)
41. Park, H.S., Jain, E., Sheikh, Y.: Predicting Primary Gaze Behavior Using Social Saliency Fields. In: IEEE International Conference on Computer Vision. pp. 3503–3510 (2013). https://doi.org/10.1109/ICCV.2013.435
42. Parks, D., Borji, A., Itti, L.: Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. Vision Research **116**, 113–126 (2015). https://doi.org/10.1016/j.visres.2014.10.027
43. Patacchiola, M., Cangelosi, A.: Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. Pattern Recognition **71**, 132–143 (2017). https://doi.org/10.1016/j.patcog.2017.06.009
44. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.: Context encoders: Feature learning by inpainting. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2536–2544 (2016). https://doi.org/10.1109/CVPR.2016.278
45. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. ACM Transactions on Graphics **22**(3), 313–318 (2003). https://doi.org/10.1145/882262.882269
46. Philips, G.R., Catellier, A.A., Barrett, S.F., Wright, C.: Electrooculogram wheelchair control. Biomedical sciences instrumentation **43**, 164–169 (2007), https://europepmc.org/abstract/med/17487075

47. Rasouli, A., Kotseruba, I., Tsotsos, J.K.: Agreeing to cross: How drivers and pedestrians communicate. In: IEEE Intelligent Vehicles Symposium. pp. 264–269 (2017). https://doi.org/10.1109/IVS.2017.7995730
48. Rudoy, D., Goldman, D.B., Shechtman, E., Zelnik-Manor, L.: Learning video saliency from human gaze using candidate selection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1147–1154 (2013). https://doi.org/10.1109/CVPR.2013.152
49. Shapovalova, N., Raptis, M., Sigal, L., Mori, G.: Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization. In: Advances in Neural Information Processing Systems. pp. 2409–2417 (2013), https://dl.acm.org/citation.cfm?id=2999881
50. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from Simulated and Unsupervised Images through Adversarial Training. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2107–2116 (2017). https://doi.org/10.1109/CVPR.2017.241
51. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015), https://arxiv.org/abs/1409.1556
52. Smith, B.A., Yin, Q., Feiner, S.K., Nayar, S.K.: Gaze Locking: Passive Eye Contact Detection for Human-Object Interaction. In: ACM Symposium on User Interface Software and Technology. pp. 271–280 (2013). https://doi.org/10.1145/2501988.2501994
53. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-Synthesis for Appearance-based 3D Gaze Estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1821–1828 (2014). https://doi.org/10.1109/CVPR.2014.235
54. Wilczkowiak, M., Brostow, G.J., Tordoff, B., Cipolla, R.: Hole filling through photomontage. In: British Machine Vision Conference. pp. 492–501 (2005), http://www.bmva.org/bmvc/2005/papers/55/paper.pdf
55. Wood, E., Baltrušaitis, T., Morency, L.P., Robinson, P., Bulling, A.: Learning an appearance-based gaze estimator from one million synthesised images. In: ACM Symposium on Eye Tracking Research & Applications. pp. 131–138 (2016). https://doi.org/10.1145/2857491.2857492
56. Wood, E., Baltrusaitis, T., Zhang, X., Sugano, Y., Robinson, P., Bulling, A.: Rendering of Eyes for Eye-Shape Registration and Gaze Estimation. In: IEEE International Conference on Computer Vision. pp. 3756–3764 (2015). https://doi.org/10.1109/ICCV.2015.428
57. Wood, E., Robinson, P., Bulling, A.: A 3D morphable eye region model for gaze estimation. In: European Conference on Computer Vision. pp. 297–313 (2016). https://doi.org/10.1007/978-3-319-46448-0_18
58. Yeh, R.A., Chen, C., Lim, T.Y., G., S.A., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with deep generative models. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5485–5493 (2017). https://doi.org/10.1109/CVPR.2017.728
59. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016). https://doi.org/10.1109/LSP.2016.2603342
60. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-Based Gaze Estimation in the Wild. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4511–4520 (2015). https://doi.org/10.1109/CVPR.2015.7299081

61. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 51–60 (2017). https://doi.org/10.1109/CVPRW.2017.284
62. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In: International Conference on Computer Vision. pp. 2223–2232 (2017). https://doi.org/10.1109/ICCV.2017.244