3DGazeNet: Generalizing 3D Gaze Estimation with Weak-Supervision from Synthetic Views Supplementary Material

Evangelos Ververas[1,2], Polydefkis Gkagkos[2], Jiankang Deng[1,2], Michail Christos Doukas[1], Jia Guo[3], and Stefanos Zafeiriou[1]

[1] Imperial College London, UK
[2] Huawei Noah's Ark Lab, UK
[3] InsightFace
https://ververas.github.io/3DGazeNet/

1 Implementation Details

Data Augmentation As model input we center crop 3 patches (left eye, right eye and face) resize them to shape $128 \times 128 \times 3$ and stack them channel-wise. Before stacking the image patches we randomly scale, translate, flip, and add noise on the color channel with probability 0.5. All geometric augmentations are also applied to the coordinates of eyes given as ground truth or pseudo-ground truth. Due to different image quality on each dataset we also add Gaussian blur to the images. The intensity of the blur is randomly selected from a set of kernels. Note that we augment each patch with the same augmentation parameters.

Training Details We train our method using a Adam optimizer (weight decay at 0.0005, and batch size of 128) on a single Tesla V100-PCIE (32GB) GPU. The learning rate starts from 1e-6, linearly warming up to 1e-4 in the first 3 epochs and then divided by 10 at 60 and 80 epochs. The training process is terminated at 100 epochs.

2D Iris Landmark Localization To localize 2D iris landmarks from images we started by using the model from [7]. Even though the available pre-trained models of [7] perform well on high resolution datasets, such as ETH-XGaze, they perform poorly on images with low resolution, such as the ones from Gaze360, AVA and CMU datsets. To overcome this problem we trained a version of our mesh-based model using predictions of [7] on high resolution images. In fact, we first applied [7] on ETH-XGaze and FFHQ datasets which resulted on 2D iris landmarks for these images. Then, using our pseudo-label generation pipeline (described in Sec. 3.2 of the paper, fig. 3(c)) we fit 3D eyeballs on these images. Next, we trained a 3D eyeball mesh reconstruction model, with the same architecture as the one described in the paper but without including the gaze head. Because of the data augmentation we applied during training (as described above), this model performs well on images with low resolution, occlusions and low-light conditions and gives us reliable 2D iris landmarks to either fit 3D eye meshes on existing gaze datasets using ground truth (fig. 3(b) of the paper) and

generate psuedo-labels (fig. 3(c) of the paper). Note that we only use this model to acquire 2D iris landmarks and do not care about its performance in gaze estimation.

Calculating Gaze Direction from 3D Eye Meshes Having recovered a 3D eyeball mesh with topology adhering to our 3D eyeball template, we calculate gaze from the orientation of the central axis of the eyeball. Particularly, we calculate 3D gaze vectors using the centre of the eyeball and the centre of the iris as shown in Fig. 1(c). After obtaining 3D gaze vectors from both left and right eyes, we add and normalize the two vectors to retrieve a mean eyeball-based gaze prediction, Fig. 1(d). Lastly, we add and normalize the mean eyeball-based gaze prediction with the gaze vector predicted by the gaze head of our model. This vector constitutes the final gaze output of our model.

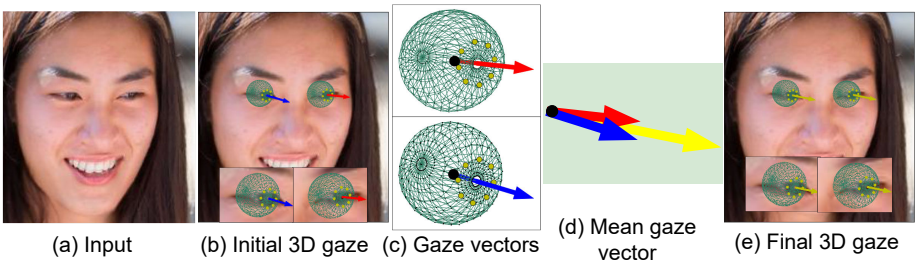


Fig. 1: Calculating 3D gaze from eye meshes. Given 3D eye meshes extracted by our method, we calculate gaze direction as the mean of the two independent gaze vectors from the left and right eyes.

Using the central axis of the 3D eye meshes to calculate the gaze direction is a reasonable approximation for our model. In practice, as seen in Fig. 2, an offset angle (the kappa coefficient) exists between the central (optical) and visual axes of eyes, which is subject dependent and varies between -2^o to 2^o across the population [15]. Even though accounting for this offset is crucial for person specific gaze estimation [2, 4, 5, 15], in the case of cross-dataset and in-the-wild gaze generalization, errors are much larger than the possible offset as can be seen from our all our cross-dataset experiments. Therefore, in that case data diversity is more important than anatomical precision. Besides, by combining the results of the eye mesh and gaze head of our model, we benefit both from the robustness offered from dense coordinate prediction and the accuracy of directly predicting gaze labels. This is supported by the results of Sec. 4.3 of the paper.

2 Application on Gaze Redirection

One of the main objectives of this work is to train gaze tracking models that generalize well to unseen domains and in-the-wild conditions. In this way models can be employed in a plug-n-play fashion by applications, offering reliable

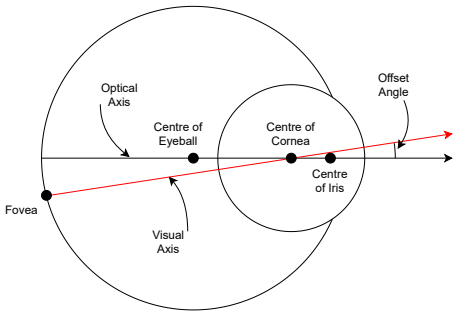


Fig. 2: Eyeball anatomy demonstrating the offset between the optical and visual axis.

gaze predictions without the need for knowledge of a specific target domain or parameter fine-tuning. As a practical use case, we test the effect of 3DGazeNet in the tasks of gaze redirection and gaze correction (gaze redirection to the camera, i.e. $(0^o, 0^o)$ pitch and yaw angles). To that end, we employ VFHQ [13] a high resolution video dataset of talking faces, initially built for video-based face super-resolution, which contains 16K clips and a wide variation of environments, illumination conditions and subject nationalities.

We design the gaze redirection experiment as follows. First, we split VFHQ in a training and a test set containing 12K and 4K videos each, and sample a total of 200K and 20K frames from each subset. Then, we extract gaze labels using 3DGazeNet trained on all datasets (MPII, GC, EXG, G360, ITWG-MV) and employ them for training an image translation network as in [3]. For clarity in comparisons we name the image translation network GazeRet-ITW. To evaluate the above model we generate 5 images with random gaze labels in the range $[-50^o, 50^o]$ yaw and $[-30^o, 30^o]$ pitch from each image of the test set. We additionally generate an image with gaze direction $(0^o, 0^o)$. We predict the gaze of the translated images using the initial gaze estimation model and measure the redirection error between the predictions and the known target labels.

To highlight the benefits of 3DGazeNet in the above task, we repeat the experiment using VFHQ gaze labels extracted from a model trained with all gaze datasets except for ITWG-MV (thus, using only ground truth supervision) which we name 3DGazeNet-GT. Also, we name this version of the image translation network as GazeRet-GT. Tab. 1 presents the redirection errors in the above two cases, as well as in cross-evaluation scenarios, for both redirection and correction. In all cases, images produced with GazeRet-ITW lead to lower errors showcasing the benefits of augmenting gaze datasets with ITWG-MV and weakly-supervising training with our multi-view consistency loss. This is due to the fact that 3DGazeNet can produce reliable gaze labels in the unseen domains of the VFHQ video clips, while the simple gaze tracking model cannot. Images with manipulated gaze direction using GazeRet-ITW can be seen in Fig. 3.

Implementation details For image-to-image translation we adapt the model architecture from StarGANv2 [1], removing the style encoder and the latent-to-style mapping network, and replacing the style codes with gaze vectors as conditions to the generator. We also adjust the discriminator to predict continu-

ous gaze labels as in [8,11], inducing precise gaze in image translation. Similarly to the above models, we operate using a cyclic reconstruction loss and a gaze prediction loss, without image pairs. As model input, we crop image patches of size 256×512 containing both eyes and perform image translation on them. We sample target gaze labels (generator conditioning labels) from the training set, as well as randomly. During testing, we blend the full face input images with the output patches. Lastly, we train the model with learning rate 0.0001 and batch size 64, using Adam optimizer for a total of 300K steps.

Table 1: Gaze Redirection error on images of our test split of VFHQ. Images generated with the GazeRet-ITW method lead to lower errors when evaluated by either of the gaze estimation models for both redirection and correction. The gaze errors are in degrees (lower is better).

	Redirection		Correction	
Method	3DGazeNet-GT	3DGazeNet	3DGazeNet-GT	3DGazeNet
GazeRet-GT	9.4	10.1	7.1	7.6
GazeRet-ITW	**6.7**	**8.4**	**3.8**	**5.2**

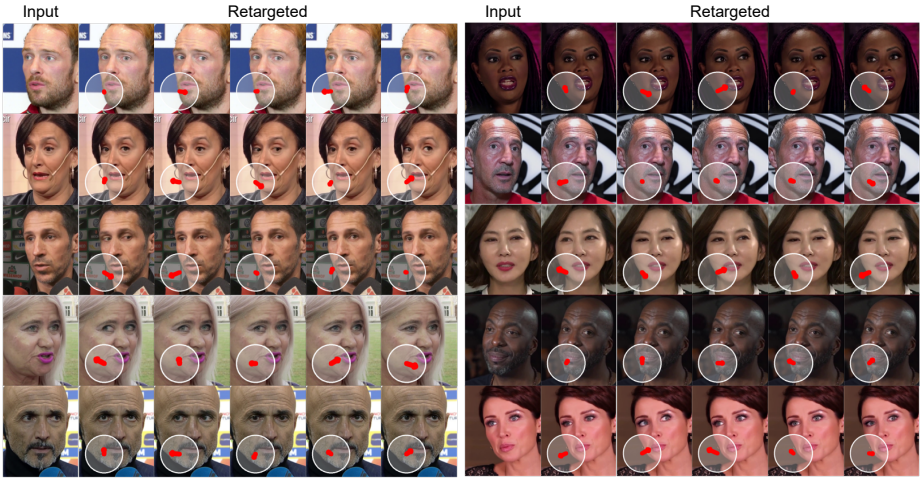


Fig. 3: Gaze redirection on images of VFHQ. The redirected images have been produced by GazeRet-ITW. The red arrow at the bottom left of each redirected image indicates the target gaze label used to generate that particular image.

3 Comparison with Model-Based Methods

As 3DGazeNet infers gaze based on 3D models of the eyes, it is sensible to mention related methods that use shape representations of the eyes. In particular, CrtCLGM+MTL [14] learns to predict gaze simultaneously with sparse 2D landmarks of the eye region and find their multi-task objective to be beneficial in comparison to just predicting gaze. CrtCLGM+MTL achieves 5.7^o error on the within-dataset experiment on UTMV dataset [10] (which includes 64K images and 50 subjects), while our method trained just with ground truth meshes and gaze vectors achieves 5.5^o, following the same experimental settings as [14]. Another related method, DPG [6], learns 2D segmentation maps of the eyeball and iris and employ those to infer gaze. DPG achieves error 4.5^o on MPII and 3.6^o on Columbia [9] (5880 images and 56 subjects) on within-dataset experiments, while our method achieves 4.0^o and 3.1^o respectively.

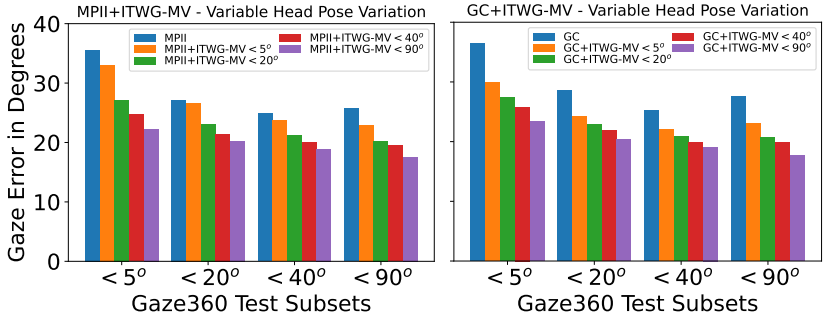
Other than the above methods which regress gaze as an output of a network, [7] and [12] infer gaze using reconstructions of the 3D eyeball, similarly to our method. In particular, [7] infers 2D landmarks of the eye region as well as the eyeball center and radius, which allows the reconstruction of a 3D eyeball and thus the prediction of a 3D gaze vector. [7] is trained on rendered images using a 3D model of the eye region and achieves error 7.1^o on the Columbia dataset without any annotation from it. Our method trained only with ITWG achieves error 5.6^o on the same dataset. Moreover, [12] have proposed a parametric model fitting approach in which they fit a 3D morphable model of the eye region and eyeball shape, texture and illumination and infer gaze based on the 3D eyeballs. Their model-fitting approach has given 7.5^o error on the Columbia dataset.

4 Additional Ablation on the effect of ITWG's Head Pose Variation

In this section we present further analysis on the effect of head pose variation of ITWG on the model's generalization ability. For training we employ different subsets of ITWG based on head pose ($< 5^o$, $< 20^o$, $< 40^o$, $< 90^o$(all)) and additional ground truth supervision from MPIIFaceGaze (MPII) or GazeCapture (GC). We present results of our models for different subsets of G360's test set, based again on the head pose yaw angle ($< 5^o$, $< 20^o$, $< 40^o$, $< 90^o$). Results reported in Tab. 2 and Fig. 4 demonstrate that an improvement of 6^o to 13^o (24% to 37%) is achieved in all cases (for all subsets of G360) between the baselines of training just with gaze datasets (MPII, GC) and our full method of including ITWG and multi-view supervision in training. It is also worth noticing that performance consistently increases when training with more diverse subsets of ITWG in all cases. Lastly, it is worth noticing that even though performance increase is expected for the subsets of G360 with large head pose values ($> 40^o$), as MPII and GC do not include such images at all, the larger increase in performance is seen for near-frontal images. This fact, validates the effectiveness of our pseudo-labelling method and our multi-view supervision algorithm.

Table 2: The effect of head pose variation of ITWG. Starting from either MPII or GC ground truth datasets, which include much smaller head pose and gaze variation than G360, incorporating data from ITWG with increasingly more diverse head pose, leads to lower gaze error (measured in degrees, lower is better). The error decreases significantly even in the case of G360 $< 5^o$ which indicates that the pseudo-labels of ITWG are meaningful and its face and environmental variation useful.

	D=MPII				D=GC			
Training Datasets	Gaze360 Test Subsets				Gaze360 Test Subsets			
	$< 5^o$	$< 20^o$	$< 40^o$	$< 90^o$	$< 5^o$	$< 20^o$	$< 40^o$	$< 90^o$
D	35.6	27.2	24.9	25.7	36.4	28.5	25.2	27.5
D+ITWG-MV$< 5^o$	33.0	26.6	23.7	22.9	29.9	24.2	22.1	23.1
D+ITWG-MV$< 20^o$	27.2	23.1	21.2	20.3	27.4	22.8	20.8	20.7
D+ITWG-MV$< 40^o$	24.7	21.4	20.0	19.5	25.8	21.8	19.9	19.8
D+ITWG-MV all	**22.3**	**20.2**	**18.9**	**17.6**	**23.3**	**20.4**	**19.1**	**17.6**



(a) MPII (b) GC

Fig. 4: Bar plots showing the effect of head pose variation of ITWG in cross-dataset experiments for the case of training with ground truth from (a) MPII and (b) GC and testing on G360. The data are taken from Tab. 2.

5 Additional Ablation on the effect of Pseudo-Labels and Multi-View Supervision

In this section we present further evaluations on the effect of our pseudo-labels and multi-view consistency constraints during training. To that end, we repeat the experiments described in Sec. 4.3 of the main paper for 3 additional cases of ground truth supervision. In particular, we utilize G360, EXG and MPII as source datasets with valid ground truth. From the results we can draw the conclusion that our method is always effective when there are large differences between the source and target dataset head pose and gaze variation. In such cases ITWG helps to close the gap and results in significant improvement. Only small improvement is noticed for the within-dataset experiment on Gaze360, while results do not improve for within-dataset experiments on EXG and MPII. EXG has been captured in lab conditions and already includes a wide variety of head poses and gaze directions, thus augmenting our model's training with

more in-the-wild data from ITWG does not benefit within dataset evaluation. Similarly for MPII, which is very restricted in terms of environments and pose variation, the best performance is achieved with ground truth supervision only. Additional results are presented in Tab. 3.

Table 3: The effect of incorporating pseudo-ground truth and multi-view supervision during training. Both components contribute towards improving results in cross-dataset gaze estimation experiments. Gaze error is in degrees (lower is better).

			D=G360				D=EXG				D=MPII			
Dataset	$\mathcal{L}_{PGT}$	$\mathcal{L}_{MV}$	G360	GC	EXG	MPII	G360	GC	EXG	MPII	G360	GC	EXG	MPII
D	-	-	9.6	12.1	18.3	9.1	22.1	10.7	**4.3**	7.7	23.6	**6.3**	26.3	**4.0**
D+ITWG	✓	-	9.4	10.2	16.4	8.2	19.4	11.1	4.7	6.8	19.8	7.2	21.7	4.4
D+ITWG-MV	-	✓	9.6	11.7	18.1	9.1	21.6	12.0	4.5	7.4	22.9	7.4	25.1	4.2
D+ITWG-MV	✓	✓	**9.3**	**8.0**	**14.6**	**6.3**	**15.4**	**7.8**	**4.3**	**6.0**	**17.6**	6.8	**14.9**	4.2

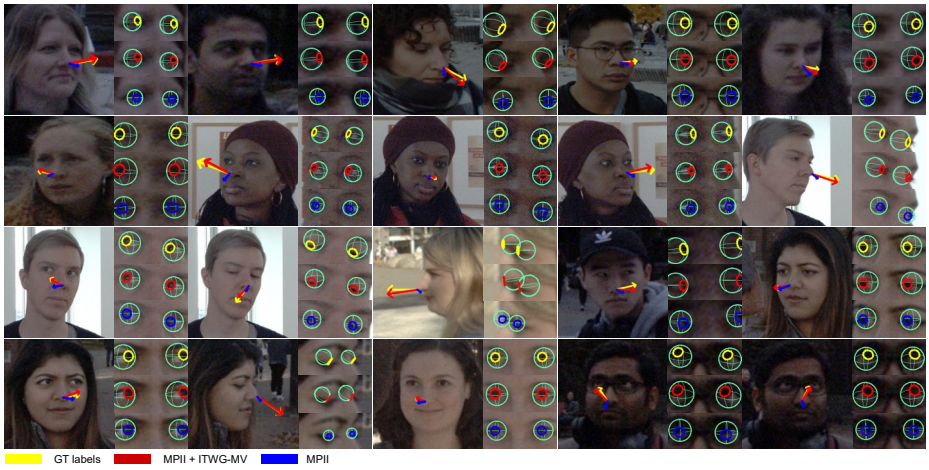


Fig. 5: Results of our models trained with MPII (blue vectors) and combined MPII and ITWG with multi-view supervision (red vectors), applied on the test set of G360 (yellow vectors). The predicted gaze directions are closer to the ground truth when ITWG is included in training. Especially for side and profile views, the effect of the pseudo-labels is significant.

6 Qualitative Results for 3D Gaze Estimation

Here we visualize gaze predictions of our model for training scenarios discussed in Sec. 4.3 of the main paper (The Effect of Head Pose Distribution of ITWG).

In particular, we present the results of our model in the two edge cases of Fig. 6 of the paper, i.e. a) only MPII is employed for training and b) MPII and the whole ITWG with multi-view supervision. Testing is performed on the images of G360. Fig. 5 includes results for the two cases as well as the ground truth labels of G360. As can be seen, the predicted gaze directions are much closer to the real ones when in-the-wild face data from ITWG are employed for supervision. Especially for profile views, the effect of the pseudo-labels is significant.

We also apply the above two models on in-the-wild face images (from VFHQ and AFLW) and present the results on Fig. 7 and Fig. 6. As actual gaze accuracy cannot be measured for such images (ground truth data are not available), we attempt to draw conclusions based on observation. From the visualizations it can be seen that for side and profile views, our multi-view supervision method (MPII + ITWG-MV) performs substantially better, while for near-frontal ones the predictions improve. Especially, in Fig. 6 we have included cases which we consider difficult for gaze estimation. These include images in profile views, images with occluded eyes or eye glasses as well as images with low resolution and bad illumination. Results demonstrate that 3DGazeNet can produce reliable gaze labels for all of the above cases, which makes it ideal for real applications operating in unrestricted environments.

Fig. 6: Results from applying our model on difficult cases including faces in profile pose, faces with glasses and faces with occlusions or low resolution. Our model can successfully handle these difficult scenarios and produce reliable gaze predictions.

Fig. 7: Results of our models trained with MPII (blue vectors) and combined MPII and ITWG with multi-view supervision (red vectors), applied on images of VFHQ. Our full model predicts robust gaze labels across all head pose angles, especially for profile ones the effect of the pseudo-labels is significant.

References

1. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: CVPR. pp. 8188–8197 (2020)
2. He, J., Pham, K., Valliappan, N., Xu, P., Roberts, C., Lagun, D., Navalpakkam, V.: On-device few-shot personalization for real-time gaze estimation. In: ICCV Workshops (2019)
3. He, Z., Spurr, A., Zhang, X., Hilliges, O.: Photo-realistic monocular gaze redirection using generative adversarial networks. In: ICCV. pp. 6932–6941 (2019)
4. Liu, G., Yu, Y., Mora, K., Odobez, J.: A differential approach for gaze estimation with calibration. In: BMVC (2018)
5. Park, S., Mello, S.D., Molchanov, P., Iqbal, U., Hilliges, O., Kautz, J.: Few-shot adaptive gaze estimation. In: ICCV (2019)
6. Park, S., Spurr, A., Hilliges, O.: Deep pictorial gaze estimation. In: ECCV (2018)
7. Park, S., Zhang, X., Bulling, A., Hilliges, O.: Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In: ACM ETRA (2018)
8. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: Anatomically-aware facial animation from a single image. In: ECCV. pp. 818–833 (2018)
9. Smith, B., Yin, Q., Feiner, S., Nayar, S.: Gaze Locking: Passive Eye Contact Detection for Human?Object Interaction. In: ACM UIST (2013)
10. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based 3d gaze estimation. In: CVPR (2014)
11. Ververas, E., Zafeiriou, S.: Slidergan: Synthesizing expressive face images by sliding 3d blendshape parameters. IJCV **128**(10-11), 2629–2650 (2020)
12. Wood, E., Baltrušaitis, T., Morency, L.P., Robinson, P., Bulling, A.: A 3d morphable eye region model for gaze estimation. In: ECCV (2016)

13. Xie, L., Wang, X., Zhang, H., Dong, C., Shan, Y.: Vfhq: A high-quality dataset and benchmark for video face super-resolution. In: CVPR. pp. 657–666 (2022)
14. Yu, Y., Liu, G., Odobez, J.M.: Deep multitask gaze estimation with a constrained landmark-gaze model. In: ECCV Workshops (2018)
15. Yu, Y., Liu, G., Odobez, J.M.: Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In: CVPR (2019)