Supplementary Material: CUDA-GHR: Controllable Unsupervised Domain Adaptation for Gaze and Head Redirection

Swati Jindal, Xin Eric Wang
University of California, Santa Cruz
{swjindal, xwang366}@ucsc.edu

1. Data Pre-processing

We follow the same data pre-processing pipeline as done in Park *et al*. [12]. The pipeline consists of a normalization technique [15] initially introduced by Sugano *et al*. [13]. It is followed by face detection [4] and facial landmarks detection [2] modules for which open-source implementations are publicly available. The Surrey Face Model [7] is used as a reference 3D face model. Further details can be found in Park *et al*. [12]. To summarize, we use the public code[1] provided by Park *et al*. [12] to produce image patches of size 256×64 containing both eyes.

2. Architecture Details

Our framework CUDA-GHR. We use DenseNet architecture [5] to implement image encoder $\mathbf{E}_a$. The DenseNet is formed with a growth rate of 32, 4 dense blocks (each with four composite layers), and a compression factor of 1. We use instance normalization [14] and leaky ReLU activation function (α = 0.01) for all layers in the network. We remove dropout and 1×1 convolution layers. The dimension of latent factor z^a is set to be equal to 16. Thus, to project CNN features to the latent features, we use global-average pooling and pass through a fully-connected layer to output 16-dimensional feature from $\mathbf{E}_a$. The gaze encoder $\mathbf{E}_g$ is a MLP-based block whose architecture is shown in Table 1. The dimension of z^g is set as 8.

For the generator network $\mathbf{G}$, we use HoloGAN [11] architecture shown in Table 5. The latent vector z for each AdaIN [6] input is processed by a 1-layer MLP, and the rotation layer is the same as the one used in the original paper [11]. The latent domain discriminator $\mathbf{D}_F$ consists of 4 MLP layers as shown in Table 2. It takes the input of dimension 24 and gives 1-dimensional output. Both image discriminators $\mathbf{D}_T$ and $\mathbf{D}_S$ are PatchGAN [9] based networks as used in Zheng *et al*. [16]. The architecture of the discriminator is described in Table 4.

[1] https://github.com/swook/faze_preprocess

Table 1: Architecture of gaze encoder $\mathbf{E}_g$

Layer name	Activation	Output shape
Fully connected	LeakyReLU (α = 0.01)	2
Fully connected	LeakyReLU (α = 0.01)	2
Fully connected	LeakyReLU (α = 0.01)	2
Fully connected	None	8

The task network $\mathcal{T}$ is a ResNet-50 [3] model with batch normalization [8] replaced by instance normalization [14] layers. It takes an input of 256×64 and gives a 4-dimensional output describing pitch and yaw angles for gaze and head directions. It is initialized with ImageNet [1] pre-trained weights and is fine-tuned on the GazeCapture training subset for around 190K iterations. The GazeCapture validation subset is used to select the best-performing model. The initial learning rate is 0.0016, decayed by a factor of 0.8 after about 34K iterations. Adam [10] optimizer is used for optimization with a weight decay coefficient of 10^{-4}. The architecture of $\mathcal{T}$ is summarized in Table 3.

Downstream Tasks. For gaze and head pose estimation, we use similar architecture as employed for $\mathcal{T}$ shown in Table 3. For all the experiments, the initial learning rate is 0.0001 decayed by a factor of 0.5 after every 1500 iterations. The pre-trained models are trained for 10 epochs with a batch size of 64 while fine-tuning is done for 5 epochs with a batch size of 32.

State-of-the-art Baselines. We re-implement the ST-ED [16] on images containing both eyes for a fair comparison with our method using the public code[2] available. We use the same hyperparameters as provided by the original implementation. For the accurate comparison, we replaced *tanh* non-linearity with an identity function and removed a constant factor of 0.5π in all the modules.

[2] https://github.com/zhengyuf/STED-gaze

Table 2: Architecture of latent domain discriminator $\mathbf{D}_F$

Layer name	Activation	Output shape
Fully connected	LeakyReLU ($\alpha = 0.01$)	24
Fully connected	LeakyReLU ($\alpha = 0.01$)	24
Fully connected	LeakyReLU ($\alpha = 0.01$)	24
Fully connected	None	1

Table 3: Architecture of the task network $\mathcal{T}$

Module/Layer name	Output shape
ResNet-50 layers with MaxPool stride=1	$2048 \times 1 \times 1$
Fully connected	4

3. Additional Results

In Figures 1 and 2, we show additional qualitative results for both target datasets, namely, MPIIGaze and Columbia. Figure 1a and 2a represent gaze redirected images and Figure 1b and 2b show head redirected images.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[2] Jiankang Deng, Yuxiang Zhou, Shiyang Cheng, and Stefanos Zaferiou. Cascade multi-view hourglass model for robust 3d face alignment. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 399–403. IEEE, 2018. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[4] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 951–959, 2017. 1

[5] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1

[6] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 1

[7] Patrik Huber, Guosheng Hu, Rafael Tena, Pouria Mortazavian, P Koppen, William J Christmas, Matthias Ratsch, and Josef Kittler. A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016. 1

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 1

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[11] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 1

[12] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9368–9377, 2019. 1, 4, 5, 6, 7

[13] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1821–1828, 2014. 1

[14] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 1

[15] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–9, 2018. 1

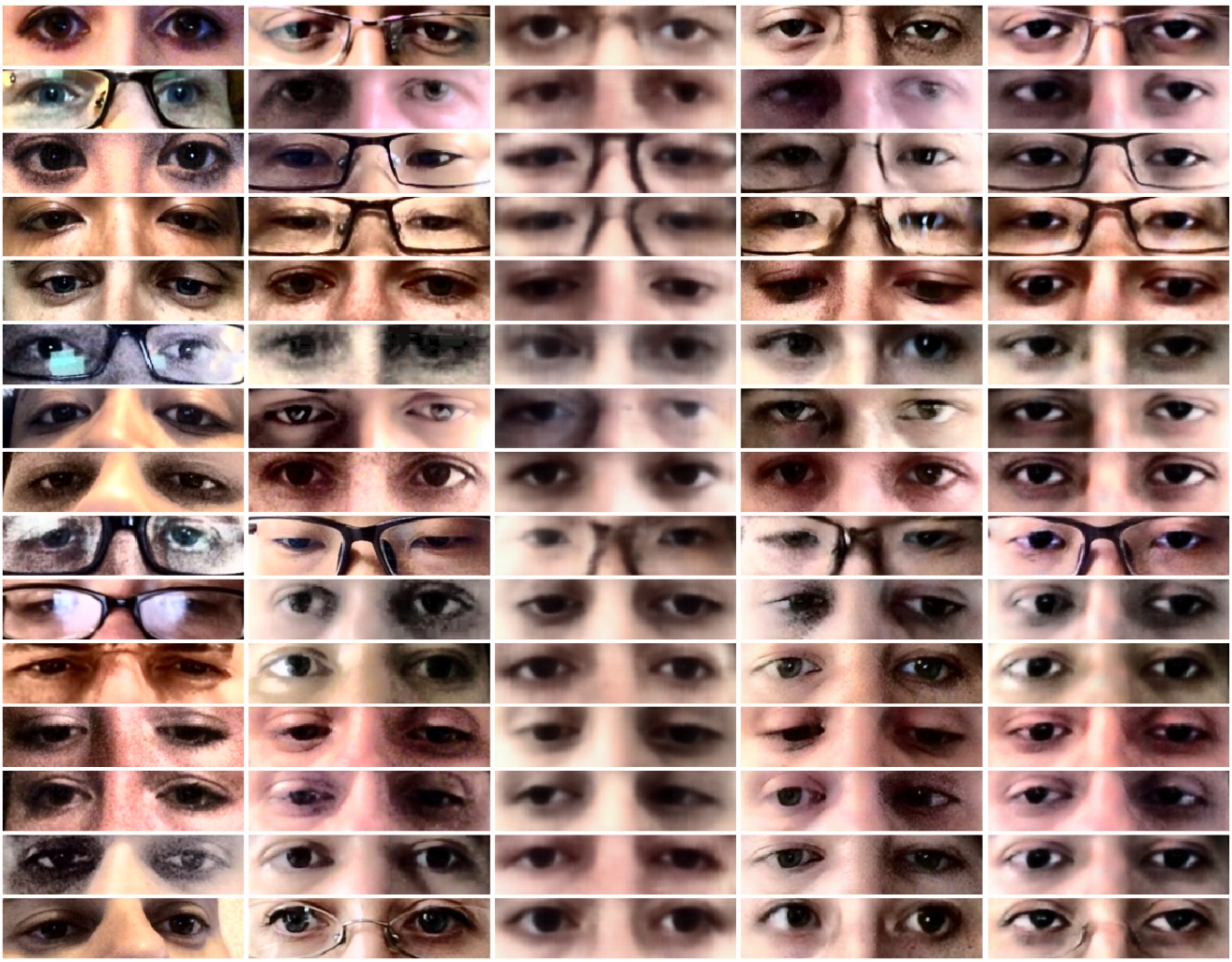
[16] Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, and Otmar Hilliges. Self-learning transformations for improving gaze and head redirection. *Advances in Neural Information Processing Systems*, 33:13127–13138, 2020. 1, 4, 5, 6, 7

Table 4: Architecture of the image discriminator networks $\mathbf{D}_T$ and $\mathbf{D}_S$. Note that, both the discriminators has the same architecture.

Layer name	Kernel, Stride, Padding	Activation	Normalization	Output shape
Conv2d	4×4, 2, 1	LeakyReLU (α = 0.2)	-	64× 32×128
Conv2d	4×4, 2, 1	LeakyReLU (α = 0.2)	InstanceNorm	128×16×64
Conv2d	4×4, 2, 1	LeakyReLU (α = 0.2)	InstanceNorm	256×8×32
Conv2d	4×4, 1, 1	LeakyReLU (α = 0.2)	InstanceNorm	512×7×31
Conv2d	4×4, 1, 1	-	-	1×6×30

Table 5: Architecture of the generator network $\mathbf{G}$

Layer name	Kernel	Activation	Normalization	Output shape
Learned Constant Input	-	-	-	512×4×2×8
Upsampling	-	-	-	512×8×4×16
Conv3d	3×3×3	LeakyReLU	AdaIN	256×8× 4×16
Upsampling	-	-	-	256×16×8×32
Conv3d	3×3×3	LeakyReLU	AdaIN	128×16×8×32
Volume Rotation	-	-	-	128×16×8×32
Conv3d	3×3×3	LeakyReLU	-	64×16×8×32
Conv3d	3×3×3	LeakyReLU	-	64×16×8×32
Reshape	-	-	-	(64 · 16)×8×32
Conv2d	1×1	LeakyReLU	-	512×8×32
Conv2d	4×4	LeakyReLU	AdaIN	256×8×32
Upsampling	-	-	-	256×16×32
Conv2d	4×4	LeakyReLU	AdaIN	64×16×64
Upsampling	-	-	-	64×32×128
Conv2d	4×4	LeakyReLU	AdaIN	32×32×128
Upsampling	-	-	-	32×64×256
Conv2d	4×4	Tanh	-	3×64×256



Gaze Source | Input Image | FAZE [12] | ST-ED [16] | CUDA-GHR

(a) Gaze Redirected images for MPIIGaze dataset (*GazeCapture→MPIIGaze*)

Head Source | Input Image | FAZE [12] | ST-ED [16] | CUDA-GHR

(b) Head Redirected images for MPIIGaze dataset (*GazeCapture→MPIIGaze)*

Figure 1: **Additional Qualitative Results (*GazeCapture→MPIIGaze*):** More qualitative results on the MPIIGaze dataset. 1a shows the gaze redirected images and 1b shows the head redirected images. The first column shows the gaze/head pose source image from which gaze/head pose information is used to redirect. The second column shows the input image from the MPIIGaze dataset. Best viewed in color.

Gaze Source | Input Image | FAZE [12] | ST-ED [16] | CUDA-GHR

(a) Gaze Redirected images for Columbia dataset (*GazeCapture→Columbia*)

Head Source | Input Image | FAZE [12] | ST-ED [16] | CUDA-GHR

(b) Head Redirected images for Columbia dataset (*GazeCapture→Columbia)*

Figure 2: **Additional Qualitative Results (*GazeCapture→Columbia*):** Qualitative results on the Columbia dataset. 2a shows the gaze redirected images and 2b shows the head redirected images. The first column shows the gaze/head pose source image from which gaze/head pose information is used to redirect. The second column shows the input image from the Columbia dataset. Best viewed in color.