

NVGaze: An Anatomically-Informed Dataset for Low-Latency, Near-Eye Gaze Estimation

Supplementary Material

Joohwan Kim*
NVIDIA

David Dunn
UNC

Michael Stengel*
NVIDIA

Samuli Laine
NVIDIA

Alexander Majercik
NVIDIA

Shalini De Mello
NVIDIA

Morgan McGuire
NVIDIA

David Luebke
NVIDIA

ACM Reference Format:

Joohwan Kim*, Michael Stengel*, Alexander Majercik, Shalini De Mello, David Dunn, Samuli Laine, Morgan McGuire, and David Luebke. 2019. NVGaze: An Anatomically-Informed Dataset, for Low-Latency, Near-Eye Gaze Estimation, Supplementary Material. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 20 pages. <https://doi.org/10.1145/3290605.3300780>

Dataset Publication Page

The NVGaze eye ball model and datasets are available on the project page <https://sites.google.com/nvidia.com/nvgaze>.

Resolution and Network Complexity

Though neural network architectures for gaze estimation vary widely in form and complexity [3, 11, 15], the minimum required network sophistication for accurately estimating a viewer’s gaze has not previously been analyzed. For a camera that is stably fixed with respect to the eye, gaze direction might be a simple function of the position of iris or pupil, easily computed by a simple network. However, blinks, different users, and slippage are common factors that often challenge such simple algorithms and make a more complex mechanism necessary. This section examines to what extent these variations have an effect on the size and complexity of the neural network required for accurate gaze estimation.

Network Architecture and Training. Our template neural network consists of a variable number of convolutional layers, followed by a single fully-connected layer that outputs the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-5970-2/19/05...\$15.00

<https://doi.org/10.1145/3290605.3300780>

two result values for horizontal and vertical gaze. All convolutional layers use 3×3 kernels and have a stride of 2×2 pixels, i.e., they approximately halve the image resolution at each layer. No pooling or padding were used. Motivated by Laine et al. [9], the output channel count is increased by a factor of 1.5 at each convolutional layer. The convolutional layers use ReLU activation, whereas the final fully-connected layer uses linear activation. Dropout layers [13] with $p = 0.1$ were added after each convolutional layer to prevent overfitting. Finally, when calibrating for multiple subjects, a learned, per-subject affine transformation is performed as a post-processing step.

To evaluate how well a trained network generalizes on a novel subject, we define *generalization error* as the absolute error between the test labels and inferred values transformed according to an optimal affine calibration transform, computed between the set of inferred values and the set of test labels.

When training, the network weights were randomly initialized following He et al. [5], and Adam optimizer [8] was run for 1500 epochs with parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$ and $\epsilon = 10^{-8}$. Every minibatch contained 10 images from each training subject, allowing the minibatch size to vary between different experimental setups, but keeping the number of updates to the network constant. Learning rate was ramped up to $\lambda = 10^{-3}$ during the first 10 training epochs, and ramped down to zero during the last 150 epochs to ensure convergence to a local optimum.

Experimental Conditions. There are three conditions, each of which was characterized by a dataset representing different degrees of input data complexity. The first and simplest dataset consists of data generated using a single synthetic head model. The distance and rotation between the camera and the face are constant for all samples and the eye has no animated blinks. The second and modestly complex dataset was again generated by using a single synthetic model but includes randomized blinks and randomized slippage between the camera and the face for each image sample. The last and most complex dataset was generated using all 10 synthetic

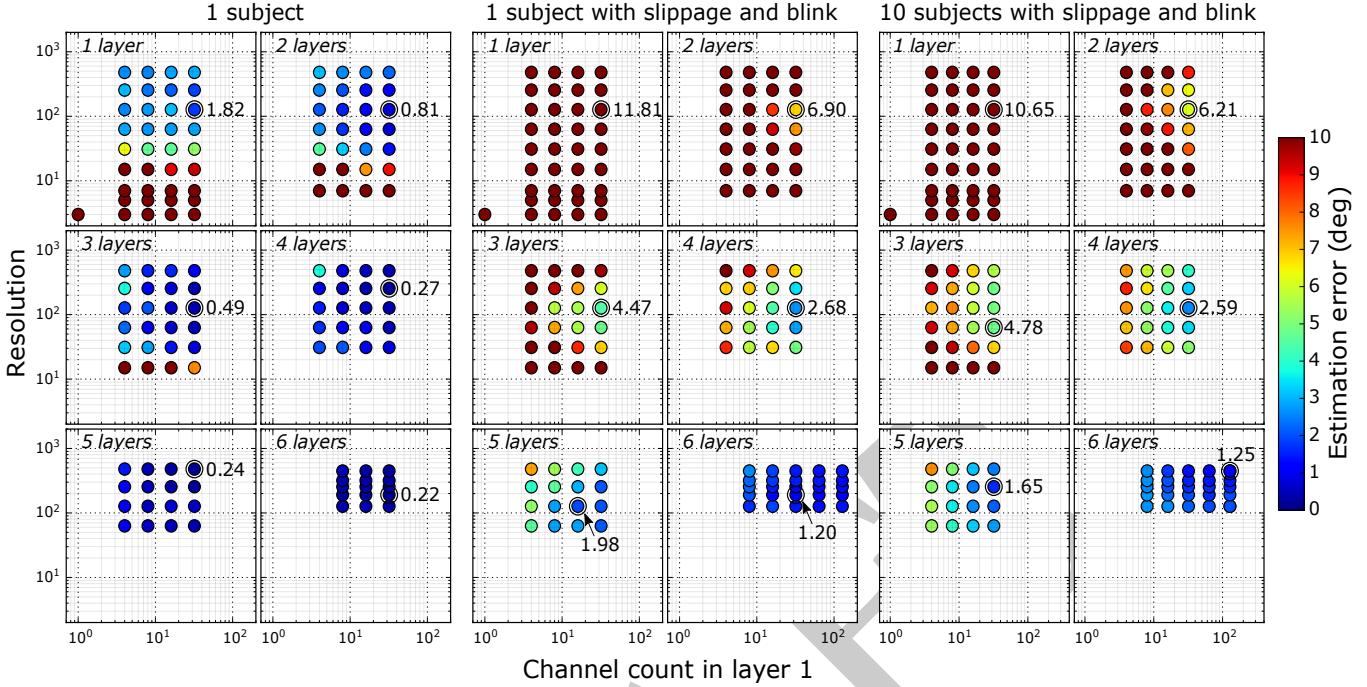


Figure 1: Effect of data and network complexity on gaze estimation accuracy. Using our synthetic data, we explore the effect of complexity in input data on the required amount of information and size of network for accurate gaze estimation. Left, middle, and right columns are, respectively, when input data consist of 1 subject’s images without simulating slippage and blink, 1 subject’s images with slippage and blink simulated, and 10 subjects’ images with slippage and blink simulated. Each plot represents a set of gaze estimation accuracy with a network with certain number of layers. The network yielding the best accuracy in each plot is denoted by a black circle with its accuracy labelled next to it. The unit of estimation error was degree in visual angle.

head models and included blinks and slippage randomization. The slippage value is randomly selected within $[-0.5, 0.5]$ cm around the original camera position in horizontal and vertical directions with uniform distribution. The pupil size is uniformly randomized within the normal range (2 to 8 mm) for all the datasets.

In each experimental condition, we trained approximately 150 networks with different degrees of network complexity and input resolution. The number of convolutional layers varies from 1 to 6 and, independently, the number of output channels of the first convolutional layer is varied. As we grow the number of channels by a fixed factor of 1.5 for each layer, the total number of activations in the network scales linearly with respect to the output channel count of the first layer. Input resolution R is selected among the values that meet the criterion in Eq. 1, which describes the condition where every pixel in the input data contributes to the output for a given stride size s , number of layers l , kernel size k , and an arbitrary positive integer N .

$$R = s^l \times N + (k - s) \frac{s^l - 1}{s - 1} \quad (1)$$

Experimental Results. Fig. 1 summarizes the results of our investigation. Note that the supported minimum resolution of our network architecture grows with the number of used convolutional layers due to the criterion of Eq. 1. As expected, rather shallow network architectures are sufficient for the simplest case (Fig. 1, left) to perform gaze estimation with a surprisingly high degree of accuracy. Even a single-layer network can estimate gaze with less than 2 degrees angular error. A two-layer network brings the error down to below 1 deg. Increasing the number of convolutional layers improves accuracy, but the improvement is marginal after 4 layers. Interestingly, higher input resolution was not helpful for the simple single subject case.

The middle plots in Fig. 1 show the results for modestly complex dataset (1 subject with slippage and blink). Note the dramatic increase in the estimation error in shallow networks. In this case, we need at least 5 convolutional layers in order to get down below 2 deg of estimation error. Such complex network architecture is probably required for dealing with the complexity in input induced by the simulated slippage and blink. As before, higher resolution input does not guarantee improvement in estimation accuracy.

Finally, the right plots in Fig. 1 show the result from the most complex dataset (10 subjects with slippage and blink). Note that we report generalization error in this condition. As for a single subject with slippage and blink, we clearly need a more complex architecture for an accurate gaze estimation. The best network configurations are quite consistent as in the single subject condition when the number of convolutional layers are small. However, deeper networks benefit from high input resolutions more than in the previous test.

We conclude from our investigation that a multi-layer convolutional network with high feature count is necessary to become robust to realistic challenges such as slippage and blinks. A shallow network architecture can accurately perform gaze estimation only when there is not much variation in input (e.g. no slippage or blinks). More channels in each convolutional layer generally improve results, suggesting a trade-off between performance and computational resources. However, increasing the input resolution does not guarantee improvement in accuracy; we advise choosing the optimal resolution through experiments. We hope our synthetic dataset is helpful in executing such experiments.

A good reference for computational complexity analysis in terms of FLOPS and milliwatts can be found in Zemblys et al. [18].

Synthetic Dataset Evaluation

In addition to the evaluation in the main manuscript, we additionally compared our network trained on the UnityEyes model of Wood et al.’s [16] with rasterized images and path traced images generated from our anatomically-informed extensions to the SynthesEyes model of Wood et al. [17], holding both resolution and number of images (1M) constant. In this experiment, path tracing the images and improved anatomical accuracy decreased the generalization error from 3.7° to 3.1° when validated on real images. This result provides additional evidence of the superior performance of our proposed synthetic model for the task of near-eye gaze estimation under IR lighting.

Pupil Localization: Training

The network architecture is equivalent to the gaze estimation experiment, except that we use 7 convolutional layers as shown below.

| Layer index | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|-----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Kernel size | 9×9 | 7×7 | 5×5 | 5×5 | 3×3 | 3×3 | 3×3 |
| Output channels | 24 | 36 | 52 | 80 | 124 | 256 | 512 |

We perform various augmentation steps during training as we did for the gaze estimation network, making a subset of our synthetic data sufficient for convergence of our network. We use a resolution of 293x293 as input for pupil location

estimation and always rescale the input to this resolution using bicubic filtering.

We train on 80% of all dataset samples for 150 epochs using Adam [8] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$ and use the remaining samples for testing. The learning rate is kept at $\lambda = 10^{-4}$ until a fairly long ramp-down to zero during the last 50 epochs. There is no learning rate ramp-up.

Our augmentations are as follows:

First, we transform the image and label using a random affine transformation (translation up to 50 pixels, rotation about origin up to 10° , rescale with a randomized factor within the range [0.9,1.5]) while making sure that pupil location is within a reasonable range of [0.1,0.9] of horizontal and vertical image size. Second, we apply pixel-wise intensity noise with a maximum offset of ± 10 (probability $p=0.5$). Third, we apply a global intensity offset of ± 40 ($p=0.5$). Fourth, we apply Gaussian filtering with $\sigma \in [0.6,1.6]$ and kernel size of 7 ($p=0.5$). Fifth, using bicubic filtering we apply random shrinking with a scale factor $s \in [0.25,1.0]$ followed by upscaling again to the full input resolution ($p=0.5$). Sixth, to simulate environment reflections in the eye we randomly overlay the image with images out of 326 natural photographs from the dataset published in [7] ($p=0.25$). Before blending, the typically larger overlay image B is converted to grayscale and randomly cropped the eye image resolution. For eye image E and a randomly chosen opacity value $o \in [0.1, 0.2]$ we use a soft blending function $(1 - ((1 - E) * (1 - B))) * o + E * (1 - o)$ and clip the resulting pixel intensities to a maximum of 255. Seventh, we apply histogram equalization implemented in OpenCV [1]. As a last step, we rescale the image intensities to a range of [-1,1] and randomly shift the mean about ± 0.15 and the min and max intensities about ± 0.1 . During inference of test images only histogram equalization and normalization to [-1,1] are applied.

Pupil Localization: Comparison to previous methods

We compare performance of our network for pupil center detection against Park et al. (Fig.6 on the PupilNet datasets from Fuhl et al. (Fig.2,3,4,5; Table 1,2).

Our comparison shows performance of each approach as a percentage of samples within a give pixel range of values from 0.1–15, increasing by 0.1 pixels at each step. To measure Park et al.’s approach, we used their pretrained network with an input resolution of 180x108. To make the most charitable comparison, we cropped images from the PupilNet dataset in a 180x108 crop centered on the pupil position label and clipped to the edges of the image. This approach provides the highest resolution possible and the clearest pupil image, thus ensuring the best performance of Park et al.’s method on the PupilNet dataset.

REFERENCES

- [1] Ivan Culjak, David Abram, Tomislav Pribanic, Hrvoje Dzapo, and Mario Cifrek. 2012. A brief introduction to OpenCV. In *MIPRO, 2012 proceedings of the 35th international convention*. IEEE, 1725–1730.
- [2] Wolfgang Fuhl, Thomas Kübler, Katrin Sippel, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2015. Excuse: Robust pupil detection in real-world scenarios. In *International Conference on Computer Analysis of Images and Patterns*. Springer, 39–51.
- [3] Wolfgang Fuhl, Thiago Santini, Gjergji Kasneci, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2017. PupilNet v2.0: Convolutional Neural Networks for CPU based real time Robust Pupil Detection. *CoRR* abs/1711.00112 (2017). arXiv:1711.00112 <http://arxiv.org/abs/1711.00112>
- [4] Wolfgang Fuhl, Thiago C Santini, Thomas Kübler, and Enkelejda Kasneci. 2016. Else: Ellipse selection for robust pupil detection in real-world environments. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. ACM, 123–130.
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *CoRR* abs/1502.01852 (2015).
- [6] Amir-Homayoun Javadi, Zahra Hakimi, Morteza Barati, Vincent Walsh, and Lili Tcheang. 2015. SET: a pupil detection method using sinusoidal approximation. *Frontiers in neuroengineering* 8 (2015), 4.
- [7] Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2008. Hamming embedding and weak geometric consistency for large scale image search. In *European conference on computer vision*. Springer, 304–317.
- [8] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014).
- [9] Samuli Laine, Tero Karras, Timo Aila, Antti Herva, Shunsuke Saito, Ronald Yu, Hao Li, and Jaakko Lehtinen. 2017. Production-Level Facial Performance Capture Using Deep Convolutional Neural Networks. *Proc. Symposium on Computer Animation (SCA)*.
- [10] Dongheng Li, David Winfield, and Derrick J Parkhurst. 2005. Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In *null*. IEEE, 79.
- [11] Seonwook Park, Adrian Spurr, and Otmar Hilliges. 2018. Deep Pictorial Gaze Estimation. *arXiv preprint arXiv:1807.10002* (2018).
- [12] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. 2018. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. *arXiv preprint arXiv:1805.04771* (2018).
- [13] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15 (2014), 1929–1958.
- [14] Lech Świrski and Neil A. Dodgson. 2014. Rendering synthetic ground truth images for eye tracker evaluation. In *Proceedings of ETRA 2014*. 219–222. <http://www.cl.cam.ac.uk/research/rainbow/projects/eyerender/>
- [15] Marc Tonsen, Julian Steil, Yusuke Sugano, and Andreas Bulling. 2017. InvisibleEye: Mobile Eye Tracking Using Multiple Low-Resolution Cameras and Learning-Based Gaze Estimation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 106 (Sept. 2017), 21 pages. <https://doi.org/10.1145/3130971>
- [16] Erroll Wood, Tadas Baltrusaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. 2016. Learning an appearance-based gaze estimator from one million synthesised images. In *ETRA*.
- [17] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. 2015. Rendering of Eyes for Eye-Shape Registration and Gaze Estimation. In *Proc. of the IEEE International Conference on Computer Vision (ICCV 2015)* (2015-12-12).
- [18] Raimondas Zemblys and Oleg Komogortsev. 2018. Making stand-alone PS-OG technology tolerant to the equipment shifts. In *Proceedings of the 7th Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction*. ACM, 2.

| Dataset | ElSe[4] | ExCuSe[2] | Fuhl et al.[3] SK_8P_8 | Fuhl et al.[3] F_{CK_XPY} | Fuhl et al.[3] F_{SK_XPY} | Park et al.[12] | Ours |
|---------|-------------|-----------|-----------------------------|--------------------------------|--------------------------------|-----------------|-------------|
| I | 0.86 | 0.72 | 0.77 | 0.78 | 0.82 | 0.35 | 0.76 |
| II | 0.65 | 0.40 | 0.80 | 0.79 | 0.79 | 0.65 | 0.78 |
| III | 0.64 | 0.38 | 0.62 | 0.60 | 0.66 | 0.40 | 0.87 |
| IV | 0.83 | 0.80 | 0.90 | 0.90 | 0.92 | 0.72 | 0.92 |
| V | 0.85 | 0.76 | 0.91 | 0.89 | 0.92 | 0.42 | 0.95 |
| VI | 0.78 | 0.60 | 0.73 | 0.78 | 0.79 | 0.60 | 0.94 |
| VII | 0.60 | 0.49 | 0.73 | 0.80 | 0.73 | 0.39 | 0.78 |
| VIII | 0.68 | 0.55 | 0.84 | 0.83 | 0.81 | 0.47 | 0.91 |
| IX | 0.87 | 0.76 | 0.86 | 0.86 | 0.86 | 0.41 | 0.76 |
| X | 0.79 | 0.79 | 0.80 | 0.78 | 0.81 | 0.53 | 0.88 |
| XI | 0.75 | 0.58 | 0.85 | 0.74 | 0.91 | 0.77 | 0.73 |
| XII | 0.79 | 0.80 | 0.87 | 0.85 | 0.85 | 0.61 | 0.85 |
| XIII | 0.74 | 0.69 | 0.79 | 0.81 | 0.83 | 0.66 | 0.74 |
| XIV | 0.84 | 0.68 | 0.91 | 0.94 | 0.95 | 0.53 | 0.93 |
| XV | 0.57 | 0.56 | 0.81 | 0.71 | 0.81 | 0.35 | 0.88 |
| XVI | 0.60 | 0.35 | 0.80 | 0.72 | 0.80 | 0.65 | 0.86 |
| XVII | 0.90 | 0.79 | 0.99 | 0.87 | 0.97 | 0.66 | 0.96 |
| XVIII | 0.57 | 0.24 | 0.55 | 0.44 | 0.62 | 0.08 | 0.85 |
| XIX | 0.33 | 0.23 | 0.34 | 0.20 | 0.37 | 0.19 | 0.68 |
| XX | 0.78 | 0.58 | 0.79 | 0.73 | 0.79 | 0.31 | 0.85 |
| XXI | 0.47 | 0.52 | 0.81 | 0.67 | 0.83 | 0.25 | 0.91 |
| XXII | 0.53 | 0.26 | 0.50 | 0.52 | 0.58 | 0.30 | 0.82 |
| XXIII | 0.94 | 0.93 | 0.86 | 0.87 | 0.90 | 0.87 | 0.88 |
| XXIV | 0.53 | 0.46 | 0.46 | 0.55 | 0.55 | 0.30 | 0.71 |
| New I | 0.62 | 0.22 | 0.69 | 0.56 | 0.69 | 0.31 | 0.69 |
| New II | 0.26 | 0.16 | 0.44 | 0.35 | 0.45 | 0.08 | 0.69 |
| New III | 0.39 | 0.34 | 0.45 | 0.44 | 0.49 | 0.21 | 0.74 |
| New IV | 0.54 | 0.48 | 0.83 | 0.77 | 0.82 | 0.34 | 0.92 |
| New V | 0.75 | 0.59 | 0.78 | 0.76 | 0.81 | 0.25 | 0.88 |
| Average | 0.67 | 0.54 | 0.74 | 0.71 | 0.76 | 0.44 | 0.83 |

Table 1: Five pixel error on PupilNet dataset. We estimate the probability of finding the pupil within a distance of max. 5 pixels with respect to the ground truth label of the 384x288 input images. Values are estimated using our 293x293 network images rescaled to 293x293. Images for Park et al.[12] are cropped to 180x108 as described in the text. Other values are directly copied from original PupilNet paper by Fuhl et al. [3].

| Pixel Error | Starburst [10] 384x288 | SET [6] 384x288 | Swirski [14] 384x288 | ExCuSe [2] 384x288 | ElSe [4] 384x288 | Fuhl et al. [3] F_{CK_XPY} 384x288 | Fuhl et al. [3] F_{SK_XPY} 384x288 | Park et al. [3] cropped 180x108 | Ours 384x288 | Ours 293x293 |
|-------------|---------------------------|--------------------|-------------------------|-----------------------|---------------------|--|--|---------------------------------------|-----------------|-----------------|
| 0 | 0 | 0 | 0 | 0 | 0.01 | 0 | 0.01 | 0 | 0 | 0 |
| 1 | 0.007 | 0.0278 | 0.0528 | 0.108 | 0.1306 | 0.105 | 0.1287 | 0.01284 | 0.1040 | 0.1296 |
| 2 | 0.03 | 0.0853 | 0.1548 | 0.2959 | 0.3693 | 0.2945 | 0.3477 | 0.09298 | 0.3295 | 0.3963 |
| 3 | 0.06 | 0.148 | 0.2343 | 0.4376 | 0.541 | 0.499 | 0.5564 | 0.19988 | 0.5565 | 0.6304 |
| 4 | 0.09 | 0.1849 | 0.2757 | 0.5045 | 0.6236 | 0.6279 | 0.6856 | 0.35068 | 0.7259 | 0.7876 |
| 5 | 0.115 | 0.2068 | 0.3012 | 0.5418 | 0.6709 | 0.7059 | 0.7671 | 0.43719 | 0.8308 | 0.8736 |
| 6 | 0.135 | 0.2159 | 0.318 | 0.5656 | 0.6973 | 0.7481 | 0.7991 | 0.48127 | 0.8910 | 0.9218 |
| 7 | 0.151 | 0.2229 | 0.3348 | 0.5824 | 0.717 | 0.7709 | 0.8225 | 0.52939 | 0.9282 | 0.9532 |
| 8 | 0.162 | 0.2279 | 0.3463 | 0.5933 | 0.7307 | 0.7857 | 0.8367 | 0.59343 | 0.9528 | 0.9719 |
| 9 | 0.173 | 0.231 | 0.3599 | 0.6034 | 0.742 | 0.7961 | 0.8465 | 0.67657 | 0.9708 | 0.9848 |
| 10 | 0.1819 | 0.2324 | 0.3694 | 0.6113 | 0.7502 | 0.8028 | 0.854 | 0.71770 | 0.9829 | 0.9931 |
| 11 | 0.1891 | 0.2332 | 0.3794 | 0.6187 | 0.756 | 0.8052 | 0.8589 | 0.73676 | 0.9906 | 0.9952 |
| 12 | 0.1965 | 0.2349 | 0.3888 | 0.6248 | 0.7606 | 0.8067 | 0.8624 | 0.75653 | 0.9950 | 0.9963 |
| 13 | 0.205 | 0.2368 | 0.4005 | 0.632 | 0.7637 | 0.8101 | 0.8652 | 0.78478 | 0.9972 | 0.9971 |
| 14 | 0.21 | 0.237 | 0.4093 | 0.6384 | 0.7689 | 0.8143 | 0.8679 | 0.81807 | 0.9975 | 0.9977 |
| 15 | 0.215 | 0.24 | 0.4171 | 0.6406 | 0.7738 | 0.8167 | 0.8711 | 0.83605 | 0.9978 | 0.9979 |

Table 2: Average Pupil Tracking Error on PupilNet dataset. The average pixel error distribution is given for our network and other pupil trackers. Error is computed with respect to the given input resolution. PupilNet images have native size of 348x288. For fair comparison to Park et al. we crop the image to the network input size of 180x108. Crop is performed as centric as possible to the ground truth pupil location. No rescaling of image is performed. For our network we rescale the image to 293x293 and compute the error with respect to 384x288 as well as 293x293. The results show that our network is performing better even with respect to 384x288 although inference is only using 293x293 input.

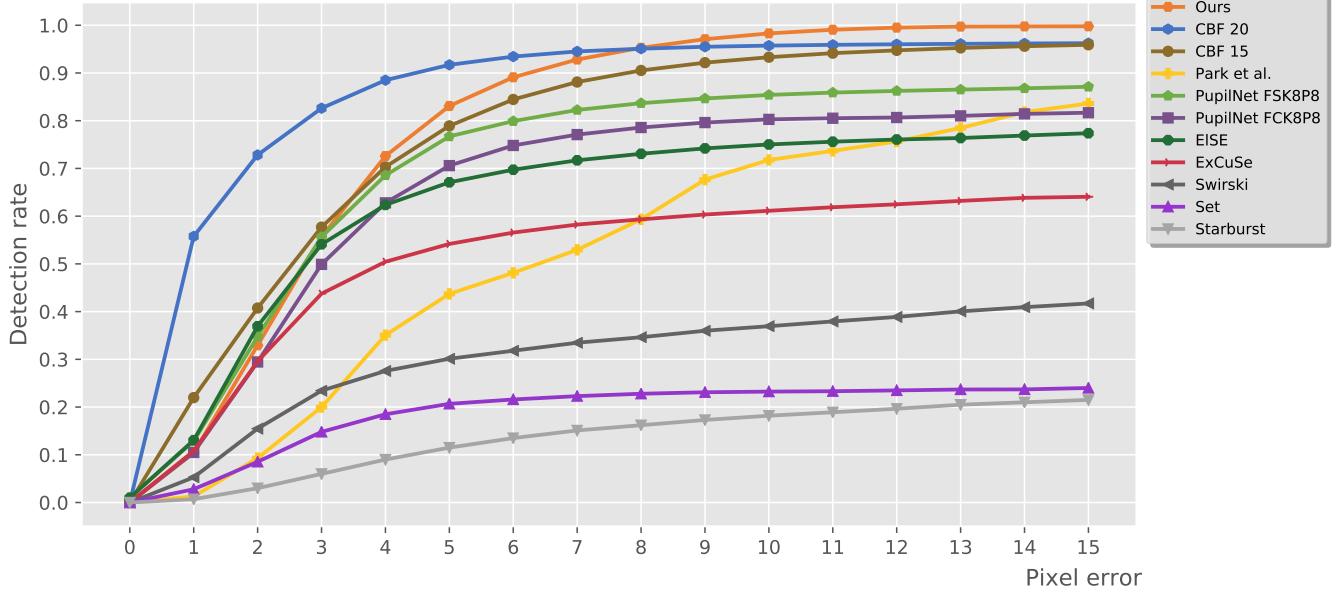


Figure 2: Average Pupil Tracking Error on PupilNet dataset with respect to 384x288 input image resolution. Our 7-layer 293x293 pupil tracking network provides consistently high accuracy and exceeds previous work in terms of robustness. Please consider explanation in Table 2 for details on the shown values.

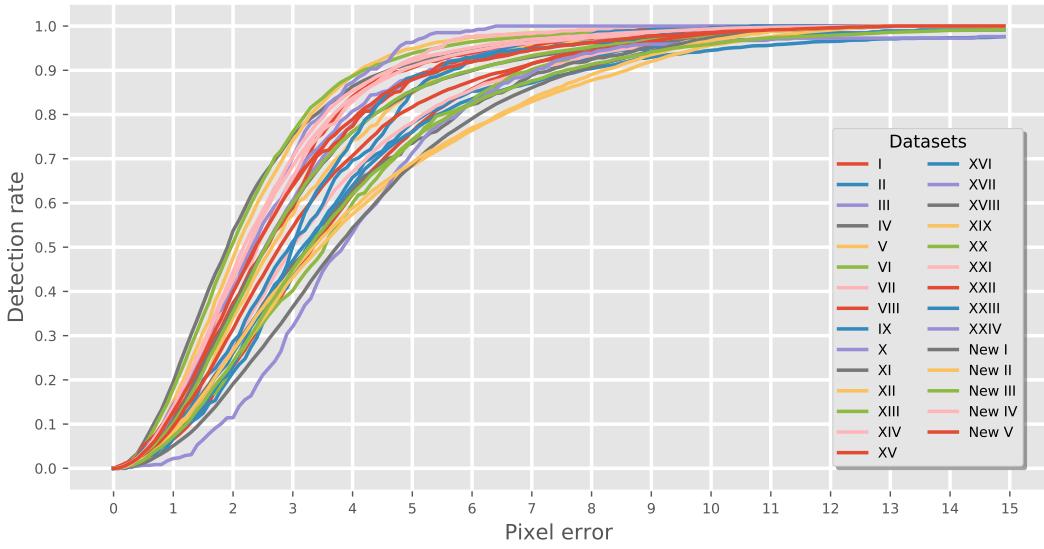


Figure 3: Pupil Tracking Error per individual dataset of PupilNet. Input rescaled to 293x293 for our pupil tracking network. Error computed with respect to original dataset resolution of 384x288 pixels.

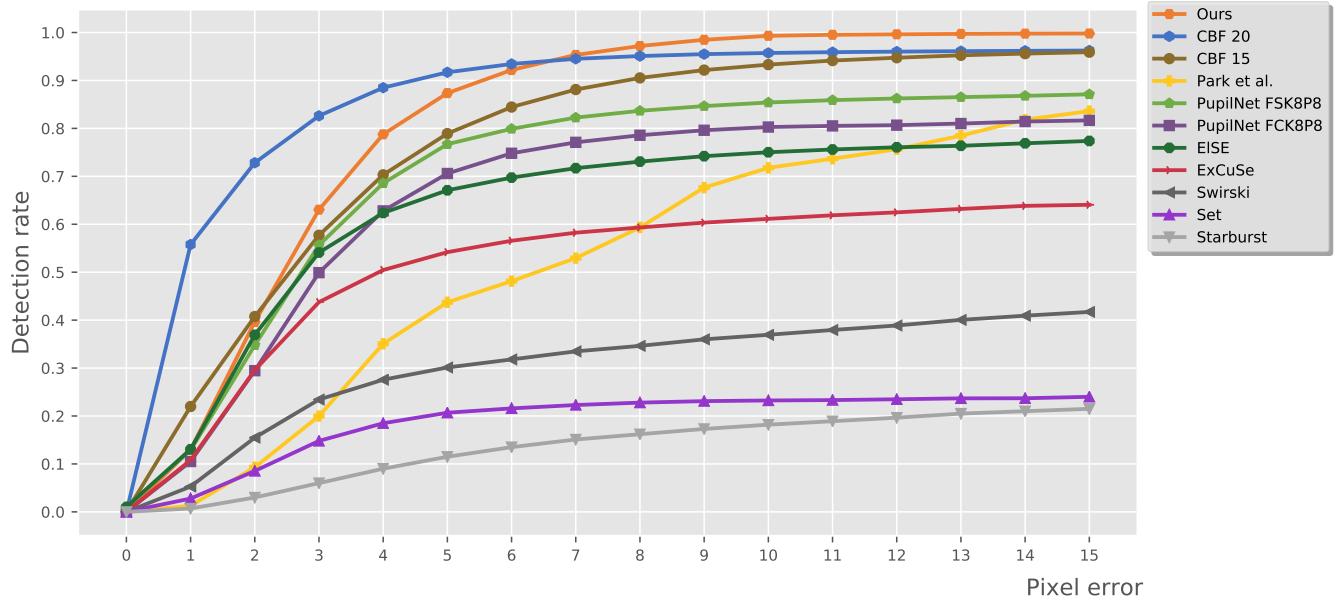


Figure 4: Average Pupil Tracking Error on PupilNet dataset with respect to native network resolution. Results shown for our 7-layer 293x293 pupil tracking network. Please consider explanation in Table 2 for details on the shown values.

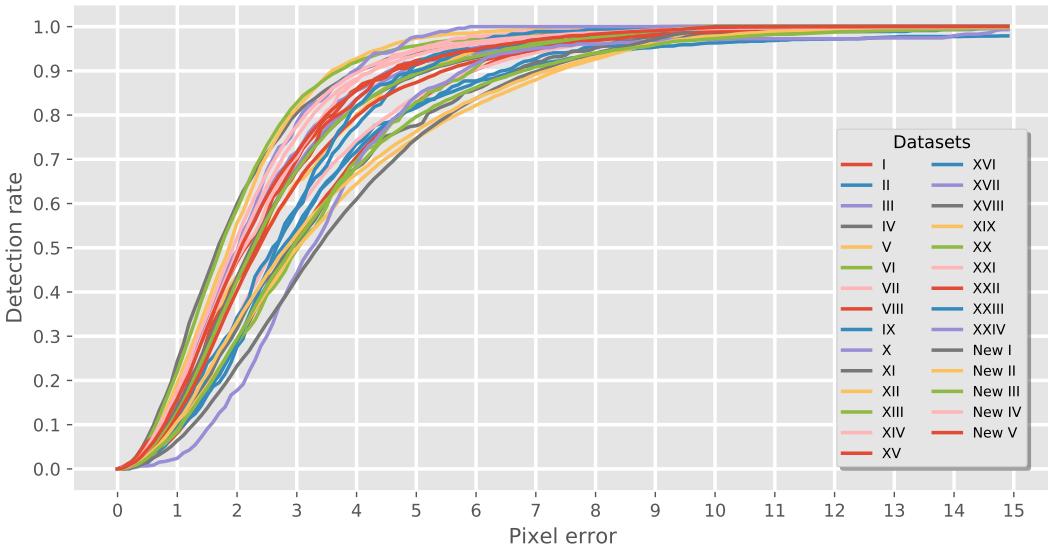


Figure 5: Pupil Tracking Error per individual dataset of PupilNet. Input rescaled to 293x293 for our pupil tracking network. Error computed with respect to the same resolution (293x293).

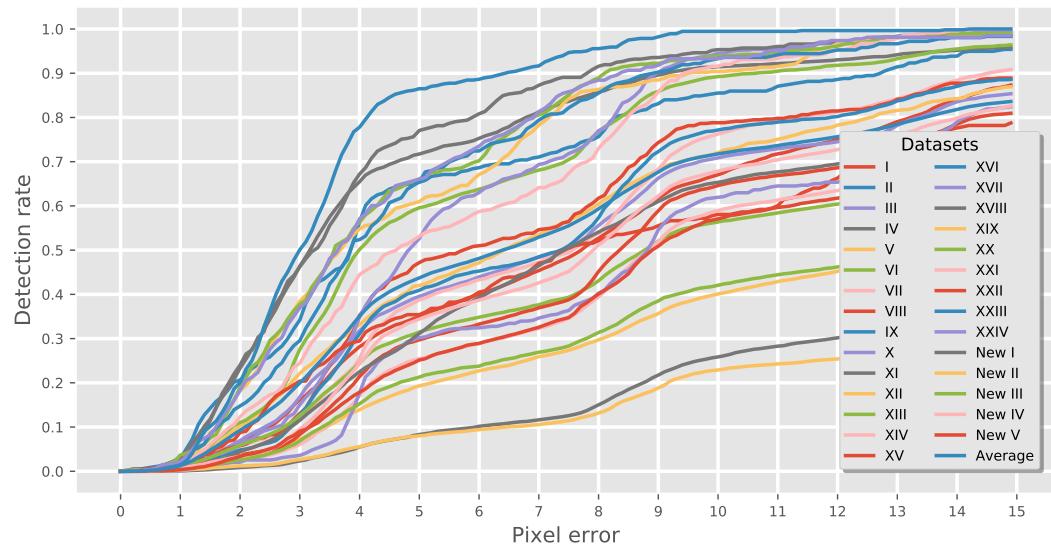


Figure 6: Park et al. Average Pupil/Iris Tracking Error on PupilNet dataset with respect to dataset resolution. Results shown for 180x108 network of Park et al. for Pupil/Iris Center Tracking. Please consider explanation in Table 2 for details on the shown values.



Figure 7: Successful pupil estimation using 7-layer network with 293x239 input resolution. The dataset includes challenging cases such as bad lighting conditions, occluded and partly occluded pupils, head motion, mascara darkening eye lashes, refraction and reflections off glasses and contacts. The estimated pupil location is shown in red.



Figure 8: Augmentation samples for pupil estimation network. Image augmentations include rescaling, affine transformation, brightness and contrast variation, gaussian noise, randomly overlayed images. The estimated pupil location is shown in red.

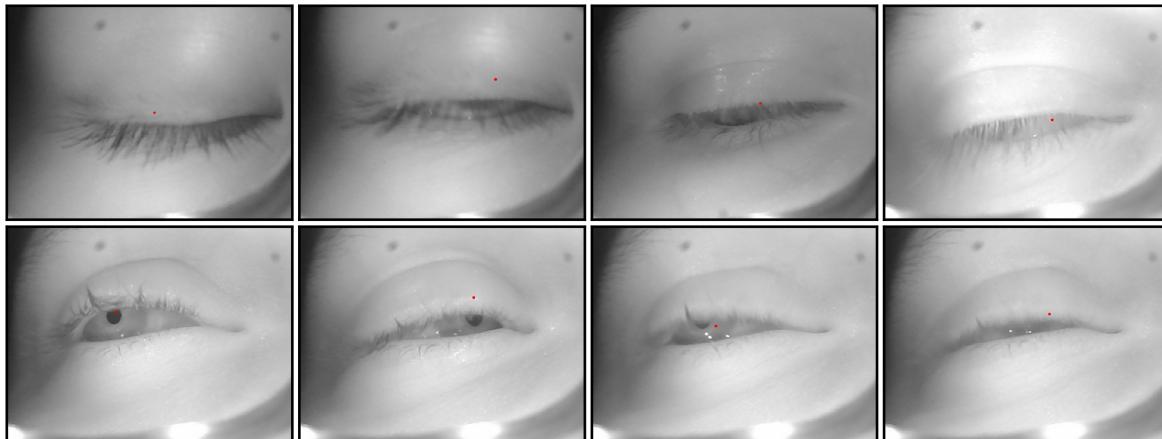


Figure 9: Limiting samples for pupil estimation network. Pupil estimation failed due to insufficient pupil visibility. The estimated pupil location is shown in red.

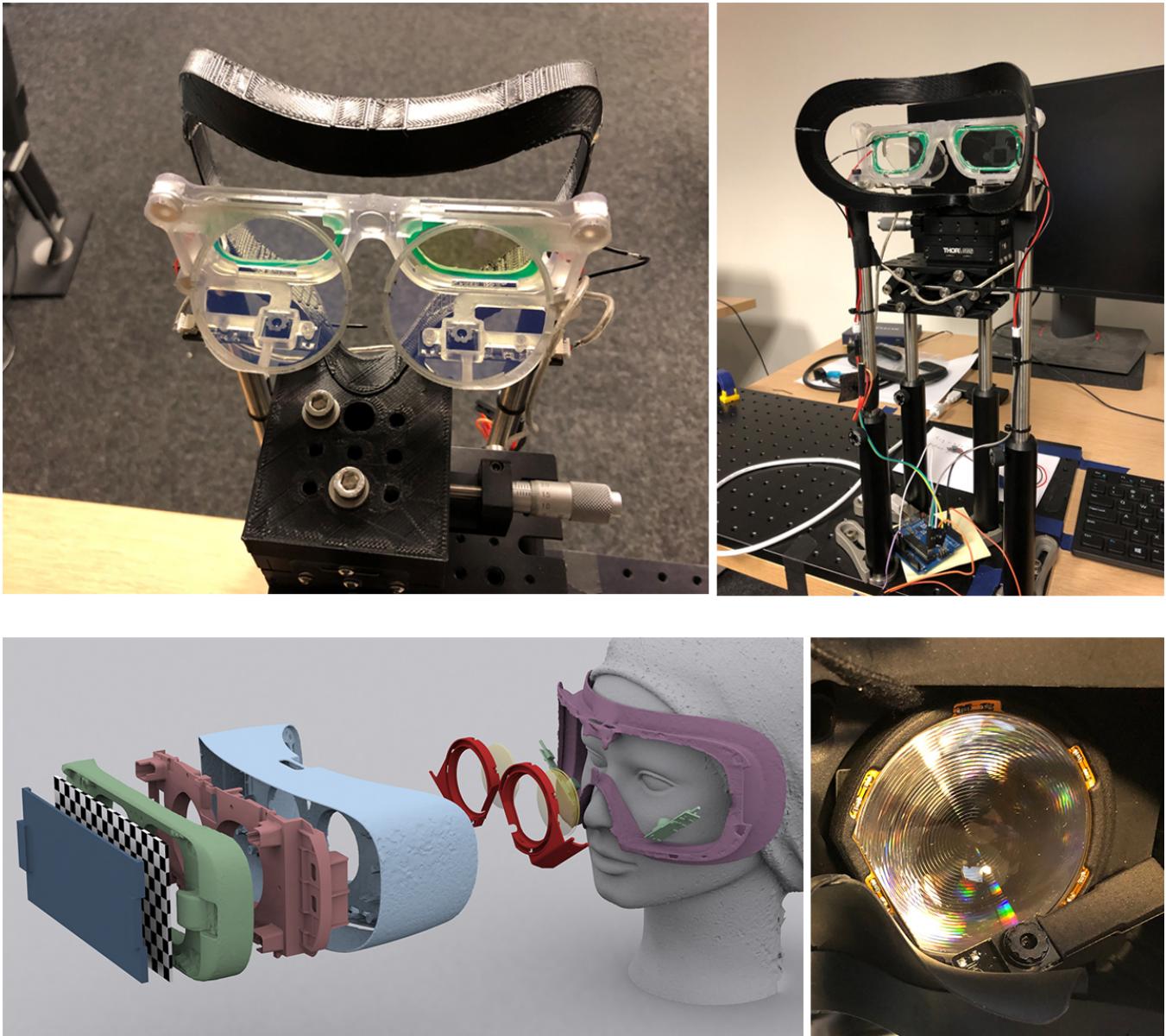


Figure 10: Eye Capture Hardware setup. Top row: The mounted AR glasses setup used for the first real dataset includes beam splitters to enable the on-axis view for the binocular eye tracking cameras. The infrared lighting units is driven by an Arduino microcontroller to enable temporally varying lighting conditions. The face mask stabilizes the user's head during the capture experiment. The overall geometric setup is also used for generating the rendered synthetic dataset. Bottom row: Custom VR headset based on Samsung GearVR and Pupil Labs cameras. The headset is used for creating the off-axis eye image dataset. The headset includes clips attached to the lenses to hold the binocular eye cameras and infrared LEDs in place.

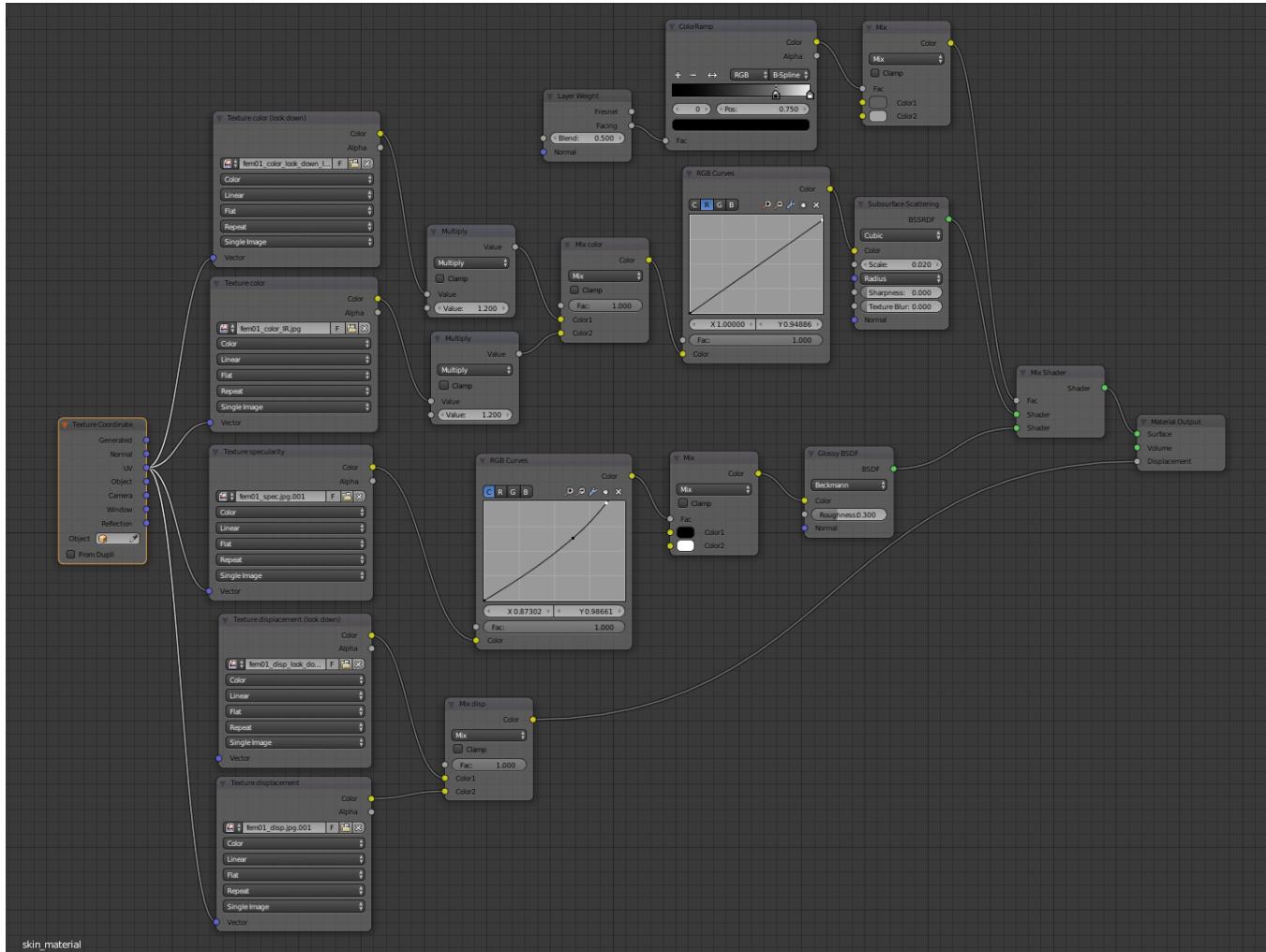


Figure 11: Shader graph of skin material for Cycles renderer in Blender.

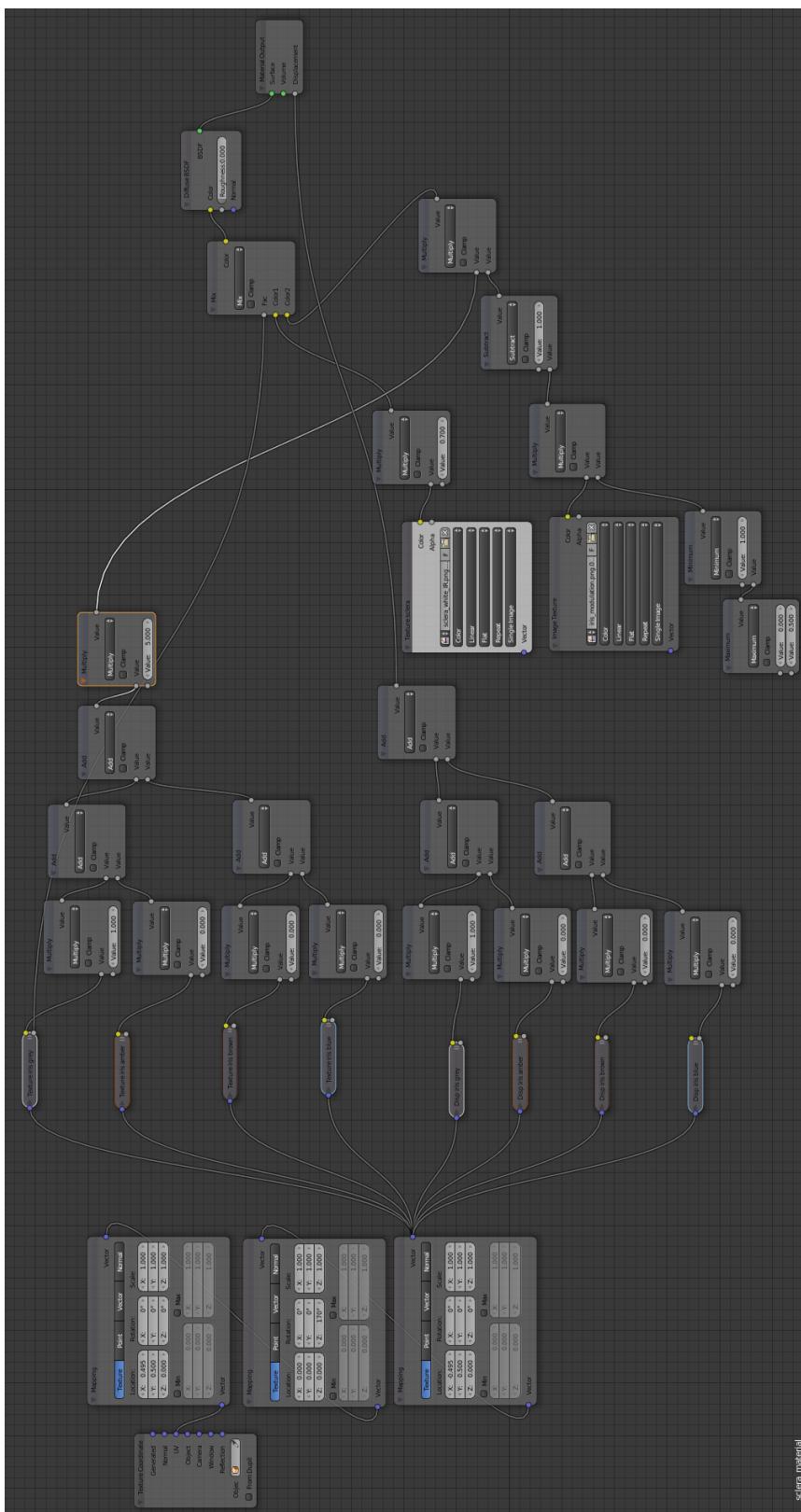


Figure 12: Shader graph of sclera material for Cycles renderer in Blender.

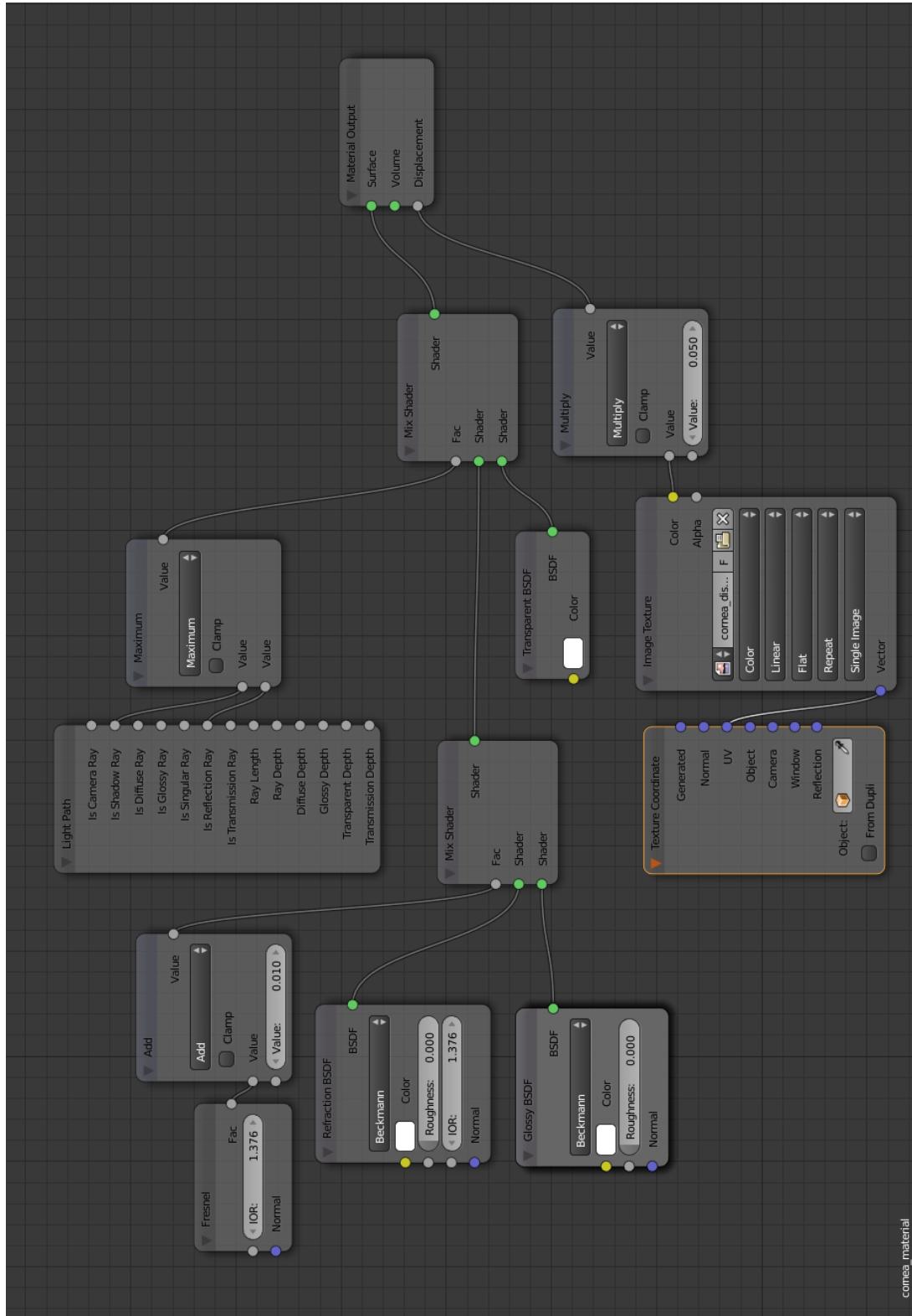


Figure 13: Shader graph of cornea material for Cycles renderer in Blender.