Improving Physical Parameters for Image Registration of Retinal Images

Retinal Image Registration is an important task used to help physicians track, monitor, and diagnose many retinal diseases. A key point detection deep learning network, SuperRetina, has been applied to the Fundus Image Registration (FIRE) dataset with moderate success. In this paper, we analyze how physical parameters like aperture diameter, band pass filters, and physical lenses can affect the training of SuperRetina. The results show that changing physical parameters can affect the performance of SuperRetina and suggest how optimizing physical parameters of image acquisition can assist in the task of image registration.

1. Introduction

Retinal Image Registration (RIR) is a technique in which, given a pair consisting of a test and a reference image, the test image is transformed so that its points are co-located with the corresponding points in the reference image. Often these images in the pair can different with respect to their viewpoint, acquisition time and acquisition device.[1] More specifically, in this paper, we discuss retinal image matching (RIM), which is to match color fundus photographs based on their visual content. The criteria used for matching is often task-dependent.[2] RIM is nontrivial. This is due to various factors such as illumination conditions, abnormal retinal changes, and natural motions of the eye.[2] Thus, the use of computer vision and machine learning have become crucial for this task.

This image modality is important because it aids in the monitoring and diagnosis of diabetic retinopathy, glaucoma, and other retinal diseases. Diabetic retinopathy (DR) is a complication of diabetes mellitus and the second leading cause of visual loss in the United States.[3] Diabetes mellitus is projected to affect approximately 600 million people by 2035.[4] Because of this our team plans to explore different image parameters and decide which apertures could best improve image registration of retinal images.

The FIRE Dataset which consists of 134 retinal fundus image pairs will be used.[5] The images will then be fed through physical layers with different apertures. We will focus on how each aperture affects the loss after a certain number of epochs. Moreover, we attempt to find the optimal apertures and mask to help with registration. These being either the aperture, band pass mask, or a phase mask. Then, we will feed the optimal weights from these layers to train a CNN that will provide both a bitmask with locations and features vectors with features such as color and contrast about each location. Determining which image parameters and apertures improve the image registration will help physicians better diagnose and treat patients with retinal diseases.

2. Related Work

Both research papers and review papers have focused on the topic of RIR and RIM. Many iterations of machine learning models such as a SIFT detector and GLAMpoints were used in order to achieve these tasks. A SIFT detector finds corners and blobs but tends to respond around the lesions and the boundary between the circular foreground and dark background. GLAMpoints was a trainable detector that learned key points and was self-supervised. The major flaw in this detection was the self-supervision. It had many detections on non-vascular areas which were unreliable. Specifically in this paper, we modify the SuperRetina Network Architecture. In a paper written by Liu et. al. , they use a semi self-supervised training model for key point expansion. They concluded that it resolved many incomplete annotations and also compares favorably against prior methods. We use this SuperRetina Network Architecture with the physical layers to conduct labeling of image pairs.[2]

3. Methods

3.1 SuperRetina Network Architecture

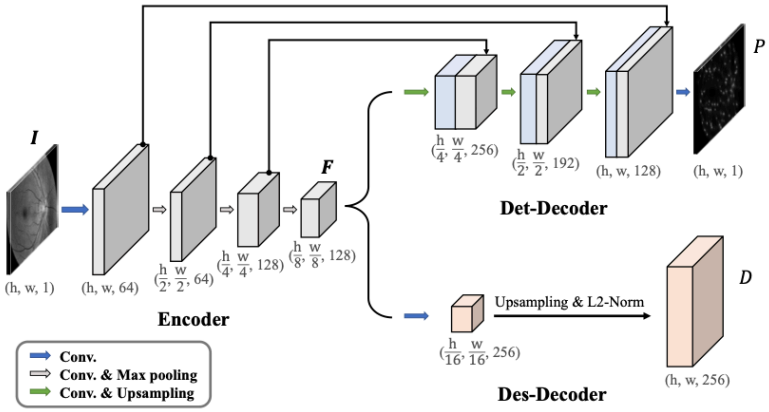

Figure 1. SuperRetina Network Architecture (PKE Removed)

Our physical layer experiments are built on top of a modified version of SuperRetina Net by Liu et al.[2]. The original network involved self-supervised training for progressive key point expansion (PKE) based on intermediate model results. We created a baseline without the PKE pipeline to minimize the number of variables being changed in our physical layer experiments (Figure 1). The model comprises a detection decoder to create segmentation maps using a U-Net like architecture to find key points within the input image. It also contains a description decoder to build 1x256 feature vectors for each input pixel, which should be invariant to affine transformations of the image (shift, scale, rotation, etc.). After the model is run on each image within an image pair, the predicted keypoints and their corresponding feature vectors can be used to find matching keypoint pairs (based on feature similarity).

3.2 Loss Functions

The overall model loss consists of l_{clf} for the detection-decoder plus l_{des} for the description decoder as defined below.

$$\ell_{clf}(I;Y) = 1 - \frac{2 \cdot \sum_{i,j} (P \circ \tilde{Y})_{i,j}}{\sum_{i,j} (P \circ P)_{i,j} + \sum_{i,j} (\tilde{Y} \circ \tilde{Y})_{i,j}}$$

Equation 1. Loss function for detection-decoder.

$$\ell_{des}(I;\mathcal{H}) = \sum_{(i,j)\in\widehat{P}} \max(0, m + \phi_{i,j} - \frac{1}{2}(\phi_{i,j}^{rand} + \phi_{i,j}^{hard}))$$

Equation 2. Loss function for description-decoder.

The detection-decoder loss is a soft-dice loss between the expected keypoint bitmap and the model's segmentation outputs (P) over all pixels (i,j). Soft-dice loss is used to smooth out the expected key points into gaussian peaks ($\tilde{Y}$) to smooth the model's loss gradient due to the sparse keypoint labels.

The description-decoder uses a triplet loss to compare the geometric distance between each pixel's feature vector and a positive sample (the sample pixel affine transformed) as well as a negative sample (a randomly selected pixel that should not match). $\Phi_{i,j}$ and $\Phi_{i,j}^{rand}$ represent the distance of pixel (i,j) to the positive and negative samples respectively. During training $\Phi_{i,j}$ is minimized and $\Phi_{i,j}^{rand}$ is maximized. By combining the detector and descriptor losses, the model optimizes for finding key points which are likely to have consistent feature vectors across images.

3.3 Data Set

The original dataset for the SuperRetina Network was not publicly available, so the FIRE (Fundus Image Registration) Dataset was used for model training instead[5]. The dataset consists of 129 retinal images forming 134 image pairs, with approximately 10 key point annotations per image. The images were acquired at a resolution of 2912x2912 pixels and a FOV of 45° both in the x and y dimensions. In addition, ground truth data or control points for the calculation of registration error between images in the pair was provided.

3.4 Data Preprocessing

To ensure compatibility with the SuperRetina Network, the original RGB images were converted to grayscale and were resized to 768x768 pixels. The matching annotations were also scaled down proportionally, as they were encoded in raw pixel coordinates. Image augmentations of random gaussian blurring, contrast shifts and brightness shifts were introduced to compensate for the small training set size.

3.5 Physical Layers

3.5.1 Concentric Overlapping Apertures

To study the effects of aperture choice on retinal key point detection, a trainable layer was created to take the weighted sum of 6 circular apertures of different radii (h/16, h/8, h/6, h/4, h/3, h/2). The combination of these apertures created a weighted mask, which was multiplied element-wise with the input image's FFT. The result was then inverted to create a filtered image to pass through the baseline SuperRetina Network (Figure 2). By analyzing the final aperture weight distribution, we may potentially find better methods for physical aperture design and how to combine multiple aperture images for optimal image registration results.

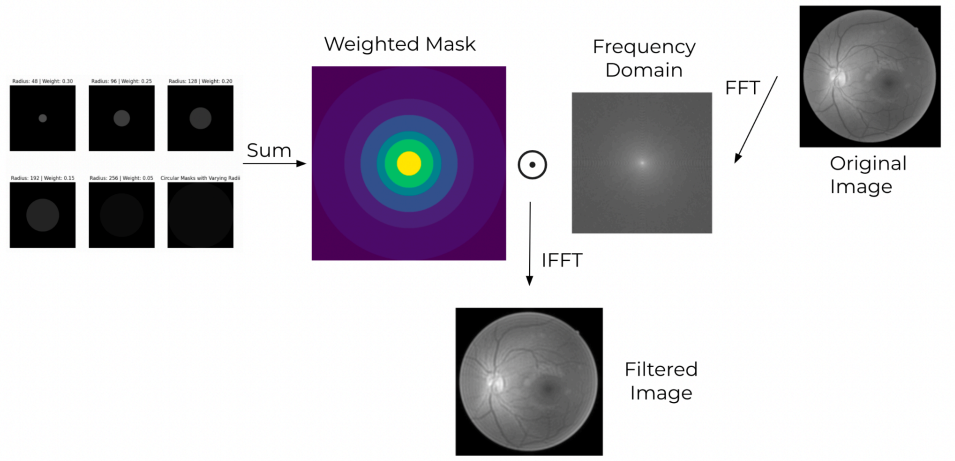

Figure 2. Concentric overlapping aperture masking pipeline

3.5.2 Concentric Non-Overlapping Ring Apertures

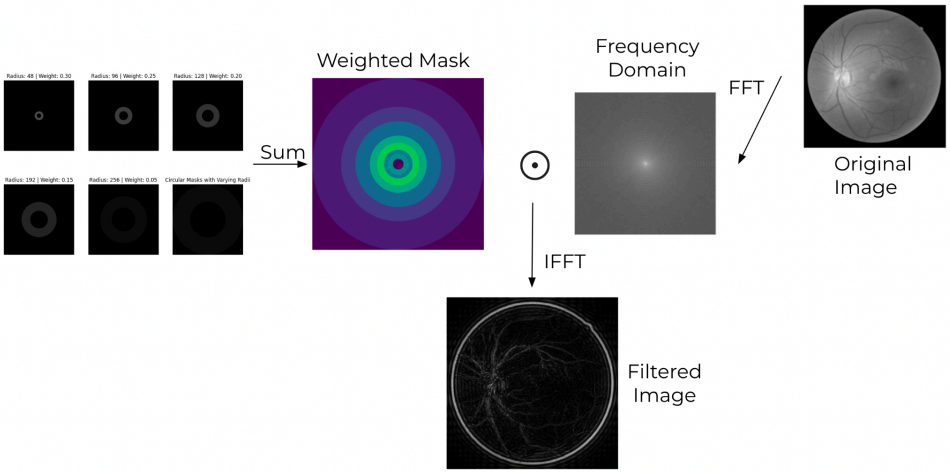

Figure 3. Concentric non-overlapping ring aperture masking pipeline

A similar physical layer was implemented, but with concentric, non-overlapping ring apertures rather than circles. Each previously used circle had an inner radius half its outer one. This would allow the model to have more fine-grained discrimination against the filtering strategy in the fourier domain, with the possibility of creating multiple passbands of varying weights (Figure 3). Similar to physical layer I, analysis of the final aperture weight distribution may potentially suggest improvements for physical aperture design and how to combine multiple apertures for image registration tasks.

3.5.3 Physical Layer III

Another similar physical layer was implemented with three examples of physical lenses which could possibly be added to color fundus cameras to improve machine learning results. Each mask is based upon examples which are seen in image acquisition. The cubic phase mask is designed to reduce the effect of defocus and aberration in the image.[6] The spiral phase mask is used as a type of filter that modifies the amplitude of the input image. It is used to enhance the high-frequency content of the image, which can improve the accuracy of feature extraction.[7] Finally, the Fresnel phase mask represents the focal length of

the camera. This parameter determines the curvature of the wavefront of the incoming light, which affects the phase of the input image. By modifying the phase of the input image in a specific way, the Fresnel mask can improve the accuracy of feature extraction. Similar to physical layer I, analysis of the final phase mask weights distribution may potentially suggest improvements for physical lens design and how to combine multiple lenses for image registration tasks. These masks are shown in Figure 4.

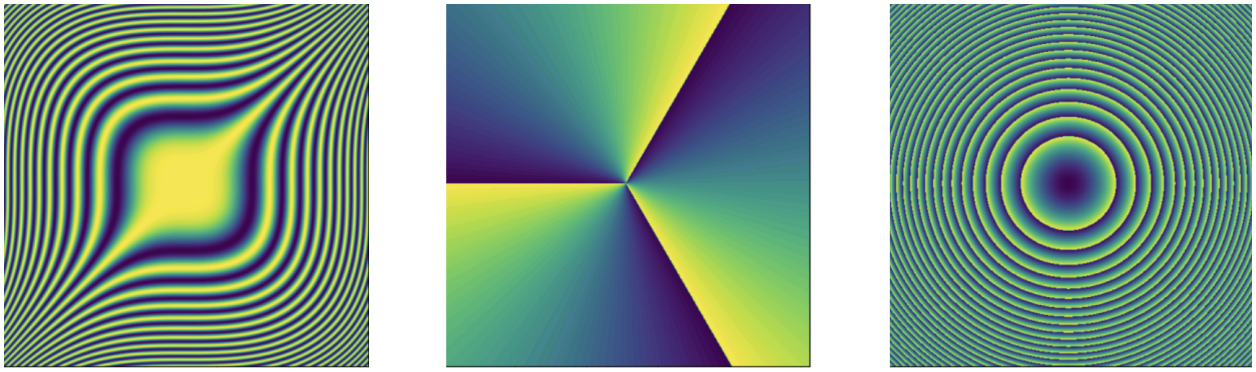

Figure 4. Depiction of the three phase masks experimented on

4. Results

4.1 Baseline Model

The performance of the baseline model with PKE removed is shown in Figure 5 below. Of note, the description-decoder trained much faster than the detector-decoder, leading to more overfitting of the former before the 100 epoch cutoff. The overall loss converged to 0.68 and 0.8 for training and validation sets respectively. Based on performance of the model on keypoint registration on the validation set, the final mAUC was 0.755.

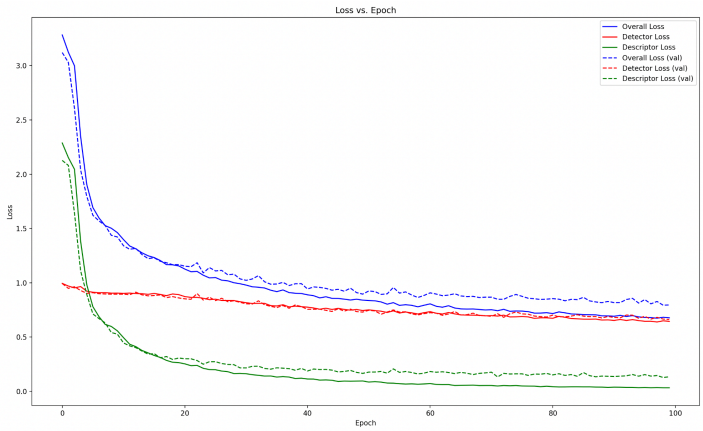

Figure 5. Baseline loss graph (PKE removed)

4.2 Concentric Overlapping Apertures

The performance of the modified model with physical layer I prior to the SuperRetina Network removed is shown in Figure 6 below. Of note, the model was stuck in a local minima for around 30 epochs and then the loss began dropping rapidly. At the 100 epoch cutoff, the overall losses were 1.02 and 1.18 for training and validation respectively. While these results are slightly higher than the baseline model, it seems like

the loss was on a sharper downwards trajectory and may have continued descending if the training cutoff was extended longer. The final validation mAUC on key point matching was 0.64.

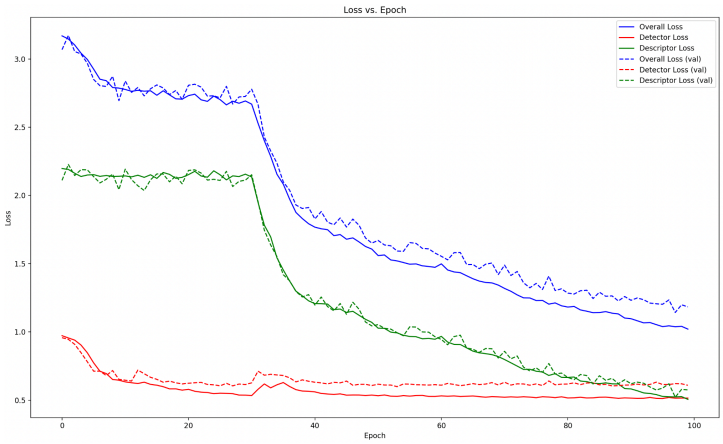

Figure 6. Concentric overlapping apertures loss graph

Random normal and uniform weight initialization strategies were tried for the aperture weights. Uniform weight initialization did not seem to result in much distribution shift after training. However, random initialization yielded the result shown in Figure 7. The model preferred all aperture masks equally except for the one with radius h/8. This seems peculiar and may be due to an oscillation due to over-correcting this weight from its high starting point (~0.3) to match the others.

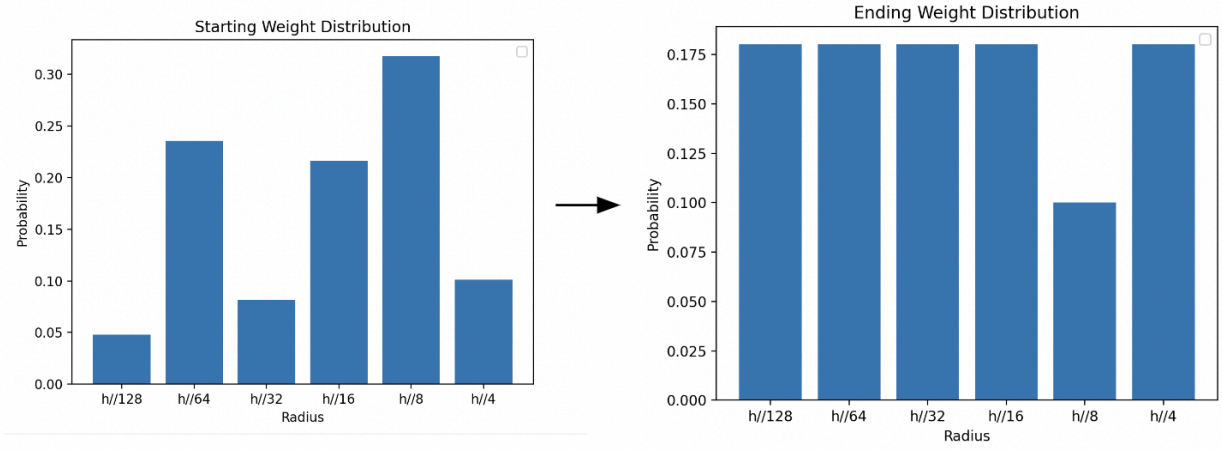

Figure 7. Weight distribution shift for concentric overlapping apertures

4.3 Concentric Non-Overlapping Ring Apertures

The performance of the modified model with physical layer II prior to the SuperRetina Network removed is shown in Figure 8 below. Of note, the model was stuck in a local minima for around 70 epochs and then the loss began dropping rapidly. At the 100 epoch cutoff, the overall losses were 1.33 and 1.65 for training and validation respectively. These losses are higher than the baseline, but still have a sharp downward trend at the 100 epoch cutoff. Due to being caught in the local minima, it seems like the loss functions cannot be compared definitively. The final validation mAUC was 0.55.

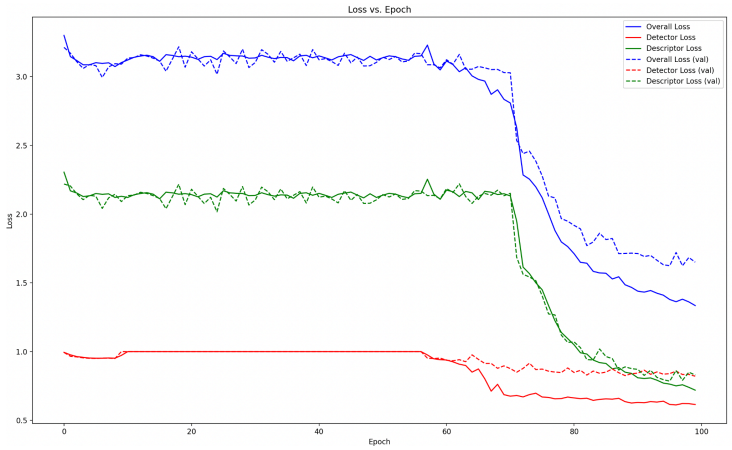

Figure 8. Concentric non-overlapping ring apertures loss graph

Random weight initialization did not seem to result in much distribution shift after training. However, uniform initialization yielded the result shown in Figure 9. The model diminished the effect of the outer two rings and preferred all other aperture rings equally.. This seems logical as some high frequency noise content could be filtered out in the fourier space using this distribution of weights. Unexpectedly, the model did not create any bandpass filters, only using lowpass instead. A sample output of the masking layer is shown below in Figure 10.

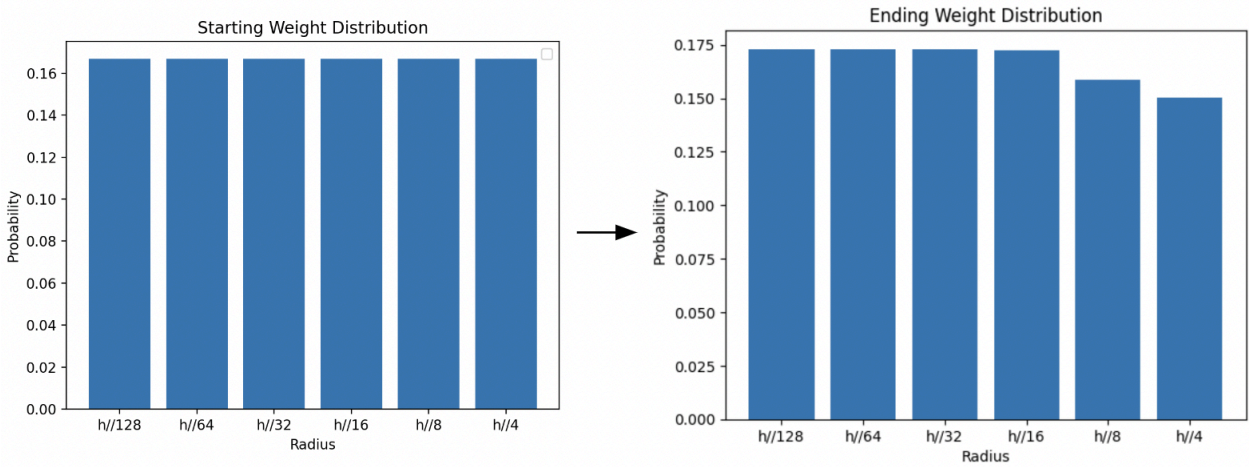

Figure 9. Weight distribution shift for concentric non-overlapping ring apertures

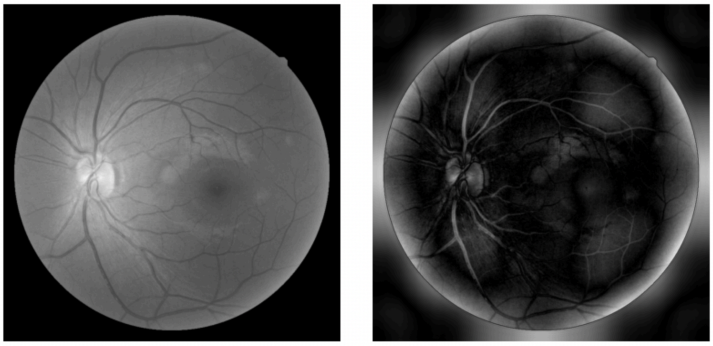

Figure 10. Sample output comparison with (right) and without (left) the trained filtering layer

4.4 Weighted Phase Masks

The performance of the modified network when introducing the three masks is shown below in figure 11. The network appeared to be caught in a local minima without changing the overall loss at all. It appears that the network does, in fact, have no preference. This is seen as the weights of each of the masks are randomly initialized, but rapidly approach 0.33 during training. As the network is not training, it may appear that introducing the phase masks increases the complexity of the decision space for the network.

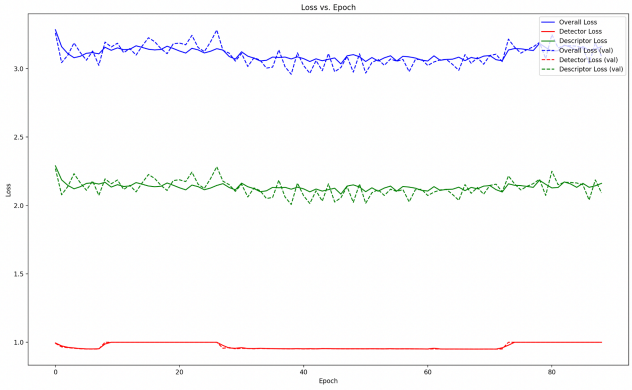

Figure 11. Weighted Phase Mask loss graph

5. Discussion

Our model produced promising results. First, our base model without a physical layer proves that the model can train with the original dataset as shown in Figure 5 as it gives a baseline mAUC of 0.755. For the first physical layer of concentric overlapping ring apertures, we found that after around 60 epochs the model began to train. However, we found that the losses for both the descriptor and the detector after training was still greater than the losses with the baseline model. This was also seen in Figure 7 with the all probabilities being the same except for the h//8 radius. Second, for the second physical layer of non-overlapping, we had similar results to the first physical layer. It took both models approximately 60 epochs to start training. However, the major difference is that it does not do a better job than the first physical layer. As shown in figure 9, the weight distributions are heavier on the smaller radius circles. Given this information, we can conclude that these apertures do not help the model with image registration. Finally, for the selection between phase masks, it appears that the model has no preference

for any phase mask. During training, we note that as the modifications to the frequency data increases in complexity, the model appears to be stuck at a local minima for an increasing number of epochs. As the phase masks modify the frequency domain of the image in a more significant manner, this may mean that the local minima is harder to escape.

There are next steps that could improve our results. First, we would need to change the choice of network architecture. This is because our last experiment was stuck on a local minima and was not able to train properly. We may also need to allow the model to train for longer. In our experiments, we let the model train for 100 epochs. We suspect that longer training would allow for better results. Finally, we would also need to explore other image parameters such as contrast and run experiments on how they impact image registration. Our initial findings show us that the baseline model is better than the concentric overlapping ring aperture's physical layer which is better than the non-overlapping ring aperture's physical layer. Perhaps there are other image parameters that would improve the baseline model that could be explored.

References

[1] Hernandez-Matas, C., Zabulis, X., & Argyros, A. A. (2021). Retinal image registration as a tool for supporting clinical applications. *Computer methods and programs in biomedicine*, *199*, 105900. https://doi.org/10.1016/j.cmpb.2020.105900

[2] Liu, J., Li, X., Wei, Q., Xu, J., & Ding, D. (2022, October). Semi-supervised Keypoint Detector and Descriptor for Retinal Image Matching. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI* (pp. 593-609). Cham: Springer Nature Switzerland.

[3] Abràmoff, M. D., Garvin, M. K., & Sonka, M. (2010). Retinal imaging and image analysis. *IEEE reviews in biomedical engineering*, *3*, 169–208. https://doi.org/10.1109/RBME.2010.2084567

[4] Goh, J. K., Cheung, C. Y., Sim, S. S., Tan, P. C., Tan, G. S., & Wong, T. Y. (2016). Retinal Imaging Techniques for Diabetic Retinopathy Screening. *Journal of diabetes science and technology*, *10*(2), 282–294. https://doi.org/10.1177/1932296816629491

[5] Hernandez-Matas, Carlos & Zabulis, Xenophon & Triantafyllou, Areti & Anyfanti, Panagiota & Douma, Stella & Argyros, Antonis. (2017). FIRE: Fundus Image Registration dataset. Journal for Modeling in Opthalmology (to appear). 1. 10.35119/maio.v1i4.42.

[6] S. Prasad, V. P. Pauca, R. J. Plemmons, T. C. Torgersen, and J. van der Gracht, “Pupil-phase optimization or extended focus, aberration corrected imaging systems,” Proc. SPIE 5559, 335–345 (2004).

[7] Martin Teich, Michael Mattern, Jeremy Sturm, Lars Büttner, and Jürgen W. Czarske, "Spiral phase mask shadow-imaging for 3D-measurement of flow fields," Opt. Express 24, 27371-27381 (2016)