A Hybrid Pipeline For The Segmentation Of Eye Regions From Video Frames

Adish Rao, Aniruddha Mysore, Abhishek Guragol, Rajath Shetty, Siddhanth Ajri, Poulami Sarkar and Gowri Srinivasa

Abstract An accurate tracking and, in turn, an accurate segmentation of key regions of the eye is a *sine qua non* to provide users with superior quality of immersive experiences in augmented reality (AR), virtual reality (VR) and mixed reality (MR) applications. In this paper, we present the detailed rationale and research behind the design of an image processing pipeline to perform pixel-wise segmentation of eye from the background and labeling key structures, viz., sclera, iris and pupil, of the eye. The images used in this study are from a data set provided by Facebook as a part of the OpenEDS Challenge 2020, and are sampled frames from a video capture of the eye. The pipeline we present is a hybrid of traditional image preprocessing

Adish Rao
PES Center for Pattern Recognition, PES University Electronic City Campus, Bengaluru-560100, India. e-mail: contact.adishrao@gmail.com

Aniruddha Mysore
PES Center for Pattern Recognition, PES University Electronic City Campus, Bengaluru-560100, India. e-mail: aniruddha.mysore@gmail.com

Abhishek Guragol
PES Center for Pattern Recognition, PES University Electronic City Campus, Bengaluru-560100, India. e-mail: abhishekguragol@gmail.com

Rajath Shetty
PES Center for Pattern Recognition, PES University Electronic City Campus, Bengaluru-560100, India. e-mail: rajathshetty1999@gmail.com

Siddhanth Ajri
PES Center for Pattern Recognition, PES University Electronic City Campus, Bengaluru-560100, India. e-mail: contactsiddajri@gmail.com

Poulami Sarkar
PES Center for Pattern Recognition, PES University Electronic City Campus, Bengaluru-560100, India. e-mail: poulamisarkar101@gmail.com

Gowri Srinivasa
PES Center for Pattern Recognition and the Department of Computer Science and Engineering, PES University, Bengaluru, India. e-mail: gsrinivasa@pes.edu

techniques (such as histogram equalization) with application-specific augmentation (such as emulating a glare pattern) and an ensemble of five powerful, deep-learning based segmentation networks derived from the U-net and Linknet, followed by postprocessing that harnesses temporal information. With this pipeline, we obtain a final evaluation score of 0.9641, which is well above the baseline score of 0.840 (provided by Facebook as a part of the challenge) and comparable to the top scores reported in the public domain.

1 Introduction

The future of human-computer interaction is headed in the direction of Augmented, Virtual and Mixed Reality technologies (AR, VR, MR). A key component of virtual and mixed reality is the head mounted display, which allows users to interact with digital content and in the case of MR the real world as well. Eye-movement based tracking has been identified as a possible solution to not only control and interact with head mounted displays but also to enhance the quality of the virtual experience by increasing user immersion and engagement. This calls for accurate, efficient, and real-time computation of the human gaze [1].

Several eye-tracking methods can be found in the literature [2–6]. In almost all of these, the efficacy of tracking the eye is dependent on an accurate segmentation of key regions of the eye. Such a sensitivity to the accuracy of segmentation motivates a per pixel segmentation. Each pixel is thus classified as one of four different regions: the background, the sclera, the iris, and the pupil. There have been myriad research undertakings using segmentation in images, even in the specific field of segmenting the key regions of the eye. However, one of the more nascent applications in this domain is the use of segmentation to identify key features in video datasets from AR/ VR headsets, which offers an additional temporal dimension.

In this work, we utilize the dataset published by Facebook as part of the OpenEDS Challenge 2020. This dataset contains sequential frames of footage captured from a head-mounted camera [7]. We explore numerous techniques to segment the aforementioned regions of the eye and present a complete pipeline to preprocess video frames, train and validate the segmentation model, and fine-tune the predictions with image-processing methods. Our model, a hybrid of traditional image processing techniques, deep-learning based segmentation networks and postprocessing that incorporates temporal data, achieves a high score for the Mean Intersection-over-Union (mIoU) metric that significantly beats the baseline score achieved by the original OpenEDS paper and is on par with other state-of-the-art techniques for sparse semantic segmentation of features of the eye.

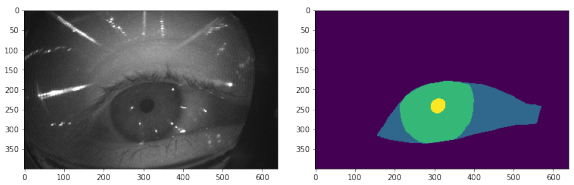
2 Related Work

Semantic segmentation of images and, in particular, semantic segmentation of the eye is a well researched field. Previous work carried out in this domain explored various aspects of segmentation of the different regions of the eye, such as the segmentation of the iris region of the eye under visible and near infrared light [2], and the segmentation of the sclera region of the eye [3]. Both these efforts, however, focus primarily only on segmenting out one out of the many different regions of the eye, rather than all regions simultaneously.

Research on the segmentation of multiple regions of the eye, including eyelashes as one of the segments, was not geared towards deployment on VR/MR headsets and thus, do not delve into leveraging the temporal information present in the image data available or techniques to deal with the noise present due to light emitted from the headsets [4].

Other efforts in this direction, such as those undertaken as a part of the original OpenEDS challenge, address the task of segmentation of the different eye regions factoring in their use case in the VR/MR domain. The focus of these efforts is on fulfilling the segmentation challenge under the constraint of having a low complexity model for improved runs times [5,6]. This constraint leads to faster prediction times but poorer accuracy of prediction. Furthermore, the original challenge made no mention of a temporal dimension. The need for a more accurate segmentation and a component in the pipeline that can harness the temporal dimension serve as the motivation behind the present work.

3 Dataset



(A) (B)

Table 1: A sample image from the dataset: (a) a video frame (image) showing structures of the eye and (b) the corresponding mask

The dataset provided by Facebook consists of video frames depicting the front view of the eye. The footage was gathered from a sample of 74 participants. The

frames were extracted from 200 video sequences sampled at 5 Hz with about 150 frames/sequence, thereby resulting in 29,476 extracted frames, for the segmentation task. Five percent of these images are labelled and constitute the base set. We further split the data, with a split ratio of 75 : 25, for model training and validation sets respectively. The remaining unlabelled data was sampled by Facebook to generate a test set for evaluation purposes. The masks for the images were of dimensions 400×640 to represent the class each pixel in the image belongs to. The labels for the pixels are as follows: 0 for the background regions exterior to the eye, 1 for the sclera, 2 for the iris, and 3 for the pupil. These numbers are used to color code regions and are represented as masks. For representations of the image and mask refer Table. 1

Some of the challenges in this data include extensive variations across the images. To begin with, not all images provide a view of the eyes being wide open that would make the task of semantic segmentation an easy one. Rather, many of the images depict the eye mid-blink or completely closed. Another major challenge is a saturation noise or glare on account of users wearing a pair of spectacles in addition to the headset, which creates patterns that pose a significant hurdle for segmentation. Further, some of the images are not sharp on account of motion blur. There are also images where the user has a noticeable amount of makeup on, mainly eyeliner and mascara, which adversely impacts the segmentation task. A variation ascribed to the temporal aspect of the data set is the transition between open and closed eye images, in the event of a blink, not being seemless. It should also be noted that the images are in grayscale and are not full color, i.e., Red, Blue, Green (RGB) images, which makes an accurate prediction of boundaries a challenging task. These variations in the dataset can be seen in the images depicted in Table. 2.

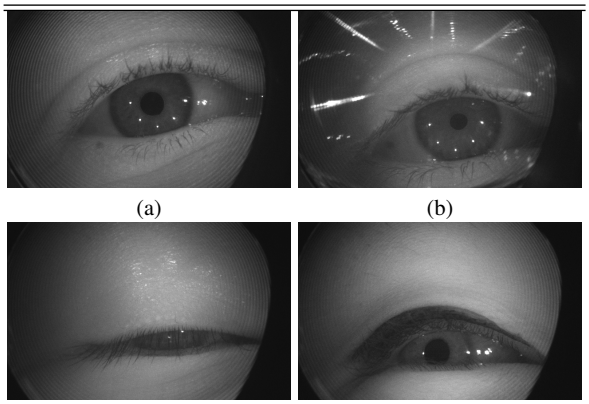


Table 2: Types of Noise in Images: (a) a normal image (b) an image with glare (c) mid-blink and (d) an image with eyeliner/mascara.

4 Design of the Hybrid Pipeline for Segmentation

The design of the hybrid pipeline for segmentation of the regions of the eye evolved from exploring the most proven, off-the-shelf deep learning models to incorporating traditional image processing techniques to increase the accuracy and incorporating state-of-the-art deep learning based segmentation networks with post processing that harnesses temporal information. This evolution is explicated below with an explanation of the strength of the model and limitations, leading to the design of the final pipeline.

4.1 Dual Model Approach and Image Augmentation

Our initial approach sought to leverage the Convolutional Neural Network (CNN) model described in [8] used in the previous iteration of the competition. The authors demonstrated the efficacy of a dual model approach to achieve greater accuracy while still utilizing an optimized model. We utilized the same approach, wherein the first model predicted the first three values of the segmentation masks that describe the relatively larger regions (the background, sclera and iris), while the second model distinguished between the third and fourth values that describe more detailed structures (the iris and pupil). The results obtained were stitched together to obtain the final mask, delineating all four regions of interest. Images used for this experiment were scaled down from a dimension of 400×640 to an image size of 96×160. The value chosen was to ensure the dimensions were $M\%32 \times N\%32$ while still trying to retain the aspect ratio. This model, however, incorporated no preprocessing nor image augmentation techniques achieved an mIoU of only 0.84 on the validation set.

In an attempt to improve the performance we integrated the ImageDataGenerator module from Keras to perform the inbuilt flip, rotate, shift, and zoom operations along with various custom augmentations such as adding noise in the form of random bright spots, image brightening, Gaussian blurring, and random white lines in a circular pattern to emulate the pattern caused due to the glare present in few images. With the addition of these augmentations the model performance improved marginally to achieve an mIoU of 0.9 on the validation set. In a final attempt to improve performance, a technique known as Attention Gating [9] was embedded into the models which further increased the mIoU to 0.95 on the validation set. Although the model achieved good results on the validation set, the score achieved on the test set was only a marginal improvement over the benchmark score. This prompted a move from a dual model structure to a single unified U-Net architecture as described in the next subsection.

4.2 Image Preprocessing and The Switch to U-Net

The dual-model approach in [8] was tailored for the constraints of the previous iteration of the OpenEDS challenge which imposed a size penalty on models. Given that this model parameter size restriction is not applicable for the data set under consideration, we switched architectures to the U-Net architecture mentioned earlier [10].

The U-Net approach resulted in a significant increase in accuracy over the dual-model approach previously described. We further substituted the white-line augmentation used in the previous approach with a glare augmentation layer that was extracted from the original eye images to better represent the patterns present in the data. A key difficulty experienced by the model was the reflection present in images where the participants wore a pair of spectacles. This caused the sclera region to be misclassified. In order to amend this, we tried additional image preprocessing approaches.

We experimented with a variant of histogram equalization known as Contrast Limited Adaptive Histogram Equalization, shortened to CLAHE, to reduce the sharp glare. CLAHE operates on tiles (sub-regions of the image) rather than the whole image. However, we found that the traditional histogram equalization from the OpenCV package resulted in better overall performance. Another noticeable aspect of the dataset was that most of the images were dark. To tackle this problem, we made use of image brightening methods which helped highlight the sclera of the eye more prominently, although it should be noted that the performance on the other regions of the eye had a negative impact. Finally, Gaussian Blurring technique was applied in order to reduce the effect of noise present in almost all of the images.

Along with the vanilla U-Net architecture, we also tried a variation of U-Net that used the aforementioned Attention Gating mechanism. This technique, while providing a noticeable performance jump when compared to the vanilla U-Net, with scores considerably higher than the baseline score, still failed to match the performance of other state of the art models. This prompted research into other models which yielded in the discovery of the segmentation models package that provided a variety of inbuilt models with multiple options that can be customized, such as altering the backbone architecture, varying the activation functions, etc. [11].

A key experimentation performed at this stage was leveraging the temporal information present in the data. The final prediction was a combination of the current prediction along with a weighted influence of the previous prediction. This was done as the images are sequential frames taken from a video and thus most of the regions locations from one image to another do not change drastically.

To achieve high performance, finding the optimal value of current to previous prediction ratio was imperative. To that extent, various weights ranging from 0.1 to 0.5 were experimented with and it was found that a weight of 0.2 to the previous image prediction gave the best accuracy.

$$X_n = X_n + (0.2 * X_{n-1}) \tag{1}$$

4.3 Segmentation Models, Experimentation with Loss Functions and Model Ensembling

Using the segmentation models package, we experimented with the U-net architecture with a ResNet34 backbone. Data set properties were also varied, with images being resized to 640×384 being the closet dimensions to the original data which is still $M\%32 \times N\%32$. This is done to avoid a loss of information during the process of scaling. Other changes made were moving to single channel images, as opposed to the three channel images that were being used for training up until this point. The loss function was also changed to use more sophisticated losses, given the complexity of the problem, making a combination of Dice loss and Jaccard loss as the final loss function of choice [12].

With the segmentation models package, it was now possible to use model ensembling given the variety of models offered. Our experimentation at this stage made use of an ensemble comprising U-Net and LinkNet architectures, both having ResNet34 backbone. Multiple U-Net and LinkNet models were trained on two sets of grayscale images, the original and an inverted counterpart (with pixel values being inverted), with one of the U-Net models in the ensemble being additionally equipped with CLAHE to preprocess the images.

Finally, we looked into freezing encoder weights by training the model for a few initial epochs and then training for a few more epochs where only decoder weights are tuned during training, but this specific technique did not yield any improvement in performance. All these experimentations led to the final model implementation which has been described in Section 5.

4.4 Experimentation with Nascent Techniques

There was no significant improvement in performance with the addition of other augmentation techniques and changes in model architecture which prompted us to further explore more nascent pre-processing and post-processing techniques such as Conditional Random Fields (CRF) with Superpixeling.

Superpixeling is the process of partitioning the image into multiple segments (superpixel) based on some common characteristics like pixel intensity. This helps in creating a clear delineation between the different regions present in an image which can potentially improve segmentation performance. Conditional Random Fields on the other hand are a discriminative statistical modeling method used when the class labels for the various inputs are not independent. CRF provides predictions for the data considering the input features along with the labels of all other inputs which are dependent on the current input. In CRF, the data is structured as a graph consisting of a set of nodes V and edges between the nodes E. An edge between a node i and a node j in the graph denotes the output label of data points i and j are dependent. CNN's sometimes fail to accurately predict the complex boundaries between classes at the pixel level resulting in slightly distorted image masks. Superpixel-enhanced

pairwise conditional random fields are often used in segmentation to correct these minor inconsistencies and generate more accurate boundaries leading to a state of art accuracy. However, these techniques did not provide any boosts in performance and were thus dropped leading to the final implementation explained next.

5 Final Hybrid Pipeline and Results

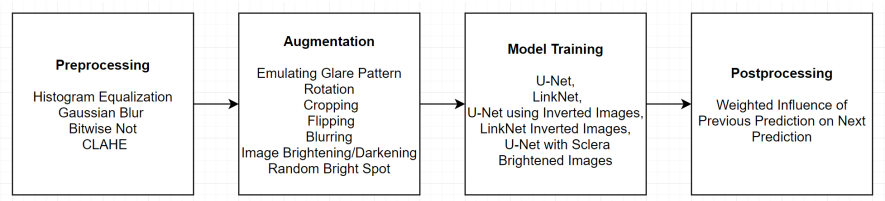


Fig. 1: A hybrid pipeline for the segmentation eye regions from video frames

A schematic diagram of the final hybrid pipeline is depicted in Fig. 1. The first step involves preprocessing using the best methods identified in the experimentation steps described above. The second step is to perform image augmentation, as shown in Table. 3, to increase the size of the training set. Then, the model ensemble is trained on the images. Finally, to account for the temporal dimension of the data set, a postprocessing step where the output from the previous prediction is weighted and fed back to influence the next iteration of training.

The ensemble comprises of five models which are a combination of U-net and Linknet architectures provided by the segmentation models package. The base model pair consists of a U-Net and Linknet model duo trained on images pre-processed with the Histogram Equalization and Gaussian Blurring techniques mentioned earlier. The secondary model pair is trained on images processed additionally using the Inversion technique. This pair contributes to increasing inference accuracy over the iris region. The fifth component of the ensemble makes use of CLAHE and ConvertScaleAbs preprocessing methods that increase image brightness and the prominence of the sclera region. These images are trained using U-Net.

We define a custom loss function that uses the sum of Dice and Jaccard losses. All five models in the ensemble use this loss for the training process. The models are trained for 128 epochs using the Adam optimizer and fine-tuned for a further 16 epochs using SGD optimizer with following parameters - learning rate = 0.0001, decay = 10^{-6} and momentum = 0.9.

The predictions from the individual models are averaged to compute the output mask. The masks are generated sequentially, one frame at a time. At each step, the mask prediction for the previous video frame is added to the current prediction as mentioned earlier in Equation (1). This results in the final output mask for the frame.

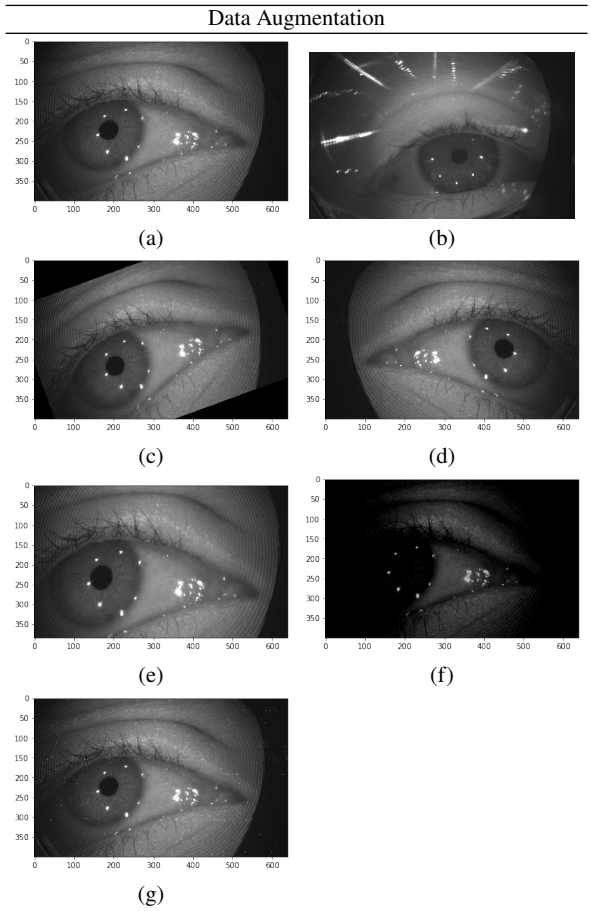


Table 3: Different transformations on the data for augmentation: (a) original image, (b) introducing a glare (starburst pattern), (c) rotating the image, (d) lateral inversion, (e) cropping (and resizing) the image, (f) reducing brightness and (g) introducing random bright spots in the image.

A visual depiction of the performance of the hybrid pipeline model is presented with representative images in Table 4.

The first column in Table 4 presents the original video frame, the second column presents the ground truth (annotated mask) and the third column presents the segmentation mask obtained using the hybrid pipeline. Visually, there is very little difference between the ground truth masks and those from the hybrid pipeline. To better understand the results, we present a quantitative comparison of the performance of various models in Table 5 to an accuracy of two decimal places. The hybrid pipeline model yields an mIoU of 0.9641, providing an accurate segmentation of

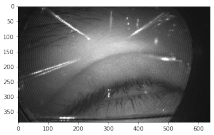
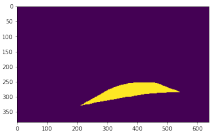
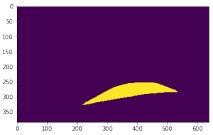
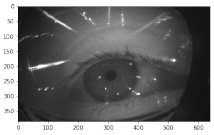
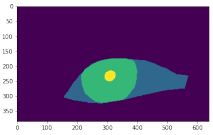
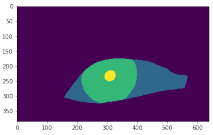
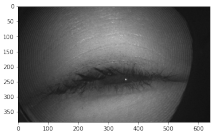
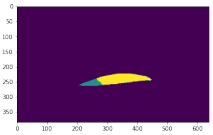
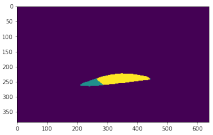
Image	Ground Truth	Hybrid Pipeline
(a) Prediction on an image with limited regions and glare		
		
(b) Prediction on an image with all regions and glare		
		
(c) Prediction on an image with limited regions		
		

Table 4: A visual analysis of segmentation performance

even challenging images and making the most of the temporal information available. When used as a part of a larger pipeline such as gaze tracking, etc., achieving as accurate a segmentation of the regions of the eye as possible, would only ensure better efficacy for the larger system.

Segmentation Model	**mIoU**
Dual Model CNN	0.84
Dual Model CNN + Data Augmentation	0.90
Dual Model CNN + Attention Gating	0.95
Hybrid Segmentation Model (see Fig. 1)	**0.96**

Table 5: A quantitative comparison of segmentation performance of various models

6 Conclusion

This paper details the rationale and evolution of models to develop a state of the art pipeline to perform semantic segmentation of eye images obtained from video

sequences recorded using VR/MR headsets. The final pipeline comprises a hybrid of traditional image preprocessing techniques with an ensemble of deep-learning based segmentation networks followed by a postprocessing component that factors in temporal data, through using the mask generated on the previous image in the sequence to correct minor inconsistencies in the prediction. The use of standard architectures from the segmentation models package in the ensembled form contributes to the ease of training and execution. The final results achieved, quantified with an mIoU score of 0.9641 is seen to be very close to the ground truth. With the high mIoU score achieved, this pipeline is ideal for use in AR/VR/MR applications, to provide high quality, immersive and interactive experiences or for any other applications, such as gaze tracking, that require a semantic segmentation of key regions of the eye.

References

1. S. J. Garbin, Y. Shen, I. Schuetz, R. Cavin, G. Hughes, and S. S. Talathi, "Openeds: Open eye dataset," 2019.
2. W. Sankowski, K. Grabowski, M. Napieralska, M. Zubert, and A. Napieralski, "Reliable algorithm for iris segmentation in eye image," *Image and Vision Computing*, vol. 28, no. 2, pp. 231 – 237, 2010, segmentation of Visible Wavelength Iris Images Captured At-a-distance and On-the-move. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885609001103
3. A. Das, U. Pal, M. A. Ferrer, M. Blumenstein, D. Stepec, P. Rot, Z. Emersic, P. Peer, V. Struc, S. V. A. Kumar, and B. S. Harish, "Sserbc 2017: Sclera segmentation and eye recognition benchmarking competition," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 742–747.
4. P. Rot, Z. Emersic, V. Struc, and P. Peer, "Deep multi-class eye segmentation for ocular biometrics," in *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, 2018, pp. 1–8.
5. F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Eye-mms: Miniature multi-scale segmentation network of key eye-regions in embedded applications," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.
6. J. Perry and A. Fernandez, "Minenet: A dilated cnn for semantic segmentation of eye features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.
7. C. Palmero, A. Sharma, K. Behrendt, K. Krishnakumar, O. V. Komogortsev, and S. S. Talathi, "Openeds2020: Open eyes dataset," 2020.
8. A. Rao, A. Mysore, S. Ajri, A. Guragol, P. Sarkar, and G. Srinivasa, "Automated segmentation of key structures of the eye using a light-weight two-step classifier," *Journal of Intelligent and Fuzzy Systems*, pp. 1–7, Mar 2021.
9. J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," 2018.
10. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597
11. P. Yakubovskiy, "Segmentation models," https://github.com/qubvel/segmentation_models, 2019.
12. C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," *CoRR*, vol. abs/1707.03237, 2017. [Online]. Available: http://arxiv.org/abs/1707.03237