Research on Biomedical Engineering manuscript No.
(will be inserted by the editor)

Better Feature Extraction using Multi-Encoder Convolutional Neural Networks for Optic Cup Segmentation from Digital Fundus Images

Ambika Sharma · Monika Agrawal ·
Sumantra Dutta Roy · Vivek Gupta

Received: date / Accepted: date

Abstract
Purpose Glaucoma is an eye disease that is chronic, asymptomatic, and cannot be cured once it progresses. An important step in clinical analysis of glaucoma is to measure the cup-to-disc ratio (CDR). Optic cup segmentation is a challenging task (as compared to detecting the optic disk, for instance), due to poor contrast on the cup boundary region, and occlusion from veins and arteries. Contemporary systems are based on image processing/computer vision and/or machine learning. However, obtaining accurate optic cup segmentation over large datasets is still a challenge.
Methods We propose a novel asymmetric 'multi-encoder U-Net'/Y-Net architecture with Inception and context blocks in the bottleneck layer. The architecture has an ResNet34-based primary encoder and a light-and-efficient EfficientNetB0 auxiliary encoder. The asymmetry involves avoiding multi-stage skip connections from the auxiliary encoder to the decoder. This avoids the complexity of feature map concatenation at different levels. The Inception block in the bottleneck layer performs feature enrichment. Different receptive fields in parallel paths result in multi-scale optic cup features. The next cascaded context block helps maintain spatial consistency of the multi-scale feature maps.
Results and Discussion We have experimented extensively on four public datasets, and the challenging AIIMS community camp (private) dataset. The proposed network outperforms the state-of-the-art with an average Dice coefficient of 91.11%

Ambika Sharma
Indian Institute of Technology Delhi
E-mail: ambika.sharma@dbst.iit.ac.in

Monika Agrawal
Indian Institute of Technology Delhi
E-mail: maggarwal@care.iitd.ac.in

Sumantra Dutta Roy
Indian Institute of Technology Delhi
E-mail: sumantra@ee.iitd.ac.in

Vivek Gupta
All India Institute of Medical Sciences, New Delhi
E-mail: vgupta@aiims.ac.in

and 87.77% on the Drishti-GS (Sivaswamy et al. (2014)) and Refugee (Maninis and Pont-Tuset (2010)) public datasets. Our ablation studies with different competing architectures also shows the proposed method achieving the highest Dice coefficient and cup overlap percentage. The training itself achieves a much lower train-validation loss, as seen over a large number of epochs.

Conclusion The novel architecture has each sub-part geared towards getting good optic cup segmentation performance across a large number of datasets. The network shows robust segmentation performance on challenging images with various retinal artifacts (blurring, poor illumination, and clinical pathologies).

Keywords Retinal Images · Optic Cup · Deep Convolutional Neural Network (DCNN) · U-Net · Image Segmentation

1 Introduction

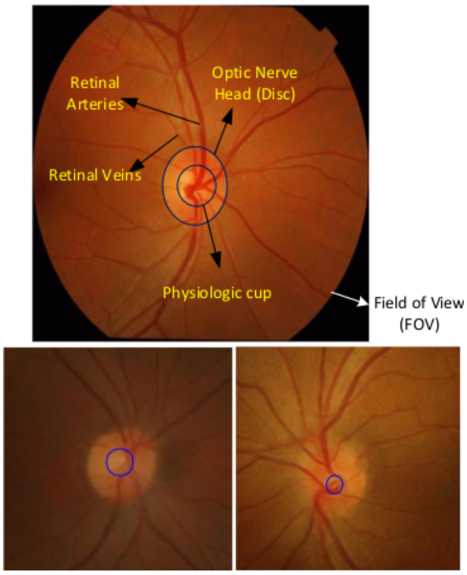


Fig. 1: The top row shows a sample retinal image with a few landmarks marked: the Optic Cup/'Physiologic Cup' (the subject of this paper), the optic nerve head (or the Optic Disc), and retinal blood vessels (arteries and veins) marked with black arrows. Optic cup segmentation is difficult due to the low contrast at the cup boundaries and occlusion from blood vessels. Sec. 1 has more details. The second row shows two images with varying optic cup sizes, from the Drishti-GS public dataset (Sivaswamy et al. (2014)), with the the optic cup marked with blue circles.

Robust optic cup segmentation forms an important step in early diagnosis of glaucoma (an irreversible eye disease) (Muramatsu et al. (2009)), which involves the computation of the vertical cup-to-disc ratio (CDR, hereafter). Optic cup segmentation is a challenging task (compared to detecting the optic disk, for instance), due to poor contrast on the boundary of the optic cup (as in Fig. 1), and occlusion from blood vessels (the veins and arteries, as in Fig. 1)(Muramatsu et al. (2009)). Retinal images from fundus cameras or ophthalmoscopes are the only cost-effective means in this regard, as compared to the expensive and more informative Optical coherence tomography (OCT), Scanning Laser Ophthalmoscopy (SLO), and Confocal microscopy. OCT for instance, gives depth information and intra-ocular pressure (IOP) estimates, which are useful in glaucoma identification. However, these three techniques need specialized setup arrangements, and cannot be used in the field, in community eye camps, for instance.

Fig. 1 shows the difficulty in optic cup segmentation. Only in a healthy retina, the optic cup is generally horizontally oval in shape (8% greater than the vertical extent) and yellowish-white in color (Algazi et al. (1985)). Color alone cannot form a basis for optic cup segmentation (Dada and Coote (2010)). It does not help that the local contrast in the region is often poor. It is difficult to estimate the blood vessel bends at the cup and those at other regions, owing to the poor contrast (Dada and Coote (2010)). Moreover, there is occlusion from the blood vessels themselves. Optic cup segmentation approaches generally fall into two major categories namely, Image processing/computer vision-based techniques (Sec. 1.1) and deep convolutional network-based pixel-level classification (Sec. 1.2).

1.1 **Image Processing/Computer Vision-based Approaches**

Wong et al. (2009) seek to detect blood vessel kinks. in retinal images. The authors use a level set-based method with Canny edge detection and wavelet transform techniques in a probabilistic framework. Thakur and Juneja (2019) use a level set-based Adaptively Regularized Kernel-Based Intuitionistic Fuzzy C Means clustering-based approach. A related approach is a Fuzzy c-Means (FCM) clustering algorithm with morphological operations (Khalid et al. (2014)). For glaucoma diagnosis, Mittapalli and Kande (2016) use the gray level change near the boundary. They extract bends in small blood vessels using spatially weighted fuzzy c means (SWFCM) clustering-based thresholding. Examples of superpixel classification-based approaches are Tan et al. (2015), Mohamed et al. (2019), and Xu et al. (2014). These methods extract use histogram and textural image features. Sanfilippo et al. (2010) use geometric morphometric methods of elliptic Fourier analysis and sliding semi-landmark analysis with a minimum bending energy criterion. They eliminate the variation unrelated to shape (i.e., location, size, and orientation) and obtain a series of PCA-summarized shape variables. Stereo-based approaches need a specialized setup. Chakravarty and Sivaswamy (2017) perform depth-based cup extraction using a boundary-based conditional random field (CRF) representation of depth between optic disc and cup. A similar method uses color difference and vessel bends (Hu et al. (2017)) to locate the optic cup boundaries. The aggregation uses confidence values.

1.2 **Machine Learning-based Approaches**

In biomedical image segmentation problems, U-Net architectures (an encoder and decoder with skip connections) are very common. Yu et al. (2019) use a ResNet-34 based pre-trained an encoder along with classical U-Net decoding layers. A joint optic disc and cup segmentation (Liu et al. (2019b)) uses a conditional Generative Adversarial Network (GAN) framework. It uses a segmentation net S, a generator G, and discriminator D network. Both the segmentation and generator networks are trained to learn the bidirectional mappings between fundus image and segmentation maps. The segmentation net learns the mapping from the fundus to the binary image, and generator learns the mapping from binary image to the fundus image. All networks are trained simultaneously with adversarial and segmentation generator network reconstruction losses (segmentation and generator networks). Kamble et al. (2020) use an EfficientNet as an encoder in a U-Net++ framework. Hybrid methods such as Liu et al. (2019a) propose novel spatial distribution-aware maximum conditional probability framework for joint optic disc and cup segmentation. The methodology is based on the explicit variance of the spatial layout of vessels, and the spatial sparsity of blood vessel kinks at a small scale. The classification neural network consists of an atrous CNN module, a pyramid filtering module comprising of M parallel pyramid filtering blocks. and a spatial-aware segmentation module of M parallel spatial-aware segmentation blocks.

We summarize the main features of our work as follows:

- We propose a novel asymmetric Y-Net (multi-encoder U-Net)-based architecture with an Inception block and a multi-kernel context block in the bottleneck layer. A pre-trained ResNet34 acts as the primary encoder, and an EfficientNetB0, the auxiliary encoder. The decoder uses the skip connections only from the primary encoder and up-samples the concatenated feature maps to the original image dimension.
- The bottleneck layer has a cascade of a multi-scale Inception block and multi-kernel context block. The multi-scale Inception block captures multi-scale features through different receptive fields in parallel paths. The context block maintains spatial contextual information from the Inception block.
- Retinal images from community camps and healthcare setups are taken with low-cost hand-held ophthalmoscopes. Blurriness and poor illumination are common in such cases. Retinal pathologies (such as exudates, haemorrhages, lesions and atrophy) are also common since the subjects often have poor access to even basic healthcare. In addition to 4 public datasets (Drishti-GS (Sivaswamy et al. (2014)), DRIVE, DRIONS and Refugee (Maninis and Pont-Tuset (2010)), we have also experimented with the challenging (private) AIIMS community camp dataset.
- We show the results of extensive experiments with different segmentation networks and their combinations, vis-a-vis our proposed network. We show qualitative and quantitative results of the proposed network outperforming the state-of-the-art.

The organization of the rest of the paper is as follows. Sec. 2 gives details of the proposed architecture: the asymmetric 'multi-encoder U-Net'/Y-Net with Inception and context blocks in the bottleneck layer. This section explains the motivation behind the choice of each architectural sub-part in the proposed network. Sec. 3 presents results of detailed experiments with the proposed architecture, with four public datasets, and the challenging AIIMS community camp (private) dataset. The section describes suitable metrics for optic cup segmentation, system implementation details, compares performance parameters with the state-of-the-art in the area, a comparison of competing architectural structures, and the results of the proposed method across datasets. Sec. 5 puts the work into perspective, and concludes the paper.

2 Methods

2.1 A Novel Asymmetric 'Multi-Encoder U-Net' (Y-Net) with an Inception Block and a Context Block in the Bottleneck Layer

The paper proposes a novel asymmetric 'multi-encoder U-Net'/Y-Net with an Inception block and a context block in the bottleneck layer. Fig. 2 gives an overview of the proposed architecture, highlighting its main components. In what follows, we explain features of each component, and illustrate their suitability to the task of optic cup segmentation from retinal images. The architecture proposes a novel feature extraction algorithm using multi-scale context modeling, with two independent and asymmetric top-level segmentation networks. This is quite different from two networks with similar names. The Y-Net of (Mehta et al. (2018)) is not a multi-encoder network: it is a network for joint segmentation and classification. Our asymmetric multi-encoder Y-Net differs from the symmetric Y-Net of (Mohammed et al. (2018)), and has additional bottleneck layer Inception and context blocks.

2.1.1 Motivation for a Multi-Encoder Structure

The conventional U-Net architecture (Ronneberger et al. (2015)) is now commonplace for biomedical segmentation tasks. The U-Net is an encoder-decoder structure with skip connections between the encoder (contracting path) and the decoder (expanding path, to the original image dimensions). However, with an increase in the depth (levels) of the network, the number of weights increases due to an increase in the number of filters. During backpropagation-based learning methods, these weights do not get updated much due to low gradients. Convergence will be slower at the deeper layers. During backpropagation, by the time we reach the initial layers, the weights will be almost the same, implying bypassing the initial layers (which are responsible for 'abstract' features, in the first place). One can have pre-trained architectures at the encoder side to alleviate training data scarcity issues commonly associated with medical datasets. (In other words, one trains only the decoder side). In fact, in our results section (Sec. 3.6), we show comparative segmentation results of optic cup segmentation of a basic U-Net, and variants with pre-trained encoder structures (Table 4). Having a single encoder biases the segmentation network into using only one type of features. A multi-encoder structure can take advantage of information from different sets of features.

2.1.2 A ResNet34-based Primary Encoder, and a EfficientNetB0-based Auxiliary Encoder, in an Asymmetric Setting

We propose a dual encoder set up, to take advantage of different sets of input features. This is inspired by ensemble learning models, which select diverse non-

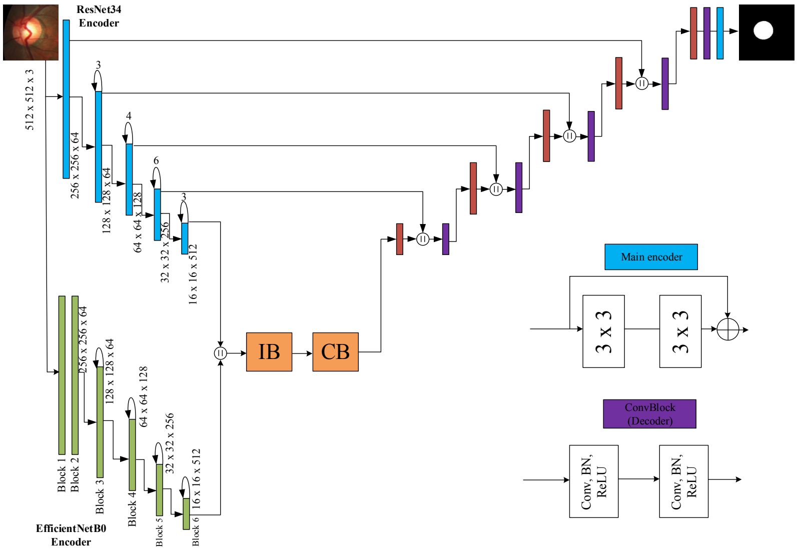


Fig. 2: The proposed asymmetric 'multi-encoder U-Net'/Y-Net architecture with Inception and context blocks in the bottleneck layer. The multi-encoder part takes features from two independent architectures (A ResNet34 primary encoder and a light-and-efficient EfficientNetB0 auxiliary encoder). The asymmetry in the two encoders is to avoid the complexity of feature map concatenation at different levels. The bottleneck layer has a cascade of an Inception block and a context block. The role of the inception block is to capture multi-scale features through different receptive fields in parallel paths. The context block maintains spatial contextual information from the Inception block. The right bottom of the figure shows some details of an encoder and decoder block (in blue and violet, respectively). 'BN' represents batch normalization, 'Conv' convolution, and 'ReLU', the rectified linear unit activation function. Further, the dark brown sub-blocks in the decoder path represent the up-sampling operation. Sec. 2.1 has the relevant details of the network and sub-blocks. Fig. 4 shows the Inception and context blocks in detail. Fig. 3 shows a conceptual representation of our asymmetric multi-encoder structure.

redundant features from two or more different architectures. We use a ResNet (He et al. (2016)) architecture-based encoder (ResNet34) as a primary encoder. This uses pre-trained weights from the large and diverse ImageNet dataset (Deng et al. (2009)). The choice of a ResNet architecture is governed by its deeper structure with comparatively fewer parameters, owing to residual connections. The previous section mentions the U-Net issue of bypassing information from the initial layers. A pre-trained ResNet-based U-Net in the encoding layers not only provides high-level features, it also maintains the abstract features from the initial layers. Another interesting facet is the presence of skip connections not just in the encoder-decoder structure: skip connections are there in the ResNet34 primary encoder as well.

We choose an EfficientNet (Tan and Le (2019)) architecture-based encoder (EfficientNetB0) as the auxiliary encoder. The EfficientNet family of pre-trained

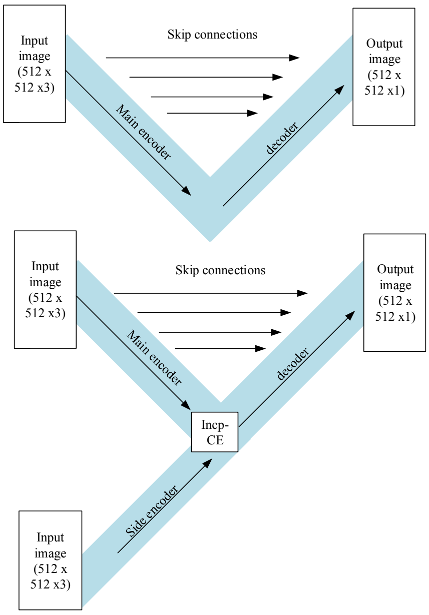

Input image (512 x 512 x3)
Skip connections
Output image (512 x 512 x1)
Main encoder
decoder
Input image (512 x 512 x3)
Skip connections
Output image (512 x 512 x1)
Main encoder
decoder
Incp-CE
Side encoder
Input image (512 x 512 x3)

Fig. 3: A conceptual description of a standard U-Net (above) and our asymmetric 'multi-encoder U-Net' (Y-Net) (below), with two encoders. Our asymmetric Y-Net architecture has a primary encoder, and a lighter auxiliary encoder with an independent set of features. The auxiliary encoder (with no skip connections to the decoder side) avoids the complexity of feature map concatenation at different levels. Sec. 2.1.2 has the details.

networks (EfficientNetB0-B7) provides a disciplined approach to scaling network dimensions: the width, depth and image resolution. EfficientNet architectures are light and efficient compared to contemporary convolutional neural network models (Tan and Le (2019)). In a standalone EfficientNet, the small number of parameters needs less training time and acts as a light model, without compromising much on overall performance. The EfficientNetB0 architecture has MBConv as the main building block (which consists of seven inverted residual blocks, each with a different setting). Further, these blocks are composed of squeeze and excitation blocks, and have Swish activation functions (Tan and Le (2019)). We choose the light-and-efficient pre-trained EfficientNetB0 as the auxiliary encoder. (The pre-training is on the ImageNet dataset (Deng et al. (2009)), again.)

The concatenation of feature sets from the primary and auxiliary encoders happens in a multi-scale context block at the bottleneck level (after the last down-sampling step). Fig. 3 shows this concatenation and the asymmetric part of the architecture as well. (Fig. 3 compares the basic philosophy of our asymmetric Y-Net architecture to that of a U-Net, as well.) A separate section (Sec. 2.1.3 explains the Inception block and the context block in detail. In the decoder part, the primary encoder (ResNet34) skip connections remain the same as the conventional U-Net. This re-localizes high-level features to the full-resolution optic cup binary mask. The asymmetry in the architecture comes from the auxiliary EfficientNetB0

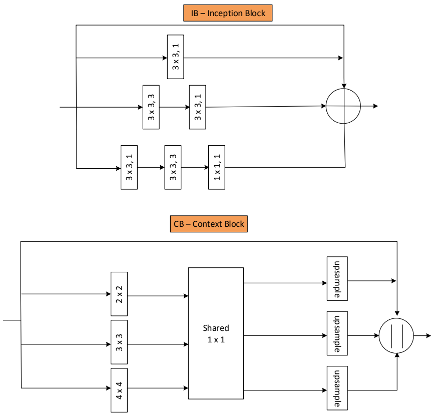


Fig. 4: Details of the Inception and context blocks. The Inception block (at the top) consists of three parallel convolution operations, with each path having a different receptive field dimensions in order to capture the multi-resolution features for the optic cup region. The context block (at the bottom) extracts the contextual information from the above features and maintains the spatial consistency of the feature map. Sec. 2.1.3 has the details of the two blocks.

encoder not having skip connections on the decoder side. (Fig. 2 and Fig. 3 show this asymmetry in detail.) This avoids the complexity of feature map concatenation at different levels. Fig. 3 presents this asymmetry, and the basic conceptual difference between a conventional U-Net structure, and the proposed asymmetric 'multi-encoder U-Net'/Y-Net (with Inception and context blocks in the bottleneck layer). A separate section (Sec. 2.1.4) describes the decoder structure in detail.

2.1.3 ***Multi-scale Contextual Features in the Bottleneck Layer: An Inception Block and a Context Block in Cascade***

As mentioned in the previous section, the feature concatenation from the primary and auxiliary encoders takes place at the bottleneck layer. This has the Inception block and the context block. Fig. 4 shows a diagrammatic representation of the architecture of these two blocks. The concatenated $16 \times 16 \times 512$ bottleneck feature maps from the primary and auxiliary encoders are fed to a multi-scale Inception block that extracts the more detailed high-level features. The choice of an Inception architecture (Szegedy et al. (2015)) is to have parallel convolution paths with different-sized receptive fields, to capture multi-resolution features for the optic cup region. Even though the primary and auxiliary encoder networks capture multi-level feature maps, the process is limited by the scaling and pooling operations with just a kernel size of 3×3 and pooling by a factor of 2. To alleviate this constraint, we use a multi-scale atrous/dilated convolution with residual connections. This step not only enhances the detailed information, this also prevents the loss of semantic segmentation to some extent (Chen et al. (2017)). In addition to this, the retinal image dataset used in this work has been generated from different regions and patients, where the cup may have different sizes and shapes. To cover all sorts of cup sizes, Inception blocks can bring out most of the multi-scale semantic features for segmentation. The top part of Fig. 4 shows the detailed description of the Inception block, where multiple atrous convolutions are performed and finally integrated with an input residual connection. This module accounts for small to large optic cup sizes by providing a range of receptive fields (larger receptive fields for larger optic cups, and vice-versa).

The bottom part of Fig. 4 shows a diagrammatic representation of the architecture of the context block. The high-level features from the Inception block are fed to a context block. The context block generates contextual information from the bottleneck feature maps and preserves spatial information from precise cup segmentation. The block consists of multi-pool kernel configurations ranging from size 2×2 to 4×4. Further, the output from each pooling operation is gathered by a shared convolutional layer with a kernel size of 1×1 to generate a one-channel image map. Finally, these one-channel maps are up-sampled to the original feature map dimension and concatenated all together with the original input feature map using a skip connection. Fig. 4 gives a pictorial representation of the process, with three different kernel sizes.

2.1.4 ***The Decoder Network***

The decoder module (Fig. 2 and Fig. 3) recovers the binary cup map (with the same resolution as the RGB input image) from the deep encoded features. The high-level features from the multi-scale context block (Sec. 2.1.3) are fed as an input to the decoder. The skip connections from the primary encoder (ResNet34) are fused with the deep-semantic features. The EfficientB0 network is excluded from the decoder side, due to the high complexity (parameters) of feature maps when concatenated at different levels. (In our experiments, putting these in increased the training time, without any significant improvement in the performance.) In our system work, the decoder network consists of five decoder blocks. Each block comprises of an up-sampling layer of kernel size 2 followed by a concatenation operation with the skip connections. Lastly, a ConvBlock is added to decode the much larger feature maps from the previous layer, to the binary cup mask. The ConvBlock is a series of two 3×3 convolutional operations followed by batch normalization operation and ReLU activation function. Fig. 2 shows all decoder blocks along with their sub-blocks. In the last step of the decoder network, the convolution operation with one feature map and sigmoid activation function is employed to predict the cup probability of each pixel. It is interesting to note that instead of using residual skip connections, we employed concatenated skip connections when fusing the deep semantic features with shallow high-resolution features. This seems to improve the performance of optic cup segmentation. Further ahead, Sec. 3.6 (Table 4 summarizes the results) presents a detailed ablation study of different encoder and decoder setups.

3 Results

3.1 **Datasets**

For the evaluation of the proposed work, we use four public datasets namely Drishti-GS (Sivaswamy et al. (2014)), DRIONS, DRIVE and Refugee (Maninis and Pont-Tuset (2010)). We have also experimented with the challenging AIIMS community camp (private) dataset. Table 1 shows details about all the datasets used in this work. The Drishti-GS (Sivaswamy et al. (2014)) and Refugee (Maninis and Pont-Tuset (2010)) datasets have the optic disk margin ground truth available. For the others: DRIVE, DRIONS and the challenging AIIMS community camp (private) dataset, we had the optic cup ground truth segmentation given by experienced ophthalmologists from AIIMS New Delhi. The Drishti-GS (Sivaswamy et al. (2014)) and Refugee (Maninis and Pont-Tuset (2010)) datasets have been divided into a 50:50 train-test ratio (which we have tweaked, to have a 40:10:50 train-validate-test ratio), whereas the rest are divided in a 80:10:10 train-validate-test ratio. There are no particular reasons for these split ratios. For the Drishti-GS dataset (Sivaswamy et al. (2014)), this division is already done by the dataset provider. Further, others researchers have also tested their work on 51 images (the dataset has 101 images in all). We have tweaked this to have a 40:10:50 train-validate-test split. The situation is similar for the Refugee dataset (Maninis and Pont-Tuset (2010)) in the MICCAI competition: the train-validate-test split is given as 400:400:400. For the other datasets DRIVE, DRIONS and the challenging AIIMS community camp dataset, there is no *a priori* train-test (or train-validate-test) split. For these, we choose an 80:10:10 train-validate-test split, a fairly common method method used for small-sized datasets. Further, these datasets were collected for other ophthalmological tasks, and are not ground-truthed. For these cases as mentioned above, experienced ophthalmologists (at AIIMS New Delhi) manually marked the optic cup ground truth for our work.

Table 1: Retinal image datasets, and their details. We use four public datasets Drishti-GS (Sivaswamy et al. (2014)), Refugee, DRIONS, DRIVE (Maninis and Pont-Tuset (2010)), and the challenging AIIMS community camp (private) dataset. Sec. 3.1 has the details, including the choice of the train-test split for these datasets.

S.No	Dataset Name	Total Images	Image Dimension
1	DRIONS	110	600 × 400
2	DRIVE	40	768 × 584
3	Drishti-GS	101	2896 × 1944
4	Refugee	1200	1634 × 1634
5	AIIMS (private)	364	1536 × 1584

3.2 Evaluation Metrics for Optic Cup Segmentation

The standard representation of the statistical performance of a particular method is in terms of the ROC parameters/the confusion matrix parameters, in various forms. These parameters are true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). For our cup segmentation task, TP indicates those pixels which the segmentation method identifies as being from the cup region, and which are corroborated by the ground truth as well. TN counts the number of non-cup pixels correctly classified as background (or non-cup) pixels. FP indicates the number of background pixels misclassified as being from the cup region. FN represents the ground-truth cup pixels that are misclassified as background (or non-cup) pixels.

The Dice coefficient (F1 score) (Eq. 1) and the Jaccard coefficient (Intersection-over-Union (IOU), or the percentage overlap) (Eq. 2) are the most commonly used performance measures in segmentation problems.

$$\textit{Dice coefficient (F1 score)} = 2 \times \frac{(TP)}{TP + FN + FP} \tag{1}$$

$$\textit{Intersection over Union (Jaccard coefficient, Overlap)} = \frac{(TP)}{TP + FN + FP} \tag{2}$$

For instance, the IOU/Jaccard coefficient also represents the percentage overlap between between the predicted cup segmentation map, and the ground-truth binary mask.

$$\textit{Sensitivity (Recall)} = \frac{(TP)}{TP + FN} \tag{3}$$

$$\textit{Specificity} = \frac{(TN)}{TN + FP} \tag{4}$$

As mentioned above, the Dice coefficient/F1 score and the Jaccard coefficient/IOU/cup overlap percentage are suitable parameters. ROC analysis also cites sensitivity and specificity values. In the medical informatics tasks, sensitivity is the more significant of the two, since it gives a measures of false negatives (which are crucial in disease diagnosis). To this end, Tables 2, 3, 4 and 5 all mention the Dice coefficient, the cup overlap percentage and the sensitivity values. We also note that the notion of specificity is intrinsically captured through the use of the Dice coefficient.

3.3 **System Implementation Details**

The input images to the system are semi-automatically processed as follows. Unnecessary background removal happens through a process of optic disk localization (Meyer et al. (2018)). (As mentioned before, optic disk localization is a much easier task, and can be performed in a much more robust manner (Muramatsu et al. (2009)).) The system extracts the region-of-interest (the optic disk), crops this, and resizes images of all datasets (Sec. 3.1) to a uniform 512 × 512 pixels. The data augmentation (to avoid over-fitting issues) happens through the following geometric operations: zooming in steps of 0.2, vertical flipping, shear within steps of 0.1, and rotation within 90°. We use a heuristic of keeping the augmented training set as 8 times the size of the training set. (Fig. 6 clearly indicates that the proposed Y-Net architecture has no over-fitting issues.

For training the proposed Y-Net model (the asymmetric ‘multi-encoder U-Net’/Y-Net with an Inception block and a context block in the bottleneck layer), we show some representative experiments here with the Drishti-GS dataset (Sivaswamy et al. (2014)). The Drishti-GS dataset often serves a standard glaucoma dataset with marked optic cup positions. This has 50 training and 51 test images. The segmentation converges within 100 epochs with a training batch size set to 4 input

RGB images (with their corresponding binary masks). We use an Adam optimizer with accuracy as the training metric. For the purpose of image segmentation loss function we use a pixel-wise cross-entropy loss, which compares the class predictions for each pixel individually. In addition to this, we have experimented with other segmentation losses such as the Dice loss and the IOU loss. The model uses a initial learning rate of 1e-4. This is dynamically updated after 10 epochs. The system implementation is on an Intel i7 Windows 10 System with an NVIDIA Quadro P5000 GPU card with 2560 CUDA cores. An important part of the learning process is a check for over-fitting. Fig. 6 shows that the proposed Y-Net model does not suffer from over-fitting issues. In addition to the above training regimen, we perform a post-processing step while testing the image. We use an empirically set threshold of 0.7 for the predicted optic cup map. All pixel values greater than this are set to 1 (the optic cup region), and rest (the background) are set to 0. As mentioned in Sec. 2.1.4, the proposed Y-net architecture has a sigmoid activation function after the last fully connected layer of the network. Optic cup detection is a binary classification problem, hence a sigmoid is quite apt (instead of a softmax, which is better-suited for multi-class classification). The sigmoid gives a probability value between 0 to 1, to predict how much the pixel belongs to the optic cup region. This explains the use of the threshold, above.

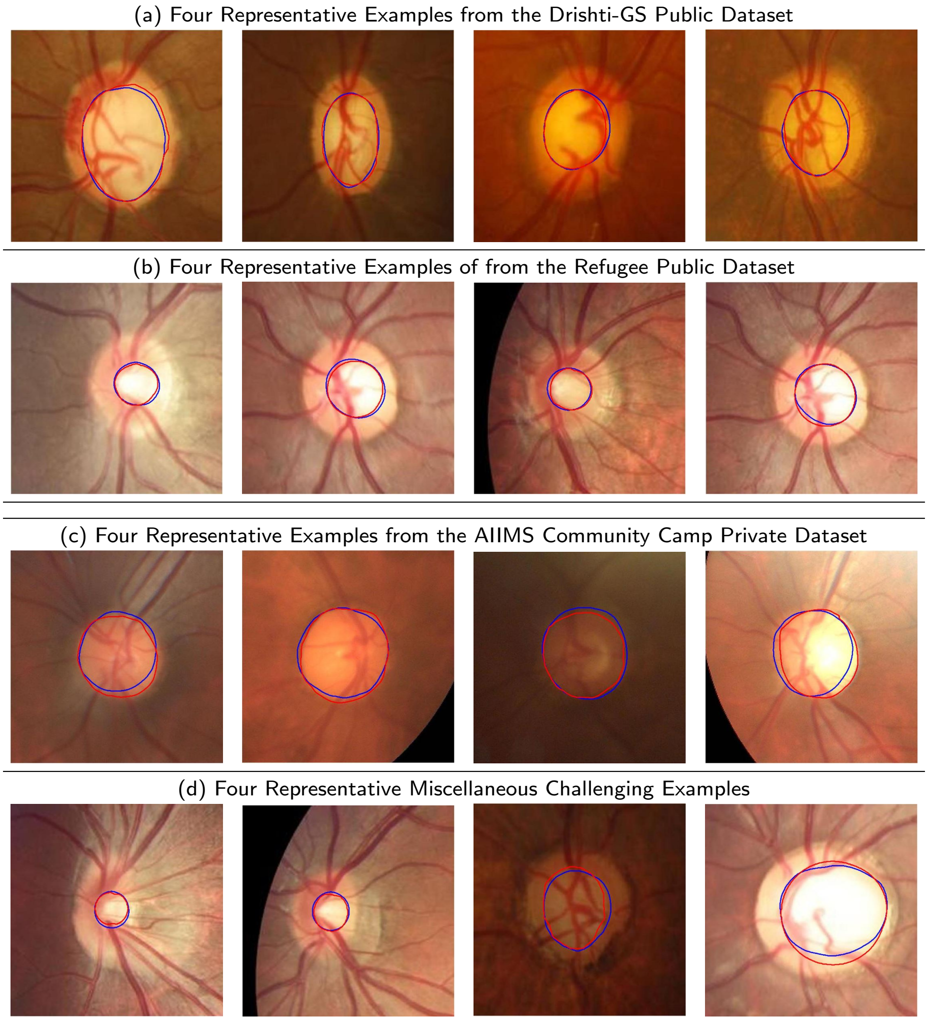
3.4 **Some Representative Results with Different Datasets**

Fig. 5 shows some representative optic cup segmentation results of images from the Drishti-GS (Sivaswamy et al. (2014)) Refugee (Maninis and Pont-Tuset (2010)) public datasets, and the extremely challenging AIIMS community camp (private) dataset. As mentioned before, the AIIMS community camp dataset is extremely challenging since this contains images taken in poor lightning conditions using low-cost hand-held ophthalmoscopes. The red circles represent the ground truth, and the blue ones represent the predicted optic cup boundaries.

3.5 **Quantitative Comparison with the State-of-the-Art across Datasets**

This section compares the performance of the proposed method (asymmetric 'multi-encoder-based U-Net'/Y-Net with an Inception block and a context block) with the state-of-the-art methods for two popular public datasets. Table 2 shows a comparison of all performance parameters (Dice coefficient, cup overlap percentage and the sensitivity) for the Drishti-GS dataset (Sivaswamy et al. (2014)) with six state-of-art optic cup segmentation methods. The proposed method scores 1.0% over the closest (Wang et al. (2019)) method, has a very high cup overlap percentage, at an acceptable high sensitivity. Our sensitivity values are less than that of (Fu et al. (2018)), indicating a higher relative number of false negatives. However, the much larger Dice coefficient indicates a better overall relation between the sensitivity and specificity.

Table 3 shows a similar comparison of the proposed method with five state-of-the-art methods, for the Refugee dataset (Maninis and Pont-Tuset (2010)). In this case as well, our approach clearly outperforms the state-of-art methods in terms of the Dice coefficient and overlap, with values 87.77% and 78.60%, respectively.



(a) Four Representative Examples from the Drishti-GS Public Dataset

(b) Four Representative Examples of from the Refugee Public Dataset

(c) Four Representative Examples from the AIIMS Community Camp Private Dataset

(d) Four Representative Miscellaneous Challenging Examples

Fig. 5: Some representative optic cup segmentation results with different datasets: (a) Drishti-GS (Sivaswamy et al. (2014)), (b) Refugee (Maninis and Pont-Tuset (2010)), (c) the challenging AIIMS community camp dataset, and (d) other miscellaneous challenging examples across different datasets. The AIIMS community camp (private) dataset has images taken in poor lighting conditions, using low-cost hand-held ophthalmoscopes. The red and blue circles represent the ground-truth and predicted cup margins respectively. Sec. 3.4 presents a visual description of representative results, while Sec. 3.5 presents quantitative results of cup segmentation across datasets, compared with the state-of-the-art. Sec. 3.6 compares quantitative results with different architectures keeping the dataset the same. Sec. 3.7 presents quantitative results of the proposed method (asymmetric 'multi-encoder U-Net'/Y-Net with an Inception block and a context block in the bottleneck layer) for different datasets.

Table 2: A comparison of the proposed approach with the state-of-the-art, keeping the dataset the same: representative results for the Drishti-GS dataset (Sivaswamy et al. (2014)). The proposed approach (asymmetric 'multi-encoder U-Net'/"Y-Net", with Inception and context blocks in the bottleneck layer) scores the highest in terms of the Dice coefficient, has an very high cup overlap percentage, at an acceptable high sensitivity. Sec. 3.5 has the details. Table 3 has the corresponding details for the Refugee (Maninis and Pont-Tuset (2010)) dataset.

S.No	Detection method	Dice coefficient (%)	Cup overlap (%)	Sensitivity (%)
1	(Al-Bander et al. (2018))	82.82	71.13	-
2	(Fu et al. (2018))	88.60	85.88	97.38
3	GL-Net (Jiang et al. (2019))	90.50	-	-
4	(Wang et al. (2019)	90.1	-	-
5	(Yu et al. (2019))	88.77	80.42	-
6	GlaucoNet (Panda et al. (2021))	89.99	82.29	-
7	**Proposed Approach (Y-Net)**	**91.11**	**83.38**	**86.29**

Table 3: A comparison of the proposed approach with the state-of-the-art, keeping the dataset the same: representative results for the Refugee dataset (Maninis and Pont-Tuset (2010)). Just as it was for the Drishti-GS dataset (Sivaswamy et al. (2014)) in Table 2, the proposed approach (asymmetric multi-encoder U-Net/"Y-Net", with Inception and context blocks in the bottleneck layer) scores the highest in terms of the Dice coefficient, has a very high cup overlap percentage, at a very high sensitivity. Sec. 3.5 has the details. Here, 'GAN' refers to a representative conditional GAN with 4 convolutional layers and an 8-level encoder-decoder setup.

S.No	Detection method	Dice coefficient (%)	Cup overlap (%)	Sensitivity (%)
1	(Liu et al. (2019b))	-	80	-
2	GAN	-	74.49	-
3	SegNet (Badrinarayanan et al. (2017))	-	79.06	-
4	(Wang et al. (2019))	87.5	-	-
5	(Fu et al. (2018))	86.48	84.02	-
6	**Proposed Approach (Y-Net)**	87.77	78.60	97.69

Most state-of-the-art methods choose these two datasets for their experimentation: this explains our choice of the two datasets for a comparative analysis, above. In the corresponding tables (Table 2 and Table 3), we have quoted statistical figures from these contemporary systems based on what the authors state in their papers. Contemporary work has not experimented with other datasets such as the DRIVE and DRIONS datasets (Maninis and Pont-Tuset (2010)).

3.6 **A Comparison of Different Architectures for the same Dataset**

In this section, we show results of a comparison of the comparative performance of different optic cup segmentation architectures, on the same dataset. Table 4 shows a comparison of the performance of optic cup segmentation with architectures, starting from a sample representative conditional GAN (with 4 convolutional layers, and an 8-level encoder-decoder setup). The next 5 rows (rows 2-6)

Table 4: Representative validation results with different architectures: results for the Drishti-GS dataset Sivaswamy et al. (2014). As shown in the table, the proposed method 'Y-Net Res34-EffB0 IC' (asymmetric 'multi-encoder U-Net'/Y-Net with an Inception block and a context block at the bottleneck layer: ResNet34 as the primary encoder, and EfficientNetB0 as the auxiliary encoder) has the largest cup coverage percentage, and the highest Dice coefficient, at an acceptable level of sensitivity. It is interesting to note the good performance of a similar architecture with a different primary encoder 'Y-Net Res50-EffB0' (which replaces the primary encoder with a deeper ResNet50, but is otherwise architecturally similar). This gives a larger sensitivity with a similar cup overlap percentage and Dice coefficient. Further details are in Sec. 3.6. The cup overlap percentage is the Intersection-over-Union Jaccard coefficient of Eq. 2. 'GAN' refers to a representative conditional GAN with 4 conditional layers and an 8-level encoder-decoder setup. Rows 2-6 in the table represent different U-Net variants, ranging from a sample representative convolutional U-Net in row 2, to variants with a particular pre-trained structure in the encoder, for the others.

S.No	Detection method	Dice coefficient (%)	Cup overlap (%)	Sensitivity (%)
1	GAN	86.34	76.87	88.86
2	U-Net (seven levels)	84.79	74.51	85.03
3	Res34-U-Net	89.94	82.11	90.62
4	Res50-U-Net	88.24	79.79	92.69
5	EfficientNetB0 U-Net	84.01	73.89	94.90
6	EfficientNetB4 U-Net	71.62	58.53	86.55
7	Y-Net (Res50-EffB0) with IC	89.39	81.32	90.87
8	**Y-Net (Res34-EffB0) with IC**	**91.11**	**83.38**	86.29

represent U-Net variants. Row 2 represents a sample convolutional U-Net. Rows 3-6 represent U-Nets which have specific pre-trained networks on their encoder side: ResNet34 (row 3), ResNet50 (row 4), EfficientNetB0 (row 5) and EfficientNetB4 (row 6). The first observation from the table is the relative improvement in performance of the basic U-Net architecture, through the addition of a suitable pre-trained deep network in the encoder part. The performance parameters in rows 3-6 represent a considerable improvement over those for a basic U-Net structure (row 2). Among these variants, ResNet34-based U-Net alone achieves an average Dice coefficient value of 89.94% which has outperformed the ResNet50-based U-Net architecture by 1.70%. This shows that increasing the depth of a network component does not necessarily lead to better performance (owing to the larger number of hyper-parameters for a relatively small dataset).

The proposed structure (the last row) achieves the best optic cup segmentation, as borne by the highest values of the Dice coefficient and the cup overlap percentage, at an acceptably high sensitivity. The proposed basic structure is an asymmetric 'multi-encoder U-Net'/Y-Net with an Inception block and a context block in the bottleneck layer. The particular architecture proposed in the paper has a ResNet34-based primary encoder, and an EfficientNetB0-based auxiliary encoder. Table 4 shows that the proposed Y-net improves the Dice coefficient by 5.32% over that of the conventional U-Net model. It is interesting to note the good performance of a similar basic architecture (row 7), but the primary encoder with a deeper structure (ResNet50). This also outperforms the U-Net variants in performance on an average, by a fair margin. Interestingly, this gives a larger sensitivity value, with similar values of the Dice coefficient and the cup overlap percentage. The table shows that the EfficientNetB0 works better than EfficientNetB4. The observation extends from using the two in a U-Net-based architecture, to the proposed asymmetric 'multi-encoder U-Net'/Y-Net. Performance improvement is not guaranteed with the deeper (or larger) variants of EfficientNet, especially for tasks with less data or fewer classes. Moreover, for larger EfficientNet variants, hyper-parameters tuning is not easy. Further, the Res34-U-Net performs better than EfficientNet on smaller datasets (especially those with lower resolution images), owing to over-fitting challenges.

We have also compared the proposed asymmetric 'multi-encoder U-Net'/"Y-Net" architecture with a sample representative baseline U-Net model, with regard to training and validation losses. Fig. 6 shows that the proposed model architecture achieves a much lower train-validation loss (cross-entropy loss) over a large number of epochs as compared to a representative U-Net model. Another take-home point from the above figure is the following (as briefly mentioned before in Sec. 3.3): neither the baseline U-Net, nor the proposed Y-Net model have over-fitting issues.

It is also interesting to see a confusion matrix comparison between a sample baseline U-Net model, and the proposed Y-Net architecture. Fig. 7 shows this comparison for a representative image of the Drishti-GS dataset (Sivaswamy et al. (2014)). For this representative case, the number of true positives is quite comparable in both cases (both correctly classify optic cup regions quite well). The proposed Y-Net classifies non-optic cup regions considerably better than the U-Net. Further, the U-Net misclassifies a considerably larger number of optic cup pixels as background pixels as compared to the proposed Y-Net. This indicates a higher sensitivity of the proposed Y-Net for instance, as borne out by statistics of a complete dataset. Table 4 in Sec. 3.6 shows the proposed Y-Net architecture

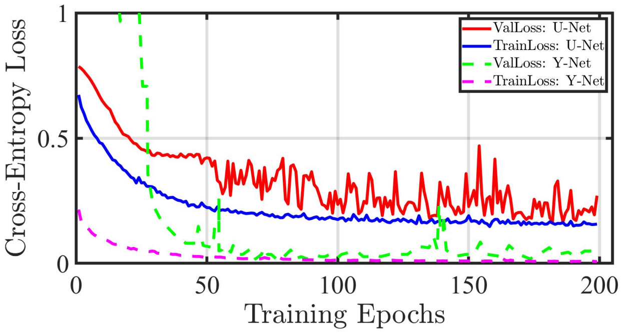
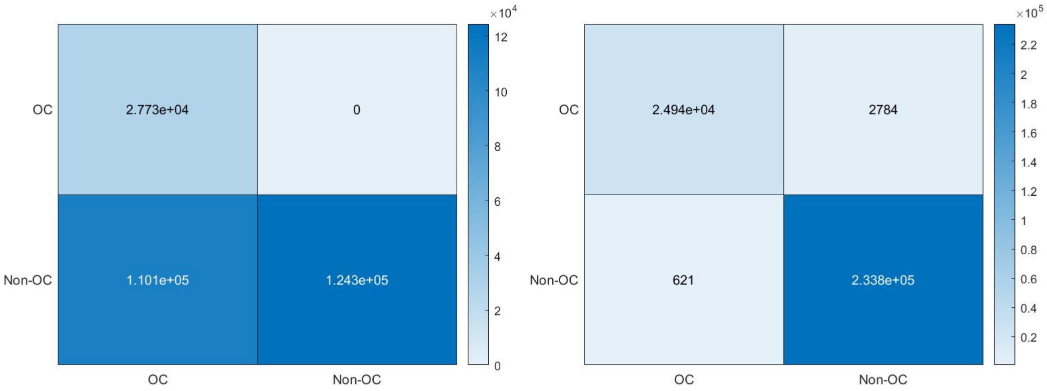


Fig. 6: The proposed Y-Net architecture achieves a much lower train-validation loss over a large number of epochs, as compared to a sample representative baseline U-Net model. Further, neither the sample baseline U-Net model, nor the proposed Y-Net model have over-fitting issues. Sec. 3.6 has the details.



(a) (b)

Fig. 7: A confusion matrix comparison between (a) a sample baseline U-Net model and (b) the proposed Y-Net architecture, for a representative image of the Drishti-GS dataset (Sivaswamy et al. (2014)). The rows represent the output of the architecture, while the columns represent the ground truth. The total number of true positives in both cases is quite comparable. The proposed Y-Net behaves considerably better for true negatives (non-optic cup regions). The proposed Y-Net also significantly reduces the number of false negatives (optic cup region identified as the background), as compared to a U-Net. This indicates a higher sensitivity of the Y-Net over a U-Net, for instance. Table 4 has the overall statistics (Dice coefficient, cup overlap and sensitivity) for a complete representative dataset: Drishti-GS (Sivaswamy et al. (2014)). The proposed Y-Net architecture outscores competing architectures on all these counts. Sec. 3.6 has the relevant discussion.

outperforming competing architectures across a set of suitable parameters: the Dice coefficient, cup overlap, and sensitivity.

3.7 Performance of the Proposed Architecture over Different Datasets

Table 5: The proposed approach (asymmetric 'multi-encoder U-Net'/"Y-Net", with Inception and context blocks in the bottleneck layer) scores very high across performance measures for four (public) datasets i.e., Drishti-GS (Sivaswamy et al. (2014)), Refugee, DRIVE, DRIONS (Maninis and Pont-Tuset (2010)) and the challenging AIIMS community camp (private) dataset. We have taken a 50:50 train-test split for the Drishti-GS and Refugee datasets, and an 80:20 split for the others. Sec. 3.7 has the details. Sec. 3.1 has details about the retinal image datasets, including the choice of the train-test split ratios.

S.No	Dataset	No. of Train-Validate-Test Images	Dice coefficient (%)	Cup overlap (%)	Sensitivity (%)
1	Drishti-GS	40-10-51	91.11	83.38	86.29
2	Refugee	400-400-400	86.34	76.87	88.87
3	DRIVE	30-5-5	83.90	72.64	81.71
4	DRIONS	88-11-11	84.91	74.21	85.86
5	AIIMS (private)	229-49-49	86.66	77.14	89.08

Table 5 shows the results of the proposed method on four (public) datasets Drishti-GS (Sivaswamy et al. (2014)), Refugee, DRIVE, DRIONS (Maninis and Pont-Tuset (2010)). Sec. 3.1 has details about the retinal image datasets, including the choice of the train-test split ratios. As can be seen from the table, the proposed architecture performs admirably across datasets. Even on the challenging AIIMS community camp (private) dataset, the proposed method reports an average Dice coefficient of 86.66, average cup overlap of 77.14% and a sensitivity of 89.08%. The AIIMS community camp dataset is full of blurred and noisy retinal images taken in poor illumination conditions, with low-cost hand-held ophthalmoscopes.

4 Declarations

Conflict of interest The authors declare no competing interests.

5 Conclusion

We propose a novel architecture for optic cup segmentation. Our novel asymmetric 'multi-encoder U-Net'/Y-Net has pre-trained ResNet34 and EfficientNetB0 primary and auxiliary encoders. The Inception and context blocks in the bottleneck layer maintain multi-scale details, and spatial consistency across the optic cup feature map. The asymmetry in the auxiliary encoder (no skip connections to the decoder side) avoids the complexity of feature map concatenation at different levels. We have validated the performance of the proposed network on four public datasets, and the challenging AIIMS community camp (private) dataset. We achieve an average Dice coefficient of 91.11% and 87.77% on the Drishti-GS (Sivaswamy et al. (2014)) and Refugee (Maninis and Pont-Tuset (2010)) public datasets, which outperforms state-of-art methods.

References

Al-Bander B, Williams B, Al-Nuaimy W, Al-Taee M, Pratt H, Zheng Y (2018) Dense Fully Convolutional Segmentation of the Optic Disc and Cup in Colour Fundus for Glaucoma Diagnosis. Symmetry 10:1 – 16

Algazi VR, Keltner JL, Johnson CA (1985) Computer Analysis of the Optic Cup in Glaucoma. Investigative Ophthalmology and Visual Science

Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(12):2481 – 2495

Chakravarty A, Sivaswamy J (2017) Joint Optic Disc and Cup Boundary Extraction from Monocular Fundus Images. Computer Methods and Programs in Biomedicine 147:51 – 61

Chen LC, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation. CoRR

Dada T, Coote M (2010) Clinical Evaluation of Optic Nerve Head. International Society of Glaucoma Surgery

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: A Large-Scale Hierarchical Image Database. In: Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp 248 – 255

Fu H, Cheng J, Xu Y, Zhang C, Wong DWK, Liu J, Cao X (2018) Disc-Aware Ensemble Network for Glaucoma Screening from Fundus Image. IEEE Transactions on Medical Imaging 37:2493 – 2501

He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In: Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp 770 – 778

Hu M, Zhu C, Li X, Xu Y (2017) Optic Cup Segmentation from Fundus Images for Glaucoma Diagnosis. Bioengineered 8(1):21 – 28

Jiang Y, Tan N, Peng T (2019) Optic Disc and Cup Segmentation based on Deep Convolutional Generative Adversarial Networks. IEEE Access 7:64483 – 64493

Kamble R, Samanta P, Singhal N (2020) Optic Disc, Cup and Fovea Detection from Retinal Images using U-Net++ with EfficientNet Encoder. In: Ophthalmic Medical Image Analysis, Springer International Publishing, pp 93 – 103

Khalid NEA, Noor NM, Ariff NM (2014) Fuzzy c-Means (FCM) for Optic Cup and Disc Segmentation with Morphological Operation. Procedia Computer Science 42:255 – 262

Liu Q, Hong X, Li S, Chen Z, Zhao G, Zou B (2019a) A Spatial-Aware Joint Optic Disc and Cup Segmentation Method. Neurocomputing 359:285 – 297

Liu S, Hong J, Lu X, Jia X, Lin Z, Zhou Y, Liu Y, Zhang H (2019b) Joint Optic Disc and Cup Segmentation using Semi-Supervised Conditional GANs. Computers in Biology and Medicine 115:1 – 12

Maninis KK, Pont-Tuset J (2010) Retinal Databases. http://www.vision.ee.ethz.ch/ cvlsegmentation/driu/downloads.html

Mehta S, Mercan E, Bartlett J, Weaver D, Elmore JG, Shapiro L (2018) Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images. In: Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp 893 – 901

Meyer MI, Galdran A, Medonça AM, Campilho A (2018) A Pixel-Wise Distance Regression Approach for Joint Retinal Optical Disc and Fovea Detection. In: Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp 39 – 47

Mittapalli PS, Kande GB (2016) Segmentation of Optic Disk and Optic Cup from Digital Fundus Images for the Assessment of Glaucoma. Biomedical Signal Processing and Control 24:34 – 46

Mohamed NA, Zulkifley MA, Zaki WMDW, Hussain A (2019) An Automated Glaucoma Screening System using Cup-to-Disc ratio via Simple Linear Iterative Clustering Superpixel Approach. Biomedical Signal Processing and Control 53:101 – 454

Mohammed A, Yildirim S, Farup I, Pedersen M, Hovde O (2018) Y-Net: A Deep Convolutional Neural Network for Polyp Detection. In: Proc. British Machine Vision Conference (BMVC), pp 1 – 11

Muramatsu C, Nakagawa T, Sawada A, Hatanaka Y, Hara T, Yamamoto T, Fujita H (2009) Determination of Cup-to-Disc Ratio of Optical Nerve Head for Diagnosis of Glaucoma on Stereo Retinal Fundus Image Pairs. In: Medical Imaging

Panda R, Puhan NB, Mandal B, Panda G (2021) GlaucoNet: Patch-Based Residual Deep Learning Network for Optic Disc and Cup Segmentation Towards Glaucoma Assessment. SN Computer Science 2:1 – 17

Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp 234 – 241

Sanfilippo PG, Cardini A, Sigal IA, Ruddle JB, Chua BE, Hewitt AW, Mackey DA (2010) A Geometric Morphometric Assessment of the Optic Cup in Glaucoma. Experimental Eye Research 91(3):405 – 414

Sivaswamy J, Krishnadas SR, Joshi GD, Jain M, Ujjwal, Syed Tabish A (2014) Drishti-GS: Retinal Image Dataset for Optic Nerve Head (ONH) Segmentation. In: Proc. IEEE International Symposium on Biomedical Imaging (ISBI)

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going Deeper with Convolutions. In: Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp 1 – 9

Tan M, Le QV (2019) EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Proc. International Conference on Machine Learning (ICML), pp 1 – 11

Tan NM, Xu Y, Goh WB, Liu J (2015) Robust Multi-Scale Superpixel Classification for Optic Cup Localization. Computerized Medical Imaging and Graphics 40:182 – 193

Thakur N, Juneja M (2019) Optic Disc and Optic Cup Segmentation from Retinal Images using Hybrid Approach. Expert Systems with Applications 127:308 – 322

Wang S, Yu L, Yang X, Fu CW, Heng PA (2019) Patch-Based Output Space Adversarial Learning for Joint Optic Disc and Cup Segmentation. IEEE Transactions on Medical Imaging 38(11):2485 — 2495

Wong DWK, Liu J, Lim JH, Li H, Wong TY (2009) Automated Detection of Kinks from Blood Vessels for Optic Cup Segmentation in Retinal Images. In: Proc. SPIE Medical Imaging, vol 7260, pp 1 – 9

Xu Y, Duan L, Lin S, Chen X, Wong DWK, Wong TY, Liu J (2014) Optic Cup Segmentation for Glaucoma Detection Using Low-Rank Superpixel Representation. In: Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp 788 – 795

Yu S, Xiao D, Frost S, Kanagasingam Y (2019) Robust Optic Disc and Cup Segmentation with Deep Learning for Glaucoma Detection. Computerized Medical Imaging and Graphics 74:61 – 71