IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 13, 2021, accepted November 8, 2021, date of publication November 10, 2021,
date of current version December 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3127280

FundusPosNet: A Deep Learning Driven Heatmap Regression Model for the Joint Localization of Optic Disc and Fovea Centers in Color Fundus Images

BHARGAV J. BHATKALKAR[1], (Senior Member, IEEE), S. VIGHNESH NAYAK[2], SATHVIK V. SHENOY[3], AND R. VIJAYA ARJUNAN[1]

[1]Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education (MAHE), Manipal, Karnataka 576104, India
[2]Department of Computer and Communication Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education (MAHE), Manipal 576104, India
[3]Department of Electronics and Communication Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education (MAHE), Manipal 576104, India

Corresponding author: R. Vijaya Arjunan (arjun.connects@gmail.com)

ABSTRACT The localization of the optic disc and fovea is crucial in the automated diagnosis of various retinal diseases. We propose a novel deep learning driven heatmap regression model based on the encoder-decoder architecture for the joint detection of optic disc and fovea centers in color fundus images. To train the regression model, we transform the ground-truth center coordinates of optic disc and fovea of the IDRiD dataset to heatmaps using a 2D-Gaussian equation. The model is capable of pinpointing any single pixel in a vast 2D image space. The model is tested on IDRiD test dataset, Messidor, and G1020 datasets. The model outperforms the state-of-the-art methods on these datasets. The model is very robust and generic, which can be trained and used for the simultaneous localization of multiple landmarks in different medical image datasets. The full implementation code and the trained model with weights (based on Keras) are available for reuse at **https://github.com/bhargav-jb/FundusPosNet**.

INDEX TERMS Fundus image, optic disc, fovea, deep learning, heatmap, regression neural network, Gaussian blob.

I. INTRODUCTION

Automated diagnosis of retinal diseases saves a considerable amount of human labor and other resources involved in contrast to manual diagnosis. It also increases the accuracy and efficiency of the screening process. The automated diagnosis of most of the diseases in fundus images requires the precise localization of Optic Disc (OD) and fovea. These two landmarks define the reference points for detecting other crucial anatomical and pathological structures in the retina [1]. The important anatomical structures and their relative distances in a color fundus image is shown in Fig. 1.

The precise detection of OD and fovea center coordinates is challenging because of the vast fundus image space. The problem of finding the centers of OD and fovea can be related to human pose estimation tasks where a trained Convolutional Neural Network (CNN) generates N heatmaps corresponding to different key-joints in the human body [2]– [4]. Further, these key-joints are used to predict the actions performed by human beings. The heatmap regression technique is applied to locate the specific coordinates in X-ray images [5], which can also be extended to determine the landmarks' coordinates in fundus images. The initial models designed for the landmark localization using the regression technique were directly regressing the coordinates using fully-connected layers on top of a CNN feature extractor [6]. This method primarily suffered from localization error as the task of precisely mapping the coordinates in a very large image space was highly non-linear in the context. The recent advancements in human pose estimations have shown remarkable accuracy

The associate editor coordinating the review of this manuscript and approving it for publication was G. R. Sinha.

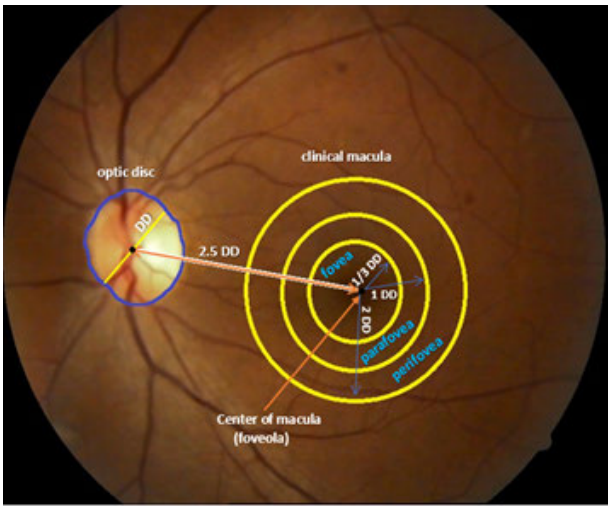


FIGURE 1. Fundus image illustrating important anatomical structures.

in landmark's detection using a heatmap based coordinate regression technique. The heatmap based regression technique achieves a very low localization error as it also considers spatial context during regression.

Image augmentation techniques are commonly used to artificially increase the size of the training and validation datasets in deep learning [7]. We propose a novel deep learning regression model for jointly regressing the fovea and optic disc centers' coordinates using encoder-decoder architecture. We name our model, the *FundusPosNet*. A unique regression technique is proposed to automatically generate the heatmap labels for the OD and fovea regions in fundus images. The proposed network is trained on these labels to generate heatmaps for OD and fovea landmarks. The proposed model achieves state-of-art results on IDRiD [8], Messidor [9], and G1020 [10] datasets.

II. STATE-OF-THE-ART

Fundus image analysis is an active research area, and recently there have been many significant works related to the joint detection of the optic disc and fovea region centers. In [11] the authors presented a milestone in the joint detection and segmentation of crucial retinal structures. Their method jointly detected the three major anatomical structures, the macula, the OD, and the vascular arch, in the retinal images. The algorithm outputs 16 distinct points in the retinal image representing the three major anatomical structures by fitting a single point-distribution-model. The method uses a cost function to find the correct positions of anatomical structures based on a combination of local and global cues fetched from the reference images. This method has a limitation that both macula and the optic disc regions should be at the center of the fundus images. In their next work, the same authors proposed a regression method to localize the fovea and optic disc using a kNN regressor [12]. The proposed technique uses two templates to extract the features to localize the optic disc and fovea. It requires the availability of vessel extraction as a prerequisite for the input images. The method used training images manually marked for optic disc and fovea center to train the regressor. The regressor first selects the point with the lowest predicted distance to the optic disc as the optic disc center and based on this, the search area for the fovea is defined. The regressor selects the location with the least predicted distance to the fovea as the fovea center location from this search area.

A faster method for the optic disc and fovea localization was proposed using template matching and directional matched filters in CIElab color space [13]. The method also uses the vessel characteristics in the optic disc to avoid false positives. A search area for locating the fovea is defined based on the optic disc location and its diameter. The point of lowest matched filter response within the search area is selected as the fovea center.

The usage of a saliency region detection algorithm to detect the optic disc in CIElab color space was proposed in [14]. A saliency map may identify multiple image regions as the possible optic disc due to pathological symptoms. For validating the optic disc detection, an unsupervised, probabilistic latent semantic analysis classification algorithm was used, which uses the specific structure of vasculature in the detected region. After detecting the optic disc, the estimation of possible fovea region was done using the prior knowledge about the distance of the fovea center from the optic disc center along the axis of symmetry of a parabola whose vertex is at the center of the optic disc. The proposed method fails if the OD region is damaged or lacks saliency concerning vessel structure, color, and illuminance in the image.

The semi-elliptical convex shapes like the OD in the fundus image can be detected using Super-Elliptical Filters [SEF] [15]. The authors have also proposed a setup for the simultaneous OD and fovea detection using two individual SEF filters located at a fixed distance from each other according to the vertical and horizontal distances between the OD and fovea mentioned in [16].

A unified approach for detecting OD and fovea based on normalized cross-correlation [NCC] technique was presented in [17]. The method performs NCC on fundus images using the OD and fovea templates obtained from cropping the specific regions in the sample fundus images by the experts. To optimize the traditional NCC technique, the authors have replaced the conventional mean and variance operations with vector inner products and norms. To further increase the detection speed, they have performed NCC on down-sampled templates and down-sampled fundus images, and finally mapping the Region-of-Interest (ROI) obtained as a result back to the original fundus image.

The signal and intensity domain information from the fundus images are used to detect OD and fovea locations. The method proposed in [18] uses 1-D projections of the image feature set in which 19 scanned lines were used to identify the landmarks' precise locations. For the detection of OD, the method uses the intensity variation information of the central optic nerve and retinal vessels emerging from

OD. This variation in the intensity is very different from any other variations in the intensities resulting from other image pathologies. The peak-valley analysis is performed on the scanned intensity lines to select OD's center coordinates, followed by choosing the reduced search space to detect the fovea. The signal-valley analysis is then performed on the reduced search space to precisely detect the fovea center.

CNN are recently gaining much popularity in medical image analysis. A 7-layer CNN to jointly segment OD, fovea, retinal vasculature, and background regions was proposed by [19]. In this method, the three channels of the input are passed to the CNN for every pixel location (x, y) in the ROI for the classification. The first channel of input is a 7×7 neighborhood of the pixel (x, y) scaled to a size of 33×33. The second channel of input is the 33×33 neighborhood of the pixels. The third channel of input is a 165×165 neighborhood with the pixel (x, y) as its center but scaled down to a size of 33×33. The CNN used for classification has five hidden layers and an output layer with four neurons for 4-class classification.

A multi-stage faster-RCNN network for the OD and fovea detection is given in [20]. In the first stage, OD detection is performed using the traditional faster-RCNN [21], followed by OD segmentation using SVM. The second stage uses an RPI-based faster-RCNN to segment the fovea, followed by its center regression.

In most fundus images, the relative spatial positions of optic disc and fovea size are constant. The authors in [22] exploited this constant relative geometry to jointly detect the centers of optic disc and fovea in their two-stage proposed method. In the first stage, a relation network draws bounding boxes which is the ROI around the OD and fovea. The relation network uses a Faster-RCNN with Resnet-101 [23]. In the second stage, a simple regressor implemented as a two-layer CNN was used to jointly regress the center of OD and fovea inside the bounding boxes.

The first regression model to perform a pixel-wise regression task to jointly detect OD and fovea was proposed by [9]. The method employs a fully convolutional deep neural network for jointly regressing the centers of OD and fovea. The network learns on the entire image to assess the global features for predicting two minimal distances as OD and fovea centers instead of learning on specific cropped ROI representing them.

A deep multi-scale sequential CNN was used to regress OD and fovea centers in [1]. The proposed method is fast and robust, which does not depend on the relative geometry information between the landmarks in the fundus image. The network has two stages of CNNs. In the first stage, the ROIs of both OD and fovea are extracted from the input image. The second stage takes these ROIs as the input and performs regression to detect both fovea and OD centers.

III. PROPOSED MODEL

In this section, we emphasize the details of the FundusPosNet design and its implementation. We explain the network architecture, dataset preparation, setup used for training the network, and the details of the new method used for extracting the landmark's center from the predicted heatmaps.

A. NETWORK ARCHITECTURE

FundusPosNet uses encoder-decoder architecture as its backbone inspired by the revolutionary U-Net model [24] designed for biomedical image segmentation. The network takes $128 \times 128 \times 3$ fundus image as its input and outputs two 128×128 heatmaps having pixel values in the range between 0-1. These two heatmaps represent the two crucial landmarks OD and fovea in the fundus image.

During network training, the encoder path learns to map the input fundus image to a vector in the latent space, and the decoder path learns to map this latent space vector to heatmaps representing the OD and fovea regions. Table 1 lists the detailed layer-wise information of FundusPosNet architecture. The following subsections explain the design details of each type of layer used in FundusPosNet.

1) CONV-BN-LeakyReLU

In this layer, we first perform convolution using a $K \times K$ kernel with a stride of (1, 1) and a dilation rate of D. The convolution is followed by the Batch Normalization (BN) operation applied to the feature maps along the channel-axis. Finally, the Leaky ReLU activation function is applied to the output after the BN.

2) CONV-BN-SIGMOID

In this layer, we first perform convolution using a $K \times K$ kernel with a stride of (1, 1) and a dilation rate of D. The convolution is followed by the BN operation applied to the feature maps along the channel-axis. Finally, the Sigmoid activation function is applied to the output after the BN.

3) TransposeConv

This is a 2×2 transpose convolution layer where we perform the up-convolutions on the input vector with a stride of 2 to up-sample its height and width by a factor of 2.

4) MaxPool

This is a $Ps \times Ps$ pooling layer with a stride S of 2×2 used to down-sample height and width of input by a factor of 2.

5) CONCATENATE

This layer performs a channel-wise concatenation of encoder and decoder output. Skip connections are used in the encoder-decoder networks to avoid the problem of vanishing gradients. Skip connections concatenate the up-sampled vector in the decoder path with the symmetrically opposite output vector in the encoder path along the channel-axis.

The output of the decoder path is subjected to a 1×1 convolution operation, followed by batch normalization. Finally, a sigmoid activation function is applied to this normalized output of the decoder to predict the two heatmaps.

TABLE 1. Layer-wise details of FundusPosNet architecture.

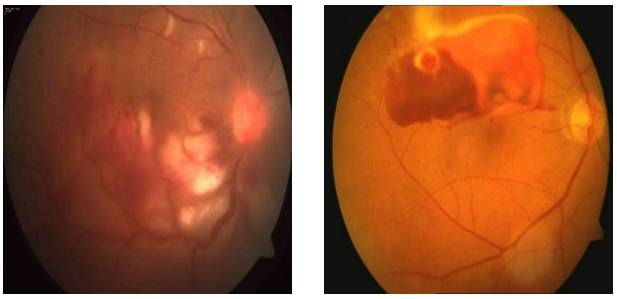
Layer name	Input shape	Layer setting	Output shape
Conv-BN-LeakyReLU_1	(128, 128, 3)	K = (7, 7), D = (1, 1)	(128, 128, 32)
Conv-BN-LeakyReLU_2	(128, 128, 32)	K = (7, 7), D = (1, 1)	(128, 128, 16)
Conv-BN-LeakyReLU_3	(128, 128, 16)	K = (7, 7), D = (1, 1)	(128, 128, 32)
MaxPool_1	(128, 128, 32)	Ps = (2, 2), S = (2, 2)	(64, 64, 32)
Conv-BN-LeakyReLU_4	(64, 64, 32)	K = (5, 5), D = (1, 1)	(64, 64, 64)
Conv-BN-LeakyReLU_5	(64, 64, 64)	K = (5, 5), D = (1, 1)	(64, 64, 32)
Conv-BN-LeakyReLU_6	(64, 64, 32)	K = (5, 5), D = (1, 1)	(64, 64, 64)
MaxPool_2	(64, 64, 64)	Ps = (2, 2), S = (2, 2)	(32, 32, 64)
Conv-BN-LeakyReLU_7	(32, 32, 64)	K = (3, 3) , D = (1, 1)	(32, 32, 128)
Conv-BN-LeakyReLU_8	(32, 32, 128)	K = (3, 3), D = (1, 1)	(32, 32, 64)
Conv-BN-LeakyReLU_9	(32, 32, 64)	K = (3, 3), D = (1, 1)	(32, 32, 128)
MaxPool_3	(32, 32, 128)	Ps = (2, 2), S = (2, 2)	(16, 16, 128)
Conv-BN-LeakyReLU_10	(16, 16, 128)	K = (3, 3), D = (2, 2)	(16, 16, 256)
Conv-BN-LeakyReLU_11	(16, 16, 256)	K = (3, 3), D = (2, 2)	(16, 16, 128)
Conv-BN-LeakyReLU_12	(16, 16, 128)	K = (3, 3), D = (2, 2)	(16, 16, 256)
MaxPool_4	(16, 16, 256)	Ps = (2, 2), S = (2, 2)	(8, 8, 256)
Conv-BN-LeakyReLU_13	(8, 8, 256)	K = (3, 3), D = (2, 2)	(8, 8, 512)
Conv-BN-LeakyReLU_14	(8, 8, 512)	K = (3, 3), D = (2, 2)	(8, 8, 256)
Conv-BN-LeakyReLU_15	(8, 8, 256)	K = (3, 3), D = (2, 2)	(8, 8, 512)
TransposeConv_1	(8, 8, 512)	K = (2, 2), S = (2, 2)	(16, 16, 256)
Concatenate_1	—	Output of Conv-BN-LeakyReLU_12 + Output of TransposeConv_1	(16, 16, 512)
Conv-BN-LeakyReLU_16	(16, 16, 512)	K = (3, 3), D = (2, 2)	(16, 16, 512)
Conv-BN-LeakyReLU_17	(16, 16, 512)	K = (3, 3), D = (2, 2)	(16, 16, 256)
Conv-BN-LeakyReLU_18	(16, 16, 256)	K = (3, 3), D = (2, 2)	(16, 16, 512)
TransposeConv_2	(16, 16, 512)	K = (2, 2), S = (2, 2)	(32, 32, 128)
Concatenate_2	—	Output of Conv-BN-LeakyReLU_9 + Output of TransposeConv_2	(32, 32, 256)
Conv-BN-LeakyReLU_19	(32, 32, 256)	K = (3, 3), D = (1, 1)	(32, 32, 256)
Conv-BN-LeakyReLU_20	(32, 32, 256)	K = (3, 3), D = (1, 1)	(32, 32, 128)
Conv-BN-LeakyReLU_21	(32, 32, 128)	K = (3, 3), D = (1, 1)	(32, 32, 256)
TransposeConv_3	(32, 32, 256)	K = (2, 2), S = (2, 2)	(64, 64, 64)
Concatenate_3	—	Output of Conv-BN-LeakyReLU_6 + Output of TransposeConv_3	(64, 64, 128)
Conv-BN-LeakyReLU_22	(64, 64, 128)	K = (5, 5), D = (1, 1)	(64, 64, 128)
Conv-BN-LeakyReLU_23	(64, 64, 128)	K = (5, 5), D = (1, 1)	(64, 64, 64)
Conv-BN-LeakyReLU_24	(64, 64, 64)	K = (5, 5), D = (1, 1)	(64, 64, 128)
TransposeConv_4	(64, 64, 128)	K = (2, 2), S = (2, 2)	(128, 128, 32)
Concatenate_4	—	Output of Conv-BN-LeakyReLU_3 + Output of TransposeConv_4	(128, 128, 64)
Conv-BN-LeakyReLU_25	(128, 128, 64)	K = (7, 7), D = (1, 1)	(128, 128, 64)
Conv-BN-LeakyReLU_26	(128, 128, 64)	K = (7, 7), D = (1, 1)	(128, 128, 32)
Conv-BN-LeakyReLU_27	(128, 128, 32)	K = (7, 7), D = (1, 1)	(128, 128, 64)
Conv-BN-Sigmoid_1	(128, 128, 64)	K = (1, 1), D = (1, 1)	(128, 128, 2)

K: Kernel size, D: Dilation rate, Ps: Pool size, S: Stride

In some fundus images, the OD and fovea centers are tough to locate due to the lack of adequate image quality or disease pathology in the images. Retinal diseases often produce dark and bright patches in the retinal layer, which may be mistaken for the macula and OD regions, as shown in Fig. 2. To avoid interference from the pathological anomalies during OD and fovea centers' regression, we consider the unique geometrical distance relationship between OD and fovea centers. The center of the fovea is located in the darkest region of a fundus image, approximately 2.5 times the OD diameter from the OD region [1]. The usage of dilated convolutions [25] in the deeper layers of the architecture and the application of different kernel sizes increase the receptive field. This larger receptive field further aids the network to learn more parameters based on the OD and fovea center distance relation.

B. PREPARING DATASET FOR NETWORK TRAINING

We have used the IDRiD grand challenge fundus dataset is used for network training. This dataset has 413 images for training and 103 images for testing. All the images in the dataset are 4288 × 2848 dimension RGB images. The experts have labeled each image in the dataset for the OD center (Ox, Oy) and fovea center (Fx, Fy) by considering the top-left pixel as the origin (0, 0). We have resized all the IDRiD dataset images to 128 × 128 dimensions for training the network. For better convergence while training, the images are also normalized.



(a) White patches brighter than OD

(b) Dark patch darker than macula

FIGURE 2. **Fundus images with retinal diseases.**

Since there are only 413 images available for training, we use image augmentation to artificially increase the dataset size to avoid network overfitting during training and better generalization. The following image transformations are used to augment the training data.

1) *Random horizontal & vertical flips:* Each image is subjected to horizontal and vertical flip with a probability of 0.5.
2) *Random scaling:* Images are scaled by randomly selecting a scaling factor between -0.4 to $+0.4$ range.
3) *Randomly varying brightness:* The brightness of each image is varied by randomly selecting a factor from -0.5 to $+0.5$ range.

C. THE HEATMAP LABEL GENERATION

Although the pixel coordinates regression is similar to image segmentation, we do not want the output of the network to be a binary mask; instead, we need the output to be a heatmap with continuous pixel values in the range between 0 - 1. If we use ground truth as binary mask with only a single bright pixel to label target location without any spatial context, in that case, the task of predicting such heatmaps becomes extremely difficult as there can be multiple locations in fundus image which have pixel values similar to that of fovea or OD pixel values. A 2D Gaussian kernel has the ability to focus on target location as well as provide spatial context. The mean of 2D Gaussian kernel is centered around the ground truth coordinate, and the remaining part of the kernel provides the spatial context. The amount of spatial context can be controlled by varying the σ value. The spatial context will allow the model to learn the important features more effectively. We generate the heatmap using the Gaussian function given by Eq. (1) in the paper [26].

$$H(x, y) = \exp\left(-\frac{(x-\alpha)^2 + (y-\beta)^2}{2\sigma^2}\right) \tag{1}$$

where, (α, β) is the actual (annotated) center of the landmark, (x, y) is a coordinate in 128×128 image vector representing the heatmap label, and σ is a constant used to control the size of the Gaussian blob in the generated heatmap label.

The following steps are performed to generate the heatmap labels for each landmark (OD and fovea) in the fundus image.

Step 1: Initialize a 128×128 image vector with all pixel values set to zero. This image vector will be the heatmap label produced for the input fundus image. The pair (x, y) in (1) represent the coordinates of each pixel in this image vector. The size of this image vector is chosen to be 128×128 to keep it the same as the dimension of the network output.

Step 2: The generated pixel value $H(x, y)$ in the heatmap label is obtained by plugging the values of the coordinates (x, y) of every pixel in (1).

Step 3: Repeat Step 2 for all the pixels in the image vector represented by the set $\{(0, 1), (0, 2), \ldots\ldots (127, 127)\}$.

In the generated heatmap labels, the pixel value around the center of the landmark will be high, and it smoothly decreases as we go away from the center. The constant σ in (1) controls the size of the Gaussian blobs in heatmap labels. The higher the value of σ, the wider will be the Gaussian blob and vice-versa. The value of σ chosen for the localization of OD and fovea is different as the size of the OD region is smaller compared to the macular region. It is essential to ensure that the Gaussian blob covers the entire landmark and its related surrounding region to prevent false localization. Suppose we choose a very small value for σ. In that case, the size of the Gaussian blob will be relatively small, and there is a minimal global and spatial context present in it, making it vulnerable to localization errors. After thorough experimentation with the network's different parameter values, we have selected the σ value of 2.0 and 2.5 for OD and fovea landmarks, respectively. Fig. 3 shows the samples of generated heatmap labels for OD and fovea landmarks.

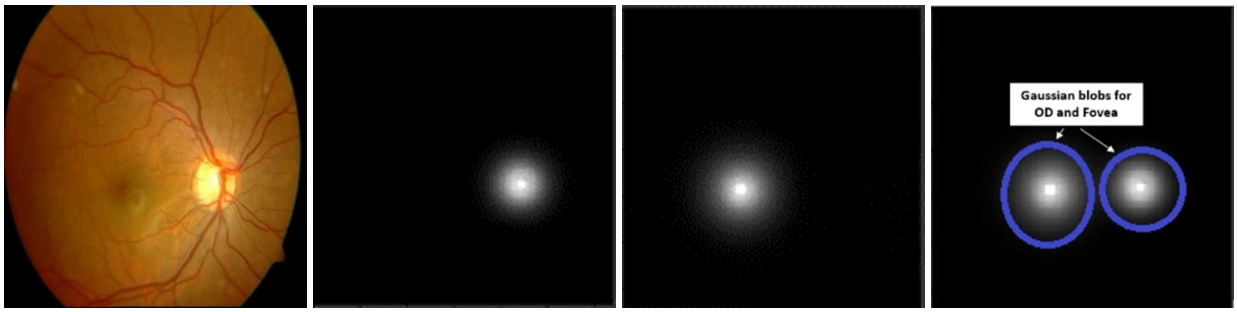
D. NETWORK TRAINING

The network is trained using 413 fundus images from the IDRiD training set and the augmentation and respective heatmap labels. The network learns to regress every pixel in the heatmap for the given input fundus image. Since the network generates heatmaps for the two landmarks and all the pixels in these heatmaps are always between the ranges 0 - 1, we use the binary cross-entropy loss function given in Eq. (2) to train the network towards an optimum state. The binary cross-entropy loss function is also best suited for the logistic regression that we use. Since we are trying to reduce the error in heatmap generation and not the Euclidean distance, binary cross-entropy is best suited for our purpose. Glorot uniform initializer is used for the weight initialization for all the kernels, and all the bias variables are initially set to zero.

$$L_{BCE} = \frac{-1}{N}\sum_{i=1}^{N} \hat{y} log\,(y) + \left(1-\hat{y}\right) log\,(1-y) \tag{2}$$

where, $\hat{y}$ is the ground-truth heatmap and y is the predicted pixel value in the generated heatmap.

For optimization, Adam optimizer is used with an initial learning rate of 0.001. A batch size of 12 is selected, and



(a) Input fundus image (b) Heatmap label for OD (c) Heatmap label for fovea (d) Concatenated heatmap label for OD and fovea along the channel-axis

FIGURE 3. Heatmap label generation.

TABLE 2. OD and fovea center localization on Messidor dataset (1200/1200).

Method	No. images	$1/8R$	$1/4R$	$1/2R$	$1R$	ED	$\overline{D}_R$
OD detection							
Al-Bander et al. [1]	1200	-	83.6	95.00	97.00	-	-
Yu et al. [13]	1200	-	-	99.08	98.24	-	-
Gegundez et al. [28]	1200	**87.33**	97.75	99.50	99.75	-	**7.03**
Meyer et al. [9]	1136	65.58	93.57	97.10	98.94	-	15.01
FundusPosNet	1136	71.19	**97.78**	**99.56**	**99.77**	**9.24**	10.68
Fovea detection							
Al-Bander et al. [1]	1200	-	66.80	91.40	96.60	-	-
Yu et al. [13]	800	23.63	64.88	94.00	98.00	-	-
Niemeijer et al. [12]	800	76.88	93.25	96.00	97.38	-	-
Gegundez-Arias et al. [28]	800	**82.00**	94.25	95.88	96.50	-	-
Dashtbozorg et al. [15]	1200	-	66.50	93.75	98.87	-	-
Meyer et al. [9]	1136	70.33	94.01	97.71	99.74	-	12.55
FundusPosNet	1136	63.26	**95.33**	**99.74**	**100**	**10.12**	**11.62**

TABLE 3. OD center localization on IDRiD test dataset (103/103).

Team name	ED
DeepDR	21.072
Relation Network Regressor [22]	26.12
CBER	29.183
VRT	33.538
ZJU-BII-SGEX	33.875
SDNU	36.220
FundusPosNet	**16.760**

TABLE 4. Fovea center localization on IDRiD test dataset (103/103).

Team name	ED
Relation Network Regressor	43.460
CBER	59.751
DeepDR	64.492
VRT	68.466
SDNU	85.400
ZJU-BII-SGEX	570.133
FundusPosNet	**40.13**

the network is trained for 800 epochs. We save the weights having the least validation loss on the IDRiD validation set. Design of architecture and the training is carried out using Keras framework on Tesla V80 GPU provided by the Google colaboratory platform.

E. EXTRACTING THE LANDMARKS' CENTER COORDINATES FROM THE GENERATED HEATMAPS

Ideally, the center pixel of the landmark should be the brightest pixel in the heatmap generated by the network. But sometimes, because of severe pathologies, there can be more than one brightest pixel detected in the generated heatmaps. If we choose the criterion for selecting the brightest pixel in the heatmap as the center of the landmark, it can lead to localization errors.

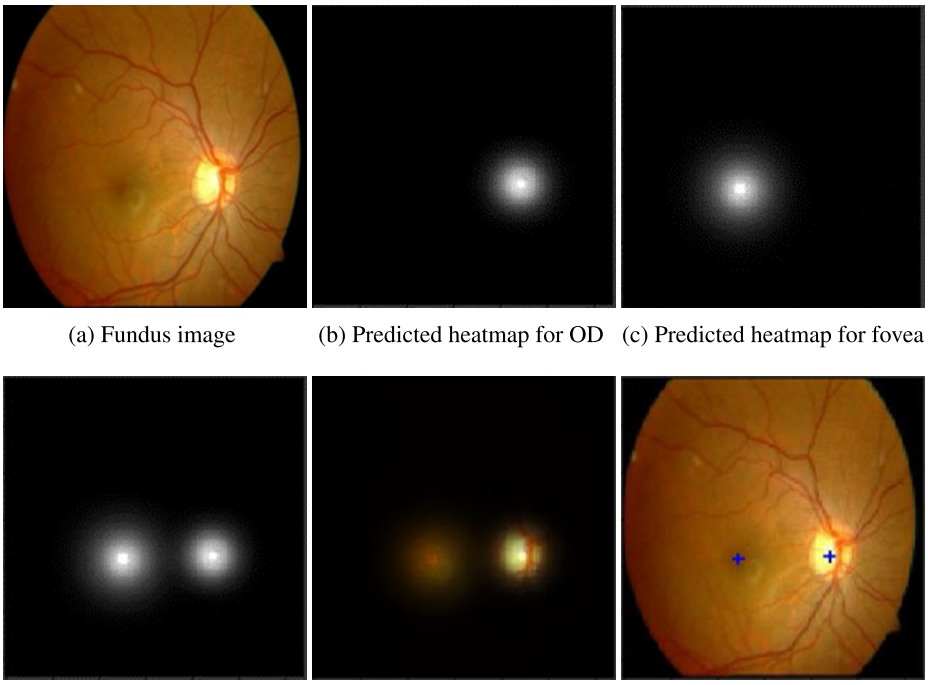
We introduce a method to accurately approximate the center coordinates of generated heatmaps. First, we determine the contour in the generated heatmap using the method given in [27]. The center of this contour determines the center of the landmark. The center $(\bar{x}, \bar{y})$ of the contour is found by first computing the moment of the contour using Eq. (3) followed by the detection of the centroid of this moment using Eq. (4). Finally, the coordinates of the landmark's detected center are mapped back from 128×128 image space to the original dimension of the input image using Eq. (5).

$$M_{ij} = \sum_{x=0}^{W} \sum_{y=0}^{H} x^i y^j I(x, y) \tag{3}$$

$$(\bar{x}, \bar{y}) = \left\{ \frac{M_{10}}{M_{00}}, \frac{M_{01}}{M_{00}} \right\} \tag{4}$$

$$X = \frac{(X_{128} \times W)}{128.0}, \quad Y = \frac{(Y_{128} \times H)}{128.0} \tag{5}$$

where, $I(x, y)$ represents pixel intensity, W and H represent the width and height of the image respectively. The complete process of OD and fovea center detection using the proposed heatmap regression technique is shown in Fig. 4.

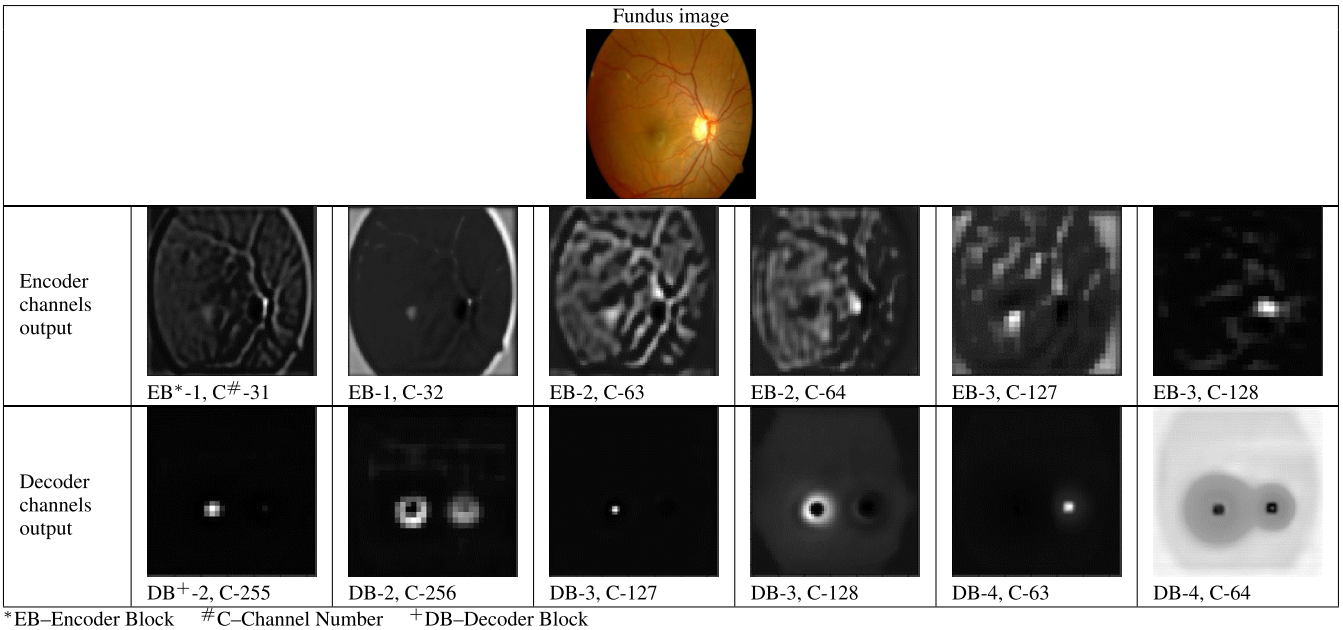


(a) Fundus image (b) Predicted heatmap for OD (c) Predicted heatmap for fovea

(d) Concatenation of OD and fovea heatmaps (e) Mapping of predicted landmarks on the fundus image (f) Landmarks' center detection

FIGURE 4. Steps in OD and fovea localization.

TABLE 5. Encoder-decoder block outputs for the last two channels.



*EB–Encoder Block #C–Channel Number +DB–Decoder Block

IV. RESULTS AND COMPARISON

To perform a fair comparison with state-of-the-art methods on the Messidor dataset, we use the *R-criterion* evaluation metric given in [28] to determine the accuracy of the proposed model. Accordingly, we use the value of R = 68, R = 103, and R = 109 for the Messidor images of resolution 1440 × 960, 2240 × 1488, and 2304 × 1536, respectively. R is the radius of the optic disc, and we compare the results in terms of (1/8)R, (1/4)R, (1/2)R, and 1R Euclidean distances between the predicted centers and the actual centers (annotated) of the OD and fovea landmarks. Following [9], we also compute the Mean Euclidean Distance ($\bar{D}_R$) between the predicted and the actual centers normalized by the OD radius as given in Eq. (6).

$$\bar{D}_R = (D(P_p,\ P_r)/R.100) \tag{6}$$

TABLE 6. Localization of OD and fovea centers in different visual quality fundus images.

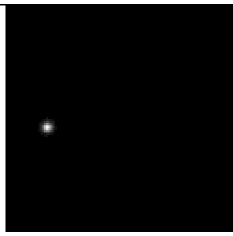
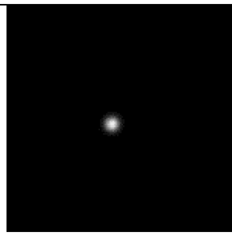
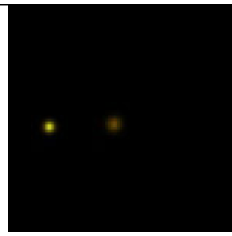
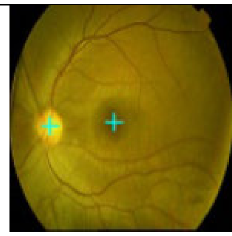
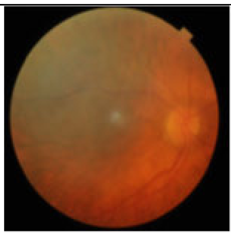
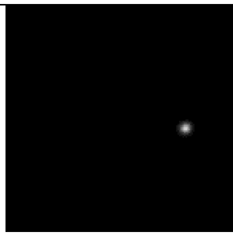
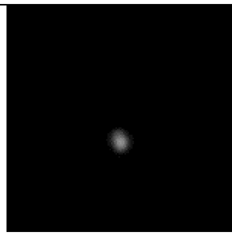
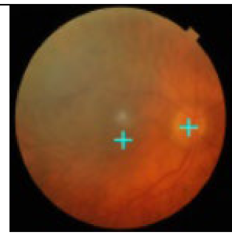
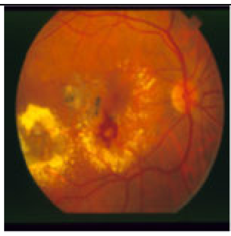
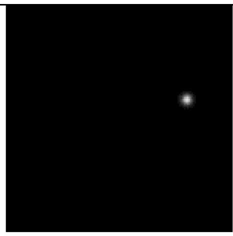
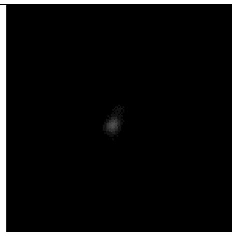
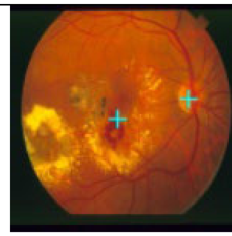
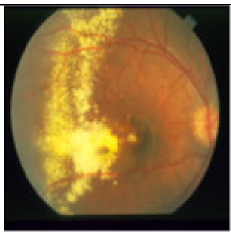
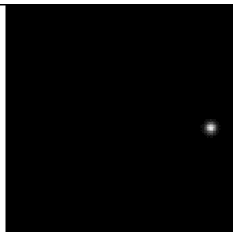
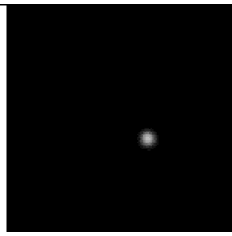
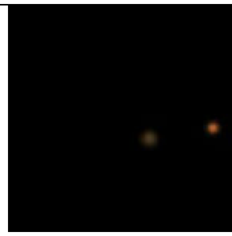
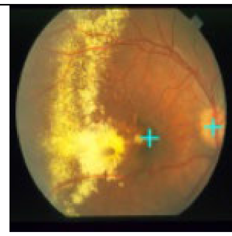
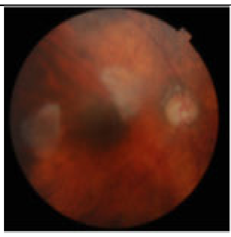
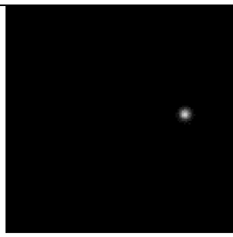
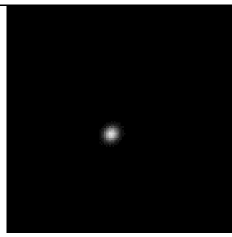
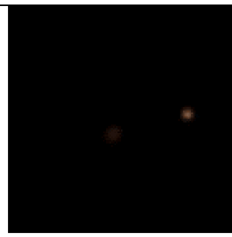
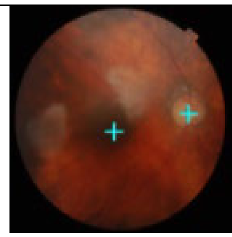
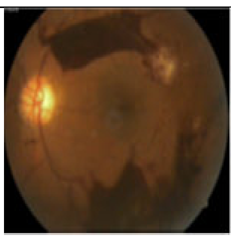
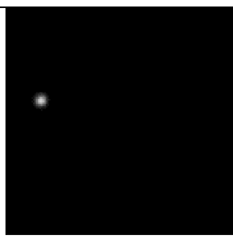
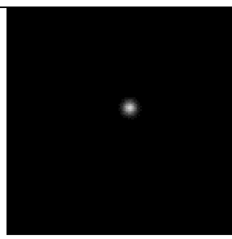
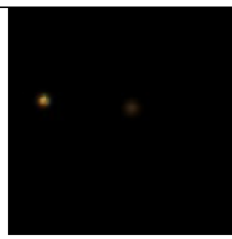
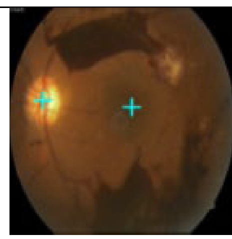
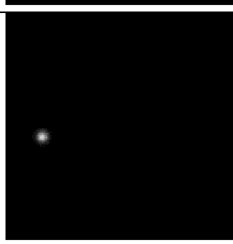
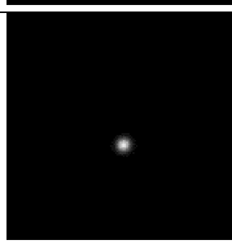
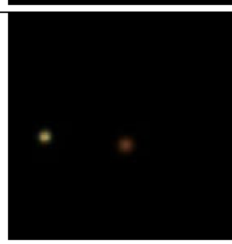
Fundus image	Heatmap for OD	Heatmap for fovea	Mapping of OD and fovea regions in fundus image	OD and fovea centers
				
				
				
				
				
				
				

TABLE 7. Architectural differences between FundusPosNet and U-Net.

Batch Normalization (BN)	Unlike U-Net, which does not use BN, FundusPosNet uses BN after the convolutional layer but before applying the activation function.
Activation function	In FundusPosNet, every convolutional layer is followed by a BN layer, which is followed by Leaky ReLU activation except for the last convolutional layer where the sigmoid activation function is used after the BN layer. In U-Net, ReLU is used as a standard activation function after every convolutional layer.
Number of parameters	FundusPosNet has 12,865,562 parameters, whereas U-Net has about 7,759,521 [29].
Copy and Crop	In U-Net, while concatenating vectors in the decoding path, the encoding path's output is cropped and copied. In FundusPosNet, we don't crop the vector because the network architecture is symmetrical, i.e., the shape of vectors along the encoding and decoding path is symmetrical.
Kernel size	U-Net uses 3 x 3 kernels for all encoder and decoder blocks, except for the last decoder block, which uses 1 x 1 convolutions. In FundusPosNet, we use different kernel sizes at different levels of encoding and decoding, as shown in Table 1.
Dilation rate	In U-Net, dilated convolution is not used, whereas in FundusPosNet, we use dilated convolutions in the last two encoder block and first decoder block, with dilation rate = 2.
Number of filters	In FundusPosNet, within each block, the number of filters follows a bottleneck scheme shown in Table 1. U-Net doesn't follow this bottleneck scheme within each block.
Loss function	The U-Net uses the soft-max loss function, whereas FundusPosNet uses a binary cross-entropy loss function.
Optimization method	U-Net is trained using SGD optimizer, whereas FundusPosNet is trained using Adam optimizer with learning rate = 0.001.

where D is the Euclidean distance and P_p and P_r are the predicted and actual pixels.

We have also considered the Euclidean Distance (ED) between the predicted center and the actual center as a metric for assessing our proposed model's performance with other state-of-the-art methods on Messidor and IDRiD test dataset. The Euclidean Distance is computed using Eq. (7) is also known as the L2 norm.

$$ED = \sqrt{\left(X_g - X_p\right)^2 + \left(Y_g - Y_p\right)^2} \tag{7}$$

where, $\left(X_g, Y_g\right)$ is the ground-truth center and $\left(X_p, Y_p\right)$ is the predicted center, respectively.

Table 2 shows the accuracy of FundusPosNet compared to other state-of-the-art methods on the Messidor dataset [9]. Table 3 and Table 4 show the accuracy of FundusPosNet compared to other models on IDRiD grand challenge test dataset [8] for the OD and fovea center detection, respectively.

Furthermore, we test FundusPosNet on G1020 dataset [10] for the optic disc localization. There are 1020 images in this dataset and each image is of 2423 $\times$ 3003 resolution. FundusPosNet detects OD centers for 979 images with a mean Euclidean distance of **54.59**. The model failed to predict OD heatmap for remaining 41 images due to the poor visuality of OD in these images.

V. DISCUSSION AND CONCLUSION

In this paper, we show that our proposed deep learning regression model performs exceptionally well on IDRiD and Messidor datasets for the OD and fovea landmarks localization and their center detection. The model is very generic, and with ground-truth data, it can be efficiently trained to localize multiple landmarks in different medical image datasets simultaneously.

Every layer in a robust deep learning model should contribute effectively while generating the desired output. Table 5 shows the encoder-decoder blocks' attention in their last two channels during OD and fovea landmarks prediction. It is evident from the output that the regression model is optimally regressing the pixels of landmark regions with proper attention given in each encoder and decoder blocks. We skip the demonstration of channel outputs for the last two encoder blocks and the first decoder block as they represent the deepest layers in the encoder-decoder path of the network, and their output is hard to interpret as well as significantly less intuitive.

In very minimal cases, we have observed that if the fundus images have severe anatomical pathologies, FundusPosNet fails to detect the OD and fovea centers precisely. It is because the network is mistakenly identifying the pathological deformations as the region of interest. Table 6 shows the result of FundusPosNet for different visual quality fundus images like good quality fundus image, inadequate quality fundus image, fundus image with not clearly visible OD and macular regions, fundus images having pathology regions brighter than the OD region, and fundus images having pathologies with darker patches than the macular region.

Although the proposed model is inspired by the revolutionary U-Net model, there are significant architectural differences between them. Table 7 lists the fundamental differences between FundusPosNet and U-Net in terms of design and implementation.

REFERENCES

[1] B. Al-Bander, W. Al-Nuaimy, B. M. Williams, and Y. Zheng, "Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc," *Biomed. Signal Process. Control*, vol. 40, pp. 91–101, Feb. 2018.

[2] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Integrating spatial configuration into heatmap regression based CNNs for landmark localization," *Med. Image Anal.*, vol. 54, pp. 207–219, May 2019.

[3] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2016, pp. 717–732.

[4] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 648–656.

[5] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Regressing heatmaps for multiple landmark localization using CNNs," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 230–238.

[6] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1799–1807.

[7] J. Wang and L. Perez, "The effectiveness of data augmentation in image classification using deep learning," *Convolutional Neural Netw. Vis. Recognit*, vol. 11, pp. 1–8, Dec. 2017.

[8] P. Porwal, S. Pachade, M. Kokare, G. Deshmukh, J. Son, W. Bae, L. Liu, J. Wang, X. Liu, L. Gao, and T. Wu, "IDRiD: Diabetic retinopathy–segmentation and grading challenge," *Med. Image Anal.*, vol. 59, Jan. 2020, Art. no. 101561.
[9] M. I. Meyer, A. Galdran, A. M. Mendonça, and A. Campilho, "A pixel-wise distance regression approach for joint retinal optical disc and fovea detection," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 39–47.
[10] M. N. Bajwa, G. A. P. Singh, W. Neumeier, M. I. Malik, A. Dengel, and S. Ahmed, "G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.
[11] M. Niemeijer, M. D. Abràmoff, and B. Van Ginneken, "Segmentation of the optic disc, macula and vascular arch in fundus photographs," *IEEE Trans. Med. Imag.*, vol. 26, no. 1, pp. 116–127, Dec. 2006.
[12] M. Niemeijer, M. D. Abràmoff, and B. van Ginneken, "Fast detection of the optic disc and fovea in color fundus photographs," *Med. Image Anal.*, vol. 13, no. 6, pp. 859–870, Dec. 2009.
[13] H. Yu, S. Barriga, C. Agurto, S. Echegaray, M. Pattichis, G. Zamora, W. Bauman, and P. Soliz, "Fast localization of optic disc and fovea in retinal images for eye disease screening," *Proc. SPIE*, vol. 7963, Mar. 2011, Art. no. 796317.
[14] M. Haloi, S. Dandapat, and R. Sinha, "An unsupervised method for detection and validation of the optic disc and the Fovea," 2016, *arXiv:1601.06608*.
[15] B. Dashtbozorg, J. Zhang, F. Huang, and B. M. ter Haar Romeny, "Automatic optic disc and fovea detection in retinal images using super-elliptical convergence index filters," in *Proc. Int. Conf. Image Anal. Recognit.* Cham, Switzerland: Springer, 2016, pp. 697–706.
[16] T. D. Williams and J. M. Wilkinson, "Position of the fovea centralis with respect to the optic nerve head," *Optometry Vis. Sci.*, vol. 69, no. 5, pp. 369–377, May 1992.
[17] J. R. H. Kumar, S. Sachi, K. Chaudhury, S. Harsha, and B. K. Singh, "A unified approach for detection of diagnostically significant regions-of-interest in retinal fundus images," in *Proc. IEEE Region 10 Conf. (TENCON)*, Nov. 2017, pp. 19–24.
[18] R. Kamble, M. Kokare, G. Deshmukh, F. A. Hussin, and F. Mériaudeau, "Localization of optic disc and fovea in retinal images using intensity based line scanning analysis," *Comput. Biol. Med.*, vol. 87, pp. 382–396, Aug. 2017.
[19] J. H. Tan, U. R. Acharya, S. V. Bhandary, K. C. Chua, and S. Sivaprasad, "Segmentation of optic disc, fovea and retinal vasculature using a single convolutional neural network," *J. Comput. Sci.*, vol. 20, pp. 70–79, May 2017.
[20] X. Li, L. Shen, and J. Duan, "Optic disc and fovea detection using multi-stage region-based convolutional neural network," in *Proc. 2nd Int. Symp. Image Comput. Digit. Med. (ISICDM)*, 2018, pp. 7–11.
[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[22] S. C. Babu, S. R. Maiya, and S. Elango, "Relation networks for optic disc and fovea localization in retinal images," 2018, *arXiv:1812.00883*.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[24] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.
[25] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
[26] G. Moon, J. Y. Chang, and K. M. Lee, "PoseFix: model-agnostic general human pose refinement network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7773–7781.
[27] S. Suzuki and K. Be, "Topological structural analysis of digitized binary images by border following," *Comput. Vis., Graph., Image Process.*, vol. 30, no. 1, pp. 32–46, 1985.
[28] M. E. Gegundez-Arias, D. Marin, J. M. Bravo, and A. Suero, "Locating the fovea center position in digital fundus images using thresholding and feature extraction techniques," *Comput. Med. Imag. Graph.*, vol. 37, nos. 5-6, pp. 386–393, 2013.
[29] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, Jan. 2020.

BHARGAV J. BHATKALKAR (Senior Member, IEEE) received the Diploma degree, bachelor's degree, the master's degree in computer science and engineering, and the Ph.D. degree in computer science discipline. He is currently a Senior Assistant Professor with the Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal. His research area includes computer vision, and he has a specific interest in computer-aided diagnosis and network/data security. He is a machine learning enthusiast and has developed efficient deep learning models for the automated detection and staging of retinal diseases. He has several publications in SCOPUS and Web of Science indexed reviewed journals. He started his career as a software developer, and later, he opted for the teaching profession. He has more than 13 years of teaching experience for UG/PG students at the university level. He is also a lifetime Associate Member of the Institution of Engineers (AMIE), India.

S. VIGHNESH NAYAK is currently pursuing the B.Tech. degree in computer and communication engineering with the Manipal Institute of Technology, Manipal, India. He has worked on semantic segmentation for many indigenous traffic datasets. His works also include neural style transfer. His areas of interests include biomedical image processing and machine learning.

SATHVIK V. SHENOY is currently pursuing the B.Tech. degree in electronics and communication engineering with the Manipal Institute of Technology, Manipal, India. He has mentored young students in the field of computer vision and machine learning. He has collaboratively worked on several other medical imaging datasets focusing on diseases, like pneumothorax and osteoporosis. His areas of interests include biomedical image processing and machine learning.

R. VIJAYA ARJUNAN received the master's degree in computer science and engineering from the Sathyabama Institute of Science, Technology, in 2005, and the Ph.D. degree in computer science and engineering from Sankara University, in 2013. He is currently working as an Associate Professor with the Department of Computer Science and Engineering, Manipal Institute of Technology, MAHE, Manipal. He had worked on deputation with the School of Engineering and IT, Manipal, Dubai Campus, from 2014 to 2017. He has published over 40 research articles in various international conferences and journals. His research interests include image processing, machine learning, deep learning, and data mining. He is a Life Member of various professional societies, like Indian Society for Technical Education, Broadcast Engineering Society, International Association of Computer Science, and Information Technology and Computer Society of India.