End-to-End Fovea Localisation in Colour Fundus Images with a Hierarchical Deep Regression Network

Ruitao Xie, Jingxin Liu, Rui Cao, Connor S. Qiu, Jiang Duan, Jon Garibaldi and Guoping Qiu

***Abstract*—Accurately locating the fovea is a prerequisite for developing computer aided diagnosis (CAD) of retinal diseases. In colour fundus images of the retina, the fovea is a fuzzy region lacking prominent visual features and this makes it difficult to directly locate the fovea. While traditional methods rely on explicitly extracting image features from the surrounding structures such as the optic disc and various vessels to infer the position of the fovea, deep learning based regression technique can implicitly model the relation between the fovea and other nearby anatomical structures to determine the location of the fovea in an end-to-end fashion. Although promising, using deep learning for fovea localisation also has many unsolved challenges. In this paper, we present a new end-to-end fovea localisation method based on a hierarchical coarse-to-fine deep regression neural network. The innovative features of the new method include a multi-scale feature fusion technique and a self-attention technique to exploit location, semantic, and contextual information in an integrated framework, a multi-field-of-view (multi-FOV) feature fusion technique for context-aware feature learning and a Gaussian-shift-cropping method for augmenting effective training data. We present extensive experimental results on two public databases and show that our new method achieved state-of-the-art performances. We also present a comprehensive ablation study and analysis to demonstrate the technical soundness and effectiveness of the overall framework and its various constituent components.**

***Index Terms*—Fovea localisation, Coarse-to-fine framework, Three-stage network, Deep learning, Data fusion, Data augmentation.**

I. Introduction

RETINAL diseases are widespread among the population and early diagnosis is important for successful treatment. In recent years, many computer aided diagnosis (CAD) techniques based on the analysis of colour fundus images have been developed [34], [14], [15], [12], [20], [7]. However, due to the complex and varied nature of fundus images, there are still many problems that remained unsolved. One of the most challenging problems is the accurate localisation of the fovea in retinal fundus images.

The fovea is a very important anatomic landmark in the retina situated at the centre of the macula in the posterior pole of the human eye. The fovea is of vital importance for our visual function [43]. If lesions appear near the fovea, the visual function of our eyes could be affected, which in severe cases could lead to blindness [38], [16]. The severity of many retinal diseases, such as maculopathy and diabetic retinopathy, is often related to the distance of the lesions from the fovea [19], [39], [11], [28], [33]. Therefore, detecting the location of the fovea in the images of the eye is a prerequisite for developing automatic diagnosis of many retinal diseases.

In early works, handcrafted image features are used to explicitly exploit the positional relationship between the fovea and the optic disc (OD), the blood vessel information, and the colour characteristics for fovea localisation [31], [37], [6], [5], [12], [27], [13]. With the development of deep learning [23], great breakthroughs have been made on image recognition using deep learning techniques [21], [18]. It is especially powerful in learning representative hierarchical features progressively from large amounts of data in an end-to-end manner. Applying deep learning methods to detect the fovea location has received increasing attention in recent years [42], [4], [50]. However, due to the lack of sufficient data and the intrinsically difficult nature of the task, the fovea localisation problem is far from solved. There are several major challenges. Many existing methods rely on auxiliary structures such as the OD for fovea localisation. However, when these auxiliary parts are damaged, the predicted results will be seriously affected. The visual appearance of the fovea is relatively fuzzy in most cases, the lack of distinctive visual features around the fovea region often leads to a large degree of uncertainty in the predicted location. As it is commonly recognized that one of the major challenges of using deep learning architecture for medical image analysis is the lack of sufficient labeled data. The fundus image analysis problem we are tackling here has the same issue, the lack of training data will lead to model over-fitting and low generalization capability.

To address the aforementioned issues, in this paper, we propose a hierarchical deep regression neural network architecture for end-to-end fovea localisation in fundus images. The end-to-end architecture of the new method works without having to explicitly extract features from other anatomic landmarks such as blood vessels and OD. Besides, unlike most previous deep learning-based works using one-stage [42], [29] or two-

R. Xie, J. Liu, R. Cao and G. Qiu are with the College of Electronics and Information Engineering, Shenzhen University, Shenzhen Insitute of Artificial Intelligence and Robotics for Society, and Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen 518060, P.R.China.

J.Liu is with Histo Pathology Diagnostic Center, Shanghai, China.

C. Qiu is with the Wexham Park Hospital, Frimley Health NHS Foundation Trust, Slough, UK, and Faculty of Medicine, Imperial College London, UK.

J. Duan (Corresponding author) is with the School of Economic Information Engineering, Southwestern University of Finance and Economics, China (e-mail: duanj_t@swufe.edu.cn) and Chengdu Everimaging Ltd.

J. Garibaldi and G. Qiu are with the School of Computer Science, University of Nottingham, Nottingham NG8 1BB, U.K.

This work is partially supported by the Education Department of Guangdong Province, PR China, under project No 2019KZDZX1028.

stage [36], [4] approaches, the proposed network is composed of hierarchical coarse-to-fine three stages. Furthermore, we fuse multi-scale features to obtain rich feature maps and introduce a self-attention mechanism [49] to acquire enhanced contextual features for improving the performance of the coarse localisation step. Additionally, we introduce two techniques to improve the fine localisation networks. We propose to extract multiple field-of-view (multi-FOV) features as the regions of interest (ROIs) in the feature maps. We further propose a novel Gaussian-shift-cropping technique to obtain a diverse set of ROIs to create a rich set of training samples for improving the training of deep neural network. A preliminary version of our model competed at the *Pathologic Myopia Challenge* (PALM) [1] held at *ISBI 2019*, and won the **1st place** in the fovea localisation task. We have also tested the performance of our network on the widely used *Messidor* dataset [10] and compared with state-of-the-art methods. The results show that our proposed method achieved new state-of-the-art performances.

In summary, we have made following contributions in this paper:

- We have developed a hierarchical coarse-to-fine deep regression neural network architecture for accurate end-to-end fovea localisation and we demonstrate that the new technique achieved state-of-the-art performances.
- We have developed an integrated framework of multi-scale feature fusion and self-attention module to exploit location, semantic, and contextual information.
- We have developed a multi-field-of-view (multi-FOV) feature fusion technique for context-aware feature learning and a Gaussian-shift-cropping method for augmenting effective training data.

II. Related work

A. Handcrafted feature-based methods

Most handcrafted feature-based methods leveraged anatomical features to determine the region of interest (ROI), and then further located the fovea in the ROI based on the darker visual features of the fovea. These methods can be classified into four categories in terms of the auxiliary information used, i.e. 1) only using blood vessel information, 2) only using OD information, 3) using both blood vessel and OD information, and 4) using none of the information of blood vessel and OD.

Some works only utilize blood vessel information to identify ROIs. For example, Deka et al. [12] firstly detected the blood vessels and then segmented out the macula region; Medhi et al. [27] applied horizontal canny edge detector to the blood vessels to find the macula. Other works exploited OD information to determine ROIs. Narasimha-Iyer et al. [31] selected a square centred at a point that is 1.75 OD diameter temporal, and 0.5 OD diameter below the OD centre as the ROI. Sekhar et al. [37] defined the ROI as the portion of a sector subtended at the the OD centre by a 30 degree angle above and below the line between the OD and image centre. Similarly, based on the location of the OD, the ROI was identified by Asim et al. [6]. Then all the minimum intensity value pixels in the ROI were found, and the median pixel of all the minimum values was considered as the fovea if there were no blood vessels around it.

There is research using both blood vessels and OD information to identify ROIs. Li et al. [24] extracted the points on the main blood vessels in the fundus image, and the extraction result was fitted to a parabola with the OD as the focus. Aquino et al. [5] segmented the OD and vascular tree to obtain a more accurate fovea location. There are also a few algorithms that do not explicitly leverage any information of blood vessels and optic disc. Sinthanayothin et al. [41] directly designed a template of intensities that can be used to approximate a typical fovea. GeethaRamani et al. [13] applied data mining through heuristic based clustering to obtain the macular candidate regions. Pachade et al. [32] derived the field of view (FOV) mask of the fundus image based on a mathematical method and identified the ROI according to the centre and the diameter of the FOV.

As can be seen above, most handcrafted feature-based methods rely on information such as blood vessels and OD for ROI identification, and dark visual features of fovea for final localisation. However, these methods would fail to work when damage occurs around the fovea or the auxiliary parts which can significantly affect the effectiveness of related features.

B. Deep learning-based methods

With its powerful abilities in learning representative features automatically, the development of deep learning has provided a new solution to the fovea localisation problem. Currently, most existing methods based on deep learning can be divided into two types, one is one-stage localisation, and the other is two-stage localisation.

Some works treat the fovea localisation problem as a one-stage issue. Tan et al. [42] proposed a 7-layer convolutional neural network and simultaneously realized the segmentation of fovea, OD and blood vessels. For every effective point in the fundus image, three different size neighborhoods of points were extracted as the input of the network. The final input size to this network was settled on 33×33 by some comparative experiments. Meyer et al. [29] used a pixel-wise MTL-like (Multi-Task Learning) strategy and reformulated the problem as regressing the distance from each image location to the closest of the OD and fovea of interest. Then, a U-Net [35] based architecture was employed for distance regression and a good result was obtained on the Messidor dataset.

Some other works address the problem by using two-stage framework. Sedai et al. [36] designed two fully convolutional neural networks with 5 convolutional blocks and skip connections. One is a coarse network used to generate the ROI and the other is a fine network. For the coarse network, side feature maps were taken from the last four blocks whereas for the fine network they were taken from the last two blocks. Then the authors concatenated these side feature maps and used them for the coarse segmentation and the fine segmentation. Similarly, AI-Bander et al. [4] proposed a two-stage system. In the first stage, the whole resized image along with the scaled centre was fed to the convolutional neural network and it would generate the coarse-localisation results of both OD

and fovea centre. Two subimages were obtained by cropping the regions around the OD centre and fovea centre. In the 2nd stage, these subimages along with the scaled centres were used to train CNNs to obtain the final results.

III. Methodology

A. Overview of the hierarchical deep regression network

The overall framework of the proposed three-stage network is shown in Fig. 1, which is mainly composed of three parts, i.e. the Main-Net, Sub-Net1 and Sub-Net2. The Main-Net is used for coarse localisation, while Sub-Net1 and Sub-Net2 are used for fine localisation. The output of the Main-Net is a vector $\boldsymbol{X_1}$, which is a predicted point of coarse localisation. We crop the feature maps in the Main-Net centered around $\boldsymbol{X_1}$ and obtain the extended ROI (E-ROI-1), which is composed of 3 ROIs. We use E-ROI-1 as the input of Sub-Net1 which produces another vector $\boldsymbol{X_2}$ as the result of the second stage. We again crop the same feature maps in the Main-Net centered around $\boldsymbol{X_2}$ and obtain another extended ROI (E-ROI-2). We take E-ROI-2 as the input of Sub-Net2 and obtain the output $\boldsymbol{X_3}$ as the final result. All three networks target the location coordinate of the fovea and perform regression. In the following, we will describe each part of the system in detail.

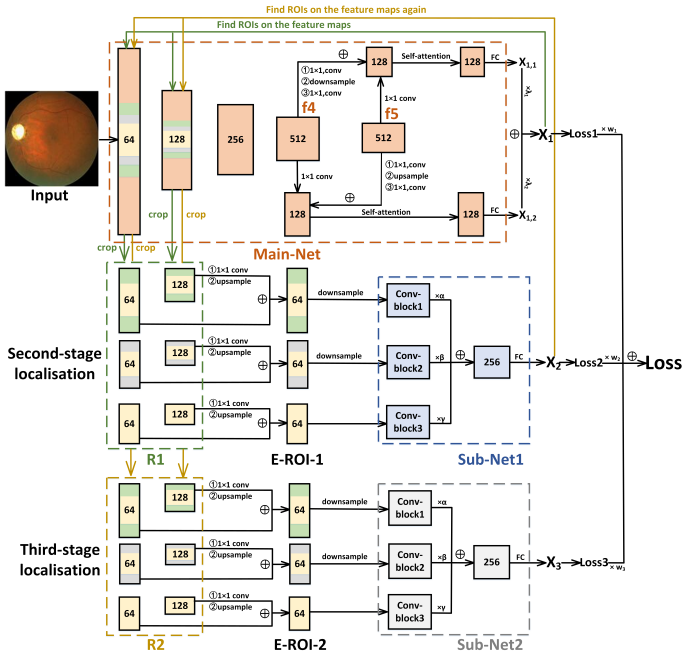

Fig. 1: The framework of the proposed three-stage network for fovea localisation. The numbers in the blocks represent the number of channels of the feature maps. Details of the self-attention module are illustrated in Fig. 2, and details of the conv-blocks in Sub-Net1 and Sub-Net2 are illustrated in Fig. 3. R1 and R2 represent the feature blocks with three different fields of view. Firstly, the fundus image is fed into the Main-Net to produce coarse prediction $\boldsymbol{X_1}$. Then, based on $\boldsymbol{X_1}$, E-ROI-1 is generated and fed into Sub-Net1 to produce $\boldsymbol{X_2}$. Finally, E-ROI-2, is obtained based on $\boldsymbol{X_2}$ and put into Sub-Net2 to generate final fine prediction $\boldsymbol{X_3}$. Please see main text for detailed operations.

B. Main-Net for coarse localisation

As shown in Fig. 1, the Main-Net is the first stage in our proposed three-stage network and it produces a vector $\boldsymbol{X_1}$ as the result of the coarse localisation. We first use a pre-trained VGG19 [40] to extract initial features. Then, we leverage multi-scale feature fusion technique and the self-attention module [49] to enhance the extracted features with richer location and contextual information. As the location of the fovea is related to many other features on the fundus image, such as the location of the OD and the distribution of blood vessels, it is very important to capture the contextual spatial features. Specifically, the initial features of the 4th and 5th blocks of VGG19 are fused and then put into the self-attention module [49] to obtain two sets of features, which are further fed into two fully connected layers and produce two fovea location vectors, $\boldsymbol{X_{1,1}}$ and $\boldsymbol{X_{1,2}}$, respectively. The weighted sum of the two locations are taken as the coarse location output of the Main-Net: $\boldsymbol{X_1} = \lambda_1 \boldsymbol{X_{1,1}} + \lambda_2 \boldsymbol{X_{1,2}}$. In practice, we only regress $\boldsymbol{X_1}$ towards the fovea location, and fix λ_1 and λ_2 to simplify the model. In this paper, these two parameters are set as $\lambda_1 = 1$ and $\lambda_2 = 3$, please see experiment section for details.

1) Multi-scale feature fusion: In order to obtain more semantic location information for coarse localisation, we fuse multi-scale features as shown in the Main-Net in Fig. 1. In general, the feature maps in shallower layers contain more location details, while features of deeper layers present more abstract semantic information [26]. The contour and location features of other structures on the fundus image is related to the location of the fovea. For example, the contour and location features of the optic disc are helpful for fovea localisation because the location of the fovea is about 2.5 OD diameters away from the OD centre [22], [46]. The feature maps in the 4th block are twice as large as that of the 5th block, therefore, the former contains more contour and location information than the latter. Therefore, we additionally extracted the features in the 4th block of the VGG19 to add more location information.

Here, we marked the features of the 4th and 5th block as f4 and f5 respectively, and they are fused to integrate more information. As shown in the Main-Net in Fig. 1, we first let f4 and f5 pass two 1×1 convolutional layers to reduce the feature dimension, which can reduce the parameters and make the features more compact. Since the spatial size of f4 is larger than f5, in order to match them, we downsample the size of f4 by average pooling to match f5, and upsample the size of f5 via bilinear interpolation to match f4. Then we let these two sets of scaled feature maps pass through two 1×1 convolutional layers respectively, which will generate the weights to mix f4 and f5. The scaled f5 is merged into f4 so that f4 obtains more semantic information, while the scaled f4 is merged into f5 and then f5 can include more location details. Finally, we obtain two sets of enhanced features of the fundus image, which are more descriptive for fovea localisation.

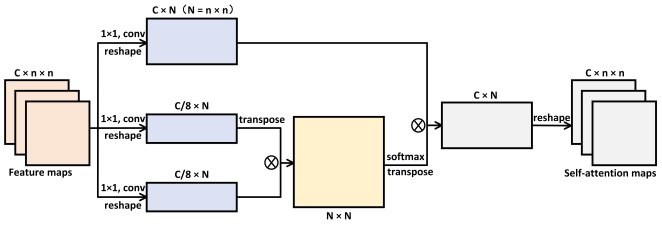

Fig. 2: Illustration of the self-attention module. C represents the number of channels of the feature maps and n represents the width and height of the feature maps.

2) Self-attention module for long-range dependencies modelling: As shown in Fig. 1, self-attention [49] mechanism is applied to our Main-Net after we obtain the enhanced feature maps as described above. The location of the fovea is closely related to other parts of the fundus images, such as the OD which is far from the fovea. Due to the limited receptive field of the convolutional layers, it is often difficult to model the long-range dependencies, making it not easy to link the information between the fovea and the OD. Self-attention is a mechanism that calculates the response at a position in a sequence by attending to all positions within the same sequence [8], which has been shown to achieve state-of-the-art results for machine translation models [44]. Besides, formalized as a non-local operation, the self-attention mechanism can capture long-range dependencies by computing interactions between any two positions [45]. Considering these facts, we exploit the self-attention module [49] to model the long-range dependencies between different parts of the fundus images.

Fig. 2 illustrates the self-attention module used in this paper. It can be seen that the input feature map is reshaped from $n \times n$ to $1 \times N$, and then multiplied by its transpose, which allows the information of each position to be related to all positions on the feature map. Besides, 1×1 convolutional layers in the module make the relationships between positions learnable. Fovea location is related to the information of other parts, such as OD and blood vessels on the fundus image, thus the self-attention module can help capture these critical cues for fovea localisation.

Specifically, we apply the self-attention module to our feature maps. For each pixel on the feature maps, a set of weights, with a sum of 1, are generated, and the weighted sum of the values at all pixels is set to the response of the corresponding position on the self-attention map. A larger value on a position of the self-attention map indicates that this position is more globally dependent and should be paid more attention to. We add the original feature maps to the weighted self-attention maps and obtain the new feature maps. The introduction of the self-attention module can help generate more long-range intervening information correlating the parts associated with the fovea, and produce more attention to the parts that are relevant to the detection of the fovea.

C. Subnets for location refinement

The Main-Net mainly considers the global characteristics of the fundus images, and use the self-attention module to model long-range dependencies. It can detect the coarse position of the fovea. However, for finer localisation, local information such as the dark colour of the fovea and the absence of blood vessels inside the fovea is necessary. Thus, we further design two sub-networks for localisation refinement as shown in Fig. 1, which are referred to as Sub-Net1 and Sub-Net2, respectively. The two sub-networks share the same structure as shown in Fig. 3. The input of the network is the extended ROI (E-ROI) with three different fields of view in the feature maps of Main-Net.

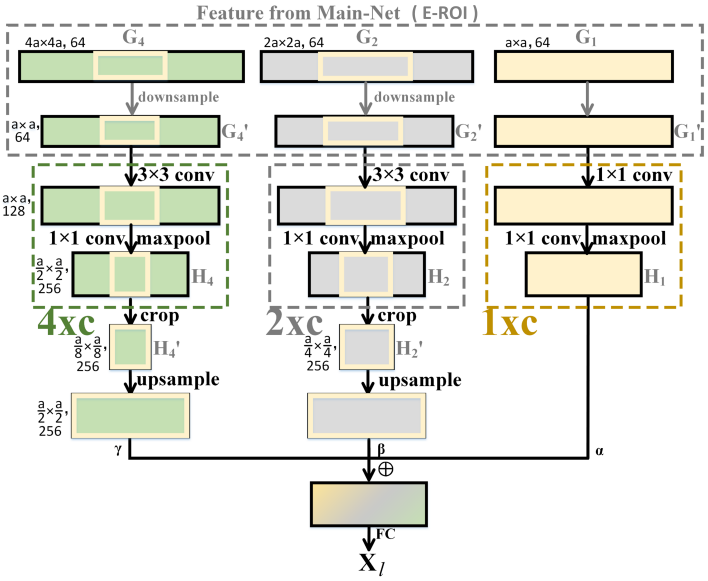

Fig. 3: The structure of the proposed subnets, i.e. Sub-Net1 and Sub-Net2, which share the same structure. a represents the $1\times$ cropping window size used to crop the feature maps of the first block of the VGG19. l represents 2 or 3.

1) E-ROI generation for fine localisation: To obtain the input to Sub-Net1, E-ROI-1 (see Fig.1), we crop the feature maps of the first and second blocks of VGG19. 3 windows of different sizes centred around $\boldsymbol{X_1}$ are cropped from the feature map of the 1st block. Another 3 windows also centred around $\boldsymbol{X_1}$ are cropped from the feature map of the 2nd block. As the feature map of the 1st block has a higher spatial resolution than that in the 2nd block, the window size used for the 1st block is larger than that of the corresponding window used for the 2nd block such that both cover the same area in the input image. The corresponding windows for the two blocks are then made to have the same number of channels and the same spatial dimension so that they can be added together to form E-ROI-1. E-ROI-1 consists of 3 windows of features each covering a different size of area of the input image centered around $\boldsymbol{X_1}$. Similarly, to obtain the input to Sub-Net2, E-ROI-2 (see Fig.1), the process is exactly the same except that this time, we replace $\boldsymbol{X_1}$ with the output of Sub-Net1, $\boldsymbol{X_2}$.

The generation of E-ROI can be expressed as Equation (1):

$$\begin{aligned} &G_k(x,y) = V_1(u,v) + U\left\{C\left[V_2(\frac{u}{2},\frac{v}{2})\right]\right\}, \\ &u = x + x_c - \frac{ka}{2},\ v = y + y_c - \frac{ka}{2},\ \ x,y \in (0,ka] \end{aligned} \tag{1}$$

where G represents the generated E-ROI (feature maps), x and y represent the abscissa and the ordinate of the position in the feature maps, V_1 and V_2 refer to the feature maps of the first and the second block from the VGG19 in the Main-Net respectively. The operation of 1×1 convolution and upsampling using bilinear interpolation for the feature maps is expressed as the functions C and U respectively. x_c and y_c represent the abscissa and the ordinate of the cropping centre position in the feature maps of the first block of the VGG19. The $1\times$ cropping window size used to crop the feature maps of the first block of the VGG19 is marked as a. k ($k = 1, 2, 4$) represents the multiplier of the $1\times$ cropping window size, $1\times$, $2\times$ and $4\times$ cropping window sizes are selected to obtain three sets of feature blocks G_1, G_2 and G_4 as our E-ROI.

2) Multi-FOV feature fusion for context-aware feature learning: Considering that there is a certain relationship between the fovea and the surrounding information, such as its darkest colour in the macula and its absence of major blood vessels [6], we also exploit the multi-FOV technique to enhance our subnets for better context-aware feature learning. Specifically, as explained above, we crop the feature maps three times using different-sized cropping windows and obtain $1\times$, $2\times$ and $4\times$ size feature blocks that are the regions of interest with different fields of view, which are marked as G_1, G_2 and G_4 respectively. We further downsample the feature blocks to the same $1\times$ size and obtain $G_1^{'}$, $G_2^{'}$ and $G_4^{'}$ respectively, which can be further explained in Equation (2):

$$G_k^{'}(x,y) = \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} G_k(kx-i, ky-j)/k^2,\ \ x,y \in (0,a]. \tag{2}$$

The feature blocks $G_2^{'}$ and $G_4^{'}$ contain more surrounding information, while the local information of interest is less obvious. As shown in Fig. 3, three parallel convolutional blocks are used to process three sets of feature blocks, and we mark them as 1xc, 2xc and 4xc respectively. The $G_1^{'}$, $G_2^{'}$ and $G_4^{'}$ feature blocks will pass through the 1xc, 2xc and 4xc convolutional blocks respectively. In the 2xc and 4xc convolutional blocks, we use 3×3 convolution kernels to link the surrounding parts to the parts we focus on.

Then, we acquire three sets of feature maps with the same size from the 1xc, 2xc and 4xc, which are marked as H_1, H_2 and H_4 respectively. Then we crop the feature maps and obtain the parts we focus on, which are corresponding to the H_1 feature maps and described in Equation (3):

$$H_k^{'}(x,y) = H_k\left[x + \frac{(k-1)a}{4k}, y + \frac{(k-1)a}{4k}\right], x,y \in (0, \frac{a}{2k}] \tag{3}$$

where $H_k^{'}$ respresents the corresponding result of sampling from H_k. Using bilinear interpolation, we upsample $H_2^{'}$ and $H_4^{'}$, which contain more surrounding information, to the same size as the H_1 feature maps. Finally, we merge these three sets of feature maps together according to a certain weight ratio. The merged feature maps are fed into a fully connected layer and a more accurate location is then obtained.

3) Gaussian-shift-cropping for effective training: As shown in Fig.1, we crop the feature maps according to the predicted results of the previous stages, but instead of cropping the feature maps centred on these predicted locations, we propose a novel Gaussian-shift-cropping mechanism to crop the feature maps, which enables us to obtain more effective training data.

When training the network and the predicted location of the coarse localisation step is relatively stable, if we use the predicted results as the centre to crop the feature maps, the acquired ROI would greatly overlap with each other, with little changes. This would lead to easy training convergence of the subnets. However, if we randomly select points around the predicted point, then the corresponding centre cropped feature maps will significantly differ from each other, which means that we will have more training data with sufficient variety.

Based on this observation, we propose a novel Gaussian-shift-cropping technique to generate ROI. Specifically, we generate a two-dimensional Gaussian probability map centred at the predicted location, and the cropped centre is obtained based on this Gaussian probability map, which can be further explained in Equation (4):

$$\begin{cases} p(x,y) = \dfrac{1}{2\pi\sigma^2 V} exp[\dfrac{(x-x_p)^2 + (y-y_p)^2}{-2\sigma^2}] \\ V = \displaystyle\iint_{(x,y)} \frac{1}{2\pi\sigma^2} exp[\frac{(x-x_p)^2 + (y-y_p)^2}{-2\sigma^2}]dxdy \\ x \in [x_p - 2, x_p + 2], y \in [y_p - 2, y_p + 2] \\ x_c = [x - (x_p - 2)]a/4 + (x_p - a/2) \\ y_c = [y - (y_p - 2)]a/4 + (y_p - a/2) \end{cases} \tag{4}$$

where the predicted location is marked as (x_p, y_p), the selected cropping centre is indicated by (x_c, y_c), and a represents the $1\times$ cropping window size used to crop the feature maps of the first block of the VGG19. A point (x, y) is randomly generated with the probability value $p(x, y)$ and the cropped centre is obtained according to this point. The cropping centre generated during the training of each iteration will be sufficiently different, which will further generate different ROI and increase the diversity of the training data.

D. Loss function

As shown in Fig. 1, the outputs of the Main-Net, Sub-Net1 and Sub-Net2 are $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ and $\boldsymbol{X_3}$ respectively, which are three predicted fovea locations of the first, second and third stages in our proposed network. One possible way to define the loss function of the network is to make each stage's output approaching the true fovea location as close as possible. Let $\boldsymbol{X}$ be the true location of the fovea, we define the overall loss function, $\mathcal{L}$, as the weighted sum of the Euclidean distance between the predicted vectors and the ground truth as shown in Equation (5):

$$\mathcal{L} = w_1\|\boldsymbol{X} - \boldsymbol{X_1}\| + w_2\|\boldsymbol{X} - \boldsymbol{X_2}\| + w_3\|\boldsymbol{X} - \boldsymbol{X_3}\|. \tag{5}$$

where w_1, w_2 and w_3 are weighting constants of different networks errors. Because the errors in the coarser network

will propagate to finer networks, the coarse errors should be punished more heavily, that is, we should in principle give more weights to coarser errors.

IV. Experiments

A. Datasets

Two datasets are used in our experiments, one is the PALM dataset [1] and the other is the Messidor dataset [10].

The PALM dataset comes from the *Pathologic Myopia Challenge (PALM)* held at *ISBI2019*. The PALM challenge published 1200 fundus images, of which 400 are labeled for training, and the remaining 800 are unlabeled. For the 400 labeled fundus images, 213 are pathological myopia fundus images and the remaining 187 are normal fundus images. 350 fundus images have a resolution of 2124×2056, while the resolution of the remaining 50 images is 1444×1444.

There are 1200 eye fundus images in the Messidor dataset. The dataset is widely used as a benchmark for fovea localisation. The resolutions of these fundus images are 2240× 1488, 1440×960, and 2304×1536 pixels. 540 cases were marked as healthy retinas, while the remaining 660 cases marked as pathological retinas. The fovea centre of 1136 fundus images in this dataset can be obtained from [14].

B. Experiment setup

1) Data split and augmentation: For PALM, we randomly divide the 400 fundus images into four equal subsets, where there are almost the same number of pathological myopia fundus images in each subset. Three of the subsets are taken as training data and the remaining part as the testing data. For Messidor, we randomly split the 1136 labeled fundus images into two equal subsets, one for training and one for testing. By exchanging training and testing data, two cross-validation experiments have been conducted. We take the average of these two experiments as the final result.

To increase the number of training data, before a fundus image was fed into the network for training, we generated a uniform random number between 0 and 1, and we would add Gaussian noise to this image if this number was greater than 0.6. The mean value of Gaussian noise was 0, and the variance value was generated from another uniform random number between 0.01 and 0.05.

2) Evaluation metrics: For PALM, the predicted location error of average Euclidean distance is used as the evaluation metric. Euclidean distance between the predicted location and the ground-truth fovea centre is also used as the criteria of performance in many related works [48], [36], [25].

For Messidor, the evaluation metric is the R rule [14], [15], [5], [9], [17], [30], [4], [29], [50], [32]. When the Euclidean distance between the predicted result and the ground truth is smaller than the radius of the OD, it is considered correct.

3) Implementation details: The GPU we used to train the networks is NVIDIA GeForce GTX 1080 Ti. The input images are resized to 224×224. Considering that feature maps with larger spatial resolution contain more location information (such as the information of the OD and the blood vessels location) that is crucial for coarse localisation, the weight λ_1 and λ_2 were set to 1 and 3 respectively, when fusing the two branches of the Main-Net as shown in Fig. 1. When mixing the features obtained from the three parallel convolutional blocks of the Sub-Nets shown in Fig. 3, considering that local features (such as the darkest colour) is more important for fine localisation, the weights α, β, and γ were set to 0.5, 0.3, and 0.2 respectively. This setting assigns more weights to the features with smaller FOV to include more local information for finer localisation. Since the first-stage coarse localisation is vital for fine localisation, we set the weights $w_1 = 2$, $w_2 = 1$, and $w_3 = 1$ in the final loss function in equation (5). For the proposed Gaussian-shift-cropping technique, the variance value (σ) of the Gaussian function as shown in Equation (4) was set to 1.

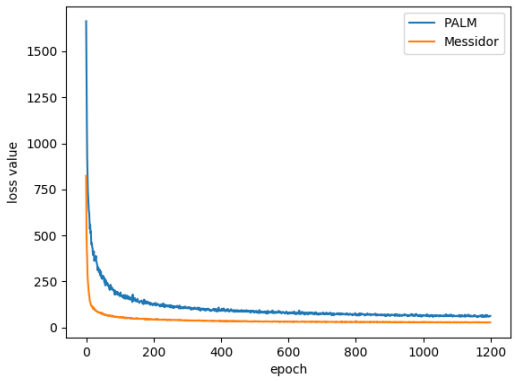

Fig. 4: The learning curves when training the proposed three-stage network on the PALM and Messidor datasets.

C. Network training

The loss funtion used to train the network is defined as Equation (5). During each iteration of training, the fundus image is fed into the Main-Net firstly and the predicted coarse location $\boldsymbol{X_1}$ is obtained, which is used to generate E-ROI-1. Then Sub-Net1 takes the E-ROI-1 as input and generates the predicted fine location $\boldsymbol{X_2}$. Finally, E-ROI-2 is generated according to $\boldsymbol{X_2}$ and fed into Sub-Net2 to produce $\boldsymbol{X_3}$ as the final predicted location. As shown in Fig. 1, using the label of fovea location, Loss1, Loss2 and Loss3 are generated according to the predicted $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ and $\boldsymbol{X_3}$ respectively, and the final loss is the weighted sum of the losses from three stages. We train the whole network (the composition of Main-Net, Sub-Net1 and Sub-Net2) end-to-end based on the final loss (Equation (5)). The learning curves when we trained the proposed three-stage network on the PALM and Messidor datasets are shown in Fig.4. It can be seen that the losses decrease with the increasing of training epochs, which suggests that the predicted results of three stages gradually get closer to the ground truth. We can also see that the learning curve on the Messidor dataset converges earlier and reaches lower loss values, which indicates that it is easier to localise the foveas in the fundus images of Messidor dataset.

D. Experiments on PALM

1) Overall results: To evaluate the performance of our proposed three-stage network, we have conducted experiments

TABLE I: Overall results of the proposed network and baselines. The best result is highlighted in bold.

Method	One-stage	Two-stage	Three-stage	Three-stage (parameter free)
Error	63.84	55.45	**50.18**	53.64

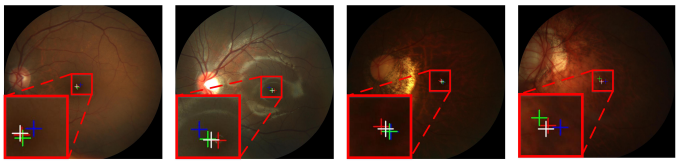

Fig. 5: Examples of the results of fovea localisation using different methods. The red cross represents the ground truth, while the blue, green, and white crosses represent the predicted results of the one/two/three-stage networks respectively.

on both the proposed network and two baselines, i.e. one/two-stage networks. For the one-stage network, only Main-Net is used for localisation. For the two-stage network, the Main-Net is used for coarse localisation while the Sub-Net1 is used for fine localisation. For the proposed three-stage network, both Sub-Net1 and Sub-Net2 are exploited for fine localisation. Additionally, we also introduce a parameter free version of the three-stage system by setting all hyperparameters (λ_1, λ_2, α, β, γ, w_1, w_2, w_3) to 1.

The overall results of the proposed three-stage network and the baselines of one/two-stage networks are presented in Table I. As can be seen, the proposed three-stage network achieves an error of 50.18 and significantly outperforms both baselines, with error reduction of 21.40% and 9.50% compared with the one-stage and two-stage networks respectively. The parameter free version of the three-stage system also performs better than the baselines even though it is slightly worse than the complete system. This result demonstrates the superiority of our method.

Some cases are presented in Fig. 5. The regions in the red box in the lower lefthand corner present the zoomed local regions of the fovea. The red cross is the ground truth, while the blue, green, and white ones indicate the predicted locations of the one/two/three-stage networks respectively. We can see that the three-stage network can always acquire more precise location than the baselines despite the significant differences of the fundus images, which demonstrates the effectiveness of our method.

2) Ablation results of the Main-Net for coarse localisation: In order to prove the validity of our proposed Main-Net for coarse localisation, we conducted ablation studies on the multi-scale data fusion and self-attention design. The results are presented in Table II, in which *Vanilla* represents original VGG19 version, while *w/ fusion* and *w/ attention* indicate the versions with our proposed multi-scale fusion and self-attention module respectively. We can see that the Main-Net achieves an error of 63.84 and significantly outperforms other baselines, with error reduction of 18.42% compared with the original VGG19 version. Examples of coarse localisation are also presented in Fig. 6. It can be seen that our proposed Main-Net performed best in all the cases despite of the obvious variance of the fundus images, which further demonstrates the effectiveness of our proposed network design.

TABLE II: The results of the ablation studies of the proposed multi-scale data fusion and self-attention design for coarse localisation. The best result is highlighted in bold.

Method	Vanilla	w/ fusion	w/ attention	w/ fusion, attention
Error	78.25	66.45	75.89	**63.84**

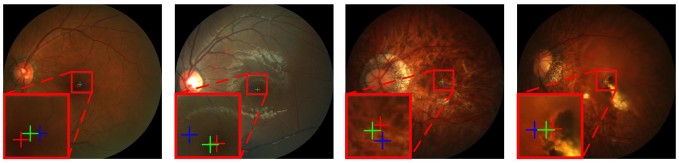

Fig. 6: Examples of the results of coarse localisation using different methods. The red cross represents the ground truth, while the blue and green crosses represent the predicted results of the VGG19 and the proposed Main-Net respectively.

TABLE III: The results of the ablation studies of the proposed multi-FOV design and Gaussian-shift-cropping technique for fine localisation. The best result is highlighted in bold.

Method	Vanilla	w/ multi-FOV	w/ GSC	w/ multi-FOV, GSC
Two-stage	63.28	61.21	60.87	**55.45**
Three-stage	58.94	54.14	55.83	**50.18**

3) Ablation results of the Sub-Nets for fine localisation: Ablation studies have also been conducted to demonstrate the effectiveness of the proposed multi-FOV design and Gaussian-shift-cropping technique for fine localisation. The experimental results are presented in Table III, in which *Vanilla* represents the network without the proposed improvements, while *w/ multi-FOV* and *w/ GSC* indicate the versions with multi-FOV network design and Gaussian-shift-cropping respectively.

As can be seen, for both networks, the ones with both multi-FOV design and Gaussian-shift-cropping achieve the lowest errors, with average Euclidean distance of 55.45 and 50.18 pixels, which brings error reduction of 12.37% and 14.86% for the two-stage and the three-stage frameworks respectively.

4) Evaluation of the impact of mixing weights in Main-Net: As shown in Fig. 1, the mixing weights of the predicted results of two different branches are set to λ_1 and λ_2 respectively. To analyse the effect of different mixing weights on the coarse localisation results, we compared the results with different λ_1 and λ_2 settings. The results are presented in Table IV. As can be seen, when $\lambda_1 = 1$ or $\lambda_2 = 1$, the average Euclidean distance error decreases first and then increases as λ_1 or λ_2 increases, which shows that mixing the information of the two branches with too much or too little is not good for our detection, and mixing the information of the two branches with an appropriate weight ratio is beneficial to our detection. The lowest average Euclidean distance error is obtained when $\lambda_1 = 1$ and $\lambda_2 = 3$, which indicates that the high-resolution feature maps are more important.

TABLE IV: The results of different mixing weights of Main-Net. The best result is highlighted in bold.

(λ_1, λ_2)	(1, 0)	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)
Error	67.32	64.72	64.34	**63.84**	64.37	65.22
(λ_1, λ_2)	(0, 1)	(1, 1)	(2, 1)	(3, 1)	(4, 1)	(5, 1)
Error	65.83	64.72	66.14	66.40	66.92	67.11

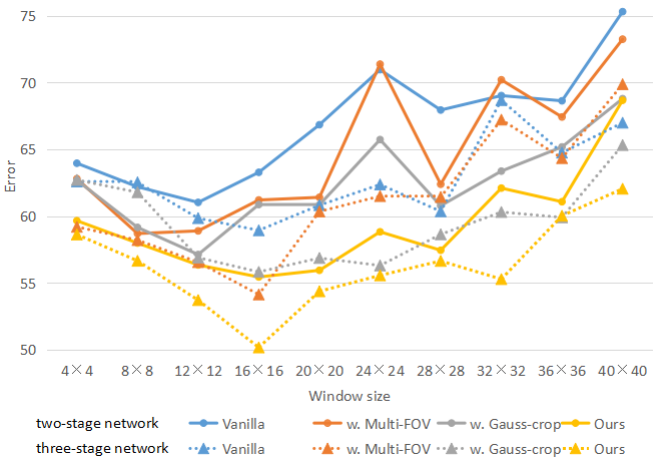

Fig. 7: Results of different methods using different cropping window sizes for the two-stage and three-stage frameworks.

5) Evaluation of the impact of cropping window size: In order to find the optimal cropping window size for ROI generation, we evaluated different methods to perform fine localisation using 10 different window sizes.

We show the results of different methods using different cropping window sizes in Fig. 7. The horizontal axis shows the $1\times$ window sizes used for Sub-Net1 and corresponding half-size windows are used for Sub-Net2 when cropping the feature maps of the first block of the VGG19 in the Main-Net, while the vertical axis shows the average Euclidean distance error obtained. We can see that for both two-stage and three-stage frameworks, the networks with both of our proposed multi-FOV and Gaussian-shift-cropping methods perform better compared with vanilla version for almost all window sizes, and our proposed method using both techniques outperforms other baselines for almost all window sizes. The optimal window size for both two-stage and three-stage frameworks is 16×16.

6) Evaluation of the impact of the number of stages: In order to analyse the effect of the number of stages for fovea localisation, we further designed a four-stage and a five-stage network which are based on the proposed three-stage network by adding one and two extra Sub-Nets (all the Sub-Nets share the same architecture) respectively. The experimental results are shown in Table V. We can see that the proposed three-stage network can achieve the best result. Besides, the more the stages, the worse the results. The results indicate that too many stages will use feature maps that are of too low resolution and contain not enough detailed information about the location of the fovea, thus leading to worse performances. Therefore, it is suggested that three stage is a good tradeoff.

TABLE V: The results of different number of stages. The best result is highlighted in bold.

Method	Three-stage	Four-stage	Five-stage
Error	**50.18**	53.69	56.91

E. Experiments on Messidor

In order to further verify the validity of our proposed method, we have also conducted experiments on the Messidor dataset. For different resolutions of 2240×1488, 1440×960, and 2304×1536 pixels, the radius of the OD was selected as 68, 103, and 109 pixels respectively [14], [5], [29]. We also calculated the proportions of the fundus images with prediction errors within $R/8, R/4, R/2$ and $2R$. And we compared with several existing methods, the results are shown in Tab. VI.

As can be seen, the accuracy of our proposed method is the highest for all the evaluation metrics. It is worth noting that our parameter free model also performed very well. For the $R/8$ rule, we achieve 13.48% and 23.42% accuracy improvement compared with the results of Meyer et al. [29] and Zheng et al. [50] respectively. For the $R/4$ rule, we attain 4.14% accuracy improvement compared with the method of Meyer et al. [29]. Furthermore, 99.74% and 99.82% accuracy have been achieved for the $R/2$ and R rules respectively. When using the $2R$ rule, an accuracy of 100.00% is obtained. We also calculated the average Euclidean distance between the predicted results and the ground truth and obtained the final result of 7.64 pixels, which is very close to the ground truth. These results further demonstrate the effectiveness of our proposed three-stage network for accurate fovea localisation.

F. Experiments on different training and testing datasets

In order to verify the robustness of the proposed model, we used different datasets for experiments. Specifically, we trained our network on the PALM dataset while tested it on the Messidor dataset. This setting is challenging, since the visual appearances of the fovea, blood vessels and OD of many fundus images are damaged in the PALM dataset, while most fundus images in Messidor are relatively complete, which results in significant differences between the training and testing datasets. We conducted comparative experiments on the proposed network and the VGG19 as baseline. We show the results in Table VII, in which *Vanilla* represents original pre-trained VGG19 version, while for the three-stage network, both Sub-Net1 and Sub-Net2 are used for fine localisation. As can be seen in Table VII, our proposed three-stage network significantly outperforms the baseline. Even though very different datasets are used for training and testing respectively, our method can achieve an accuracy of 95.26% using R rule and obtain only 22.84 pixels error using average Euclidean distance. About 22.37% accuracy improvement and 61.14% error reduction have been achieved, which demonstrates the superiority of our proposed method in terms of robustness.

We also trained our proposed three-stage network on the Messidor dataset while tested it on the PALM dataset. This task is also challenging, since almost no images of the training

TABLE VI: Comparison with the results of existing methods on the Messidor dataset. The best results are highlighted in bold.

Method	No.images	1/8R criterion (%)	1/4R criterion (%)	1/2R criterion (%)	R criterion (%)	2R criterion (%)
Gegundez-Arias et al. [14]	1136	-	76.32	93.84	98.24	99.30
Giachetti et al. [15]	1136	-	-	-	99.10	
Aquino [5]	1136	-	83.01	91.28	98.24	99.56
Dashtbozorg et al. [9]	1200	-	66.50	93.75	98.87	
Girard et al. [17]	1200	-	-	94.00	98.00	
Molina-Casado et al. [30]	1200	-	-	96.08	98.58	99.50
Al-Bander et al. [4]	1200	-	66.80	91.40	96.60	99.50
Meyer et al. [29]	1136	70.33	94.01	97.71	99.74	-
GeethaRamani et al. [13]	1200	-	85.00	94.08	99.33	-
Pachade et al. [32]	1200	-	-	-	98.66	-
Zheng et al. [50]	1136	60.39	91.36	98.32	99.03	-
Our three-stage	1136	**83.81**	**98.15**	**99.74**	**99.82**	**100.00**
Our three-stage (parameter free)	1136	81.07	97.98	99.65	99.74	99.91

TABLE VII: The results of the proposed method and VGG19 (vanilla) using PALM dataset for training and Messidor dataset for testing.

Method	Vanilla	Three-stage (ours)
Error	58.77	**22.84**
Accuracy (R rule)	72.89%	**95.26%**

dataset (Messidor dataset) present similar characteristics to the pathological fundus images of the test dataset (PALM dataset). The experiment achieved an average Euclidean distance error of 124.63 pixels, and the proportion of those images with an error within 150 pixels (close to the R rule for the PALM dataset) reaches about 80%. Considering the difficulty of the generalization task, the test results are relatively good, which further demonstrate the robustness of our proposed network.

To further verify the robustness of the proposed method on other dataset, we also used our trained model, which have been trained on the PALM dataset, to test on the Kaggle Diabetic Retinopathy dataset [2]. The dataset contains five grades of diabetic diseases, which are labelled as one of the five categories, i.e. *no*, *mild*, *moderate*, *severe*, and *proliferative* diabetic retinopathy. This dataset has complex situations around the fovea which is challenging for fovea localisation. The annotations of the fovea location are not provided, we herein show several typical visual examples of the localisation results in Fig. 8, in which two fundus images are presented for each of the five categories. It can be seen that the predicted results are roughly accurate, which further shows that our proposed method is capable of functioning well for this dataset with complex situations around the fovea.

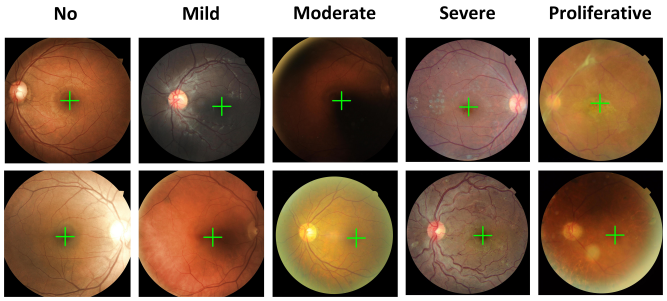

Fig. 8: Examples of fovea localisation results on the Kaggle Diabetic Retinopathy dataset. The five columns represent the fundus images from the five categories of *no*, *mild*, *moderate*, *severe*, and *proliferative* diabetic retinopathy, respectively. The green crosses represent the predicted fovea locations using our trained model.

G. Evaluation of the impact of hyperparameter setting

In order to analyze the degree of dependence of our network on hyperparameter settings, we compare two sets of hyperparameter settings, i.e. the proposed setting ($\lambda_1 = 1$, $\lambda_2 = 3$, $\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$, $w_1 = 2$, $w_2 = 1$ and $w_3 = 1$) and the equal-weight setting ($\lambda_1 = \lambda_2 = \alpha = \beta = \gamma = w_1 = w_2 = w_3 = 1$) which is equivalent to removing all the hyperparameters. We have performed experiments on both the PALM and Messidor datasets. On PALM, the network with equal-weight hyperparameter setting (parameter free version) attains average Euclidean distance error of 53.64 pixels, which is higher than that of the network with the proposed hyper-parameter setting (50.18) but still significantly lower than the baselines (see Table I). On Messidor, the network with equal-weight setting (parameter free version) only brings an error increase from 7.64 to 8.02 pixels, compared with our proposed setting; furthermore, the accuracy can still reach 99.65% using the $R/2$ rule and 99.74% using the R rule, which is better than the baselines presented in Table VI. These results show that our network is not strongly dependent on the setting of hyperparameters, which further verify the robustness of our proposed network design.

V. Discussion and Analysis

A. Evaluation on successful and failure cases

In order to further investigate into the performance of our model, we test the fundus images and analyse the results. We first set thresholds to categorize test results into successful and failed cases. For the PALM dataset, the threshold is set to be 100 pixels in terms of the Euclidean distance error, considering the fact that the OD radius of most fundus images on the PALM dataset is about 150 pixels. While for the Messidor dataset, the threshold is 0.25 OD radius ($0.25R$ rule). The success rate is about 87% on the PALM dataset and 98% on

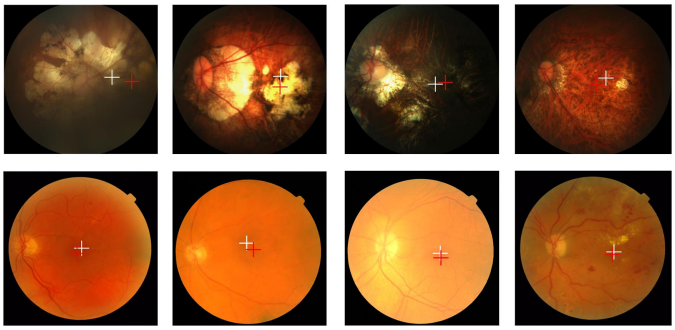

Fig. 9: Examples of failed cases where the visual features of the fovea, OD, or blood vessels are unclear. The red and white crosses represent the true and the predicted locations, respectively. The ones in the first and second rows are the results from the PALM and the Messidor datasets respectively.

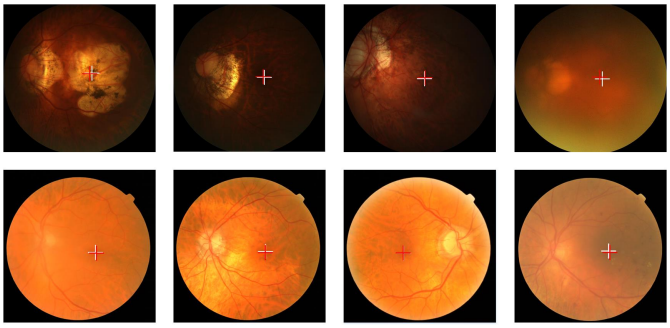

Fig. 10: Examples of successful cases where the visual features of the fovea, OD, or blood vessels are unclear. The red and white crosses represent the true and the predicted locations, respectively. The ones in the first and second rows are the results from the PALM and the Messidor datasets respectively.

the Messidor dataset respectively. If we further increase the threshold from 100 to 150 pxiels (close to the R rule) for the PALM dataset, the success rate can reach about 94%.

We further analyse the failed cases and discover that most of them are the pathological fundus images, whose visual appearance is severely affected by lesions. Some failed cases are presented in Fig. 9. It is of note that the visual features of the OD, blood vessels, and the fovea are abnormal and difficult to recognize. However, we further analyse the 213 pathological fundus images with seriously affected visual appearance on the PALM dataset and find that the proportion of good cases within 100 pixels error still reach 78%, and it will reach about 90% if we calculate the ones within 150 pixels error. While for the 1136 fundus images on the Messidor dataset, more than half of which are the images marked as pathological retinas, however, only less than 20 cases have an error exceeds 0.25R. We also show some successful cases in Fig. 10, and we can see that even though fundus images have lesions which severely affect the visual appearance of the OD, fovea, or the blood vessels, the proposed method still achieves good performance on them. These results show that our proposed method can also perform well in challenging pathological images.

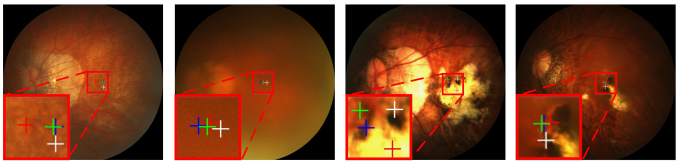

Fig. 11: Failed cases when using multiple stages for fine localisation. The red cross represents the ground truth, while the blue, green, and white crosses represent the predicted results of the one/two/three-stage networks respectively.

B. Evaluation on multi-stage fovea localisation

In order to further analyse the behaviour of our proposed three-stage network for fovea localisation, we tested on the fundus images in the PALM dataset using trained one/two/three-stage networks respectively. Although the average performance is greatly improved as shown in experiment section, there are still some cases where using more stages performed worse than using fewer stages. We show several these cases in Fig. 11. We can see that multi-stage design for fine localisation performs worse than one-stage localisation in these cases. We discover that most of them are severely damaged and the local features around the fovea are very unclear. The subnets we design mainly use local information around the fovea for fine localisation, so when these local features are severely affected, the performance of the subnets may also be affected, leading to erroneous results.

There are about 170 fundus images (most of them are normal fundus images) with clear local features around the fovea in the 400 labeled fundus images from the PALM dataset, and the remaining about 230 fundus images (most of them are pathological fundus images) are those without prominent local features around the fovea. For the 170 images with clear local features around the fovea, multi-stage networks outperformed one-stage networks for 156 (91.76%) images, with the average Euclidean distance error reducing from 34.74 to 11.48 pixels (66.95%); and for the 230 images without clear local features, multi-stage networks outperformed one-stage networks for 162 (70.43%) images, decreasing the average Euclidean distance error from 95.62 to 78.89 pixels (17.50%). This further shows that our multi-stage strategy is desirable in most cases, but also indicates that our method has scope for improvement. One possible approach is to first classify the images based on the regions around the fovea into those with clear features and those without clear features, and then use networks with fewer stages for those without and networks with more stages for those with clear features. Another possibility is to classify the image into normal or pathological and then process accordingly [47].

C. Qualitative evaluation of the self-attention module

Fig. 12 shows the self-attention map in the Main-Net that we have learned for some fundus images. The attention maps (bottom) are corresponding to fundus images (top), with red and blue indicate high and low values respectively. As can be seen, the grids near the OD and the fovea have higher values, which means that the responses corresponding to the locations

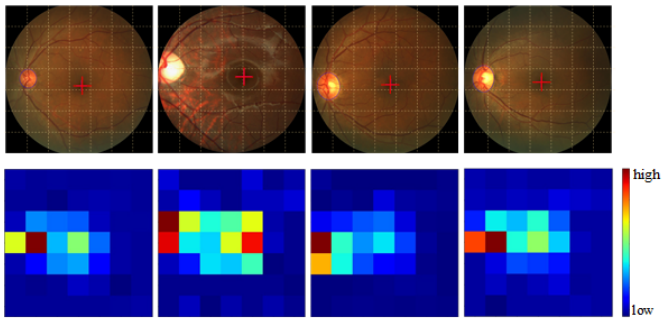

Fig. 12: Examples of self-attention maps of the feature maps obtained from the coarse localisation step for fundus images. The red cross indicates the location of the fovea.

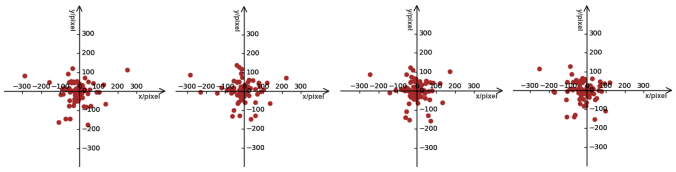

Fig. 13: The predicted results of the Main-Net for 100 images using four trained models. The coordinate origin represents the ground truth, while the brown points are the predicted results.

are larger, and as a result, these parts are more globally relevant and more attention is paid to them. This demonstrates the effectiveness of the self-attention module in associating the fovea with the OD, which will be helpful for fovea localisation since the OD is much more salient in fundus images.

D. Rationality of the Gaussian-shift-cropping

We present our observation to demonstrate the plausible rationality of the proposed Gaussian-shift-cropping technique. Specifically, we used four trained models to perform on the test data. The results of the Main-Net are shown in Fig. 13. The coordinate origin (0, 0) represents the ground truth, while the brown points are the predicted results. We can see that the distribution of the predicted results concentrates around and spreads from the fovea location. The observation suggests that the predicted locations from the Main-Net can be approximated by a 2D Gaussian-distribution centred at the fovea location. This indicates that we should pay more attention to the location around the predicted point in most cases (where the fovea may be located). Thus we can sample points within the area to simulate the possible predicted locations to generate ROIs of sufficient variation.

E. Generalization to other applications

To further evaluate the generalization ability of our proposed method in other applications, we used similar method for the scleral spur localisation task at the Angle closure Glaucoma Evaluation Challenge [3] held in the 22nd International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2019). This challenge provides a large dataset of 4800 anterior segment optical coherence tomography (AS-OCT) images, which consists of three equal subsets

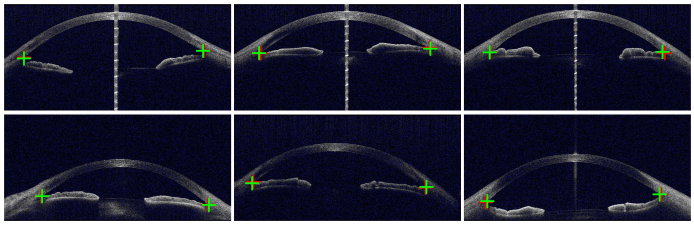

Fig. 14: Several cases of scleral spur localisation. The red crosses represent the groundtruth, and the green ones represent the predicted results of our proposed method.

(the number of the images of each subset is 1600) for training, validation and testing repectively. The resolution of the images is 2130 $\times$ 998. We obtained the average Euclidean distance error of 15.18 and 14.00 pixels in the validation dataset and the test dataset respectively, which won the fourth place in the competition participated by more than 200 teams worldwide. Fig. 14 shows some visual examples of this application, and it can be seen that the predicted results are very close to the real results although these OCT images differ substantially from the colour fungus images. These results show that our method can be generalised to other applications and different image modalities.

F. Analysis of computational complexity

To analyse computational complexity of our network, we calculate the floating point operations (FLOPs) when we use the three-stage network to predict the fovea location. Many works treated the fovea localisation as a segmentation problem [42], [36], [29], which often means that higher resolution of the input of the network is needed. We directly treat it as the regression problem between the predicted point and the ground truth and reduce the resolution of the image to 224$\times$224 before we put it to the network, which greatly reduces the amount of calculation and needs much less FLOPs than these works [42], [36], [29]. Since we use 1$\times$1 convolution operation to reduce the feature dimension when extracting the feature maps of 4th and 5th blocks of VGG19 in the Main-Net and downsample the features before the E-ROIs are put in the subnets as shown in Fig. 1, furthermore, the parallel convolution blocks of the subnets consist of only 6 convolutional layers, and the downsampling using max pooling is applied to reduce the scale of the feature maps as shown in Fig. 3, therefore, the FLOPs is only increased by about 2.3% compared to VGG19. For our proposed three-stage network, it only takes about 0.3 second to test a fundus image using our GPU, and one epoch of training using 300 fundus images of PALM dataset only cost about one minute.

VI. Conclusions

Fovea localisation is important for diagnosing and treating retinal diseases. However, the problem is very challenging due to the vague appearance of the fovea and its vulnerability to damage nearby. To address the issues, in this article, we proposed a coarse-to-fine three-stage regression network for accurate fovea localisation. The network consists of three

steps, including one coarse localisation step and two localisation refinement steps. For the coarse localisation step, we fused the multi-scale features and introduced the self-attention [49] mechanism to acquire enhanced features. For the localisation refinement steps, we extracted multi-FOV features as the E-ROIs in the feature maps. To reduce the dependence of training models on specific data, we further proposed a novel Gaussian-shift-cropping technique to obtain ROIs with more variance for fine localisation, which can make the network training more effective. The proposed approach has achieved the **1st place** in the fovea localisation task at the *Pathologic Myopia Challenge (PALM)* [1] held at *ISBI 2019*. We have also tested the performance of our network on the *Messidor* dataset [10] and compared with state-of-the-art methods. The results show that our proposed method achieve new state-of-the-art results with accuracy of 99.82% on the *Messidor* dataset.

REFERENCES

[1] https://palm.grand-challenge.org/.
[2] https://www.kaggle.com/c/diabetic-retinopathy-detection.
[3] https://age.grand-challenge.org/.
[4] 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. *Biomedical Signal Processing and Control*, 40:91–101, 2018.
[5] A. Aquino. Establishing the macular grading grid by means of fovea centre detection using anatomical-based and visual-based features. *Computers in biology and medicine*, 55:61–73, 2014.
[6] K. M. Asim, A. Basit, and A. Jalil. Detection and localization of fovea in human retinal fundus images. In *2012 International Conference on Emerging Technologies*, pages 1–5. IEEE, 2012.
[7] R. J. Chalakkal, W. H. Abdulla, and S. S. Thulaseedharan. Automatic detection and segmentation of optic disc and fovea in retinal images. *IET Image Processing*, 12(11):2100–2110, 2018.
[8] J. Cheng, L. Dong, and M. Lapata. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*, 2016.
[9] 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 *International Conference on Image Analysis and Recognition*, pages 697–706. Springer, 2016.
[10] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014.
[11] K. S. Deepak and J. Sivaswamy. Automatic assessment of macular edema from color retinal images. *IEEE Transactions on medical imaging*, 31(3):766–776, 2011.
[12] D. Deka, J. P. Medhi, and S. Nirmala. Detection of macula and fovea for disease analysis in color fundus images. In *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*, pages 231–236. IEEE, 2015.
[13] R. GeethaRamani and L. Balasubramanian. Macula segmentation and fovea localization employing image processing and heuristic based clustering for automated retinal screening. *Computer methods and programs in biomedicine*, 160:153–163, 2018.
[14] 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. *Computerized Medical Imaging and Graphics*, 37(5-6):386–393, 2013.
[15] A. Giachetti, L. Ballerini, E. Trucco, and P. J. Wilson. The use of radial symmetry to localize retinal landmarks. *Computerized Medical Imaging and Graphics*, 37(5-6):369–376, 2013.
[16] L. Giancardo, F. Meriaudeau, T. P. Karnowski, Y. Li, S. Garg, K. W. Tobin Jr, and E. Chaum. Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. *Medical image analysis*, 16(1):216–226, 2012.
[17] F. Girard, C. Kavalec, S. Grenier, H. B. Tahar, and F. Cheriet. Simultaneous macula detection and optic disc boundary segmentation in retinal fundus images. In *Medical Imaging 2016: Image Processing*, volume 9784, page 97841F. International Society for Optics and Photonics, 2016.
[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
[19] E. T. D. R. S. R. Group et al. Early photocoagulation for diabetic retinopathy: Etdrs report number 9. *Ophthalmology*, 98(5):766–785, 1991.
[20] 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. *Computers in biology and medicine*, 87:382–396, 2017.
[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
[22] H. W. Larsen. *The ocular fundus: a color atlas*. WB Saunders Company, 1976.
[23] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
[24] H. Li and O. Chutatape. Automated feature extraction in color retinal images by a model based approach. *IEEE Transactions on biomedical engineering*, 51(2):246–254, 2004.
[25] X. Li, L. Shen, and J. Duan. Optic disc and fovea detection using multi-stage region-based convolutional neural network. In *Proceedings of the 2nd International Symposium on Image Computing and Digital Medicine*, pages 7–11. ACM, 2018.
[26] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
[27] J. P. Medhi and S. Dandapat. An effective fovea detection and automatic assessment of diabetic maculopathy in color fundus images. *Computers in biology and medicine*, 74:30–44, 2016.
[28] J. P. Medhi, M. K. Nath, and S. Dandapat. Automatic grading of macular degeneration from color fundus images. In *2012 World Congress on Information and Communication Technologies*, pages 511–514. IEEE, 2012.
[29] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 39–47. Springer, 2018.
[30] J. M. Molina-Casado, E. J. Carmona, and J. García-Feijoó. Fast detection of the main anatomical structures in digital retinal images based on intra-and inter-structure relational knowledge. *Computer methods and programs in biomedicine*, 149:55–68, 2017.
[31] H. Narasimha-Iyer, A. Can, B. Roysam, V. Stewart, H. L. Tanenbaum, A. Majerovics, and H. Singh. Robust detection and classification of longitudinal changes in color retinal fundus images for monitoring diabetic retinopathy. *IEEE transactions on biomedical engineering*, 53(6):1084–1098, 2006.
[32] S. Pachade, P. Porwal, and M. Kokare. A novel method to detect fovea from color fundus images. In *Computing, Communication and Signal Processing*, pages 957–965. Springer, 2019.
[33] A. Punnolil. A novel approach for diagnosis and severity grading of diabetic maculopathy. In *2013 international conference on advances in computing, communications and informatics (ICACCI)*, pages 1230–1235. IEEE, 2013.
[34] R. J. Qureshi, L. Kovacs, B. Harangi, B. Nagy, T. Peto, and A. Hajdu. Combining algorithms for automatic detection of optic disc and macula in fundus images. *Computer Vision and Image Understanding*, 116(1):138–145, 2012.
[35] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
[36] S. Sedai, R. Tennakoon, P. Roy, K. Cao, and R. Garnavi. Multi-stage segmentation of the fovea in retinal fundus images using fully convolutional neural networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 1083–1086. IEEE, 2017.
[37] S. Sekhar, W. Al-Nuaimy, and A. K. Nandi. Automated localisation of optic disk and fovea in retinal fundus images. In *2008 16th European Signal Processing Conference*, pages 1–5. IEEE, 2008.
[38] P. F. Sharp, J. Olson, F. Strachan, J. Hipwell, A. Ludbrook, M. O'Donnell, S. Wallace, K. Goatman, A. Grant, N. Waugh, et al. The value of digital imaging in diabetic retinopathy. *Health technology assessment (Winchester, England)*, 7(30):1–119, 2003.
[39] P. Siddalingaswamy and K. G. Prabhu. Automatic grading of diabetic maculopathy severity levels. In *2010 International Conference on Systems in Medicine and Biology*, pages 331–334. IEEE, 2010.
[40] K. Simonyan and A. Zisserman. Very deep convolutional networks for

large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
[41] C. Sinthanayothin, J. F. Boyce, H. L. Cook, and T. H. Williamson. Automated localisation of the optic disc, fovea, and retinal blood vessels from digital colour fundus images. *British journal of ophthalmology*, 83(8):902–910, 1999.
[42] 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. *Journal of Computational Science*, 20:70–79, 2017.
[43] J. Tombran-Tink and C. J. Barnstable. *Retinal degenerations: biology, diagnostics, and therapeutics*. Springer Science & Business Media, 2007.
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
[45] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.
[46] D. Welfer, J. Scharcanski, and D. R. Marinho. Fovea center detection based on the retina anatomy and mathematical morphology. *Computer methods and programs in biomedicine*, 104(3):397–409, 2011.
[47] R. Xie, L. Liu, J. Liu, and C. S. Qiu. Pathological myopic image analysis with transfer learning. *International Conference on Medical Imaging with Deep Learning – Extended Abstract Track*, 2019.
[48] C. Y. Yu, C. C. Liu, and S. S. Yu. A novel scheme for the fovea localization on retinal images. In *2014 International Symposium on Computer, Consumer and Control*, pages 609–612. IEEE, 2014.
[49] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.
[50] S. Zheng, Y. Zhu, L. Pan, and T. Zhou. New simplified fovea and optic disc localization method for retinal images. *Journal of Medical Imaging and Health Informatics*, 9(4):847–855, 2019.