information

Article

Image Stitching of Low-Resolution Retinography Using Fundus Blur Filter and Homography Convolutional Neural Network

Levi Santos [1], Maurício Almeida [1], João Almeida [1,*], Geraldo Braz [1], José Camara [2,3] and António Cunha [2,3]

[1] Applied Computing Group (NCA—UFMA), Federal University of Maranhão, Av. dos Portugueses, 1966—Vila Bacanga, Saint Louis 65080-805, MA, Brazil; levi.cs@discente.ufma.br (L.S.); mauricio.ma@discente.ufma.br (M.A.); geraldo@nca.ufma.br (G.B.)
[2] School of Science and Technology, University of Trás-os-Montes e Alto Douro, Quinta de Prados, 5000-801 Vila Real, Portugal; jcrcamara@hotmail.com (J.C.); acunha@utad.pt (A.C.)
[3] ALGORITMI Research Centre, University of Minho, 4800-058 Guimaraes, Portugal
* Correspondence: jdallyson@nca.ufma.br

Abstract: Great advances in stitching high-quality retinal images have been made in recent years. On the other hand, very few studies have been carried out on low-resolution retinal imaging. This work investigates the challenges of low-resolution retinal images obtained by the D-EYE smartphone-based fundus camera. The proposed method uses homography estimation to register and stitch low-quality retinal images into a cohesive mosaic. First, a Siamese neural network extracts features from a pair of images, after which the correlation of their feature maps is computed. This correlation map is fed through four independent CNNs to estimate the homography parameters, each specializing in different corner coordinates. Our model was trained on a synthetic dataset generated from the Microsoft Common Objects in Context (MSCOCO) dataset; this work added an important data augmentation phase to improve the quality of the model. Then, the same is evaluated on the FIRE retina and D-EYE datasets for performance measurement using the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). The obtained results are promising: the average PSNR was 26.14 dB, with an SSIM of 0.96 on the D-EYE dataset. Compared to the method that uses a single neural network for homography calculations, our approach improves the PSNR by 7.96 dB and achieves a 7.86% higher SSIM score.

Keywords: image stitching; retinography; low resolution; homography; convolutional neural network

Citation: Santos, L.; Almeida, M.; Almeida, J.; Braz, G.; Camara, J.; Cunha, A. Image Stitching of Low-Resolution Retinography Using Fundus Blur Filter and Homography Convolutional Neural Network. *Information* **2024**, *15*, 652. https://doi.org/10.3390/info15100652

Academic Editors: Francesco Fontanella and Shuohong Wang

Received: 16 August 2024
Revised: 7 October 2024
Accepted: 9 October 2024
Published: 17 October 2024

1. Introduction

While the World Health Organization, as estimated by the 2019 World Sight Report [1], says that 2.2 billion people worldwide suffer from some domain of visual impairment, nearly 1 billion of the cases are avoidable or treatable. In this scenario, early detection and accurate diagnosis of eye pathologies become critical for the prevention of loss of vision and the maintenance of patients' quality of life. One of the biggest challenges to eye health is the lack of ophthalmologists, especially in remote distanced places and developing countries. This is causing a deficiency in access to timely diagnosis and high-quality management of eye disorders. On the other hand, advances in technology—like the one represented by the empowerment of hand-held mobile phones equipped with the D-EYE device for digital imaging of the retina [2]—open up new opportunities for early diagnosis.

Eye screening by visualizing the retina through fundus imaging offers an opportunity to non-invasively examine the systemic microcirculation in the human retina. Detailed clinical observations of the characteristics of the retinal fundus contribute not only to detecting eye diseases but also to identifying early indicators of a wide range of pathologies, such as diabetes, stroke, hypertension, arteriosclerosis, cardiovascular, neurodegenerative, renal and fatty liver diseases [3].

One challenge that limits the early detection of vision pathologies is the need for more professionals in regions far from large urban centers, especially in developing countries. A

limited amount of the literature exists pertaining to the number of optometrists and related ophthalmic personnel. According to a recent study conducted in 2019, which examined the ophthalmology workforce in 198 countries, representing 94% of the global population, it was found that despite the increasing number of practicing ophthalmologists in most countries, there exists an uneven distribution and a significant shortfall in the current and projected number of ophthalmologists [4]. According to the International Agency for the Prevention of Blindness (IAPB), there are critical human resource shortages for allied ophthalmic personnel, especially in sub-Saharan Africa [5].

Considering this gap, methods that can help in the screening process for ocular pathologies can be great allies for ophthalmologists and patients. One viable option for retinal screening is using mobile devices, such as the D-EYE device [6].

Although devices like the D-EYE have been developed as a low-cost tool to capture images of the fundus with just a simple attachment to a smartphone, these still have limitations in image quality. Figure 1 shows some frames of a video obtained with the D-EYE device, and as can be observed, the images contain a perceptible amount of noise and cannot always allow a complete visualization of the optic disk, which is the region of interest for a possible diagnosis, and it is already just a tiny part of the image.

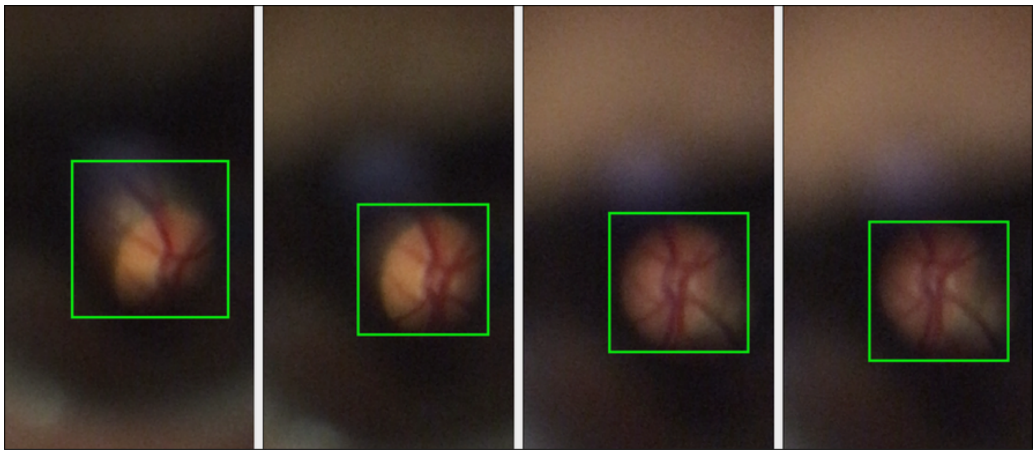

Figure 1. Different frames of the same video obtained by the D-EYE; the green squares indicate the visible contents of the optic disk and retina.

Among the few studies in the literature related to the problem of low-resolution retinographies, image stitching is the one developed by [7], presenting a comprehensive method that includes the steps of retinal segmentation, identification and matching of reference points, image alignment, and fusion. Variation in the video capture conditions, for example, due to changes in the retinal illumination, impacts the consistency of the results realized with this method. Furthermore, investigating the detection and montage of low-quality retinal images [8], brought out the approach of Super Retina by [9]. This method used semi-supervised deep learning model training techniques for keypoint detectors and descriptors involving the proposed Keypoint Progressive Expansion (KPE) technique, which addresses incompleteness issues arising from the manual labeling of data used in training. Although Super Retina was mostly tested on high-resolution images, it achieved remarkable feature matching and produced well-aligned results. Another technique reviewed by [8] was the stitching of several images suggested by [10], expected to give a complete image with captures obtained from an attached smartphone. This may be a very promising approach for retrieving a complete exposure of the retina, using Uniform Robust Scale-Invariant Feature Transform (UR-SIFT) [11] for feature detection and combining techniques for feature matching and refinement. However, the effectiveness of low-quality images, such as those obtained with D-EYE, may require the application of image pre-processing and image stitching techniques to improve visualization and analysis, both by experts and by automatic methods.

Therefore, this paper aims to address the problem of low-resolution retinal images by employing the image stitching technique to improve the visualization of images captured by the D-EYE. Thus, a method for image stitching between low-resolution retinographies via homography estimation is proposed based on the homography estimation model proposed by [12] with relevant modifications. Thus, we can highlight the following contributions: (1) we proposed a Fundus Blur Filter (FBF) originally designed in this study to mitigate the variations arising from the image capture conditions, which is fundamental to the success of the method; (2) we proposed a modification to the homography estimation network architecture in [12], utilizing four independent semi-siamese convolutional neural networks for more precise coordinate regression, instead of the original single regression.

The rest of this paper is organized as follows. Section 2 presents the proposed method, including the method for generating the synthetic dataset used to train the deep learning model used in homography estimation and the FBF applied to the input images during the model training, originally created for and used in this work, which is essential for the model to work. Section 3 presents the results of the proposed method and discusses the results. Section 4 presents the conclusions and future work.

2. Materials and Method

This section describes the proposed method for image stitching on low-resolution retinal images. First, we describe the dataset used to train the deep learning model. Next, we discuss the different stages of the proposed method, including data augmentation, feature extraction, and homography estimation, as well as the evaluation metrics and experimental setup used to validate the method.

2.1. Synthetic Dataset

Training a neural network for image stitching typically necessitates a large dataset [13,14]. Due to the scarcity of datasets specifically related to stitching low-resolution retinal images, the approach presented by [13] was used as a base for the generation of the synthetic dataset used in this work, modified to suit the needs of this application. For this, 80,000 images from the MS COCO dataset [15] were used.

Each image in the MS COCO dataset is cropped, adjusted to a square format, and resized to 274 $\times$ 274 pixels. Subsequently, a new crop is taken from the center of each image, represented in Figure 2A by the white square, which becomes the first input to the neural network. A position—blue square in Figure 2A—for the second crop is marked to ensure overlap with the first. The corners of this position are adjusted to simulate slight variations in the viewing angle, causing deformation relative to the initial marking; the deformed blue square is represented by the orange square in Figure 2B. The pixel distances between the corners of the original second crop position and the deformed marking are recorded as the label for the deep learning model, used in calculating the displacement needed in each corner of the second image to align with the first. The image is then warped so that the corners of the second crop form a square (orange square in Figure 2C), which serves as the second input to the network. The label for the stitching is the union of white and blue square contents in Figure 2B, as can be seen in Figure 2D. Experiments were conducted by adjusting the overlap between images within 50% to 100%. Such adjustments consider the characteristics of images obtained through D-EYE, which are captured from videos, allowing for increased overlap when selecting adjacent frames.

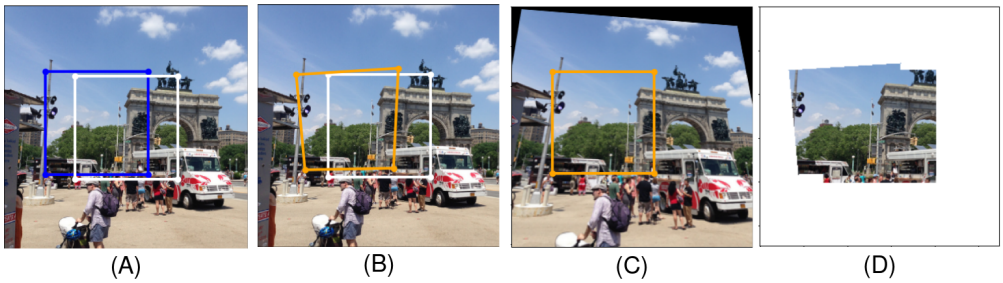

Figure 2. Synthetic dataset generation steps. (**A**) Delimitation of the first image pair used to train the network. (**B**) Small angle variations are applied to the image delimited in (**A**). (**C**) Deformed image for using as the second input to the network (**D**). Label for the stitching, resulting from the union of the contents of the white and orange squares in (**B**).

2.2. Proposed Method

The proposed method for estimating homography in low-resolution retinal images was inspired by the work of [12], with some adaptations.

A Fundus Blur Filter is applied exclusively during the training phase to deal with variations in the image capture condition, a Siamese convolutional neural network [16,17] for feature extraction, a feature correlation layer, and a regression network to predict the points that will be used to calculate homography using the Direct Linear Transformation (DLT) algorithm [18]. Figure 3 shows the proposed method pipeline.

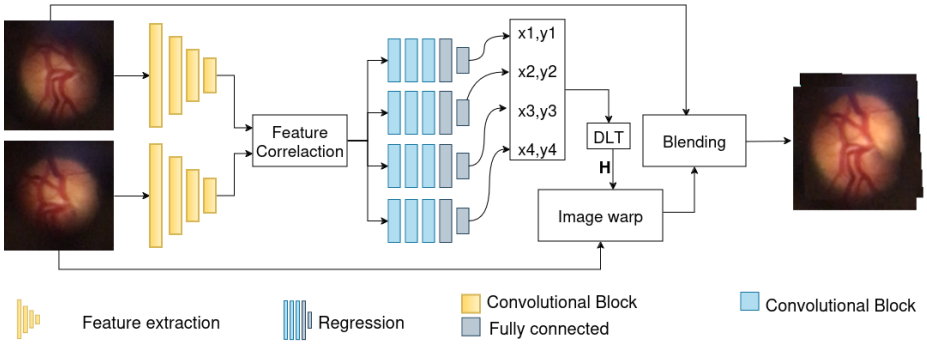

Figure 3. Proposed method pipeline.

2.2.1. Data Augmentation

Similar approaches for synthetic dataset generation are common in training homography estimation models [12,19]. However, it has disadvantages for predicting homography between image crops obtained under different lighting conditions, as a deep learning model usually performs optimally when it is fed with data closely similar to those samples seen during training [20]. In order to avoid this limitation and achieve generalization, especially with low-resolution images coming from the same retina captured at different moments, a significant innovation was introduced. Thus, we proposed a filter, denominated Fundus Blur Filter (FBF) (Figure 4), consisting of a black image with a white ellipse and a circle surrounding the ellipse, which was applied randomly in terms of position and size around the center of the image. This filter varies the input images' intensity toward the ellipse, making the model more robust to variations that D-EYE images may contain. Such modification generalizes the model between crops of the same image and between images of the same retina captured at different moments, making the approach adaptable to varying image acquisition conditions.

An improvement was introduced to overcome the limitations in predicting homography between images obtained under different lighting conditions in low-resolution retinal images, such as those faced by [7]. This improvement provides an effective strategy for ensuring the robustness and generalizability of the homography estimation model.

The proposed solution involves applying a blur filter consisting of a black mask containing a white ellipse and a surrounding circle, with random positions and sizes around the center of the image. This filter is then applied to the input images during training, gradually adjusting the intensity towards the ellipse, as shown in Figure 4. To do this, Equation (1) is used, where f_0 is the input image at position x, $f_1(x)$ is the filter at position x, $g(x)$ is the resulting image at position x, and α controls the blending ratio [21].

$$g(x) = (1 - \alpha) f_0(x) + \alpha f_1(x) \quad (1)$$

By gradually adjusting the intensity of the images towards the ellipse, we improve the model's ability to deal with the nuances in the images captured by the D-EYE device. The FBF allows the model to generalize between clippings of the same image and between images captured at different times with variations in retinal illumination, thus improving the quality and consistency of the stitching process. The implementation details of the FBF are outlined in Algorithm 1, while the parameters used in the experiments conducted in this work are presented in Table 1.

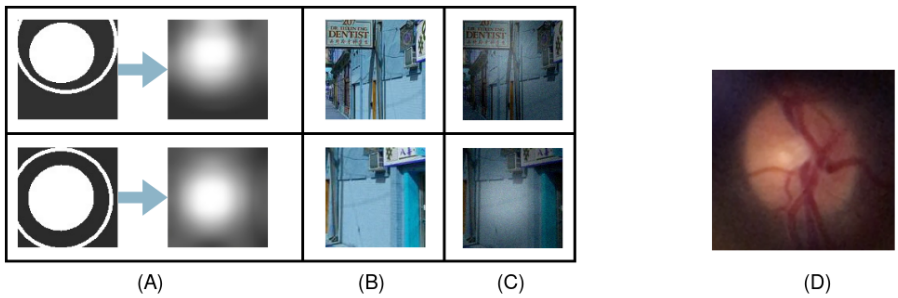

Figure 4. Example of Fundus Blur Filter application. (**A**) Filter's creation. (**B**) Input images. (**C**) Inputs with filters applied. (**D**) Sample from the D-EYE image dataset.

Algorithm 1. Fundus_Blur_Filter

Initialize the filter:
filter ← matrix(image_size, image_size) filled with 0

Define the center for the ellipse with some offset:
$ellipse\_center\_x \leftarrow (image\_center\_x + \text{ellipse\_offset\_x})$
$ellipse\_center\_y \leftarrow (image\_center\_y + \text{ellipse\_offset\_y})$
$ellipse\_center \leftarrow (ellipse\_center\_x, ellipse\_center\_y)$

Defines the circle center around the ellipse:
circle_center_x ← ellipse_center_x + circle_offset_range_x
circle_center_y ← ellipse_center_y + circle_offset_range_y

Define the axes of the ellipse with random values within the axis ranges:
ellipse_axes ← (ellipse_axis_range_x, ellipse_axis_range_y)

Randomly select the angle for the ellipse:
ellipse_angle ← ellipse_angle_range

Draw the ellipse on the filter at the defined center, axes, and angle:
draw_ellipse(filter, ellipse_center, ellipse_axes, ellipse_angle, color = 255)

Draw the circle on the filter at the defined center:
draw_circle(filter, circle_center, radius = image_size/2, color = 255, thickness = 3)

Apply Gaussian blur to the filter:
filter ← GaussianBlur$(filter, \text{blur_kernel}, \text{blur_sigma_range})$

Table 1. Parameters of the Fundus Blur Filter.

Parameter	Description	Variation
image_size	Size of the image	128
image_center	Center of the image	(64,64)
ellipse_offset_x	Random integer offset in the X-axis	−25 to 25
ellipse_offset_y	Random integer offset in the Y-axis	−25 to 25
circle_offset_x	Random integer offset in the X-axis	−40 to 40
circle_offset_y	Random integer offset in the Y-axis	−40 to 40
ellipse_axis_range_x	Random integer X-axis of the ellipse	38 to 42
ellipse_axis_range_y	Random integer Y-axis of the ellipse	35 to 42
ellipse_angle_range	Random integer angle of the ellipse	0° to 360°
blur_kernel	Size of the blur kernel	(51, 51)
blur_sigma_range	Range of sigma for Gaussian blur	20 to 30
center_threshold_range	Range of center offset	−5 to 5

2.2.2. Feature Extraction

Feature extraction is one of the significant steps in image stitching, which helps match the features between images to estimate the geometric relationship. Therefore, an approach motivated by the work in [12] is used, in which a Siamese convolutional neural network was used for feature extraction along with a feature correlation layer.

In this work, the input images are transformed to grayscale, following most of the practices on training homography estimation models [19,22,23]. The architecture of the network consists of two identical sub-networks with the same weights, each one being composed of four blocks, where each block has a convolutional layer with kernel size 3×3 and a stride of 1. The number of output filters in the first two layers is 64, and 128 in the last two layers. Batch normalization and max-pooling 2D with a pool size of 2×2 are applied after each convolutional layer to reduce the data dimension, followed by a ReLU activation function. The sub-networks process the input images in parallel and extract their feature maps, which are then passed to the subsequent Feature Correlation phase. Figure 5 depicts this architecture, where Feature Map F_A contains the features extracted from Input Image 1, and Feature Map F_B contains the features extracted from Input Image 2.

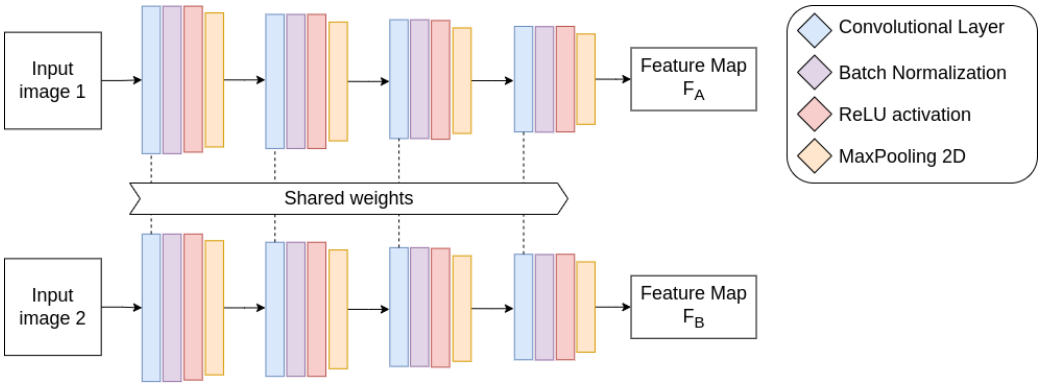

Figure 5. Feature extraction network architecture.

2.2.3. Feature Correlation

The next step in the network is the feature correlation layer, following the feature extraction phase, where the correspondences between feature maps F_A and F_B extracted from Input Images 1 and 2, are calculated using the proposed method in [12].

Here, L2 normalization is applied to the feature maps F_A and F_B, $\in W \times H \times C$; then, correspondences $D \in W \times H \times (W \times H)$ between these two maps are obtained. W, H, and C represent the dimensions of the feature maps, respectively, as well as the width, height, and number of filters (channels). The correspondence at every position in F_A with regard to all positions in F_B is calculated as follows:

$$D(x_1, x_2) = \frac{\langle F_A(x_1), F_B(x_2) \rangle}{|F_A(x_1)||F_B(x_2)|}, \quad (x_1, x_2) \in \mathbb{Z}^2 \tag{2}$$

where (x_1, x_2) are positions in feature maps, while $F_A(x_1)$ is a one-dimensional feature vector at position x_1 in the feature map F_A. $D(x_1, x_2) \in [0, 1]$, where a value closer to 1 it means that the features in the corresponding positions of the F_A and F_B maps correspond better, thus highlighting the regions where features are more alike.

2.2.4. Regression

To calculate the homography that maps points from one image to another, it is necessary to identify corresponding points between the two images. In this case, we perform the regression of the distances between the corners of an image and themselves in a position where the image aligns with the other image.

For the regression, four semi-Siamese convolutional neural sub-networks were employed. Unlike traditional Siamese networks, where branches share identical weights, semi-Siamese networks have the same structure but allow for different weights in each branch. Each sub-network calculates an ordered pair (x, y), which represents the distances between corners of the second input of the model regarding the deformed marking described in Section 2.1. These distances are the displacement required for each corner of the first input image to align with the second one. Figure 6 shows the Regression Network Architecture proposed. Unlike the [12] method that uses a single neural network for this task, this would allow the homography between low-resolution retinal images to be estimated with more precision since each coordinate is estimated individually.

These sub-networks take the tensor containing the correspondences of features as the input and are structured with four convolutional layers with 512 output filters each. Batch normalization and the ReLU activation function were applied after each convolutional layer, followed by two fully connected layers with a 50% dropout between them. The four networks will have their outputs concatenated to obtain a vector of size 8, which will be the distance between the four corners of the markings for the second input to the network. These are added to the initial coordinates of the corners and provide the final coordinates. The initial and final coordinates are used in computing the homography via the DLT algorithm, reported in the literature by [18].

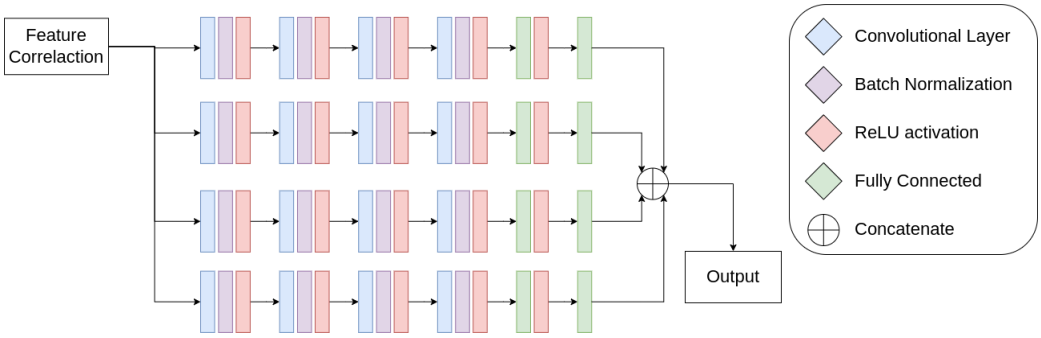

Figure 6. Regression Network Architecture.

2.2.5. Image Stitching

Image stitching, also known as image mosaicking, is a process that combines images with overlapping areas to form a wide-view, high-resolution image [24], often referred to as a panorama. The image stitching process involves five main steps: acquiring the images with overlapping regions, detecting distinctive features such as corners or edges in each image, matching corresponding features between the overlapping images, and aligning them based on those matches. After alignment, the images are blended to minimize visible seams and differences in exposure, ending in the generation of the final stitched image [21]. The resulting images must have overlapping areas and represent the same scene.

To perform the image stitching process, we put each 128 × 128 input image in the center of a black frame of size 274 × 274 pixels. Then, we use the inverse of the homography matrix **H** to project each pixel (u, v) of the second image to its new position (u', v') in the resulting image. In Equation (3), w is the normalization or scaling factor that appears when we represent coordinates in a projective space. Just like w, w′ is used to transform the homogeneous coordinates back to Cartesian coordinates by dividing x' and y' by w'. This ensures that the scale is removed and the points are represented correctly in the 2D plane [21]. The transformation is provided by Equation (3) in homogeneous coordinates.

$$\begin{pmatrix} u' \\ v' \\ w' \end{pmatrix} = \mathbf{H}^{-1} \begin{pmatrix} u \\ v \\ w \end{pmatrix} \tag{3}$$

Subsequently, after applying the homography matrix, image fusion is carried out to blend the overlapping images seamlessly so that minimal differences show up at the places where they meet. Fusion is dealt with, pixel by pixel, in an elementary way just by picking out each color channel from both images for the maximum value. The implementation details of our proposed method are available at https://gitlab.com/jdallyson/fbf_homography_cnn.

2.3. Evaluation Metrics

To compare the quality between the result of stitching and the expected image, metrics that are widely used in the literature [25] were adopted: Structural Similarity Index, Peak Signal-to-Noise Ratio, and Root Mean Squared Error.

Natural image signals have strong spatial dependencies that convey essential structural information, which traditional metrics like RMSE and PSNR fail to capture due to their reliance on pointwise differences [26]. These metrics can yield the same error value for different types of distortions, making them less effective in assessing structural integrity [27]. In contrast, the Structural Similarity Index (SSIM) directly compares image structures by evaluating luminance, contrast, and correlation, aligning more closely with human perception [28]. While SSIM is better suited for capturing the perceptual quality critical to image stitching, RMSE and PSNR still provide valuable insight into pixel-level errors.

The Structural Similarity Index (SSIM) between two images α and β is defined as

$$SSIM(\alpha, \beta) = \frac{(2\mu_\alpha\mu_\beta + C_1)(2\sigma_{\alpha\beta} + C_2)}{(\mu_\alpha^2 + \mu_\beta^2 + C_1)(\sigma_\alpha^2 + \sigma_\beta^2 + C_2)} \tag{4}$$

where μ_α and μ_β represent the mean pixel values of images α and β, respectively, providing an average intensity for each image. The terms σ_α^2 and σ_β^2 denote the variances of the pixel values in images α and β. The covariance between the images is represented by $\sigma_{\alpha\beta}$. Constants C_1 and C_2 are included to stabilize the division when the denominators are close to zero, typically defined as $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$, where K_1 and K_2 are small constants, and L is the dynamic range of pixel values (e.g., 255). The SSIM index ranges from -1 to 1, with 1 indicating perfect similarity between the images.

The Peak Signal-to-Noise Ratio (PSNR) measures the quality between the label and the image resulting from the proposed method; the higher the PSNR, the better the resulting quality, while lower PSNR values mean more significant differences between the images. It is defined as

$$\text{PSNR} = 20 log_{10}\left(\frac{MAX_f}{\sqrt{MSE}}\right) \tag{5}$$

where MAX_f is the maximum pixel value in the image, and MSE is the Mean Squared Error, which is defined as

$$\text{MSE} = \frac{1}{n}\sum(y - \hat{y})^2 \tag{6}$$

where n is the number of pixels in each image, y is a pixel from the label image, and $\hat{y}$ is the corresponding pixel in the resulting image.

The RMSE measures the difference between the result and the label and is defined as follows:

$$\text{RMSE} = \sqrt{MSE} \tag{7}$$

2.4. Experiments

To estimate homography, we used the model proposed in this work, trained with 120,000 pairs of 128 × 128 pixel images, the same number of 274 × 274 pixel label images and distance vectors described in Section 2.1, for validation. Additionally, 12,000 pairs of images, labels and distance vectors were used for validation during training. These images were generated from a set of 80,000 images from the MSCOCO dataset [15], following the method described in Section 2.1. During the training, the model reached its best weights after 9 epochs with a learning rate of 0.0001 [29–31], a common value used to train deep learning image processing models, and a batch size of 7 as it was the maximum batch value that the machine used could support. The machine configuration used included an Intel i5-11400 processor, 16 GB of RAM and an NVIDIA GeForce RTX 3060 GPU with 12 GB of VRAM.

The tests were carried out on 40 pairs of images extracted manually from the D-EYE video dataset [2]. During extraction, care was taken to avoid frames with little or no relevant information for diagnosis, similar to those illustrated in Figure 7. The green channel of the retinal images contained more relevant information and was therefore used to estimate homography. The videos were taken without dilating the pupil, giving the patient greater comfort, but sacrificing image quality due to eye movements, pupil size and media opacity. The selected images only show the limits of the optic disc, which hinders the demonstration of the method in other structures of the posterior pole.

The label for each image was defined in GIMP [32] by manually aligning the image pairs. Afterward, the images were aligned again, swapping their positions. The resulting composite images were merged pixel by pixel, selecting the highest value for each channel.

Also, 49 pairs of high-resolution images from the FIRE dataset [33] were added to the study, specifically from subset 'P', which contains image pairs that could be useful for mosaicking applications. As far as these images are concerned, two kinds of tests were conducted: on the whole, FIRE 'P' images to test the alignment technique and, in a second phase, Regions Of Interest corresponding to the optical disk in each image were cut, producing a set of pairs of cut-outs that were then joined together. This was carried out just like with the D-EYE images in GIMP. The choice to include both the complete images and the optical disk cut-outs was aimed at verifying whether the proposed model is effective at different scales and levels of detail. Figure 8 shows one example of each dataset used in the tests.

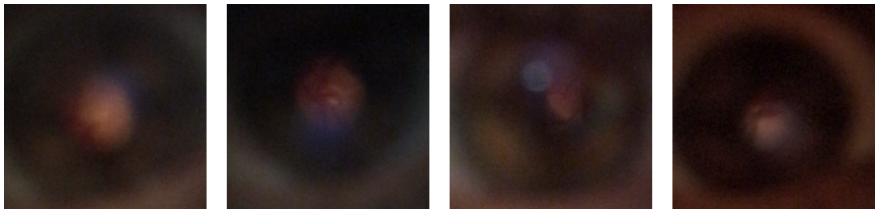

Figure 7. Images with no relevant information for stitching or diagnosis.

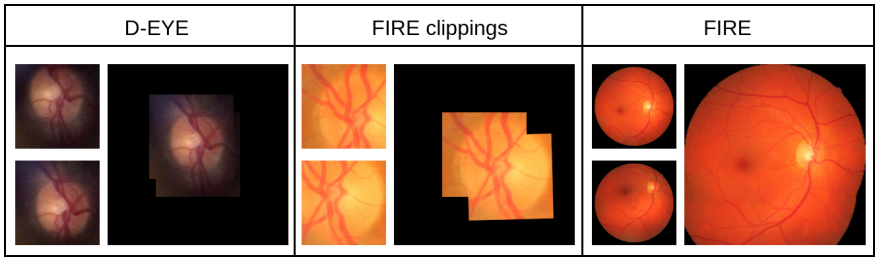

Figure 8. Examples of input and label images from the D-EYE and FIRE datasets used in the tests. The input images are the smaller images, and the labels are the larger ones.

3. Results and Discussion

This section describes and discusses the results obtained by the proposed method. To this end, the SSIM, PSN, MSE and RMSE metrics are presented in Section 2.3.

Table 2 shows the result with the homography estimation method of [12]. While it may seem that the values of the metrics are close to those in Table 3, the original approach was unhelpful in predicting the homography in the test cases coarsely. Mostly, the approach just centers the images and overlays them over each other without accurate alignment. It was not until the FBF proposed in this work that the network aligned the images correctly. The introduction of the FBF was a breakthrough for the homography estimation model, representing an improvement over previous results.

Table 2 reflects the limits of the former method, and Tables 3 and 4 represent a qualitative leap in the precision of the estimates. The introduction of the blur mask increased PSNR significantly from 18.18 dB to 22.90 dB, indicating a big drop in the image reconstruction error. The SSIM improved from 0.82 to 0.89, indicating enhanced structural fidelity after alignment. Moreover, the reduction in the RMSE from 4.89 to 3.92 shows improved accuracy of estimated distances between images.

Table 2. Results using the homography estimation method proposed by [12].

Dataset	Metric	Average	Standard Deviation	Minimum	Median	Maximum
D-EYE	PSNR↑	18.18	2.56	13.00	18.67	23.09
	SSIM↑	0.89	0.03	0.77	0.83	0.89
	RMSE↓	4.89	0.33	4.09	4.91	5.81
Fire P	PSNR↑	19.01	2.60	15.26	18.33	25.85
	SSIM↑	0.80	0.04	0.71	0.81	0.90
	RMSE↓	51.52	13.94	22.52	53.52	76.22
Fire P crop	PSNR↑	17.84	2.36	14.11	17.15	22.71
	SSIM↑	0.85	0.03	0.75	0.85	0.92
	RMSE↓	58.53	14.64	32.29	61.27	86.95

Table 3. Results using the homography estimation method proposed by [12] with the addition of the Fundus Blur Filter during training.

Dataset	Metric	Average	Standard Deviation	Minimum	Median	Maximum
D-EYE	PSNR↑	22.9	3.29	18.04	22.69	32.4
	SSIM↑	0.89	0.04	0.77	0.90	0.97
	RMSE↓	3.92	0.69	2.45	3.90	5.43
Fire P	PSNR↑	23.62	5.02	14.63	23.64	36.03
	SSIM↑	0.82	0.09	0.57	0.85	0.96
	RMSE↓	33.99	19.11	6.98	29.06	81.94
Clipping Fire P	PSNR↑	20.29	3.10	14.11	20.23	27.65
	SSIM↑	0.89	0.04	0.77	0.89	0.97
	RMSE↓	45.35	15.93	18.29	42.99	86.98

The proposed method goes further ahead in the improvements made. The average PSNR increased to 26.14 dB, the SSIM to 0.926, and the RMSE reduced to 3.21 for the D-EYE dataset. As observed in Table 4, it is standing apart and additionally proving that the blur mask is not some incremental improvement; rather, it is designed for playing a very critical role in the accurate and effective aligning of retinal images.

Figure 9 shows four examples of results where the input images have substantial illumination differences between them. Although the results are generally more accurate, minor displacements between the label and the result image are still present, as reflected in

the metric values shown in Table 4. Table 5 provides a summary of the average values of each metric across the datasets, facilitating a clearer comparison between the methods.

Table 4. Results obtained from our method.

Dataset	Metric	Average	Standard Deviation	Minimum	Median	Maximum
D-EYE	PSNR↑	26.14	3.78	16.93	26.17	33.15
	SSIM↑	0.96	0.04	0.80	0.93	0.97
	RMSE↓	3.21	0.85	1.74	3.09	5.19
Fire P	PSNR↑	24.08	6.27	14.86	23.49	36.99
	SSIM↑	0.80	0.10	0.56	0.82	0.95
	RMSE↓	34.84	22.25	6.25	29.57	79.81
Clippling Fire P	PSNR↑	25.46	3.80	17.82	25.19	36.89
	SSIM↑	0.94	0.03	0.83	0.94	0.98
	RMSE↓	25.66	10.45	6.31	24.29	56.76

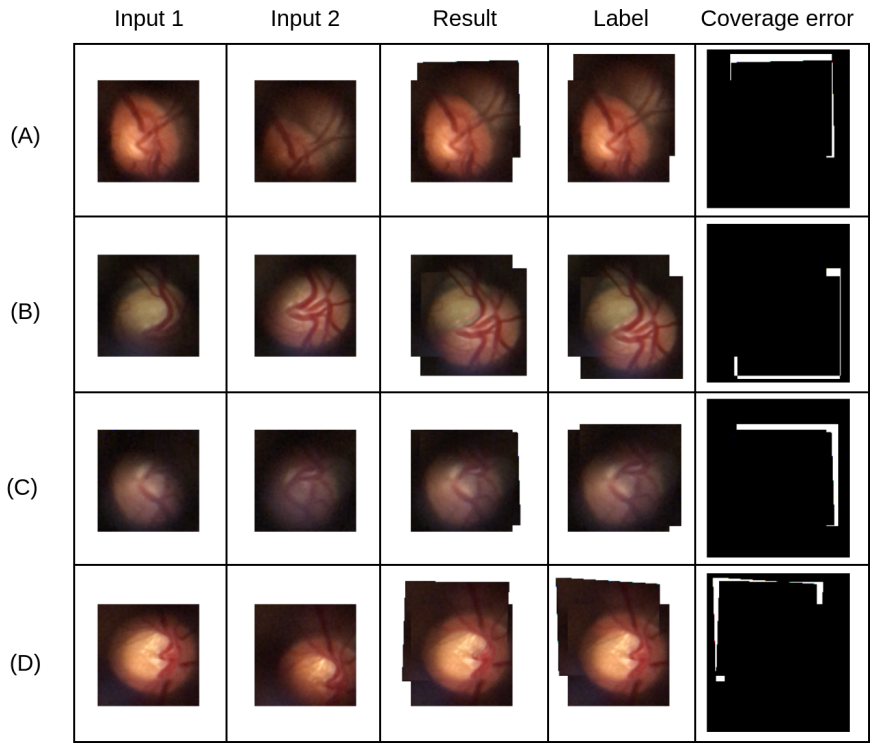

Figure 9. Example results of the proposed method for low-resolution retinography image stitching. White areas indicate misaligned regions, where the result either incorrectly covers or fails to cover the reference image. (**A**–**D**) Test case.

Table 5. Result by average metric for each method.

	Proposed by [12]			Proposed by [12] + FBF			Our Method		
	D-EYE	FIRE P	C FIRE P	D-EYE	FIRE P	C FIRE P	D-EYE	FIRE P	C FIRE P
PSNR↑	18.68	19.01	17.84	22.90	23.62	20.29	**26.14**	**24.08**	**25.46**
SSIM↑	0.89	0.80	0.85	0.89	**0.82**	0.89	**0.96**	0.80	**0.94**
RMSE↓	4.89	51.52	58.53	3.92	**33.99**	45.35	**3.21**	34.84	**25.66**

Highest values are indicated in bold.

Still, it was found that the individualized regression strategy for each point was really useful for correcting inaccuracies and improving alignment, as exemplified in Figure 10, where it is easy to see that the multiple regression method can provide results closer to the label than single regression.

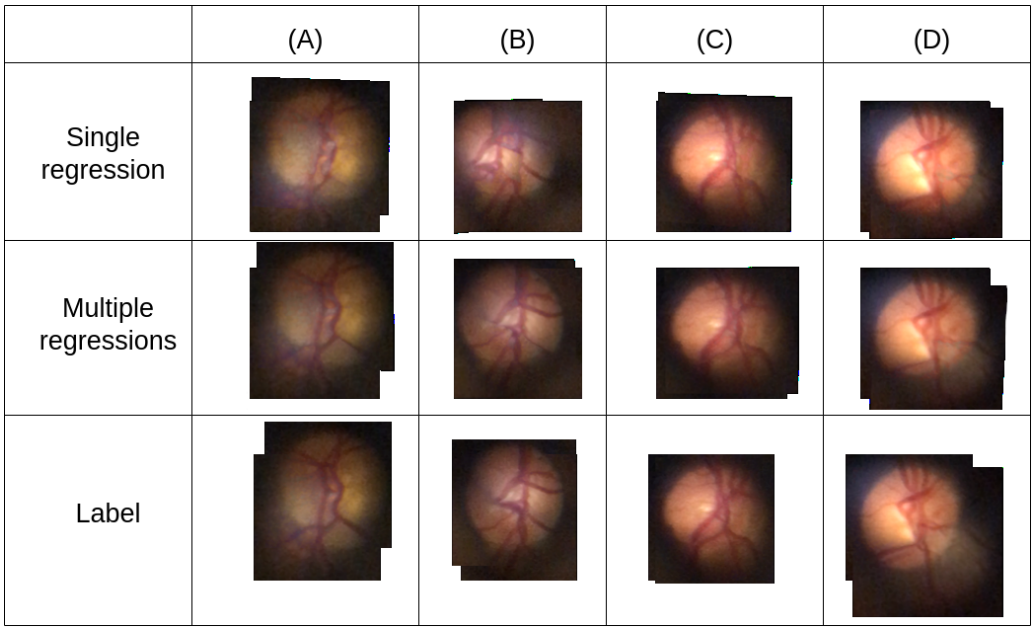

Figure 10. Visual comparison of results before and after applying individual regressions. Columns (**A**–**D**) each represent a test case.

Figure 11 presents comparative examples of the 3 methods; it is evident in the first row, where FBF is not used during model training, that the model is unable to estimate a homography matrix that can be used to align the images. In the second line, with the single regression that used FBF during the training in some images, such as (A), (D), (E) and (F), it is possible to observe inferior alignment quality compared to the proposed method.

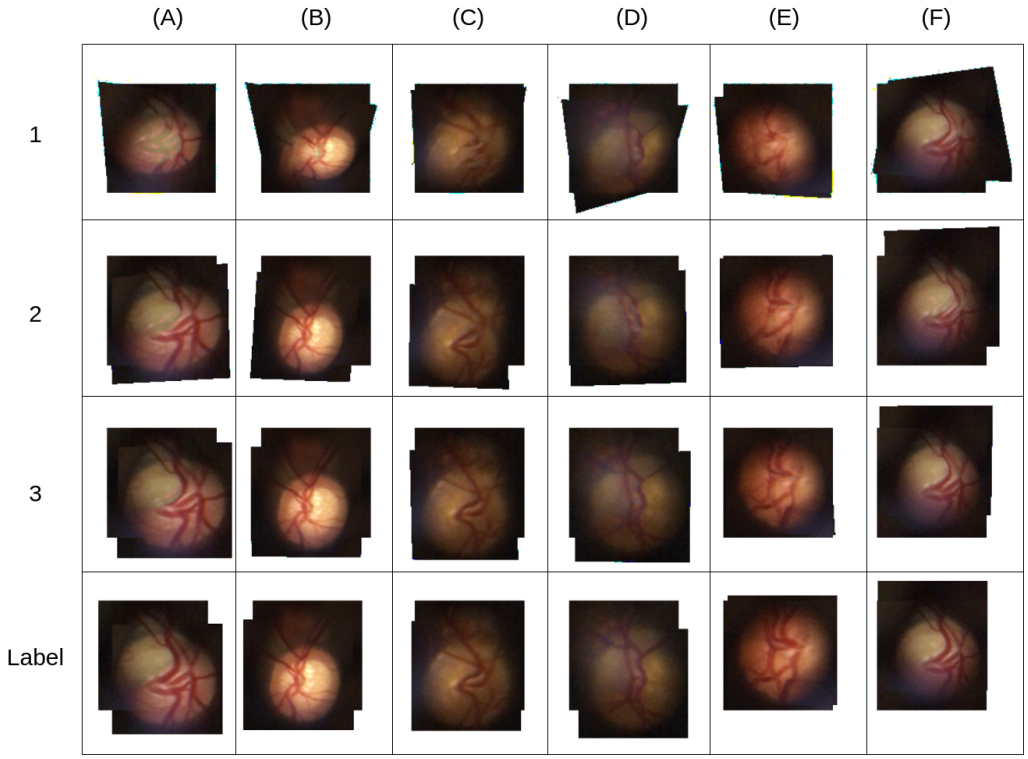

Figure 11. Comparative examples of different methods for test cases (**A**–**F**). The numbered lines show the results of the following methods: (1) method proposed by [12], (2) method proposed by [12] with the addition of FBF, and (3) our method. The "Label" line contains the expected results.

Evaluating the proposed method on high-resolution images from the P subset of the FIRE dataset showed that the model could return a sufficiently accurate homography in 21 cases out of 49 image pairs. Most of the failures occurred with images with little overlap between them, an essential factor for accurate homography estimation. However, this is not a significant obstacle to the problem presented in this work since it is possible to use neighboring frames of the videos for stitching, ensuring sufficient overlap. Figure 12A shows a pair of images from subsection "P" of the FIRE dataset, where, with sufficient overlap between the pictures, the proposed method was able to estimate an approximate homography matrix. Figure 12B shows a pair of images with little overlap, in which the model could not estimate a homography matrix that aligned with the figures.

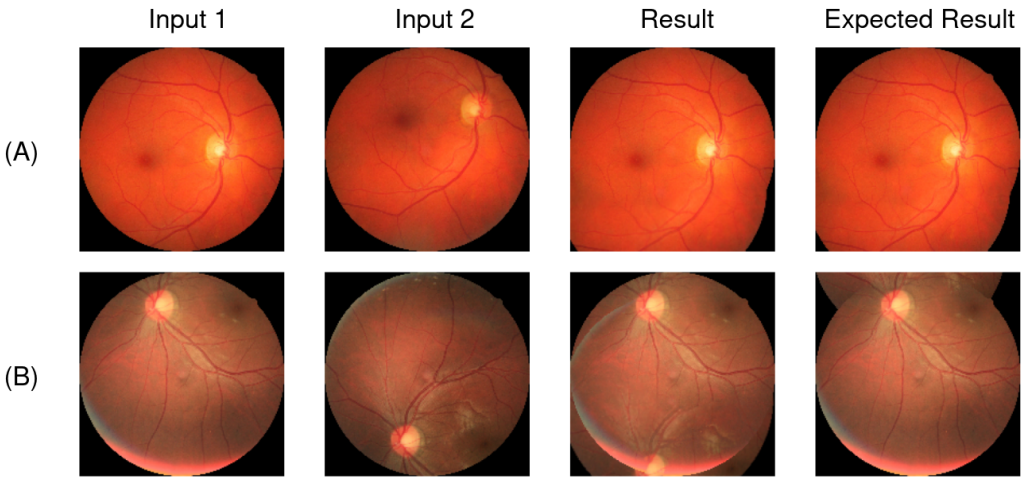

Figure 12. Test case examples of the proposed method with the subset P from the FIRE dataset. (**A**) Case where there is enough overlap. (**B**) Not enough overlap.

4. Conclusions

This research proposed a homography estimation method for stitching low-resolution retinal images. The method could extract relevant features from the input images and calculate a correlation feature, enabling the alignment of retinal images.

From the results discussed, we can see that the introduction of the Fundus Blur Filter (FBF) into the training process of the homography estimation model brought important improvements with regard to image alignment accuracy. The increase in PSNR and SSIM, as well as the reduction in RMSE, indicate that the proposed model was able to consistently improve the quality of the reconstruction, outperforming previously tested methods. Although the method proved effective, especially in the D-EYE dataset, when compared to the method proposed by [12], improving the PSNR by 7.96 dB and achieves a 7.86% higher SSIM score, there are still challenges in situations where there is little overlap between the images, as observed in the FIRE dataset. Despite these limitations, we believe that the use of neighboring frames, as discussed, can mitigate some of these difficulties, ensuring sufficient overlap for proper alignment.

As this current approach requires manual intervention in selecting frames from the videos obtained through D-EYE, acting as a drawback for use in a clinical setting, future work is expected to overcome the identified limitations, aiming at automating the selection of sections of the area of interest in the video frames captured using D-EYE to develop automatic algorithms for ordering frames to enable alignment and stitching of multiple images. This will enable better visualization of the retina by reconstructing low-resolution retinographies in a panorama, helping ophthalmologists identify eye diseases early enough to treat them.

Author Contributions: Conceptualization, L.S., A.C. and J.A.; methodology L.S., M.A. and J.A.; validation, J.A. and M.A.; formal analysis, L.S. and J.A.; investigation, L.S. and J.A.; writing—original draft preparation, L.S. and J.A.; writing—review and editing, A.C., J.C., M.A. and G.B.; supervision, J.A.; project administration, J.A. All authors have read and agreed to the published version of the manuscript.

Funding: The authors acknowledge the Fundação para a Ciência e Tecnologia, IP (FCT) within the R&D Units Project Scope: UIDB/00319/2020 (ALGORITMI), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil—Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil, and Fundação de Amparo à Pesquisa Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA) Brazil (Grant number 000527/2024) and Empresa Brasileira de Serviços Hospitalares (Ebserh) Brazil (Grant number 409593/2021-4) for providing financial support.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. WHO. *World Report on Vision*; World Health Organization: Geneva, Switzerland, 2019.
2. Neto, A.; Camara, J.; Cunha, A. Evaluations of deep learning approaches for glaucoma screening using retinal images from mobile device. *Sensors* **2022**, *22*, 1449. [CrossRef]
3. Pachade, S.; Porwal, P.; Thulkar, D.; Kokare, M.; Deshmukh, G.; Sahasrabuddhe, V.; Giancardo, L.; Quellec, G.; Mériaudeau, F. Retinal fundus multi-disease image dataset (RFMiD): A dataset for multi-disease detection research. *Data* **2021**, *6*, 14. [CrossRef]
4. Resnikoff, S.; Lansingh, V.C.; Washburn, L.; Felch, W.; Gauthier, T.M.; Taylor, H.R.; Eckert, K.; Parke, D.; Wiedemann, P. Estimated number of ophthalmologists worldwide (International Council of Ophthalmology update): Will we meet the needs? *Br. J. Ophthalmol.* **2020**, *104*, 588–592. [CrossRef]
5. Abdulhussein, D.; Abdul Hussein, M. WHO Vision 2020: Have we done it? *Ophthalmic Epidemiol.* **2023**, *30*, 331–339. [CrossRef]
6. Pihlblad, M.S.; Stockslager, S. D-EYE: A portable and inexpensive option for fundus photography and videography in the pediatric population with telemedicine potential. *J. Am. Assoc. Pediatr. Ophthalmol. Strabismus* **2016**, *20*, e21. [CrossRef]
7. Barritt, N.; Pilon, L.; MacLean, A.; Lin, A.; Cole, A.; Faruq, I.; Lakshminarayanan, V. Development and testing of a stabilization and image processing system for improvement of mobile fundus camera image quality. In Proceedings of the Novel Optical Systems, Methods, and Applications XXIII, SPIE, Virtual, 24 August–4 September 2020; Volume 11483, pp. 79–88.
8. Correia, T.V.S. Detection and Mosaicing through Deep Learning Models for Low-Quality Retinal Images. Master's Thesis, School of Technology and Management of the Polytechnic Institute of Leiria, Leiria, Portugal, 2023. Available online: https://iconline.ipleiria.pt/handle/10400.8/8892 (accessed on 7 August 2023).
9. Liu, J.; Li, X.; Wei, Q.; Xu, J.; Ding, D. Semi-supervised keypoint detector and descriptor for retinal image matching. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 593–609.
10. Hu, R.; Chalakkal, R.; Linde, G.; Dhupia, J.S. Multi-image stitching for smartphone-based retinal fundus stitching. In Proceedings of the 2022 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), IEEE, Sapporo, Japan, 11–15 July 2022; pp. 179–184.
11. Sedaghat, A.; Mokhtarzade, M.; Ebadi, H. Uniform Robust Scale-Invariant Feature Matching for Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote. Sens.* **2011**, *49*, 4516–4527. [CrossRef]
12. Nie, L.; Lin, C.; Liao, K.; Liu, M.; Zhao, Y. A view-free image stitching network based on global homography. *J. Vis. Commun. Image Represent.* **2020**, *73*, 102950. [CrossRef]
13. DeTone, D.; Malisiewicz, T.; Rabinovich, A. Deep image homography estimation. *arXiv* **2016**, arXiv:1606.03798.
14. Huang, R.; Chang, Q.; Zhang, Y. Unsupervised Oral Endoscope Image Stitching Algorithm. *J. Shanghai Jiaotong Univ. Sci.* **2024**, *29*, 81–90. [CrossRef]
15. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context. *arXiv* **2015**, arXiv:1405.0312.
16. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), IEEE, San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 539–546.
17. Chicco, D. Siamese neural networks: An overview. *Artif. Neural Netw.* **2021**, *2190*, 73–94.
18. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*; Cambridge University Press: Cambridge, UK, 2003.

19. Kang, L.; Wei, Y.; Xie, Y.; Jiang, J.; Guo, Y. Combining convolutional neural network and photometric refinement for accurate homography estimation. *IEEE Access* **2019**, *7*, 109460–109473. [CrossRef]
20. O'shea, K.; Nash, R. An introduction to convolutional neural networks. *arXiv* **2015**, arXiv:1511.08458.
21. Szeliski, R. *Computer Vision: Algorithms and Applications*; Springer Nature: Berlin/Heidelberg, Germany, 2022.
22. Liu, S.; Lu, Y.; Jiang, H.; Ye, N.; Wang, C.; Zeng, B. Unsupervised global and local homography estimation with motion basis learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 7885–7899. [CrossRef]
23. Zhou, Q.; Li, X. STN-Homography: Direct estimation of homography parameters for image pairs. *Appl. Sci.* **2019**, *9*, 5187. [CrossRef]
24. Wang, Z.; Yang, Z. Review on image-stitching techniques. *Multimed. Syst.* **2020**, *26*, 413–430. [CrossRef]
25. Jagalingam, P.; Hegde, A.V. A review of quality metrics for fused image. *Aquat. Procedia* **2015**, *4*, 133–142. [CrossRef]
26. Dissanayake, V.; Herath, S.; Rasnayaka, S.; Seneviratne, S.; Vidanaarachchi, R.; Gamage, C. Quantitative and Qualitative Evaluation of Performance and Robustness of Image Stitching Algorithms. In Proceedings of the 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Adelaide, Australia, 23–25 November 2015; pp. 1–6. [CrossRef]
27. Wang, Z.; Bovik, A.C. Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. *IEEE Signal Process. Mag.* **2009**, *26*, 98–117. [CrossRef]
28. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
29. Zhu, M.; Li, C.; He, X.; Xiao, X. Unsupervised deep learning image stitching model assisted with infrared images. In Proceedings of the International Conference on Algorithm, Imaging Processing, and Machine Vision (AIPMV 2023), Qingdao, China, 15–17 September 2023; SPIE: Bellingham, WA, USA, 2024; Volume 12969, pp. 249–255.
30. Duan, H.; Min, X.; Sun, W.; Zhu, Y.; Zhang, X.P.; Zhai, G. Attentive deep image quality assessment for omnidirectional stitching. *IEEE J. Sel. Top. Signal Process.* **2023**, *17*, 1150–1164. [CrossRef]
31. Ni, J.; Li, Y.; Ke, C.; Zhang, Z.; Cao, W.; Yang, S.X. A Fast Unsupervised Image Stitching Model Based on Homography Estimation. *IEEE Sens. J.* **2024**, *24*, 29452–29467. [CrossRef]
32. The GIMP Development Team. GIMP: GNU Image Manipulation Program. Version 2.10.36. The GIMP Development Team. 2023. Available online: https://www.gimp.org (accessed on 10 February 2024).
33. Hernandez-Matas, C.; Zabulis, X.; Triantafyllou, A.; Anyfanti, P.; Douma, S.; Argyros, A.A. FIRE: Fundus image registration dataset. *Model. Artif. Intell. Ophthalmol.* **2017**, *1*, 16–28. [CrossRef]