DEEP IMAGE REGISTRATION FOR COMPOSING SPECTRAL RETINAL IMAGES

Lappeenranta-Lahti University of Technology LUT

Master's Program in Computational Engineering, Master's Thesis

2022

Mikhail Farmakovskii

Examiner: Professor Lasse Lensu
D.Sc. (Tech.) Lauri Laaksonen

ABSTRACT

Lappeenranta-Lahti University of Technology LUT
School of Engineering Science
Computational Engineering

Mikhail Farmakovskii

Deep image registration for composing spectral retinal images

Master's thesis
2022
52 pages, 32 figures, 3 tables, 1 appendices
Examiners: Professor Lasse Lensu and D.Sc. (Tech.) Lauri Laaksonen

Keywords: computer vision, deep image registration, spectral retinal images

Eye diseases are very common among the population, which directly affect the quality of life. The anatomy of the eye is the complex structure that determines how the tissues will interact with light. Optical filters transmitting light at specific wavelengths can be used to obtain multispectral images of the eye fundus. This allows to observe various details of the organ such as blood vessels, nerves, and the retina. The issue of image registration is relevant since the combination of several color channels allows to expand useful information of one image which makes it possible to to improve the diagnosis of eye diseases. Due to a significant increase in computing power, availability of data, and methodological advances artificial neural networks have become the main tool for analyzing medical images. This work studies a deep image registration method based on the U-net architecture that produces high-quality spectral retinal images, to evaluate the alignment with a subsequent comparison with the existing dual-bootstrap iterative closest point algorithm using basic statistical methods. The studied technique is able to compete with existing ones.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to the LUT University, which organized the educational process very well, allowed me to touch the most modern knowledge in the field of machine learning, and introduced people with whom I could share my scientific interest. I am also grateful for the university building for a comfortable stay with all amenities.

I could not have undertaken this journey without my supervisor Lasse Lensu, who from the very beginning helped with this thesis, introduced me to the scientific world. His support, guidance and overall insights have made this an inspiring experience for me.

I would like to thank my family who provided a financial opportunity to study abroad. I would be remiss in not mentioning friends and their sense of humor which makes life more fun.

Lappeenranta, June 21, 2022

Mikhail Farmakovskii

LIST OF ABBREVIATIONS

ADAM	adaptive moment estimation
AF	activation function
AMD	age-related macular degeneration
CCD	charge-coupled device
CFP	color fundus photography
CNN	convolutional neural network
CPs	control points
DA	data augmentation
DL	deep learning
DR	diabetic retinopathy
FAF	fundus autofluorescence
GCL	ganglion cell layer
GDBICP	generalized dual bootstrap iterative closest point algorithm
HSI	hyperspectral imaging
ILM	inner limiting membrane
INL	inner nuclear layer
IPL	inner plexiform layer
KL	Kullback-Leibler
MI	mutual information
ML	maximum likelihood
MSI	multispectral imaging
NCC	normalized cross correlation
NFL	nerve fibre layer
NMI	normalized mutual information
OLM	outer limiting membrane
ONL	outer nuclear layer
OPL	outer plexiform layer
PE	pigment epithelium layer
R-CL	rods and cones layer
ReLU	rectified linear unit
RMSE	root mean square error
SA	spherical aberrations
SEFIG	sharpness estimation from image gradients
TL	Tukey loss
TRE	target registration error
ULF	unsupervised loss function

CONTENTS

1 INTRODUCTION

1.1 Background

The prevalence of eye diseases is growing rapidly [1]. For example, age-related macular degeneration (AMD), diabetic retinopathy (DR) are the most common eye diseases [2]. Researches aimed at studying AMD mainly evaluate the optical quality of the eye. There is a direct correlation between age group and focus mismatch for rays of light [2]. It was found that spherical aberrations (SA) regularly increase with age and possibly this correlates with the appearance of irregularities on the cornea and the hardening of the lens which reduces the sensitivity to contrast [3]. DR is a severe complication of diabetes mellitus [4]. This occurs due to microangiopathy which damages capillaries and small blood vessels [5]. Hemorrhages in the eye are attributed to a very serious manifestation of the disease which can lead to blindness.

Some national programs are already aimed at diagnosing DR [4, 6, 7]. It possible to diagnose the disease at an early stage, however, it increases the workload on the health-care system. Technologies for analyzing the fundus image are the key to the successful treatment of ophthalmic diseases. For example, color fundus photography (CFP), fundus autofluorescence (FAF), and multispectral imaging (MSI) have been used in recent years. MSI technique considers the concept of channel image which consists in depicting the fundus using a certain filter to fix the specific wavelength to highlight small details in images. For instance, by combining them, the hemoglobin level can be determined will allow for assessing the saturation of tissues with oxygen.

Due to insufficient data and the novelty of automatic analysis of the fundus image, no method solves all the problems of modern ophthalmology. Depending on the system, different stages of image processing can be investigated including data preprocessing. Obviously, the data is not always of good quality, in particular, humans can not remain motionless, and the eyes always make involuntary saccadic movements, that interfere with obtaining accurate images. Therefore, the problem of alignment is in comparing images, quantifying the differences between the real and processed image [8]. Researchers are faced with the issue of assessing the quality of raw data to select the optimal preprocessing algorithm.

In deep convolutional neural networks, which are used for many computer vision tasks a single very unrepresentative image can affect the final model performance. In conditions

of lack of data, data augmentation (DA) is used which can heavily decrease overfitting and boost model performance. The DA methods include geometric transformations, flipping, color augmentations, kernel filters, and others [9]. In conditions of poor focus, which is often the case with raw data, it is simply necessary to somehow evaluate the sharpness of the image, and sharpness estimation from image gradients (SEFIG) algorithm for grayscale images is used for mentioned purpose. SEFIG is based on the edge map generation [10]. A regular histogram suits for brightness estimation, followed by a mean and standard deviation calculation to highlight anomalous images. Due to a significant increase in computing power and data availability, convolutional neural network (CNN) has become the main tool for analyzing medical images [11]. Scientific articles [12–14] based on the use of deep learning (DL) for spectral image registration have been widely publicized and currently trained models indicate good accuracy.

To illustrate the proposed method, a workflow is shown in Fig. 1.

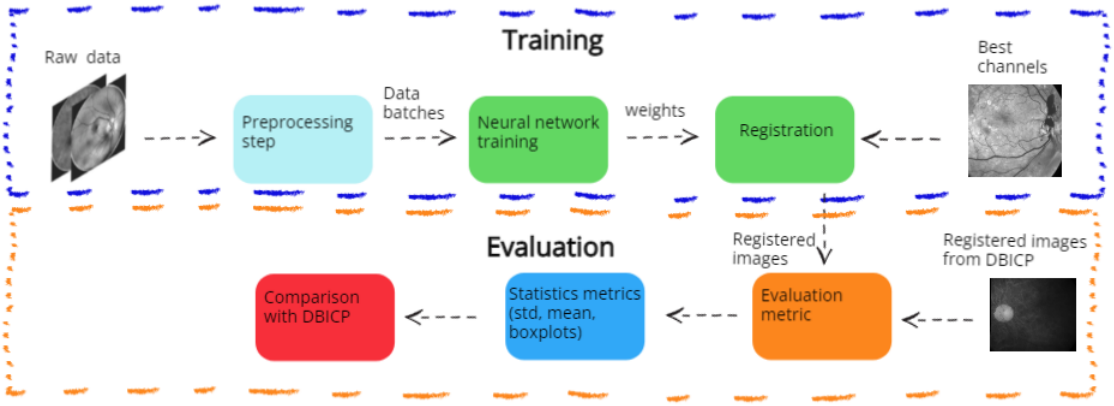


Figure 1. High level workflow.

1.2 Objectives and delimitations

This research aims to develop a deep image registration method that produces high-quality spectral retinal images and to evaluate the alignment. The detailed objectives of this research are as follows:

- Study image preprocessing for improved channel image alignment.
- Implement a deep image registration method to obtain spectral retinal images.
- Quantitatively evaluate image alignment by using a suitable measure.

- Compare the implemented channel image registration method performance with a benchmark.

Considering the applications, the developed approach for composing spectral retinal images can be used by ophthalmologists for deeper visualization of abnormalities in the eye and integration into an automatic eye disease recognition system. Evaluating and interpreting different eye objects from spectral images is not the purpose of this research.

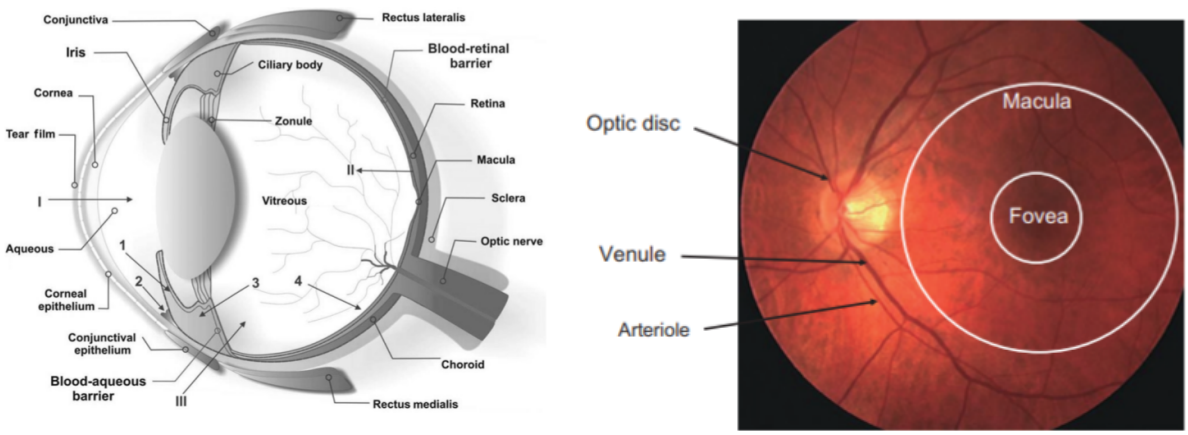
1.3 Structure of the thesis

The thesis is organized as follows. The chapter 2 illustrates a structure of the eye, imaging of ocular fundus, and spectral retinal image analysis. Chapter 3 reflects a concept of convolutional neural network, gradient descent, and activation function. The chapter 4 introduces a review of image registration methods. The chapter 5 describes a deep neural network for image registration. The chapter 6 shows the system evaluation and experimental results. The chapter 7 indicates a considerations about current and future studies. Conclusions are given in chapter 8.

2 OCULAR FUNDUS IMAGING

2.1 Structure of eye

The anatomy of the eye is a complex structure presented in Fig. 2a and consists of three main layers: outer layer, middle layer, inner layer [15]. Optical properties are determined by how the tissue of an organ interacts with light. The eye structure does not allow fixing all the structures in one shot and different backlight settings are used for this. The outer layer consists of the cornea and the sclera. The sclera forms a shell that protects the eye from internal and external influences and maintains its shape [15]. The cornea plays a significant role which consists in the refraction of light and protection of the organ from infections and structural damage, represents the front part of the eye and is located by the pupil. The middle layer is represented by the iris, the ciliary body, and the choroid. The last one provides oxygen to the retina and nourishes it, the ciliary body controls the shape of the lens, and the iris controls the amount of light falling on the retina by the size of the pupil. The fundus structure is presented in Fig. 2.



(a) (b)

Figure 2. The eye: (a) structure; (b) fundus [15].

The inner layer plays a key role in the signal formation and the retina with many types of neurons is responsible for this. It consists of 6 major classes of cells: photoreceptors, bipolar cells, horizontal cells, amacrine cells, and ganglion cells [15]. Human eyes contain two types of photoreceptors converting light into an electrical signal: cones include special pigments with absorption peaks in the green, blue or red parts of the spectrum; rods are not able to recognize colors, they transform light irritation into nervous excitement.

The rods allow seeing in low light without color vision. The density of photosensitive receptors varies in different areas of the retina with an offset to the center. Detailed retinal layers and cells are shown in Fig. 3.

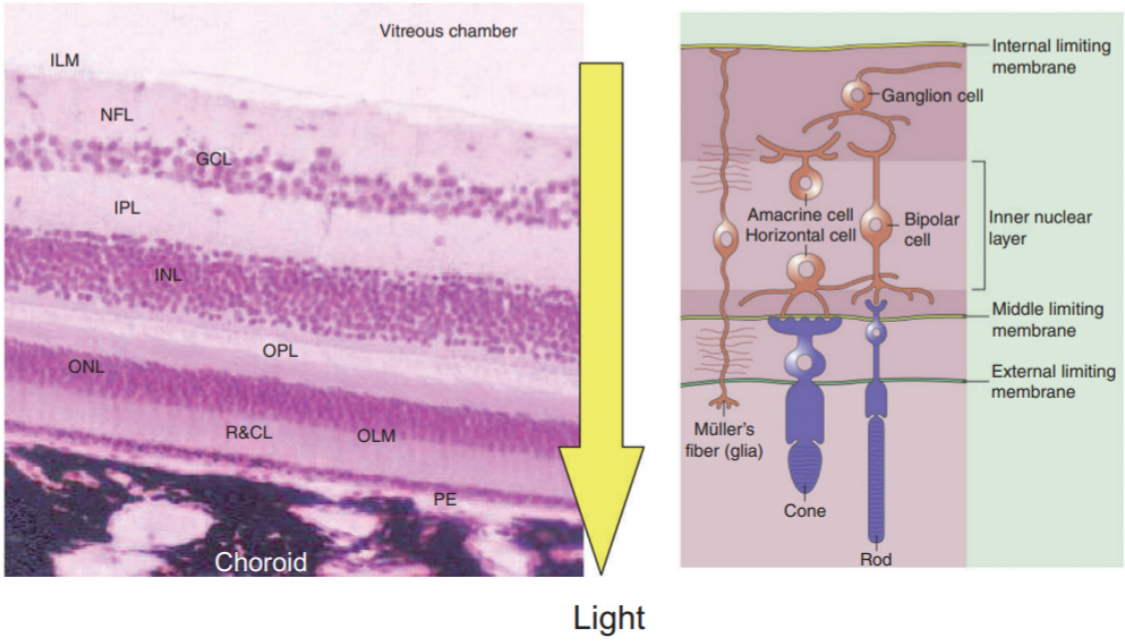
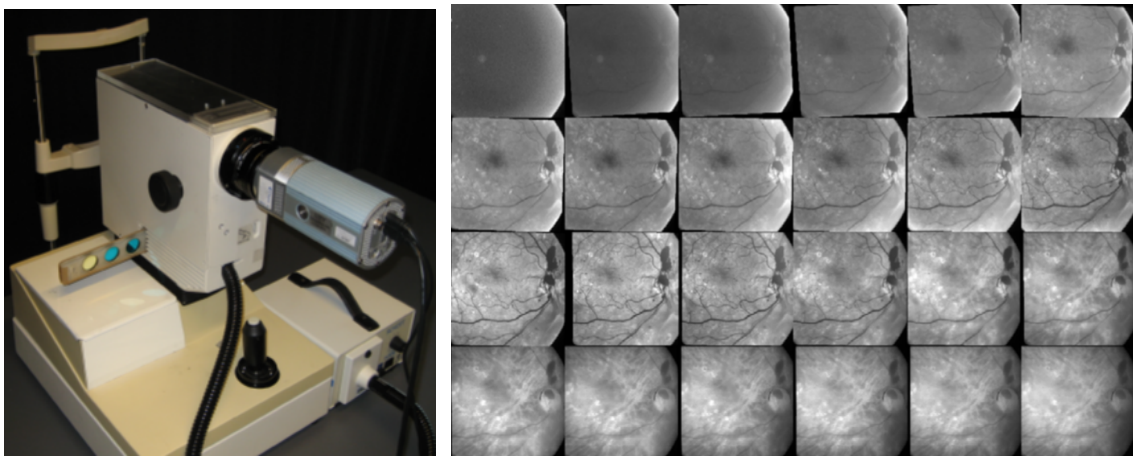


Figure 3. Layers of the retina: inner limiting membrane (ILM), nerve fibre layer (NFL), ganglion cell layer (GCL), inner plexiform layer (IPL), inner nuclear layer (INL), outer plexiform layer (OPL), outer nuclear layer (ONL), outer limiting membrane (OLM), pigment epithelium layer (PE), rods and cones layer (R-CL) [15].

2.2 Imaging of the ocular fundus

Visualization of the inner part of the human eye, called the fundus, is the main tool in the diagnosis and monitoring of a number of common diseases. Various visualization methods are used to study the fundus, however, the most used technology nowadays is MSI. For the system, which is shown in Fig. 4a, Canon CR5-45NM was modified by removing unnecessary components followed by installation a QImaging Retiga 4000RV digital grayscale charge-coupled device (CCD) camera and rails for Edmund Optics narrow bandpass filters with central wavelengths in the range $400\,\mathrm{nm}$ to $700\,\mathrm{nm}$ and $10\,\mathrm{nm}$ step, the light source is a $150\,\mathrm{W}$ OSRAM halogen lamp with a daylight-simulating filter [16]. During imaging, for the best quality, sets of several images are taken for each filter. Received sets are presented in Fig. 4b.



(a) (b)

Figure 4. Installation for obtaining images: (a) 30-channel fundus camera [17]; (b) Montage of channel images obtained from camera system [16].

2.3 Spectral retinal image analysis

It has been recognized that the monochrome or RGB color imaging methods have limitations in the early detection and identification of tissue abnormalities [18]. These methods essentially do not allow distinguishing healthy tissues from abnormal ones due to the limitations of color characteristics [19]. Currently, spectral images of the retina are being used to diagnose tissue perfusion [20, 21], blood oximetry which compares different spectral characteristics of oxygenated (HbO2) vs deoxygenated haemoglobin (Hb) [22], as well as for diabetes detection, however, the spectral images, whose pixels are captured over a selected wavelength interval, allow to evaluate how the tissue interacts with a particular spectrum to detect various diseases at an early stage. The developed spatial models containing neighboring different channels are used for more accurate classification and segmentation tasks [23].

One of the most promising research is the hyperspectral imaging (HSI) [24] which captures information from multiple wavelengths, generating a four-dimensional hyperspectral cube consisting of wavelength bands, orthogonal spatial data, and absorbance/reflectance intensities at each wavelength as dimensions. HSI technology allows more spectral information to be collected compared to common retinal imaging. Many articles [25–27] have shown the potential of this method, but none of the applications have entered clinical practice due to the lack of large enough validation studies [24]. HSI is presented in Fig. 5.

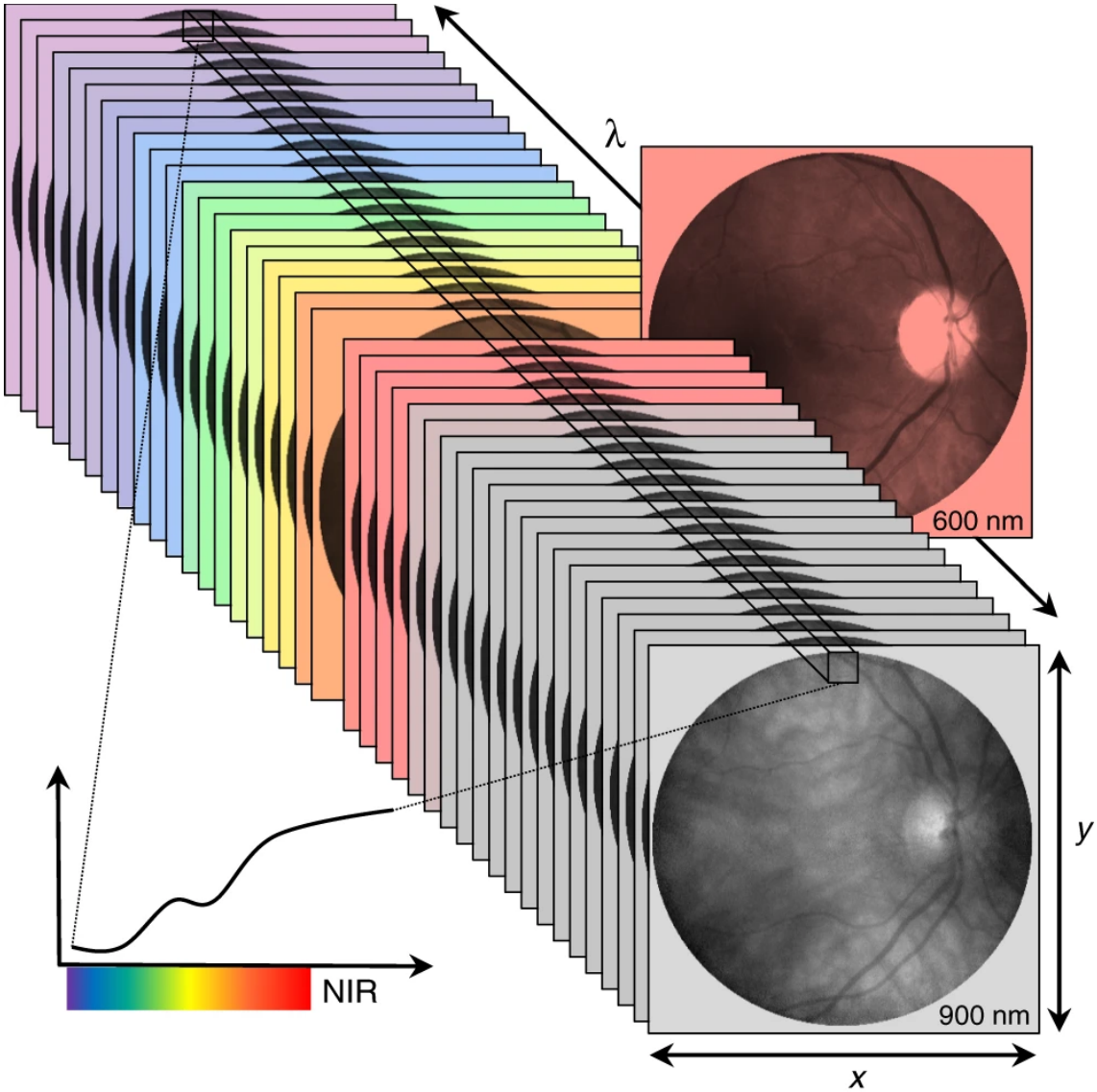


Figure 5. Principle of retinal hyperspectral imaging. A narrow bandwidth tunable light source illuminates the retina and the reflected light from the retina is collected by an image sensor. NIR = near-infra-red [28].

3 CONVOLUTIONAL NEURAL NETWORK

CNN has become dominant in various computer vision tasks. It is widely used in such areas as classification [29], segmentation [30], as well as diabetic retinopathy screening [31]. The CNN architecture includes several building blocks, such as convolution layers, pooling layers, and fully connected layers.

A convolution layer is a main component of CNN that performs feature extraction. A filter or kernel is used to calculate an element-wise product between each element of the kernel and the input tensor to obtain a feature map. Convolution of an image with different filters can perform operations such as edge detection, blur and sharpen by applying filters. After each convolution block, an activation function is used to introduce non-linearity. A pooling layer provides a typical dimensionality reduction to decrease the number of subsequent learnable parameters. The output feature maps of the final pooling layer is transformed into a one-dimensional (1D) array and connected to one or more fully connected layers, also known as dense layers. Fully connected layers together represent a common multilayer perceptron. The final fully connected layer has the same number of output nodes as the number of classes. The architecture of CNN is presented in Fig. 6.

A loss or cost function measures the error between output of the network and given ground truth labels. A type of loss function is one of the hyperparameters and needs to be determined according to the given tasks. A gradient descent algorithm is used to minimize the loss function by updating the parameters in the opposite direction of the gradient of the objective function. This technique is called backpropagation. The gradient descent can be defined as

$$w := w - \alpha \frac{\partial L}{\partial w} \tag{1}$$

where w is a learnable parameter, α is a learning rate, L is a loss function. The learning rate determines the size of the steps we take to reach a (local) minimum. In other words, the gradient descent follows the direction of the slope of the surface created by the loss function to reach the minimum.

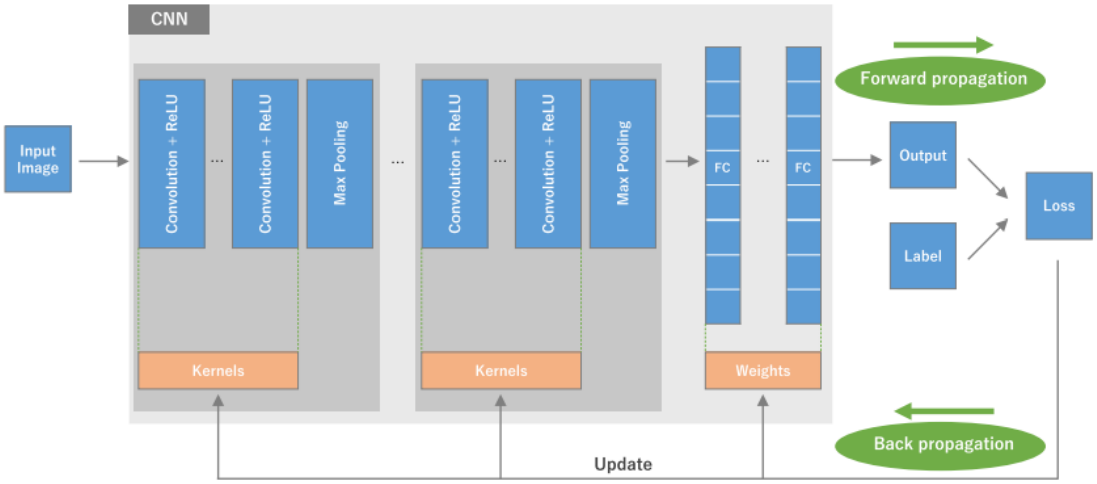


Figure 6. An overview of a convolutional neural network architecture and the training process [32].

Activation function

One of the most paramount part of the neural network is an activation function (AF) which can activate the feature of neurons to solve nonlinear problems. Without the AF the network is just a linear regression model that is not capable to learn and perform more complex tasks [33]. No matter how many layers it has, the network would behave just like a single-layer perceptron. Consider a two layer neural network without AF. A single layer without an activation function can be defined as

$$Ax + b \tag{2}$$

where A is a weight matrix, b is a bias. A second layer can be defined as

$$C(Ax + b) + d \tag{3}$$

where C, d are the corresponding weight matrix and bias. Eq.3 is equivalent to a single layer neural network with weight matrix CA and bias vector $Cb + d$. For clarity, a single layer perceptron, which is shown in Fig. 7, reflects the structure of the neuron within the AF to solve the classification task.

In Fig. 7, the output of the model can be defined as:

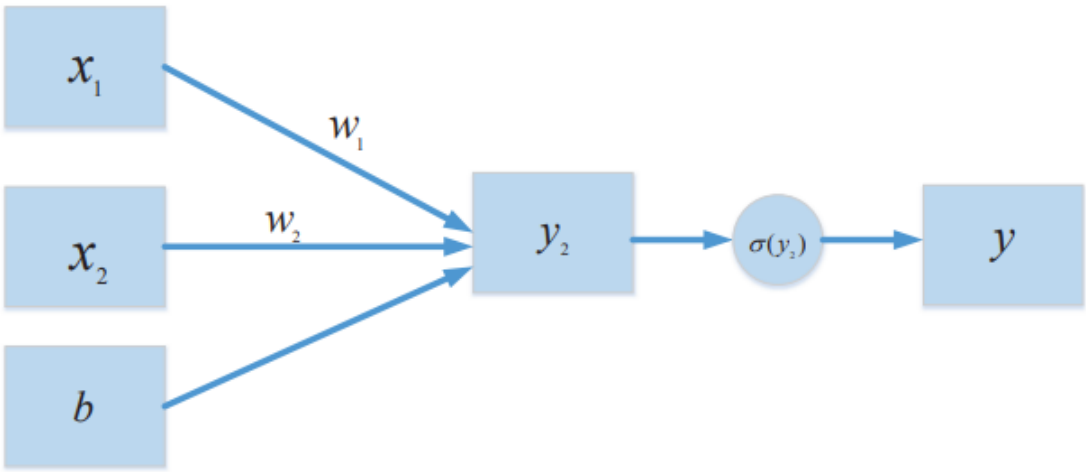


Figure 7. A single layer perceptron with an activation function [34].

$$y_2 = w_1 x_1 + w_2 x_2 + b \tag{4}$$

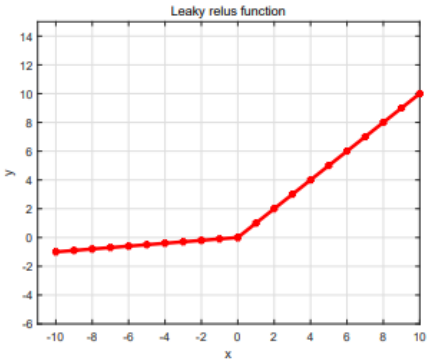
$$y = \sigma(y_2) \tag{5}$$

where $\sigma(y_2)$ is the AF which solves non-linear problems. Thus, several types of AF will be considered in the following sections to choose the most suitable option for deep image registration.

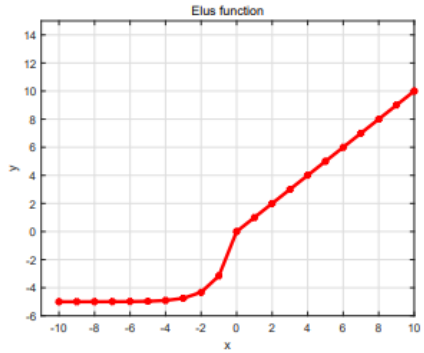
The dying rectified linear unit (ReLU) problem refers to the scenario when many ReLU neurons only output values of 0. The worst-case scenario is when the entire network dies, meaning that it becomes just a constant function. Thus, to solve this problem, the negative part of the ReLU has been changed. Several articles [35–37] have proposed some improved activation functions based on the ReLU such as LeakyReLU, ELU, Tanh-ReLU, Soft-ReLU, and others. The curves of these variations can be seen in Fig. 8. In fact, there are more than twenty variations, but in this work only several ones will be considered.

The equations of the ReLU variations mentioned in Fig. 8 are accordingly defined as:

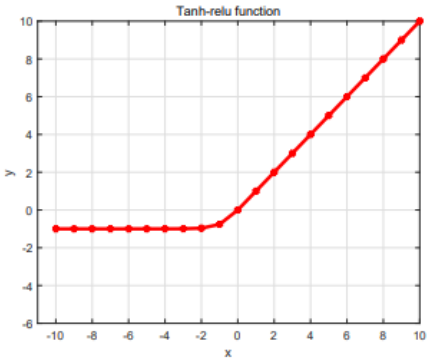
$$f_{\mathrm{LReLU}}(x) = \begin{cases} x & \text{if } (x > 0) \\ a \cdot x & \text{otherwise} \end{cases} \tag{6}$$



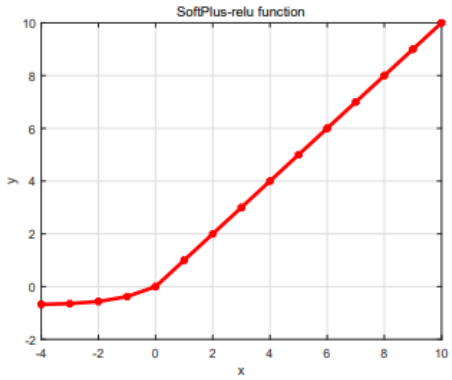
(a) The curve of leaky ReLus function



(b) The curve of Elus function



(c) The curve of Tanh-ReLu function



(d) The curve of softplus–ReLu function

Figure 8. Variants of the ReLU AF [34].

$$f_{\mathrm{ELU}}(x) = \begin{cases} x & \text{if } (x \geq 0) \\ a \cdot (e^x - 1) & \text{otherwise} \end{cases} \tag{7}$$

$$f_{\mathrm{Tanh-ReLU}}(x) = \begin{cases} \frac{1-e^{-2x}}{1+e^{-2x}} & \text{if } (x < 0) \\ \max(0, x) & \text{otherwise} \end{cases} \tag{8}$$

$$f_{\mathrm{Soft-ReLU}}(x) = \begin{cases} \ln(1 + e^x) - \ln(2) & \text{if } (x < 0) \\ \max(0, x) & \text{otherwise} \end{cases} \tag{9}$$

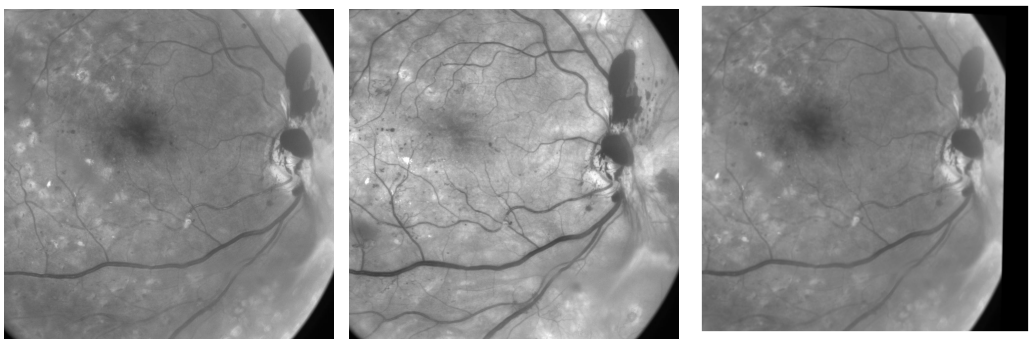
LeakyRelu AF, which is presented in Fig. 8, is solving "dying" of ReLU, but it still computes quickly. The LeakyRelu AF was chosen as the activation function for the deep registration method.

4 IMAGE REGISTRATION

Image registration, which is widely used in medical imaging, is the process of overlaying two or more images of the same scene taken at different times, from different viewpoints, with different sensors [38]. The majority of the registration methods consist of the following steps:

- Feature detection. Distinctive objects (regions, edges, contours, etc.) are automatically detected. Further, these features can be represented as centers of gravity, line endings, distinctive points, which are called control points (CPs).
- Feature matching. Spatial relationships and similarity measures are used to find correspondence between the features detected from different images.
- Transform model estimation. At this stage, the parameters of the functions aligning the two images are evaluated.
- Image resampling and transformation. With appropriate interpolation technique the new image is calculated.

Registration process is presented in Fig. 9.



(a) (b) (c)

Figure 9. Registration process: (a) Channel $500\,\mathrm{nm}$; (b) Channel $568\,\mathrm{nm}$; (c) Registered image.

Several methods were chosen for the comparative analysis of medical images. The purpose of this section is to compare the methods of deep learning and alternative methods, compare the advantages and disadvantages and evaluate the modernity of the method. Scientific topics present a wide range of different approaches to the registration of medical images [16].

4.1 Deformable registration by similarity

In local similarity based approach the image is deformed in such a way to maximize a set of local similarities. The maximum likelihood (ML) estimation combines most similarity measures which is to maximize the joint probabilities of images [8, 39]. The similarity function is estimated by pixels, therefore the location of the pixels does not matter and means that the pixels are independent and stationary. With the pixel intensity, a measure of proximity is determined which in turn leads to the emergence of two approaches: using complex Bayesian models and image registration with intensity correction [39]. Comparison measure of proximity, the most accurate is the residual complexity technique, where the root mean square error (RMSE) is the lowest [39].

4.2 Registration by demons

James Maxwell introduced the concept of demons to describe a paradox in thermodynamics which is a reduction of entropy of two gasses, where is the membrane with a set of demons which can to distinguish 'a' and 'b' particles. In theory, these particles are divided equally along the edges by diffusion, in contradiction with the second principle of thermodynamics [40]. Further, the concept was transferred to image analysis for the task of matching two images. In this case, we consider the contour of an object as a membrane with demons and model filtering points according to their polarity (inside or outside). In diffusing model each point is a demon and the force of the demon, depending on the position of the disk, can be directed outward or inward which is illustrated in Fig. 10.

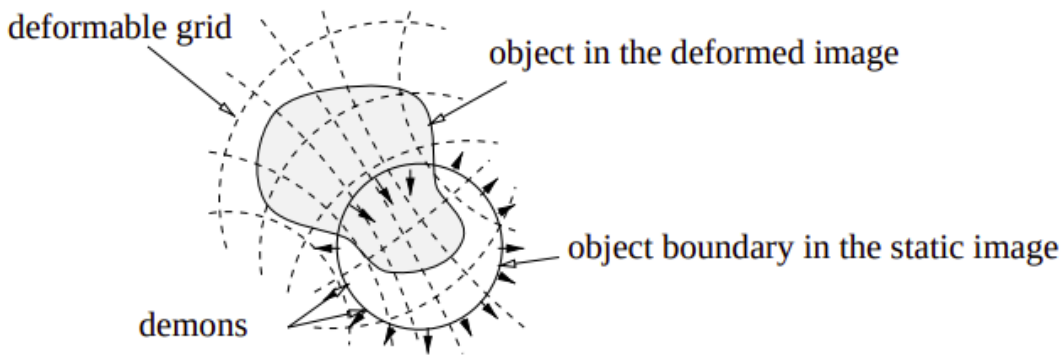


Figure 10. Diffusing models [40].

4.3 Feature-based registration

The generalized dual bootstrap iterative closest point algorithm (GDBICP) was fabricated almost simultaneously by several scientists [41], [42], [43], [44]. Subsequently, it found its application for the registration of medical images, and the article [45] that will be further considered is aimed at retinal image registration. With each iteration, the region increases and stops when the entire image is viewed and R defines the set points to be examined.

The approach shown in Fig. 11 is based on three paramount notions:

- The Bootstrap Region. The region the algorithm is working on is called the bootstrap region. Region R defines the area of the image over which the transformation is considered accurate [45].
- Robust ICP which includes a carefully estimated error scale.
- Bootstrapping the Model. The model evolves to a higher-order model as the domain expands.

At the very beginning, it is necessary to determine two sets of point vectors: $P = (p_i)$ and $Q = (q_j)$ for the first and second images, respectively. The problem is to find transformation parameters θ. Set of correspondences $C \subset P \times Q$ could be found to minimize error distance. Loss function of GDBICP can be defined as

$$E(\theta, \sigma; C) = \sum_{(p_i, q_j) \in C} \rho(d(M(\theta; p_i), q_j)/\sigma) \tag{10}$$

where $M(\theta; p_i)$ is a mapping of p_i, $d(M(\theta; p_i), q_j)$ is a distance metric between the mapped and corresponding vectors, σ is the standard deviation of error distances, ρ is a robust loss function. Being an iterative algorithm and therefore procedures outside the algorithm should provide the starting estimate. The following description is presented in Algorithm 1.

Overall, GDBICP uses the covariance matrix of the estimated transformation parameters. Without model selection reduces accuracy slightly while removing robustness leads to very significant deterioration. The combination of each of the three major components takes the algorithm to compete with other methods. The result of the implementation is described in Fig. 12 which confirms the above conclusions.

Algorithm 1 GDBICP

1. Define the initial bootstrap region R, define point P_t.
2. $t = 1$.
3. While the estimate has not converged
   (a) Using an iterative closest point algorithm to determine the scale estimate $\widehat{\sigma_t}$, the transformation $\widehat{\theta_m t}$, the correspondence set C_t. Obtain covariance matrix $\sum_{m\,t}$.
   (b) Bootstrap the model: Choose the new model M_{m_t+1} and define a new covariance matrix.
   (c) Bootstrap the region: Use the covariance matrix, $\sum_{m\,t}$, and the new model M_{m_t+1} to expand the region based on the covariance propagation error by using the new model and covariance matrix.
   (d) Check for convergence. The algorithm will converge just like normal ICP after the bootstrap region, R_t, stabilizes.
   (e) $t = t + 1$.

4.4 Registration with deep learning

Scientific articles [12–14, 46, 47] describing the application of DL for medical image registration are based on the application of CNN. The architecture discussed in this section is based on U-net [48] which is shown in Fig. 13. It contains two parts: an encoder and a decoder. At the first stage, the network corresponds to a typical CNN architecture and consists of the repeated application of convolution layers, each followed by a ReLU and a max-pooling operation to reduce the dimension, together with the doubling the number of feature channels. In the second stage, each step consists of an upsampling of the feature map followed by an up-convolution layer with simultaneous reduction of the feature channels [48]. It is worth noting that such a network contains twice as many parameters as a common CNN which in turn brings about increasing the complexity of calculations. When compared with a very similar network SegNet [49], the differences are in the transmission of complete features maps in U-net instead of the use of pooling indices in SegNet [48].

From the article [12], it can be concluded that the U-net architecture identifies patterns between multispectral images of the fundus and obtained retinal images.

In article [12] to validate the power of alignment the target registration error (TRE) [50]

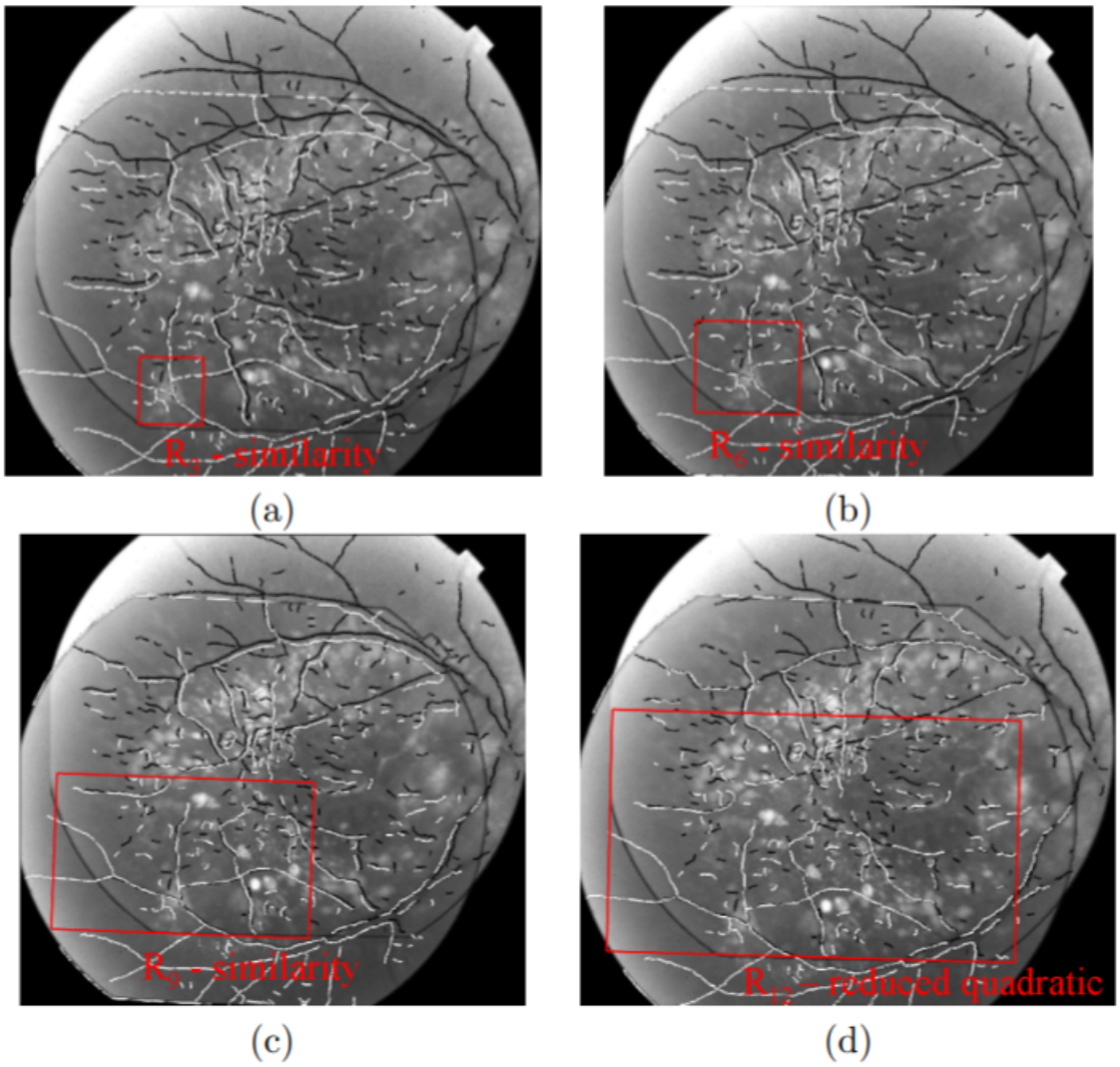


Figure 11. Expansion of the region R in GDBICP for retinal image registration [45].

can be calculated between the manually marked points and their matched points. TRE can be defined as

$$TRE = \frac{1}{m(m-1)} \sum_{a=1}^{m} \sum_{b=1}^{m} |I_a - I_{ba}|, \; a \neq b, \tag{11}$$

where the number of images in a group represents by m and $|I_a - I_{ba}|$ represents the absolute error between the manually marked point and their matched point.

The average TRE on validation and test sets is 2,33 and 2,97 correspondingly. The proposed approach allows to accurately solve the problem of registering spectral retina images and the result of the algorithm can be seen in Fig. 14.

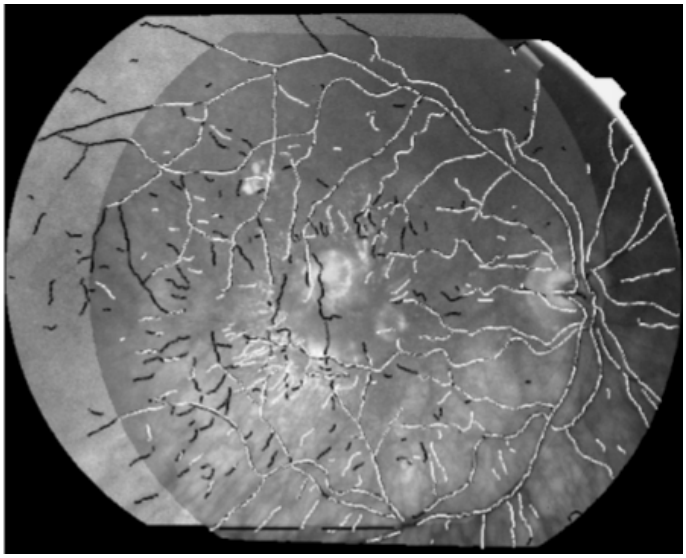


Figure 12. The result of the GDBICP [45].

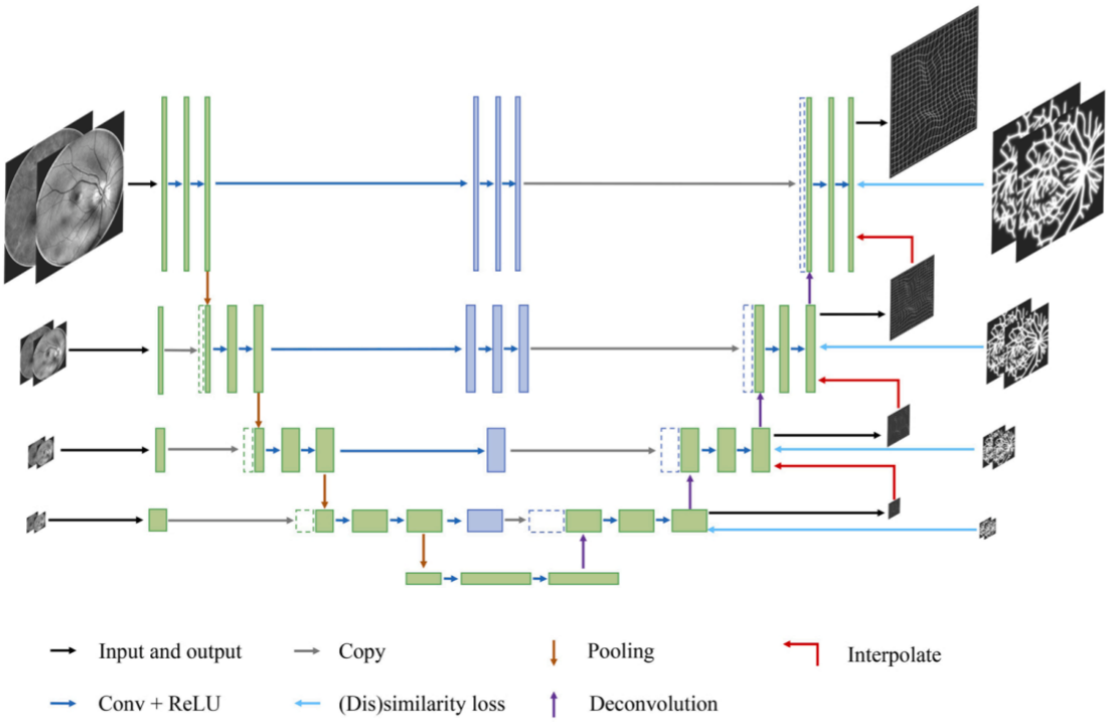


Figure 13. The architecture of weakly supervised registration network [12].

There are several recent papers [51, 52] based on ground truth warp fields for deep image registration, which either rely on classical image registrations on pairs of scans, or by using deformed images. A few papers [53, 54] firstly introduced unsupervised image registration techniques using spatial transformation function to warp an image pair.

The authors demonstrate that a suitable loss function for image similarity leads to high accuracy as in the supervised case.

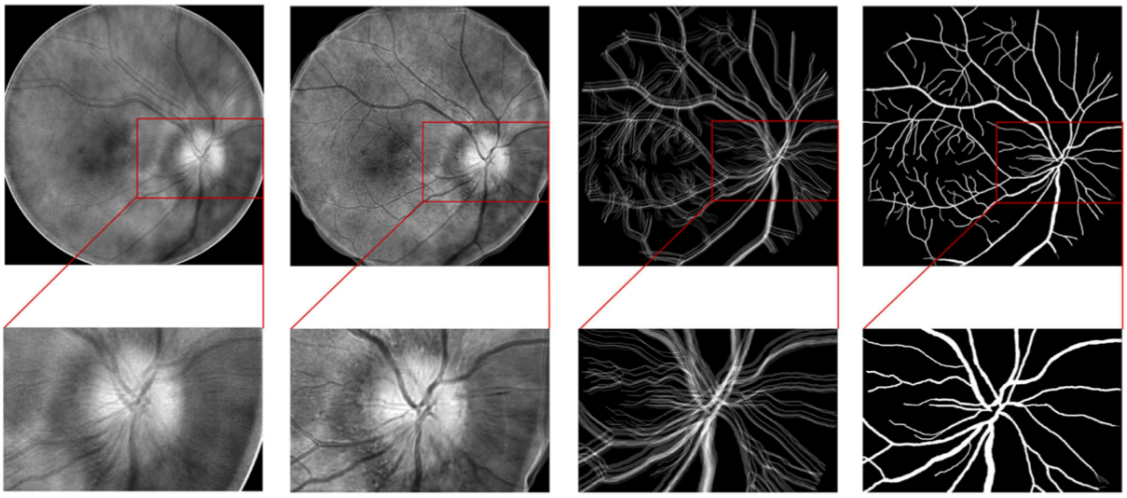


Figure 14. Results of the experiment, from left to right: image obtained with MSI before registration, transformed image, retinal vessels map before registration and retinal vessels after registration [12].

As a result of the analysis of the literature, it was decided to compare two techniques: DL approach, which is rapidly gaining popularity, and the GDBICP, a common choice for retinal image registration. A detailed mathematical description of the deep image registration method is presented in Chapter 5.

5 DEEP NEURAL NETWORK FOR IMAGE REGISTRATION

5.1 Method overview

A weakly-supervised image registration network architecture [12], which is presented in Fig. 15, can be used when there is no intensity based loss for the image pair. In this case the neural network can take a pair of corresponding moving and fixed labels (ground truth), represented by binary masks, to compute a label dissimilarity to register two images. Thus, loss function can be defined as

$$\mathcal{L} = \mathcal{L}_M + \mathcal{L}_R, \tag{12}$$

$$\mathcal{L}_M(L_T, L_S, \phi) = -M(L_T, L_S(\phi)), \tag{13}$$

$$L_R(\phi) = \frac{\alpha}{N}\sum_u ||\bigtriangledown^2 \phi(u)||_2^2 + \frac{\beta}{N}\sum_u ||\phi(u)||_2^2 \tag{14}$$

where $\mathcal{L}_M$ is a (dis)similarity loss which minimizes the difference in appearance [47] between the warped moving label $L_S(\phi)$ and the fixed label L_T. ϕ is the displacement map to be estimated, and L_S, L_T are ground truth labels. $L_R(\phi)$ is a regularization loss which is used to constrain the regularization to obtain a smooth deformation field. α, β are regularization parameters.

Combined registration network [55], presented in Fig. 16, is alternative approach, which combines intensity based, feature based, and deformation based losses together, which in theory should increase the accuracy of image registration.

Following the unsupervised registration method overview, presented in Fig. 17, the algorithm takes two images as input: moving image m and fixed image f to calculate ϕ. Parameters of CNN can be defined as θ, the kernels of the convolutional layers. For similarity evaluation, m should be warped to $m \circ \phi$ using a spatial transformation function. The aim is to find optimal parameters θ by minimizing differences between $m \circ \phi$ and f using fundus images. In this work, unsupervised method proposed two loss functions which will be discussed further (see chapter 5.2)

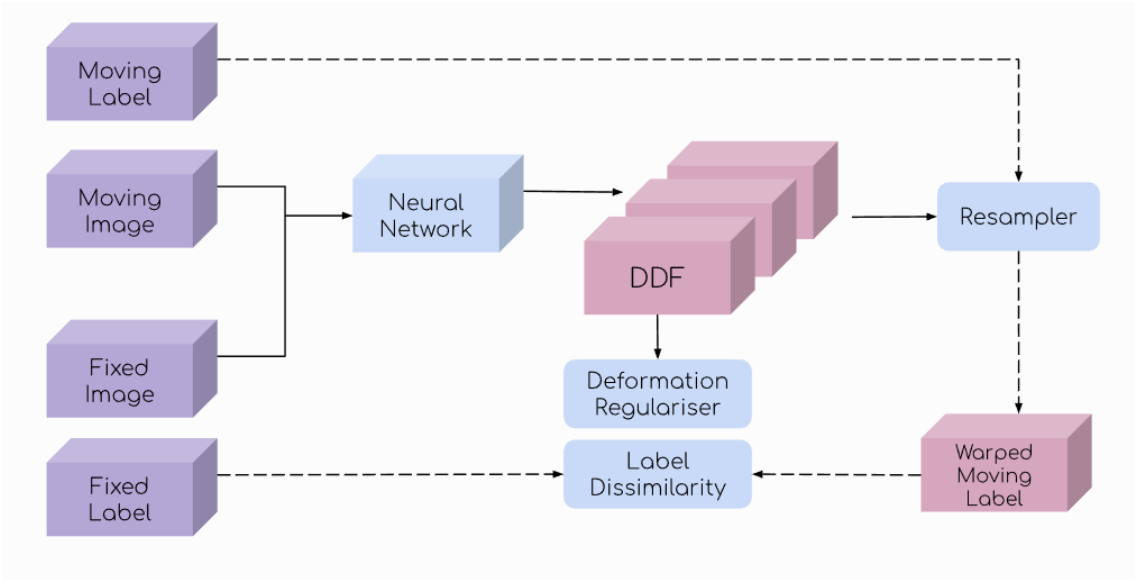


Figure 15. Architecture of weakly-supervised registration network.

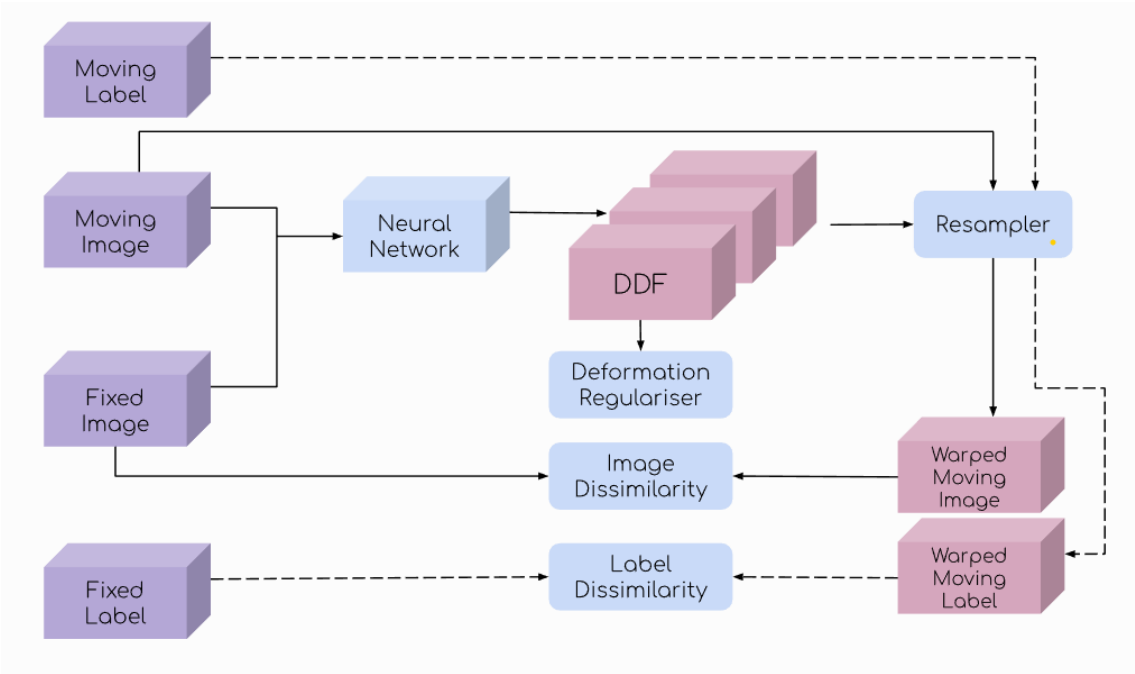


Figure 16. Architecture of combined registration network.

The CNN structure shown in Fig. 18 is similar to U-net [48] which contains an encoder and a decoder. m and f are represented as one input with a concatenation layer. Each convolution is followed by a LeakyReLU layer. The convolutional layers, that capture features, were used to estimate ϕ for each registration pair. MaxPooling layers reduce the spatial dimension in half at each convolutional-pooling layer. At the decoder step, for precise alignment, there are concatenations with layers from the encoder which generates

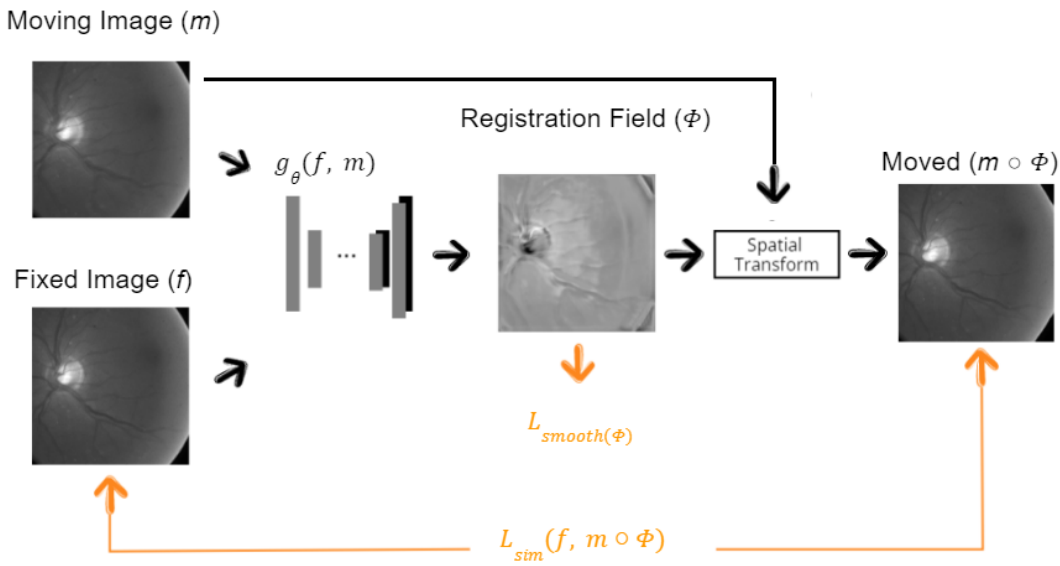


Figure 17. Unsupervised method overview [56].

the registration. Adaptive moment estimation (ADAM) [57] optimizer has been used as a minimization technique that works with momentums of first and second order for rapid convergence.

Spatial transformation is a differentiable module that applies a spatial transformations to a feature map during a single forward pass, where the transformation is conditioned on the particular input, producing a single output feature map [54]. The inclusion of spatial transformers within a CNN allows transforming the feature space to help minimize the overall loss. To make sure that $m \circ \phi$ is close to f we need to warp the input image m. Dense spatial transformations [58] and global parametric image alignment [54] is performed by spatial transformation layer using unsupervised learning. This work relies on an unsupervised approach using a suitable cost function.

5.2 Loss function

Solving the optimization problem of the loss function, allows us to estimate how far the predicted value is from the true value and eventually minimize the error. Depending on the objective loss functions differ and this thesis considers unsupervised techniques. Any of the mentioned losses can be used for registration purposes.

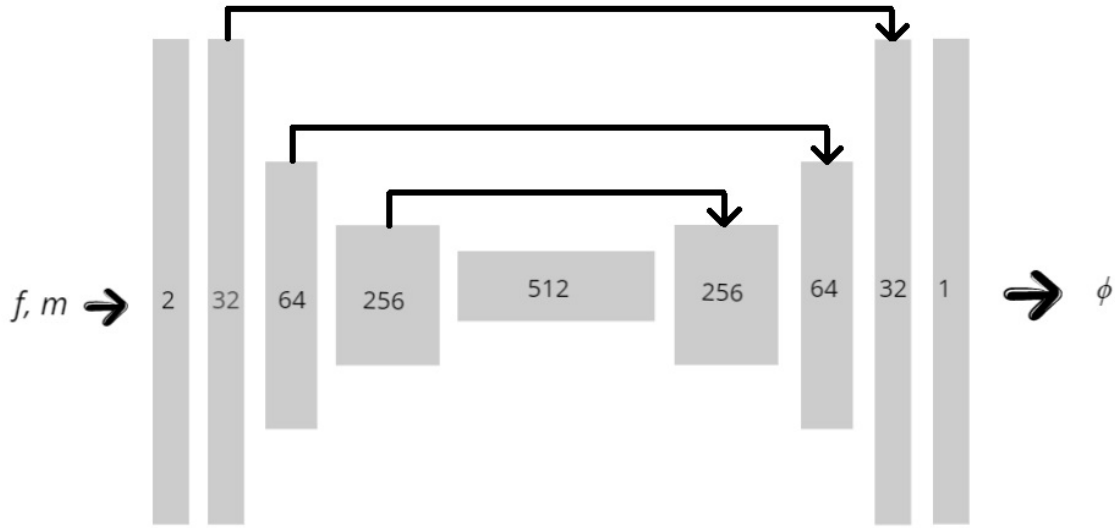


Figure 18. U-Net architecture implementing $g_\theta(f, m)$. Each rectangle represents a 3d volume of layers. The numbers in the architecture represent the dimensions of the convolutional filters. Arrows represent skip connections, which concatenate encoder and decoder features.

Unsupervised loss function

The unsupervised loss function (ULF) [56] L_{us} consists of two components: L_{sim} that penalizes differences in appearance, and L_{smooth} that penalizes local spatial variations in registration field:

$$L_{\text{us}}(f, m, \phi) = L_{\text{sim}}(f, m \circ \phi) + \lambda L_{\text{smooth}}(\phi), \tag{15}$$

$$L_{\text{sim}}(f, m \circ \phi) = \frac{1}{|\Omega|} \sum_{p \in \Omega} [f(p) - [m \circ \phi](p)]^2 \tag{16}$$

where λ is a regularization parameter and Ω is an n-D spatial domain. Minimizing L_{sim} will induce $m \circ \phi$ to approximate *f*, but may generate a non-smooth ϕ that is not physically realistic. We encourage a smooth displacement field ϕ using a diffusion regularizer on the spatial gradients of displacement *u*:

$$L_{\text{smooth}}(\phi) = \sum_{p \in \Omega} || \bigtriangledown u(p)||^2 \tag{17}$$

Mutual information

Mutual information (MI) is the amount of information one image gives about the other. Firstly, it was introduced by [8], has become a common loss function for image registration [59] and it can be defined as the following [60]:

$$I(A, B) = \sum_{a,b} p(a, b) \log \frac{p(a, b)}{p(a)p(b)} \tag{18}$$

where $p(a)$, $p(b)$ are probabilities calculated by constructing a histogram of pixel intensities for each of the two images A and B. $p(a, b)$ is the joint distribution of the intensities of these images. $p(a, b)$ must be the least similar to $p(a)p(b)$ in case of perfect alignment then MI is maximized.

Kullback-Leibler divergence

Relative entropy or Kullback-Leibler (KL) divergence [61] is another similarity measure, one of the fundamental quantities in information theory and its applications [62]. KL is a measure of how one probability distribution P is different from a second, reference probability distribution Q and it can be defined as

$$D_{\mathrm{KL}}(P||Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} = -\sum_{x \in X} P(x) \log \frac{Q(x)}{P(x)} \tag{19}$$

where P, Q are probability distributions. It explains the expectation of the logarithmic difference between the probabilities. It should be mentioned that MI is a kind of KL, thus Eq. 18 is equal to KL [61] between the probability distributions.

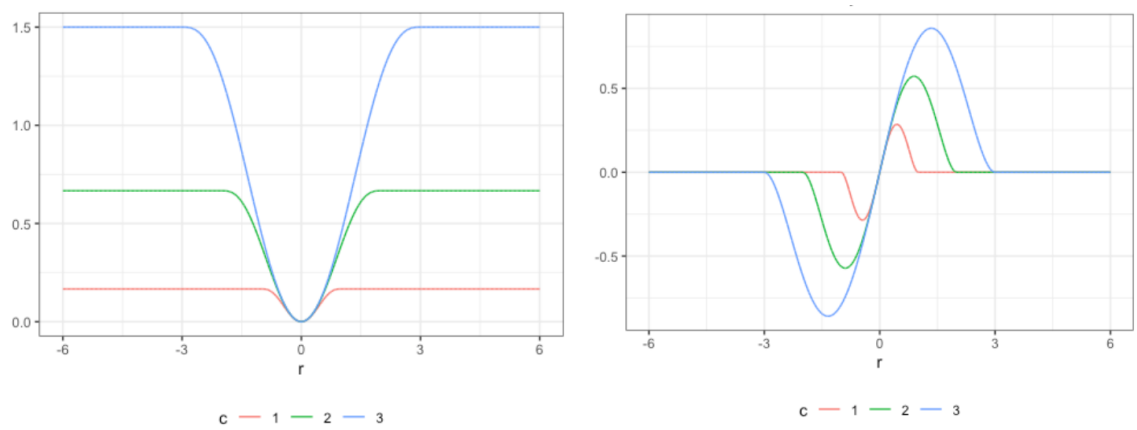
Tukey loss

Another function to be considered is Tukey loss (TL) or Tukey's biweight function which is used in robust statistics [63]. It demonstrates quadratic behavior near the origin which makes it resistant to outliers. It is possible to draw a parallel between Tukey loss and squared error under conditions of the small error until some predetermined limit c is

reached, it flattens out. Therefore large errors do not have an arbitrarily large influence on the result [64]. Tukey loss is shown in Fig. 19.

It can be defined as the following [64]:

$$p(x) := \begin{cases} \frac{c^2}{2}\left(1 - \left(1 - \frac{x^2}{c^2}\right)^3\right) & \text{if } |x| \leq c \\ \frac{c^2}{2} & \text{otherwise} \end{cases} \tag{20}$$



(a) (b)

Figure 19. Tukey function: (a) Tukey loss; (b) Tukey derivative.

Normalized cross-correlation

Numerous techniques have been proposed for image matching [65] such as sum-of-squared difference, closest point, and MI that follow a well-established trend to make the algorithm independent of significant scale and rotation changes. The normalized cross correlation (NCC) cost function [66, 67] being an intensity-based similarity measure was the basis for several image matching methods [68–70] and it has found application in a variety of computer vision and pattern recognition tasks. The aim is to maximize the NCC between two images by taking into account the correlations of intensities in local neighborhoods of the currently processed image pair [71]. The NCC between two images can be defined as

$$NCC(I_1, I_2, x) = \frac{\left(\sum \bar{I}_1(y)\bar{I}_2(y)\right)^2}{\sqrt{\sum \bar{I}_1(y)\bar{I}_1(y)}\sqrt{\sum \bar{I}_2(y)\bar{I}_2(y)}} \tag{21}$$

where x is a point in the target image, $I_{1,2}$ - matching images. Eventually, the NCC similarity cost function which could be used as a loss function in the neural network can be defined as

$$NCC(I_1, I_2) = \int_{\Omega} NCC(I_1, I_2, x)dx. \tag{22}$$

The combined loss function is preferred due to faster convergence and better accuracy [56]. As a result, the unsupervised loss function was chosen for the deep image registration method.

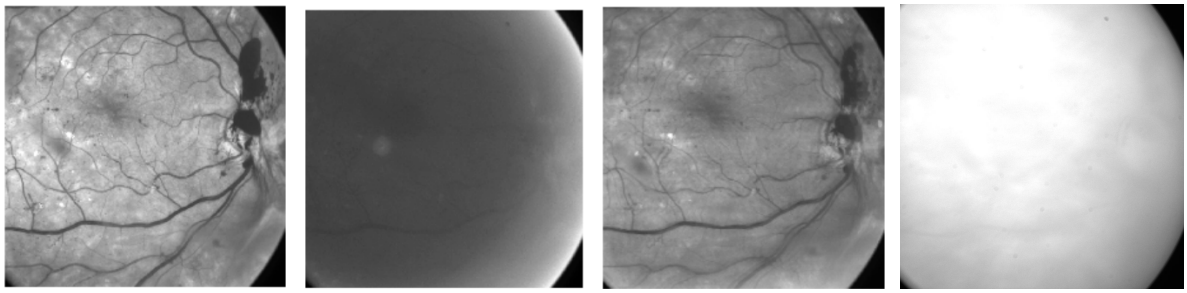
6 EXPERIMENTS

The implementation for the experiments was prepared in Python 3.9.12 with VS code IDE. Tensorflow 2.8 was used for deep learning implementation and the training process was performed on a computing server using an Nvidia Titan RTX with 24 gigabytes of memory. An essential task is to choose the optimal parameters of the neural network and GPU for training since it is not always possible to select model hyperparameters due to the very high demands on hardware. The time involved in this work to train the model was several hours per one epoch for such a relatively small amount of data.

6.1 Data

For the training and evaluation purposes of the neural network, a set of raw channel images was used. The DiaRetDB2 is a public database for benchmarking diabetic retinopathy detection algorithms. The set contains 7191 digital gray images of the eye fundus for 55 patients within 1024x1024 resolution for each image and expert annotated ground truth for several well-known diabetic fundus lesions. Channels vary from $430\,\mathrm{nm}$ to $700\,\mathrm{nm}$ with 5 instances for each channel to be able to choose the best one.

Predictably, the dataset contains many inaccurate instances which are shown in Fig. 20. The histogram of brightness presented in Fig. 21 reflects a necessity to evaluate the brightness of each image at the preprocessing stage since such images contain a minimum of relevant information. Using the standard deviation of brightness, insufficient or excessive brightness can be defined. During the experiments, the best range $[mean - 1.4std; mean + 3std]$ was determined to exclude the worst samples.



(a) (b) (c) (d)

Figure 20. Image samples: (a) Satisfactory; (b) Lack of information; (c) Blurry; (d) Overbright.

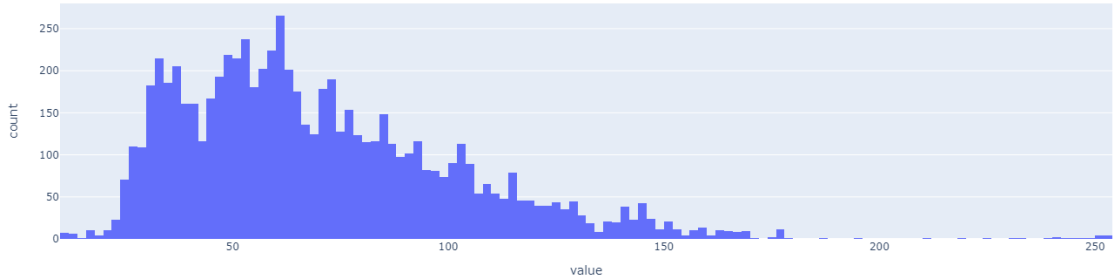


Figure 21. Histogram of brightness of raw images.

6.2 Description of experiments and evaluation

To evaluate the similarity of registered images, functions mentioned in chapter 5.2 can be used. Following the evaluation process, which is presented in Fig. 22, for each patient, for each channel, the best instance is known in advance which is determined by a photographer. For each registration, the best example among all channels is used to further evaluate how accurately the registration process occurs to a certain, pre-selected best instance. The next step is to compare neighboring registered channels using normalized mutual information (NMI) [72]:

$$NMI(A, B) = \frac{H(A) + H(B)}{H(A, B)} \tag{23}$$

To do this, each image is divided into 9 parts of equal size, and a NMI is calculated between sub-images of the same location to obtain a sequence with metric values for subsequent graphical and statistical analysis.

NMI can be substituted by NCC which returns a 2-dimensional array with values scaled to the range [-1,1]. This process is shown Fig. 23. The most paramount step is to evaluate the resulting array which allows us to figure out how similar the selected sub-images are to each other. For this, the Euclidean distance from the center of the array to the index with the largest values was calculated which provides information on the possible misalignment between the two sub-images. Further, the Euclidean distance will be replaced by the number of pixels, which are the same values for images.

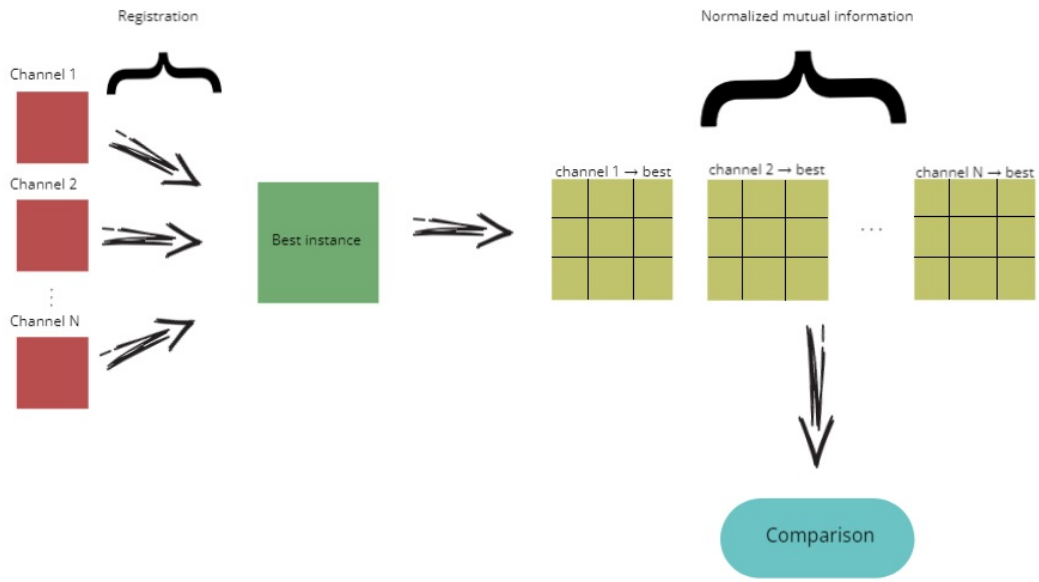
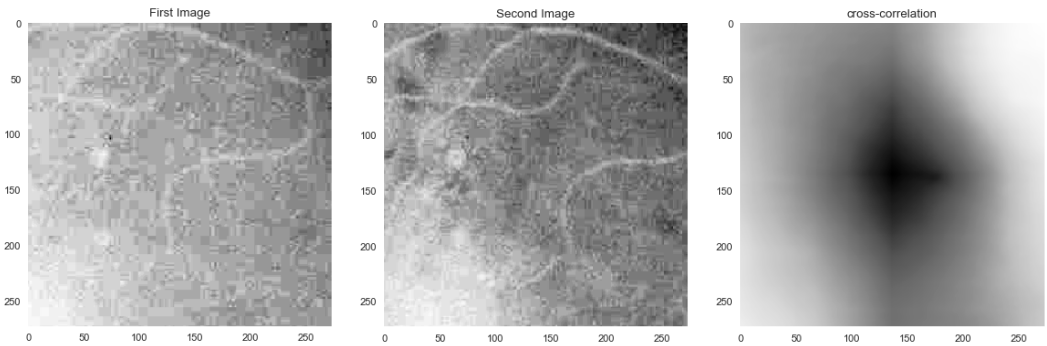


Figure 22. Schematic of the evaluation process.



(a) (b) (c)

Figure 23. Example of evaluation: (a) Sub-image from the first instance; (b) Sub-image from the second instance; (c) Normalized cross-correlation array.

6.3 Results

Unlike neural network training, image registration takes only 3 minutes for 55 patients, what makes the process fast relative to the learning process. Examples of registered channels are shown in figures 24 and 25. Overlaid images, which are presented in Fig. 25, contain many artifacts which indicates a failed registration for channels with low wavelengths, however, Fig. 24 demonstrates small differences between registered channels. These conclusions correlate with the values presented in figures 27 and 28.

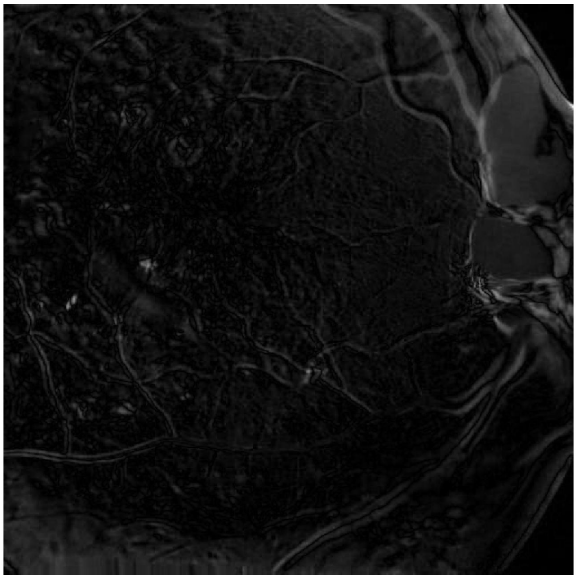


Figure 24. Registered overlaid neighboring channels $500\,\mathrm{nm}$ and $510\,\mathrm{nm}$.

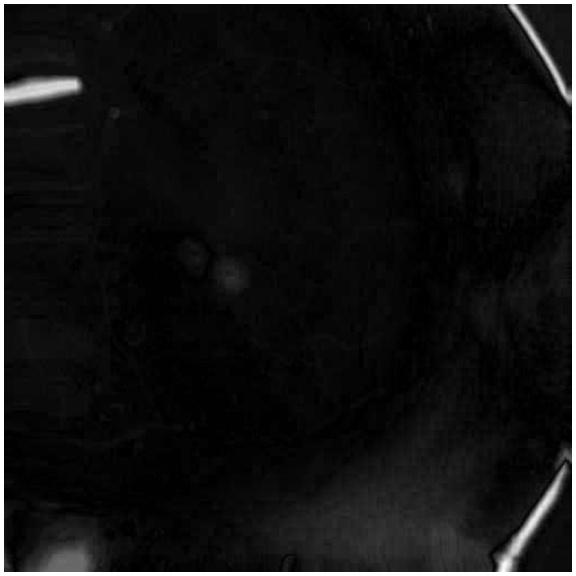
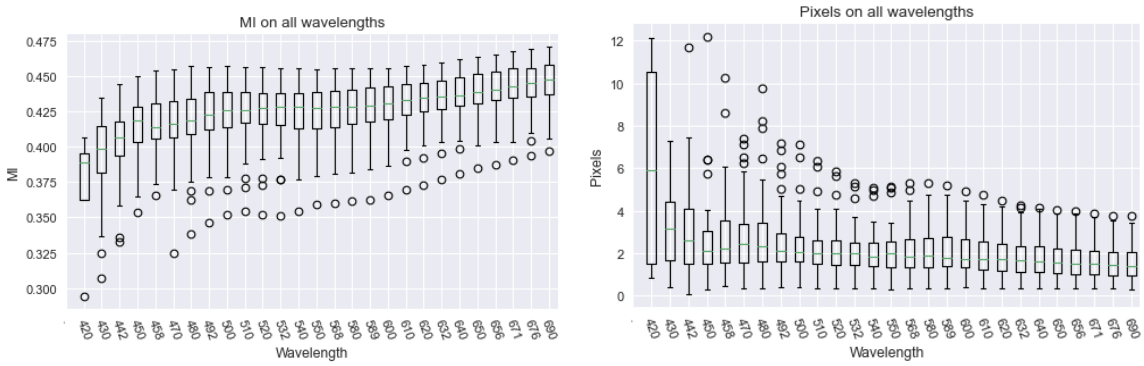


Figure 25. Registered overlaid neighboring channels $442\,\mathrm{nm}$ and $450\,\mathrm{nm}$.

Examples of the obtained NMI and NCC values for one patient are shown in figures 27 and 28, also makes it clear how the metrics vary. The range of captured channel images varied per patient. However, as the wavelengths of the channel images were known during the registration process, it is possible to aggregate the data into one graph, which is shown in Fig. 26. Based on these boxplots, it can be concluded that with increasing wavelength, NMI increases, which is consistent with the the fact that the channels captured at the shorter wavelengths are noisy and more difficult to register.



(a) (b)

Figure 26. Metrics for each channel: (a) Normalized mutual information; (b) Number of pixels.

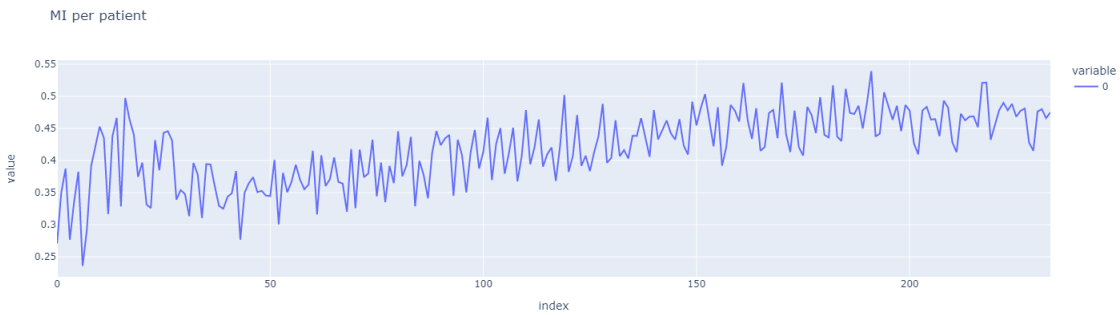


Figure 27. Mutual information per one patient.

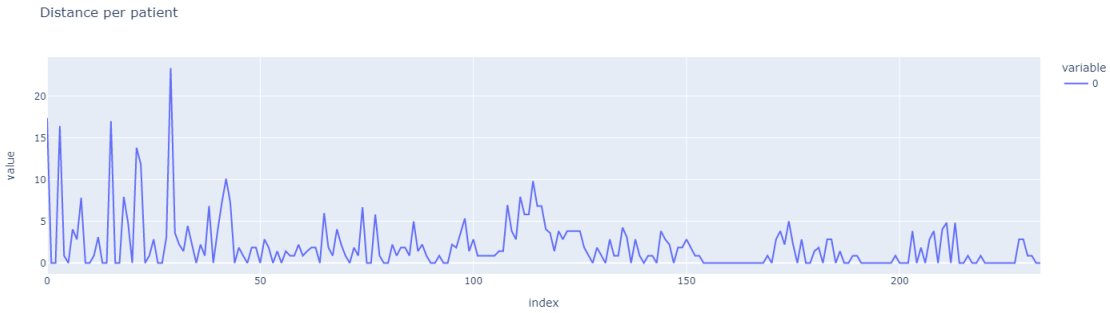


Figure 28. Euclidean distances calculated from the center of the normalized cross-correlation arrays to the elements with the largest value per one patient.

Another way that allows to display the distribution of evaluation metrics (NMI, NCC distances) among all patients is using boxplots which are shown in figures 29 and 31. Each plot contains minimum, first quartile (Q1), median, third quartile (Q3), and maximum values which allows one to look at the data in more detail and draw the appropriate conclusions: registration quality varies from patient to patient. For some patients (№ 7, 13,

29, 51) mutual information values are very high, however, outliers on other patients (№ 3, 17, 24, 35) define registrations as below average. Detailed results for each patient are presented in Appendix 1. Figures 29 and 31 can be used to assess the quality of each patient's registration, either to use the following formula

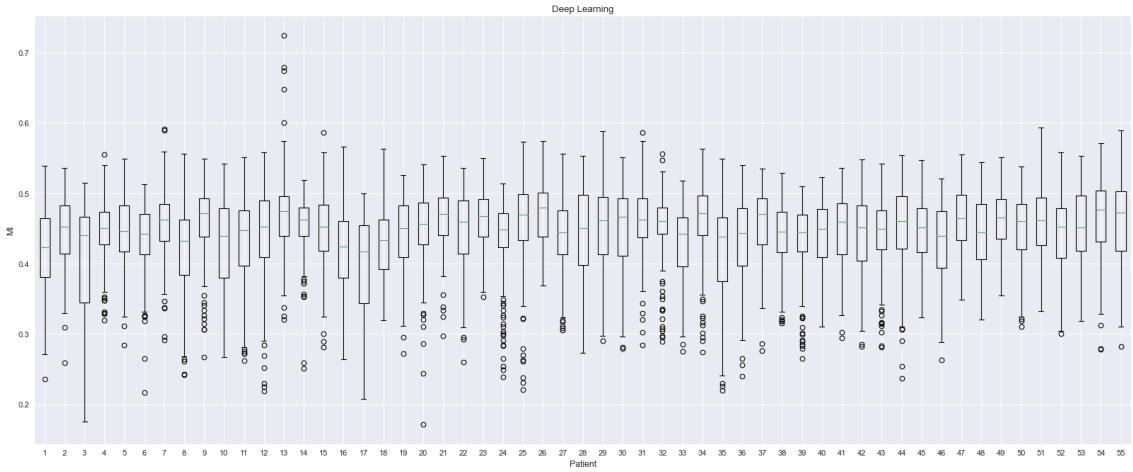


Figure 29. Mutual information among all patients.

where C is a coefficient showing the ratio of the selected metric of a particular patient to the average value of the metric among all patients. The example for mean value, which is presented in Fig. 30, shows how each patient's mean MI differs from the total average MI among all patients. This technique allows to determine how the results of each patient's evaluation differ from the average.

$$C = \frac{X_i}{\overline{X}} \tag{24}$$

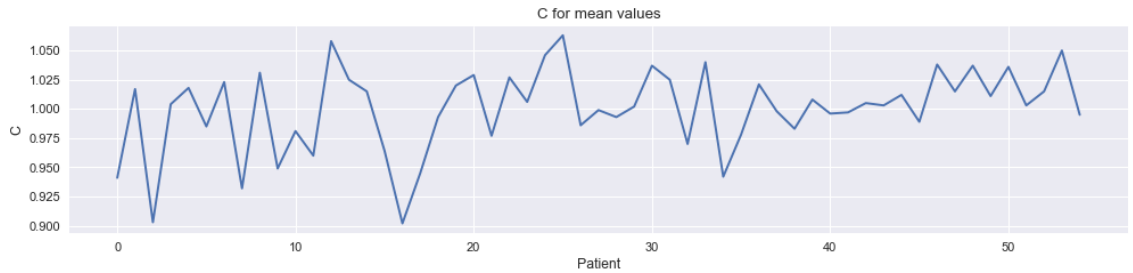


Figure 30. Graph comparing the mean values of each patient's mutual information against the overall mean.

To evaluate the performance of the model, four identical models were used to determine how different the final loss is. And also one model with fewer encoder (decoder) features

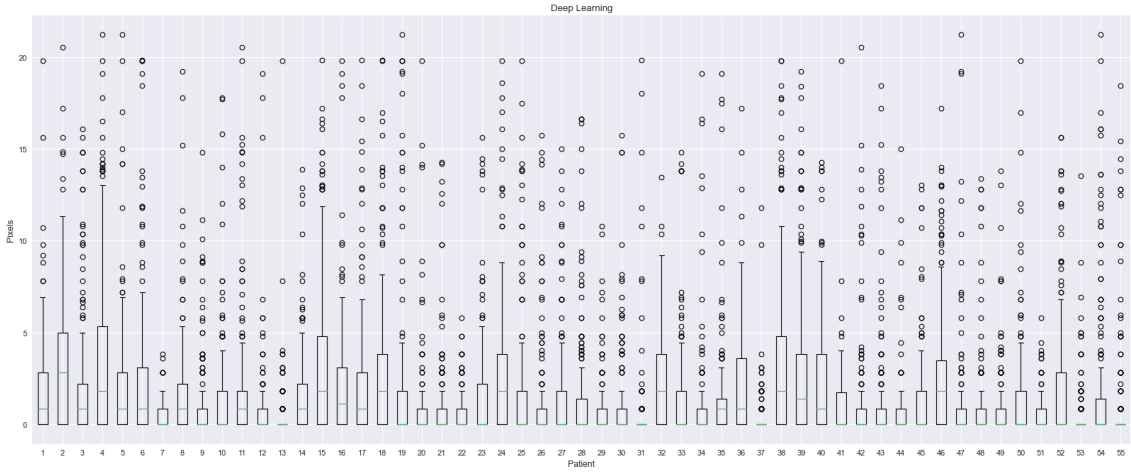


Figure 31. Euclidean distances obtained from the normalized cross-correlation measure among all patients.

as mutable hyperparameters was trained. Getting results, that are listed in Table 2, make it clear that the same parameters of the neural network do not give the same or very close final loss function after 100 epochs. More epochs do not reduce the loss function, which means that the training process converges with 100 epochs. Model № 5 with the fewest parameters showed the highest loss at the end of training. The convergence of the training process is shown in Fig. 32. It follows from the graph that the loss function drops quite quickly during the first 20 epochs, then the fall slows down and remains exactly until the end of training. There is a problem with determinism. Different models showed different accuracy which was confirmed by loss function.

Table 2. Comparison of different models performance.

N_{model}	N_{epoch}	$N_{\text{parameters}}$	final loss
1	100	1,753,730	0.0093
2	100	1,753,730	0.0075
3	100	1,753,730	0.0060
4	100	1,753,730	0.0087
5	100	116,370	0.0113

To evaluate the method performance, a comparison was made of the main statistical indicators with GDBICP which are shown in Table 3. Based on the results of MI, it can be concluded that the developed method of deep image registration reflects a higher mean value (0.445) and a lower standard deviation (0.052) compared to GDBICP metrics. Euclidean distances obtained from NCC confirm the regularity mentioned above.

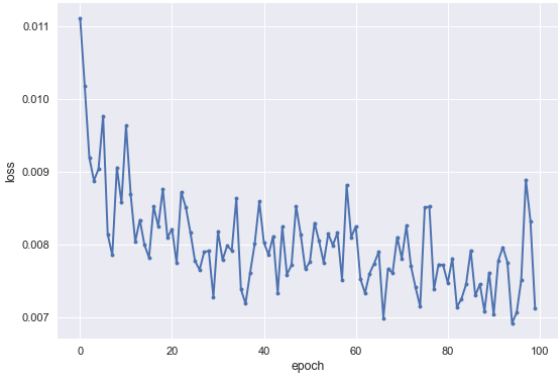


Figure 32. Loss function changes during the training. After 70 epoch the loss function fluctuates at the same level, which indicates the convergence.

Table 3. Algorithms comparison. Normalized cross-correlation is performed in pixels.

	NMI_{DL}	NMI_{GDBICP}	$NMI_{\triangle}$	NCC_{DL}	NCC_{GDBICP}	$NCC_{\triangle}$
min	0.171	0.0	0.171	0.0	0.0	0
max	0.724	0.758	-0.034	24.042	193.747	-169.705
mean	0.445	0.283	0.162	2.096	7.576	-5.48
std	0.052	0.106	-0.054	3.786	28.627	-24.841

7 DISCUSSION

7.1 Current study

Image preprocessing plays a key role in ML systems which improves the quality of raw data. Steps such as image normalization, data cleaning, image to tensor conversion, and data batches preparation are the initial stages of learning-based methods development.

This work indicates that CNN can solve medical image registration problems which are confirmed by the used similarity metrics such as NMI and NCC. NMI comparison with an already existing registration algorithm (GDBICP) showed a superiority in the mean (0.445) against the mean value (0.283) of the existing algorithm as well as a smaller standard deviation. Each registered pair allows to determine a registration success by comparing with all registered pairs among all patients using mean value and standard deviation. The evaluation process revealed a correlation between the registration error and the channels are used. Low wavelengths do not contain sufficient information for successful registration. Aligned images can be used in disease monitoring, guiding treatment, and in other areas where multispectral fundus images are required.

7.2 Future work

In the course of numerous experiments with the model, it was found that tuning the hyperparameters directly affects the convergence rate and the final registration accuracy. Therefore, further work can be aimed at determining the optimal neural network parameters such as weights of the loss function, the number of layers, and the regularization parameter, as well as testing other cost functions that can indicate better quality of the model. Also, the problem of determinism can be considered because the implemented models reflect different results from each other which makes some of them worse than existing methods. The NMI values are very sensitive to any misalignment. Further study could be aimed at understanding the cause of this sensitivity.

Given how much data is available, a combination of various deep image registration techniques suggests itself, such as supervised or weakly-supervised approaches [73,74] which rely on ground truth warp fields. Image registration methods making use of combined losses have shown superior registration accuracy.

Estimation of the performance of learned models is a key issue in the design of efficient ML systems. Working with unsupervised learning we need to formalize the notion that the method may generalize data without model overfitting. Possible extensions include evidence-based metrics [75] to avoid overfitting and to gain generalization. Hence there is a need to increase the number of patients to expand the dataset for fundus image registration. In fact, the developed algorithm can be used in other problems where multispectral images are.

8 CONCLUSION

In this thesis, the problems of architecture selection and activation functions were studied. A deep image registration method with an unsupervised loss function, that produces high-quality spectral retinal images, has been studied. Data preprocessing was made by cleaning and normalizing images. For the registered channels, evaluation metrics were calculated such as NMI and NCC followed by comparison with GDBICP. The evaluation method shows the superiority of retinal registration using the DL approach. The quality of image registration for each patient was determined by comparison with the mean value.

REFERENCES

[1] N. Congdon, B. O'Colmain, C. C. W. Klaver, R. Klein, B. Muñoz, D. S. Friedman, J. Kempen, H. R. Taylor, P. Mitchell, and L. Hyman. Causes and prevalence of visual impairment among adults in the United States. *Archives of Ophthalmology*, 122(4):477–485, 2004.

[2] R. Klein and B. E. K. Klein. The prevalence of age-related eye diseases and visual impairment in aging: current estimates. *Investigative Ophthalmology & Visual Science*, 54(14):ORSF5–ORSF13, 2013.

[3] J. S. McLellan, S. Marcos, and S. A. Burns. Age-related changes in monochromatic wave aberrations of the human eye. *Investigative Ophthalmology & Visual Science*, 42(6):1390–1395, 2001.

[4] P. H. Scanlon. The english national screening programme for sight-threatening diabetic retinopathy. *Journal of Medical Screening*, 15(1):1–4, 2008.

[5] P. J. Watkins. Retinopathy. *British Medical Journal*, 326(7395):924–926, 2003.

[6] J. Choremis and D. R. Chow. Use of telemedicine in screening for diabetic retinopathy. *Canadian Journal of Ophthalmology*, 38(7):575–579, 2003.

[7] S. Vujosevic, S. J. Aldington, P. Silva, C. Hernández, P. Scanlon, T. Peto, and R. Simó. Screening for diabetic retinopathy: new perspectives and challenges. *The Lancet Diabetes & Endocrinology*, 8(4):337–347, 2020.

[8] P. Viola. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.

[9] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[10] Christoph Feichtenhofer, Hannes Fassold, and Peter Schallauer. A perceptual image sharpness metric based on local edge gradient analysis. *IEEE Signal Processing Letters*, 20(4):379–382, 2013.

[11] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. De Vries, M. J.N.L. Benders, and I. Išgum. Automatic segmentation of mr brain images with a convolutional neural network. *IEEE Transactions on Medical Imaging*, 35(5):1252–1261, 2016.

[12] X. Sui, Y. Zheng, Y. Jiang, W. Jiao, and Y. Ding. Deep multispectral image registration network. *Computerized Medical Imaging and Graphics*, 87:101815, 2021.

[13] Y. Wang, J. Zhang, M. Cavichini, D. G. Bartsch, W. R. Freeman, T. Q. Nguyen, and C. An. Robust content-adaptive global registration for multimodal retinal images using weakly supervised deep-learning framework. *IEEE Transactions on Image Processing*, 30:3167–3178, 2021.

[14] J. A. Lee, P. Liu, J. Cheng, and H. Fu. A deep step pattern representation for multimodal retinal image registration. In *IEEE/CVF International Conference on Computer Vision*, pages 5077–5086, 2019.

[15] C. E. Willoughby, D. Ponzin, S. Ferrari, A. Lobo, K. Landau, and Y. Omidi. Anatomy and physiology of the human eye: effects of mucopolysaccharidoses disease on structure and function–a review. *Clinical & Experimental Ophthalmology*, 38:2–11, 2010.

[16] L. Laaksonen. *Spectral retinal image processing and analysis for ophthalmology*. PhD thesis, Lappeenranta University of Technology, 2016.

[17] P. Fält, J. Hiltunen, M. Hauta-Kasari, I. Sorri, V. Kalesnykiene, and H. Uusitalo. Extending diabetic retinopathy imaging from color to spectra. In *Scandinavian Conference on Image Analysis*, pages 149–158. Springer, 2009.

[18] Qingli Li, Xiaofu He, Yiting Wang, Hongying Liu, Dongrong Xu, and Fangmin Guo. Review of spectral imaging technology in biomedical engineering: achievements and challenges. *Journal of Biomedical Optics*, 18(10):100901, 2013.

[19] Costas Balas, Vassilis Papadakis, Nicolas Papadakis, Antonis Papadakis, Eleftheria Vazgiouraki, and George Themelis. A novel hyper-spectral imaging apparatus for the non-destructive analysis of objects of artistic and historic value. *Journal of Cultural Heritage*, 4:330–337, 2003.

[20] Leopoldo C Cancio, Andriy I Batchinsky, James R Mansfield, Svetlana Panasyuk, Katherine Hetz, David Martini, Bryan S Jordan, Brian Tracey, and Jenny E Freeman. Hyperspectral imaging: a new approach to the diagnosis of hemorrhagic shock. *Journal of Trauma and Acute Care Surgery*, 60(5):1087–1095, 2006.

[21] Karel J Zuzak, Michael D Schaeberle, Mark T Gladwin, Richard O Cannon III, and Ira W Levin. Noninvasive determination of spatially resolved and time-resolved tissue perfusion in humans during nitric oxide inhibition and inhalation by use of a visible-reflectance hyperspectral imaging technique. *Circulation*, 104(24):2905–2910, 2001.

[22] DJ Mordant, I Al-Abboud, G Muyo, A Gorman, A Sallam, P Ritchie, AR Harvey, and AI McNaught. Spectral imaging of the retina. *Eye*, 25(3):309–320, 2011.

[23] Artzai Picón, Ovidiu Ghita, Paul F Whelan, and Pedro M Iriondo. Fuzzy spectral and spatial feature integration for classification of nonferrous materials in hyperspectral data. *IEEE Transactions on Industrial Informatics*, 5(4):483–494, 2009.

[24] Sophie Lemmens, Jan Van Eijgen, Karel Van Keer, Julie Jacob, Sinéad Moylett, Lies De Groef, Toon Vancraenendonck, Patrick De Boever, and Ingeborg Stalmans. Hyperspectral imaging and the retina: worth the wave? *Translational Vision Science & Technology*, 9(9):9–9, 2020.

[25] Guolan Lu and Baowei Fei. Medical hyperspectral imaging: a review. *Journal of Biomedical Optics*, 19(1):010901, 2014.

[26] Robert Koprowski, Sławomir Wilczyński, Zygmunt Wróbel, Sławomir Kasperczyk, and Barbara Błońska-Fajfrowska. Automatic method for the dermatological diagnosis of selected hand skin features in hyperspectral imaging. *BioMedical Engineering Online*, 13(1):1–16, 2014.

[27] Siti Salwa Md Noor, Kaleena Michael, Stephen Marshall, and Jinchang Ren. Hyperspectral image enhancement and mixture deep-learning classification of corneal epithelium injuries. *Sensors*, 17(11):2644, 2017.

[28] Xavier Hadoux, Flora Hui, Jeremiah KH Lim, Colin L Masters, Alice Pébay, Sophie Chevalier, Jason Ha, Samantha Loi, Christopher J Fowler, Christopher Rowe, et al. Non-invasive in vivo hyperspectral imaging of the retina for potential biomarker use in alzheimer's disease. *Nature Communications*, 10(1):1–12, 2019.

[29] Koichiro Yasaka, Hiroyuki Akai, Osamu Abe, and Shigeru Kiryu. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced ct: a preliminary study. *Radiology*, 286(3):887–896, 2018.

[30] Patrick Ferdinand Christ, Mohamed Ezzeldin A Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D'Anastasi, et al. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 415–423. Springer, 2016.

[31] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[32] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights Into Imaging*, 9(4):611–629, 2018.

[33] Tomasz Szandała. Review and comparison of commonly used activation functions for deep neural networks. In *Bio-inspired Neurocomputing*, pages 203–224. Springer, 2021.

[34] Yingying Wang, Yibin Li, Yong Song, and Xuewen Rong. The influence of the activation function in a convolution neural network model of facial expression recognition. *Applied Sciences*, 10(5):1897, 2020.

[35] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Citeseer, 2013.

[36] Stamatis Mastromichalakis. Alrelu: A different approach on leaky relu activation function to improve neural networks performance. *ArXiv*, 2020.

[37] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *ArXiv:1710.05941*, 2017.

[38] Barbara Zitova and Jan Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, 2003.

[39] A. Myronenko and X. Song. Intensity-based image registration by minimizing residual complexity. *IEEE Transactions on Medical Imaging*, 29(11):1882–1891, 2010.

[40] J.-P. Thirion. Image matching as a diffusion process: an analogy with Maxwell's demons. *Medical Image Analysis*, 2(3):243–260, 1998.

[41] P. J. Besl and N. D. McKay. Method for registration of 3-d shapes. In *Sensor Fusion IV: Control Paradigms and Data Structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992.

[42] G. Champleboux, S. Lavallee, R. Szeliski, and L. Brunie. From accurate range imaging sensor calibration to accurate model-based 3d object localization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 92, pages 926–927, 1992.

[43] Y. Chen and G. Medioni. Object modelling by registration of multiple range images. *Image and Vision Computing*, 10(3):145–155, 1992.

[44] Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *International Journal of Computer Vision*, 13(2):119–152, 1994.

[45] C. V. Stewart, C.-L. Tsai, and B. Roysam. The dual-bootstrap iterative closest point algorithm with application to retinal image registration. *IEEE Transactions on Medical Imaging*, 22(11):1379–1394, 2003.

[46] Hamid Reza Boveiri, Raouf Khayami, Reza Javidan, and Alireza Mehdizadeh. Medical image registration using deep neural networks: a comprehensive review. *Computers & Electrical Engineering*, 87, 2020.

[47] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9252–9260, 2018.

[48] 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.

[49] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[50] Jihye Kim, Jeongjin Lee, Jin Wook Chung, and Yeong-Gil Shin. Locally adaptive 2d–3d registration using vascular structure model for liver catheterization. *Computers in Biology and Medicine*, 70:119–130, 2016.

[51] Xiaohuan Cao, Jianhua Yang, Jun Zhang, Dong Nie, Minjeong Kim, Qian Wang, and Dinggang Shen. Deformable image registration based on similarity-steered cnn regression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 300–308. Springer, 2017.

[52] Julian Krebs, Tommaso Mansi, Hervé Delingette, Li Zhang, Florin C Ghesu, Shun Miao, Andreas K Maier, Nicholas Ayache, Rui Liao, and Ali Kamen. Robust non-rigid registration through agent-based action learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 344–352. Springer, 2017.

[53] Bob D de Vos, Floris F Berendsen, Max A Viergever, Marius Staring, and Ivana Išgum. End-to-end unsupervised deformable image registration with a convolutional neural network. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 204–212. Springer, 2017.

[54] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[55] Yunguan Fu, Nina Montaña Brown, Shaheer U. Saeed, Adrià Casamitjana, Zachary M. C. Baum, Rémi Delaunay, Qianye Yang, Alexander Grimwood, Zhe Min, Stefano B. Blumberg, Juan Eugenio Iglesias, Dean C. Barratt, Ester Bonmati, Daniel C. Alexander, Matthew J. Clarkson, Tom Vercauteren, and Yipeng Hu. Deepreg: a deep learning toolkit for medical image registration. *Journal of Open Source Software*, 5(55):2705, 2020.

[56] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019.

[57] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ArXiv:1412.6980*, 2014.

[58] Jason J Yu, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *European Conference on Computer Vision*, pages 3–10. Springer, 2016.

[59] Courtney K Guo. *Multi-modal image registration with unsupervised deep learning*. PhD thesis, Massachusetts Institute of Technology, 2019.

[60] Josien PW Pluim, JB Antoine Maintz, and Max A Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, 22(8):986–1004, 2003.

[61] Igor Vajda. *Theory of statistical inference and information*, volume 11. Springer, 1989.

[62] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[63] Charles V Stewart. Robust parameter estimation in computer vision. *SIAM Review*, 41(3):513–537, 1999.

[64] Martin Reuter, H Diana Rosas, and Bruce Fischl. Highly accurate inverse consistent registration: a robust approach. *Neuroimage*, 53(4):1181–1196, 2010.

[65] Christian Heipke. Overview of image matching techniques. In *OEEPE Workshop on the Application of Digital Photogrammetric Workstations*, 1996.

[66] Colin Studholme, Derek LG Hill, and David J Hawkes. Automated three-dimensional registration of magnetic resonance and positron emission tomography brain images by multiresolution optimization of voxel similarity measures. *Medical Physics*, 24(1):25–35, 1997.

[67] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, 2008.

[68] Zhengyou Zhang, Rachid Deriche, Olivier Faugeras, and Quang-Tuan Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78(1-2):87–119, 1995.

[69] Wolfgang Förstner. A feature based correspondence algorithm for image matching. *ISPRS ComIII*, pages 150–166, 1986.

[70] Maurizio Pilu. A direct method for stereo correspondence based on singular value decomposition. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 261–266. IEEE, 1997.

[71] Joo Hyun Song. *Methods for evaluating image registration*. The University of Iowa, 2017.

[72] Colin Studholme, Derek LG Hill, and David J Hawkes. An overlap invariant entropy measure of 3d medical image alignment. *Pattern Recognition*, 32(1):71–86, 1999.

[73] Hessam Sokooti, Bob de Vos, Floris Berendsen, Boudewijn PF Lelieveldt, Ivana Išgum, and Marius Staring. Nonrigid image registration using multi-scale 3d convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 232–239. Springer, 2017.

[74] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. Quicksilver: Fast predictive image registration–a deep learning approach. *NeuroImage*, 158:378–396, 2017.

[75] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020.

Appendix 1. Detailed evaluation results

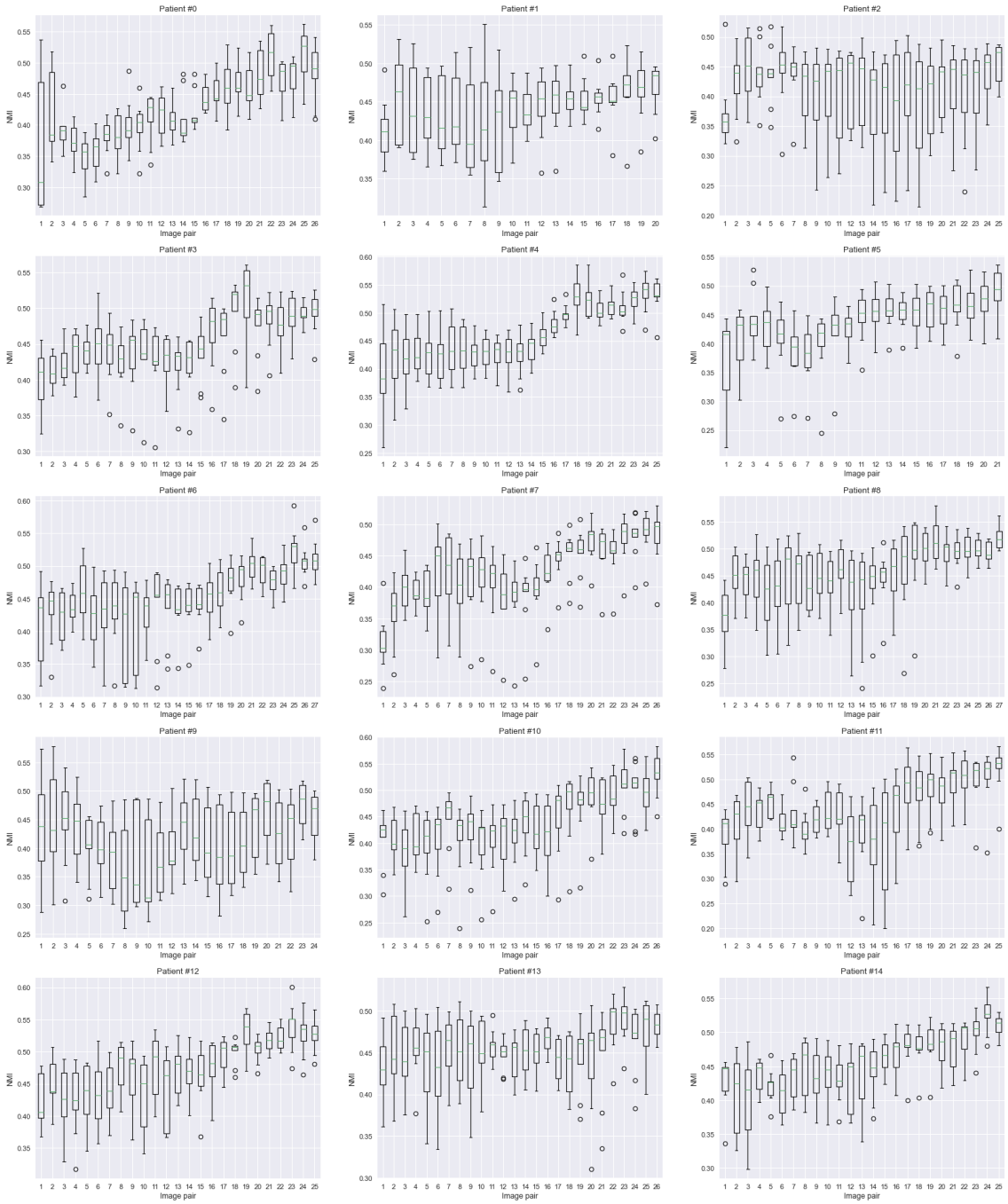


Figure A1.1. Normalized mutual information boxplots for patients № 0-14.

Appendix 1. (continued)

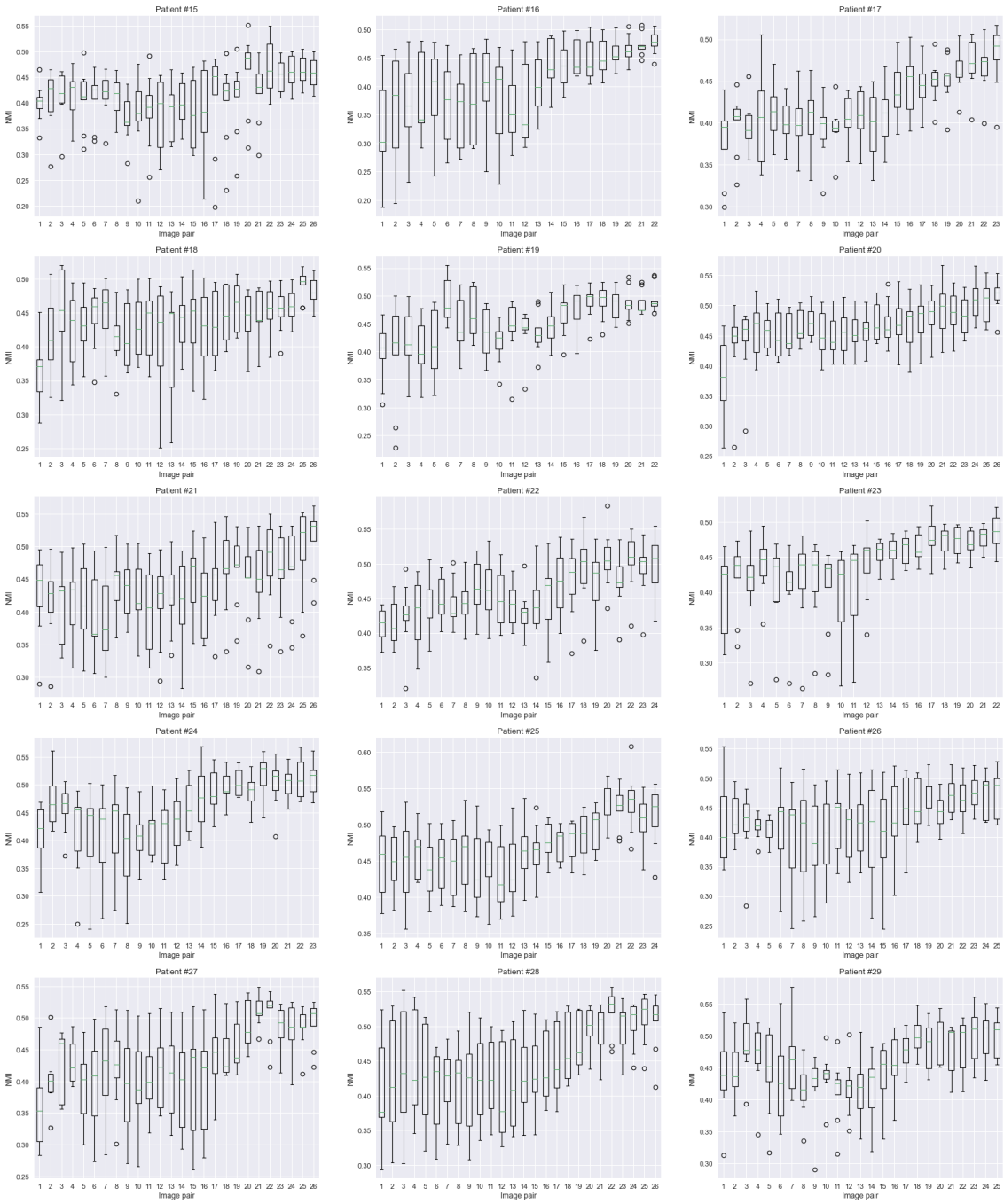


Figure A1.2. Normalized mutual information boxplots for patients № 15-29.

Appendix 1. (continued)

Figure A1.3. Normalized mutual information boxplots for patients № 30-44.

Appendix 1. (continued)

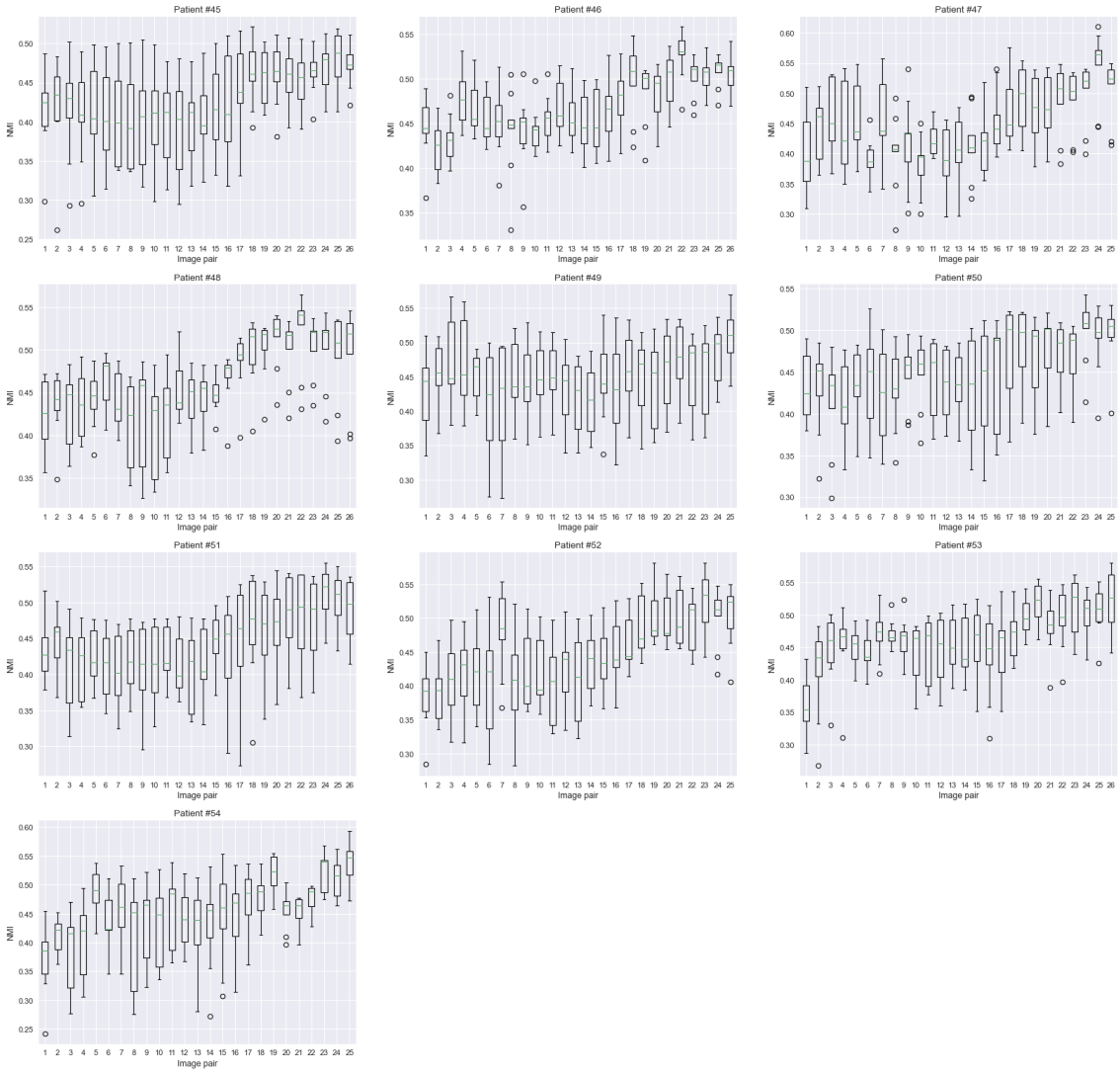


Figure A1.4. Normalized mutual information boxplots for patients № 45-54.