Framework for Deep Learning Based Multi-Modality Image Registration of Snapshot and Pathology Images

Ryan A. L. Schoop, Lotte M. de Roode, Lisanne L. de Boer, and Behdad Dashtbozorg

***Abstract*—Multi-modality image registration is an important task in medical imaging because it allows for information from different domains to be correlated. Histopathology plays a crucial role in oncologic surgery as it is the gold standard for investigating tissue composition from surgically excised specimens. Research studies are increasingly focused on registering medical imaging modalities such as white light cameras, magnetic resonance imaging, computed tomography, and ultrasound to pathology images. The main challenge in registration tasks involving pathology images comes from addressing the considerable amount of deformation present. This work provides a framework for deep learning-based multi-modality registration of microscopic pathology images to another imaging modality. The proposed framework is validated on the registration of prostate ex-vivo white light camera snapshot images to pathology hematoxylin-eosin images of the same specimen. A pipeline is presented detailing data acquisition, protocol considerations, image dissimilarity, training experiments, and validation. A comprehensive analysis is done on the impact of pre-processing, data augmentation, loss functions, and regularization. This analysis is supplemented by clinically motivated evaluation metrics to avoid the pitfalls of only using ubiquitous image comparison metrics. Consequently, a robust training configuration capable of performing the desired registration task is found. Utilizing the proposed approach, we achieved a dice similarity coefficient of 0.96, a mutual information score of 0.54, a target registration error of 2.4 mm, and a regional dice similarity coefficient of 0.70.**

***Index Terms*—Camera snapshot image, deep learning, deformation, image registration, multi modality, pathology image.**

I. Introduction

IN ONCOLOGIC surgery, pathological analysis is the gold standard for investigating surgical specimens and determining to what extent excised tissue contains tumor tissue [1]. However, a pathology image is the result of extensive tissue processing that results in a microscopic image of a tissue slice. This tissue processing includes cutting, fixation in formalin, embedding in paraffin, slicing, sectioning, and staining. As a result of these processing steps, the pathology image of a tissue specimen is highly deformed with respect to any other prior imaging done of the same tissue specimen [2], [3]. This makes image registration of diagnostic, pre-surgical, or intra-operative imaging modalities such as magnetic resonance imaging (MRI), computed tomography (CT), ultrasound, or other experimental technologies to pathology images very challenging.

Classical multi-modal image registration methods can be roughly characterized into two groups: intensity-based methods and feature-based methods [4], [5], [6]. Feature-based methods are based on identifying and matching specific features, such as landmarks, or structures in the images to establish a transformation model. Such an approach, by identifying landmark points, has for instance been done with white-light images and pathology images [7], and also CT/PET imaging and pathology images [8]. However, a large amount of manual point pairs need to be selected for reliable registration. Alternatively, intensity-based models use the intensity values of the pixels in the images with the guidance of a cost function to establish a transformation that aligns the images. This approach usually requires extensive pre-processing or a suitable deformation model for registration with pathology images [9], [10], [11].

In recent years, most methods for image registration have been using deep learning methods based on convolutional neural networks (CNNs) [12], [13]. However, a common problem faced by these learning approaches is the limited amount of medical training data and the lack of ground truth information. The limited amount of training data is often addressed by using data augmentation, whereas the lack of ground truth information

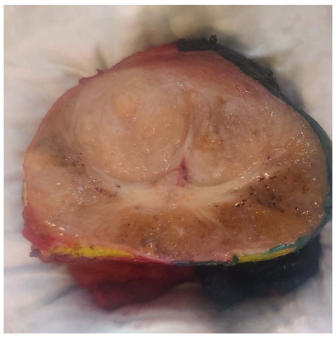
Received 7 May 2024; revised 8 July 2024; accepted 9 August 2024. Date of publication 16 August 2024; date of current version 7 November 2024. This work was supported in part by the Dutch Cancer Society under Grant KWF 13727 and in part by Institutional Grants of the Dutch Cancer Society and of the Dutch Ministry of Health, Welfare and Sport at the Netherlands Cancer Institute. *(Ryan A.L. Schoop and Lotte M. de Roode contributed equally to this work.) (Corresponding authors: Ryan A. L. Schoop; Lotte M. de Roode.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of The Netherlands Cancer Institute/Antoni van Leeuwenhoek (Amsterdam, the Netherlands) under Application No. IRBm 19-124, and performed in line with the Declaration of Helsinki.

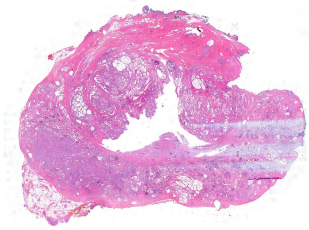
Ryan A. L. Schoop and Lotte M. de Roode are with the Image-Guided Surgery, Department of Surgery, Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands, and also with the Faculty of Science and Technology, University of Twente, 7522 NB Enschede, The Netherlands (e-mail: r.schoop@nki.nl; l.de.roode@nki.nl).

Lisanne L. de Boer and Behdad Dashtbozorg are with the Image-Guided Surgery, Department of Surgery, Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands (e-mail: l.d.boer@nki.nl; b.dasht.bozorg@nki.nl).

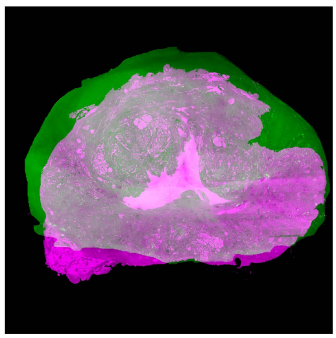
Digital Object Identifier 10.1109/JBHI.2024.3444908



(a) Original snapshot



(b) Original HE



(c) Original snapshot over original HE

Fig. 1. Example of tissue deformation occurring during histopathological processing. (a) A snapshot of a cleaved prostate. (b) The corresponding Hematoxylin and Eosin-stained pathology slide of the tissue is depicted in (a). (c) The original snapshot projected over the original HE to illustrate image dissimilarity.

stimulated the use of weakly-supervised, self-supervised, and unsupervised training schemes [14], [15], [16]. When it comes to image registration with pathology images, most methods are developed using varying strategies in the context of pre-surgical MRI images. For instance, predicting homologous points on the pathology image and the corresponding MR image to infer an image transform [17], or training deep learning models such as SPCNet, U-Net, ResNet, Vgg16, and DeepLabv3+ with patient data [18], [19], [20], [21]. However, the training strategies of these methods can be quite different. The strategy employed in [20] borrows ideas from generative adversarial networks, whereas the method in [21] estimates parameters for an affine and a thin-plate spline transform.

This work aims to describe and tackle the challenges and pitfalls that arise in the registration task of ex-vivo imaging of surgically excised tissue to its associated pathology images using a deep learning method. To investigate this, a prostate dataset was prepared, which consists of macroscopic white-light ex-vivo snapshot images with corresponding microscopic hematoxylin-eosin (HE) stained pathology images. Fig. 1 demonstrates an example of paired snapshot and HE images before registration, thus illustrating the image dissimilarities. In this work, an unsupervised spatial transformer network based on VoxelMorph is used [16], [22]. Although there have been recent advancements in the field of image registration, we chose Voxelmorph because this network architecture, in part, outputs a dense deformation field, and does not require additional registration data other than the images to be registered [23], [24]. This is beneficial due to the lack of consistent paired structures or landmarks on the pathology and snapshot images. In a recent study, our team utilized the unmodified VoxelMorph to correlate optical measurements with accurate pathology labels in breast lumpectomy specimens [25]. Besides this work, to the best of our knowledge, VoxelMorph has not been used for multimodal microscopic/macroscopic registration before. Despite achieving promising results, several limitations were identified, particularly regarding evaluation, which left room for further improvement.

By building on the previous findings, this pipeline first aims to reduce image dissimilarity through several pre-processing steps. Additionally, various training experiments have been conducted to better understand the capabilities of the chosen network architecture and to investigate the effect of a set of key parameters on network performance. These key training experiments involved different modes of data augmentation, loss functions, and regularization factors. No other modifications were made to the neural network architecture. Furthermore, the significance of choosing the right evaluation metrics for such a task when the ground truth is not available has been considered. The effect of pre-processing, as well as the effect of the selected key parameters, was evaluated using both image comparison metrics and clinically established evaluation metrics. Compared to our previous study, this is a comprehensive approach aiming to advance the methodology and generalize it for use in different applications with a similar problem description. Hence, our aim is to provide a framework of considerations and methodological steps that are generally applicable regardless of tissue type or registration context.

Overall, our primary contributions are summarized as follows:

- Collection of a dataset consisting of paired snapshots and HE images of prostatectomy specimens.
- Development of a pipeline for multi-modal image registration utilizing an unsupervised deep learning model, specifically addressing the significant deformations and dissimilarities between microscopic and macroscopic images.
- Introduction of a new loss function and evaluation metrics.
- Comprehensive analysis of the impact of selecting pre-processing steps, loss functions, and evaluation metrics in applications of multi-modal image registration.

II. Materials and Methods

A. Data Acquisition

The data set in this study consisted of pairs of photographed images from cleaved prostates and hematoxylin and eosin (HE) stained histology coupes from those same cleaved prostates Fig. 1.

The study was conducted in accordance with the Declaration of Helsinki and received approval from the Institutional Review Board (IRBm 19-124) of The Netherlands Cancer Institute/Antoni van Leeuwenhoek (Amsterdam, the Netherlands). In compliance with Dutch law (WMO), written informed consent from patients was not required.

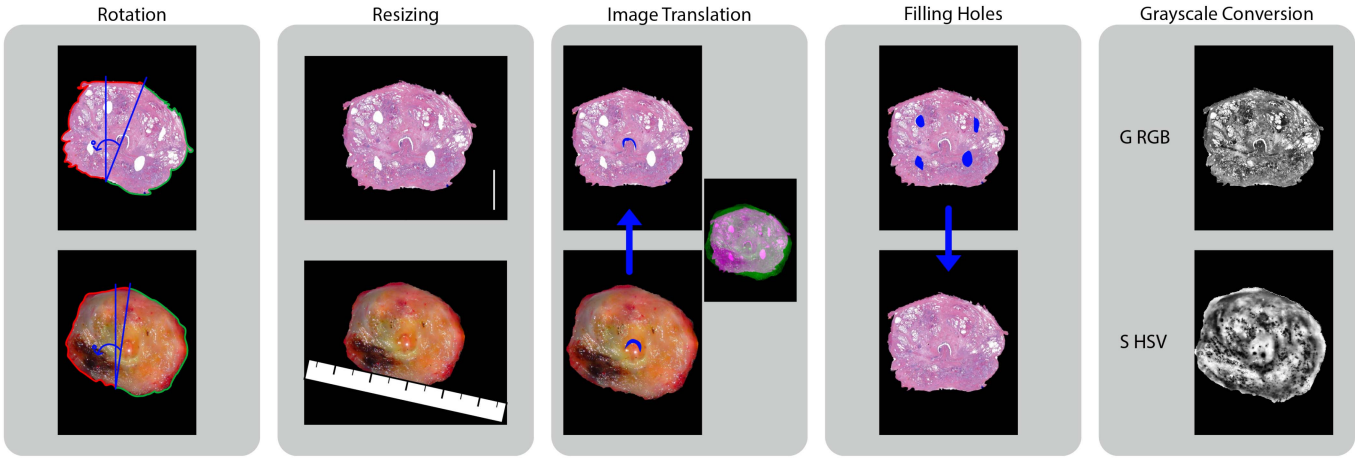


Fig. 2. Overview of data preprocessing steps. 1) Rotation; green and red ink margins are used to rotate and align the snapshot, and HE-image, 2) Resizing; a pixel scale was used to rescale and resize the snapshot, and HE- image, 3) Image translation; the urethra of the snapshot was translated to the urethra of the HE-image, 4) Filling holes; biopsy holes and small defects were filled in the HE-image, 5) Gray conversion; both images were converted to the same color space.

The prostate specimens were collected directly from the operating room after the radical prostatectomy. After collecting the prostate, the prostate capsule was first inked as per standard protocol for histological evaluation (green ink for the left lobe; red ink for the right lobe). Subsequently, the prostates were cleaved in the transversal plane, guided by the preoperative MRI to identify the potential tumor location. The basal half of the cleaved prostate was then placed on a flat surface, with the cutting surface facing upwards, and a regular phone camera (12 MP) was used to capture an image of this cutting surface. This image served as the reference for the multi-modal registration. Additionally, a picture was taken with a ruler positioned adjacent to the prostate at the level of the cleaved surface. This image was used to scale the images correctly during the pre-processing procedures.

Additionally, in five prostates, four 18-gauge intravenous cannulae (BD Venflon) were inserted into the cut prostate surface, one in each quadrant. The inner needles were removed, and the cannulae were trimmed to protrude approximately 2 mm above the cleaved prostate surface. Another picture was captured of the cleaved prostate surface, this time including the protruding cannulae. This image was employed during the "Target Registration Error" evaluation procedure.

The background of both the snapshot and HE images was eliminated using the MATLAB 2023 Image Segmenter App (MathWorks Inc., Natick, Massachusetts). Any areas in the image that did not include the cutting surface of the prostate or contained fatty tissue were considered part of the background. These background areas were removed from the image and filled in with black color.

B. Data Set

The data set consisted of a total of 166 images, which included 83 snapshots of prostate specimens and 83 corresponding HE images. For five out of these 83 image pairs, images containing the protruding intravenous cannulae were available. The prostate slice was too large in fourteen specimens to fit in one coupe, and those specimens were sliced in half. For those cases, the HE image was constructed by digitally stitching the two separated coupes.

C. Data Pre-Processing

The introduced pipeline aims to register macroscopic prostate snapshots to microscopic HE images and thus correct for any changes and deformation the prostate tissue undergoes as part of pathology processing. To ensure the network focuses on correcting deformations due to histopathology processing, five pre-processing steps were conducted to overcome difficulties related to the background, scaling, rotation, present defects, and color space of the images. In Fig. 2 an overview of data pre-processing steps is demonstrated. These pre-processing steps will be elaborated on in this subsection.

1) Rotation: As part of the pathology protocol, the left lobe of the prostate was inked green, and the right lobe was inked red. This inked border was visible in both the snapshots and HE images on the specimen edge. The rotational differences between the snapshots and HE images were corrected by using the border where green and red ink meet. To rotate the snapshot, a vertical line was drawn from the prostate's dorsal (lower) side to the ventral (upper) side, starting at the border of green and red ink. Another line was drawn from the same starting point to the point at the border of the ventral side where the red and green ink meet. The angle between these two lines was used to rotate the snapshot, aligning the second line with the first. This pre-processing step addresses the orientation dissimilarity of the image pair. The same process was repeated for the HE image, ensuring that its orientation matched the snapshot's, as illustrated in Fig. 2.

2) Resizing: The snapshot and the HE image are acquired in different resolutions. To compensate for the scale difference, a

resizing step was applied in which the width of the prostate in the snapshot was measured in millimeters with a referenced ruler, which was included in another snapshot. These measurements were used to calculate the amount of millimeters per pixel on the snapshot. Similarly, the scaling bar in millimeters in the HE image was used to determine the amount of millimeters per pixel on the HE image. By comparing these calculated pixel sizes, the image with the lower amount of millimeters per pixel was down-scaled to match the pixel size of the image with the higher amount of millimeters per pixel. This form of scaling, where the underlying physical distance per pixel is matched, ensures that the relative size of the sample in the HE image and the snapshot image is correct. Despite this, all the images must have the same image dimension for the application of the neural network. Therefore, whilst keeping the underlying physical distance (millimeters) per pixel consistent, the final HE and snapshot images were resized to 512×384 pixels.

3) Image Translation: The snapshot was aligned with the HE images by using the position of the urethra as a reference point. This alignment process ensured that both images were centrally positioned. Additionally it serves as an initial registration step. After this step, it is primarily near the border of the specimen where the image dis-alignment needs to be addressed. The urethra was selected as the reference point because these were easily identifiable and visible in all snapshots and HE images, as can be seen in Fig. 2. The urethra is marked in blue in the figure.

4) Filling Holes: In some cases, the HE images were damaged or contained holes where biopsies had been taken by the histopathology department. These types of distortions can introduce new structures in the HE image with no matching structure in the snapshot image, affecting the registration performance. To furthermore prevent this from affecting the neural network performance on those select few samples, a patch-based inpainting approach was applied to the images. This algorithm fills the holes by computing patch priority and finding the bast-matching path using the sum of squared differences. This was done using the *inpaintExemplar* function in MATLAB 2023 (MathWorks Inc., Natick, Massachusetts).

5) Grayscale Conversion: In order to get the images in a similar color space, both the snapshot and the HE image were transformed into grayscale images. This enables the use of image intensity-based loss functions to facilitate the registration when training the neural network. During the conversion, special attention was given to maintaining consistent intensity level scales for corresponding areas in both images. This ensured that similar regions in the snapshots and HE images appeared with similar shades of gray. To determine the best grayscale conversion, we selected one sample in which we manually registered the snapshot and HE images. Then several grayscale conversions were applied to both the registered snapshot and the HE image of that prostate. For grayscale conversion, we used individual channels in RGB, HSV, and CIELAB color spaces. To compensate for intensity level and structural differences, contrast-limited adaptive histogram equalization (CLAHE) and a 2-D Gaussian smoothing kernel with a standard deviation of 1 were applied on paired images. For each grayscale conversion pair, the mutual information (MI) was calculated. The grayscale conversion pair with the highest MI score was chosen for the final grayscaling in all images in the dataset. For the snapshot images a *HSV* color space was used with enhanced contrast of the saturation channel, and for HE a *RGB* color space was used with enhanced contrast of the green channel. The final result of the image processing can be seen in Fig. 2.

D. Network Architecture

The network architecture used in this work is an unsupervised neural network based on VoxelMorph [16] as shown in Fig. 3. The network takes as input the concatenation of both a moving image (m) and a fixed image (f), resulting in a two-channel 2D image. The moving image is the one which will be deformed to match the fixed image. The first part of this network is a U-Net based convolutional neural network (CNN). This U-Net consists of ten 2D convolutional layers, where all layers have 32 channels per layer except the last convolutional layer which has 16 channels. Each convolutional layer operates with a convolution kernel size of 3×3 with a stride of 1 and zero padding around the edges of the image. Each of the convolutional layers is followed by a Leaky Rectified Linear Unit (ReLU) activation function with a negative slope parameter of 0.2. The encoder part of the U-Net consists of four layers with 2×2 max-pooling, consequently, the input is down-sampled to half its spatial dimensions in each dimension after each layer. The decoder part of the U-Net consists of the remaining six layers with 2×2 up-sampling and skip connections from the same-sized encoder layers.

The main idea is that the output of this U-Net based network is a registration field ϕ, which, once applied to the moving image, yields the moved and registered image. By including the spatial transformer layer in the network architecture, it becomes possible to use the U-Net based CNN output as a registration field. In this registration field, each pixel contains a displacement vector. The spatial transformer layer then applies this output to the moving image by linear interpolation, which allows for differentiability of the spatial transformer layer. Consequently, a registration field is learned by the U-Net based CNN when it's trained end-to-end, and the final output of the complete architecture is a moved and registered image. Note that the spatial transformer layer does not contain any learnable parameters.

E. Training Experiments

Several training experiments were conducted to explore different training options that can alter network performance when using the aforementioned network architecture. The training options considered were: 1. data augmentation, 2. the loss function, and 3. the amount of regularization. All deep learning training experiments were trained for 250 epochs with a learning rate of 0.001 using the Adam optimizer with $(\beta_1, \beta_2) = (0.9, 0.999)$. The dataset was split into training and test sets where 21 samples were used for testing, and the remaining 62 samples were used for training.

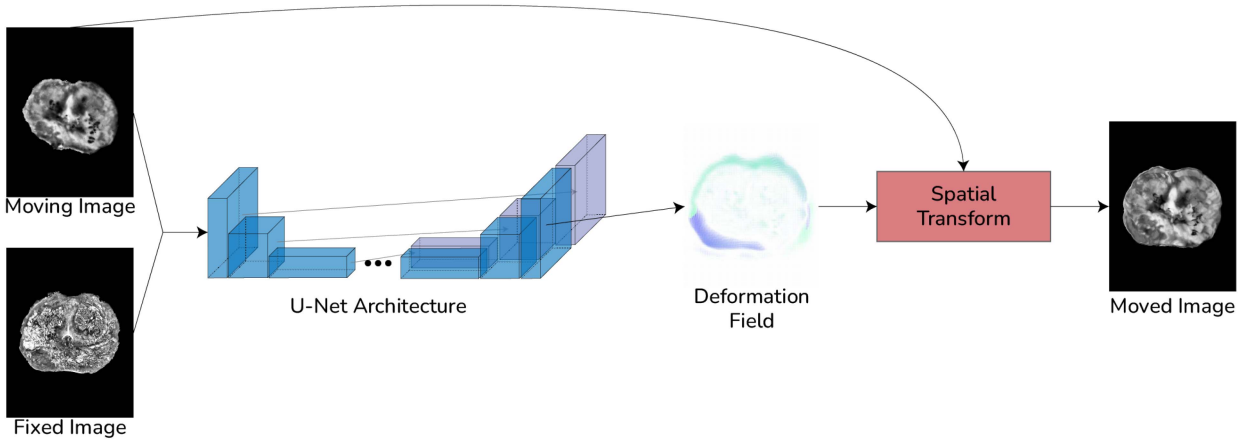


Fig. 3. Overview of the VoxelMorph-based neural network architecture applied to the registration task of white light snapshot images and HE images. The moving and fixed images are concatenated and given as input to the U-Net based convolutional neural network (CNN). It consists of 10 layers utilizing a 3 × 3 kernel with 2 × 2 max pooling and 32 channels per layer, except the last layer has 16 channels. Skip connections are depicted by the blue cuboids of the encoder becoming purple cuboids in the decoder which are added to the blue cuboids of the decoder. This CNN will output a registration field, which describes for each pixel in the image a displacement vector indicating where that pixel should be moved towards. The next step in the pipeline is the spatial transformer layer. This is where the registration field is applied to the moving image by means of linear interpolation to obtain the moved image.

F. Data Augmentation

The available data was limited; thus, to enlarge the size of the dataset, data augmentation was explored. Elastic deformations were used for data augmentation to increase the amount of training data. These elastic deformations are very general and flexible, however, the severity of the deformations has to be limited in order to make the deformations physically plausible. These elastic deformations were implemented using a library for elastic grid-based deformations for N-dimensional images [26]. The method employed produces a coarse grid with random displacement vectors sampled from a Gaussian distribution. In this case, the grid was chosen to be a 5×5 grid, with the Gaussian distribution having $\sigma = 7$. The grid is then interpolated to obtain a displacement vector for each pixel in the image. Afterward, the displacement matrix was applied to the input image to obtain an augmented image.

The network essentially has two images as input: the moving image and the fixed image. Hence, there are four ways in which one can do the augmentation: 1) augment only the moving image, 2) augment only the fixed image, 3) augment both the moving and fixed image with the same deformation, and 4) augment both the moving and fixed image with a different deformation. Training experiments were conducted to determine which of these options is most desirable for this application.

G. Loss Functions

The selection of loss functions in image registration is crucial as it directly influences the accuracy and robustness of the registration process. The loss function serves as a quantitative measure of the dissimilarity between the reference and transformed images, guiding the optimization process to find the optimal transformation parameters for accurate alignment. The general form of the loss function used for training the networks is

$$\mathcal{L}_{\text{total}}\left(m, f, \phi\right) = \mathcal{L}\left(m \circ \phi, f\right) + \lambda \mathcal{R}\left(\phi\right). \tag{1}$$

Here $\mathcal{L}_{\text{total}}$ is the complete loss function comprised out of $\mathcal{L}$ which computes the similarity between the moved and fixed image, and $\mathcal{R}$ which applies regularization to the registration field. The significance of this regularization to the overall loss is determined by the regularization parameter λ. We investigated four different loss functions for training the registration network: 1) Mutual information ($\mathcal{L}_{MI}$), 2) no background mutual information ($\mathcal{L}_{NoBGMI}$), 3) dice similarity coefficient ($\mathcal{L}_{DSC}$), and 4) mean squared error ($\mathcal{L}_{MSE}$). Additionally, the linear combination of two loss functions was also investigated.

1) Mutual Information: Mutual information is defined as

$$\mathcal{L}_{MI}(g, f) = -\sum_{v,w} h_{gf}(v, w) \cdot \log\left(\frac{h_{gf}(v, w)}{h_g(v) \cdot h_f(w)}\right). \tag{2}$$

Here h_{gf} is the joint 2D histogram of the images g and f. The symbols h_g and h_f are the marginal histograms defined by $h_g(v) = \sum_w h_{gf}(v, w)$ and $h_f(w) = \sum_v h_{gf}(v, w)$. The arguments v and w describe the bins of the of the joint histograms h_{gf} along each dimension. The amount of bins taken for each dimension of the histogram h_{gf} was 48. The mutual information is a measure that expresses the dependence of the joint histogram of f and g relative to their marginal histograms. Intuitively, for a correct registration between the images f and g, this relative dependence is expected to be large. Note, however, the minus sign in the definition of the mutual information loss; this means that a lower value corresponds to a larger relative dependence of the joint histogram to the marginal histograms, and that, in turn, implies a better similarity between f and g.

2) No Background Mutual Information: If a large part of the image contains background pixels, this could affect the histogram and the mutual information, thus potentially assigning

more weight to the background pixels than relevant. To investigate this, we proposed a new mutual information calculation, in which the background pixels are disregarded. By disregarding the background pixels, it is expected that the registration map outputted by the network will primarily focus on (foreground) interior pixels. The no background mutual information loss $\mathcal{L}_{\text{NoBG MI}}$ is defined by using the same equation as the regular mutual information defined in (2). However, instead of using the regular joint 2D histogram h_{gf}, a modified version h^*_{gf} is used where the first bin is disregarded from the histogram. The first bin contains the count of the number of pixels with pixel value 0, which corresponds to the background pixels after having scaled the image pixel values to be within $[0, 1]$. Consequently, in this way, the mutual information is calculated by disregarding the contribution of the background pixels. The number of bins taken for each dimension of the histogram h_{gf} was again 48, therefore the number of bins for h^*_{gf} was 47.

3) Dice Similarity Coefficient: The dice similarity coefficient loss ($\mathcal{L}_{DSC}$)is defined as

$$\mathcal{L}_{DSC}(g, f) = -\frac{2\,|g \cap f|}{|g| + |f|}. \tag{3}$$

Where $|g|$ and $|f|$ gives the number of nonzero pixels in the images g and f, respectively. And $|g \cap f|$ denotes the number of pixels that are nonzero in the intersection of images g and f. Note that for the purposes of this loss function, the image dimensions of g and f are assumed to be the same. Furthermore, similar to the mutual information loss definition, there is a minus sign in the definition so that a lower value implies a better similarity. Since this loss function measures the amount of overlap between two images, it is expected that using it will result in increased overlap between the moving and fixed image.

4) Mean Squared Error: Finally, the mean squared error loss is defined as

$$\mathcal{L}_{MSE}(g, f) = \frac{1}{N} \cdot \sum_{n=1}^{N} (g_n - f_n)^2. \tag{4}$$

Where g_n and f_n represent the pixel intensity values of the images g and f respectively, with N being the total number of pixels. With this loss function a better similarity is directly implied by a lower value. This loss function directly compares the pixel intensities at corresponding locations. Because of that, it is due to the grayscale pre-processing step that using this loss function is meaningful in this context.

5) Combination of Loss Functions: Each loss function has its own implications on what the network ultimately learns. Hence, it was also considered to investigate the linear combination of two loss functions. Given two of these loss functions $\mathcal{L}_1$ and $\mathcal{L}_2$, the combined loss function shown in (5) was formed

$$\mathcal{L} = a\mathcal{L}_1 + b\,(\gamma\mathcal{L}_2). \tag{5}$$

In combining these loss functions one has to keep in mind that the loss values of $\mathcal{L}_1$ and $\mathcal{L}_2$ are not necessarily the same in absolute value. Consequently, $\mathcal{L}_1 + \mathcal{L}_2$ does not necessarily imply an approximate $1:1$ contribution of $\mathcal{L}_1$ and $\mathcal{L}_2$. This is what the γ variable aims to solve. Consider the $\mathcal{L}_1$ loss value after convergence: $\mathcal{L}_{1,\,\text{conv}}$, and also consider the loss value of $\mathcal{L}_2$ after convergence: $\mathcal{L}_{2,\,\text{conv}}$. If one takes $\gamma \approx |\mathcal{L}_{1,\,\text{conv}}/\mathcal{L}_{2,\,\text{conv}}|$, then it will approximately hold that $|\mathcal{L}_{1,\,\text{conv}}| \approx |\gamma\mathcal{L}_{2,\,\text{conv}}|$. Which then, by using this γ factor, does give $\mathcal{L}_1$ and $\mathcal{L}_2$ a similar contribution to the total loss. Hence γ is a scale factor that aims to get the absolute value of the two loss functions approximately the same. With that, a and b are scale factors that decide how much relative weight each loss function gets in the calculation of the total loss. In light of the regularization function also playing a role in the calculation of the complete loss in (1), one might desire to have $a + b = 1$. This will imply that the relative contribution of the regularization parameter is approximately the same for $\mathcal{L}$ as it is for $\mathcal{L}_1$. This is, however, not strictly necessary because not doing it only changes the relative contribution of the regularization on $\mathcal{L}$ compared to what that contribution would have been on $\mathcal{L}_1$.

6) Regularization: In this context, regularization reduces the possibility of the registration field ϕ being physically implausible. The regularization function $\mathcal{R}$ is defined on the spatial gradients of ϕ by

$$\begin{aligned} \mathcal{R}(\phi) &= \sum_{n=1}^{N} ||\nabla\phi(x_n, y_n)||^2, \text{ where} \\ \nabla\phi(x_n, y_n) &= \left(\frac{\partial\phi(x_n, y_n)}{\partial x}, \frac{\partial\phi(x_n, y_n)}{\partial y}\right). \end{aligned} \tag{6}$$

where x_n and y_n are respectively the x and y coordinates of the n-th pixel with N being the total amount of pixels. The effect of this regularization function is that it smooths the spatial gradients of the registration field. The relative contribution of this regularization to the total loss is determined by the regularization parameter λ as specified in (1). Training was done by taking the following regularization parameters for each loss function: 0.05, 0.5, 1, and 2.

H. Evaluation

A set of four evaluation metrics was employed to assess the network's performance in image registration. Two of them were intensity-based, one based on a regional shape within the moving and the fixed image, and one based on deformation. Besides these four quantitative evaluation metrics, we also took note of the visual output of the network.

1) Dice Similarity Coefficient: Firstly the dice similarity coefficient (DSC), also known as the dice score, was utilized to gauge the similarity between the images generated by the network and those of the fixed image. The DSC is defined as

$$\text{DSC}(g, f) = \frac{2\,|g \cap f|}{|g| + |f|}. \tag{7}$$

Where $|g|$ and $|f|$ gives the number of nonzero pixels in the images g and f respectively. And $|g \cap f|$ denotes the number of nonzero pixels in the intersection of images g and f.

2) Mutual Information Score: Mutual information score (MI) measures the statistical dependence between pixel intensities of corresponding locations in two images, providing a quantitative assessment of their similarity by evaluating the amount of shared information. A higher mutual information score indicates greater

similarity between the images. The Mutual Information Score was adopted to quantify shared information between the two images. The MI is defined as

$$\mathrm{MI}(g,f) = \sum_{v,w} h_{gf}(v,w) \cdot \log\left(\frac{h_{gf}(v,w)}{h_g(v)\cdot h_f(w)}\right). \quad (8)$$

Here h_{gf} is the joint 2D histogram of the images g and f. The symbols h_g and h_f are the marginal histograms defined by $h_g(v) = \sum_w h_{gf}(v,w)$ and $h_f(w) = \sum_v h_{gf}(v,w)$.

3) Target Registration Error: In the context of medical image registration, the Target Registration Error (TRE) measures the accuracy of aligning one set of points to another. The TRE between two points (x_1, y_1) and (x_2, y_2) is defined as

$$\mathrm{TRE}\left((x_1,y_1),(x_2,y_2)\right) = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2} \quad (9)$$

The TRE was calculated with the snapshot images of the cleaved prostate including protruding cannulae, and HE images depicting the holes created by these cannulae. First, the area of the protruding cannulae in the snapshot images were segmented. These areas were then moved by using the registration field derived from the network output. For each of these moved areas, the center of mass was determined to serve as the center point. Additionally, the corresponding center points of the cannula holes were selected in the HE images. The TRE is then calculated by calculating (9) for the pairs of corresponding points of transformed snapshot image points and selected HE image points. The TRE is calculated from 17 positions across five prostate samples.

4) Regional Dice Similarity Coefficient: In order to calculate the regional DSC, a selection of samples was identified with visibly corresponding sub-areas in the snapshot and HE image. These areas correspond for instance to benign prostate hyperplasia, or in some cases to tumor areas. These areas were then segmented for both the snapshot and the HE image resulting in two binary masks. The binary mask of the snapshot image was consequently moved by applying the registration field derived from the network output. After this the DSC was calculated between the transformed binary mask of the snapshot image and the binary mask of the HE image. This regional DSC is calculated from nine samples, each having at least one notable region that can be identified both on the snapshot and HE image.

III. Results

A. Effect of Data Pre-Processing

To show the effect of pre-processing on the outcome of registration, we have trained and tested the U-Net based CNN using an MSE loss function with a regularization parameter of 0.05 on the unprocessed images and the pre-processed images. The unprocessed input images were converted to grayscale in order for the network to process them. The evaluation metrics from the network output can be seen in Table I.

The improvement in evaluation metrics is mainly visible in the TRE and regional DSC, as there is little to no improvement in MI and DSC after utilizing data pre-processing. It has been observed that the network which used pre-processed input gives better visual output than the network which used unprocessed input. This is, among others, because the latter network does not account for rotational differences, as evident from Fig. 4.

TABLE I
Raw Input Evaluation Metrics Compared to Pre-Processed Evaluation Metrics

Evaluation Metric	Unprocessed Input Mean±SD	Pre-processed Input Mean±SD
MI	0.64 ± 0.03	0.60± 0.06
DSC	0.98 ± 0.01	0.98 ± 0.01
TRE (mm)	6.6 ± 5.9	2.6 ± 1.6
Regional DSC	0.44 ± 0.23	0.69 ± 0.05

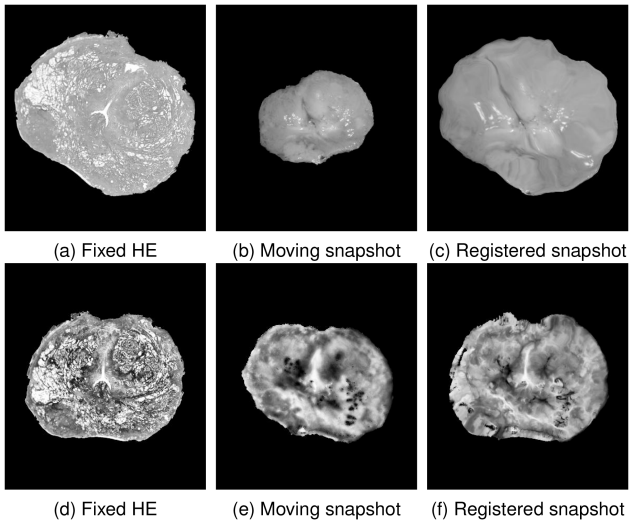


Fig. 4. Visual output for a single sample. (a), (b), and (c) show the fixed HE, the moving snapshot and the network output for the network that used unprocessed input. (d), (e), and (f) show the fixed HE, the moving snapshot and the network output for the network that used pre-processed input.

B. Data Augmentation

The effect of data augmentation on the training performance was investigated using the MSE loss function at a regularization parameter of 0.05. The original amount of image pairs is 83 samples and this amount was increased by adding two augmentations per image pair, which means that 166 image pairs were added. From the total of 249 image pairs: 186 samples were used for training and 63 samples for testing. Fig. 5 shows the performance of the different augmentation options in terms of the evaluation metrics: DSC, MI, TRE, and Regional DSC. Five augmentation options were considered: no augmentation (None), augmentation of the moving and fixed image with the same transformation (Same), augmentation of the moving image (Moving), augmentation of the fixed image (Fixed), and augmentation of the moving and fixed image with different transformations (Both). The evaluation metrics DSC and MI give a slight indication to the favorable performance of no augmentation and augmentation of only the moving image.

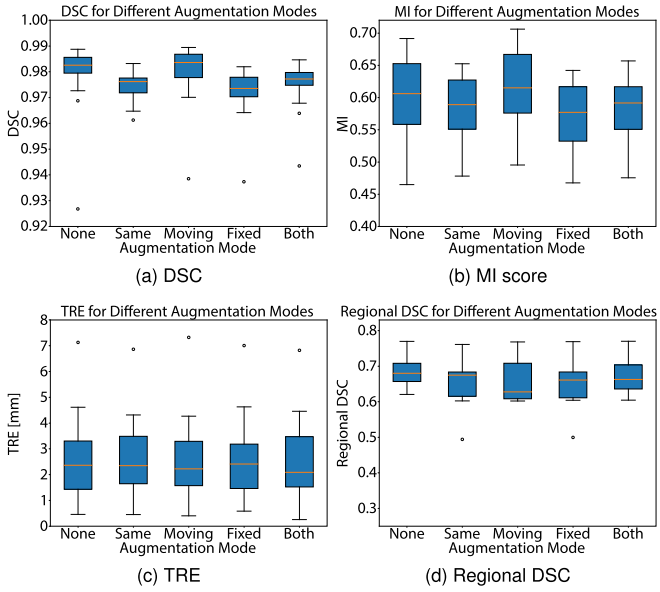


Fig. 5. Augmentation effect: (a) DSC, (b) MI, (c) TRE, and (d) Regional DSC, calculated for the different data augmentation options. None means no augmentation, same means the moving and fixed image are augmented with the same transformation, moving means only the moving image is augmented, fixed means only the fixed image is augmented, and both means that the moving and fixed image are augmented with different transformations. Network models were trained using the mean squared error (MSE) loss function with a regularization parameter of 0.05. The training data was increased to three times its original size for each of the augmentation options except for no augmentation. The image comparison metrics DSC and MI are calculated using all 21 non-augmented samples in the test set. The TRE is calculated from 17 positions across five samples. And the Regional DSC is calculated from nine samples, each having at least one notable region that can be identified both on the RGB and HE image.

However, the performance is comparable between these two options. Furthermore, since the other evaluation metrics, TRE and Regional DSC, show a similar performance of all options, it follows that, in this case, data augmentation is of no benefit to network performance.

C. Loss Functions

The implication of different loss functions on the network output was investigated by considering the four loss functions: mutual information (MI), no background mutual information (NoBG MI), dice similarity coefficient (DSC), and mean squared error (MSE), all trained with a regularization parameter of 0.05. Fig. 6 shows the visual output of the different loss functions. Qualitatively, all loss functions except NoBG MI show the capability to approximate the general shape of the fixed HE image from the moving snapshot image. However, as is visible in Fig. 6(b), the MSE loss function may produce tears in the model output, which reduces the quality of the registration. Similarly, albeit on another note, the network output of the DSC loss shown in Fig. 6(c) can be considered poor in quality, because the image structure and intensities are neither matching the moving nor the fixed image. The network trained with the MI-loss function shows in Fig. 6(d) an image output that qualitatively looks most stable. On the other hand, in Fig. 6(e) the NoBG MI loss function shows a visual output that resembles the fixed HE image with much of the granular structure maintained.

To better understand how the visual output of the trained networks is formed, Fig. 6(g)–(j) shows the registration fields to the network outputs of Fig. 6(b)–(e). The deformation fields corresponding with the MSE and MI loss functions, displayed in Fig. 6(g) and (i) respectively, show that most deformation is concentrated around the non-overlapping region of the fixed and moving image. This is consistent with the observed improvement in general overlap. Contrasting this is the NoBG MI loss, displayed in Fig. 6(j), which predominantly shows deformations of smaller magnitude in the interior of the prostate image. These finer interior deformations allow the network output to resemble the interior of the fixed HE image. Finally the DSC loss, displayed in Fig. 6(h), has both large deformations at the non-overlapping regions between the moving and fixed image, as well as large deformations within the prostate interior. It is unclear what causes these large deformations within the prostate interior.

Fig. 7 shows the evaluation metrics DSC, MI, TRE, and Regional DSC for each loss function. The qualitative observation of the MSE, DSC, and MI loss functions approximating the shape of the fixed image from the moving image can be seen in the performance of these losses with respect to the DSC evaluation metric. In line with this, the DSC metric shows that the NoBG MI loss does not perform well at approximating the shape of the fixed image. However, it should be noted that both the DSC and NoBG MI losses are performing relatively poorly with respect to the Regional DSC evaluation metric. Which can be attributed to the qualitative assessments made regarding their registration fields. Meanwhile, the MSE loss shows higher performance despite noticeable visual tearing in the output image. Its top performance in the Regional DSC score is evident in Fig. 7(d).

D. Regularization

The effect of regularization can be understood by (6) as making neighboring deformations similar to each other. The more weight one gives $\mathcal{R}$ to the total loss by a higher regularization parameter, the more value is given to the spatial gradients of ϕ being small. This implies that locally, the deformations will resemble each other. Here, the effect of the amount of regularization is investigated for a selection of loss functions. The losses MSE, DSC, and MI have been shown to improve the shape and overlap of the moving image to the fixed image. Whereas the NoBG MI loss improves the structural detail of the prostate interior to that of the HE image. These are both desirable properties to the registration, henceforth the following combined losses are considered: MSE + NoBG MI, DSC + NoBG MI, and MI + NoBG MI as described in (5) with a 1:1 ratio.

Fig. 8 and Table II demonstrate the evaluation metrics: DSC, MI, TRE, and Regional DSC of the aforementioned combined losses as a function of the regularization parameter. The MI evaluation metric clearly shows a decrease in performance for all losses for an increasing amount of regularization. As well as the MSE + NoBG MI loss achieving the best performance.

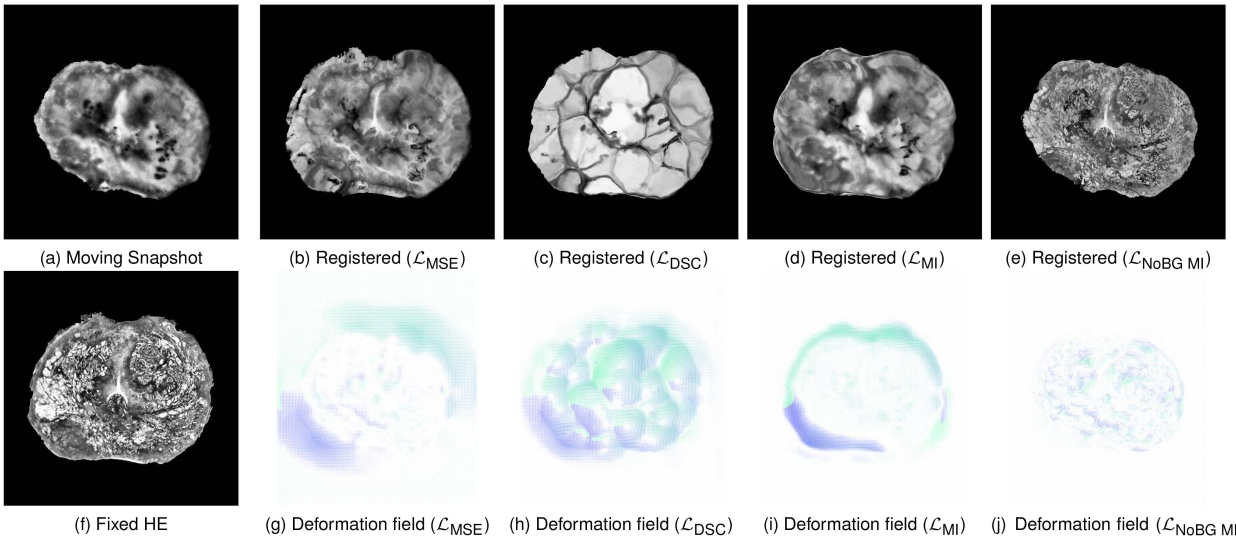


Fig. 6. Visual output of networks trained with a different loss function for a single sample. (a) and (f) show the moving snapshot and the fixed HE image respectively. These are the network inputs. The trained network outputs for different loss functions are shown by: (b) mean squared error (MSE) loss, (c) dice similarity coefficient (DSC) loss, (d) mutual information (MI) loss, and (e) no background mutual information (NoBG MI) loss. All these networks were trained with a regularization parameter of 0.05. Below these are the deformation fields for the networks trained for these different loss functions: (g) mean squared error (MSE), (h) dice similarity coefficient (DSC), (i) mutual information (MI), and (j) no background mutual information (NoBG MI). A blue color indicates deformation to the left, whereas green indicates deformation to the right. The intensity of the color is proportional to the amount of deformation at that location. All these networks have been trained with a regularization parameter of 0.05.

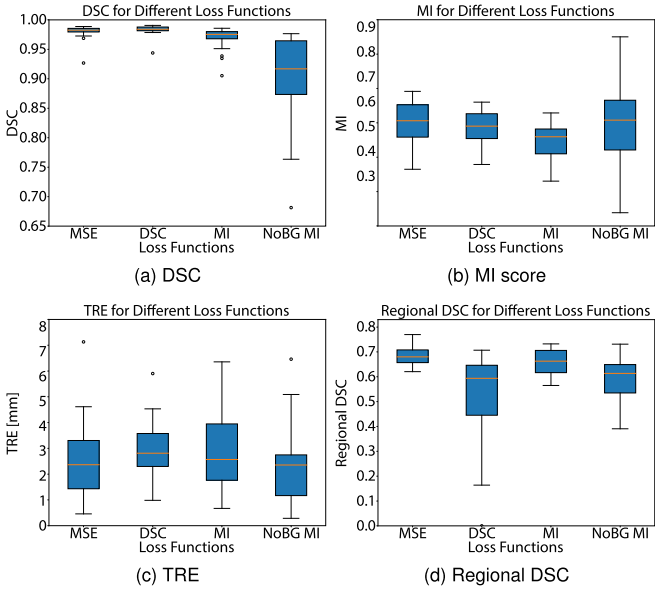


Fig. 7. Loss functions comparison: (a) DSC, (b) MI, (c) TRE, and (d) Regional DSC, calculated for the different loss functions. All losses were trained with a regularization parameter of 0.05. The image comparison metrics DSC and MI are calculated using all 21 samples in the test set. The TRE is calculated from 17 positions across five samples. And the Regional DSC is calculated from nine samples, each having at least one notable region that can be identified both on the RGB and HE image.

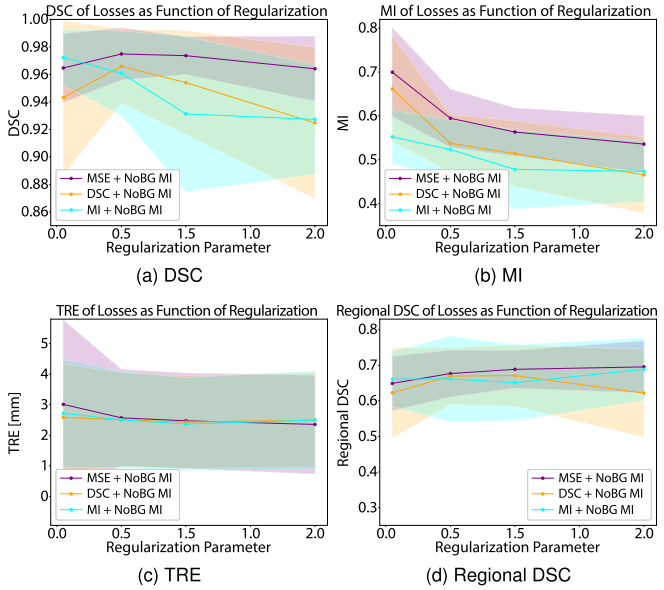


Fig. 8. Regularization parameter effect for different loss functions: (a) DSC, (b) MI, (c) TRE, and (d) Regional DSC, calculated as a function of the regularization parameter for a selection of combined loss functions. The image comparison metrics DSC and MI are calculated using all 21 samples in the test set. The TRE is calculated from 17 positions across five samples. And the Regional DSC is calculated from nine samples, each having at least one notable region that can be identified both on the RGB and HE image.

With respect to the DSC evaluation metric, it is also the MSE + NoBG MI loss that achieves the best performance with a high mean value and the smallest error when averaged over all the test samples. Here, the loss functions perform best with low regularization parameters of 0.5 or 0.05. For the Regional DSC evaluation metric, the MSE + NoBG MI loss performs best, showing the smallest errors and achieving the highest mean values. It is worth mentioning that the top performance

TABLE II
EVALUATION METRICS OF COMBINED LOSS FUNCTIONS ($\mathcal{L}$) FOR VARYING REGULARIZATION (λ)

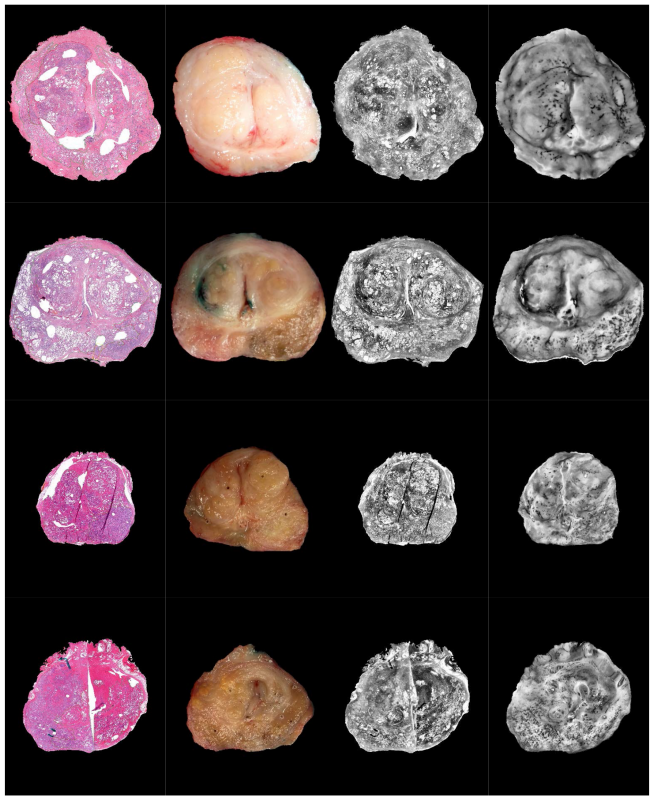
	$\lambda = 0.05$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$
	$\mathcal{L}$: MSE + NoBG MI			
DSC	0.96 ± 0.02	0.97 ± 0.02	0.97 ± 0.01	0.96 ± 0.02
MI	0.70 ± 0.10	0.59 ± 0.07	0.56 ± 0.05	0.54 ± 0.06
TRE	3.0 ± 2.8	2.6 ± 1.6	2.5 ± 1.6	2.4 ± 1.6
RegDSC	0.65 ± 0.08	0.68 ± 0.07	0.69 ± 0.05	0.70 ± 0.07
	$\mathcal{L}$: DSC + NoBG MI			
DSC	0.94 ± 0.06	0.97 ± 0.03	0.95 ± 0.04	0.92 ± 0.05
MI	0.66 ± 0.12	0.54 ± 0.07	0.51 ± 0.07	0.47 ± 0.09
TRE	2.6 ± 1.8	2.5 ± 1.5	2.4 ± 1.5	2.5 ± 1.6
RegDSC	0.62 ± 0.13	0.67 ± 0.08	0.67 ± 0.09	0.62 ± 0.12
	$\mathcal{L}$: MI + NoBG MI			
DSC	0.97 ± 0.02	0.96 ± 0.03	0.93 ± 0.06	0.93 ± 0.04
MI	0.55 ± 0.06	0.52 ± 0.07	0.48 ± 0.09	0.47 ± 0.07
TRE	2.7 ± 1.8	2.5 ± 1.6	2.4 ± 1.5	2.5 ± 1.6
RegDSC	0.66 ± 0.08	0.66 ± 0.12	0.65 ± 0.11	0.69 ± 0.09

of MSE + NoBG MI for the Regional DSC is obtained at a high regularization parameter of 2. With respect to the target registration error, there is less useful information available because all the models have a similar performance with large error bars.

E. Training Experiments

As can be seen from the wide standard deviations in Fig. 8, some network variants perform better in some cases than others. This section will provide some examples.

Fig. 9 shows examples of the visual output of one network configuration (MSE loss with regularization of 1.00) for four different cases. The first two rows of Fig. 9 show two examples in which the network had high MI scores of 0.65 and 0.64, respectively. Both these samples contained biopsy holes, as can be seen in Fig. 9(a). The third and fourth rows in Fig. 9 show two examples in which the network had lower MI scores (0.48 and 0.56, respectively). The sample in the third row had been folded during histopathological processing, as can be seen in Fig. 9(a), row 3. The sample in the fourth row was too large to fit into one histology cassette and, therefore, had to be split in two, resulting in the need to stitch the two HE coupes together for image registration, as can be seen in the HE image. The model still performed well even when the input was sub-optimal due to the HE image being folded in or stitched. However, the performance was poorer compared to situations where this did not occur.



(a) Preprocessed HE image (b) Snapshot (moving) (c) Gray-scale HE (fixed) (d) Registered snapshots

Fig. 9. Visual output from different prostate examples all using the same network configuration (MSE loss with regularization of 1.00). (a) shows the preprocessed HE prior to filling any holes. (b) shows the input snapshot images after preprocessing, but prior to gray conversion (moving image). (c) shows the input HE after gray conversion (fixed image). (d) shows the network output (registered snapshot images). The first and the second rows are well-performed registration examples. The third row is from an example in which the HE images were folded in the histopathological processing. The fourth row was from an example in which the prostate was too large to fit into one HE coupe, and thus two HE coupes had to be stitched together to create one input HE image.

IV. DISCUSSION

In this work, a pipeline for using an unsupervised neural network was investigated to register snapshot images to pathology hematoxylin-eosin (HE) stained images. The introduced pipeline aimed to tackle several arising issues including dissimilarity in the visible structure between macroscopic and microscopic images, extreme deformation due to tissue processing, utilizing clinically relevant evaluation metrics, and the importance of hyperparameters in training the model.

The first step in this pipeline involved pre-processing the data in a way that was tailored to the application. The model was trained twice, initially with unprocessed images and subsequently with pre-processed images, resulting in comparable averages for MI and DSC metrics. However, as shown in Table I, pre-processing significantly improves the registration performance, particularly in terms of TRE and regional DSC. Specifically, the pre-processing step resulted in a reduction in the average TRE from 6.6 mm to 2.6 mm, alongside an increase in regional DSC from 0.44 to 0.69.

Furthermore, variations in data augmentation, loss functions, and the weight of regularization were investigated to determine their effect on registration performance. From Fig. 5, it is observed that data augmentation is of no performance benefit to the network being trained. The target registration error scores and regional DSC are overall similar, albeit with varying medians and errors. Regarding the image comparison metrics DSC and mutual information (MI), there is still considerable overlap in the performance metrics of all augmentation modes. However,

from these results, one would choose either no augmentation or augmentation of the moving image over the other variants. These deep learning methods are data-driven approaches, implying the requirement of data to achieve high performance and to generalize well to unseen data. Often, the lack of data from the medical field limits the use of deep learning methods. However, this investigation shows that the current method does not suffer from this limitation. A reason for this could be the extensive pre-processing. As a result of the pre-processing, all the images have canonical orientation due to rotation, comparable size differences due to scaling, and aligned urethra due to translation. Therefore the deformation fields that the network produces are more alike. Consequently, there are no additional details that the network can learn from using more data. On another note, in the cases where augmentation was applied to HE images (fixed images), we observed a slight decrease in performance. This decline could be attributed to the excessive detail present in the structures of HE images, as well as the possibility that our augmentation approach may not accurately reflect the realistic deformations tissues undergo during pathological processes.

The elastic deformation employed in this work for data augmentation encapsulates a general form of tissue deformation. However, the generalizability of this registration method can benefit from a more detailed investigation into the different kinds and severity of data augmentation. However, one should be careful within this framework, since augmentation in the form of rotation, scaling or translation can go against part of the pre-processing. Future work on data augmentation can focus on finding methods that better reflect the severity of tissue deformation originating from its pathological processing. This can then include augmentation methods to increase the robustness on data of lesser quality by, for instance, mimicking missing tissue areas on the HE image. In terms of acquired data, the 83 image pairs is a relatively small dataset for training a neural network. However, it is a sizable dataset considering that it comes from 83 unique patients. Nonetheless, a larger dataset, or rather, a multi-institutional dataset can much better test the generalizability of this method. Furthermore, we expect this method to work on similar problem descriptions which allow for a similar form of pre-processing to be done. But the extent of this generalization is best explored in future research considering different imaging modalities or different tissue types.

The deformation fields visualized in Fig. 6 provide the most detailed information regarding the use of different loss functions. As demonstrated in this figure, the choice of loss function has a significant influence on the predicted deformation field which could result in a substantial difference in registration outcome. By considering the deformation fields, one sees that the loss functions MSE and MI with high regional DSC scores have deformations focused on the non-overlapping area of the moving and fixed image. This result could be due to both the MSE and MI loss functions being dependent on differences in pixel intensities, albeit in a different way. Indeed, the largest area of discrepancy of the image intensities between the moving and fixed image is the non-overlapping area between them. In contrast, the DSC and NoBG MI losses have most of their deformations in various directions within the prostate interior, which results in overall lower Regional DSC scores with a larger variance. For the NoBG MI loss function this deformation field makes sense, since the loss was constructed to focus on the surface interior. However, the DSC loss is expected to be a measure of overlap, so the large magnitude of internal deformation is unexpected. A possible explanation for this is the fact that the VoxelMorph framework implements the intersection of the DSC loss as a multiplication. Which works, since multiplication with 0 will disregard the contribution of the pixel. However, it also implies that a higher pixel intensity will improve the loss score. This could explain why the model output with the DSC loss is much brighter and tries to minimize the dark areas of the surface interior. Furthermore, regarding the TRE evaluation metric, it can be observed that there is no clear difference in performance across the different loss functions. In fact, all training experiments have shown a similar performance regarding the TRE. It follows that within this framework the improvement of the TRE is solely achieved as a result of the pre-processing.

Something that the MSE, MI, and DSC loss functions do have control over is the overlap of the moving and the fixed image. This is most easily observed for the loss functions MSE and MI in Fig. 6(g) and (i) respectively, because there is not much interior deformation for these loss functions. The important observation to make here is that the deformation accounting for this overlap is very concentrated around the borders of the non-overlapping area of the moving and fixed image. To see this, compare regions with a lot of deformation from Fig. 6(g) or (i) to the same region in the moving and fixed image shown in Fig. 6(a) and (f). The bottom-right has little deformation because there is already a lot of overlap there. However, the bottom-left and top-right have a lot more deformation because there is more non-overlapping area.

For the loss functions MSE and MI, the visualization of the deformation fields in Fig. 6 also show that the deformation magnitude vanishes quickly in the transition from an area of non-overlap to an area of overlap. There is a clearly delineated inner border within the prostate sample in Fig. 6(g) and (i), where the deformation magnitude vanishes compared to the outer border. This abrupt decrease in the amount of deformation does not agree with the elasticity that one would expect from tissue samples.

The way to mitigate this in the current framework is by increasing the weight of regularization because the function used for regularization, shown in (6), implies that neighboring deformation vectors will look like each other. However, one cannot increase the regularization parameter too much either. The regularization function will be applied to all vectors, which means it is also applied to the vectors in the overlapping area that have little deformation. And there are a lot more of those vectors than vectors with large deformation around the non-overlapping area of the image. The effect of increasing the regularization weight will then be that more deformation vectors in the non-overlapping area will be limited in magnitude in order to look like the vectors in the overlapping area that are small in magnitude. Add to this the deformation vectors from the black background of the image, which are all also small or nearly zero in magnitude. By this reasoning, there is a limit to how high

the regularization parameter can be set to still be meaningful. An interesting consideration could be having a regularization of the kind of (6) only applied to vectors of the non-overlapping area while also excluding the background pixels. Alternatively, a different kind of regularization function could be investigated that allows for a physically plausible deformation field. In the same spirit, future research can consider more complex loss functions that try to leverage more anatomical information.

In this work, we have used three well-known quantitative evaluation metrics including dice similarity coefficient, mutual information, target registration error, and additionally, we introduced a new metric called regional DSC. From these metrics, DSC and MI are strict image comparison metrics aimed at evaluating image similarity. The DSC is an important evaluation metric for this purpose because it can quantify the amount of overlap. Since pixel-wise comparison holds less relevance, due to comparing images from different modalities, the MI evaluation metric provides a better way of comparing pixel intensities in terms of the overall histograms. On the other hand, the TRE and regional DSC are evaluation metrics aimed at providing information that is more clinically relevant. The TRE gives a measure of the distance that the registration is off, and the regional DSC gives the amount of overlap from common areas from the prostate interior identified in both the snapshot image and the HE image. An important observation to make regarding these evaluation metrics is that high performance of the image comparison metrics does not imply that the registration is clinically relevant. Table I shows that both the unprocessed and pre-processed inputs can result in similar performance regarding the image comparison metrics (DSC and MI) after training. However, the clinically motivated evaluation metrics (TRE and regional DSC) show that the registration is not accurate, and Fig. 4 provides an example of how it fails to be accurate. Here, the moving and fixed images are rotated with respect to each other, but the model output was not able to correct for this rotation. Instead, the moving image is radially stretched out to improve the overlap with the fixed image. It follows that the clinically motivated evaluation metrics are more relevant than the image comparison evaluation metrics.

From the comprehensive analysis of this framework, it rests to specify the preferred training configuration for the current application of prostate images. It followed from Fig. 5 that the effect of data augmentation has no significant influence on the registration performance, hence training can be done without data augmentation. With respect to the loss function, it was established that using one of the combined loss functions as investigated in Fig. 8 is most desirable. From these loss functions, the combined loss MSE + NoBG MI achieves the best overall performance. Finally, given the interpretation of the regularization and the importance of clinically relevant evaluation metrics, it can be concluded from Fig. 8 that a regularization factor of 2 is most desired with the MSE + NoBG MI loss function. Using such reasoning an appropriate configuration may be found for a variety of applications.

There is limited research on image registration of camera snapshot images to pathology images, making it difficult to compare the proposed method with others in the same application. However similar work has been done in image registration of MRI images to pathology images [10], [20], [21]. Particularly, the research setup of ProsRegNet [21] is similar to the one of this work. Here the authors report an achieved Dice (DSC) score of 0.98 (±0.01) and a landmark error of 2.68 (±0.68) mm. Which are comparable to the results obtained in this work. A big difference in the landmark point selection is that the landmark pairs in MRI and pathology images are annotated by radiologists and pathologists, respectively. For a white light snapshot image of the prostate interior, there is no medical specialist who can annotate landmarks in conjunction with a pathologist.

The results of the preferred training configuration are quite satisfactory. However, the use of VoxelMorph has inherent limitations that cannot be resolved by any training configuration. The main limitation is that the spatial transformer layer in the VoxelMorph framework essentially only allows for the movement of pixel intensities. This means that any pixel intensity in the moved image, originates from an area in the moving image with a similar pixel intensity. Suppose for instance that there is a tear, a hole, or some missing tissue in the HE image with respect to the snapshot image. In order for this missing area (that is represented by a black area) in the HE to be reflected in the moved image, it will have to be corrected with some black pixel area from the moving image. This can lead to inaccurate deformations especially if the regularization factor is low. In the original application of VoxelMorph this is not a pertinent limitation, since the images are locally more consistent. Another point to consider is that the learnable part of VoxelMorph is essentially a U-Net. Which means that the features for registration are extracted through convolutional layers with down-sampling of the resolution through pooling layers. This, on the one hand, aids broad applicability of the same architecture. However, it is limited in the capability of integrating any shape or anatomical information directly into the network architecture. In contrast, a more modern transformer architecture could allow for choosing a specific method of patching the image to highlight certain anatomical consistencies between the image pairs. Future research could explore the trade-off between improved performance and generalizability of such an approach. The work of [27] provides a promising unsupervised transformer-based model to test these ideas with.

A topic that has not been investigated is the choice of moving and fixed images. Two image types are considered: snapshot images from fresh specimens and HE-stained images of the same specimen after histopathologic processing. Either one can be chosen as the moving image with the other being the fixed image. In this work, the snapshot image has been the moving image for all trained networks. Since both moving and fixed images are used as input for this unsupervised method, it is less obvious what kind of effect changing the role of the images will have. However, the learned deformation field is applied to the moving image which means that different features are learned by the network to establish the deformation fields if a different moving image is used. It is worth mentioning that the learned deformation fields are, in general, not invertible transformations meaning that one can not obtain a transformation from the HE image to the snapshot image by knowing a transformation

from the snapshot image to the HE image. This suggests that changing the role of the snapshot and HE image leads to a transformation unattainable without such a role reversal. Given these observations, it is imperative to explore the selection of moving and fixed images further.

V. Conclusion

The results show that the pipeline presented in this work is capable of multi-modal registration of camera snapshot images to HE images. This pipeline includes pre-processing of images using rotation, resizing, image translation, inpainting, and grayscale conversion followed by a deep-learning-based registration model. The results demonstrate that the best-performing network can be trained without data augmentation using the MSE + NoBG MI loss function with a regularization parameter of 2. This achieves good registration performance with a DSC of 0.96, an MI score of 0.54, a TRE of 2.4 mm and a regional DSC of 0.70.

Acknowledgment

The authors would like to thank all urologists and surgeons from the Department of Surgery at the Netherlands Cancer Institute for their assistance in collecting the specimens and all pathologists and pathologist assistants from the Department of Pathology and the Core Facility Molecular Pathology and Biobanking for processing the specimens and making the data available.

References

[1] M. Unger and J. N. Kather, "Deep learning in cancer genomics and histopathology," *Genome Med.*, vol. 16, no. 1, pp. 44–44, 2024.

[2] S. Jonmarker, A. Valdman, A. Lindberg, M. Hellström, and L. Egevad, "Tissue shrinkage after fixation with formalin injection of prostatectomy specimens," *Virchows Archiv*, vol. 449, no. 3, pp. 297–301, Sep. 2006, doi: 10.1007/s00428-006-0259-5.

[3] C. Hughes, O. Rouvière, F. Mege-Lechevallier, R. Souchon, and R. Prost, "Robust alignment of prostate histology slices with quantified accuracy," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 2, pp. 281–291, Feb. 2013. [Online]. Available: https://ieeexplore.ieee.org/document/6336793

[4] X. Jiang, J. Ma, G. Xiao, Z. Shao, and X. Guo, "A review of multimodal image matching: Methods and applications," *Inf. Fusion*, vol. 73, pp. 22–71, 2021.

[5] P. Markelj, D. Tomaževič, B. Likar, and F. Pernuš, "A review of 3D/2D registration methods for image-guided interventions," *Med. Image Anal.*, vol. 16, no. 3, pp. 642–661, Apr. 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841510000368

[6] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1153–1190, Jul. 2013. [Online]. Available: https://ieeexplore.ieee.org/document/6522524/

[7] L. L. de Boer et al., "Method for coregistration of optical measurements of breast tissue with histopathology: The importance of accounting for tissue deformations," *J. Biomed. Opt.*, vol. 24, no. 7, Jul. 2019, Art. no. 075002. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6995961/

[8] A.-L. Grosu et al., "11c-choline pet/pathology image coregistration in primary localized prostate cancer," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 41, no. 12, pp. 2242–2248, 2014.

[9] J. Albers et al., "Elastic transformation of histological slices allows precise co-registration with microct data sets for a refined virtual histology approach," *Sci. Rep.*, vol. 11, no. 1, pp. 10846–10846, 2021.

[10] L. Li et al., "Multi-scale statistical deformation based co-registration of prostate MRI and post-surgical whole mount histopathology," *Med. Phys.*, vol. 51, no. 4, pp. 2549–2562, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.16753

[11] I. N. Huszar et al., "Tensor image registration library: Deformable registration of stand-alone histology images to whole-brain post-mortem mri data," *NeuroImage*, (Orlando, FL), vol. 265, pp. 119792–119792, 2023.

[12] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "Deep learning in medical image registration: A review," *Phys. Med. Biol.*, vol. 65, no. 20, pp. 20TR01–20TR01, 2020.

[13] K. Drukker, P. Yan, A. Sibley, and G. Wang, "Chapter 4 - biomedical imaging and analysis through deep learning," in *Artificial Intelligence in Medicine*, L. Xing, M. L. Giger, and J. K. Min, Eds. New York, NY, USA: Academic Press, Jan. 2021, pp. 49–74. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128212592000041

[14] I. Galić, M. Habijan, H. Leventić, and K. Romić, "Machine learning empowering personalized medicine: A comprehensive review of medical image analysis methods," *Electron. (Basel)*, vol. 12, no. 21, 2023, Art. no. 4411.

[15] M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, "Training CNNs for image registration from few samples with model-based data augmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Switzerland, 2017, pp. 223–231.

[16] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: A learning framework for deformable medical image registration," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1788–1800, Aug. 2019.

[17] A. Ruchti, A. Neuwirth, A. K. Lowman, S. R. Duenweg, P. S. LaViolette, and J. D. Bukowy, "Homologous point transformer for multi-modality prostate image registration," *PeerJ. Comput. Sci.*, vol. 8, pp. e1155–e1155, 2022.

[18] I. Bhattacharya et al., "Bridging the gap between prostate radiology and pathology through machine learning," *Med. Phys. (Lancaster)*, vol. 49, no. 8, pp. 5160–5181, 2022.

[19] X. Yu et al., "Deep attentive panoptic model for prostate cancer detection using biparametric MRI scans," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2020, pp. 594–604.

[20] W. Shao et al., "Raphia: A deep learning pipeline for the registration of MRI and whole-mount histopathology images of the prostate," *Comput. Biol. Med.*, vol. 173, pp. 108318–108318, 2024.

[21] W. Shao et al., "ProsRegNet: A deep learning framework for registration of MRI and histopathology images of the prostate," *Med. Image Anal.*, vol. 68, pp. 101919–101919, 2021.

[22] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[23] T. Han, J. Wu, W. Luo, H. Wang, Z. Jin, and L. Qu, "Review of generative adversarial networks in mono- and cross-modal biomedical image registration," *Front. Neuroinform.*, vol. 16, pp. 933230–933230, 2022.

[24] X. Song et al., "Cross-modal attention for multi-modal image registration," *Med. Image Anal.*, vol. 82, pp. 102612–102612, 2022.

[25] L. Feenstra, M. Lambregts, T. J. Ruers, and B. Dashtbozorg, "Deformable multi-modal image registration for the correlation between optical measurements and histology images," *J. Biomed. Opt.*, vol. 29, no. 6, 2023, doi: 10.1117/1.JBO.29.6.066007.

[26] G. van Tulder, "Elasticdeform: Elastic deformations for N-dimensional images," Sep. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7102577

[27] Z. Chen, Y. Zheng, and J. C. Gee, "Transmatch: A transformer-based multilevel dual-stream feature matching network for unsupervised deformable image registration," *IEEE Trans. Med. Imag.*, vol. 43, no. 1, pp. 15–27, Jan. 2024.