Segmentation of glandular epithelium in colorectal tumours to automatically compartmentalise IHC biomarker quantification: a deep learning approach

Yves-Rémi Van Eycke[a,b,*], Cédric Balsat[a], Laurine Verset[c], Olivier Debeir[b,d], Isabelle Salmon[a,c], Christine Decaestecker[a,b,**]

[a] *DIAPath, Center for Microscopy and Molecular Imaging, Université Libre de Bruxelles (ULB), CPI 305/1, Rue Adrienne Bolland, 8, 6041 Gosselies, Belgium.*
[b] *Laboratories of Image, Signal processing & Acoustics, Université Libre de Bruxelles (ULB), CPI 165/57, Avenue Franklin Roosevelt 50, 1050 Brussels, Belgium.*
[c] *Department of Pathology, Erasme Hospital, Université Libre de Bruxelles (ULB), Route de Lennik 808, 1070 Brussels, Belgium.*
[d] *MIP, Center for Microscopy and Molecular Imaging, Université Libre de Bruxelles (ULB), CPI 305/1, Rue Adrienne Bolland, 8, 6041 Gosselies, Belgium.*

Abstract

In this paper, we propose a method for automatically annotating slide images from colorectal tissue samples. Our objective is to segment glandular epithelium in histological images from tissue slides submitted to different staining techniques, including usual haematoxylin-eosin (H&E) as well as immunohistochemistry (IHC). The proposed method makes use of Deep Learning and is based on a new convolutional network architecture. Our method achieves better performances than the state of the art on the H&E images of the GlaS challenge contest, whereas it uses only the haematoxylin colour channel extracted by colour deconvolution from the RGB images in order to extend its applicability to IHC. The network only needs to be fine-tuned on a small number of additional examples to be accurate on a new IHC dataset. Our approach also includes a new method of data augmentation to achieve good generalisation when working with different experimental conditions and different IHC markers. We show that our methodology enables to automate the compartmentalisation of the IHC biomarker analysis, results concurring highly with manual annotations.

Keywords: computational pathology, data augmentation, deep learning, gland, image segmentation, immunohistochemistry

*yveycke@ulb.ac.be
**cdecaes@ulb.ac.be

1. Introduction

Immunohistochemistry (IHC) is an efficient and routinely used technique to localise a specific antigen in a tissue sample and its cell components. This staining technique is commonly employed for diagnostic and prognostic purposes in histopathology as well as for biomarker validation in clinical research. Whole slide scanning and image analysis tools now enable to objectively and quantitatively evaluate IHC biomarkers in a whole tissue slide or a specific region of interest delineated by a pathologist. Compartmentalising the quantitative evaluation of IHC biomarkers in a specific histological structure, such as glandular epithelium in colorectal, breast or prostatic lesions, is often required in histopathology to provide more relevant and informative measurements for clinical research, as evidenced by Verset et al. (2013, 2015). For this purpose, pathologists have to annotate thousands of structures present in histological slide series, a long, tedious and potentially biased task that would greatly benefit from automation. Through different public challenges, such as GLaS (2015), CAMELYON16 (2016), TUPAC16 (2016) and CAMELYON17 (2017), deep learning approaches recently demonstrated their efficiency to automatise such annotation tasks performed on haematoxylin-eosin (H&E)-stained slides in the context of computer-aided diagnosis as shown in CAMELYON16 (2016), CAMELYON17 (2017) and Sirinukunwattana et al. (2017). In the present work, we propose a new method, also based on deep learning, for automatically annotating glandular epithelium in colorectal tissue samples. Regarding the state-of-the-art, we designed our method to be more generic, i.e. to be applicable on H&E as well as any IHC staining. To validate our method and to compare its performance with the state of the art, we first used the public data set provided by the GlaS challenge and based on H&E staining. Secondly, to apply our method in the context of IHC staining, we created a second dataset in our lab by making use of tissue microarray slides that we submitted to IHC to reveal the expression of different antigens on colorectal tumour samples. An expert manually annotated the whole slide images to delineate the glandular epithelium. Using Lin's concordance correlation score we analyzed the concordance between a quantification made on the expert annotations and that obtained on the automatic annotations. In addition to be efficient in terms of segmentation, our approach aims to be sufficiently fast to enable its use on large series of whole slide images with a typical size of 1 gigapixel (at $20\times$ magnification), such as those that we generally use for IHC biomarker quantification.

2. Previous work and novel contributions

The automated delineation of histological structures, such as glandular tissue, in tissue slide images is a problem that was first raised in by Doyle et al. (2007); Naik et al. (2008). These first methods relied exclusively on handcrafted image characteristics and conventional supervised machine learning methods. Later on, graph-based approaches were used in supervised (Altunbay et al. (2010); Tosun and Gunduz-Demir (2011)) and unsupervised methods (Simsek

et al. (2012)). These approaches have the advantage of taking into account the positioning of the various structures in relation to each other within the tissue. Similarly, Olgun et al. (2014) proposed a method based on "local object pattern" in order to take into account the neighbourhood of each object detected. By 2015, Sirinukunwattana et al. (2015) proposed a method using a Monte Carlo simulation to more accurately detect the glands. During this same year we saw a huge increase in the number of methods using Deep Learning, especially on the occasion of the GlaS Challenge Contest where 106 teams proposed methods for gland segmentation (Sirinukunwattana et al. (2017)). Chen et al. (2017) presented the most effective methods at that time as DCAN. This deep convolutional network uses a network inspired by the "VGG" architecture proposed by Simonyan and Zisserman (2014) and for which particular care was taken for edge detection in order to separate the different glandular structures. U-Net from Ronneberger et al. (2015), which was proven efficient on other medical image segmentation problems, also yielded very good results in the GlaS Challenge Contest. Likewise, the ExB team proposed a network composed of two branches, one, deep and working on a local part of the image, and the other, shallower, working on larger image areas (Sirinukunwattana et al. (2017)). More recently, Xu et al. (2017b) proposed a deep learning method, which was the most accurate at the moment of its publication on the data used in the GlaS segmentation challenge. This method makes use of an additional channel to detect a bounding box around objects in addition to the two channels used by DCAN to segment the edges of objects and the objects themselves. In contrast to DCAN, each of these channels is the output of a particular convolutional network. Using as inputs the information provided by the three channels, an additional convolutional network is then required to produce the final segmentation. All the introduced changes resulted in a wider and deeper network than DCAN.

To solve image-related problems deep learning usually uses data augmentation in order to avoid overfitting and to increase generalisability and robustness with respect to spatial and colour variations. In the context of histological image processing, current practices often focus on geometric variations applied to the training images, such as affine transformation (e.g., flip, rotation and translation) and blurring, in order to make the model invariant for these transforms (Sirinukunwattana et al. (2017)). Recently, Xu et al. (2017b) show that the additional use of elastic transformations is beneficial for gland instance segmentation. Colour augmentation is also investigated to take into account stain variations. It usually consists in random transformations applied in the RGB or HSV colour space (see e.g., Sirinukunwattana et al. (2016, 2017); Lafarge et al. (2017)) or on principal components (Xu et al. (2017a); Mishra et al. (2017)). It should be noted that random variations in the RGB space should be small to prevent from producing aberrant colours out of the range of the H&E or IHC histological staining. Concerning principal components, studies on colour normalisation show that the principal components do not provide an appropriate representation of the colour space for H&E (Rabinovich et al. (2004)) and IHC staining (Van Eycke et al. (2017)). Furthermore, the colour alterations proposed in previous studies are generally based on linear transformations applied to the

whole image without specifically targeting the stained tissue. As detailed below (section 3.2.2), studies on colour normalisation suggest more realistic ways to alter the colours of histological images. Alternatively to colour augmentation, image preprocessing for colour normalisation is applied using simple techniques like in Xu et al. (2017b) using per channel zero mean or more complex ones like proposed by Ciompi et al. (2017).

All the previous works on histological structure segmentation mentioned above were designed to work with H&E staining only. Our first contribution aims to increase the application field by developing a more generic method that applies directly to H&E and IHC staining using haematoxylin (HEM) for tissue counterstaining, regardless of the targeted protein expression and the used chromogen. This contribution includes a new method of realistic data augmentation, which was developed to tackle the issue of stain and acquisition variability. Our second contribution is a new network architecture, which integrates the efficient properties proposed in three methods, namely DCAN (Chen et al. (2017)), U-Net (Ronneberger et al. (2015)) and the identity mappings proposed by He et al. (2016a,b) in the Residual Neural Network architectures. In addition, the resulting network was adapted to be much less resource intensive than the state of the art (Xu et al. (2017b)) by paying special attention to merge the layers that can be, while aiming to be more accurate than DCAN and U-Net. The third contribution consists in a quantitative comparison of compartmentalised characterisation of IHC staining carried out on the basis of manual vs. automatic segmentation.

3. Methods

3.1. Network Architecture

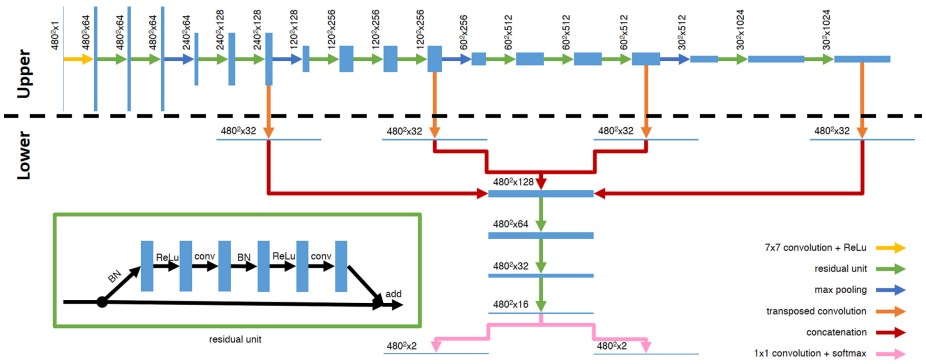

Figure 1: Network architecture: The fully convolutional network that we implemented.

Figure 1 shows the fully convolutional network that we implemented where two parts can be distinguished. The upper part carries out pixel classification. It allows the network to extract information related to the 2D structure of the input

data, while the lower part propagates and combines the information obtained at increasing levels in the upper part. The input resolution of 480×480 pixels was chosen because it can easily be divided by powers of 2 (for the successive network layers) while being large enough to cover most of the areas of the GlaS challenge images with reasonable memory consumption. If the input image is smaller than 480×480 pixels (e.g. after random crops carried out in the data augmentation process, see below), the image is padded by reflecting the borders. Each convolutional layer in the network, including those in the residual units, uses 3 by 3 kernels with the stride equal to 1, except the very first layer which uses a 7 by 7 kernel. A Rectified Linear Unit layer (ReLu) follows each convolutional layers, except the last two ones which are followed by a softmax unit layer. Each max-pooling layer uses 2 by 2 kernels with a stride equal to two. The residual units are those proposed by He et al. (2016b).

The upper part is composed mainly of max-pooling layers and residual units. The deepest layers of this part contain more accurate information at the expense of spatial resolution. This part of the network is very close to the architecture proposed in the VGG architecture (Simonyan and Zisserman (2014)) except for the inclusion of residual units. These units avoid the problem of vanishing gradient descent and thus allow us to propose a network that is much deeper than those proposed in DCAN and U-Net, i.e. 25 convolutional layers solely in the upper part of the network. The upper part of our network can also be seen as a simplified version of the Residual Network proposed by He et al. (2016a) but with the residual units they introduced later (He et al. (2016b)).

Contrary to the other networks that use 3 colour channels to process H&E staining images, our network has only one input channel corresponding to blue HEM extracted from the RGB images by colour deconvolution, using colour vectors set according to Ruifrok et al. (2001), as detailed in section 3.2.2 below. Focusing on HEM, which is common to both H&E and IHC staining, thus enables us to make the network invariant to any staining component additional to HEM (eosin in H&E or any chromogen, such as DAB, in IHC) and thus to any IHC target. One challenging task for the method developed in this study is therefore to identify glandular structures using the HEM channel only.

The lower part of the network is inspired by DCAN and combines four channels that have their source at different information levels provided by the upper part. Each channel starts with a transposed convolution whose kernel size is chosen so that the spatial resolution of the resulting feature map corresponds to the spatial resolution of the original image. In DCAN the set of channels coming from the upper part are duplicated in order to output two different segmentation masks, one for objects and one for contours. In our architecture the four paths are first concatenated and a series of residual units then reduce the number of features (from 128 to 16). The network finally separates into two parts, both composed of a 1 by 1 convolution and a softmax function as activation. In the first channel, the gland edges are segmented while in the second, the complete glandular objects are segmented. As already shown with DCAN (Chen et al. (2017)), this double segmentation has the advantage of allowing a better separation of the individual objects. By comparison, the method of Xu

et al. (2017b) uses different networks that are trained independently for different segmentation tasks before being combined using an additional network. Our architecture has the advantage to be common for the two segmentation tasks up to the output layers and thus computationally thrifty. In addition, during training, the errors in edge segmentation impact the common network parameters and thus contribute to improving object segmentation (see section 3.2.3).

3.2. Training

3.2.1. Optimisation algorithm

Training is done using Adam (Adaptive Moment Estimation, proposed by Kingma and Ba (2014)) as optimiser with a learning rate set to 0.0003 and mini-batches of size 2. We kept the default values proposed by the authors for the other parameters (i.e., $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$). The mini-batch size resulted from hardware limitations whereas the learning rate was chosen empirically. The configuration of the computer we used is given in Appendix E.

3.2.2. Realistic data augmentation

For training, we heavily rely on data augmentation for two reasons. First because only very small data sets are available for our application. Second because it enables us to develop a robust and generalisable method with respect to different staining and acquisition conditions. To this end, we propose a data augmentation methodology able to mimic the physical reality encountered in the specific context of our application. This method is divided into 4 stages involving random crops, deformations at the tissue level (using rotations and elastic transformations) and transformations at both the staining and the acquisition levels. We first crop a random image area of 480×480 pixels as our network is optimised for this input size. To transform the staining at colour and intensity levels, we mimic the staining variations that we previously observed and quantitatively characterised between different IHC staining batches in a series of IHC experiments targeting five different proteins expressed in different tissue types (Van Eycke et al. (2017)). These variations were characterized at the level of the colour vectors representative of the standard IHC stains (i.e. DAB and HEM). An adaptive colour decomposition method was used to extract the colour vectors characterizing each IHC batch. The inter-batch variations concerned both the vector directions and the distribution of the OD values. In terms of the HEM vector direction, we observed small variations similar to those shown in Figure 2a. After matching the colour vectors representative of different IHC batches, we also evidenced nonlinear deformations of the deconvoluted OD value distribution in the realistic case where different lots of reagents were used between the IHC batches. We were able to correct these deformations using B-spline regression on quantile-quantile plots (Van Eycke et al. (2017)). Similar variations were reported for the colour vectors representative of the H&E staining (e.g. see Macenko et al. (2009); Khan et al. (2014)). Our colour augmentation method aims to simulate all these variations. In fact,

the proposed method reverses the normalisation technique that we developed to correct staining variations (Van Eycke et al. (2017)). Our approach is thus based on alterations of the colour vectors extracted by colour deconvolution with the aim to introduce realistic variations in terms of staining. As our network only uses the HEM channel, we only process this one to alter the colour and the intensity of this stain.

In the present study, the original colour vectors are defined according to Ruifrok et al. (2001). Two specific deconvolution matrices are used for H&E and IHC staining (see Appendix A). The first two transposed vectors (i.e. lines in the matrices) are the specific colour vectors of the staining, the first one corresponding to HEM in both cases and the second to eosin (A.1) or DAB (A.2), respectively. The third vector is the cross-product of the two others. Unlike what we previously developed for image normalization (Van Eycke et al. (2017)), a perfect matching between the used colour vectors and the colours of the processed image is not necessary since they will be modified by our colour augmentation methodology.

To generate modified HEM vectors, we pick two angles: the first one, g, according to a Gaussian distribution centered on the origin (with $\sigma = \frac{\pi}{180}$ rad) and the second one, u, according to an uniform distribution in $[0, 2\pi[$ rad. The original HEM vector is first rotated about the origin through angle g in an arbitrarily set direction and then through angle u about the original HEM vector (see Figure 2a). The OD values in the HEM channel are also modified as follows. According to a uniform distribution in the interval $[2, 15]$, we select a random number of quantiles from the observed OD distribution in the original HEM channel. A Gaussian noise ($\sigma = 0.2$) is added to each of these selected values which are then multiplied by a common factor chosen randomly from another Gaussian distribution ($\mu = 1$, $\sigma = 0.03$). All the OD values are then sorted and considered as the quantiles of a new OD distribution (see Figure 2b). We then map the initial quantile distributions to the new one using B-spline regression that we then apply on all the pixel OD of the image (Van Eycke et al. (2017)). All the parameter values were experimentally chosen, with visual controls of the results, to mimic the variations previously observed (Van Eycke et al. (2017)).

Elastic deformations followed by random rotations are also used to produce morphological variations that aim to mimic the deformations observed in the glandular structures of colorectal cancers. To this end, we create a grid with a random number of meshes according to a uniform distribution between $[1, 10[$. We then distort each mesh randomly around its origin according to a 2D Gaussian distribution ($\sigma = 15$ pixels in each direction). The pixels are then mapped according to the grid using cubic splines for the interpolation of the coordinates. The tissue deformation so implemented is similar to the one proposed for U-Net (Ronneberger et al. (2015)) except that we randomly set the number of meshes in the grid.

To simulate different acquisition conditions, the exposure and temperature values of the images are randomly changed and image blurring is introduced. New values are chosen using Gaussian distributions centered on the origin ($\sigma_{\text{exposure}} = 0.1$, $\sigma_{\text{white balance}} = 100$ Kelvin, $\sigma_{\text{Gaussian blur}} = 1$).

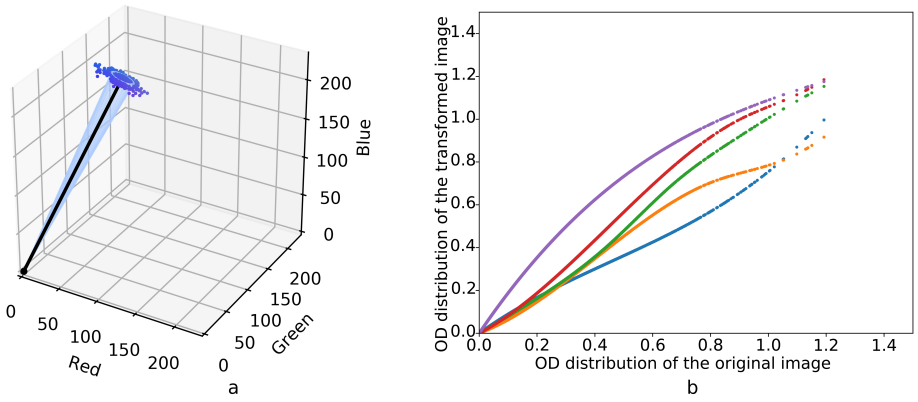

Figure 2: Colour augmentation: a) An example of the HEM colour vectors automatically generated using our data augmentation algorithm. The initial HEM vector is in black and the blue dots show the generated vector extremities. The pale blue cone represents the standard deviation of the random sampling. b) Quantile-Quantile plots exemplify the modifications done on the optical density (OD) distribution in the HEM channel. The X-axis represents the OD distribution of the original image and the Y-axis the transformed image one. The coloured lines show different distribution modifications generated from the same image.

Fig. 3 illustrates the impact of these image transformations including some strong geometric distortions to ease visualization. Since the elastic deformations involve random perturbations based on a Gaussian distribution, most of the images generated by our algorithm exhibit smaller distortions. It should also be noted that a new transformed version of the training set is generated for each training epoch.

3.2.3. Cost function

The used cost function corresponds to a modified version of the cross-entropy between the network and the desired outputs. As introduced in the U-net method, a term is added to increase the contribution of edge touches. In the present work, we propose a simplified version of this function to allow a faster calculation (see eq. 1–7).

$$c_{tot} = \frac{c_a + c_b}{2} \tag{1}$$

$$c_a = -\sum_i w_a(\mathbf{Y}_{ai}) \log(\hat{\mathbf{Y}}_{ai}) \quad \text{with any } \hat{\mathbf{Y}}_{ai} < 10^{-10} \text{ clipped to } 10^{-10} \tag{2}$$

$$c_b = -\sum_i w_b(\mathbf{Y}_{bi}) \log(\hat{\mathbf{Y}}_{bi}) \quad \text{with any } \hat{\mathbf{Y}}_{bi} < 10^{-10} \text{ clipped to } 10^{-10} \tag{3}$$

$$w_a(\mathbf{Y}_{ai}) = w_c(\mathbf{Y}_{ai}) + w_d(\mathbf{Y}_{ai}) \tag{4}$$

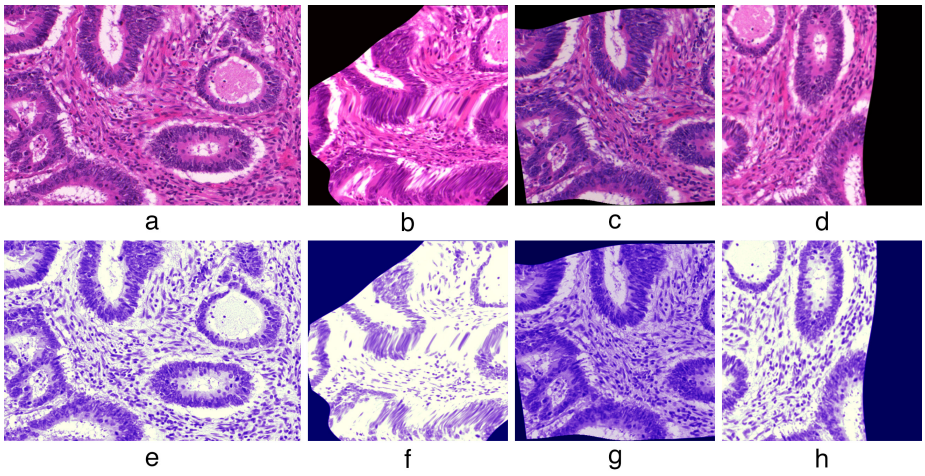

Figure 3: Data augmentation illustration on H&E images: a) the original image. b–d) Examples of transformation applied on the original image. e–h) Images showing the corresponding HEM channel used by our method. Colours were remapped to RGB for visualization purpose. To ease the visualization, images showing large distortions were selected.

$$w_b(\mathbf{Y}_{bi}) = \begin{bmatrix} \frac{y_{bi0}}{2(1-p)} & \frac{y_{bi1}}{2(p)} \end{bmatrix} \quad \text{with } p \text{ clipped into the interval } [0.01, 0.99] \tag{5}$$

$$w_c(\mathbf{Y}_{ai}) = \begin{bmatrix} \frac{y_{ai0}}{2(1-p)} & \frac{y_{ai1}}{2(p)} \end{bmatrix} \tag{6}$$

$$w_d(\mathbf{Y}_{ai}) = \begin{cases} \begin{bmatrix} \frac{-k}{2(1-p)} & \frac{k}{2(p)} \end{bmatrix} & \text{If pixel } i \text{ belongs to class 0 and is at less than } l \text{ pixels from pixels belonging to class 1} \\ \begin{bmatrix} 0 & 0 \end{bmatrix} & \text{Otherwise} \end{cases} \tag{7}$$

Where c_{tot} is the average of two cost functions, labelled c_a and c_b, which are weighted cross-entropy criteria with respect to the different targeted objects, i.e. the glands and the gland edges, respectively. For c_a class 1 indicates the glandular epithelium and class 0 is the rest of the image. $\hat{\mathbf{Y}}_{ai}$ and $\mathbf{Y}_{ai}$ are 1×2 matrices. It should be noted that we use indices 0 and 1 to indicate the matrix elements in direct relation to the class membership. $\mathbf{Y}_{ai}$ indicates the true label for pixel i using the conventional binary coding, i.e. (1,0) for class 0 and (0,1) for class 1, whereas $\hat{\mathbf{Y}}_{ai}$ specifies the class prediction made by the network for pixel i, with the two elements representing the probabilities to belong to class 0 and class 1, respectively, and summing to 1 thanks to the softmax function. p is the percentage of the image pixels classified in class 1 by the network. Equation 7 implements touching objects regularisation, where k and l are constants chosen experimentally as $k = 4$ and $l = 16$. l corresponds to the resolution difference

between the shallowest and the deepest layers of the network (480×480 and 30×30 in our case). This expression increases the weight of the pixels close to the edges of the glands and of the less represented class. For c_b class 1 is the gland edges and class 0 is the rest of the image. True edge labels are computed by applying morphological operations on the true gland label image comparable to what has been done for the DCAN architecture (Chen et al. (2017)) but with a larger disk radius, since the resolution of the deepest layers of the upper part of our network is lower than for DCAN. The edges are 16 pixels large. $\hat{\mathbf{Y}}_{bi}$ and $\mathbf{Y}_{bi}$ are similarly defined as $\hat{\mathbf{Y}}_{ai}$ and $\mathbf{Y}_{ai}$ but for the edge objects.

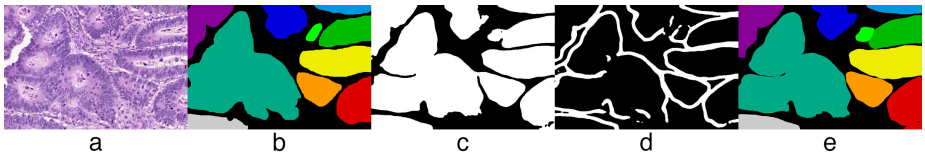

Figure 4: Gland segmentation steps: a) the original image, b) the annotations from the expert, c) the thresholded output of the channel of our network detecting the glands where arrows identify defects in gland division, d) the thresholded output of the other channel of our network detecting the edges, e) the final labels after post-processing with improvements in gland boundaries compared to c).

3.2.4. Output combination and post-processing

If the input image is larger than 480x480 pixels, it is cut into overlapping tiles of 480x480 pixels, which are submitted to the network. The outputs of each tile are then merged to recreate an output image of the same size than the input one. For the overlapping part, we assign the value of the corresponding pixel in the tile with the nearest center.

A threshold of 0.5 is applied to the two outputs of the network to determine the class membership of each pixel (see Fig. 4c–d). The thresholded edge output is subtracted to the thresholded gland output to improve the division of the touching glands. Holes are filled using morphological operations. Each object is then numbered and this unique value is used to label each pixel of the object. The obtained segmentation mask is padded by 32 pixels on each side so that its new padded border is the reflection of the object close to the image border. This helps to mitigate problems caused by glands that are cut by the image border. A morphological dilation is applied to each label to compensate the edge subtraction. Finally, the objects that are too small to be a gland (i.e. size < 2048 pixels, experimentally chosen and corresponding to about 45 µm of diameter at the image resolution used in the GlaS challenge) are removed. The padding is finally removed. Figure 4e illustrates the final result showing improvements in gland segmentation for some touching objects, whereas the network output detecting the glands already provides a very good result (Fig. 4c), probably due to the combined training targeting both the glands and the gland edges. In section 4.2 we evaluate the contribution of the edge output in the network performance and investigate the possibility to optimise the output combination

using an additional layer or network.

4. Evaluation methodology and results

4.1. Comparison with the state of the art on H&E images

We used the data and the evaluation methodology used in the GlaS segmentation challenge to compare our method to those which competed in the contest (Sirinukunwattana et al. (2017)). The data consist of a training set (n=85), labelled WQUD, and two test sets labelled Part A (n=60, composed of normal and tumour areas) and Part B (n=20, with a large majority of tumour areas). These sets are made of images of various sizes (from 567x430 to 775x522 pixels) but with a common resolution of 0.620 µm.

For the performance evaluation, three different scores are computed as previously detailed (Sirinukunwattana et al. (2017)). Briefly, the standard F1 score (combining precision and recall) is adapted to compute the detection accuracy for each individual gland. To be considered as "true/false positive" a segmented object has to overlap a ground truth annotation with at least/less than 50% of areas. The Object Dice index reflects the segmentation accuracy. This index combines an evaluation of how well each ground truth object is overlapped by a segmented object and of how well each segmented object is overlapped by a ground truth object. The Object Hausdorff score is based on the Hausdorff distance, which is the average distance for each pixel of an object with the closest pixel of another object. The distance is computed at the object level, similarly to the Object Dice index, and evaluates boundary correspondence between the ground truth and segmented objects. High values (i.e. near to 1) for the F1 score and the object Dice index indicate high accuracy and inversely for the object Hausdorff score.

The results are detailed in Table 1, which includes the best challenge results provided by Sirinukunwattana et al. (2017) and completed by new results recently published in the literature by Xu et al. (2017b). After ranking each method with respect to the others for a specific criterion, the rank-sum is computed to characterise each method. The partial rank-sums obtained on the two parts of the test set are also provided because Part B, which includes a majority of tumour areas where the gland architecture may be strongly deformed, is more difficult. Alternatively, the weighted rank-sum weights the ranks of each criterion according to the number of images in datasets A and B (i.e. using 3/4 and 1/4, respectively), as also used by Xu et al. (2017b). Our method comes first for all the rank-sums, whereas it uses only the HEM channel of the images, contrasting with all the other methods that use multiple colour channels. Compared to DCAN (labeled CUMedVision2 in Table 1), our method is more accurate on all the criteria except a slight decrease in the F1 score on Part A. The improvements on Part B are substantial as shown by the rank-sum computed on this specific dataset only, i.e. 20 for DCAN and 4 for our method. With regard to the results obtained by Xu et al. (2017b), our far simpler network produces similar results with an advantage for the Hausdorff scores showing an improvement of 10% on

Part B. In the next section we analyse the contributions provided by different components of our approach, such as the data augmentation strategy and some network features. Figure 5 qualitatively illustrates the results obtained with our method. In agreement with the quantitative results in Table 1, this figure shows the accurate segmentation obtained on different images including objects with very diversified aspects, except in the 3rd line where several complex and partially visible objects are erroneously merged.

Method	F1 score				Object Dice				Object Hausdorff				Rk sum			Weighted Rk Sum[3]
	Part A		Part B		Part A		Part B		Part A		Part B		Part	Part	Part	
	Score	Rk	Score	Rk	Score	Rk	Score	Rk	Score	Rk	Score	Rk	A	B	A&B	
Proposed	0.895	3	0.788	2	0.902	2	0.841	**1**	42.94	**1**	105.93	**1**	6	4	10	5.5
Xu et al. (2017b)	0.893	4	0.843	**1**	0.908	**1**	0.833	2	44.13	2	116.82	2	7	5	12	6.5
CUMedVision2[1]	0.912	**1**	0.716	5	0.897	3	0.781	7	45.42	3	160.35	8	7	20	27	10.25
ExB3	0.896	2	0.719	4	0.886	4	0.765	8	57.36	7	159.87	7	13	19	32	14.5
ExB1	0.891	6	0.703	6	0.882	6	0.786	4	57.41	8	145.58	3	20	13	33	18.25
Freiburg2[2]	0.870	7	0.695	7	0.876	7	0.786	4	57.09	5	148.47	5	19	16	35	18.25
CUMedVision1[1]	0.868	8	0.769	3	0.867	9	0.800	3	74.60	9	153.65	6	26	12	38	22.50
ExB2	0.892	5	0.686	8	0.884	5	0.754	9	54.79	4	187.44	10	14	27	41	17.25
Freiburg1[2]	0.834	9	0.605	9	0.875	8	0.783	6	57.19	6	146.61	4	23	19	42	22.00
CVML	0.652	11	0.541	10	0.644	12	0.654	10	155.43	12	176.24	9	35	29	64	33.50
LIB	0.777	10	0.306	12	0.781	10	0.617	11	112.71	11	190.45	11	31	34	65	31.75
vision4GlaS	0.635	12	0.527	11	0.737	11	0.610	12	107.49	10	210.10	12	33	35	68	33.50

Table 1: Performance on the GlaS challenge images in comparison with the methods obtaining the best results (Sirinukunwattana et al. (2017)), including on top one a more recent study in addition to ours. [1]Chen et al. (2017), [2]Ronneberger et al. (2015),[3]Xu et al. (2017b).

It should be noted that the GlaS challenge focused attention on individual gland object segmentation in order to provide relevant diagnostic information based on gland size, shape, etc. Consequently, all the evaluation criteria provided in Table 1 are object-based variants of standard pixel-based criteria (Sirinukunwattana et al. (2017)). Our aim is quite different and more specifically targets pixel membership to gland object. Indeed, the objective is to delineate the glandular epithelium area in order to quantify the expression of an IHC biomarker in (or out of) this specific histological component without requiring the separate identification of each gland. It is thus interesting to note that our method obtains very good performances in terms of standard F1 and Dice scores computed for pixel membership to gland class. Indeed, the F1 scores are 0.922 for Part A and 0.912 for Part B and the Dice scores are 0.921 and 0.878, respectively.

4.2. Contributions of different features of the proposed methodology

We conducted additional experiments on the GlaS challenge images to assess the contributions of different features and reported all the results in Table 2.

First, we clarified the contribution of the "edge" output of our network after training. To this end, we only used the "object" output and its post-processing as detailed in section 3.2.4 (but without output combination) to segment the glands in the H&E test set. The results evidence a substantial decrease in performance for each criterion that results in strongly higher rank-sums as compared to those of our initial method. The "edge" output thus contributes to segmentation accuracy.

Second, we evaluated the contribution of the different components of our augmentation strategy. For this purpose, we distinguished the augmentation

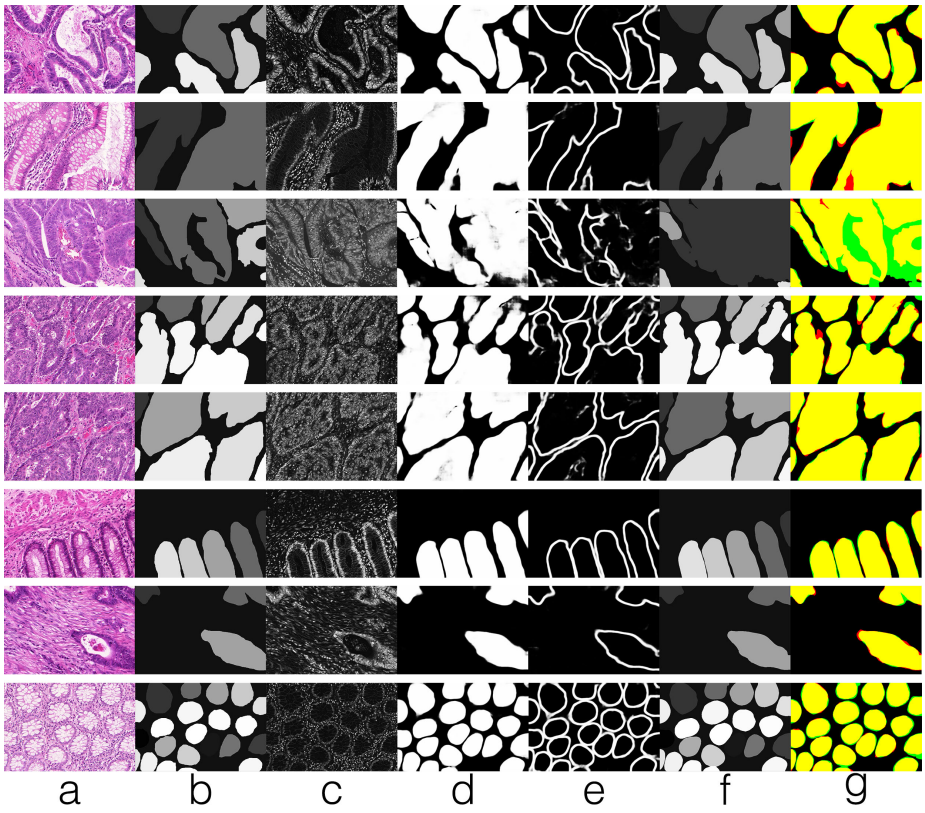

Figure 5: Segmentation results on H&E images from the GlaS challenge (Sirinukunwattana et al. (2017)): a) The original image, b) the annotation from the expert, c) the deconvoluted input of our network where the values are mapped between 0 and 255 and brightness and contrast are enhanced for visualisation purpose. d-e) The two outputs of our network (the values are mapped between 0 and 255 for visualisation purpose). f) The segmentation finally produced with our method, g) the overlap of the expert annotation (red) and the produced segmentation (green). Overlapping regions are in yellow.

method related to either acquisition, staining or morphology characteristics. We then trained our network with augmented image sets of the same size but in which one component was systematically omitted to increase the data. To further evaluate our augmentation strategy we also tested simplified versions concerning morphology and colour. For morphology we implemented simpler deformations consisting of horizontal and vertical symmetries, random translations (according to a 2D Gaussian distribution with $\sigma = 15$ pixels in each direction) and random rotations (using uniform distribution between 0 and π). Concerning colour, the simpler method consisted of random modifications of the hue and saturation of the image, according to a Gaussian distribution ($\sigma = 0.05$) for each parameter. We did not modify the intensity because it was already modified in the acquisition transform. The results show that the omission of each

component strongly impacts the performances with the largest effect observed for the morphological deformations using elastic transformations. The simplified augmentation methods also degrade performance, this time with greater impact for simplified colour augmentation. These data evidence the usefulness of the complete data augmentation strategy that we propose. Interestingly, the comparison with the DCAN results in Table 1 evidences that all the variants of our method remain more accurate on Part B except if elastic deformations, which are also used by DCAN (Chen et al. (2017)), are not included in the augmentation strategy. In contrast, only our complete approach is able to compete advantageously with the more complex architecture used by Xu et al. (2017b). These latter authors used elastic deformations for data augmentation and also image normalisation by performing per channel zero mean.

Method	F1 score				Object Dice				Object Hausdorff				Rk sum			Weighted Rk Sum
	Part A		Part B		Part A		Part B		Part A		Part B		Part A	Part B	Part A&B	
	Score	Rk	Score	Rk	Score	Rk	Score	Rk	Score	Rk	Score	Rk				
Proposed	0.895	2	0.788	3	0.902	3	0.841	**1**	42.943	2	105.926	**1**	7	5	12	6.5
Xu et al. (2017b)	0.893	4	0.843	**1**	0.908	**1**	0.833	2	44.129	4	116.821	3	9	6	15	8.25
Proposed w/o edge output	0.869	8	0.775	4	0.883	9	0.813	7	60.539	9	137.808	8	26	19	45	24.25
Proposed w/o acquisition aug.	0.888	6	0.735	9	0.892	7	0.813	6	48.972	6	118.851	4	19	19	38	19
Proposed w/o colour aug.	0.867	9	0.757	7	0.885	8	0.793	9	51.711	7	128.290	6	24	22	46	23.5
Proposed w/o morphology aug.	0.759	10	0.728	10	0.763	10	0.769	10	131.115	10	154.244	10	30	30	60	30
Proposed with simpler colour aug.	0.890	5	0.771	5	0.895	5	0.823	5	43.066	3	133.055	7	13	17	30	14
Proposed with simpler morphology aug.	0.895	3	0.754	8	0.903	2	0.826	4	38.572	**1**	126.123	5	6	17	23	8.75
Proposed with 1x1 convolution for fusion	0.885	7	0.794	2	0.893	6	0.833	3	57.872	8	116.720	2	21	7	28	17.5
Best scores from the original challenge	0.912	**1**	0.769	6	0.897	4	0.8	8	45.418	5	145.575	9	10	23	33	13.25

Table 2: Contribution of different features of the proposed approach. Performances on the GlaS challenge images when removing the contribution of the "edge" output, removing or simplifying a data augmentation component, or when adding a fusion layer to combine the outputs. The last line exhibits the best result which was independently obtained for each score in the original GlaS challenge across all the methods.

Finally, we tried to optimize the processing of the network outputs in place of using the simple output combination described in section 3.2.4. To this end, we added at the bottom of our trained network a 1x1 convolution layer to combine the two raw outputs, i.e. without thresholding. This layer was subsequently trained with the cost function defined by equation 2. Post-processing of the final output was finally applied as described in section 3.2.4, except the steps which directly concern the output combination (i.e., subtraction and dilatation). As shown in Table 2, a result deterioration can be observed when comparing this latter methodology with our initial approach, ranking the new results between the best ones and those obtained when the edge output is not considered. These data suggest that the last 1x1 layer did not efficiently learn how to combine the two outputs of our network and is not able to outclass the logical output combination that we propose in section 3.2.4. We also tested the method proposed by Xu et al. (2017b), which consists in using a 7-layer "fusion" network to combine the two network outputs, without obtaining more satisfactory results. It should be noted that Xu et al. (2017b) observed a degradation of the performances

of their method when the third output (targeting a bounding box around each gland object) was not considered, as it was the case in our experiments.

All the previous data evidence that the different features that we investigated positively contribute to colonic gland segmentation accuracy and justify their use for the next experiments on the IHC images.

4.3. Evaluation on IHC images

Since our method uses only the HEM channel, it is able to work on any type of staining that uses HEM as tissue counterstaining, especially IHC regardless of the targeted antigen. To evaluate this ability, we built new datasets based on colorectal tissue microarray (TMA) slices on which three different biomarkers were revealed by means of IHC. The three targeted proteins are Bcl-2-associated X protein (BAX), insulin-like growth factor binding protein 2 (IGFBP2) and α-smooth muscle actin (α-SMA). They have been chosen because they are expressed by different cell types from the glandular and/or stromal compartments. BAX and IGFBP2 are mostly expressed in the glandular epithelium, while α-SMA is mostly present in the stroma. In addition, the staining was carried out using different automata and different HEM compounds for counterstaining (see Appendix B for details). The TMA slides were digitised using a Nanozoomer 2.0 HT slide scanner (Hamamatsu, Hamamatsu City, Japan), which is different from the one used for the GlaS challenge. From this material, we designed four datasets, two for BAX and one for each of the two other proteins. Each set consists of slightly more than 30 TMA cores of 600 µm in diameter, taken from normal and tumoural colorectal tissue samples. There is no intersection between the four core sets, meaning that the glands from one dataset are all different from those in the other sets. All the glands have been annotated by an expert using the NDP.View 2 Software (Hamamatsu) and a Cintiq 13 HD Tablet (Wacom, Portland, USA) for increased accuracy.

We tested our method on the IHC images in three different conditions of network training. The first two conditions are previous training with the H&E images without or with the addition of fine-tuning using a small IHC set; the third one is training from scratch with this small IHC set. Fine-tuning was investigated because of changes in expert annotations as well as deep changes in the tissue processing and morphology, such as TMA core vs standard tissue slice images, the presence of tissue tearing in the IHC images unlike that of the GlaS dataset, etc. The small IHC set used for fine-tuning or training from scratch is the first BAX dataset (counting 43 cores) that we artificially augmented (Figure 6). The other three independent sets were used for testing (see Appendix C for details). All the images were used with a 10× magnification (resolution of 0.904 µm) in order to be close, without being equal, to the GlaS challenge image resolution (0.620 µm). At each training epoch a 480x480 image was randomly cropped from each core of the training set and submitted to full augmentation.

According to the above-mentioned specificity of our application, we computed pixel-based F1 and Dice scores to evaluate segmentation accuracy. The results detailed in Table 3 indicate that the network pre-trained with the H&E images and fine-tuned with a small IHC set provides the best performances

with few variability across the 3 test sets. In addition, these data exhibit no advantage for the marker used in the training set, i.e. BAX, and thus evidence a good ability to generalise across different experimental conditions, including different IHC biomarkers with different expression levels as quantified below. These results also evidence the substantial impact of the fine-tuning step on the network pre-trained with the H&E images. The network trained from scratch on IHC images also exhibits very good performances.

Most of the differences with the expert annotations concern lumen and possible intraluminal secretions occurring in tumour glands. Indeed, while our expert usually included these areas in their annotations, our algorithm excluded them as illustrated in Figure D.1 in Appendix. In doing so, our algorithm is more accurate in epithelium segmentation than the expert is. This precision is beneficial for IHC staining quantification because secretions can introduce artefacts in staining or can erroneously be considered as tissue area and so biases the computation of quantitative staining features. To decrease the bias due to the expert annotations, we recomputed the scores of the best network by excluding the "empty" areas (i.e. without tissue) from both the expert and the automated annotations. The results added in Table 3 (see the first line) show slight performance increases for BAX and IGFBP2 resulting in a clear decrease in the score variability across the 3 markers (see STD values). We can conclude that the expert annotations mainly contribute to the results variability across the markers.

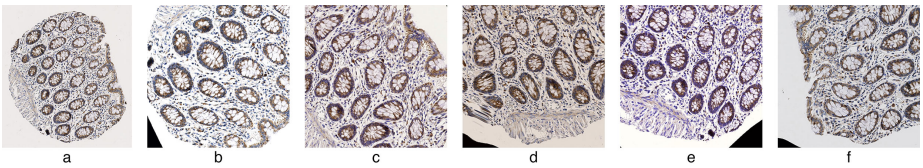

Figure 6: Data augmentation illustration on IHC images: a) Original image. b–f) Examples of transformation applied on the original image.

	F1 Pix						Dice Pix						STD	STD	Rk
	BAX	Rk	IGFBP2	Rk	a-SMA	Rk	BAX	Rk	IGFBP2	Rk	a-SMA	Rk	F1 Pix	Dice Pix	Sum
Pretrained with fine-tuning w/o empty areas	0.844		0.842		0.857		0.909		0.906		0.912		0.008	0.003	
Pretrained with fine-tuning	0.825	1	0.830	1	0.863	1	0.895	1	0.895	1	0.914	1	0.021	0.011	6
Trained from scratch on IHC only	0.803	2	0.827	2	0.858	2	0.882	2	0.891	2	0.909	2	0.027	0.014	12
Pretrained w/o fine-tuning	0.684	3	0.684	3	0.716	3	0.804	3	0.793	3	0.817	3	0.018	0.012	18

Table 3: Segmentation scores obtained on the IHC test sets with our CNN trained in different conditions. STD = standard deviation computed on the 3 score values.

Figure 7 shows the concordance obtained for the compartmentalised IHC staining evaluation computed on the basis of either the expert or the automatic annotations. The measure used for this evaluation is the labelling index, which is the percentage of immunoreactive tissue area computed in and/or out the epithelial compartment. This evaluation requires identifying the immunoreactive

(i.e. DAB-stained) pixels and the tissue ones to compute the ratio of the two respective areas, as previously detailed in Van Eycke et al. (2017). Briefly, three segmentation parameters were manually set in the deconvoluted (HEM-DAB) plane, which was fine-tuned for each IHC biomarker. The tissue pixels have a distance to the origin larger than a small value arbitrarily chosen at 0.015. DAB staining consists of pixels for which their brown value exceeds both a given threshold and their blue value by a given factor. These two parameters were set by a pathologist for each IHC biomarker.

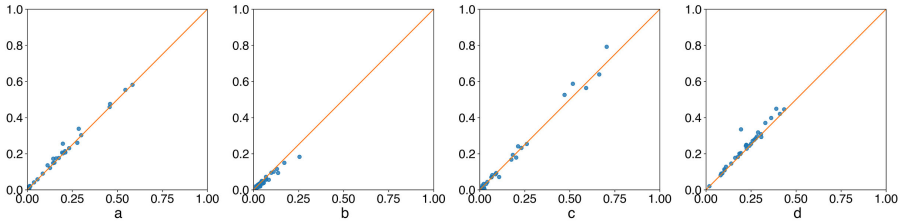

Figure 7: Concordance of quantitative IHC evaluations: Scatterplots of the labelling indices (i.e. percentages of immunoreactive tissue area) computed per core using either the expert (X axis) or machine annotations (Y axis). a) BAX marker evaluated in the epithelium (r_c=0.995). b) BAX marker outside the epithelium (r_c=0.932). c) IGFBP2 marker in the epithelium (r_c=0.993). d) α-SMA marker outside the epithelium (r_c=0.956). r_c is the concordance correlation coefficient (Lin (1989)).

The concordance levels between the two series of measurements were quantified using Lin's concordance correlation (r_c), for which values near 1 indicate strong fitting to the Y=X line and thus high concordance, where agreement occurring by chance is neutralised (Lin (1989)). The data illustrated in Figure 7 show high concordance level ($r_c > 0.90$) between the measurements while some lack of fit occurs in complex tumour tissue where the exact delimitation of the gland is not clear.

Additionally, we carried out qualitative tests on prostate tissue samples where different proteins were evidenced by IHC. Like in colon, prostate tissue includes glandular epithelium but with a different morphology. The IHC markers are also different from those in the colon datasets and include nuclear proteins. In Appendix Figure D.2 illustrates the promising results which were obtained with our network fine-tuned on the BAX colon image set. They could most likely be further improved by fine-tuning the network on annotated images of prostate tissue samples.

4.4. Convergence and computation time

With the configuration detailed in Appendix E, training with the WQUD dataset (from the GlaS challenge) was completed in a bit more than a day and took around 200,000 iterations to converge. Figure 8 shows the evolution of the c_{tot} criterion values with respect to the iteration numbers. It evidences the negative impact on the convergence of removing the skip connections in

the residual units (i.e. the shortcuts without any convolution). These results agree with those previously reported by Szegedy et al. (2017). Concerning the IHC experiments, the fine-tuning stage with the (first) BAX dataset required only 4,000 additional iterations, whereas training from scratch with this dataset took 20,000 iterations. These numbers of iterations were determined by using the second BAX dataset as validation set and applied to the IGFBP2 and α-SMA datasets.

Prediction time for an image of about 250,000 pixels without the overhead necessary to read the image was 0.09 second in average on GPU, with additional 0.43 second for the post-processing on CPU using scikit-image. Prior to that, 32 seconds were required, once for each prediction run, to implement the network on GPU using Tensorflow. In contrast, 4 to 5 hours was required for manually annotating the images of 75 TMA cores of 600 µm of diameter, i.e. an average of about 3 to 4 minutes per core against less than 1 second for automatic segmentation.

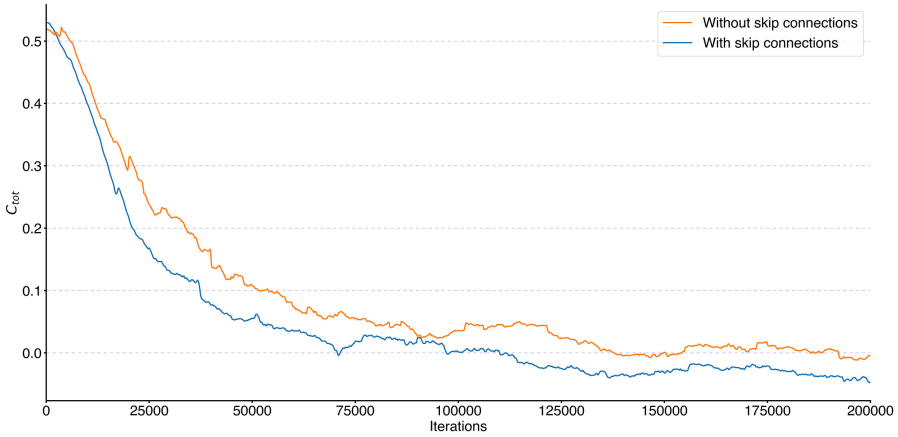

Figure 8: Network convergence on the H&E dataset: c_{tot} values were smoothed using a moving average to reduce noise. The negative values are due to regularisation for touching objects (see equation 7).

5. Discussion and conclusion

In the present study we propose a new deep learning method for epithelium segmentation in colorectal tissue slide images. We opted for a pixel-based method instead of patch classification, whereas this latter approach was also used for performing epithelium-stroma distinction (Kather et al. (2016); Ciompi et al. (2017); Huang et al. (2017)). One advantage of the patch-based approach is that labeling patches is far less heavy for experts than manually segmenting all the regions of interest present in an image. However, patch classification provides less precise results than pixel-based segmentation and this lack of precision

could strongly affect IHC quantification. To address this shortcoming, patches sliding across the image can be used but may strongly increase the processing time. In addition, working on small image patches limits deep learning ability to extract informative high-level features. It should be noted that in the top-ten results shown in Table 1, only the CVML method, which is ranked 10th, uses small (19x19 pixels) patches sliding across the image undergoing processing as input.

Our contribution consists in three major points. First, in the matter of robustness and generalisation, we propose a new method for increasing training data. The most novel aspect concerns the introduction of realistic variations affecting colour vectors extracted by colour deconvolution combined with changes in exposure and temperature values. Our experiments evidence that each component of our augmentation strategy contributes to the performances and that the complete augmentation strategy is required to be able to compete advantageously with the state-of-the-art. Our results suggest that this strategy successfully limits the effect of changes in tissue processing, staining and image acquisition, such as those encountered between the different datasets tested. By introducing blur and elastic transformations, our method also effectively limits the impact of image resolution changes.

Second, we developed a new convolutional network architecture, which integrates efficient properties proposed in previous works and some simplifications to avoid an excessive increase in the network size. Because DCAN achieved the best results in the GlaS challenge, we based our architecture on this network in which we introduced elements from two other networks (ResNet and U-net). In particular, the introduction of residual units enabled us to improve the training convergence of our deeper network in agreement with previous studies (He et al. (2016a); Szegedy et al. (2017)). Our results clearly point out the advantageous impacts of the architecture changes on gland segmentation accuracy. It should also be noted that during our developments we tried to include gradual upsampling like in U-net. However, this approach did not improve the segmentation performances while consuming a lot of memory. From our experiments, we can hypothesise that the good performances of U-net (but worse than DCAN in the GlaS challenge) would rather be due to concatenations between layers on either side of the network that actually act as skip connections like in the residual units. The last architecture modification that we investigated was to replace the output combination step that we initially proposed by a 1×1 convolution layer or the more complex fusion network proposed by Xu et al. (2017b), which were submitted to subsequent training. The deteriorated results obtained with those modifications evidence the efficiency of our simple but logical output combination step. A distinctive feature of our approach is that it makes use of the HEM channel only. In preliminary tests we did not observe significant variations in the GlaS challenge scores when the 3 deconvoluted channels were used, probably because of the hematoxylin that stains the cell nuclei, making the HEM channel sufficiently representative of the tissue structure. By using the HEM channel only, combined with strong data augmentation, our approach reduces the re-training time when a learning transfer is required. Indeed, we

illustrate application transfer from the H&E dataset to the IHC one, where numerous characteristics had changed, such as tissue processing, staining, acquisition device, and resolution, as well as the expert for annotation. However, this transfer only required 4,000 additional training iterations whereas about 200,000 iterations were necessary for the initial training. Due to our intensive data augmentation strategy, a quite similar accuracy can be obtained by training our network from scratch on the small IHC dataset used for fine-tuning. This latter result suggests that intensive data augmentation is able to compete with transfer learning.

Third, with regard to speed, our automatic segmentation method takes less than 3 minutes for processing 150 TMA cores, i.e. less than 1.2s per image of about 950×950 pixels. This is slightly faster than the speed reported for DCAN, i.e. about 1.5s for an image of 755×522 pixels (Chen et al. (2017)), whilst being more accurate. (No processing time is provided by Xu et al. (2017b).) These results allow the efficient analysis of a large amount of tissue within a reasonable time frame.

A potential limitation of our approach concerns the visibility of the cell nuclei in the HEM channel. This visibility requires sufficiently strong counterstaining in the case of nuclear DAB staining, in particular in the epithelial cells. Nevertheless, our additional tests on the prostate tissue show that the nuclear DAB staining evidencing ERG expression in numerous epithelial cells does not prevent gland segmentation. If the HEM channel exhibits clear defects in the detection of epithelial cell nuclei, image preprocessing could be applied to enrich the HEM channel with the positive cell nuclei identified in the DAB channel.

The present study focuses on colonic gland segmentation in normal and cancer tissue samples. Preliminary tests on epithelium segmentation in prostate tissue are encouraging. In future works our approach could be applied to other segmentation tasks on histological images (e.g., nuclei segmentation such as in the MICCAI 2017 challenge) or on other biomedical imaging modalities such as in Ronneberger et al. (2015).

In conclusion, we proposed a new network architecture and a new data augmentation algorithm. When coupled, we achieved state-of-the-art performances on the GlaS challenge datasets as well as high concordance in IHC staining quantification when comparing the data obtained on the basis of either the expert annotation or the automatic segmentation. Being based exclusively on the HEM channel, the method effectively processes IHC images regardless of the targeted antigen and without marker-specific training. Finally, the method is fast and able to process large batches of images in a reasonable amount of time.

6. Acknowledgment

This research is supported by the Fonds Yvonne Boël (Brussels, Belgium); the Fonds Erasme; and the European Regional Development Fund and the Walloon Region [Wallonia-Biomed grant 411132-957270]. The authors thank Martin Teller (LISA, ULB) for English writing support and Thomas Gevaert, MD (Laboratory of Experimental Urology, KU Leuven) for providing prostate tissue

images. C. Decaestecker is a senior research associate with the National (Belgian) Fund for Scientific Research (F.R.S.-FNRS). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

7. Author contributions

Y.-R.V.E. and C.D. conceived the study objectives. Y.-R.V.E., O.D. and C.D. conceived the methodology and the experiments. C.B., L.V. and I.S. contributed materials. Y.-R.V.E. implemented the methods and conducted the experiments. Y.-R.V.E and C.D. analyzed the data and wrote the manuscript. All authors reviewed the manuscript.

8. References

Altunbay, D., Cigir, C., Sokmensuer, C., Gunduz-Demir, C., mar 2010. Color graphs for automated cancer diagnosis and grading. IEEE Trans. Biomed. Eng. 57 (3), 665–74.

Ben-Shmuel, A., Shvab, A., Gavert, N., Brabletz, T., Ben-Ze'ev, A., jul 2013. Global analysis of L1-transcriptomes identified IGFBP-2 as a target of ezrin and NF-κB signaling that promotes colon cancer progression. Oncogene 32 (27), 3220–30.

CAMELYON16, 2016. ISBI Challenge on cancer metastasis detection in lymph node. `camelyon16.grand-challenge.org`, accessed: 2017-08-10.

CAMELYON17, 2017. CAMELYON17 challenge on automated detection and classification of breast cancer metastases in whole-slide images of histological lymph node sections. `camelyon17.grand-challenge.org`, accessed: 2017-12-12.

Chen, H., Qi, X., Yu, L., Dou, Q., Qin, J., Heng, P.-A., feb 2017. DCAN: Deep contour-aware networks for object instance segmentation from histology images. Med. Image Anal. 36, 135–146.

Ciompi, F., Geessink, O., Bejnordi, B. E., de Souza, G. S., Baidoshvili, A., Litjens, G., van Ginneken, B., Nagtegaal, I., van der Laak, J., apr 2017. The importance of stain normalization in colorectal tissue classification with convolutional networks. In: 2017 IEEE 14th Int. Symp. Biomed. Imaging (ISBI 2017). Vol. 42. IEEE, pp. 160–163.

Doyle, S., Hwang, M., Shah, K., Madabhushi, A., Feldman, M., Tomaszeweski, J., 2007. Automated grading of prostate cancer using architectural and textural image features. In: 2007 4th IEEE Int. Symp. Biomed. Imaging From Nano to Macro. IEEE, pp. 1284–1287.

GLaS, 2015. GlaS@MICCAI'2015: Gland Segmentation Challenge Contest. www2.warwick.ac.uk/fac/sci/dcs/research/tia/glascontest, accessed: 2017-08-10.

He, K., Zhang, X., Ren, S., Sun, J., 2016a. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778.

He, K., Zhang, X., Ren, S., Sun, J., 2016b. Identity mappings in deep residual networks. In: Proceedings of the European Conference on Computer Vision. Springer, pp. 630–645.

Huang, Y., Zheng, H., Liu, C., Ding, X., Rohde, G., apr 2017. Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images. IEEE J. Biomed. Heal. informatics.

Kather, J. N., Weis, C.-A., Bianconi, F., Melchers, S. M., Schad, L. R., Gaiser, T., Marx, A., Zöllner, F. G., jun 2016. Multi-class texture analysis in colorectal cancer histology. Sci. Rep. 6, 27988.

Khan, A. M., Rajpoot, N., Treanor, D., Magee, D., jun 2014. A Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using Image-Specific Color Deconvolution. IEEE Trans. Biomed. Eng. 61 (6), 1729–1738.

Kingma, D. P., Ba, J., dec 2014. Adam: A Method for Stochastic Optimization. Proc. 12th Annu. Conf. Genet. Evol. Comput. - GECCO '10, 103.

Kowalczyk, A. E., Krazinski, B. E., Godlewski, J., Kiewisz, J., Kwiatkowski, P., Sliwinska-Jewsiewicka, A., Kiezun, J., Sulik, M., Kmiec, Z., jul 2017. Expression of the EP300, TP53 and BAX genes in colorectal cancer: Correlations with clinicopathological parameters and survival. Oncol. Rep. 38 (1), 201–210.

Lafarge, M. W., Pluim, J. P., Eppenhof, K. A., Moeskops, P., Veta, M., 2017. Domain-adversarial neural networks to address the appearance variability of histopathology images. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 83–91.

Li, W.-W., Wang, H.-y., Nie, X., Liu, Y.-b., Han, M., Li, B.-H., jun 2017. Human colorectal cancer cells induce vascular smooth muscle cell apoptosis in an exocrine manner. Oncotarget.

Lin, L. I., mar 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45 (1), 255–68.

Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Xiaojun Guan, Schmitt, C., Thomas, N. E., jun 2009. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE 6th IEEE Int. Symp. Biomed. Imaging. IEEE, pp. 1107–1110.

Mishra, L., Bass, B., Ooi, B. S., Sidawy, A., Korman, L., dec 1998. Role of insulin-like growth factor-I (IGF-I) receptor, IGF-I, and IGF binding protein-2 in human colorectal cancers. Growth Horm. IGF Res. 8 (6), 473–9.

Mishra, R., Daescu, O., Leavey, P., Rakheja, D., Sengupta, A., 2017. Histopathological diagnosis for viable and non-viable tumor prediction for osteosarcoma using convolutional neural network. In: International Symposium on Bioinformatics Research and Applications. Springer, pp. 12–23.

Naik, S., Doyle, S., Agner, S., Madabhushi, A., Feldman, M., Tomaszewski, J., may 2008. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In: 2008 5th IEEE Int. Symp. Biomed. Imaging From Nano to Macro. IEEE, pp. 284–287.

Olgun, G., Sokmensuer, C., Gunduz-Demir, C., 2014. Local object patterns for the representation and classification of colon tissue images. IEEE J. Biomed. Heal. Informatics 18 (4), 1390–1396.

Rabinovich, A., Agarwal, S., Laris, C., Price, J. H., Belongie, S., 2004. Unsupervised Color Decomposition Of Histologically Stained Tissue Samples. In: Adv. Neural Inf. Process. Syst. 16. MIT Press, pp. 667—-674.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Med. Image Comput. Comput. Interv. - MICCAI 2015: Proceedings, Part III. Springer International Publishing, pp. 234–241.

Ruifrok, A. C., Johnston, D. A., et al., 2001. Quantification of histochemical staining by color deconvolution. Analytical and quantitative cytology and histology 23 (4), 291–299.

Simonyan, K., Zisserman, A., sep 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. Inf. Softw. Technol. 51 (4), 769–784.

Simsek, A. C., Tosun, A. B., Aykanat, C., Sokmensuer, C., Gunduz-Demir, C., 2012. Multilevel segmentation of histopathological images using cooccurrence of tissue objects. IEEE Trans. Biomed. Eng. 59 (6), 1681–1690.

Sirinukunwattana, K., Pluim, J. P., Chen, H., Qi, X., Heng, P.-A., Guo, Y. B., Wang, L. Y., Matuszewski, B. J., Bruni, E., Sanchez, U., Böhm, A., Ronneberger, O., Cheikh, B. B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead, D. R., Rajpoot, N. M., jan 2017. Gland segmentation in colon histology images: The glas challenge contest. Med. Image Anal. 35, 489–502.

Sirinukunwattana, K., Raza, S. E. A., Tsang, Y.-W., Snead, D. R., Cree, I. A., Rajpoot, N. M., 2016. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE transactions on medical imaging 35 (5), 1196–1206.

Sirinukunwattana, K., Snead, D. R. J., Rajpoot, N. M., nov 2015. A Stochastic Polygons Model for Glandular Structures in Colon Histology Images. IEEE Trans. Med. Imaging 34 (11), 2366–2378.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A. A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence. pp. 4278–4284.

Tosun, A. B., Gunduz-Demir, C., 2011. Graph run-length matrices for histopathological image segmentation. IEEE Trans. Med. Imaging 30 (3), 721–732.

TUPAC16, 2016. Tumor Proliferation Assessment Challenge 2016 | TUPAC16 | MICCAI Grand Challenge. `tupac.tue-image.nl`, accessed: 2017-12-12.

Van Eycke, Y.-R., Allard, J., Salmon, I., Debeir, O., Decaestecker, C., feb 2017. Image processing in digital pathology: an opportunity to solve inter-batch variability of immunohistochemical staining. Sci. Rep. 7, 42964.

Verset, L., Tommelein, J., Moles Lopez, X., Decaestecker, C., Boterberg, T., De Vlieghere, E., Salmon, I., Mareel, M., Bracke, M., De Wever, O., Demetter, P., sep 2015. Impact of neoadjuvant therapy on cancer-associated fibroblasts in rectal cancer. Radiother. Oncol. 116 (3), 449–454.

Verset, L., Tommelein, J., Moles Lopez, X., Decaestecker, C., Mareel, M., Bracke, M., Salmon, I., De Wever, O., Demetter, P., jul 2013. Epithelial expression of FHL2 is negatively associated with metastasis-free and overall survival in colorectal cancer. Br. J. Cancer 109 (1), 114–120.

Xu, Y., Jia, Z., Wang, L.-B., Ai, Y., Zhang, F., Lai, M., Chang, E. I.-C., dec 2017a. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinformatics 18 (1), 281.

Xu, Y., Li, Y., Wang, Y., Liu, M., Fan, Y., Lai, M., Chang, E., 2017b. Gland Instance Segmentation Using Deep Multichannel Neural Networks. IEEE Trans. Biomed. Eng. 64 (12), 2901–2912.

9. Appendices

A. Supplementary information on deconvolution

A.1 and A.2 are the matrices respectively used for the H&E and the IHC deconvolution. The vectors are expressed in optical density (OD), computed with a background intensity set to 235 in each channel, and are normalized.

$$\begin{bmatrix} 0.65 & 0.70 & 0.29 \\ 0.07 & 0.99 & 0.11 \\ -0.33 & -0.08 & 0.94 \end{bmatrix} \tag{A.1}$$

$$\begin{bmatrix} 0.65 & 0.70 & 0.29 \\ 0.27 & 0.57 & 0.78 \\ 0.63 & -0.71 & 0.30 \end{bmatrix} \tag{A.2}$$

B. Supplementary information on the IHC staining

Protein	Automata	Antibody brand	AB Concentration	Haematoxylin	Protein expression location
BAX	Leica BOND-MAX	Zymed	1/250	Haematoxylin included in the Leica Bond detection Kit (Leica Biosystem)	Mostly in epithelial cells and possibly also in endothelial cells Kowalczyk et al. (2017); Li et al. (2017)
IGFBP2	Ventana Discovery XT	Santa Cruz Biotechnology	1/150	Haematoxylin Gill's formula (Vector Laboratories)	Epithelial cells Ben-Shmuel et al. (2013); Mishra et al. (1998)
α-SMA	Leica BOND-MAX	Menarini	1/100	Haematoxylin included in the Leica Bond detection Kit (Leica Biosystem)	Stromal cells such as cancer-associated fibroblasts and vascular smooth muscle cells Verset et al. (2015)

Table B.1: IHC characteristics

C. Supplementary information on the IHC test sets

Protein	Core number	Percentage of tumour samples
BAX	34	68%
IGFBP2	37	46%
α-SMA	32	59%

Table C.1: Description of the three IHC test sets.

D. Supplementary information on the segmentation results

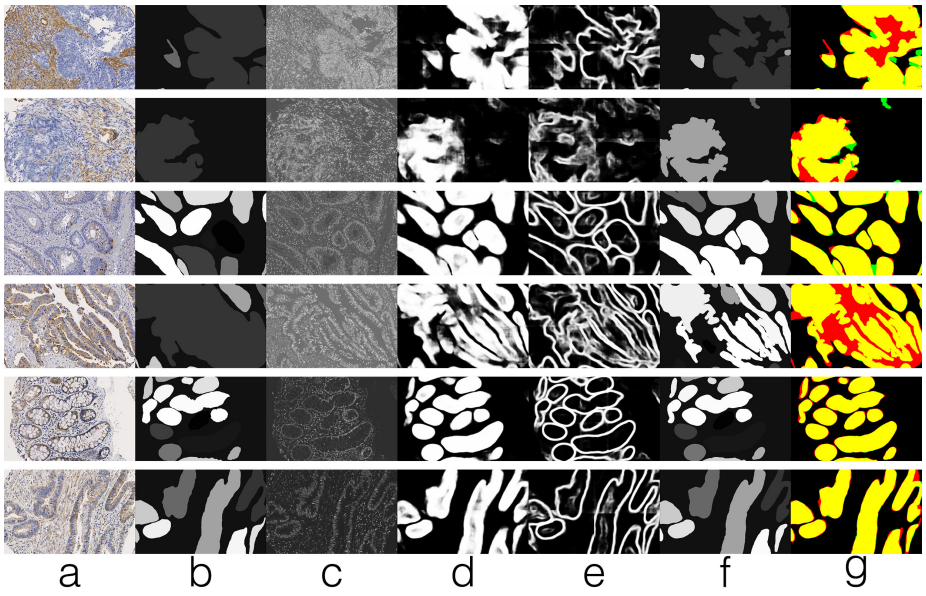

Figure D.1: Segmentation results for the IHC datasets: a) the original image, b) the annotation from the expert, c) the deconvolutated input feeded into our network (the values are mapped between 0 and 255 and brightness and contrast are enhanced for visualisation purpose). d-e) The two outputs of our network (where the values are mapped between 0 and 255 for visualisation purpose). f) The segmentation obtained with our method, g) the overlap of the expert annotation (red) and the produced segmentation (green). Overlapping regions are in yellow. There are two images per protein, from top to bottom α-SMA, IGFBP2, BAX.

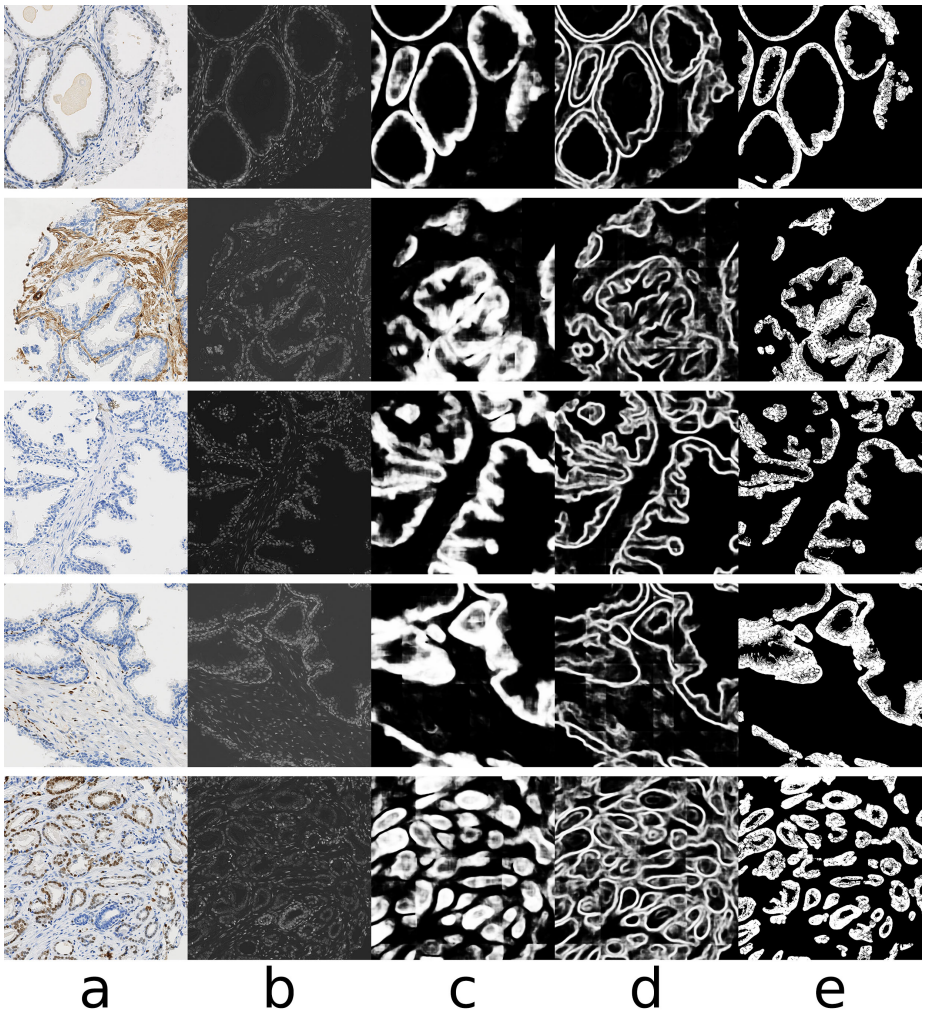

Figure D.2: Segmentation of IHC images from prostate tissue samples using the network fine-tuned on IHC images from the colon tissue: a) The original image, b) the deconvoluted input of our network where the values are mapped between 0 and 255 and brightness and contrast are enhanced for visualisation purpose. c-d) The two outputs of our network where the values are mapped between 0 and 255 for visualisation purpose. e) The segmentation obtained with our method where the areas without tissue have been removed to ease the comparison with a). Five different IHC markers are illustrated. From top to bottom: androgen receptor, caveolin 2, endoglin, estrogen receptor and ERG (protein encoded by the ETS-related gene). Steroid hormone (androgen and estrogen) receptors and ERG exhibit cell nucleus expression.

E. Supplementary information on the system used for our tests

Operating system	Windows 10 x64
Softwares and libraries	CUDA 8.0, CUDNN 5.1, Python 3.5.3, numpy 1.13.1, scikit-image 0.13, Tensorflow on gpu 1.2.1
CPU	Core I7-3930k
GPU	Nvidia Titan X (Pascal) with 12GB of GDDR5X
RAM	16GB
HDD	1TB@7200RPM

Table E.1: System configuration

10. Vitae

Yves-Rémi Van Eycke

Yves-Rémi Van Eycke graduated in 2013 from the Université Libre de Bruxelles (ULB, Belgium) with a Master's degree in Computer Science and is currently pursuing a Ph.D. in Engineering and Technology. He is also a researcher at DIAPath at CMMI (Gosselies, Belgium) where he develops image analysis and machine learning algorithms for the detection, characterisation, and validation of histological biomarkers.

Cédric Balsat

Cédric Balsat obtained his M.S. degree in Biomedical Sciences in 2007 from the Université catholique de Louvain (UCL, Belgium). He then received a Ph.D. in Biomedical and Pharmaceutical Sciences in 2014 from the Université de Liège (ULg, Belgium). His research was dedicated at characterizing the progression of cancers in human tissues using computer-assisted approaches. He is now the Lab Manager of DIAPath where he manages research projects aiming at the characterization and the validation of tissue-based biomarkers.

Laurine Verset

Dr. Laurine Verset obtained her M.D. (2009) and Ph.D. degrees (2016) from Université Libre de Bruxelles (ULB, Belgium). She is trained in Pathology and Cytopathology at the Pathology Department of the Erasme University Hospital (Brussels, Belgium). She is board certified in Pathology. Her research is dedicated to oncology, with a special focus on colorectal cancer. Her Ph.D. thesis (with Pr. P. Demetter as promotor) was carried out in the Pathology Department of the Erasme Hospital and focused on cancer-associated fibroblasts and the FHL2 protein in colorectal cancer. She is now appointed as a resident at this Pathology Department.

Olivier Debeir

Olivier Debeir obtained is M.S. degree in Applied Sciences in 1994 from the Université Libre de Bruxelles (ULB, Belgium), where he also received her Ph.D. in Applied Sciences in 2002. He is professor in image processing and biomedical imaging since 2009. He is responsible for the image unit of the Laboratories of Image, Signal processing and Acoustics at the Brussels School of Engineering (ULB) and the head of the Multimodal Imaging Processing unit of the CMMI (Gosselies, Belgium). His research topics are related to image processing and pattern recognition applied to various kind of imaging modalities.

Isabelle Salmon

Isabelle Salmon, M.D., Ph.D., is Professor of Anatomy Pathology at the Faculty of Medicine from the Université Libre de Bruxelles (ULB), Belgium. Since 1999 she is the head of the Surgical Pathology Department of the ULB Erasme Hospital. Since 2010, she coheads the DIAPath laboratory (Gosselies, Belgium) with Christine Decaestecker and since 2016 she is the Strategic Director of the Interregional University Centre, CurePath (Jumet, Belgium). Professor Salmon has expertise in the identification, characterization and validation of tissue-based biomarkers as well as in digital pathology. She recently developed in her department an integrated platform dedicated to digital pathology for second opinion analysis.

Christine Decaestecker

Christine Decaestecker obtained her M.S. degree in Mathematics in 1984 from the Université Libre de Bruxelles (ULB, Belgium), where she also received her Ph.D. in Pure Science in 1991 and her qualification for university professorship in 1997. She is a Senior Research Associate with the FNRS and director of the Laboratories of Image, Signal processing and Acoustics at the Brussels School of Engineering (ULB). Together with Isabelle Salmon, she coheads DIAPath in the CMMI (Gosselies, Belgium). She develops a multidisciplinary research in computational pathology, aiming at the characterization and validation of histological biomarkers involving image analysis, machine learning and biostatistics.