JMI Journal of Medical Imaging

RESEARCH PAPER

Stromal tissue segmentation in Ki67 histology images based on cytokeratin-19 stain translation

Montse Pardàs,[*,†] David Anglada-Rotger,[†] Maria Espina, Ferran Marqués, and Philippe Salembier

Universitat Politècnica de Catalunya BarcelonaTech Barcelona, Signal Theory and Communications Department, Barcelona, Spain

ABSTRACT. **Purpose:** The diagnosis and prognosis of breast cancer relies on histopathology image analysis. In this context, proliferation markers, especially Ki67, are increasingly important. The diagnosis using these markers is based on the quantification of proliferation, which implies the counting of Ki67 positive and negative tumoral cells in epithelial regions, thus excluding stromal cells. However, stromal cells are often very difficult to distinguish from negative tumoral cells in Ki67 images and often lead to errors when automatic analysis is used.

Approach: We study the use of automatic semantic segmentation based on convolutional neural networks (CNNs) to separate stromal and epithelial areas on Ki67 stained images. CNNs need to be accurately trained with extensive databases with associated ground truth. As such databases are not publicly available, we propose a method to produce them with minimal manual labelling effort. Inspired by the procedure used by pathologists, we have produced the database relying on knowledge transfer from cytokeratin-19 images to Ki67 using an image-to-image (I2I) translation network.

Results: The automatically produced stroma masks are manually corrected and used to train a CNN that predicts very accurate stroma masks for unseen Ki67 images. An *F*-score value of 0.87 is achieved. Examples of effect on the KI67 score show the importance of the stroma segmentation.

Conclusions: An I2I translation method has proved very useful for building ground-truth labeling in a task where manual labeling is unfeasible. With reduced correction effort, a dataset can be built to train neural networks for the difficult problem of separating epithelial regions from stroma in stained images where separation is very hard without additional information.

Keywords: histology image processing; Ki67; cytokeratin-19; deep learning; stromal tissue; semantic segmentation; generative adversarial network; fast contrastive unpaired translation; U-Net

Paper 22321GR received Nov. 26, 2022; revised May 9, 2023; accepted Jun. 5, 2023; published Jun. 23, 2023.

1 Introduction

Breast cancer is the most common type of cancer for women worldwide, and early detection and diagnosis are crucial for improving the survival rate. One of the primary methods for diagnosis is

*Address all correspondence to Montse Pardàs, montse.pardas@upc.edu

†These authors contributed equally to this work.

2329-4302/2023/$28.00 © 2023 SPIE

immunohistochemical (IHC) test of biopsies. In this context, pathologists first analyze the tissue obtained through the common hematoxylin-eosin (H&E) staining to detect tumoral areas and then apply additional stains for further classification of the tumor and for patient risk stratification. To predict prognosis and therapeutic response, the quantification of cell proliferation is usually required and can be assessed with the stain produced by the Ki67 biomarker. Ki67 is a monoclonal antibody that detects a nuclear antigen associated with mitosis in mammalian cells and is used to compute the tumor proliferation index.[1] This index is assessed by counting the percentage of positively stained tumoral cells over all the malignant cells (positive and negative). Usually, this involves manually counting between 500 and 1000 cells in three randomly selected high-power fields. Another option used by pathologists is to estimate by eyeballing the Ki67 Index, without formally counting. As expected, these methods, although very labor-intensive, result in important variability and low reproducibility depending on the selected zones and the used method.[2]

During the last decade, digital pathology is being deployed in an increasing number of pathology departments.[3] Digital pathology involves high-resolution digital images (whole-slide images or WSIs) obtained from biopsy samples captured with a scanning device. WSI can contain up to 40 Gb of uncompressed data and can substitute traditional light microscopes. At the same time, digital image analysis (DIA) techniques are emerging for automatic quantification of the most common stains (H&E, Ki67, ER, PR, and HER2 for breast cancer). Although cell segmentation and quantification results are usually quite accurate, in invasive breast cancer, tumor cells, stromal cells, and lymphocytes may be intricately mixed. For this reason, accurate quantification can only be achieved if pathologists outline the region of interest (ROI) excluding the stroma. This is quite difficult, depending on the structure of the tissue, and in some cases even not feasible without additional information supporting the distinction of epithelial regions from stroma. Furthermore, even if the tumoral region is outlined, it often includes stromal cells, as the outlining only provides external contours. In the particular case of Ki67, many authors point out the stromal cells misclassification errors as the main cause of errors in automatic quantification. Stromal cells are usually confused with negative Ki67 tumoral cells thus producing an incorrect proliferation index.

Furthermore, tumor–stroma ratio or spatial arrangement of stromal cells in tumors are also used as prognostic factors[4,5] thus making the distinction of stromal and epithelial tissues an important first step in digital pathology image analysis.

A common procedure used by pathologists to better detect tumoral regions is to compare the Ki67 sample with an adjacent tissue section stained with cytokeratin-19 (CK19). CK19 is an epithelial cell marker, which is expressed in more than 90% of breast tumors. The presence of CK19 helps identifying the tumoral cells, discarding the stroma, in those stains where the tumoral area is not clearly defined.

On the image analysis side, recent advances in computer vision based on convolutional neural networks (CNNs) produce outstanding results in semantic segmentation. Semantic segmentation identifies each pixel in an image as belonging to one of several predefined classes, thus generating a mask for each of these classes. In our case, the classes of interest are tumoral and stromal. However, CNN models, usually containing millions of parameters, need to be trained with an extensive labeled database. Usually, a huge task force is needed to create such a database since it requires the manual annotation of a large number of images. In this paper, we propose a methodology to create with a limited amount of manual annotation effort a database to train a CNN semantic segmentation model. In particular, we show how an image-to-image (I2I) translation model based on a generative adversarial network (GAN) can be trained to highlight information that, although present in the Ki67 images, is expressed in a subtle way. The GAN model captures this subtle expression through the observation of unpaired CK19 and Ki67 images. Note that, as previously mentioned, visual comparison of CK19 and Ki67 images is a common practice used by pathologists to help them localize epithelial area in Ki67 images. Once trained, the GAN model converts Ki67 images in what we call "fake-CK19," where epithelial regions are clearly distinguished (as in real CK19 images) and can be easily segmented by simple algorithms. We use the segmented fake-CK19 images as initial masks for the annotation of the Ki67 images. After manual correction when required, these masks are used as ground truth to train a stromal semantic segmentation CNN (e.g., in this work, based on the U-Net

architecture). Once trained, this CNN can be used to process arbitrary Ki67 images. Results for stroma segmentation on Ki67 images, evaluated through comparison with registered CK19 images, show the potential of this technique as a preprocessing step before proliferation index computation or for prognosis based on tumor–stroma analysis.

2 Related work

2.1 Automatic Ki67 Image Analysis

In March 2010, the "International Ki67 in Breast Cancer Working Group" agreed that IHC Ki67 measurement is a key point for tumor proliferation studies and developed guidelines for its analysis based on the computation of a Ki67 labeling index also named proliferation score. This score was defined as the percentage of positively stained cells among the total amount of evaluated tumor cells.[6] As previously mentioned, manual procedures to obtain this index are time-consuming and present a high variability, mainly due to limitations of the human eye and randomness in choice of the regions for cell counting. For this reason, and taking advantage of the continuous advances in computer vision, several works have focused on the automation of this process.

A recent study[7] has validated the usage of computer assisted image analysis for Ki67 stained images. It confirms that there is a significant benefit of automated image analysis as part of daily pathologists' workflow, both in the consistency of the automated results and in the time savings for pathologists. However, it also points out the unavailability of tumor/stroma segmentation tools. Although manual ROI annotation was used for the study, discrepancies among pathologists produced different interpretations of some cells as stroma or negative cells. The work of Ref. 8 also shows the importance of the ROI definition in Ki67 quantification when using computer assisted image analysis. The study compares commercial applications supporting semiautomated Ki67 quantification, many of which rely on measurements in user-defined ROIs. It was observed that results depend on the size of the ROI and that a common rejection cause of the software results was due to the confusion between tumor and stroma cells. This caused a rejection of 23% of the samples. In Ref. 9, it was also observed that although computer assisted image analysis has the advantage of measuring a much larger number of cells, it is less accurate than humans for stromal/inflammatory cells identification.

The most common approach taken for DIA systems is to rely on ROIs defined by the user in order to avoid stromal areas. For instance, in Ref. 10, an automatic approach for Ki67 index estimation is presented. The process is applied to hot-spot regions (area of high density of positive tumor cells for Ki67) where stromal cells are not observed. The system relies on color processing techniques to segment nuclei, which are then classified as "positive" or "negative" based on color and shape features. A recent work[11] proposes a pipeline for accurate automatic counting of Ki67 cells, using U-Net for nuclei segmentation, combined with a watershed algorithm to separate overlapped regions, and a final classification into positive and negative nuclei by a random forest classifier using deep features extracted from each nucleus patch. The analysis is also performed on manually selected hot-spots of small size, with little presence of stromal cells. Here also the most common false detection errors are due to the confusion with stromal cells. A study made for Ki67 in adrenocortical carcinoma[12] also shows that manual counting and DIA techniques are highly correlated in hot-spots while, in average areas, DIA overestimates the number of non-tumor cells, identifying stromal and inflammatory cells as tumoral cells, regardless of the parameter setting. In this research, features such as shape and size were used to exclude lymphocytes and stromal cells. Although most approaches work on selected ROIs, the Breast Cancer Working group recommends approaches that assess the whole section, due to the fact that intratumoral differences can be important when analyzing only hot-spots.[6] In this scenario, a stroma segmentation procedure is necessary.

In Ref. 2, the Ki67 index is estimated by color segmentation of the tumoral area. It is shown that the index can be approximated with an area-based computation. Non-tumor areas were manually excluded using stromal masks obtained by manual annotation of the tumor boundaries on HE-stained registered images. Then these masks were superposed on the Ki67 images and were manually corrected by a pathologist.

Some previous works have been done for automatic stroma-epithelium classification for other stains, such as H&E, for the classification of small regions. For instance, in Ref. 13, a

CNN is used for classifying epithelial and stromal regions for H&E stained images of breast cancer. The images are over-segmented in superpixels, which are then resized into fixed-square images for feeding a CNN. The CNN learns to classify the patches as epithelium or stroma. As superpixels include regions of several cells, the approach cannot be used to classify images at the pixel level. In Ref. 14, the classification is made with perception-based features for square patches, also covering too large areas for pixel classification. In order to minimize the labeling effort when different datasets are used, Qi et al.[15] proposed a domain adaptation scheme applied to stroma-epythelium classification. The algorithm works with blocks (size 50×50 in this case) that correspond to a single class and not at a pixel level.

In some studies,[16,17] the tumor area is identified with virtual dual staining. This procedure consists on digital merging of parallel CK and Ki67 stained slides. Rotation and local deformation are applied to the CK-stained image to detect the epithelial zone in the Ki67 image. Although this method was useful for many cases, misaligned cases were excluded from further analysis. The studies conclude that misalignment occurs when the physical distance between the selected slides increases. The limited availability of contiguous slides stained with CK and Ki67 restricts the application of this technique.

Hiary et al.[18] proposed a segmentation system with the same target as ours; that is, separation of stromal tissue in Ki67 images, but based on traditional image features and not in an end-to-end CNN. The classification is made with two Bayesian models that collaborate for the classification decision (epithelial versus stromal). Texture-based features extracted from the HSV color model of the images are used. The effect of stromal removal before a manual count of the pathologist, both in terms of time reduction and inter-pathologist variability was also studied. Compared to our study, the samples shown in this work do not exhibit the variability that we have encountered, and moreover, quantitative segmentation results are not given.

2.2 Semantic Segmentation

Semantic segmentation algorithms approach the image segmentation problem by performing pixel-level classifications. Compared to traditional image segmentation approaches, such as superpixel segmentation methods, active contour methods, or watershed segmentation, they introduce semantics in the image segmentation process by employing a classifier trained on annotated data in order to predict the semantic category of each pixel. Although handcrafted features were initially extracted to represent pixels when training the classifier,[19] CNN-based techniques have become mainstream, obtaining the best results using end-to-end networks. One of the first well known works that applies CNNs to semantic segmentation is the fully convolutional networks.[20] It popularized CNN architectures for dense predictions without any fully connected layer, as it allowed one to produce segmentation maps for images of any size while reducing the number of parameters in the architecture. Almost all the subsequent state-of-the-art approaches on semantic segmentation adopted this paradigm. The most successful model for biomedical image segmentation has been the one proposed by Ronneberger et al., U-Net.[21] It follows an encoder–decoder architecture, where the encoder gradually reduces the spatial dimension with pooling layers and the decoder gradually recovers the object details and spatial dimension. Although other semantic segmentation networks have succeeded in different tasks,[22–24] U-Net is still the state-of-the-art for biological images.

2.3 Image-to-Image Translation

Generative adversarial networks (GANs)[25] were first introduced in 2014 as a framework to estimate deep generative models via an adversarial process, and they became an important breakthrough in the field of unsupervised learning. A GAN consists of two different modules: the first one, the generator, which tries to capture the data distribution creating a new image, and a discriminator, which takes this generated image as input, and decides on its authenticity: either it is the expected target or a synthetic approximation. GANs have been used for I2I translation to transfer images from a source domain to a target domain while preserving the content of the representation.[26]

In our work, we have used an I2I translation GAN to generate the ground-truth masks for training the U-Net semantic segmentation network. Although, to the best of our knowledge, I2I translation has never been used for this purpose, there exists several researches in which this

technique has been applied for IHC images. In Refs. 27 and 28, this technology is used to transfer the images from an easy to acquire and cost effective stain to a more difficult to manipulate and a higher cost one. In Ref. 27, a conditional cycle-consistent adversarial networks (CycleGAN) translates H&E stained images into IHC stained ones. Levy et al.[28] used a CycleGAN to convert H&E stained images into trichrome stained images for providing a cost-effective way to perform computational chemical stains on tissues.

3 Methodology

The stromal segmentation proposed in this paper is based on a semantic segmentation CNN widely used for biological images, U-Net.[21] First, transfer knowledge from a natural image dataset is used by initializing the encoder weights with those obtained from the ImageNet dataset. This is done to reduce the size of the labeled Ki67 dataset required for training. For our task, we need a database of Ki67 images with the corresponding ground truth, which is the stroma mask for each dataset image. This database is used for training the U-Net, as shown in the left side of Fig. 1(b) (bottom part of this figure). Once trained, the model can be used to segment any new Ki67 image to produce the stroma mask, as shown in the right side of Fig. 1(b).

For the ground-truth creation, the approach shown in Fig. 1(a) (top part of the figure) is used: we train an I2I translation model based on a GAN to translate Ki67 images into fake-CK19. This is a style transfer mechanism, which is able to highlight epithelial areas in Ki67 images, making them appear as CK19 positives, without changing the relevant structural information of the original Ki67 images. As explained in Sec. 5, the GAN model is trained with unpaired Ki67 and CK19 images [left side of Fig. 1(a)]. So the datasets to train the GAN are two sets of images of the two stains, without any kind of labeling nor correspondence. As shown in the right side of Fig. 1(a), once trained, this GAN model is used to generate the fake-CK19 images of the Ki67 dataset necessary for the semantic segmentation. These fake-CK19 images are binarized, filtered, and manually corrected when necessary, to create the ground-truth masks corresponding to the Ki67 images.

To validate the generated ground-truth masks and to guide the manual correction, the real CK19 images corresponding to the Ki67 images are necessary, because, as previously mentioned, epithelial zones in Ki67 images are not always easy to visually identify. For this reason, adjacent sections of the Ki67 WSI stained with CK19 have been used. Then, for each selected Ki67 image, a registration procedure has been applied to search for the corresponding CK19 image. Let us mention that these adjacent sections can be used as a guide for the manual correction but often exhibit a poor correspondence of the cell structures because they do not represent the exact same tissue and may have been differently manipulated in the staining and scanning processes. As a result, these CK19 images cannot be used directly to obtain the

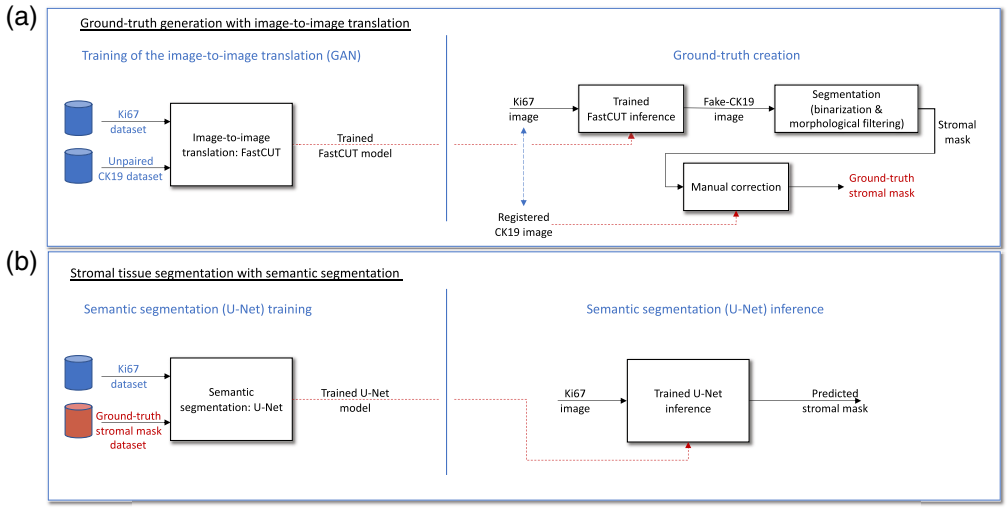


Fig. 1 Workflow diagram of the whole approach: (a) ground-truth generation with I2I translation and (b) stromal tissue segmentation with semantic segmentation.

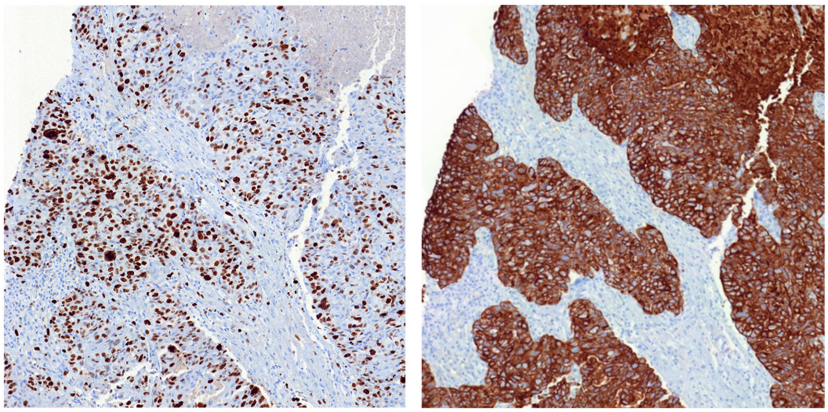
ground-truth masks. Instead, the I2I translation is used because it perfectly preserves the cell structures of Ki67 images, as needed to compute the masks.

4 Database

The database has been constructed for extracting tiles from whole slide images of 17 different patients of invasive breast carcinomas. Slides have been scanned with a "3DHISTECH Pannoramic 1000 slide scanner" with a magnification of 40× and a resolution of 0.25×0.25 μm/pixel. From each patient multiple stainings of adjacent slides are available, among them Ki67 and CK19. Although the markers are applied to parallel sections of the extracted sample, they are not always contiguous, and significant differences can be observed between them. For Ki67, we have obtained 150 tiles of size 4096×4096 pixels, corresponding to a field of view of 995×995 μm. The patients have different levels of proliferation, marked by the Ki67 stain, and the tiles reflect a wide variety of cellular structures. A registration algorithm has been applied in order to find the corresponding tiles for the CK19 marker. This produces a set of 150 couples of tiles, each one composed of original Ki67 and registered CK19 images. An example of couple of tiles is shown in Fig. 2.

Observing the CK19 WSI of the 17 patients, we can notice that there is a wide variety of cell structures and tumoral expressions (see Fig. 3). The tile selection has been done to achieve a representative and balanced database including all the different tumoral expressions.

If we focus on the CK19 tiles, where the epithelial zones can be distinguished from the stromal ones, we can observe very different structures both in the size and the distribution of the epithelial areas. This variability reflects different tumor architectures. As will be discussed



(a) (b)

Fig. 2 (a) Ki67 tile and (b) registered CK19 tile corresponding to (a).

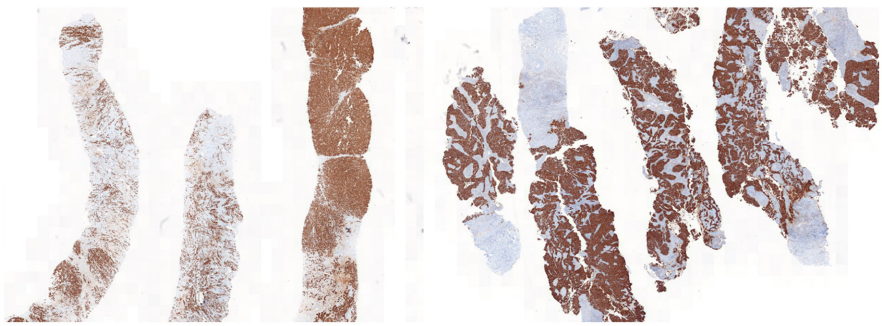


Fig. 3 Crops of various CK19 whole slide images illustrating the variety of cell structures and tumoral expressions.

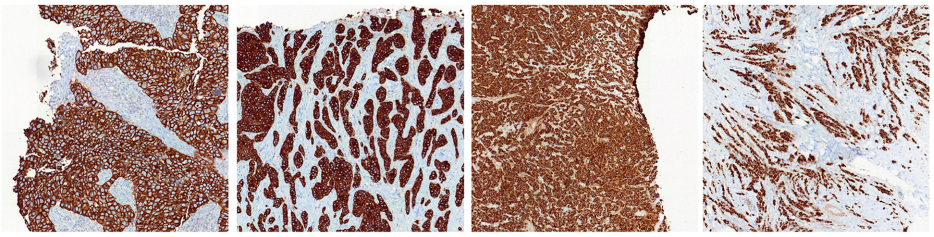


Fig. 4 The four CK19 classes: (a) class 1: nest patterns, (b) class 2: trabecular, (c) class 3: solid, and (d) class 4: cordal.

in the sequel, for ground-truth generation, it is useful to process differently these classes. As a result, we have defined four different classes, and grouped our dataset images accordingly:

- Class 1: Nest pattern, consisting of large epithelial areas. They are dense and well defined.
- Class 2: Trabecular pattern, small but well defined epithelial areas.
- Class 3: Solid pattern, epithelial zones with granular shapes with river-shaped stroma.
- Class 4: Cordal pattern, epithelial areas with a filamentous shape.

Visual examples of these classes can be seen in Fig. 4. To ensure an equal representation of the four classes, and thus a balanced dataset, we have approximately collected the same number of images for each class.

Finally, the extracted Ki67 and CK19 tiles of size 4.096×4.096 pixels have been downsampled to 512×512 pixels. We have considered that this resolution was appropriate in terms of memory requirements and computational load for the CNNs while keeping enough visual information for the stroma segmentation.

This dataset is used to train both the U-Net and the GAN. As illustrated in Fig. 1, the Ki67 dataset images are essentially used to train the U-Net semantic segmentation (bottom part of Fig. 1). The dataset has been partitioned into train, validation, and test datasets. We have used different patients for each partition in order to avoid data leakage. Table 1 describes the number of patients and images used for each partition and each class.

As discussed in Sec. 3, in order to be able to train the U-Net, we need to create the associated ground-truth images, which are masks defining the stromal areas of the Ki67 images of the database. This is the objective of the upper part of Fig. 1 and the procedure involves the training of a

Table 1 Description of the partitions of the used dataset. It must be taken into account that there are two patients with tiles assigned to both classes 3 and 4.

		Train	Validation	Test	Total
Class 1	Patients	4	2	1	7
	Images	30	2	4	39
Class 2	Patients	3	2	2	7
	Images	19	12	8	39
Class 3	Patients	1	0	1	2
	Images	30	0	4	34
Class 4	Patients	1	1	1	3
	Images	31	3	4	38
Total	Patients	10	3	4	17
	Images	110	20	20	150

GAN. The goal of the GAN is to produce initial stromal masks that will reduce the amount of manual annotation work on these specific Ki67 images. As a result, the aim of the GAN training is not to produce a model that can generalize to new unseen images but to get the initial masks of the specific set of images included in the database. Therefore, all the images included in the dataset are going to be used to train the GAN. As we do not want the GAN model to generalize to other images, there is no need to make any partition of the dataset into train, validation, and test sets as was done for the U-Net. However, the pairs of Ki67 and CK19 images have been distributed in four classes depending on the class of the CK19 images (nest, trabecular, solid, and cordal).

5 Ground-Truth Generation

5.1 Stain Translation with FastCUT

I2I translation is the process of taking images from one domain and transforming them to another domain while preserving the content representation of the source image. During these last years and thanks to the progress of deep learning technologies, many I2I translation implementations have been developed providing significant improvements in the performance and results. Many tasks, such as realistic-looking image synthesis, text-to-image translation, image semantic manipulation, or image restoration, rely on I2I translation techniques. Here we apply I2I translation to transfer Ki67 tiles into the CK19 domain, with the objective to obtain a stroma mask that can be easily segmented and manually corrected to create the ground truth. Let us note that complex epithelial-stroma structures, such as the ones shown in Figs. 4(c) and 4(d), make it almost impossible to obtain a segmentation by manual annotation, even if the corresponding registered CK19 tile is available.

The I2I translation task is based on a generative architecture capable of modeling the distribution of the target domain by generating credible fake data that resembles an original image from the target domain. Many generative models performing these I2I translation tasks are available. We have applied two popular techniques: CycleGAN[29] and fast contrastive unpaired translation (FastCUT),[30] obtaining significantly better results with the second one. Both techniques are based on GANs,[25] which are neural networks consisting on two simultaneously trained models: a generator (G) that generates fake data and a discriminator (D) that tries to distinguish the fake data from real examples. In a GAN, the generator is trained from the feedback that it receives from the classification of the discriminator. In turn, the discriminator tries to improve its task by evaluating how far its classifications are from the true labels: real or fake. During the training, the generator and the discriminator are competing because as one performs better, the other worsens. The goal is to find the equilibrium, in which the generator is capable of producing fake images that cannot be distinguished from the original training dataset, whereas the discriminator is only able to randomly guess if the generated data are real or not. This results in the adversarial loss, which is one of the main components of the loss of these networks.

There are two types of generative adversarial models: those that use paired data from both domains and those that do not require paired data. In our case, as we do not have a perfect correspondence in our pairs of tiles for Ki67 and CK19 staining, we have used a model capable of working with unpaired images. Furthermore, the objective of the selected approach is to change the appearance from an input domain to a target domain while retaining the structure of the input image. As mentioned, target appearance is enforced using the adversarial loss. In the case of CycleGAN, the content is preserved through cycle consistency. However, maximization of the mutual information between corresponding input and output patches does a better job preserving not only the content but also the structure of the input image. The FastCUT model was introduced in Ref. 30, based on the GAN architectures and adding a patch-wise contrastive learning on the discriminator. The patch-wise contrastive learning is based on maximizing the mutual information of patches from the same location while minimizing those from different locations. To compute this loss, features from different layers of the encoder, which represent patches of different sizes, are used.

To generate the fake CK19 images, the FastCUT model was selected over the CycleGAN. Note that since these I2I translation GAN are tools that simplify the generation of the ground truth, there was no available ground truth to make an objective comparison. Therefore, the

comparison between FastCUT and CycleAN was done subjectively by visual inspection of the produced fake-CK19 and the registered CK19 images.

For each one of the four CK19 style classes (nest patterns, trabecular, solid, and cordal), we trained a FastCUT model with 400 epochs and a learning rate of 0.0002. This resulted in four different trained FastCUT models. The training was stopped after subjective evaluation of the generated fake CK19 images (selecting images that would minimize the annotation work). The obtained results show the preservation of the Ki67 input structure while changing the domain. The training process is illustrated in the left side of Fig. 1(a), whereas the translation inference corresponds to the first block of the right side of Fig. 1(a).

5.2 Mask Computation

This step is referred to as "segmentation" in the diagram on the right side of Fig. 1(a). After the I2I translation process, we generate the stroma masks from the obtained fake CK19 images applying binarization and classical image processing tools, such as morphological operators.

The first step of the segmentation is the binarization of the fake CK-19 image. This binarization produces accurate results except for some noise in the stromal area and small holes in the epithelium. In the case of images of class 1 and 2, with more dense epithelial areas, such holes are removed with an opening by reconstruction of erosion, followed by closing by reconstruction of dilation. The erosion and the dilation use as structuring element a disk of radius 2. With the opening we can remove small holes in the epithelium while the closing removes noise in the stromal area. For classes 3 and 4, an opening by reconstruction of erosion is applied, with structuring element a disk of radius 1. Examples are shown in Fig. 5.

5.3 Manual Correction

The masks resulting from the translation and segmentation processes have to be checked and manually corrected before including them in the final ground truth to be used for the training of the U-Net dealing with the semantic segmentation of the Ki67 images. Registered pairs of Ki67

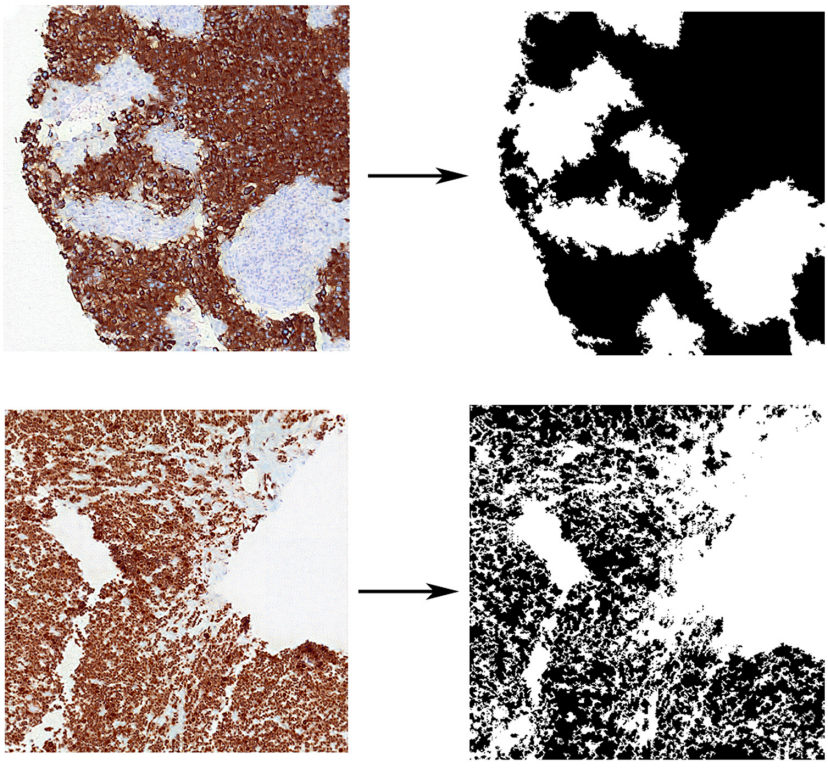


Fig. 5 Examples of masks obtained from the segmentation of fake CK19 images of classes 1 (large areas) and 3 (granular shape). (a) Fake CK19 images and (b) masks.

and CK19 tiles are used for this checking. Although there is not a perfect correspondence between these pairs of tiles, there is most of the time enough information to visually confirm whether the areas extracted as epithelium are correct. Errors in the masks are manually corrected. While errors in class 1 and 2 images (nest and trabecular, which correspond to more compact stroma areas) are easy to rectify, classes 3 and 4 (which have a filamentous pattern) are much harder, and only very clear mistakes have been corrected. Manual correction examples can be seen in the figures of Sec. 5.4. In order to measure the effect of this manual correction we have computed the Intersection over Union (Jaccard index) between the masks obtained and the ones manually corrected. Manual correction was not needed for 67% of the images, meaning that the mask obtained was considered good enough for the purpose of discarding stroma cells. For the remaining ones, the mean IoU was 0.93, with a minimum value of 0.73, and a maximum of 0.98. These numbers show the excellent CK19 translation achieved and the good segmentation result of Sec. 5.2. The differences are mainly due to areas that were incorrectly stained in the fake CK19 of images of class 1 and 2, which did not require much labeling effort. On the contrary, the detailed images of class 3 and 4, which are very hard to manually label, are usually better translated and require few changes.

5.4 Ground-Truth Generation Result

In this section, we present the results obtained from the different steps of the ground-truth generation process. The procedure is summarized in the right side of Fig. 1(a). Example of images from each one of the four classes are shown respectively in Figs. 6–9.

As a summary, 150 Ki67 images of the dataset are transformed with the FastCUT model corresponding to its class. This procedure generates the fake CK-19 images, which are segmented to obtain initial masks that are manually corrected with the help of registered CK19 images.

Note again that, as the aim of this process is actually the creation of the ground truth, we cannot rely on any objective evaluation. As a result, we have evaluated our results by subjective comparison of the pair of Ki67 and registered CK19 images with the final masks.

From this analysis, we have concluded that the FastCUT model works quite well for this task and has been the key for generating the ground truth masks for the U-Net training in an efficient

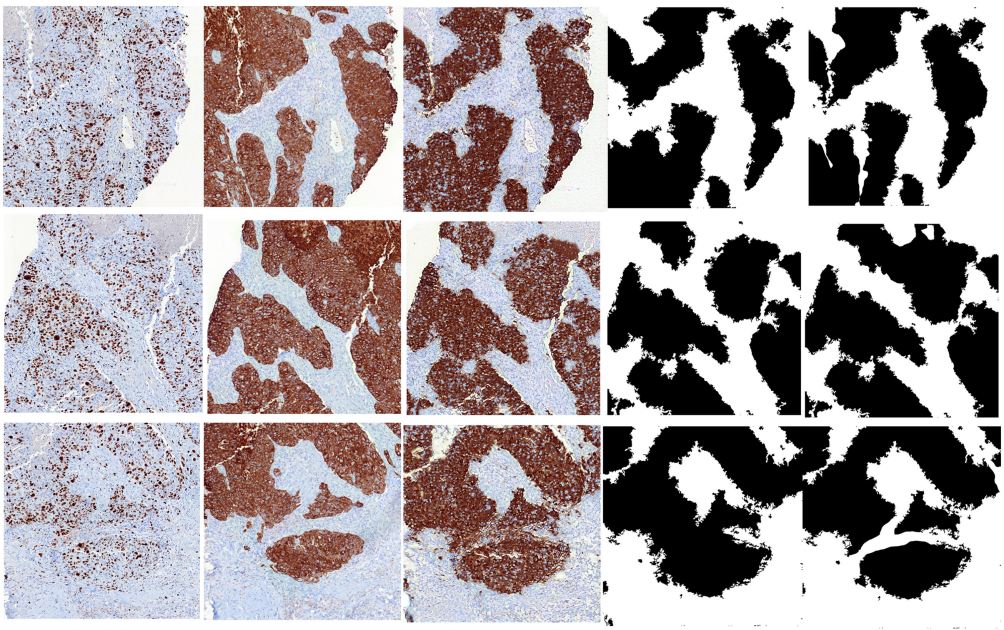


Fig. 6 Ground-truth generation results for images of class 1: nest patterns. (a) Original Ki67 image; (b) registered CK19 image; (c) fake CK19; (d) mask resulting from the translation and the segmentation; and (e) manually corrected mask.

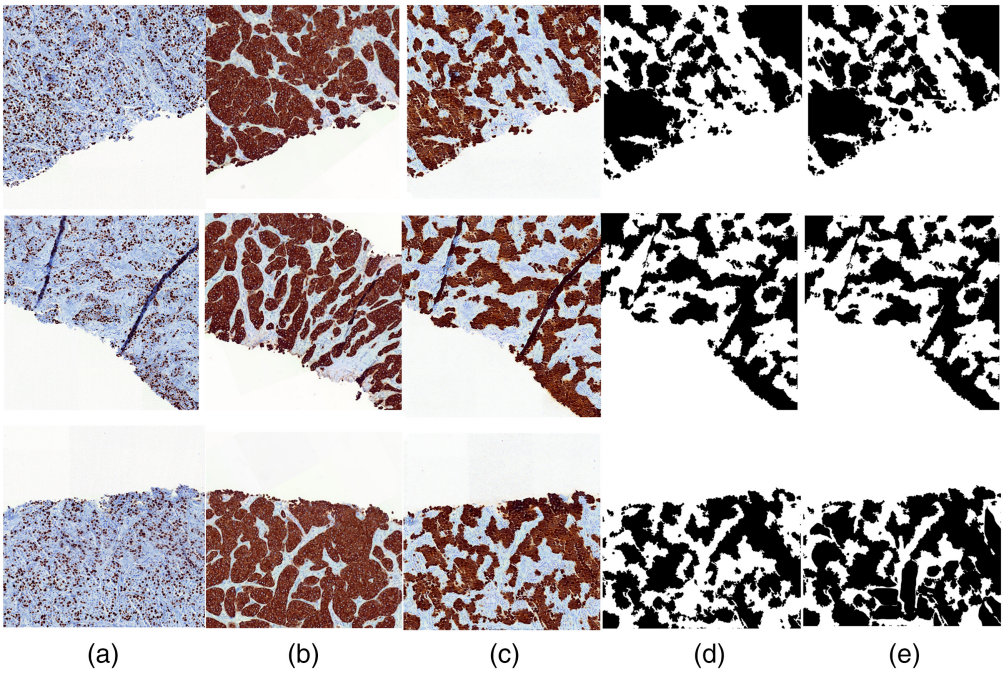


Fig. 7 Ground-truth generation results for images of class 2: trabecular. (a) Original Ki67 image; (b) registered CK19 image; (c) fake CK19; (d) mask resulting from the translation and the segmentation; and (e) manually corrected mask.

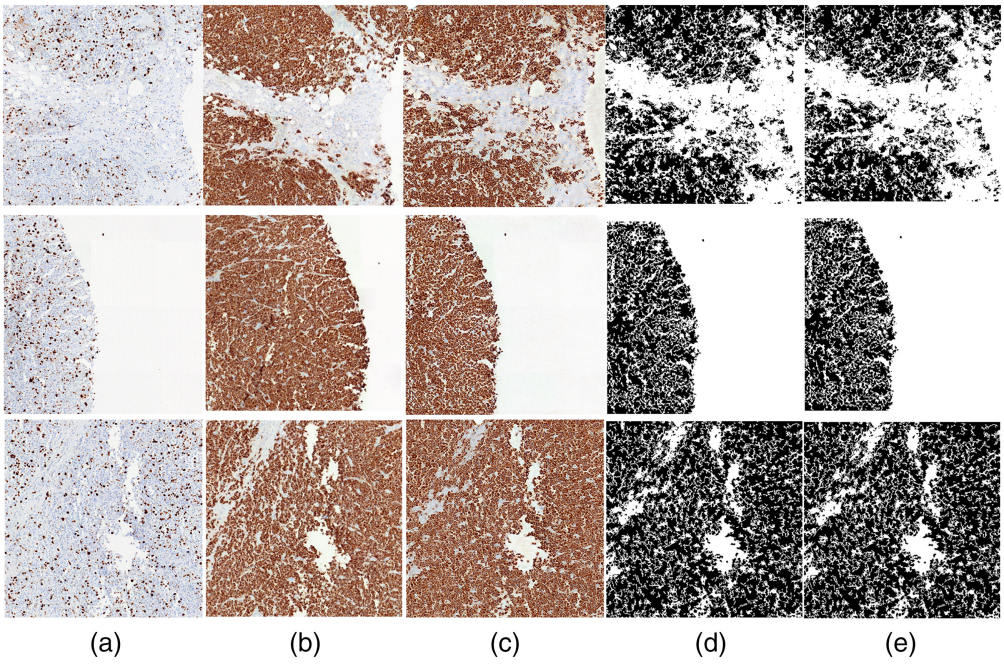


Fig. 8 Ground-truth generation results for images of class 3: solid. (a) Original Ki67 image; (b) registered CK19 image; (c) fake CK19; (d) mask resulting from the translation and the segmentation; and (e) manually corrected mask.

and easy way. The masks resulting from the translation and the segmentation are of course not always perfect and manual correction is indeed an important step. However, our experience is that the manual correction was rather easy to perform and the automatically generated masks were very good starting points.

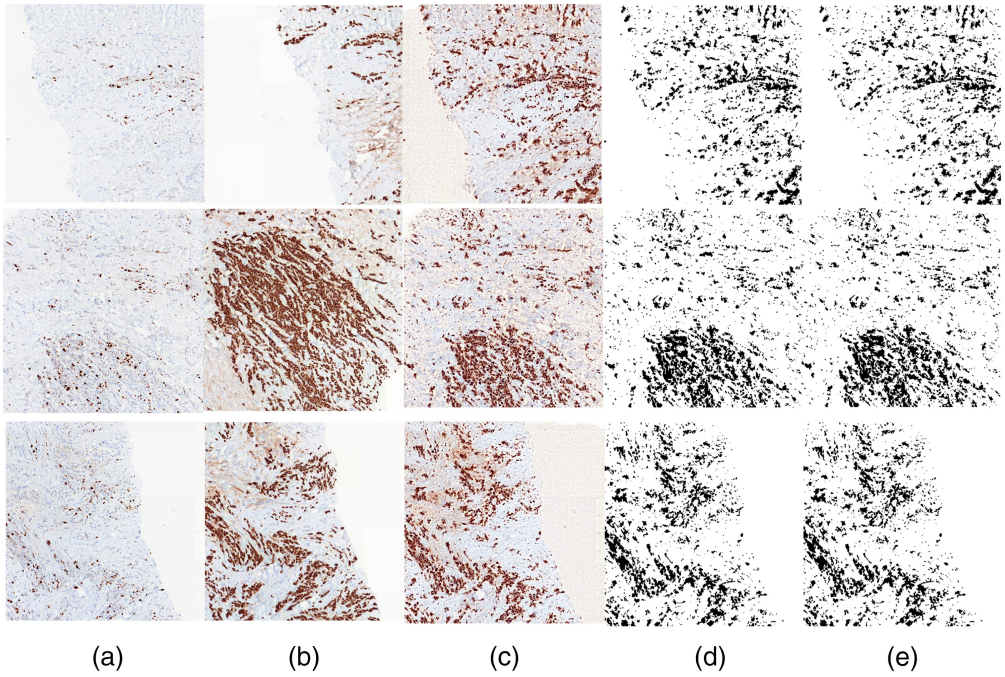


Fig. 9 Ground-truth generation results for images of class 4: cordal. (a) Original Ki67 image; (b) registered CK19 image; (c) fake CK19; (d) mask resulting from the translation and the segmentation; and (e) manually corrected mask.

6 Semantic Segmentation of Ki67 Images

6.1 U-NET-Based Segmentation

Once ground-truth masks for stroma segmentation have been obtained, we can train the semantic segmentation model to produce stroma masks for new images. This step corresponds to the left side of Fig. 1(b). The selected semantic segmentation model is the U-Net model,[21] which has proven to produce very good results for many biomedical image segmentation tasks. Of course, our model has to be able to process arbitrary Ki67 images. As the notion of cell structure classes used for the generation of the ground truth is not easy to infer automatically, the semantic segmentation algorithm cannot rely on this information. Here this notion has only been used for the partition of the database in training, validation, and test sets. The complete database is made of 150 images, distributed as explained in Table 1.

In order to increase the amount of training data, data augmentation has been applied. The following transformations were allowed: horizontal and vertical flips, translation, rotation, and small contrast variations. The encoder was the ResNeXt-50-32x4d,[31] where the blocks of the contractive path follow an architecture based on the inception's split-transform-merge strategy while using residual connections. The weights of the encoder were initialized with the ImageNet weights. We used the sigmoid function as activation function, adam as optimizer and the dice loss as loss function. The batch-size was 4 and the learning rate was 0.00068. The model was trained for 150 epochs, achieving stable results for both training and validation partitions after 60 epochs.

The best F-score for the validation partition was obtained at epoch 66. The model obtained at this epoch was used to obtain the final result for the test set, obtaining an F-score of 0.87. The importance of knowledge transfer from ImageNet is also worth mentioning. Although ImageNet is a database of natural images, the initialization of the U-Net encoder with a network pretrained on this extensive dataset allows us to achieve already good results in the very first epochs (F-score higher than 0.80 for both train and validation partitions after 5 epochs). Examples of masks resulting from the semantic segmentation for some images of the test database are shown in Figs. 10 and 11.

The field of view corresponding to the database images is relatively large: 995×995 μm. This has allowed us to capture a wide variety of structures in a limited number of images.

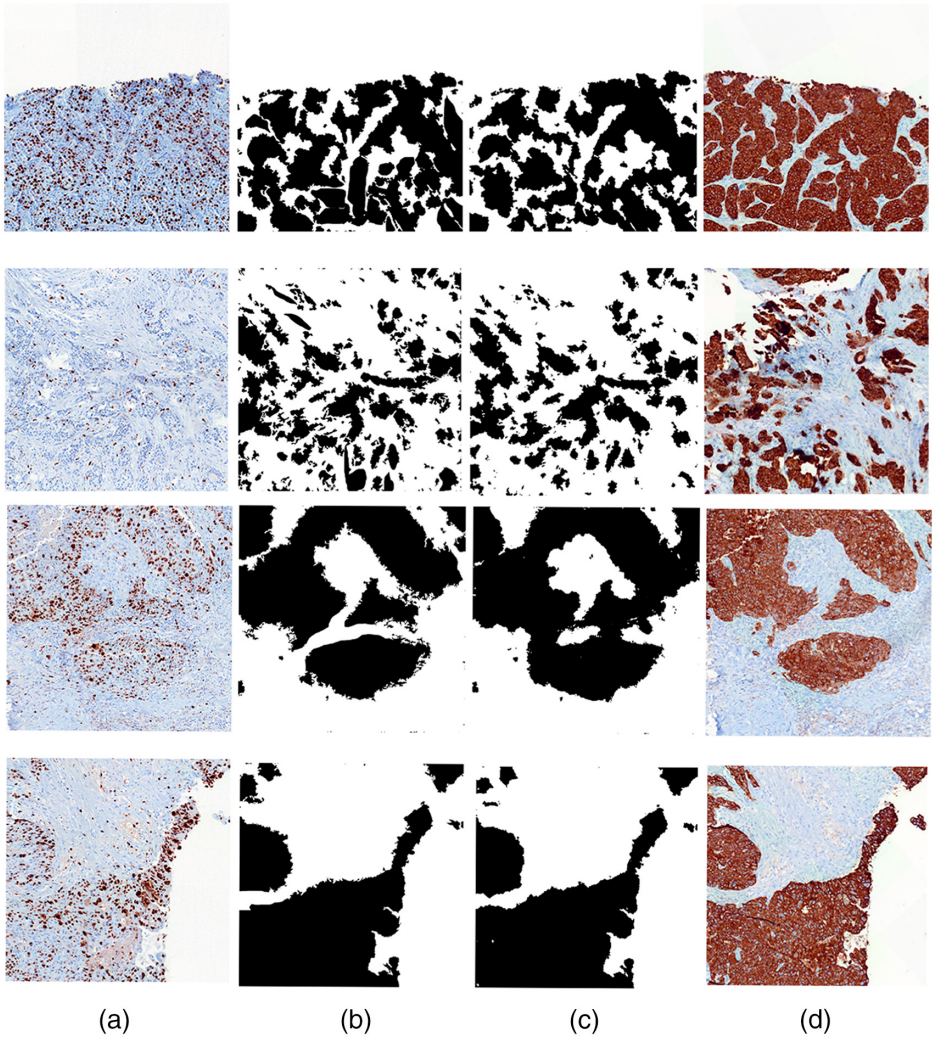


Fig. 10 Results for images of class 1 and 2. (a) KI67 image; (b) ground-truth mask; (c) mask predicted by the U-Net; and (d) registered CK-19.

However, once the ground-truth masks have been obtained, we have also tested the U-Net segmentation algorithm in smaller patches. That is, each Ki67 image and its corresponding ground-truth mask have been partitioned into 4 and 16 images, and a U-Net has been trained with 368 images in the first case (corresponding to the original 92 images of the training partition) and 1472 images in the second case. The results evaluated on the test partition provide very similar F-scores to those obtained with the initial size. This shows that U-Net semantic segmentation results are not dependent on the large field of view and so, once the ground truth has been constructed, segmentation can be applied to smaller images.

6.2 Example of Use of the Stroma Mask for Ki67 Scoring

In order to show the relevance of the stroma segmentation network for Ki67 scoring, an additional experiment is described in this section. Three images have been selected in order to perform a cell segmentation and classification task. In order to define the ground truth for this task, the three images were manually segmented on a cell basis, and each cell classified as stroma, positive or negative. Ki67 score is obtained by computing the percentage of positively stained tumor cells among the total number of malignant cells (that is positively and negatively stained cells, excluding the stroma cells).

To automatically compute this score, segmentation and classification of cells have been performed with a state-of-the art system that can only separate stained and non-stained cells. Figure 12 shows the results obtained when only positive and negative cells are obtained, and

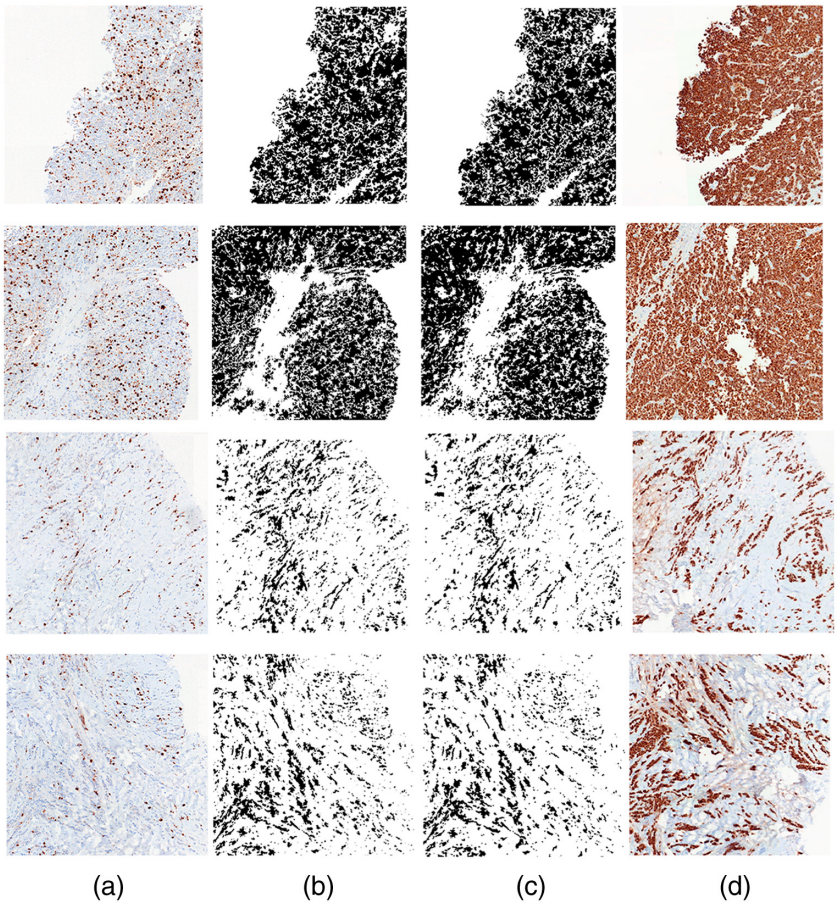


Fig. 11 Results for images of class 3 and 4. (a) KI67 image; (b) ground-truth mask; (c) mask predicted by the U-Net; and (d) registered CK-19.

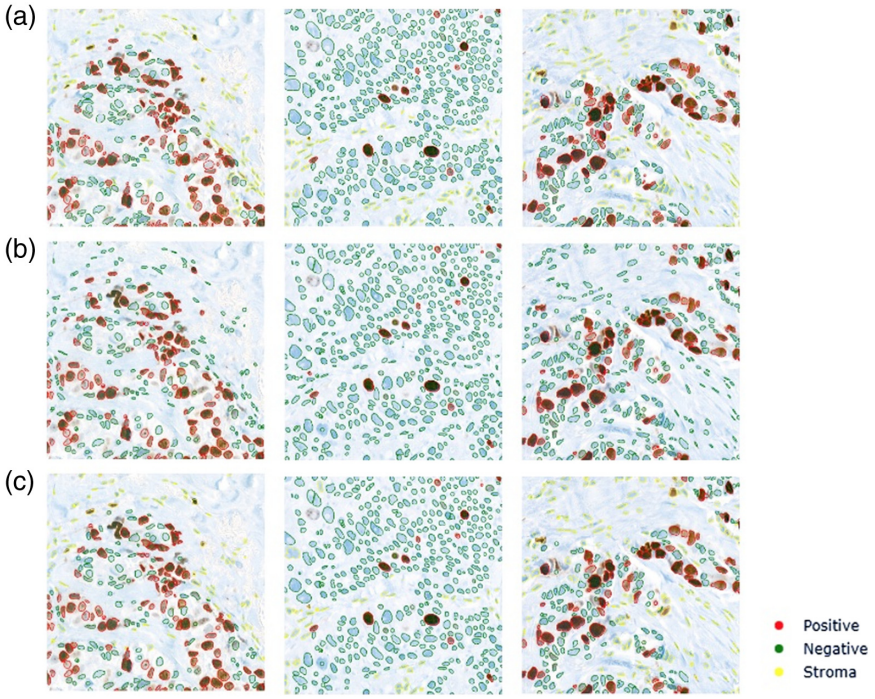


Fig. 12 Cell segmentation and classification example: (a) ground-truth for three different images; (b) cell segmentation without stroma mask; and (c) cell segmentation after applying stroma mask.

Table 2 Ki67 score [percentage of positive cells with respect to the total epithelial cells (positive + negative)] for each one of the approaches.

	Prediction (%)	Prediction + stroma mask (%)	Ground truth (%)
Image 1	3.00	3.14	3.19
Image 2	20.99	33.33	35.55
Image 3	37.75	50.00	53.64

after applying the stroma mask computed with the method proposed in this paper. The stroma mask allows to discard all cells under the mask when computing the Ki67 score. Table 2 compares the score obtained with the ground truth, the one obtained with the cell classification method, which does not consider stroma cells and the score obtained when applying our stroma mask. It can be observed that, as expected, results are much closer to ground truth since confusion of stroma with negative tumor cells is avoided.

7 Conclusions

In this paper, we have developed a semantic segmentation system for tumoral mask predictions for breast cancer images with Ki67 staining. The objective of this development was to create a tool able to distinguish between tumor and stroma regions in order to compute the Ki67 score without the interference of stroma cells. The proposed method should be used as a first step in automatic DIA, before cell counting algorithms are applied, in order to avoid errors produced by the confusion of stromal cells with Ki67 negative tumoral cells. Also, in case of manual counting, the proposed technique can generate a mask to define the areas that the pathologist needs to quantify. This strategy should lead to a reduction of the inter-observer variability and the time required for nuclei counting.

One of the essential problems we had to deal with is the lack of tumoral mask databases with associated ground truth to train the semantic segmentation. In order to generate this ground truth, we have opted for a semisupervised strategy. The first step of this strategy was to translate Ki67 images to fake CK19 images that highlight tumoral areas and are easy to segment. From the obtained results, we can conclude that the FastCUT translation model is a good option for transferring Ki67 images into CK19 sytle. These fake CK19 images are used to create initial stromal masks. We have also shown that the combination of the I2I translation model with the manual correction was a good option. The regions where the translation had clearly failed were easy to identify and manually correct. The resulting database consists of 150 images with a field of view of 995×995 μm and their corresponding stromal masks, representing a variety of structures that the stromal/epithelial regions usually adopt.

The semantic segmentation network trained with this dataset has shown excellent results. We have also proven that the results of the semantic segmentation are not dependent on the field of view used.

This system can easily be extended to produce results for other markers, such as ER, PR, or HER2. As future work, we plan to extend the semantic segmentation model so that it also predicts areas of necrosis. Necrosis areas are zones with dead tissue, which, as the stromal area, do not have to be taken into account when computing the proliferation index.

Disclosures

No conflicts of interest, financial or otherwise, are declared by the authors.

Acknowledgments

The authors would like to thank Pathology Laboratory Teams from Hospital Vall d'Hebron and Bellvitge, and Dr. Teresa Soler in particular, for the discussions and advice. This work was supported by the Spanish Research Agency (AEI) (Project No. PID2020-116907RB-I00) of

the call MCIN/AEI/10.13039/50110001103 and the European Regional Development Funds under program FEDER Catalunya 2014–2020 and SA18-014623 DIGIPATICS.

References

1. J. Gerdes et al., "Cell cycle analysis of a cell proliferation-associated human nuclear antigen defined by the monoclonal antibody Ki67," *J. Immunol.* **133**(4), 1710–1715 (1984).
2. M. K. K. Niazi et al., "Relationship between the Ki67 index and its area based approximation in breast cancer," *BMC Cancer* **18**(1), 1–9 (2018).
3. J. Temprana-Salvador et al., "DigiPatICS: digital pathology transformation of the catalan health institute network of 8 hospitals—planification, implementation, and preliminary results," *Diagnostics (Basel)* **12**(4), 852 (2022).
4. C. Downey et al., "The prognostic significance of tumour–stroma ratio in oestrogen receptor-positive breast cancer," *Br J Cancer* **110**(7), 1744–1747 (2014).
5. Y. Yuan et al., "Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling," *Sci. Transl. Med.* **4**(157), 157ra143–157ra143 (2012).
6. M. Dowsett et al., "Assessment of Ki67 in breast cancer: recommendations from the international Ki67 in breast cancer working group," *J. Natl. Cancer Inst.* **103**(22), 1656–1664 (2011).
7. Z. Volynskaya et al., "Ki67 quantitative interpretation: insights using image analysis," *J. Pathol. Inf.* **10**(1), 8 (2019).
8. M. Christgen et al., "The region-of-interest size impacts on Ki67 quantification by computer-assisted image analysis in breast cancer," *Hum. Pathol.* **46**(9), 1341–1349 (2015).
9. S. Fasanella et al., "Proliferative activity in human breast cancer: Ki67 automated evaluation and the influence of different Ki67 equivalent antibodies," *Diagn. Pathol.* **6**(Suppl 1), S7 (2011).
10. B. R. Barricelli et al., "Ki67 nuclei detection and Ki67-index estimation: a novel automatic approach based on human vision modeling," *BMC Bioinf.* **20**(1), 1–14 (2019).
11. K. Benaggoune et al., "A deep learning pipeline for breast cancer Ki67 proliferation index scoring," arXiv, https://doi.org/10.48550/arXiv.2203.07452 (2022).
12. Y. Yamazaki et al., "Comparison of the methods for measuring the Ki67 labeling index in adrenocortical carcinoma: manual versus digital image analysis," *Hum. Pathol.* **53**, 41–50 (2016).
13. J. Xu et al., "A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images," *Neurocomputing* **191**, 214–223 (2016).
14. F. Bianconi, A. Álvarez-Larrán, and A. Fernández, "Discrimination between tumour epithelium and stroma via perception-based features," *Neurocomputing* **154**, 119–126 (2015).
15. Q. Qi et al., "Curriculum feature alignment domain adaptation for epithelium-stroma classification in histopathological images," *IEEE J. Biomed. Health. Inf.* **25**(4), 1163–1172 (2020).
16. T. Koopman et al., "Digital image analysis of Ki67 proliferation index in breast cancer using virtual dual staining on whole tissue sections: clinical validation and inter-platform agreement," *Breast Cancer Res. Treat.* **169**(1), 33–42 (2018).
17. R. Røge et al., "Proliferation assessment in breast carcinomas using digital image analysis based on virtual Ki67/cytokeratin double staining," *Breast Cancer Res. Treat.* **158**(1), 11–19 (2016).
18. H. Hiary et al., "Automated segmentation of stromal tissue in histology images using a voting Bayesian model," *Signal Image Video Process.* **7**(6), 1229–1237 (2013).
19. J. Shotton, M. Johnson, and R. Cipolla, "Semantic Texton forests for image categorization and segmentation," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, IEEE, pp. 1–8 (2008).
20. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3431–3440 (2015).
21. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
22. H. Zhao et al., "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2881–2890 (2017).
23. L.-C. Chen et al., "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(4), 834–848 (2017).
24. Z. Huang et al., "CCNet: criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, pp. 603–612 (2019).
25. I. Goodfellow et al., "Generative adversarial nets," in *Adv. Neural Inf. Process. Syst. 27* (2014).
26. P. Isola et al., "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1125–1134 (2017).
27. Z. Xu et al., "GAN-based virtual re-staining: a promising solution for whole slide image analysis," arXiv, https://doi.org/10.48550/arXiv.1901.04059 (2019).
28. J. J. Levy et al., "Preliminary evaluation of the utility of deep generative histopathology image translation at a mid-sized NCI cancer center," bioRxiv, https://doi.org/10.1101/2020.01.07.897801 (2020).

29. J.-Y. Zhu et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2223–2232 (2017).
30. T. Park et al., "Contrastive learning for unpaired image-to-image translation," *Lect. Notes Comput. Sci.* **12354**, 319–345 (2020).
31. S. Xie et al., "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, pp. 1492–1500 (2017).

Biographies of the authors are not available.