DermoGAN: multi-task cycle generative adversarial networks for unsupervised automatic cell identification on in-vivo reflectance confocal microscopy images of the human epidermis

Imane Lboukili, Georgios Stamatas, Xavier Descombes

DermoGAN: Multi-task cycle-GANs for unsupervised automatic cell identification on in vivo reflectance confocal microscopy images of the human epidermis

Imane Lboukili,[a, b] Georgios Stamatas,[a] Xavier Descombes,[b]
[a]Johnson & Johnson Santé Beauté France, Paris, France
[b]UCA – INRIA – I3S/CNRS, Sophia Antipolis, France

Abstract

Significance: Accurate identification of epidermal cells on reflectance confocal microscopy (RCM) images is important in the study of epidermal architecture and topology of both healthy and diseased skin. However, analysis of these images is currently done manually, and therefore time-consuming, subject to human error and inter-expert interpretation. It is also hindered by low image quality due to noise and heterogeneity.

Aim: To design an automated pipeline for the analysis of the epidermal structure from reflectance confocal microscopy images.

Approach: Two attempts have been made at automatically localizing epidermal cells, called keratinocytes, on RCM images, the first based on a rotationally symmetric error function mask, and the second on cell morphological features. Here we propose a dual-task network to automatically identify keratinocytes on RCM images. Each task consists of a cycle Generative Adversarial Network (cycle-GAN). The first task aims to translate real RCM images into binary images thus learning the noise and texture model of RCM images, whereas the second task maps Gabor-filtered RCM images into binary images, learning the epidermal structure visible on RCM images. The combination of the two tasks allows one task to constrict the solution space of the other thus improving overall results. We refine our cell identification by applying the pre-trained StarDist algorithm to detect star-convex shapes, thus closing any incomplete membranes and separating neighboring cells.

Results: The results are evaluated both on simulated data and on manually annotated real RCM data. Accuracy is measured using recall and precision metrics, which as summarized as the F1-score.

Conclusions: We demonstrate that the proposed fully unsupervised method successfully identifies keratinocytes on RCM images of the epidermis, with an accuracy on par with experts' cell identification and is not constrained by limited annotated available data and can be extended to images acquired using various imaging techniques with no-retraining.

Keywords: cycle-GAN, identification, keratinocytes, multi-task, reflectance confocal microscopy, segmentation.

1 Introduction

Reflectance confocal microscopy (RCM) is a non-invasive *in vivo* imaging technique that allows for visualization of epidermal cells, called keratinocytes, at cellular level in the epidermis and upper layers of the dermis (150 - 200 µm in depth, depending on body site [1]). It provides information on the geometry and topology of the skin, which are key elements in skin barrier and health, thus helping in the study of infant and children skin maturation, adult skin ageing, and photo-ageing due to UV exposure. RCM can also be used to assess skin inflammatory diseases, e.g., psoriasis, allergic contact dermatitis, and skin cancer [3], providing information faster than traditional biopsies, and potentially reducing the number of unneeded biopsies, by guiding them to delimit lesion borders and helping in disease diagnosis and monitoring.

Although automated methods have been developed to identify some lesions on RCM images [4–8], most of

the analysis is performed manually which is time consuming, and subject to inter and intra-expert interpretation[9]. Hence the need for an automated method for keratinocytes identification on RCM images, which would allow for a more reproducible, unbiased, and precise analysis. Unfortunately, image quality, heterogeneity, and low signal-to-noise ratio are a hurdle to automated methods development. Attempts at automating keratinocytes identification on RCM images have been made, and were based on the identification of cell morphological features, e.g., membrane size [10,11], but are hindered by manual parametrization often different between datasets, image types, and epidermal layers. Deep learning methods could be an alternative solution to circumvent these problems.

Accurate automated cell identification on biomedical images with deep learning has been a growing research topic in computer vision but is hindered by the lack of labelled data on account of cost, time, and domain-specific skills. Unsupervised learning bypasses the labeled data scarcity problem by tapping into unlabeled data potential. One of the main developments in unsupervised learning research of the recent years are Cycle Generative Adversarial Networks (cycle-GAN) [12] for unpaired image-to-image translation, and are classically used for synthetic images generation and data augmentation [13–16].

We propose a top-down, structure aware, multi-task cycle-GANs architecture which we have named DermoGAN to automatically detect keratinocytes on RCM images. The multi-task model performs two parallel cycle-GANs, to denoise RCM images while highlighting membranes positions, and provides an incomplete cell identification, which is then refined and completed by a post-processing based on star-convex shape detection. The proposed architecture is fully unsupervised and thus not limited by training annotations, often the first limitation to the use of deep learning methods in the analysis of biomedical images. To our knowledge, this is the first use of cycle-GANs in a multi-task framework. Additionally, while generally used for synthetic images generation and data-augmentation, here we employ the cycle-GAN algorithm as an image-denoiser and cell-identifier. Indeed, we change our perspective on the cell identification problem, and move the data augmentation approach consisting in learning the noise model in RCM images in order to create synthetic images, which will then be used to augment our dataset to be used in other models, to a new

approach consisting in learning the image *denoising* model. This change in perspective makes use of the cycle-consistency property of cycle-GANs.

We compare the proposed method to seven other approaches, a supervised method based on a U-net architecture[17], a pre-trained StarDist[18] applied to Gabor-filtered images, two unsupervised approaches based on a cycle-GAN with different inputs, a tailored-pipeline based on the detection of membrane morphological features[9], the CellPose[19] algorithm for cellular segmentation, and finally DermoGAN followed by postprocessing using CellPose.

We demonstrate that the presented DermoGAN architecture performs on par with expert manual identification of cells and outperforms the seven other tested automated methods in accuracy and execution computational time. We also explore the use of DermoGAN, with no retraining, to images acquired using other image acquisition techniques, and the possibility of training it on datasets made entirely of synthetic images.

2 Methods

2.1 Identifying keratinocytes on RCM images with DermoGAN

The goal of the proposed DermoGAN model, shown in Fig. 2, is to estimate a mapping G_{A2B} from an RCM image domain (A) towards a binary domain (B). The mapping is learned using two connected complementary tasks. The first one learns RCM images noise and texture model (the likelihood of the image) from two sets of unpaired images: a set of RCM images and a set of (synthetic) binary images (obtained by simulating a prior model). The second task maps Gabor-filtered RCM images (domain C), *i.e.,* where membranes have been highlighted, into (synthetic) binary images, to learn the global geometrical structure of the epidermal tissue. The combination of the two tasks makes the overall model structure-aware, allowing us to denoise RCM images while keeping the position and integrity of the membrane.

The proposed architecture is fully unsupervised, thus circumventing the obstacle of limited labelled data. Additionally, as it does not rely on training with a manually generated ground truth, as opposed to supervised

approaches like U-net, its accuracy cannot be impacted by incorrectly labelled data, i.e., missing cells in the ground truth or wrong detections.

Each task is a cycle-GAN network, made of two generators, denoted G_{A2B} and G_{B2A} in the first task and G_{C2B} and G_{B2C} in the second task, and two discriminators, denoted D_{B_1} and D_A in the first task and D_C and D_{B_2} in the second task, making a total of 8 networks in the model.

2.1.1 Generator and Discriminator architecture

The generator and discriminator networks form pairs ($G_{A2B}/D_{B_1}$and G_{B2A}/D_A, G_{C2B}/D_{B_2} and G_{B2C}/D_C). A generator takes a 256x256 image as input, down-samples it to extract high-level features and reduce spatial resolution, applies a succession of residual (ResNet) blocks to these features, and then up-samples them to increase the spatial resolution back up and generate the output, as described in Fig. S1. Each generator aims to create realistic target images taking a source image as input. The generators are constrained by an identity loss [20], to ensure that the generator does not modify a target domain image if used as an input encouraging it to be an identity mapping, i.e., $G_{A2B}(B) \approx B$. The two generators in the network should be cycle consistent to ensure that the data is preserved during the translation process and the latter is reversible, *i.e.,* $G_{A2B}\big(G_{B2A}(B)\big) \approx B$ [21].

The weights in all generators were initiated from a Xavier (or Glorot) normal distribution [22] such that the variation of the activations are the same across all layers in order to reduce the risk of the gradient exploding or vanishing and is a random number with a normal probability distribution in the range $\pm\sqrt{\frac{6}{n_i+ n_o}}$, where n_i = 862 is the number of input images (both real RCM images and Gabor-filtered ones), and $n_o = 400$ is the number of output images (synthetic binary images). The weights of the generators were then updated by minimizing 3 loss functions (see Fig. 2).

In the case of G_{A2B},these losses are: (1) Adversarial loss calculated with a mean squared error (MSE) between the generator and its associated discriminator, here D_{B_1}, such that for a pixel at coordinates [i, j] of the generated image Gen_B, it is defined as,

$$MSE\left(D_{B_1}(Gen_B), 1\right) = \frac{1}{n_t}\sum_{i,j}^{n_t}\left(D_{B_1}(Gen_B)(i,j) - 1\right)^2 \quad (1)$$

where n_t is the size of the tensor outputted by the discriminator.

(2) Identity loss with a mean absolute error (MAE) between the input image I_B from domain B and the theoretical identity mapping $Id_{I_B} = G_{A2B}(I_B) \approx I_B$, defined for image I_B at pixel [i, j],

$$MAE\left(Id_{I_B}, I_B\right) = \frac{1}{n_i}\sum_{i,j}^{n_i}\left|Id_{I_B}(i,j) - I_B(i,j)\right| \quad (2)$$

(3) (Forwards or backwards) Cycle consistency loss with a MAE between an input image I_B from domain B and the corresponding reconstructed image $Rec_{I_B} = G_{A2B}\left(G_{B2A}(I_B)\right)$, defined at pixel of coordinates [i, j] as,

$$MAE\left(Rec_{I_B}, I_B\right) = \frac{1}{n_i}\sum_{i,j}^{n_i}\left|Rec_{I_B}(i,j) - I_B(i,j)\right| \quad (3)$$

This loss function participates 10 times more to the update of the generator weights compared to the adversarial MSE loss.

The generators were trained with the ADAM optimizer with an initial learning rate of 0.002, and a decay rate of the gradient exponential moving average of 1.

The discriminators take an image as input and output the classification results (real vs. fake) in a tensor. Each discriminator aims to distinguish between real and generated target images, thus working against its matching adversary generator, which aims to create indiscriminable generated target images. These two networks are connected through the adversarial loss in (1) [21], and the discriminator loss function is defined as,

$$\frac{1}{2}MSE\left(D_{B_1}(Gen_B), 1\right) + \frac{1}{2}MSE\left(D_{B_1}(I_B), 0\right) \quad (4)$$

Training each generator/discriminator pair simultaneously allows the cycle-GAN to learn the bidirectional image-to-image translation between two unpaired domains.

2.1.2 Multi-task approach

RCM images are noisy and heterogeneous due to tissue-induced scattering [23], and are non-specific to organelles and macro-structures. This makes the identification of keratinocytes on RCM image a challenging task, whether done manually or automatically. In this case, cell identification requires two simultaneous tasks to capture the breath of information in confocal images: noise removal and membrane identification. Multi-task learning allows for concurrent execution of these two related tasks, improving overall performance by leveraging complementary information and sharing representations [24]. This reasoning mimics expert's approach to manual cell identification on RCM images, i.e., focusing on bright tube-like membranes while ignoring the bright blob-like noise.

Noise removal was performed using a first cycle-GAN network, learning the translation between RCM images and binary images obtained by simulating the structure of keratinocytes[9] (a described in section 3.3.1). Whereas membrane identification was performed by learning the mapping between binary images and Gabor-filtered RCM images, i.e., where membranes were highlighted.

The multi-task model is optimized through soft-sharing of parameters [25], as the two tasks do not share any hidden layer, but are connected through their loss function, as shown in Fig. 2. Indeed, at each update of the loss function, those associated with the generators creating the binary images, i.e., G_{A2B} and G_{C2B}, are updated through their regular optimization, and then the maximum value of the two trios of losses is set as the loss function for both generators, in order to synchronize training across the two tasks of noise removal and membrane identification.

2.2 Refining the results with star-convex polygons

The proposed method is a top-down approach to cell detection. The DermoGAN roughly localizes individual cell locations, but post-processing is required, as shown in Fig. 3 and Fig. 4. Indeed, applying the obtained mapping G_{A2B} to a locally normalized RCM image results in an incomplete binary image. To guarantee that the outside contour of tissue where the keratinocytes are detected is closed, we compute the alpha shape [26] of the incomplete binary mask at a set level of refinement, such that the tissue comprises only one volume per external contour and is not broken down into smaller shapes, and that the alpha-shape contour matches the actual tissue area. Small holes in the membrane are then closed using a connected components analysis [27].

We assume that all cells are star-convex shapes. However non-star-convex polygons can result from the false merging of two or more cells due to the lack of contrast on the membranes. To split these shapes, we use the pre-trained convolutional neural network StarDist [18] to detect star-convex polygons within the contours detected by the DermoGAN model, consequently refining our results, countering any missed cells and reducing the number of false negatives, as shown in Fig. 3.

3 Experiments & Results

3.1 Dataset

RCM images were captured using a Vivascope 1500 (Lucid, Inc., Rochester, New York) reflectance confocal microscope, on the volar forearm of 60 children (3 months – 10 years) and 20 adult women (25 – 40 years), and on the volar forearm and cheek of 80 other adult women (40 – 80 years). All participants have Fitzpatrick types between I and III, were in good health and with no history of skin disease. The study was initiated following approval from an independent institutional review board and in accordance with the Declaration of Helsinki (studies 19.0198 and 20.0022). Subjects or their guardian gave written informed consent prior to study initiation.

The image size was 1000x1000 pixels, corresponding to 500x500µm², with a resolution of 1µm² per pixel. A region of interest (ROI) mask was generated for and applied to all images used across all six tested methods.

176 The ROI was identified by distinguishing the tissue from the dark background, due to the skin micro-relief
177 lines, using a morphological-geodesic-active-contour, and removing non-informative areas in the tissue, due
178 to low contrast and a drop in signal-to-noise ratio, through a texture classification with a support vector
179 machine on four features of the grey level co-occurrence matrix (homogeneity, contrast, dissimilarity, and
180 energy[9]).
181 Images used in DermoGAN, U-net, and both approaches using a cycle-GAN, were of size 256x256pixels
182 and obtained by splitting the full image into 9 non-overlapping square patches of 256x256 pixels. The full
183 image analysis pipeline network used full RCM images [11].
184 DermoGAN and both Cycle-GAN models were trained using the same 862 RCM images of size 256*256
185 pixels, and 400 synthetic images. The number of synthetic images used in the training of the models was
186 determined empirically. Indeed, we noticed that adding more images did not improve performance but
187 increased computational time. Using 400 images was the right balance between performance and
188 computational time and power. The RCM images represent both volar forearm and cheek, and include
189 participants ages 0 to 80 years.
190 Image classification in one of the four epidermal layers was obtained using a hybrid deep learning algorithm
191 [28], allowing to focus only on images of the *stratum granulosum* (SG) and *stratum spinosum* (SS), where
192 keratinocytes are visible and identifiable on RCM images, and arranged in a honeycomb pattern [29] inside of
193 islands surrounded by dark grooves representing micro-relief lines [9,30].
194 Ground truth used for the evaluation of all the tested approaches was generated by a single expert manually
195 pointing out cell centers on 9 RCM images of 7 subjects (children and adult women), ages 5 months to 35
196 years old. The use of ground truth generated by a sole expert in the evaluation of our results is a limitation of
197 our work but is justified by the previously documented inter-expert variation in keratinocytes identification[9].
198 *Generated binary synthetic images.* Binary (synthetic) images, of size 256x256 pixels, were created by
199 generating a random tissue mask using random Bezier curves. Within these shapes, seeds, mimicking cell
200 centers, were used to initiate a Voronoi tessellation, which have been previously used to represent both skin

cells[31] and other types of cells[32–34]. These synthetic binary images are user-controlled, and the associated ground truth is given by the seed locations. The seeds are generated with a 'hard-core' process simulation controlled by the density parameter (see Fig. 1).

Synthetic RCM images were generated by adding different levels of noise and brightness heterogeneity to the binary images, as shown in Fig. 1, and detailed in [9].

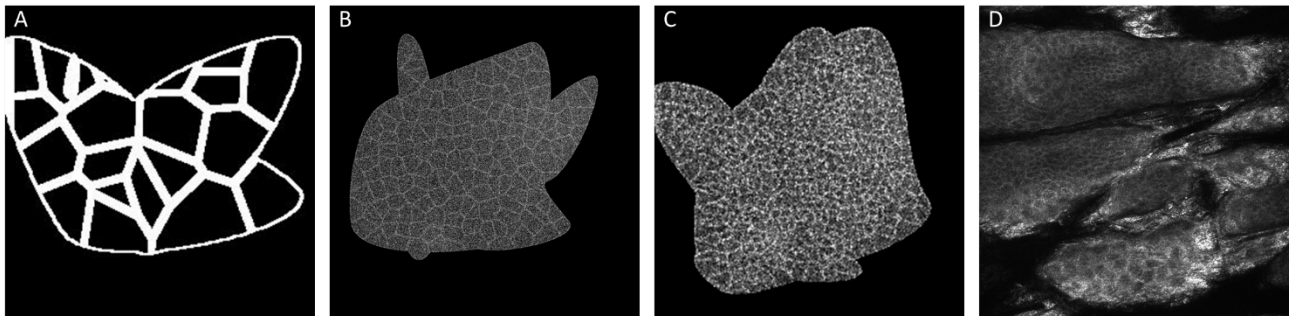

Fig. 1 Synthetic images used in the training of the models. (A) A synthetic binary image used in the DermoGAN training. (B and C) RCM images of different noise levels and cell sizes used in U-net training. (D) Real RCM image. RCM, Reflectance confocal microscopy.

3.2 DermoGAN implementation details

The used model was trained for 5172 epochs on 46.9 CPU cores and 85.4 GiB of RAM. Training took approximately 4 days. All deep learning models were implemented using PyTorch.

Intermediate models were saved every 400 epochs, and the model with best performance, i.e., accuracy metrics (precision and recall summarized into the F1-score), was chosen.

3.3 Comparison to other automated methods

The proposed method is compared to 7 other approaches: (1) a deep learning approach based on the U-net architecture [17], (2) a StarDist algorithm applied to Gabor-filtered RCM images, (3) a cycle-GAN trained to translate RCM images into binary images, (4) a second cycle-GAN trained to turn Gabor-filtered RCM images into binary images, (5) a full image analysis pipeline based on traditional image analysis

methods[9,11,35], (6) CellPose algorithm for cellular segmentation, and (7) DermoGAN followed by postprocessing with CellPose.

The proposed combination of cycle-GAN models into a multi-task approach improves results by mimicking manual expertise, disregarding noise to focus on membrane location and tissue structure.

U-net. A U-net architecture, pre-trained on the 2012 ImageNet Large Scale Visual Recognition Challenge dataset [36] with an efficientnetb3 backbone [37], was further trained on 43 real RCM images of size 256*256 (4 participants, 20 – 35 years) and 203 synthetic RCM images and tested on 13 real RCM images and 68 synthetic RCM images. The corresponding ground truths were obtained by the same expert and do not include the 9 images used in the model evaluation. The network used a combination of two loss functions: Dice loss [38] and Focal loss [39] to account for class imbalance between cell membranes and background. The model is trained with the ADAM optimizer with an initial learning rate of 0.0001, a batch size of 64, and a sigmoid activation function. The model was trained for 500 epochs on 46.9 CPU cores and 85.4 GiB of RAM. Training took approximately three days. The selected U-net model was obtained following multiple iterations, as described in [9,40].

StarDist applied to Gabor-filtered images. A Gabor-filter was applied to ROI-masked RCM images to highlight membrane positions. The result was normalized with a histogram equalization and binarized with a Gaussian adaptive thresholding. A pre-trained StarDist was then applied to the binary masked-Gabor-filtered RCM image.

Cycle-GAN based models. Two cycle-GAN models were trained on 862 RCM images and 400 binary images, each one representing a task in the DermoGAN architecture, to evaluate each model independently, and later emphasize the importance of combining the two tasks into one architecture. The first one aimed to translate RCM images into binary images, whereas the second sought to convert Gabor-filtered RCM images into binary images. Both tested cycle-GANs models were refined using star-convex shapes detection as performed in the DermoGAN architecture. Training was performed for 12068 epochs on 46.9 CPU cores and 85.4 GiB of RAM and took two days.

Full Image Analysis Pipeline (FIAP). A 3-step pipeline for keratinocytes detection [9,11,35] based on membrane detection using image filters was applied to full RCM images of size 1000x1000 pixels. First, the region of interest containing the epidermal cells is identified. Texture filters (Gabor and Sato filters) are then applied to the image to accentuate tube like structures (membranes) within the region of interest and identify individual cells within it. The detected contours are then post-processed using prior biological knowledge[41] on expected cell size to remove contours that are too small ($cell\ area < 100\ pixels$ for contours detected on RCM images of the SG and $cell\ area < 50\ pixels$ for contours detected on RCM images of the SS). The texture filters were reapplied locally on detected regions presumed to be too big to be considered as a single cell and divide them into multiple cells if needed. The FIAP is applicable to images of the SG and SS and requires a different set of parameters for each layer which were determined manually. Computational time is 7-10 minutes per image depending on image complexity and required post-processing steps.

CellPose. The pretrained CellPose model was applied, without retraining, to the full RCM images of size 1000x1000 pixels where the ROI had already been identified. CellPose is a generalist single-class instance segmentation algorithm optimized for cellular segmentation across different microscopy modalities. The model was run with the following parameters: ROI diameter = 50 for SG images, ROI diameter = 50 for SS images, and flow threshold = 0.4, cell probability threshold = 0.2, and stitch threshold = 0.

DermoGAN followed by CellPose. We replaced the previously described postprocessing step with StarDist by the TissueNet cell model available in CellPose. The model was run with the following parameters: ROI diameter = 50 for SG images, ROI diameter = 50 for SS images, and flow threshold = 0.4, cell probability threshold = 0.2, and stitch threshold = 0.

All eight tested methods were evaluated against the same RCM images. While the two cycle-GAN based approaches and DermoGAN were trained on the same images, U-net was not. Indeed, U-net is a supervised learning approach and ground truth was not available for all images used in the training of the other tested methods. This may limit the comparability of the approaches to each other, but also highlights the importance of unsupervised learning methods which are not limited by the available labelled data.

3.4 Keratinocytes identification results

The proposed DermoGAN architecture was evaluated using nine full RCM images, each divided into nine patches. Accuracy (precision and recall summarized into the F1-score) was calculated using d-accuracy [42] against a manually obtained ground truth and compared to results obtained with the eight described methods, as shown in Fig. 4 and Table 1.

The poor performance of the pre-trained U-net model augmented with real and synthetic RCM images with Focal and Dice loss functions is in part due to the limited training set. Being a supervised approach, it may also suffer from missing cells in the ground truth used for training, due to inter and intra-expert variability and subjectivity in manual keratinocytes identification on RCM image[9], and from membranes in the ground truth images created by Voronoi tessellation initiated from manually determined cell centers, not matching the actual membrane position in RCM images.

The pre-trained StarDist applied to Gabor-filtered images also performs poorly. Indeed, although the Gabor filter highlights most membranes, it may also highlight noise, due for example to organelles, leading to false positives and low precision. Although the StarDist post-processing greatly improves results by segmenting (correctly or not) the detected contours into star-convex shapes, it does not manage to correct for all missing cells, leading to false negatives and consequently low recall, and overall low F1-score.

Both cycle-GAN-based approach have low F1-scores as they fail to detect complete membranes, as shown in Fig. 3. Indeed, the cycle-GAN model trained on RCM images and binary images, struggles to distinguish between noise and microstructures making up the membranes. On the other hand, the cycle-GAN trained on Gabor-filtered images with binary images is corrupted by the spatial correlation of noise and fails to detect any structure present in the image, as seen in Fig. 3, which also hints at the reason behind the DermoGAN greater performance. Indeed, it seems that adding up the two independent cycle-GAN outputs, would close most holes in the detected membranes by focusing on membrane detection and omitting any noise visible in them.

The pretrained CellPose applied to the full RCM images without retraining has a low F1-score, and better performance for SG images compared to SS images, as shown in Table S1. Overall precision is higher than recall, showing that this method is more conservative in detecting positive instances.

Overall, the last test method using the TissueNet cell model from CellPose as a postprocessing step to the proposed DermoGAN model performs well against both SG and SS images. This method has a good F1-score but tends to have higher precision than recall. It is still outperformed by the selected method (DermoGAN followed by StarDist-based postprocessing) which has a better trade-off between recall and precision, and therefore is better suited for keratinocytes identification on RCM images.

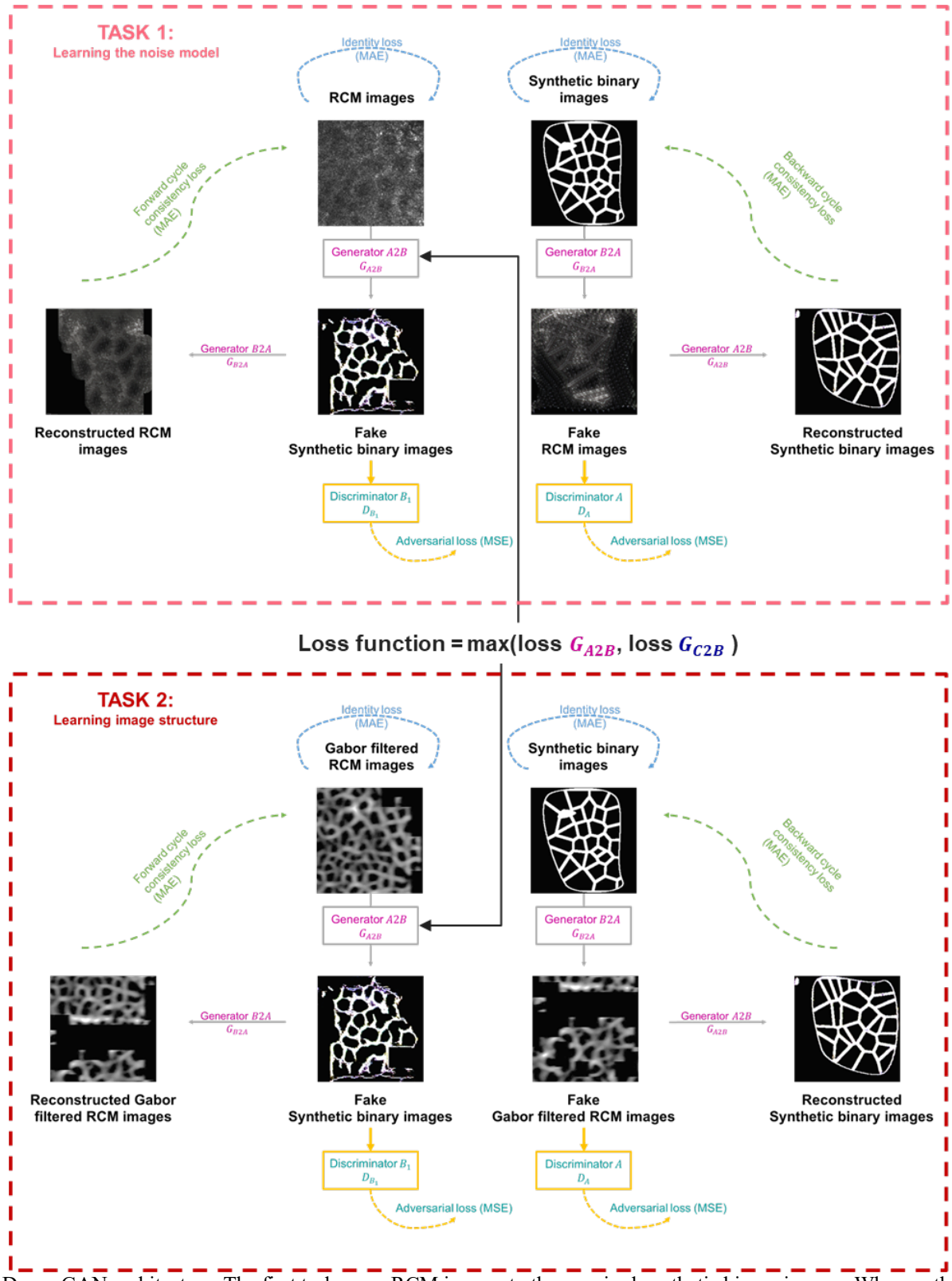

Fig. 2 DermoGAN architecture. The first task maps RCM images to the unpaired synthetic binary images. Whereas the second task learns the structure RCM images of the epidermis by translating Gabor filtered RCM images into binary images. .

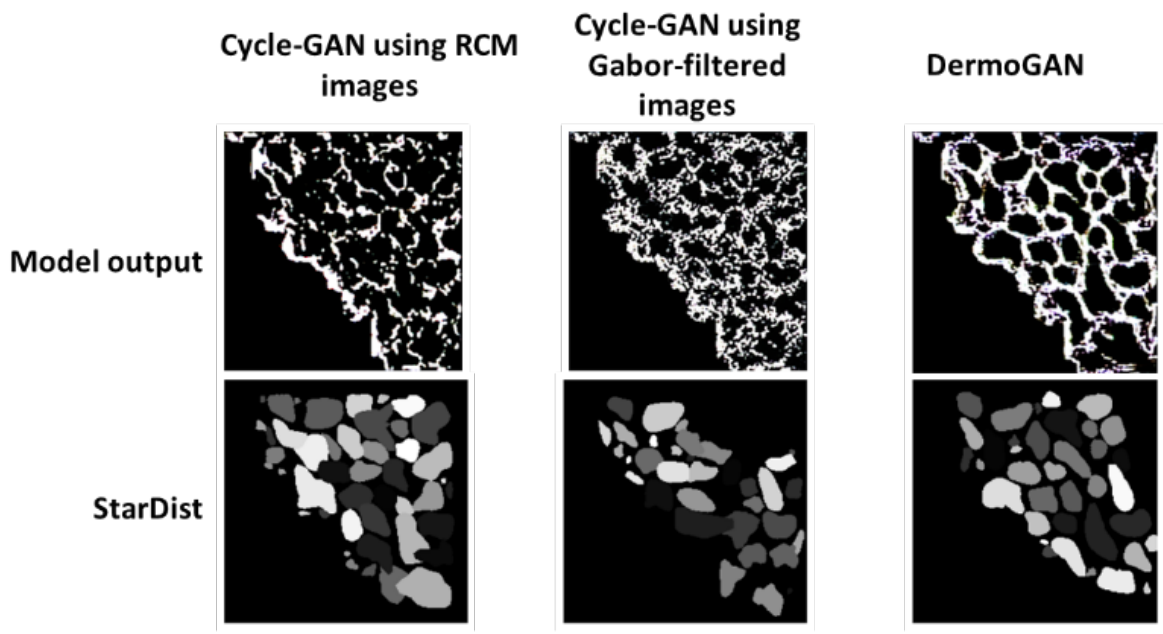

Fig. 3 Comparison of the two cycle-GAN based approaches and the proposed DermoGAN. DermoGAN outperforms both methods.

Both DermoGAN and FIAP outperform the other models, as shown in Tables 1, S1, and S2, and show a great trade-off between precision and recall. DermoGAN has a higher F1-score than FIAP for 6/9 images. The first seems to favor recall and is less likely to miss existing cells and produce false negatives, whereas the second seems to favor precision, and is less likely to invent cells and create false positive detections.

The DermoGAN architecture does not require manual parametrization nor a different set of parameters per epidermal layer, contrarily to the FIAP. This argues in favor of the DermoGAN network since multiple epidermal layers are often present in one RCM image. Once trained, its execution time is faster. It is based on the discovery of potentially unknown patterns in the image, making it less explainable than the FIAP. The latter is built on membrane detection using tubeness filters, with all its parameters being determined using general prior knowledge on the morphological features of the studied tissue. It is well documented that keratinocytes area increases with age and differs from one body site to another, and thus general parameters determined on a specific dataset may not be appropriate for all images. This point favors the

DermoGAN architecture as more adaptable to different datasets and potentially to different image acquisition techniques and/or observed tissue.

Table 1 Comparison of median f1-score (computed with d-accuracy[42]) for all eight tested approaches.

	U-net based architecture	StarDist on Gabor-filtered images	Cycle-GAN using RCM images	Cycle-GAN using Gabor-filtered images	FIAP	CellPose	DermoGAN with CellPose	DermoGAN with StarDist
Training data size	Training: 43 real RCM images 203 synthetic RCM images Testing: 13 real RCM images and 68 synthetic RCM images.	No training required	862 RCM images and 400 binary images	862 RCM images and 400 binary images	No training required	No training required	862 RCM images and 400 binary images (No training required for CellPose)	862 RCM images and 400 binary images (No training required for starDist)
Median F1-score	47,9	38,9	41,6	30,4	65,1	35,0	62,4	69
Standard deviation	12,2	12,1	6,4	8,3	7,4	16,5	8,7	4,2

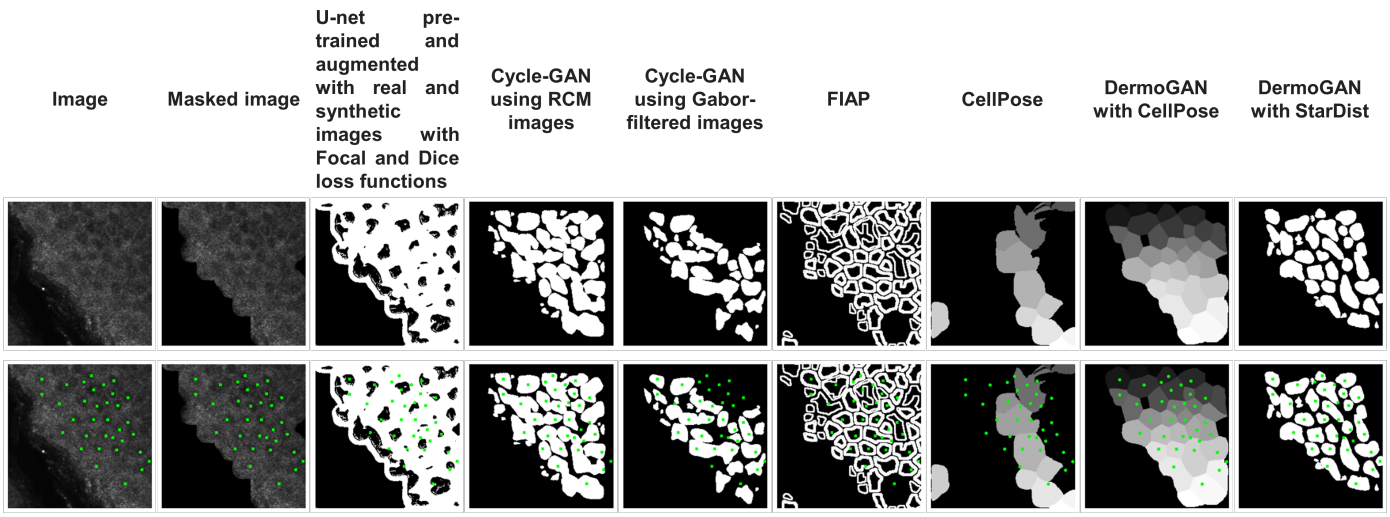

Fig. 4 RCM image analyzed with the 8 presented algorithms. DermoGAN outperforms 5 out of 6 other methods on all images, and outperforms FIAP on 6/9 images.

4 Discussion

Although the presented method was trained using RCM images for the detection of keratinocytes, we hypothesized that it can be extended without retraining to images generated by other instruments. Indeed, multi-task learning methods tend to perform well on domain adaptation and generalization and are therefore less data *dependent.* On the other hand, such adaptability can lower pixel-level segmentation, and therefore is more suited when the accuracy is calculated at object level and not at pixel level. In the following paragraphs, we aim to explore the potential applications of DermoGAN to images acquired using different imaging modalities. Accuracy metrics will not be presented, as this is a preliminary step to future research.

To explore the possibility of applying DermoGAN to other images without retraining, we applied the presented model trained on RCM images to fluorescence microscopy images and compared the obtained results to 18 thresholding methods and to the pretrained CellPose cyto model, as shown in Fig. 5. We observed that, although trained on different images of a different tissue, DermoGAN managed to identify membranes while omitting the noisy background, and outperformed traditional thresholding methods. It had similar performance to the pretrained

CellPose model which has better performance at image borders but misses a cell. It is important to note, that the fluorescence microscopy image does show a similar tissue organization, i.e., cohesive tissue with cells *sharing* membranes, to RCM images of the epidermis. However, when tested on cell culture images where cells were not always confluent, we noticed a loss in accuracy when using DermoGAN for cell identification. We therefore trained a second model using a different prior for simulating binary images and a different filter to enhance contours, that will be referred to for simplicity as DermoGAN2, on images where cells were not confluent.

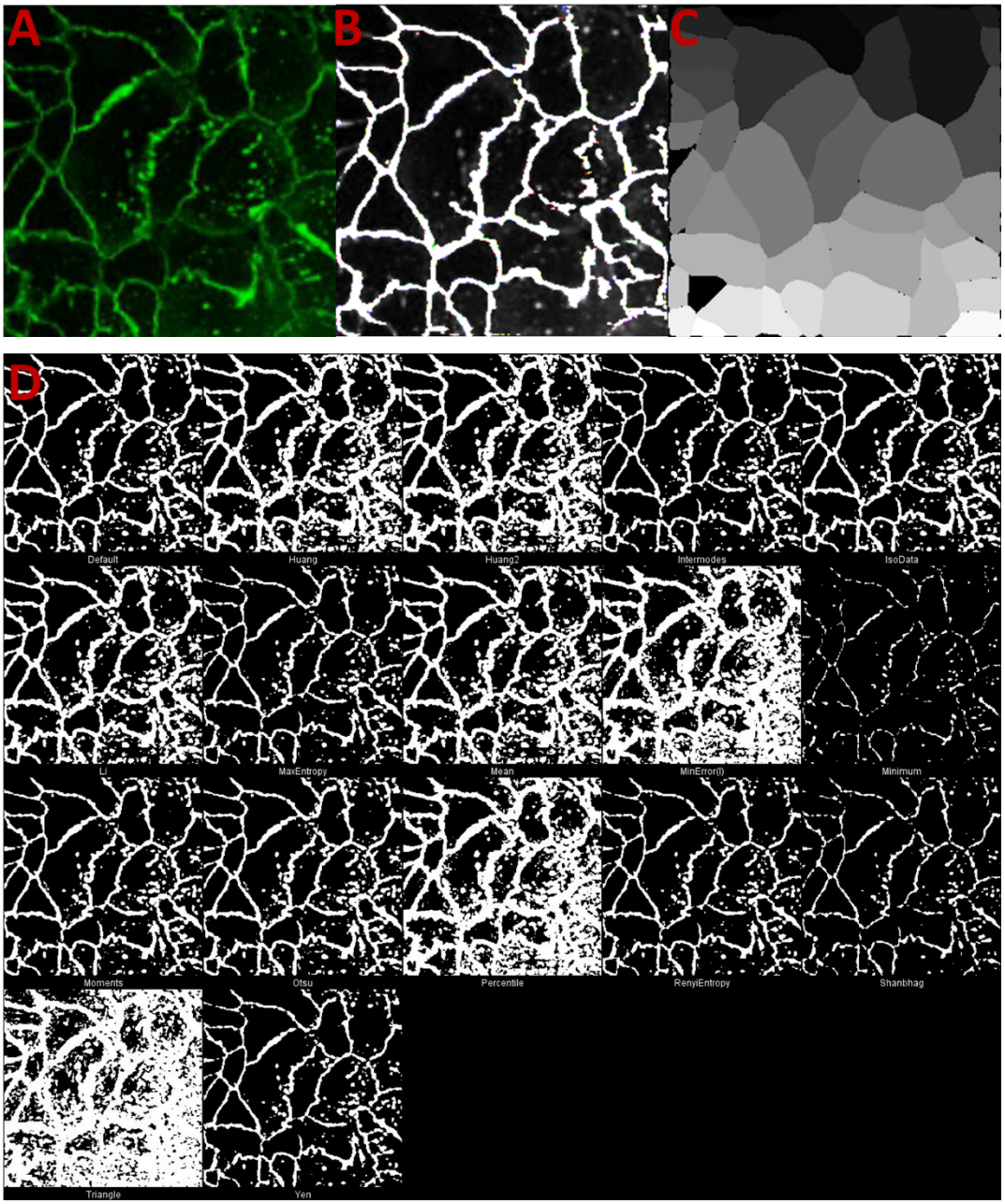

Fig. 5 DermoGAN can be extended to images acquired different imaging techniques using retraining and outperforms traditional thresholding algorithms. (A) Input florescence microscopy images. (B) Output of DermoGAN applied to image A. (C) Output of CellPose cyto model. (D) Application of 18 thresholding approaches to the same image.

4.1 Retraining the model with only synthetic images

DermoGAN2 was trained entirely on synthetic images. This served as a test of the generalization of the method when the available dataset is even more limited and serves to prove than the combination of the two tasks in the proposed model can capture general information

and therefore can be extended to different images and tissues with similar organization, architecture, or texture, even when the images of interest were not included in the training set. The first task in DermoGAN2 maps synthetic non-confluent images created using the SIMCEP software for the simulation of fluorescence microscope images of cell populations [43] (Fig. 6A) to binary non-confluent images (Fig. 6C). The binary images were obtained by simulating a marked point process embedding a constraint on overlap between objects defined by disks[44]. While the second task, aims to learn the translation of Canny-filtered synthetic non-confluent images (Fig. 6B) towards the same binary non-confluent images.

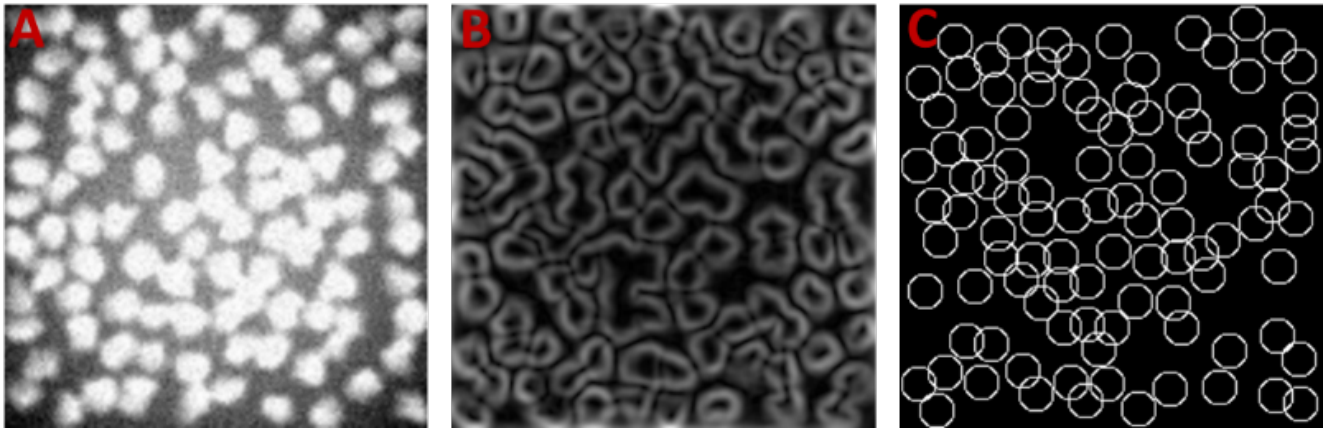

Fig. 6 DermoGAN2 was trained entirely on synthetic images. (A) Synthetic non-confluent images created using the SIMCEP. (B) Canny-filtered non-confluent images created using the SIMCEP. (C) Binary non-confluent images.

The resulting DermoGAN2 was then applied to images of cell cultures, and on mass spectroscopy images.

4.1.1 DermoGAN2 on cell culture images

We applied DermoGAN2 on an image of BV-2 microglial cells derived from C57/BL6 murine from the LIVECell dataset [45] as seen in Fig. 7. We obtained an accurate segmentation of the cells on the image. To avoid border effects in the image, a 10px frame was applied to the image. We compare DermoGAN output to the pretrained CellPose cyto model (*diameter = 10, flow*

threshold = 0.4, cell probability threshold = 0, and stitch threshold = 0). We observe a drop in performance for both methods.

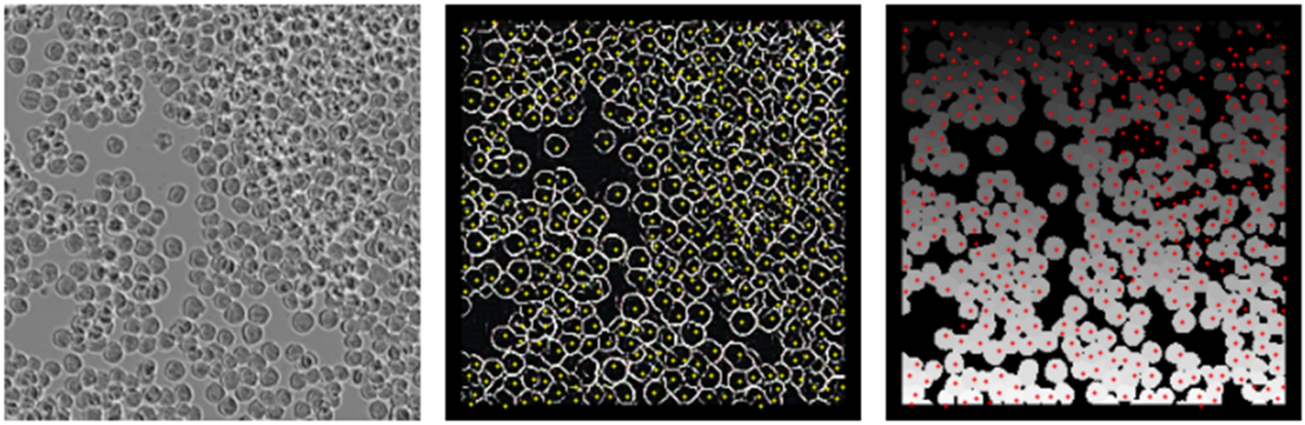

Fig. 7 DermoGAN2 applied to an image of confluent BV-2 cells (on the left), resulted in accurate detection of cells with both DermoGAN (in the middle) and CellPose (on the right). Manually determined cell centers were plotted on DermoGAN2 output in yellow, and in red on the CellPose output.

We also applied DermoGAN2 to an image of the SK-BR-3 human breast cancer cell line, where cells display morphological heterogeneity (Fig. 8A). Good cell detection was observed (Fig. 8B) on most cells when contrast is high enough. This proves that DermoGAN2 can be extended to different cell shapes and is not limited to the detection of the circular cells it was trained on, and that it is not restricted by the aspect of the cells. Indeed, the synthetic images used for training the model where more similar to fluorescence images, with high luminosity, which is not the case of the tested cell culture images. Similar results can be observed when applying the pretrained CellPose live-cell model (*diameter = 10, flow threshold = 0.4, cell probability threshold = 0, and stitch threshold = 0*) (Fig. 8D), which seems to visually outperform DermoGAN in this case.

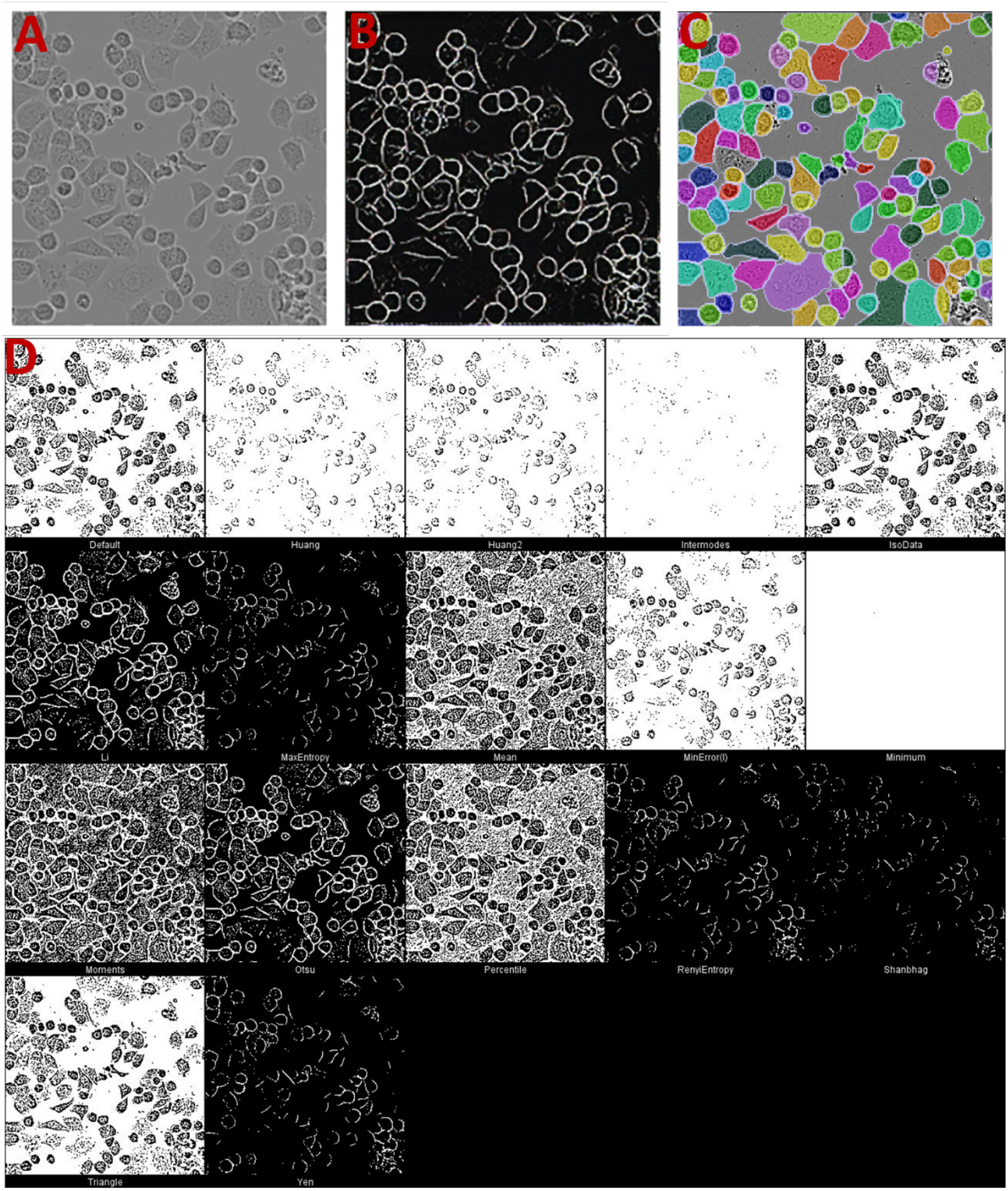

Fig. 8 DermoGAN2 applied to an image of SK-BR-3 cells (A) resulted in accurate detection with both DermoGAN (B) of cells and CellPose (C), compared to thresholding methods (D).

4.1.2 DermoGAN2 on mass spectroscopy images

Similar observations were made on mass spectroscopy images, where DermoGAN2 managed to detect hazy cell contours, with a tendency to merge close cells into one detected region, as

shown in Fig. 9. This can be solved by post-processing using StarDist as done on RCM images. When applying the pretrained CellPose cyto model to the same mass spectroscopy image, we observe a drop in detection in areas suffering from a drop in contrast, which is not the case for DermoGAN, but borders are sharper when using CellPose compared to DermoGAN.

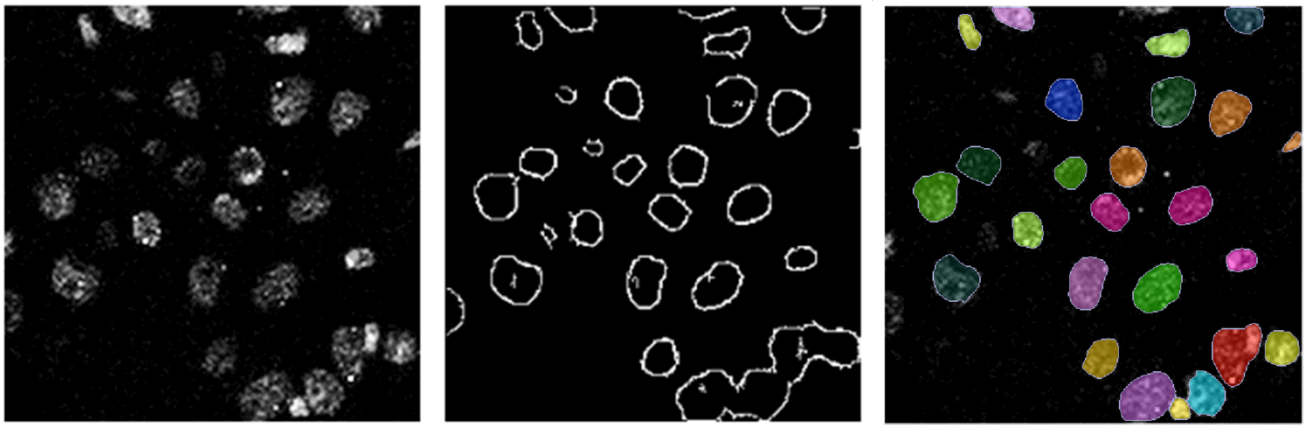

Fig. 9 DermoGAN2 applied to a mass spectroscopy image (on the left) resulted in accurate cell detection with both DermoGAN (in the middle) with merging of adjacent cells, and CellPose (on the right).

The success of DermoGAN in segmenting cohesive tissues, and that of DermoGAN2 in detecting non-confluent cells, highlights the importance of the binary denoised images domain (domain B in DermoGAN). This domain serves as a prior domain incorporating anterior knowledge in the model by describing the structure of the studied tissue. In DermoGAN, the prior is represented as a tissue island containing adjacent cells of similar size, while in DermoGAN2, this prior is represented by circular non-confluent cells. This prior domain summarizes our degree of certainty concerning the studied tissue and steers the training of the model towards the right solution. Therefore, to obtain the best results, the appearance of the tissue should guide the choice of the appropriate DermoGAN model based on the corresponding prior domain.

In all the iterations of DermoGAN so far, training was performed using cell membranes positions. It would be of interest to see if the performance of the models could be improved by replacing the images representing cell membranes in the training by images representing the

entire cell (cell membrane and inside of the cell), therefore having two classes in the data, cell vs. background instead of cell membrane vs. background as currently done.

5 Conclusion

This paper has presented a novel multitask cycle-GANs architecture for the identification of keratinocytes on RCM images and was compared to seven other methods. Supervised deep learning approaches obtained poor scores due to the lack of annotated data, even when using transfer learning. Unsupervised learning, such as cycle-GAN, failed to capture information at different scales simultaneously. Therefore, the FIAP approach outperformed these attempts. However, the proposed DermoGAN which combines two cycle-GANs to embed both local and global structure information, outperformed the classical FIAP, in terms of accuracy and execution time.

We showed that the proposed fully unsupervised architecture be used with or without retraining on other types of imaging and tissue types, bypassing the problem of required annotated data and potential label noise/missing labels provided the creation of simulated data, and that it is not limited by the training set but rather determined by the prior data domain, i.e., tissue organization and architecture in the training set. It would be of interest to generate new prior domains using marked point processes to generate more specific priors, which would help extend the use of DermoGAN to images with multiple cell types, or when the spatial dependence between different structures is important.

Disclosures

IL and GNS were employees of Johnson & Johnson Santé Beauté France at the time the research was conducted.

Acknowledgments

This research was fully funded by Johnson & Johnson Santé Beauté France.

Data and code availability

The data and code that support the findings of this study are available upon reasonable request from the corresponding author, IL.

References

1. A. Levine and O. Markowitz, "Introduction to reflectance confocal microscopy and its use in clinical practice," JAAD Case Rep **4**(10), 1014–1023 (2018) [doi:10.1016/j.jdcr.2018.09.019].
2. N. S. K. Haytoglu et al., "Assessment of skin photoaging with reflectance confocal microscopy," Skin Res Technol **20**(3), 363–372 (2014) [doi:10.1111/srt.12127].
3. Imane Lboukili, Georgios Stamatas, and Xavier Descombes, "Automating reflectance confocal microscopy image analysis for dermatological research: a review," 7, Journal of Biomedical Optics **27**(7), 070902 (2022) [doi:10.1117/1.JBO.27.7.070902].
4. S. Zorgui et al., "A Convolutional Neural Network for Lentigo Diagnosis," The Impact of Digital Technologies on Public Health in Developed and Developing Countries **12157**, 89–99 (2020) [doi:10.1007/978-3-030-51517-1_8].
5. A. Halimi et al., "An unsupervised Bayesian approach for the joint reconstruction and classification of cutaneous reflectance confocal microscopy images," 1 August 2017, 241–245 [doi:10.23919/EUSIPCO.2017.8081205].
6. A. Bozkurt et al., "A Multiresolution Convolutional Neural Network with Partial Label Training for Annotating Reflectance Confocal Microscopy Images of Skin," arXiv:1802.02213 [cs] (2018).
7. K. Kose et al., "A machine learning method for identifying morphological patterns in reflectance confocal microscopy mosaics of melanocytic skin lesions in-vivo," 968908 (2016) [doi:10.1117/12.2212978].
8. S. Koller et al., "In vivo reflectance confocal microscopy: automated diagnostic image analysis of melanocytic skin tumours," J Eur Acad Dermatol Venereol **25**(5), 554–558 (2011) [doi:10.1111/j.1468-3083.2010.03834.x].
9. I. Lboukili, G. Stamatas, and X. Descombes, "Automatic granular and spinous epidermal cell identification and analysis on in vivo reflectance confocal microscopy images using cell morphological features," J Biomed Opt **28**(4), 046003 (2023) [doi:10.1117/1.JBO.28.4.046003].
10. D. Gareau, "Automated identification of epidermal keratinocytes in reflectance confocal microscopy," J Biomed Opt **16**(3), 030502 (2011) [doi:10.1117/1.3552639].
11. I. Lboukili, G. Stamatas, and X. Descombes, "Automatic cell identification and analysis on in vivo reflectance confocal microscopy images of the human epidermis," in SPIE Photonics Europe 2022, Strasbourg, France (2022) [doi:10.1117/12.2626777].
12. J.-Y. Zhu et al., "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks" (2017) [doi:10.48550/arXiv.1703.10593].
13. R. Gulakala, B. Markert, and M. Stoffel, "Generative adversarial network based data augmentation for CNN based detection of Covid-19," 1, Sci Rep **12**(1), 19186, Nature Publishing Group (2022) [doi:10.1038/s41598-022-23692-x].
14. N. Hernandez-Cruz, D. Cato, and J. Favela, "Neural Style Transfer as Data Augmentation for Improving COVID-19 Diagnosis Classification," SN Computer Science **2**(5) (2021) [doi:10.1007/s42979-021-00795-2].

15. H. Huang, P. S. Yu, and C. Wang, “An Introduction to Image Synthesis with Generative Adversarial Nets,” arXiv:1803.04469, arXiv (2018) [doi:10.48550/arXiv.1803.04469].
16. R. Barth, J. Hemming, and E. J. van Henten, “Improved Part Segmentation Performance by Optimising Realism of Synthetic Images using Cycle Generative Adversarial Networks,” arXiv:1803.06301, arXiv (2018) [doi:10.48550/arXiv.1803.06301].
17. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, N. Navab et al., Eds., pp. 234–241, Springer International Publishing, Cham (2015) [doi:10.1007/978-3-319-24574-4_28].
18. U. Schmidt et al., “Cell Detection with Star-convex Polygons,” pp. 265–273 (2018) [doi:10.1007/978-3-030-00934-2_30].
19. C. Stringer et al., “Cellpose: a generalist algorithm for cellular segmentation,” 1, Nat Methods **18**(1), 100–106, Nature Publishing Group (2021) [doi:10.1038/s41592-020-01018-x].
20. Y. Taigman, A. Polyak, and L. Wolf, “Unsupervised Cross-Domain Image Generation” (2016) [doi:10.48550/arXiv.1611.02200].
21. I. J. Goodfellow et al., “Generative Adversarial Networks,” arXiv:1406.2661, arXiv (2014) [doi:10.48550/arXiv.1406.2661].
22. X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256, JMLR Workshop and Conference Proceedings (2010).
23. N. Ks, G. D, and R. M, “Skin imaging with reflectance confocal microscopy,” Seminars in cutaneous medicine and surgery **27**(1), Semin Cutan Med Surg (2008) [doi:10.1016/j.sder.2008.01.006].
24. M. Crawshaw, “Multi-Task Learning with Deep Neural Networks: A Survey” (2020) [doi:10.48550/arXiv.2009.09796].
25. P. Vafaeikia, K. Namdar, and F. Khalvati, “A Brief Review of Deep Multi-task Learning and Auxiliary Task Learning” (2020) [doi:10.48550/arXiv.2007.01126].
26. H. Edelsbrunner, “Smooth surfaces for multi-scale shape representation,” in Foundations of Software Technology and Theoretical Computer Science, pp. 391–412, Springer, Berlin, Heidelberg (1995) [doi:10.1007/3-540-60692-0_63].
27. C. Fiorio and J. Gustedt, “Two linear time Union-Find strategies for image processing,” Theoretical Computer Science **154**(2), 165–181 (1996) [doi:10.1016/0304-3975(94)00262-2].
28. P. Kaur et al., “Hybrid deep learning for Reflectance Confocal Microscopy skin images,” in 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 1466–1471 (2016) [doi:10.1109/ICPR.2016.7899844].
29. S. Guida et al., “Reflectance Confocal Microscopy of Aging Skin and Skin Cancer,” Dermatol Pract Concept **11**(3), e2021068 (2021) [doi:10.5826/dpc.1103a68].
30. M. Rajadhyaksha et al., “In vivo confocal scanning laser microscopy of human skin II: advances in instrumentation and comparison with histology,” J Invest Dermatol **113**(3), 293–303 (1999) [doi:10.1046/j.1523-1747.1999.00690.x].
31. J. Bensaci et al., “Geometrical and topological analysis of in vivo confocal microscopy images reveals dynamic maturation of epidermal structures during the first years of life,” J Biomed Opt **20**(9), 095004 (2015) [doi:10.1117/1.JBO.20.9.095004].
32. R. Marcelpoil and Y. Usson, “Methods for the study of cellular sociology: Voronoi diagrams and parametrization of the spatial relationships,” Journal of Theoretical Biology **154**(3), 359–369 (1992) [doi:10.1016/S0022-5193(05)80176-6].
33. F. Sheikhzadeh et al., “Quantification of confocal fluorescence microscopy for the detection of cervical intraepithelial neoplasia,” BioMedical Engineering OnLine **14**(1), 96 (2015) [doi:10.1186/s12938-015-0093-6].
34. G. Bigras et al., “Cellular sociology applied to neuroendocrine tumors of the lung: quantitative model of neoplastic architecture,” Cytometry **24**(1), 74–82 (1996) [doi:10.1002/(SICI)1097-0320(19960501)24:1<74::AID-CYTO9>3.0.CO;2-I].
35. I. Lboukili, G. N. Stamatas, and X. Descombes, “Age-dependent changes in epidermal architecture explored using an automated image analysis algorithm on in vivo reflectance confocal

microscopy images," Skin Research and Technology **29**(5), e13343 (2023) [doi:10.1111/srt.13343].
36. O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," Int J Comput Vis **115**(3), 211–252 (2015) [doi:10.1007/s11263-015-0816-y].
37. M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks" (2019) [doi:10.48550/arXiv.1905.11946].
38. C. H. Sudre et al., "Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations" (2017) [doi:10.1007/978-3-319-67558-9_28].
39. T.-Y. Lin et al., "Focal Loss for Dense Object Detection" (2017) [doi:10.48550/arXiv.1708.02002].
40. I. Lboukili, "Analysis and characterization of the tissue structure of the epidermis from confocal imaging," phdthesis, Université Côte d'Azur (2023).
41. S. Gonzalez and Y. Gilaberte, "In vivo reflectance-mode confocal microscopy in clinical dermatology and cosmetology," International journal of cosmetic science **30**, 1–17 (2008) [doi:10.1111/j.1468-2494.2008.00406.x].
42. E. Debreuve, "DAccuracy," GitLab, <https://gitlab.inria.fr/edebreuv/daccuracy/-/tree/master/daccuracy> (accessed 25 January 2022).
43. A. Lehmussola et al., "Computational framework for simulating fluorescence microscope images with cell populations," IEEE Trans Med Imaging **26**(7), 1010–1016 (2007) [doi:10.1109/TMI.2007.896925].
44. X. Descombes, "Multiple objects detection in biological images using a marked point process framework," Methods **115**, 2–8 (2017) [doi:10.1016/j.ymeth.2016.09.009].
45. C. Edlund et al., "LIVECell—A large-scale dataset for label-free live cell segmentation," 9, Nat Methods **18**(9), 1038–1045, Nature Publishing Group (2021) [doi:10.1038/s41592-021-01249-6].

Caption List

Fig. 2 Synthetic images used in the training of the models. (A) A synthetic binary image used in the DermoGAN training. (B and C) RCM images of different noise levels and cell sizes used in U-net training. (D) Real RCM image. RCM, Reflectance confocal microscopy.

Fig. 2 DermoGAN architecture. The first task maps RCM images to the unpaired synthetic binary images. Whereas the second task learns the structure RCM images of the epidermis by translating Gabor filtered RCM images into binary images. .

Fig. 3 Comparison of the two cycle-GAN based approaches and the proposed DermoGAN. DermoGAN outperforms both methods.

Fig. 4 RCM image analyzed with the 8 presented algorithms. DermoGAN outperforms 5 out of 6 other methods on all images, and outperforms FIAP on 6/9 images.

584 **Fig. 5** DermoGAN can be extended to images acquired different imaging techniques using
585 retraining and outperforms traditional thresholding algorithms. (A) Input florescence
586 microscopy images. (B) Output of DermoGAN applied to image A. (C) Output of CellPose
587 cyto model. (D) Application of 18 thresholding approaches to the same image.
588 **Fig. 6** DermoGAN2 was trained entirely on synthetic images. (A) Synthetic non-confluent
589 images created using the SIMCEP. (B) Canny-filtered non-confluent images created using the
590 SIMCEP. (C) Binary non-confluent images.
591 **Fig. 7** DermoGAN2 applied to an image of confluent BV-2 cells (on the left), resulted in
592 accurate detection of cells with both DermoGAN (in the middle) and CellPose (on the right).
593 Manually determined cell centers were plotted on DermoGAN2 output in yellow, and in red on
594 the CellPose output.
595 **Fig. 8** DermoGAN2 applied to an image of SK-BR-3 cells (A) resulted in accurate detection
596 with both DermoGAN (B) of cells and CellPose (C), compared to thresholding methods (D).
597 **Fig. 9** DermoGAN2 applied to a mass spectroscopy image (on the left) resulted in accurate cell
598 detection with both DermoGAN (in the middle) with merging of adjacent cells, and CellPose
599 (on the right).
600 **Table 1** Comparison of median f1-score (computed with d-accuracy[42]) for all eight tested
601 approaches.
602 **Fig. S1** Structure of the generator and discriminator networks used in the cycle-GAN and
603 DermoGAN approaches.
604 **Fig. S2** To obtain keratinocytes positions. We apply the $G_{A2B}: A \rightarrow B$ network to locally
605 normalized RCM image and obtain an incomplete cell identification, which is then cleaned
606 by closing any holes in the detected membrane and the outside contour, and finally the cell
607 identification is refined using StarDist algorithm. RCM, reflectance confocal microscopy.
608 **Table S1** Comparison of all accuracy metrics for all eight tested approaches (in %).

609 **Table S2** Comparison of median accuracy metrics for all eight tested approaches (in %).

Supplementary material

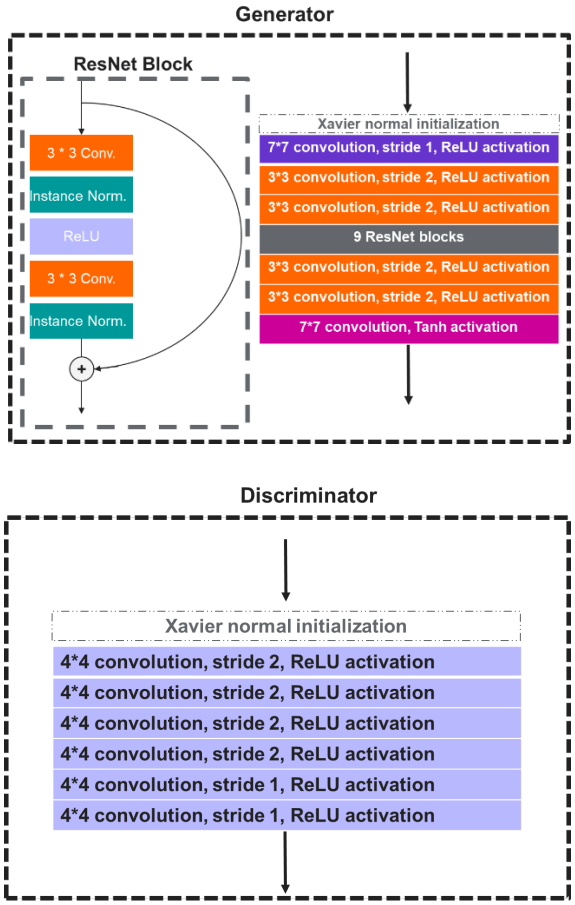

Fig. S1 Structure of the generator and discriminator networks used in the cycle-GAN and DermoGAN approaches.

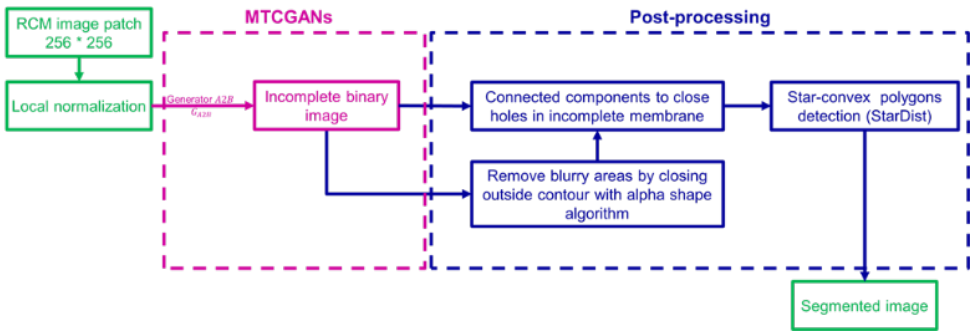

Fig. S2 To obtain keratinocytes positions. We apply the G_{A2B}: $A \rightarrow B$ network to locally normalized RCM image and obtain an incomplete cell identification, which is then cleaned by closing any holes in the detected membrane and the outside contour, and finally the cell identification is refined using StarDist algorithm. RCM, reflectance confocal microscopy.

Table S1 Comparison of all accuracy metrics for all eight tested approaches (in %).

Image	Layer	U-net pre-trained and augmented with real and synthetic images with Focal and Dice loss functions			StarDist on Gabor-filtered images			Cycle-GAN using RCM images			Cycle-GAN using Gabor-filtered images			FIAP			CellPose			DermoGAN + CellPose			DermoGAN + StarDist		
		Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
1	SS	50.3	29.8	37.5	37.1	47.5	41.6	47.8	46.8	47.3	42.9	40.8	41.8	46.2	89.4	61.0	31.3	14.1	19.5	83.6	61.5	70.9	68.4	73.6	70.9
2	SS	37.4	24.6	29.7	50.3	53.8	51.9	40.6	51.7	45.5	57.7	33.9	42.7	52.0	76.1	61.8	24.2	12.5	16.5	75.9	38.0	50.6	77.9	51.0	61.6
3	SS	40.8	24.6	30.7	50.5	44.4	47.3	45.9	42.6	44.2	50.4	26.7	34.9	48.9	79.6	60.6	25.2	11.0	15.3	65.8	40.3	50.0	72.5	60.7	66.1
3	SG	44.8	51.3	47.9	43.8	76.1	55.6	31.9	65.2	42.9	30.6	41.3	35.2	63.9	60.5	62.2	35.8	38.2	36.9	65.3	80.3	72.1	64.9	82.0	72.5
3	SG	28.8	56.8	38.2	27.7	65.5	38.9	23.1	67.3	34.4	17.5	32.7	22.8	56.4	77.0	65.1	31.9	48.6	38.5	61.6	60.8	61.2	53.4	85.1	65.6
3	SG	45.3	64.3	53.2	14.4	72.0	24.0	29.9	68.4	41.6	22.2	28.0	24.8	71.3	81.8	76.2	45.3	27.3	34.1	56.0	70.5	62.4	53.2	84.8	65.4
3	SG	42.4	70.7	53.0	14.7	78.7	24.7	19.7	58.7	29.5	16.8	31.8	22.0	71.6	90.7	80.0	52.3	60.0	55.9	62.0	81.7	70.5	63.9	92.0	75.4
3	SG	60.5	55.3	57.8	23.0	62.8	33.7	30.1	43.6	35.7	34.4	27.3	30.4	78.9	56.3	65.7	56.1	25.4	35.0	71.7	45.8	55.9	65.8	75.8	70.4
3	SG	50.5	86.6	63.8	14.8	79.5	25.0	21.0	70.9	32.4	18.8	26.1	21.8	64.1	87.4	74.0	51.5	78.7	62.3	58.8	82.3	68.6	57.4	86.6	69.0

Table S2 Comparison of median accuracy metrics for all eight tested approaches (in %).

		U-net pre-trained and augmented with real and synthetic images with Focal and Dice loss functions			StarDist on Gabor-filtered images			Cycle-GAN using RCM images			Cycle-GAN using Gabor-filtered images			FIAP			CellPose			DermoGAN + CellPose based postprocessing			DermoGAN + StarDist based postprocessing		
		Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
All	Median	44.8	55.3	47.9	27.7	65.5	38.9	30.1	58.7	41.6	30.6	31.8	30.4	63.9	79.6	65.1	35.8	27.3	35.0	65.3	61.5	62.4	64.9	82.0	69.0
	SD	8.9	21.6	12.2	15.1	13.4	12.1	10.5	11.3	6.4	15.2	5.8	8.4	11.3	12.1	7.4	12.2	23.5	16.5	8.8	17.7	8.7	8.4	13.3	4.2
SS	Median	40.8	24.6	30.7	50.3	47.5	47.3	45.9	46.8	45.5	50.4	33.9	41.8	48.9	79.6	61.0	25.2	12.5	16.5	75.9	40.3	50.6	72.5	60.7	66.1
	SD	6.7	3.0	4.2	7.7	4.7	5.2	3.7	4.6	1.5	7.4	7.1	4.3	2.9	6.9	0.6	3.8	1.6	2.1	8.9	13.0	11.9	4.8	11.4	4.6
SG	Median	45.1	60.5	53.1	18.9	74.1	29.3	26.5	66.2	35.0	20.5	29.9	23.8	67.7	79.4	69.9	48.4	43.4	37.7	61.8	75.4	65.5	60.6	85.0	69.7
	SD	10.4	13.0	8.7	11.5	7.0	12.3	5.3	10.1	5.2	7.4	5.6	5.5	7.9	14.2	7.2	9.8	20.5	12.1	5.5	14.6	6.3	5.8	5.4	3.9