POLITECNICO DI TORINO
Repository ISTITUZIONALE

Generative models for color normalization in digital pathology and dermatology: Advancing the learning paradigm

*Original*
Generative models for color normalization in digital pathology and dermatology: Advancing the learning paradigm / Salvi, M., Branciforti, F., Molinari, F., Meiburger, K.M.. - In: EXPERT SYSTEMS WITH APPLICATIONS. - ISSN 0957-4174. - STAMPA. - 245:(2024). [10.1016/j.eswa.2023.123105]

*Availability:*
This version is available at: 11583/2984887 since: 2024-01-07T19:53:09Z

*Publisher:*
Elsevier

*Published*
DOI:10.1016/j.eswa.2023.123105

*Terms of use:*

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

(Article begins on next page)

29 November 2026

Expert Systems With Applications 245 (2024) 123105

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# Generative models for color normalization in digital pathology and dermatology: Advancing the learning paradigm

Massimo Salvi *, Francesco Branciforti, Filippo Molinari, Kristen M. Meiburger

*Department of Electronics and Telecommunications, Polito[BIO]Med Lab, Politecnico di Torino, Biolab, Corso Duca degli Abruzzi 24, 10129 Turin, Italy*

ARTICLE INFO

*Keywords:*
Generative adversarial networks
Stain normalization
Color constancy
Digital pathology
Digital dermatology
Deep learning

ABSTRACT

Color medical images introduce an additional confounding factor compared to conventional grayscale medical images: color variability. This variability can lead to inconsistent evaluation by clinicians and the misinterpretation or suboptimal learning process of automatic quantitative algorithms. To mitigate the potential negative consequences of color variability, several color normalization strategies have been developed, proving to be effective in standardizing image appearance. In this paper, we present a novel paradigm for color normalization using generative adversarial networks (GANs). Our method focuses on standardizing images in the field of digital pathology (stain normalization) and dermatology (color constancy), where high color variability is consistently observed. Specifically, we formulate the color normalization task as an image-to-image translation problem, ensuring a pixel-to-pixel correspondence between the original and normalized images. Our approach outperforms existing state-of-the-art methods in both the digital pathology and dermatology fields. Extensive validation using public datasets demonstrate the effectiveness of our color normalization results on entirely external test sets. Our framework exhibits strong generalization capability on unseen data, making it suitable for inclusion in the pipeline of automatic quantitative algorithms to reduce color variability and improve segmentation and/or classification performance. Lastly, we provide the source code of our models to encourage open science.

## 1. Introduction

Color in medical imaging applications is typically handled with little standardization (Barata, Celebi, et al., 2014). The final appearance of color images heavily relies on the process of capture, processing, storage, and display, which can vary among imaging device manufacturers (Barata, Celebi, et al., 2014). In this context, color normalization aims to generate color image data with identical or similar perceptual response when evaluated by human operators. While human observers may tolerate color variability more than image analysis algorithms, data standardization is increasingly important in the era of artificial intelligence and big data (Janssen et al., 2020).

Grayscale images in the field of medicine are usually highly standardized, ensuring consistent interpretation and analysis. However, the same level of standardization does not apply to color images. Dermatology and histology, for example, share a low scale and a low conversion factor, resulting in associated variabilities. Consequently, color figures in these domains often exhibit significant variations that can affect the accuracy and reliability of image analysis algorithms. Therefore, there is a need to address the standardization of color images in medical applications to ensure consistent and reliable interpretation.

Several color normalization (or color constancy/color standardization) algorithms have been proposed in literature for two main color medical imaging modalities: digital pathology and dermatology. These algorithms aim to produce images that appear as if they were acquired under a standardized energy source, and have demonstrated improved robustness of automatic algorithms, particularly when dealing with multi-center datasets or datasets acquired using different devices (Barata, Marques, et al., 2014; Swiderska-Chadaj et al., 2020).

Digital pathology is an important clinical field where histology slides of stained biological tissue are digitized to produce high-resolution images (Janowczyk & Madabhushi, 2016). However, the manual preparation of the sample involves numerous non-standardized procedures, resulting in significant variation in the appearance of digital images. This variability poses a major challenge in designing robust systems for automated analysis of histological images (Salvi, Acharya, et al., 2020).

* Corresponding author at: Biolab, Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi, 24 – 10129 Torino, Italy.
*E-mail addresses:* massimo.salvi@polito.it (M. Salvi), francesco.branciforti@polito.it (F. Branciforti), filippo.molinari@polito.it (F. Molinari), kristen.meiburger@polito.it (K.M. Meiburger).

https://doi.org/10.1016/j.eswa.2023.123105
Received 9 December 2022; Received in revised form 23 August 2023; Accepted 29 December 2023
Available online 3 January 2024
0957-4174/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

In this field, color constancy algorithms are typically called "stain normalization" methods, as their objective is to normalize the color of the stain while maintaining good image contrast and preserving important morphological and functional information. Traditional stain normalization methods have the potential drawback of distorting tissue structures and texture in the original image (Kumar et al., 2017). Moreover, most classic techniques for stain normalization rely on a single user-selected reference image to determine the target stain (Kumar et al., 2017).

In dermatology, color images are acquired using a clinical dermatoscope enabling high-resolution and epiluminescent viewing and image acquisition of skin lesions. The color within these images depends on both image illumination and the diverse skin tones, which can vary among patients. State-of-the-art (SOA) color constancy algorithms in dermatology include Gray World (GW) (Buchsbaum, 1980), Shades of Gray (SoG) (Finlayson & Trezzi, 2004) and max-RGB (MRGB) (Land, 1977). These heuristic approaches first estimate the light source in the RGB color space and then apply a color transformation using the von Kries diagonal model (von Kries, 1970). A limitation of these traditional color constancy algorithms is their reliance on prior assumptions to estimate the color of the light source (Sidorov, 2019), which are often not met in real-life scenarios.

Generative adversarial networks (GANs) (Goodfellow et al., 2014) are the latest emerging framework for generating synthetic images and time-series data. Stain normalization of histopathological images using GANs has recently gained momentum, where the model is trained with image pairs: the input source image and the target image (Zhu et al., 2017; Zanjani et al., 2018; Sandfort et al., 2019). GAN-based approaches for color normalization, which are now considered the gold standard in digital pathology (Salehi & Chalechale, 2020), can be grouped into three categories: Pix2Pix GANs (Salehi & Chalechale, 2020), Stain Style Transfer Networks (Cho et al., 2017; Yuan & Suh, 2018; Liang et al., 2020), and Cycle GANs (Zhou et al., 2019a; Lo et al., 2021; Mahapatra et al., 2020). Pix2Pix GANs exploit a pixel-to-pixel correspondence between the original and synthetic image, producing stable results. However, it is limited by the fact that the target and source images are typically acquired with two different scanners, and the user must select one of the two images as the target image. So far, Pix2Pix GANs for color normalization have been applied to grayscale images, which may result in information loss compared to RGB images (Salehi & Chalechale, 2020). Stain Style Transfer Networks and Cycle GANs work on unpaired data and employ target and source images from two completely different datasets. With style transfer networks, a target image is defined, and its "style" is transferred to the source images as the network learns to associate spatial features in each image with their corresponding color characteristics. Hence, only the style is transferred between the target and source, without specific control over the final output, leading to a loss of pixel-to-pixel correspondence and high dependence on the statistics of the target image (Lo et al., 2021). Cycle GANs use a target group instead of a target image, making this approach more versatile and overcoming some of the limitations of Style-Transfer networks. However, challenges remain in controlling the final output of the model: since the pixel-to-pixel correspondence is lost, there is a higher risk of creating artefacts or even creating artificial objects in the output images that are not present in the input images. Another issue with GAN-based approaches is that the parameter optimization problem and the effect of the employed loss function and architecture are often overlooked. Modeling complex systems and optimization problems are a hot topic that cover various research areas (Bojan-Dragos et al., 2021; Borlea et al., 2022; Pozna et al., 2012; Precup et al., 2021; Singh & Shukla, 2022; Tan et al., 2014), and the influence of these aspects on obtained results with GAN-based techniques merits further investigation.

GAN-based normalization methods are currently considered the preferred technique in digital pathology. Most of the proposed approaches in the literature rely on the careful selection of representative template images and may struggle in regions that do not match the templates well. On the other hand, the only study in literature that utilizes GANs for color normalization in dermatology is our previous work (Salvi et al., 2022), which employed a Pix2Pix model to perform color constancy.

While generative models have achieved promising results in the field of color normalization, controlling the final output of GANs trained on unpaired data poses several challenges. These challenges arise due to the loss of pixel-to-pixel correspondence between the original image and the target image. The current main challenges are:

- *Lack of direct control*: In generative models, the mapping between the target group and the original images is learned implicitly through the adversarial training process. As a result, there is no direct control over the specific mapping between individual pixels or regions in the target and original images. This lack of direct control makes it difficult to enforce specific constraints or requirements on the final output.
- *Risk of artifacts*: Since generative models typically operate without direct pixel-level correspondence, there is a higher risk of introducing artifacts or inconsistencies in the generated images. These artifacts can manifest as unrealistic color shifts, texture distortions, or the creation of artificial structures that do not exist in the input images. Controlling and minimizing these artifacts becomes a challenge, especially when dealing with complex color normalization tasks.
- *Dependency on target image statistics*: GANs heavily rely on the statistics and characteristics of the target image group. The network learns to match the distribution of the target images in terms of color and style. This reliance on the target image statistics means that the generated output is highly influenced by the specific properties of the target group. If the target group is not representative or lacks diversity, the generated images may exhibit biased or limited variations.
- *Generalization to unseen variations*: Generative models may struggle to generalize well to unseen variations or cases that deviate significantly from the training data. If the training dataset does not adequately cover the range of possible color variations or if there are uncommon patterns, the GAN may produce suboptimal results or fail to capture the full spectrum of color normalization required in challenging scenarios.

In this work, we present a new paradigm for color normalization in digital pathology and dermatology using GANs. Our model employs a heuristic algorithm to normalize the input images, which serves as the target domain during the training process. This integration of structural information ensures a pixel-to-pixel correspondence between the original and the target images. Our research presents several key contributions:

- We introduce a new learning paradigm for color normalization in digital pathology and dermatology using GANs, addressing the limitations of existing color normalization methods.
- Our color-to-color translation and Pix2Pix-based paradigm successfully achieves a high level of visual similarity to the reference image. Unlike previous GANs designed for color normalization, our approach generates stable results even for images with various artifacts such as blur, non-uniform illumination, and tissue folds.
- The utilization of an ad-hoc image, obtained through specific heuristic algorithms, as the target domain for the GAN is crucial to ensure the method's generalizability across external datasets. We have demonstrated this through both quantitative and qualitative analysis using previously unseen data.
- We conduct a comprehensive investigation of different GAN architectures, considering parameters and loss functions, to assess the quality of the generated images. Additionally, we compute various

metrics tailored to digital pathology and dermatology applications to identify the optimal GAN configuration for each domain.
- To promote open science and facilitate direct comparisons with other research groups, we provide free access to the code of our GANs and the dataset used for testing our method. This enables researchers to download and utilize these resources for their own studies.

## 2. Materials and methods

### 2.1. GSN-GAN: Generalized stain normalization GAN for digital pathology

#### 2.1.1. Dataset

220 WSIs were collected from our previous works on prostate (Salvi et al., 2023), breast (Salvi et al., 2019) and liver (Salvi, Molinaro, et al., 2020) tissues. Histological tissues were obtained by needle biopsy, fixed in formalin, serially sectioned at 5 μm, and stained with conventional H&E staining. All biopsies were anonymized by a pathology staff member not involved in the study before any further analysis. Digital images were scanned with a 200x magnification (conversion factor: 0.467 μm/pixel) using a Hamamatsu NanoZoomer S210 Digital slide scanner. The WSIs used for training and validation were divided into tiles of equal size (2048×2028 pixels), and we underline that the images used for validation derived from the same organs as the images used for training but come from entirely different WSIs. Hence, the same WSI was only present either in the training set (200 WSIs) or validation set (20 WSIs). Starting from the extracted tiles, 2200 images (training: 2000 images, validation: 200 images) were selected for the actual development of the GAN model. The breakup of the dataset between the three tissues (i.e., prostate, breast, liver) is shown in Table 1 and more detail on how the tiles were specifically selected can be found in Section 2.1.3.

#### 2.1.2. Heuristic algorithm for stain normalization

Our previously developed stain normalization algorithm, named SCAN (Stain Color Adaptive Normalization), was used to normalize the extracted tiles (Salvi, Michielli, et al., 2020). The SCAN algorithm initially separates the histological stains by employing color deconvolution, which separates the stains based on their absorption characteristics. Following this preliminary stain separation, the algorithm focuses on detecting cellular structures within the histological image. Subsequently, a refined stain separation is performed by considering only the cellular structures, resulting in more accurate separation of the stains and reducing interference from the background. In the final phase of the algorithm, the image is normalized. The stain color appearance is standardized with respect to a reference image that exhibits an optimal and reproducible staining distribution.

The optimization problem in this context revolves around finding the optimal stain separation and normalization strategy that minimizes the variability of the output images' stains. Although SCAN is a versatile and multi-tissue normalization algorithm, it still presents some limitations. Specifically, this heuristic algorithm is not always reliable, and occasional artefacts may arise if one of the two stains is insufficiently present in the image or if the stain separation is not correctly obtained. Additionally, the processing and normalization of images using SCAN may take a few seconds. Some cases where SCAN produces a suboptimal result are displayed in Fig. 1.

**Table 1**
Data sets used to develop the GAN for digital pathology. A total of 220 WSIs of prostate, breast and liver are used to train and test GSN-GAN (Generalized Stain Normalization GAN). Tiles with a dimension of 2048×2048 pixels were extracted from each WSI, and 2200 of them were selected for the development of GSN-GAN.

| | Training | | | Validation | | |
|---|---|---|---|---|---|---|
| Tissue | WSI | Tiles | Selected | WSI | Tiles | Selected |
| Prostate | 75 | 10,578 | 700 | 7 | 987 | 80 |
| Breast | 70 | 9482 | 700 | 7 | 955 | 60 |
| Liver | 65 | 8861 | 600 | 6 | 836 | 60 |
| *Total* | *200* | *28,921* | *2000* | *20* | *2778* | *200* |

#### 2.1.3. GSN-GAN for optimal stain normalization

To overcome the current limitations of the heuristic SCAN algorithm, a GAN was trained with a specific strategy: the proposed GSN-GAN was trained only on pairs of images where the SCAN algorithm provided optimal results according to an expert pathologist. In particular, the expert pathologist selected a total of 500 tiles (each 2048×2048 pixels) in which the SCAN normalization was optimal in terms of color coherency and quality of the normalized image. The expert selected only the images in which the normalization process reproduces the target stain colors and does not introduce image artifacts that could affect the clinical pathway (Zheng et al., 2019). During this selection process, it was ensured that at least two tiles for every initial WSI were included, hence maintaining a heterogenous dataset.

For the training process, the SCAN-normalized image represents the target domain for the GSN-GAN. In this way, a Pix2Pix correspondence between domain A (original image) and domain B (normalized image) was maintained. The quality of the result can be guaranteed thanks to the careful selection of optimized images by the expert pathologist. It is important to underline that this is the first work in digital pathology that uses images that are enhanced by a heuristic algorithm as a target domain for GAN training.

The selected 2048×2048 tiles were subsequently divided into four 1024×1024 sub-images to fit the GAN input size. This process made it possible to obtain 1024×1024 tiles where only background and/or artefacts were present, giving the GAN the opportunity to be trained well even when presented with these problematic images. The entire workflow of the process is illustrated in Fig. 2 and more details of the GAN architecture are given hereafter.

The GSN-GAN is a Pix2Pix GAN that employs a U-net architecture as the generator and a three-layer fully convolutional PatchGAN (C. Li & Wand, 2016) as the discriminator network. The generator (G) learns a mapping from observed image *x* and random noise vector *z* to *y;* G(x,z) → y, and the discriminator (D) has the task of learning to classify an image as a real image from the training image (close to 1) or a fake image produced by the generator (close to 0): D(x) → [0.1]. In this context, the generator network takes the input source image and aims to transform it into a normalized image with consistent color characteristics. The discriminator network, on the other hand, is responsible for distinguishing between the generated normalized images and the real target images.

The training algorithm of the GAN involves an adversarial learning process, where the generator and discriminator networks play a competitive game. The generator aims to generate normalized images that can fool the discriminator into classifying them as real target images, while the discriminator aims to correctly distinguish between the real and generated images. This adversarial training is complemented by a pixel-wise loss which ensures that the pixel-level correspondence between the original and normalized images is preserved. Both the generator and the discriminator are trained with backpropagation and have their own loss functions. The objective function of a traditional GAN.is defined as follow:

$$L_{GAN}(G,D) = E_{x,y}[\log(D(x,y))] + E_{x,z}[\log(1 - D(x,G(x,z)))] \quad (1)$$

where G tries to maximize this objective function while D tries to maximize it, and E(•) expresses the expectation. However, traditional GANs present a certain number of disadvantages, such as the Mode Collapse problem and Vanishing Gradient (Gulrajani et al., 2017). To address these specific issues, we tested two different objective functions:

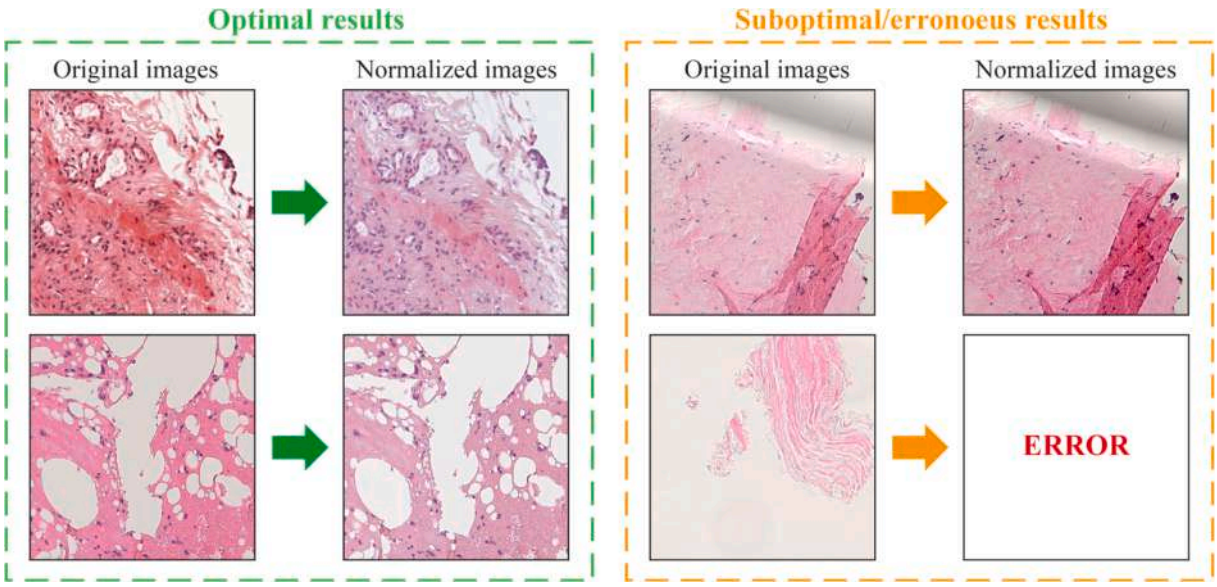


**Fig. 1.** Examples where the SCAN algorithm (Salvi, Michielli, et al., 2020) provides optimal and suboptimal or erroneous results. As can be seen, the presence of artifacts makes the normalization result suboptimal while the absence of a stain produces an error in the algorithm since stain separation cannot be performed.

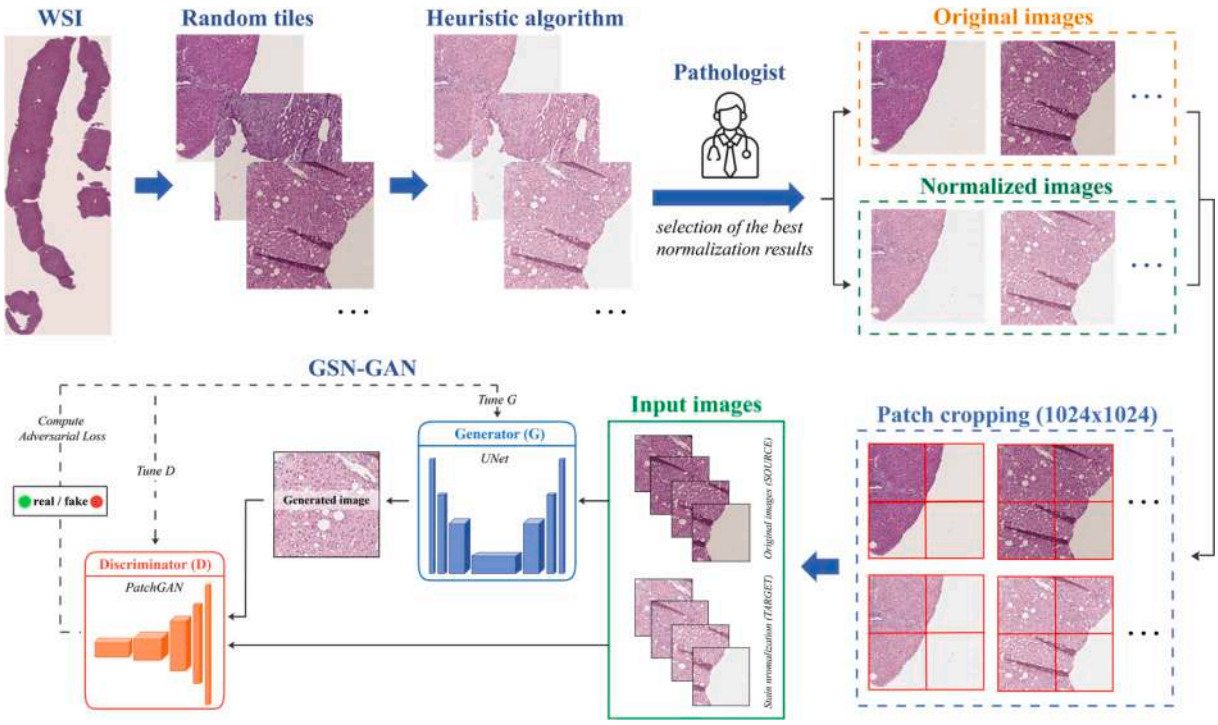


**Fig. 2.** Overview of our novel GAN-based stain normalization algorithm. From the digitized slides, only those tiles whose normalization is optimal (based on a pathologist's opinion) are selected for GAN development. Finally, the input images are provided as a condition to the Generator (G), whose output, along with a true sample is passed to the Discriminator (D). The original histological image is used as the source domain while the result of our heuristic stain normalization algorithm is employed as the target domain.

1. LS-GAN (Least Squares GAN): this is a type of GAN that adopts the least squares loss function for the discriminator (Mao et al., 2017). Minimizing the objective function of the LS-GAN minimizes the Pearson divergence. The objective function can be defined as:

$$\min_D V_{LSGAN}(D) = \frac{1}{2}E_{x,y}\left[D(x)-b^2\right] + \frac{1}{2}E_{x,z}\left[\left(D(x,G(z))-a\right)^2\right]$$

$$\min_G V_{LSGAN}(G) = \frac{1}{2}E_{x,z}\left[\left(D(x,G(x,z))-c\right)^2\right] \quad (2)$$

where a and b are the labels for fake and real data, respectively, and c denotes the value that G aims for D to believe as fake data.

2. WGAN-GP (Wasserstein GAN + Gradient Penalty) is a generative adversarial network that uses the Wasserstein loss formulation plus a gradient norm penalty to achieve Lipschitz continuity (Gulrajani et al., 2017). Instead of clipping weights like WGAN (Arjovsky et al., 2017), this network adds a penalty term to the gradient norm of the critical function. The loss used is:

$$L = E_{\tilde{x}\ Pg}[f(\tilde{x})] - E_{x\ Pr}[f(x)] + \lambda E_{\hat{x}\ P\hat{x}}\left[\left(\|\nabla_{\hat{x}} f(\hat{x})\|_2 - 1\right)^2\right] \quad (3)$$

where $P\hat{x}$ represents the distribution obtained by sampling in a uniform way along a straight line between the generated and real distributions $Pg$ and $Pr$. $\lambda$ is the penalty coefficient used to weight the gradient penalty term. As suggested by a previous work (Gulrajani et al., 2017), we set $\lambda = 10$ for all the experiments.

The GSN-GAN requires pairs of images in the training phase that consist of an original image and the corresponding transformed image, which in this case is obtained using the SCAN stain normalization algorithm. In this way, the GSN-GAN formulates the stain normalization task as an image-to-image translation problem while maintaining a pixel-to-pixel matching. The GSN-GAN is trained iteratively, with each iteration involving a forward pass through the networks, computation of the loss functions, and backpropagation of gradients to update the network parameters. The training process continues until convergence, where the generator network learns to produce high-quality normalized images that are visually similar to the target images. Our model is trained for 150 epochs using instance normalization. The learning rate is initially set at 0.0001 and reduced by a factor of 0.5 each time 50 epochs pass. To optimize the overall GSN-GAN architecture, we tested different configurations, varying the number of trainable parameters of the generator and discriminator networks, the number of filters, and the loss function. The experimental results are reported in Section 3.2.

### 2.2. GCC-GAN: Generalized color constancy GAN for dermatology

#### 2.2.1. Dataset

The dermatological images come from a set of open access datasets (Rotemberg et al., 2021) provided by isic-archive.com. To create a tool with better generalization ability, we selected dermoscopic images with different characteristics in terms of illumination, resolution, field of view (FOV) and aspect ratio. Seven different skin lesions were considered: Actinic keratoses, Basal cell carcinoma, Dermatofibroma, Keratosis-like lesions, Melanoma, Melanocytic nevi, and Vascular lesions.

Starting from an initial heterogenous pool of 30,708 images, an experienced dermatologist selected 1400 images, 200 images for each lesion type, on which a semi-automatic color constancy algorithm achieved the best possible color-constancy transformation (Section 2.2.2).

The resulting dataset (1400 images) was then randomly split into a training subset (1050 images) and validation subset (350 images), while still respecting the balance between classes. The validation set was used for testing both the performance and the generalization ability of the proposed GAN. The breakup of the dataset between the different skin lesions is shown in Table 2.

#### 2.2.2. Heuristic algorithm for color constancy

We define the reference images as a set of normalized images characterized by neutral illuminant and perfect exposure balance. Currently, there is no gold standard in literature with which to compare different color constancy methods in dermatology (Barata, Celebi, et al., 2014).

*General Gray World* (Barata, Celebi, et al., 2014), a widely used state-of-the-art algorithm, is applied semi-automatically to obtain the reference image for color constancy. This heuristic method uses image statistics to estimate the illuminant color and then transform the image:

$$\left(\frac{\int \left(I_i^{\sigma}(x,y)\right)^p dx}{\int dxdy}\right)^{\left(\frac{1}{p}\right)} = ke_i \quad (4)$$

where $I_i$ is the $i^{\text{th}}$ layer of the input image $I(x,y)$ smoothed with a Gaussian low-pass filter with standard deviation $\sigma$; $e_i$ is the illuminant of $i^{\text{th}}$ layer of the image; $k$ is a normalization constant; and $p$ is the degree of the Minkowski norm (Wang et al., 2004). The parameter $p$ determines the sensitivity of the norm to outliers, which are pixel values that deviate significantly from their neighboring pixels. When $p$ is set to 1, the algorithm essentially computes a color average, making it robust against minor color variations. However, as the value of $p$ increases, the algorithm becomes more sensitive to the brightest colors present in the image. While this can be beneficial for images dominated by specific colors, it may result in excessive corrections for images with well-balanced color distributions. Hence, the final outcome can range from a balanced color correction for lower $p$ values to potentially exaggerated corrections for higher $p$ values.

The $\sigma$ parameter in the Gaussian filter determines the scale of local variations considered by the algorithm. A smaller $\sigma$ value retains more fine details of the image, rendering the algorithm sensitive to small-scale color variations. This can be useful for images with fine color details but may lead to overcorrection in the presence of noise. Conversely, a larger $\sigma$ value smooths out these details, making the algorithm more robust against noise but potentially causing color nuances to be overlooked. Therefore, the final image can vary from a detailed but potentially noisy correction for smaller $\sigma$ values to a smoother but potentially less accurate correction for larger $\sigma$ values. Performing manual tuning of these parameters on each dermatological image is crucial because it allows for customized adjustments that cater to the specific characteristics and nuances present in individual images, ultimately leading to more accurate and visually appealing results.

Finally, the heuristic algorithm adjusts the exposure (i.e., global intensity) of the transformed image through gamma correction:

$$I_{OUT} = cI_{IN}^{\gamma} \quad (5)$$

where $I_{IN}$ is the optimized image obtained in the previous step, $c$ is a constant set to 1, and $\gamma$ is the parameter that regulates the correction. When $\gamma$ value is greater than 1, the resulting image will appear darker, enhancing the visibility of darker regions but potentially obscuring details in shadows. Conversely, when $\gamma$ value is less than 1, the image will appear brighter, which can bring out details in darker areas but may result in washed-out highlights. As a result, the final appearance of the image can vary from a darker and more contrasted image for higher $\gamma$ values to a brighter and less contrasted image for lower $\gamma$ values. Adjusting $\gamma$ parameter is particularly crucial in dermatological images, where the visibility of subtle features can significantly impact the accuracy of diagnoses.

This heuristic algorithm is applied to each image via a custom graphical user interface (GUI). Using this GUI (Fig. S1), an experienced dermatologist can manually adjust the $\sigma$, $p$, and $\gamma$ parameters to achieve the optimal transformation for each single image. A detailed description of the GUI is provided in the Supplementary Material.

This semi-automatic normalization approach provides the flexibility of manual parameter adjustment, allowing to bypass the limitations of fully automatic SOA color constancy algorithms. Such methods rely on statistical assumptions that may not always apply to a large and heterogeneous set of images with diverse lighting, shadows, reflections, and other variances. By empowering the dermatologist to fine-tune the

**Table 2**
Dataset used to develop the GCC-GAN. AKIEC: Actinic keratoses, BCC: basal cell carcinoma, DF: dermatofibroma, KL: keratosis-like lesions (benign and seborrheic), MEL: melanoma, NV: melanocytic nevi, VASC: vascular lesions.

| Lesion | Images | Selected for train | Selected for validation |
|---|---|---|---|
| AKIEC | 1066 | 150 | 50 |
| BCC | 3356 | 150 | 50 |
| DF | 246 | 150 | 50 |
| KL | 2333 | 150 | 50 |
| MEL | 5375 | 150 | 50 |
| NV | 18,079 | 150 | 50 |
| VASC | 253 | 150 | 50 |
| *Total* | *30,708* | *1050* | *350* |

parameters, the algorithm can better account for the unique characteristics of each image and create optimal reference images. The reference images were then paired with their original unprocessed version, generating the data collection used to train and validate the GCC-GAN model. This paired dataset allows the GAN to learn the type of color adjustments needed to normalize an image's color profile while still preserving its identifying visual characteristics. Manually fine-tuning the parameters for each image helps ensure that the paired images provide optimal examples for the GAN to learn from.

### 2.2.3. GCC-GAN for optimal color constancy

To develop an effective method for normalizing images to match a reference set, a Pix2Pix GAN was trained to transform dermoscopy images captured under different illumination presets (domain A) into standardized images with a standard neutral illumination profile (domain B). During the training process, the model takes as input the original image and its corresponding normalized counterpart generated by the algorithm described in the previous section. The goal of GCC-GAN is to serve as a generalized, fast and fully automatic color constancy solution with pixel-level matching between original and normalized images. By learning from paired examples of original and normalized images, the GAN learns to replicate the color adjustments needed to standardize new images. Through this pixel-to-pixel training approach, GCC-GAN aims to normalize images in a way that preserves identifying visual characteristics while achieving a standardized color profile matching the reference set.

To handle different aspect ratios and resolutions, all images underwent an automatic preprocessing stage:

Aspect ratio check: if the aspect ratio is not equal to 1, zero padding is applied to obtain an aspect ratio equal to 1.

Evaluation of dark pixels: some dermoscopic images are characterized by a black circular mask of varying size due to the acquisition instrument. If the area covered by dark pixels is less than 20 % of the image, a zero padding of 50 pixels is applied to the outer edge of the image. This step was introduced after observing artifacts produced by an intermediate GCC-GAN model on images that had no dark pixels on the outer edges.

Resampling to 1024×1024 to fit the input size of our GCC-GAN.

The training algorithm for color constancy in dermatology follows a similar adversarial learning process as described for the digital pathology. The generator network aims to generate normalized images that exhibit color constancy, while the discriminator network distinguishes between the real target images and the generated normalized images. The training is guided by adversarial loss and pixel-wise loss, ensuring that the generator produces normalized images that preserve the essential color information while reducing the variability caused by different skin tones and illumination conditions. Through the training process, the GCC-GAN optimizes the generator and discriminator networks to find a balance between generating visually appealing normalized images and fooling the discriminator network. This optimization problem is iteratively solved using the Adam optimization algorithm, which updates the network parameters based on the gradients computed during the backpropagation process. To optimize the overall architecture of our model, we tested different configurations by varying the depth of the generator and discriminator networks, the number of filters, and the objective function. The experimental results are reported in Section 3.3. The overall procedure followed for processing dermatology images (data preparation, pre-processing, and GAN training) is illustrated in Fig. 3.

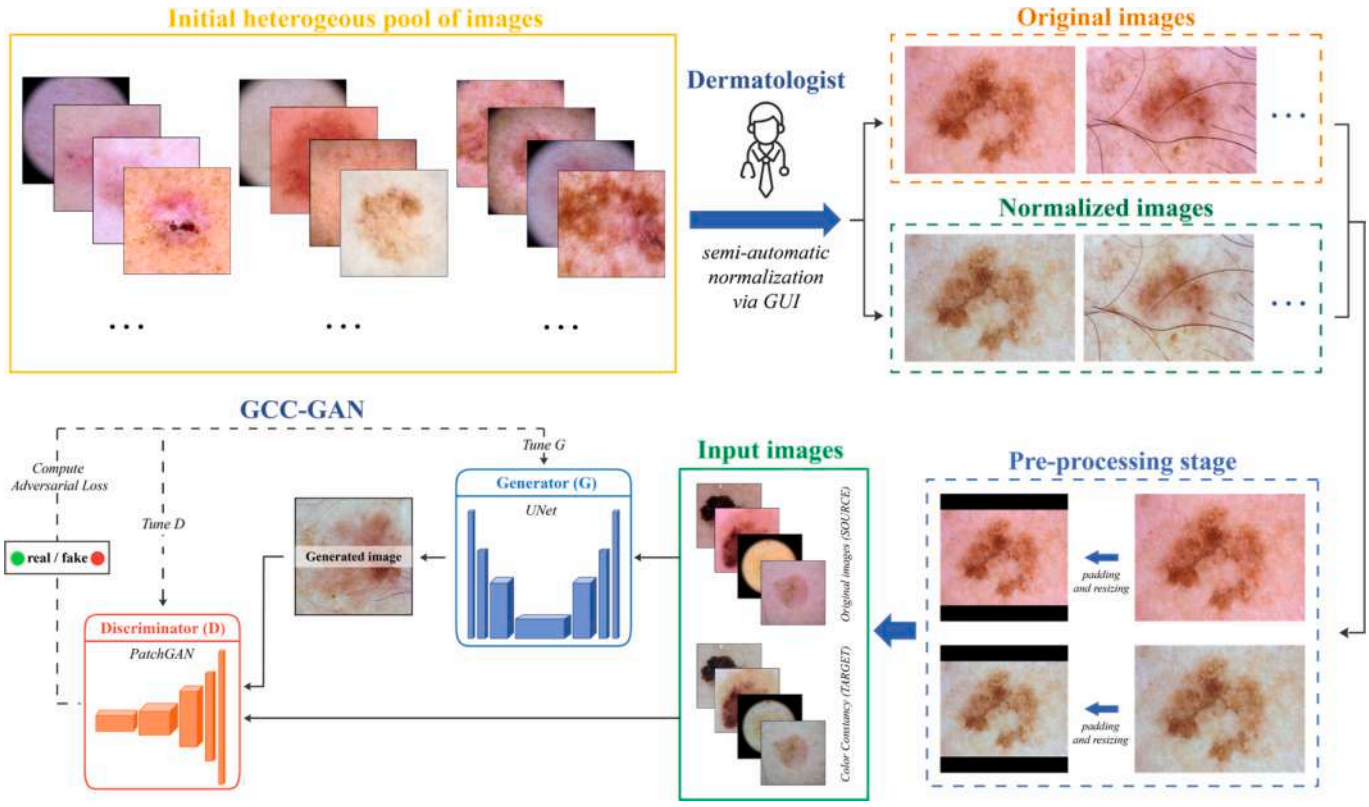


**Fig. 3.** Overview of our novel GAN-based color constancy algorithm. From an initial heterogenous pool of 30,708 images, only those images on which the General Gray World (Barata, Celebi, et al., 2014) provides optimal results based on a dermatologist's opinion are selected for GAN development. The input images are provided as a condition to the Generator (G), whose output, along with a true sample is passed to the Discriminator (D). The original dermatological image is used as the source domain while the result of the heuristic color constancy algorithm (considered as the reference image) is employed as the target domain.

### 2.3. Performance metrics

Various tests were conducted changing the model hyperparameters to determine the optimal GAN architecture for both the digital pathology and dermatology applications. In the research community, there is not a widespread agreement on how to evaluate paired image-to-image translation frameworks. Specific validation metrics were hence computed for each of the two applications:

- *Digital pathology*: the normalization process should not introduce artifacts and must retain the information present in the original image. To assess this, three metrics commonly employed in stain normalization were computed between the original image and the normalized image: SSIM (Structural Similarity Index), FSIM (Feature similarity), and PSNR (Peak signal-to-noise ratio) (Salehi & Chalechale, 2020);
- *Dermatology*: preserving data integrity in dermoscopic images during color transformation is imperative. Hence, three metrics were computed to compare the GAN-normalized image with the heuristic reference image: PCC (Pearson correlation coefficient), PSNR, and RMSE (Root mean square error).

For both applications under consideration, we compared our results with the SOA algorithms in terms of quantitative metrics and segmentations tasks. A primary challenge with artificial intelligence methods is their adaptability to covariant distribution shifts, or their capability to yield satisfactory results on samples beyond the training domain. To address this, we applied the trained models to entirely external datasets, thereby evaluating the generalization capability of our models. The GAN-based results on completely external datasets were then also compared with other SOA methods.

## 3. Results

### 3.1. Optimization of the GAN architecture

GANs have an extraordinary potential in many research fields. However, a significant drawback in the literature is the lack of hyperparameter tuning in most GAN studies to identify the optimal configuration. Here, we address this gap by presenting a comprehensive comparison of the GAN's performance using (a) two different losses, namely LSGAN and WGAN-GP, and (b) different depths of the generator and discriminator (i.e., 100 k, 500 k, 750 k, 1 M, 6 M, 12 M trainable parameters). This analysis encompasses a total of 12 different configurations. LSGAN and WGAN-GP were chosen for this comparison as they represent the most used target functions for generative models in medical imaging. For each configuration, the GAN was trained and quantitative metrics were evaluated on the train and validation sets (Table 3).

In the case of Pix2Pix GANs, the model aims to replicate the optimal results obtained with the heuristic algorithm. For this reason, the root mean square error (RMSE) between the generated image (fake B) and the image derived from the heuristic algorithm (real B) was computed.

The optimal configuration was chosen as the one that provided the lowest RMSE values on the training and validation sets. Hence, the $LRGAN_{6M}$ architecture (6 M parameters for both the generator and discriminator with LSGAN as objective function) was chosen for digital pathology, while the $LSGAN_{12M}$ architecture (12 M parameters for both the generator and discriminator with LSGAN as objective function) yielded the lowest RMSE values for the dermatology application. Notably, across both applications, the LSGAN loss function consistently resulted in lower RMSE values compared to the WGAN-GP loss function. Additionally, configurations with an equal number of trainable parameters for the generator and discriminator tended to deliver better performance results than other configurations. The detailed procedure for selecting the best epoch for the optimal configuration is reported in the Supplementary Material (Figs. S2-S3). The training is performed on a NVIDIA RTX 3090 24 GB using Pytorch framework. We have made the implementation of the GSN-GAN and GCC-GAN publicly available at https://doi.org/10.17632/32bvfw6xhj.2.

### 3.2. Digital pathology

In this section, we report the results obtained using the GSN-GAN for digital pathology applications. Quantitative results are shown and compared with SOA methods. Moreover, we tested the robustness and generalizability of the trained model on two completely external datasets with images acquired with different scanners and magnification. Finally, the stain normalization achieved with the GSN-GAN and SOA methods is applied as a preprocessing step within a segmentation framework, and the results are compared.

#### 3.2.1. Quantitative metrics

Here we compared the quantitative metrics obtained on the training and validation set using the proposed GSN-GAN and other SOA methods. We report both the average values and the percentiles (10th) to highlight the presence or absence of outliers of the various methods.

To evaluate the quality of normalization, a quantitative comparison is carried out by evaluating the SSIM, the PSNR and the FSIM. These metrics play a vital role in stain normalization as they enable objective measurements of the quality and effectiveness of the normalization methods. To provide a comprehensive comparison, we included two well-known heuristic approaches from the literature, Reinhard (Reinhard et al., 2001) and SPCN (Kumar et al., 2017), as comparative methods. Additionally, we compared the GSN-GAN with the SCAN algorithm it was trained on (Salvi, Michielli, et al., 2020), as well as a cycle-GAN method from the state-of-the-art (Zhou et al., 2019a). All

**Table 3**
RMSE values between synthetic (fake B) and target (real B) images as a function of the number of trainable parameters and the objective function of the GAN. NetG: Generator network, NetD: Discriminator network, Lsgan: Least Squares GAN, Wgan-gp: Wasserstein GAN gradient penalty.

| Name | GAN hyperparameters | | | Digital Pathology | | Dermatology | |
|---|---|---|---|---|---|---|---|
| | # of params of netD | # of params of netG | GAN objective function | Train | Val | Train | Val |
| $LSGAN_{100k}$ | ~ 100 k | ~ 1 M | lsgan | 6.19 | 6.76 | 3.21 | 8.03 |
| $LSGAN_{500K}$ | ~ 500 k | ~ 6 M | lsgan | 5.63 | 6.23 | 2.85 | 7.74 |
| $LSGAN_{750k}$ | ~ 750 k | ~ 12 M | lsgan | 5.12 | 5.94 | 2.67 | 7,38 |
| $LSGAN_{1M}$ | ~ 1 M | ~ 1 M | lsgan | 4.94 | 5.44 | 3.01 | 7.80 |
| $LSGAN_{6M}$ | ~ 6 M | ~ 6 M | lsgan | **4.29** | **4.88** | 2.72 | 7.52 |
| $LSGAN_{12M}$ | ~ 12 M | ~ 12 M | lsgan | 5.19 | 5.93 | **2.53** | **6.95** |
| $WGAN_{100k}$ | ~ 100 k | ~ 1 M | wgan-gp | 16.16 | 15.66 | 4.05 | 8.12 |
| $WGAN_{500k}$ | ~ 500 k | ~ 6 M | wgan-gp | 9.54 | 8.48 | 3.80 | 7.51 |
| $WGAN_{750k}$ | ~ 750 k | ~ 12 M | wgan-gp | 8.85 | 7.91 | 13.01 | 12.42 |
| $WGAN_{1M}$ | ~ 1 M | ~ 1 M | wgan-gp | 33.43 | 33.58 | 10.82 | 9.57 |
| $WGAN_{6M}$ | ~ 6 M | ~ 6 M | wgan-gp | 17.63 | 18.17 | 10.03 | 10.81 |
| $WGAN_{12M}$ | ~ 12 M | ~ 12 M | wgan-gp | 13.46 | 13.61 | 10.81 | 10.87 |

methods were evaluated using the default parameters specified in the respective source articles. The comparison between our model and the state-of-the-art methods Click or tap here to enter text.Click or tap here to enter text.Click or tap here to enter text. Click or tap here to enter text. is reported in Table 4. The quantitative comparison shows that GSN-GAN is the best performing method for stain normalization. To demonstrate the effectiveness of our generative model in stain normalization, a pairwise *t*-test was conducted between the performance of GSN-GAN and the compared techniques. All statistical tests were carried out with a significance level (p-value) of 0.05. The results of the paired *t*-test indicated a significant difference in both train and validation sets (Table 4).

The proposed GSN-GAN obtains the best performance for all three metrics and has the lowest computational time, being up to 20 times faster than the SOA methods. When compared to the algorithm that was used for training the GAN (i.e., SCAN), the proposed GSN-GAN obtains similar performance both in the training and in the validation. This is to be expected as the GSN-GAN aims to mimic the behavior of the heuristic algorithm. Still, the proposed method is much faster when compared to SCAN (inference time: 0.09 s vs 1.60 s) and is also able to overcome some limitations of the heuristic algorithm. The visual performances of the compared methods are reported in Fig. 4, where the results that contain typical artifacts are surrounded by an orange box. GSN-GAN reported less image color artifacts than existing approaches due to its robustness (pixel-to-pixel matching between original and normalized image) and appropriateness (trained only on optimal normalized images).

As we can see from Fig. 4, when the source image is high quality, all of the methods perform reasonably well (source 1); on the other hand, when the image presents large white areas or there is a small amount of stain present in the image, the SPCN (Kumar et al., 2017) and Reinhard (Reinhard et al., 2001) methods are suboptimal and the Cycle-GAN method (Zhou et al., 2019a) produces artefacts (source 2–3). When specific artefacts are present in the source image, such as blurred regions, the proposed GSN-GAN is the only method able to produce a coherent and robust stain normalization. As can be seen from Fig. 4, GSN-GAN is able to retain the contextual information of the source image while applying the color distribution of the target image.

Finally, the overall distribution of the dataset (pre- vs. post-normalization) was examined in the LAB color space to assess the effect of the stain normalization process. In the Supplementary Materials, Fig. S4 shows the effectiveness of the stain normalization for both training and validation sets.

**Table 4**
Quantitative metrics used to compare the GSN-GAN stain normalization with current state-of-the-art methods. The table shows the average values of the metrics and in brackets their 10th percentile (higher values are better).

| Method | Subset | Comp. Time (s) | SSIM | PSNR | FSIM |
|---|---|---|---|---|---|
| Reinhard et al. (Reinhard et al., 2001) | Train | 1.21 | 88.9 (69.7) | 20.7 (14.8) | 93.4 (81.8) |
| | Val | 1.14 | 88.4 (58.8) | 20.7 (15.1) | 92.8 (77.8) |
| SPCN (Kumar et al., 2017) | Train | 1.80 | 80.7 (53.7) | 22.7 (15.2) | 89.8 (72.1) |
| | Val | 1.82 | 81.5 (54.4) | 22.8 (14.8) | 89.9 (70.6) |
| Cycle-consistent GAN (Zhou et al., 2019a) | Train | 0.17 | 92.1 (85.7) | 25.8 (19.3) | 93.9 (88.3) |
| | Val | 0.15 | 93.7 (89.7) | **26.3** (19.2) | 95.1 (91.4) |
| SCAN algorithm (Salvi, Michielli, et al., 2020) | Train | 1.63 | 95.4 (91.1) | 25.8 (21.7) | 98.3 (96.1) |
| | Val | 1.59 | **97.1** (94.0) | 25.6 (21.3) | 98.8 (97.3) |
| GSN-GAN (proposed) | Train | 0.09 | **96.6** (**95.10**)* | **26.2** (**21.9**)* | **98.6** (**96.8**)* |
| | Val | 0.08 | 96.8 (**95.20**)* | 25.6 (**21.7**) | **98.9** (**97.4**)* |

(*) Asterisk denotes statistically significant difference ($p < 0.05$) compared to state-of-the-art methods.

### 3.2.2. Testing GSN-GAN on external datasets

To evaluate the generalization ability of the trained GAN, two open-source external datasets were chosen that contain histopathological images of organs that the GAN never saw during the training or validation process. The first dataset, named ACDC-lungHP (Z. Li et al., 2020), consists of 150 WSIs of lung tissue in which histological slides were stained with H&E and scanned by a digital slide scanner (3DHISTECH Pannoramic 250) at objective magnifications of 20x. One thousand 1024 × 1024 tiles were randomly extracted to test the GSN-GAN. The second dataset is part of the MoNuSeg Challenge (Kumar et al., 2019) which contains 30H&E-stained tissue images captured at 40x magnification from the TCGA archive. The images in this challenge come from 7 different organs: breast, liver, kidney, prostate, bladder, colon, and stomach. In Fig. 5 the visual performance of the GSN-GAN and other SOA methods is shown.

On the ACDC-lungHP dataset, the GSN-GAN performs a robust and consistent normalization relative to the source image content. In particular, our method does not produce artifacts on the dark portions of the image (Fig. 5A – first row) and correctly preserves the appearance of red blood cells (Fig. 5A – second row). Our GAN also shows excellent performance on the MoNuSeg dataset, preserving the relative contrast present within the original image (Fig. 5B – first row and second rows). On these external datasets, GSN-GAN produces high-quality images with respect to SOA methods. Finally, it can be noted that GSN-GAN provides robust results even on images containing tissue (lung, stomach, etc.) and magnification levels (40x) not used during the training process. This result demonstrates the exceptional generalization capability of GSN-GAN.

Fig. 6 compares GSN-GAN to SOA methods for all four datasets analyzed in this study (i.e., training set, validation set, ACDC-lungHP and MoNuSeg). Our technique achieves the highest SSIM for all subsets. Moreover, unlike other SOA techniques, our method shows stable performance even on external datasets without any considerable drop in performance.

### 3.2.3. Impact of stain normalization on cell nuclei segmentation

Given the diverse appearances of nuclei across multiple organs and patients, as well as the variation in staining protocols adopted by different hospitals, stain normalization becomes crucial in standardizing data and reducing staining variability. In fact, numerous studies in the literature have demonstrated the beneficial impact of using stain normalization as a preprocessing step to enhance the performance of deep learning models (Zhou et al., 2019a; Swiderska-Chadaj et al., 2020). Consistent with previous research, we addressed the impact of stain normalization on the nuclei segmentation task (Pontalba et al., 2019) using the publicly available MoNuSeg dataset (Kumar et al., 2019). The quantitative results obtained using a standard U-Net architecture, where images normalized with different stain normalization methods were used as input, are reported in Table 5. To assess accuracy in boundary delineation, we employed popular metrics for instance segmentation such as the Hausdorff distance 95th percentile (HD95). Additionally, we utilized the Aggregate Jaccard Index (AJI) to quantify both pixel-level and object-level performance (Kumar et al., 2019). The proposed GSN-GAN demonstrated excellent overall segmentation results, achieving the highest values for recall, dice coefficient, HD95 and AJI parameters. Considering the dice parameter, the proposed GSN-GAN outperforms the compared methods up to 3.3 % (0.805 GSN-GAN vs. 0.772 Heuristic algorithm). A visual comparison is also provided in Fig. 7, demonstrating that the high-quality image generated by GSN-GAN enable the segmentation network to better delineate the contours of individual cells.

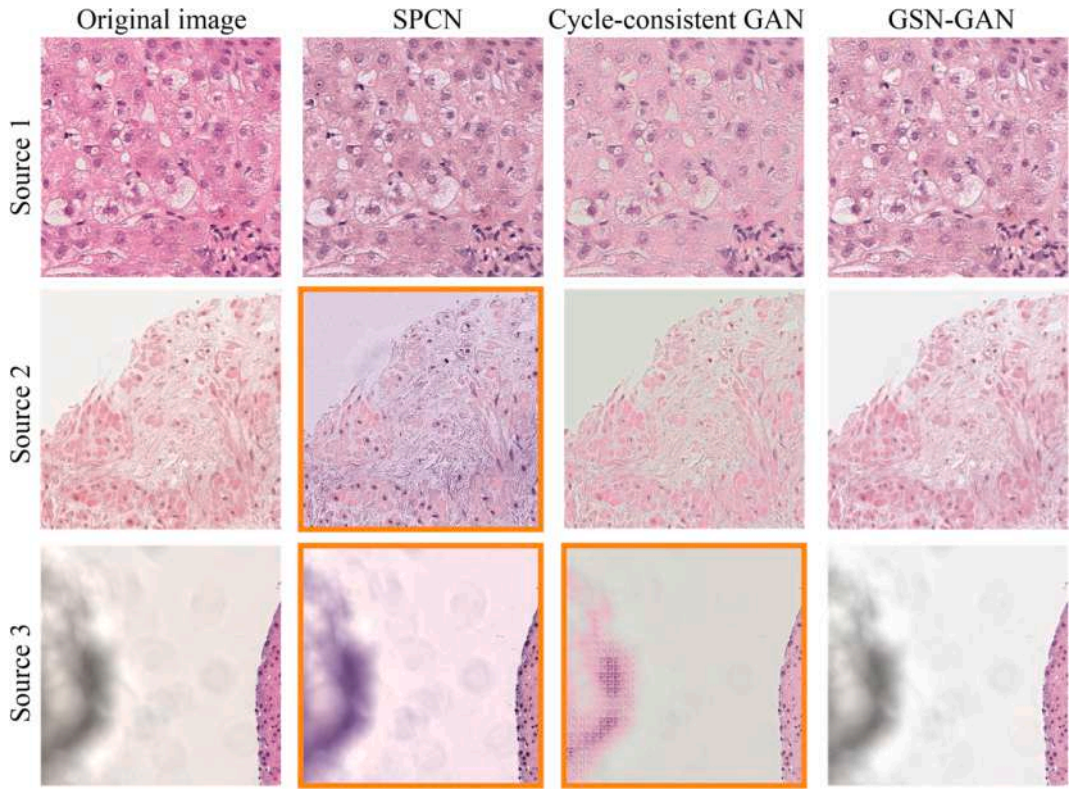


**Fig. 4.** Visual performance between the published papers (SPCN – Kumar et al., 2017; Cycle-consistent GAN - Zhou et al., 2019a) for stain normalization and the proposed method. Sub-images from the validation set are shown in columns while color normalization results are illustrated in rows. First row represents the normalized image for source image 1. The second and third row represents normalized image for source image 2 and source image 3, respectively. Results that have apparent artifacts are framed with orange boxes.

### 3.3. Dermatology

In this section, we describe the results obtained from the proposed GCC-GAN, specifically: (i) quantitative metrics in comparison with the state-of-the-art, (ii) the robustness and generalizability of our method on two external datasets, and (iii) the impact of GCC-GAN as a pre-processing step in a segmentation network.

#### 3.3.1. Quantitative metrics

Here we compared the quantitative metrics obtained on the training and validation set using the proposed GCC-GAN and other SOA methods. We report both the average values and the percentiles (10th) to highlight the presence or absence of outliers of the various methods.

To evaluate the quality of normalization, a quantitative comparison is carried out by evaluating the PCC, the PSNR and the RMSE between the normalized image and the reference image.

These metrics are extensively employed in the domain of color constancy due to their ability to quantitatively measure the effectiveness of the normalization procedure. The comparison involves three heuristic algorithms in the field of dermatology (Buchsbaum, 1980; Finlayson & Trezzi, 2004; Land, 1977) and a generative model specifically designed for color constancy (Salvi et al., 2022). The quantitative comparison showed that our GAN is the best method for color constancy (Table 6). To further validate the efficacy of our generative model in color constancy, we carried out a pairwise *t*-test between the performance of GCC-GAN with the compared techniques. All statistical analyses were conducted at a significance level (p-value) of 0.05. The outcomes of the paired *t*-test revealed a noteworthy distinction in both the training and validation sets (Table 6).

The proposed GCC-GAN achieves the best performance for PSNR and RMSE, and ranks second in PCC, with a gap of only 0.1 respect to MRGB. Comparing the proposed method with the others in terms of RMSE, a wide gap can be observed, showing how our GCC-GAN achieves a color-constancy transformation that is very close to the reference image while maintaining reliability and robustness, as evinced by the percentile values.

The visual performance of the compared methods is illustrated in Fig. 8, which clearly demonstrates how GCC-GAN achieves results that closely resemble the reference image. The use of a training pool of heterogeneous images and a custom semi-automated algorithm allows to overcome the limitations commonly encountered with statistical algorithms (Buchsbaum, 1980). As can be seen from Fig. 8, GCC-GAN is able to harmonize the appearances of healthy skin tissues, while preserving the colors and characteristics of the lesion. Moreover, our method demonstrates superior control over image contrast, as depicted in a more refined manner compared to the other methods, and effectively adjusts image exposure (Fig. 8 – first and third row).

To further demonstrate the effectiveness of our approach, a comparison was conducted in the CIELAB color-space, both before and after applying GCC-GAN for color constancy. In the Supplementary Material, Fig. S5 compares our dataset before and after applying the GCC-GAN for color constancy.

#### 3.3.2. Testing GCC-GAN on external datasets

To evaluate its generalization ability, the GCC-GAN is also applied to two external datasets with different characteristics, in terms of resolution and illuminant, when compared to the images used for the training and validation phase. The two test datasets are:

i) $PH^2$(Mendonça et al., 2013), an external dermoscopic database, including 200 images acquired using a 20x magnification, with a resolution of 768×560 pixels;

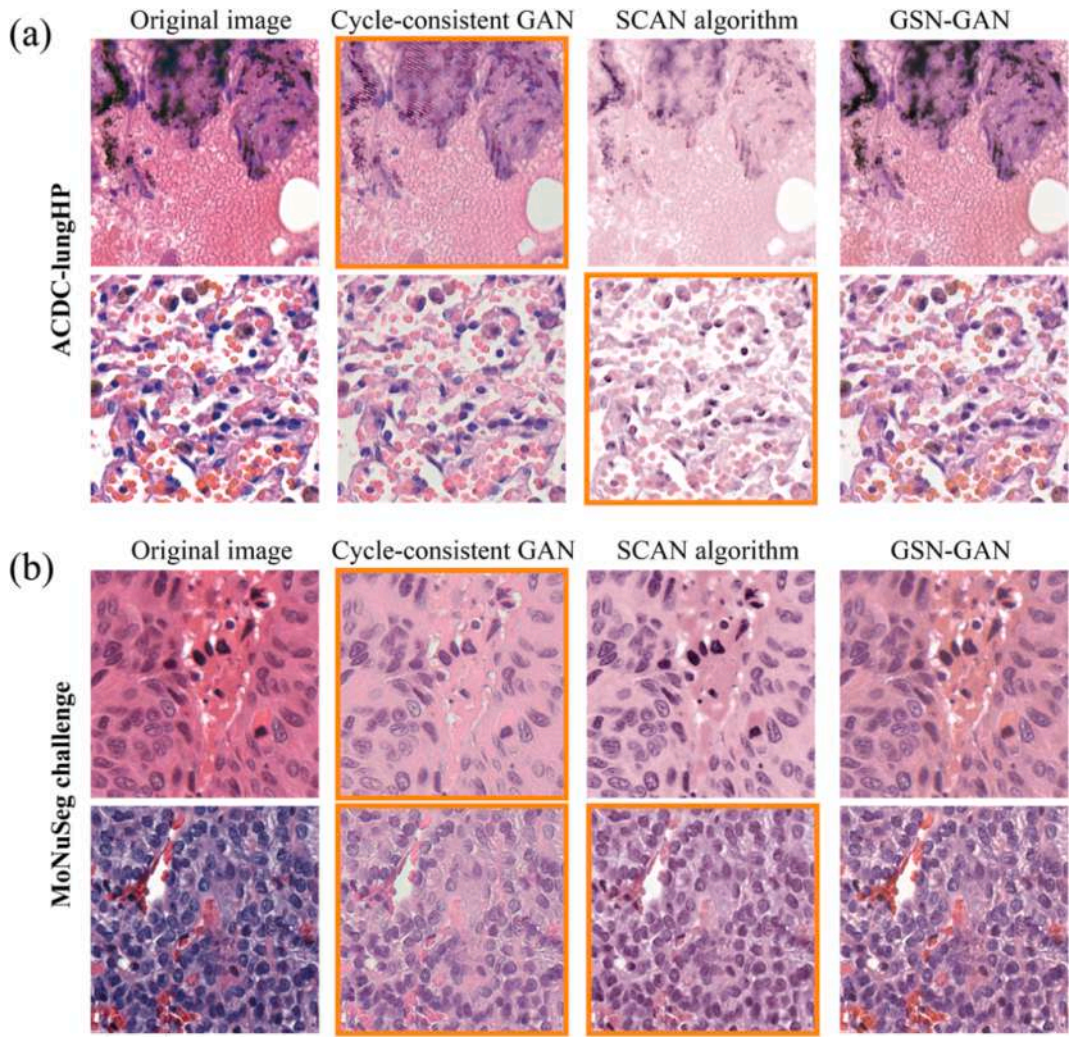


**Fig. 5.** Visual performance between the published papers for stain normalization and the proposed method on external datasets. A) Comparison between the Cycle-consistent GAN, the GSN-GAN, and the heuristic algorithm used to train our GAN (SCAN) on an external dataset (ACDC-lungHP). B) Visual comparison between our GAN and SOA methods on a second external dataset (MoNuSeg). It can be noted how the GSN-GAN provides robust results also on images containing tissue and acquired at magnification levels that were not used during the training process.

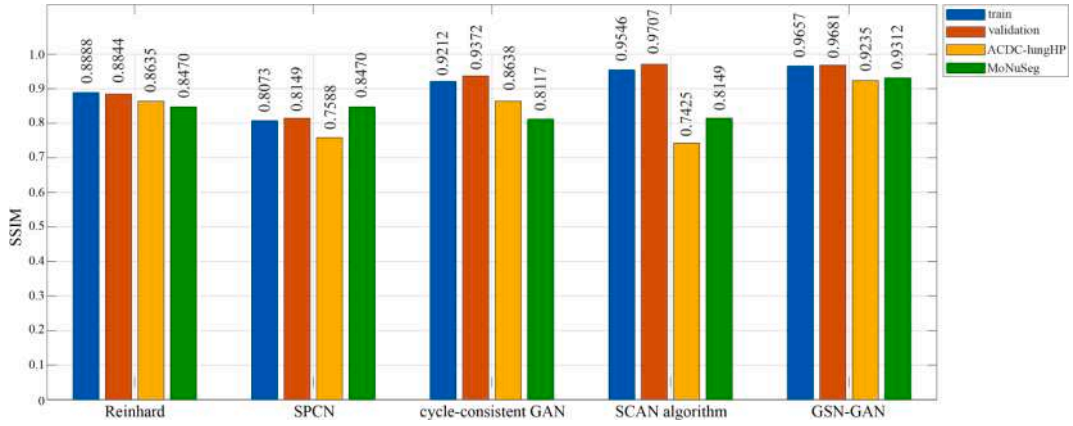


**Fig. 6.** Quantitative comparison between the GSN-GAN and the state-of-the-art methods during stain normalization. The structural similarity index SSIM is computed for all four datasets used in this work (training set, validation set, ACDC-lungHP and MoNuSeg).

ii) *NovaraDermo* (Veronese et al., 2021), a proprietary dataset, including 59 images acquired using a 10x magnification, with a resolution of 1259 × 1259 pixels.

The generalization ability is also evaluated by visually comparing the results of GCC-GAN with the main methods used in the SOA, as shown in Fig. 9(a)-(b). Our method turns out to be robust and stable even on

**Table 5**
Segmentation performance of U-Net during cell nuclei segmentation as a function of the applied pre-processing (MoNuSeg challenge – test set).

| Normalization Method | Precision | Recall | Dice | HD95 (pixels) | AJI |
|---|---|---|---|---|---|
| No normalization | 0.869 | 0.731 | 0.786 | 6.8 | 0.542 |
| Reinhard et al. (Reinhard et al., 2001) | 0.868 | 0.736 | 0.791 | 6.5 | 0.556 |
| SPCN (Kumar et al., 2017) | **0.883** | 0.712 | 0.783 | 6.6 | 0.553 |
| Heuristic SCANAlgorithm (Salvi, Michielli, et al., 2020) | 0.861 | 0.701 | 0.772 | 6.9 | 0.528 |
| Cycle-consistent GAN( Zhou et al., 2019b) | 0.841 | 0.770 | 0.796 | 7.4 | 0.532 |
| GSN-GAN (proposed) | 0.853 | **0.773** | **0.805** | **6.0** | **0.591** |

images characterized by different magnification and acquired under very different conditions than the training dataset. Fig. 9(a)-(b) shows that, even on external datasets, GCC-GAN manages to handle the color and exposure components without creating artifacts, resulting in high-contrast images with a more balanced illuminant.

### 3.3.3. *Impact of color constancy on skin lesion segmentation*

Recent studies have demonstrated that color normalization is a useful pre-processing step to improve the performance of CAD systems for skin lesion segmentation (Barata, Celebi, et al., 2014; Chabala & Jouny, 2020). To assess the impact of the color constancy provided by GCC-GAN within a deep learning framework, we employed a U-Net architecture and used the open-source HAM1000 dataset (Tschandl et al., 2018) for this segmentation task. Table 7 reports the results obtained using a standard U-Net employed to segment skin lesions when

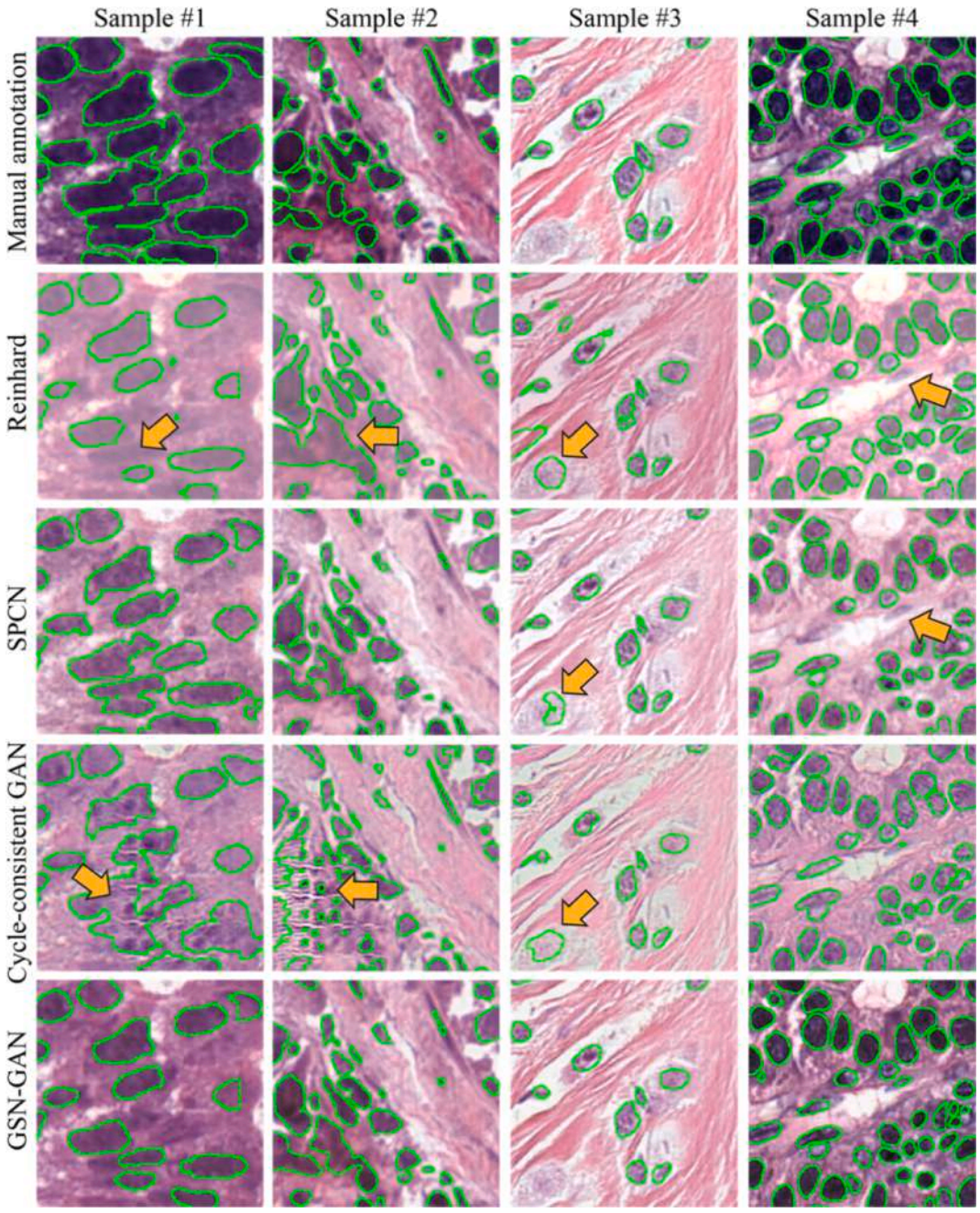


**Fig. 7.** Visual performance of the segmentation network trained with different stain normalization methods. The first row illustrates the manual annotation for four different sub-images. The compared methods are shown from the second row onwards. Orange arrows indicate segmentation errors.

**Table 6**
Quantitative metrics used to compare the GCC-GAN normalization with current state-of-the-art methods. GW: Gray World, SoG: Shades of Gray, MRGB: Max-RGB. The table shows the average values of the metrics and in brackets their percentile (10th percentile for PCC and PSNR, 90th percentile for RMSE).

| Method | Subset | Comp. Time (s) | PCC | PSNR | RMSE |
|---|---|---|---|---|---|
| GW (Buchsbaum, 1980) | Train | 0.09 | 99.5 (99.1) | 27.5 (20.5) | 13.8 (24.2) |
| | Val | 0.08 | 99.8 (99.7) | 30.0 (24.0) | 10.2 (16.9) |
| SoG (Finlayson & Trezzi, 2004) | Train | 0.04 | **99.6** (99.1) | 28.5 (20.4) | 11.9 (23.1) |
| | Val | 0.05 | 99.7 (99.6) | 31.8 (24.3) | 8.1 (14.2) |
| MRGB (Land, 1977) | Train | 0.02 | 99.5 (99.1) | 26.3 (19.9) | 15.5 (25.9) |
| | Val | 0.02 | **99.9** (**99.8**) | 29.2 (22.8) | 11.2 (18.2) |
| DermoCC-GAN (Salvi et al., 2022) | Train | 0.08 | 99.2 (98.2) | 17.6 (13.6) | 36.7 (53.2) |
| | Val | 0.08 | 99.2 (98.1) | 17.8 (13.5) | 35.8 (53.8) |
| GCC-GAN (proposed) | Train | 0.07 | 99.8 (**99.8**) | **38.6** (**35.0**)* | **2.5** (**4.3**)* |
| | Val | 0.07 | 99.8 (99.8) | **32.1** (**24.6**)* | **6.9** (**10.4**)* |

(*) Asterisk denotes statistically significant difference ($p < 0.05$) compared to state-of-the-art methods.

providing as input the images obtained with the various color constancy algorithms. The proposed GCC-GAN scores best in three of the four metrics, outperforming the other methods in terms of Hausdorff Distance, recall and dice score. Considering the dice parameter, the proposed GCC-GAN outperforms the compared methods up to 2.9 % (0.941 GCC-GAN vs. 0.912 Gray World) and achieves a 10.5 % increase when compared to the segmentation obtained on images with no color normalization applied (Fig. 10). As can be seen from the same figure, the high-quality image produced by GCC-GAN allows the segmentation network to better delineate the contour of the skin lesion.

## 4. Discussion

Color normalization plays a crucial role in reducing unwanted variability in different clinical scenarios, including the analysis of dermatological and histopathological images. Numerous studies have demonstrated the significant utility of color normalization in improving image quality for both medical experts and artificial intelligence methods (Barata, Celebi, et al., 2014; Salvi, Acharya, et al., 2020). Currently, deep learning techniques, particularly GANs, are considered the SOA methods for color normalization tasks. However, traditional GAN-based approaches have several limitations:

- Unpaired data: in dermatology and digital pathology, it is often impossible to obtain images that contain the exact same portion of tissue in both the source domain (original image) and the target domain (normalized image). Hence, models trained on unpaired data suffer from a loss of pixel-to-pixel correspondence, leading to a degradation of structural information in the source image and the generation of suboptimal images (Fig. 4);
- Lack of architecture optimization: GAN architectures are rarely optimized in terms of objective function, generator/discriminator architecture, and number of trainable parameters;
- Limited generalization capabilities: when applied to external datasets, GAN-based models often experience a drop in performance (Fig. 5). Moreover, if the distribution of the test set significantly differs from that of the training set, suboptimal or incorrect results are commonly obtained.

In this paper, we propose a novel paradigm for color normalization that addresses the limitations of current GAN-based approaches. Our method involves color style translation and a pixel-by-pixel technique, which are designed to overcome the challenges mentioned above. Notably, our model is trained on paired data and exhibits highly generalizable capabilities. The key idea is to employ as the target

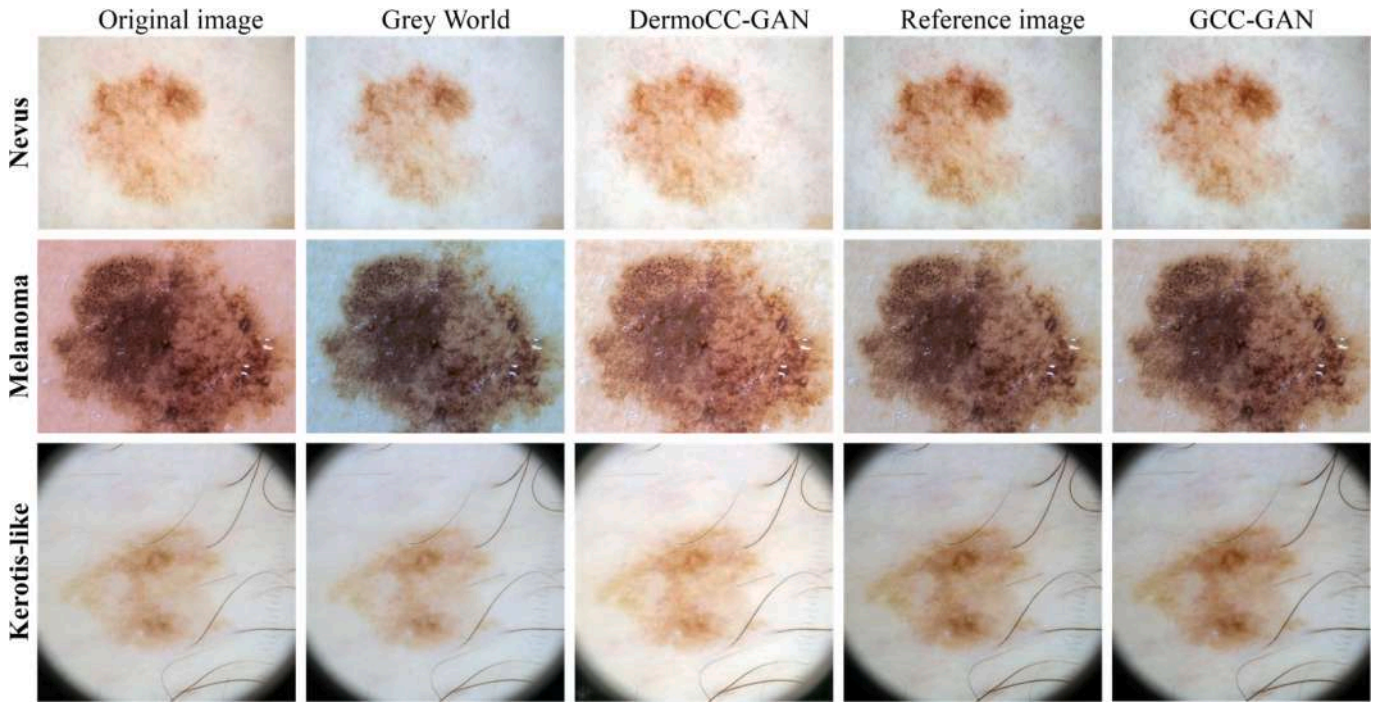


**Fig. 8.** Visual performance between the widely used color constancy in the literature, DermoCC-GAN, reference image and the proposed method. Three dermatological images of the validation set are used as an example. The original images are shown in the first colomn while compared methods are presented from the second colomn. Reference image is obtained with the semi-automatic color constancy algorithm described in Section 2.2.

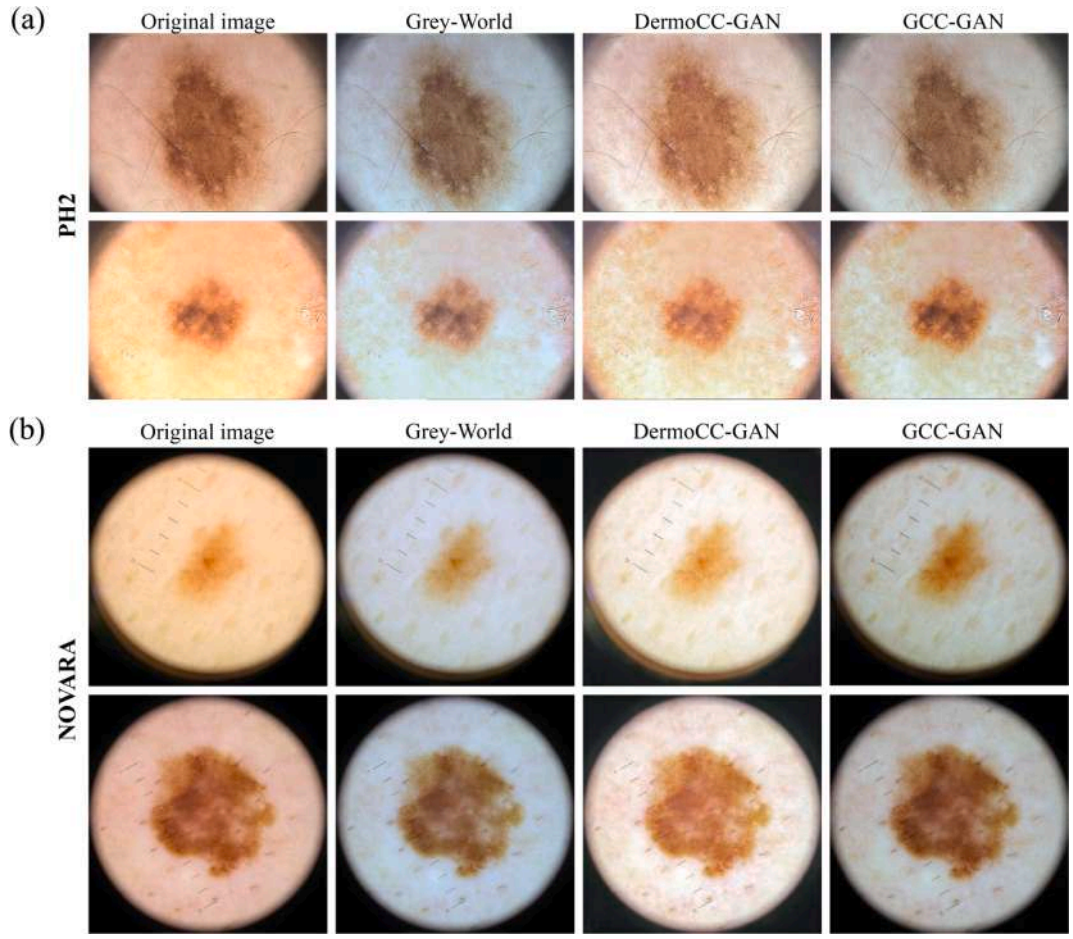


**Fig. 9.** Visual performance between the published papers for color constancy and the proposed method on external datasets. (a) Comparison between the gray-world algorithm, DermoCC-GAN, and GCC-GAN on an external dataset ($PH^2$). (b) Visual comparison between our GAN and SOA methods on an external dataset (Novara). It can be noted how the GCC-GAN provides robust results also on images characterized by different magnification and acquired under very different conditions than the training dataset.

**Table 7**
Segmentation performance of U-Net during skin lesion segmentation as a function of the applied pre-processing (HAM10000 – test set).

| Normalization Method | Precision | Recall | Dice | HD95 (pixels) |
|---|---|---|---|---|
| No normalization | 0.808 | 0.927 | 0.836 | 98.1 |
| GW (Buchsbaum, 1980) | 0.933 | 0.911 | 0.912 | 38.8 |
| SoG (Finlayson & Trezzi, 2004) | 0.951 | 0.903 | 0.918 | 34.6 |
| MRGB (Land, 1977) | 0.966 | 0.915 | 0.933 | 33.2 |
| DermoCC-GAN (Salvi et al., 2022) | **0.967** | 0.8835 | 0.917 | 31.7 |
| GCC-GAN (our) | 0.951 | **0.944** | **0.941** | **25.7** |

domain an image that has been normalized with a heuristic algorithm. In this way, a paired pixel-to-pixel correspondence is maintained between the source domain (original image) and target domain (normalized image). To ensure robustness and generalizability, our dataset includes images with various artifacts, such as unfocused areas, allowing the model to be trained on challenging images.

The proposed GSN-GAN outperforms all SOA methods both qualitatively and quantitatively for different tissues and magnifications. We demonstrate the effectiveness of our stain normalization algorithm and show quantitatively improved nuclei segmentation performance on an evaluation dataset (Table 5). Importantly, the GSN-GAN avoids introducing artifacts in the normalized image and exhibits high generalizability, producing optimal results even on images containing tissue not present in the training set (e.g., lung histopathological tissue) or acquired at different magnification levels (e.g., 40x).

The novel GCC-GAN is specifically designed for color constancy in dermatological images. Our quantitative evaluation demonstrates that our GAN can effectively generalize on unseen data collected from other patients and centers (Fig. 9(a)-(b)). The color standardization provided by GCC-GAN is independent of the starting illumination preset, allowing it to distinguish between the informative content of the skin lesion, which needs to be preserved, and the illumination component, which needs to be normalized.

The optimization of parameters and architecture in the proposed GANs significantly impacts performance outcomes. In fact, by simply changing the architecture compared to the baseline, performance results can be improved by up to 60 % (Table 3). It is surprising to note that the aspect of GAN architecture and parameter optimization is often overlooked in many studies, highlighting the need for future research to focus on this area when utilizing GANs. The pixel-to-pixel matching between source and target domains enables GANs to preserve the content of the source image, serving as an enabling technology in our approach.

While the presented paradigm holds great promise, it does have some limitations. Firstly, the fixed input size of the GAN can be a drawback: smaller images need to be resampled, and larger images must be divided into patches or undersampled. However, it is important to underline that this limitation is inherent to all GAN-based approaches. The GSN-GAN, trained on images with a 20x magnification level and tested on 40x images, yields satisfactory results. However, it may produce suboptimal

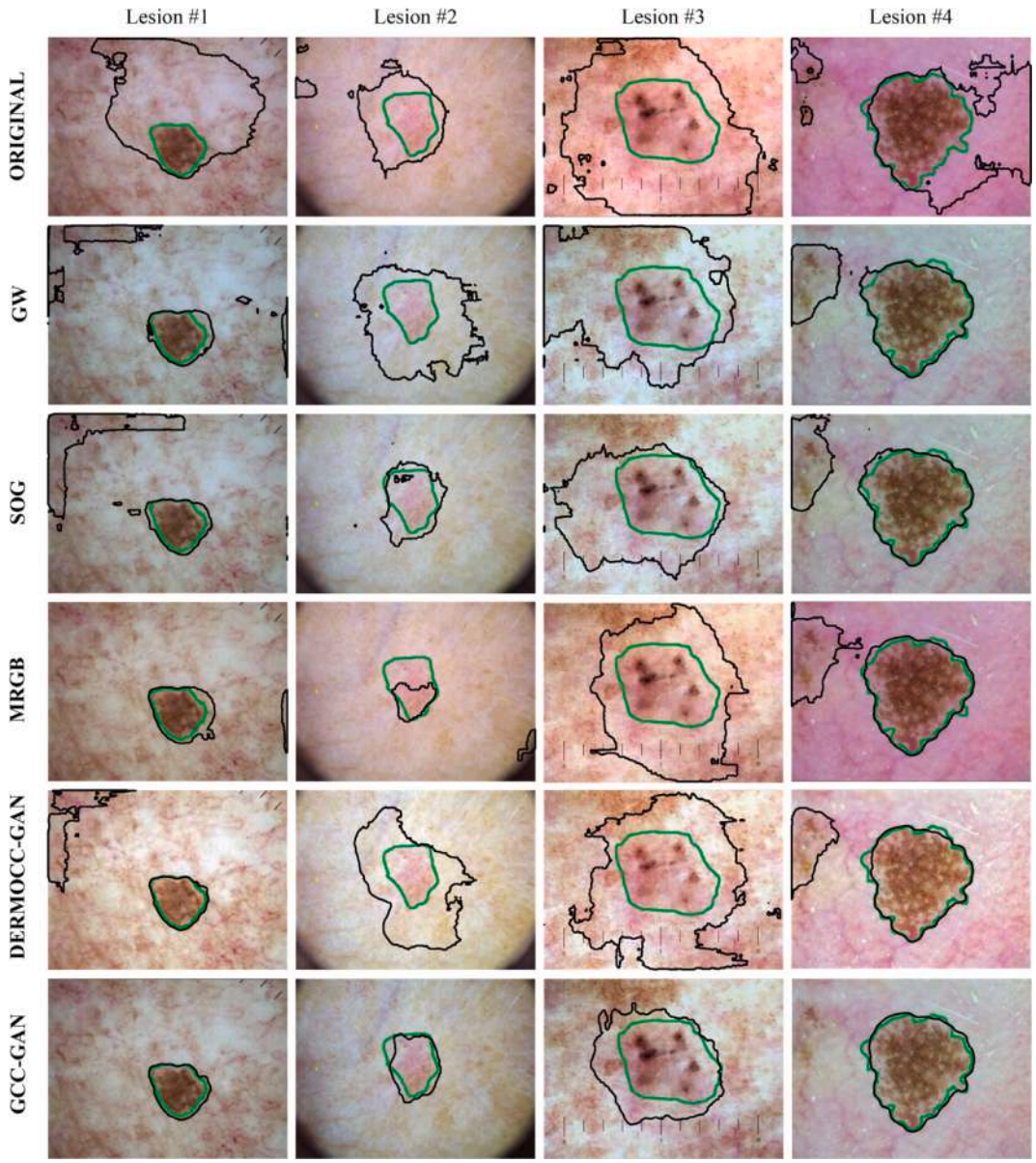


**Fig. 10.** Visual performance of the segmentation network trained with different color constancy methods. The first row illustrates the manual annotation for four different skin lesions. The compared methods are shown from the second row onwards.

results if tested on smaller magnification levels, such as 5x or 10x. For dermatological images, if the source image exceeds 1024 × 1024 pixels, undersampling is necessary, which poses the risk of losing important information due to slight resolution decrease.

We identify two points that may be promising future avenues of research. First of all, custom loss functions, such as the $SSIM_{LOSS}$ (Zhao et al., 2016), could be defined to support traditional loss functions of the LS-GAN, providing increased control over the training phase and expediting model convergence. Secondly, this approach based on a paired dataset (i.e., using a heuristic algorithm to train a GAN) could be extended to other applications, including but not limited to noise removal, artifact correction and contrast enhancement. Computationally expensive heuristic algorithms or those requiring fine-tuning of numerous parameters can be learned by the GAN-based approach presented here, enabling parameter-free and real-time reproduction.

## 5. Conclusion

This study presents a novel paradigm for color normalization of histological and dermatological images based on generative models. Our approach achieves optimal results by employing a heuristic algorithm to train the GAN specifically for the color normalization task. Extensive experiments are conducted to evaluate different GAN architectures in terms of parameters and loss function, aiming to quantify the quality of the generated images. Different metrics are calculated for both digital pathology and dermatology to identify the most suitable GAN configuration for each application.

Our models not only deliver superior qualitative and quantitative results but also offer enhanced practical usability compared to existing methods. Experimentation on publicly available dataset demonstrates that the proposed framework outperforms previous color normalization solutions by generating color-consistent images, preserving information, and obtaining high training efficiency. These results highlight the potential of the proposed paradigm as a crucial tool in the pipeline of automatic quantitative algorithms. Its implementation can effectively reduce color variability and enhance final segmentation performance.

## CRediT authorship contribution statement

**Massimo Salvi:** Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Francesco Branciforti:** Data curation, Formal analysis, Methodology, Visualization, Writing – review & editing. **Filippo Molinari:** Supervision, Writing – review & editing. **Kristen M. Meiburger:** Formal analysis, Supervision, Writing – original draft, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgements

This project has been partially funded from one of the calls under the Photonics Public Private Partnership (PPP): H2020-ICT-2020-2 with Grant Agreement ID 101016964 (REAP). We would moreover like to thank Dr. Martino Bosco (pathologist) and Dr. Elisa Zavattaro (dermatologist) for their guidance and feedback on this work.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2023.123105.

## References

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. *International Conference on Machine Learning*, 214–223.

Barata, C., Celebi, M. E., & Marques, J. S. (2014). Improving dermoscopy image classification using color constancy. *IEEE Journal of Biomedical and Health Informatics, 19*(3), 1146–1152.

Barata, C., Marques, J. S., & Celebi, M. E. (2014). Improving dermoscopy image analysis using color constancy. In *2014 IEEE International Conference on Image Processing, ICIP 2014*, *19*(3), 3527–3531. https://doi.org/10.1109/ICIP.2014.7025716.

Bojan-Dragos, C.-A., Precup, R.-E., Preitl, S., Roman, R.-C., Hedrea, E.-L., & Szedlak-Stinean, A.-I. (2021). GWO-based optimal tuning of type-1 and type-2 fuzzy controllers for electromagnetic actuated clutch systems. *IFAC-PapersOnLine, 54*(4), 189–194.

Borlea, I.-D., Precup, R.-E., & Borlea, A.-B. (2022). Improvement of K-means cluster quality by post processing resulted clusters. *Procedia Computer Science, 199*, 63–70.

Buchsbaum, G. (1980). A spatial processor model for object colour perception. *Journal of the Franklin Institute, 310*(1), 1–26.

Chabala, W. F., & Jouny, I. (2020). Comparison of Convolutional Neural Network Architectures on Dermastopic Imagery. In *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 928–931).

Cho, H., Lim, S., Choi, G., & Min, H. (2017). Neural stain-style transfer learning using gan for histopathological images. *ArXiv Preprint*. ArXiv:1710.08543.

Finlayson, G. D., & Trezzi, E. (2004). Shades of gray and colour constancy. *Color and Imaging Conference, 2004*(1), 37–41.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems, 27*.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in Neural Information Processing Systems, 30*.

Janowczyk, A., & Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics, 7*.

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. *Government Information Quarterly, 37*(3), Article 101493.

Kumar, N., Verma, R., Anand, D., Zhou, Y., Onder, O. F., Tsougenis, E., … Hu, Z. (2019). A multi-organ nucleus segmentation challenge. *IEEE Transactions on Medical Imaging, 39*(5), 1380–1391.

Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., & Sethi, A. (2017). A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging, 36*(7), 1550–1560.

Land, E. H. (1977). *The Retinex Theory of Color Vision*.

Li, C., & Wand, M. (2016). Precomputed real-time texture synthesis with markovian generative adversarial networks. *European Conference on Computer Vision*, 702–716.

Li, Z., Zhang, J., Tan, T., Teng, X., Sun, X., Zhao, H., … Li, Y. (2020). Deep learning methods for lung cancer segmentation in whole-slide histopathology images—the acdc@ lunghp challenge 2019. *IEEE Journal of Biomedical and Health Informatics, 25*(2), 429–440.

Liang, H., Plataniotis, K. N., & Li, X. (2020). Stain Style Transfer of Histopathology Images Via Structure-Preserved Generative Learning. *International Workshop on Machine Learning for Medical Image Reconstruction*, 153–162.

Lo, Y.-C., Chung, I.-F., Guo, S.-N., Wen, M.-C., & Juang, C.-F. (2021). Cycle-consistent GAN-based stain translation of renal pathology images with glomerulus detection application. *Applied Soft Computing, 98*, Article 106822.

Mahapatra, D., Bozorgtabar, B., Thiran, J.-P., & Shao, L. (2020). Structure preserving stain normalization of histopathology images using self supervised semantic guidance. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 309–319.

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., & Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2794–2802).

Mendonça, T., Ferreira, P. M., Marques, J. S., Marcal, A. R. S., & Rozeira, J. (2013). PH 2-A dermoscopic image database for research and benchmarking. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 5437–5440).

Pontalba, J. T., Gwynne-Timothy, T., David, E., Jakate, K., Androutsos, D., & Khademi, A. (2019). Assessing the impact of color normalization in convolutional neural network-based nuclei segmentation frameworks. *Frontiers in Bioengineering and Biotechnology, 7*, 300.

Pozna, C., Minculete, N., Precup, R.-E., Kóczy, L. T., & Ballagi, Á. (2012). Signatures: Definitions, operators and applications to fuzzy modelling. *Fuzzy Sets and Systems, 201*, 86–104.

Precup, R.-E., David, R.-C., Roman, R.-C., Petriu, E. M., & Szedlak-Stinean, A.-I. (2021). Slime mould algorithm-based tuning of cost-effective fuzzy controllers for servo systems. *International Journal of Computational Intelligence Systems, 14*(1), 1042–1052.

Reinhard, E., Adhikhmin, M., Gooch, B., & Shirley, P. (2001). Color transfer between images. *IEEE Computer Graphics and Applications, 21*(5), 34–41.

Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., … Gutman, D. (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data, 8*(1), 1–8.

Salehi, P., & Chalechale, A. (2020). Pix2pix-based stain-to-stain translation: A solution for robust stain normalization in histopathology images analysis. *International Conference on Machine Vision and Image Processing (MVIP), 2020*, 1–7.

Salvi, M., Acharya, U. R., Molinari, F., & Meiburger, K. M. (2020). The impact of pre-and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. *Computers in Biology and Medicine, 104129*.

Salvi, M., Branciforti, F., Veronese, F., Zavattaro, E., Tarantino, V., Savoia, P., & Meiburger, K. M. (2022). *DermoCC-GAN: A new approach for standardizing dermatological images using generative adversarial networks*. Under Review on Computer Methods and Programs in Biomedicine.

Salvi, M., Caputo, A., Balmativola, D., Scotto, M., Pennisi, O., Michielli, N., … Fraggetta, F. (2023). Impact of stain normalization on pathologist assessment of prostate cancer: A comparative study. *Cancers, 15*(5), 1503.

Salvi, M., Michielli, N., & Molinari, F. (2020). Stain Color Adaptive Normalization (SCAN) algorithm: Separation and standardization of histological stains in digital pathology. *Computer Methods and Programs in Biomedicine, 193*, Article 105506.

Salvi, M., Molinari, F., Dogliani, N., & Bosco, M. (2019). Automatic discrimination of neoplastic epithelium and stromal response in breast carcinoma. *Computers in Biology and Medicine, 110*. https://doi.org/10.1016/j.compbiomed.2019.05.009

Salvi, M., Molinaro, L., Metovic, J., Patrono, D., Romagnoli, R., Papotti, M., & Molinari, F. (2020). Fully automated quantitative assessment of hepatic steatosis in liver transplants. *Computers in Biology and Medicine, 123*, Article 103836.

Sandfort, V., Yan, K., Pickhardt, P. J., & Summers, R. M. (2019). Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Scientific Reports, 9*(1), 1–9.

Sidorov, O. (2019). Conditional gans for multi-illuminant color constancy: Revolution or yet another approach?. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Singh, D., & Shukla, A. (2022). Manifold optimization with MMSE hybrid precoder for Mm-Wave massive MIMO communication. *Science and Technology, 25*(1), 36–46.

Swiderska-Chadaj, Z., de Bel, T., Blanchet, L., Baidoshvili, A., Vossen, D., van der Laak, J., & Litjens, G. (2020). Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer. *Scientific Reports, 10*(1), 1–14.

Tan, G.-W.-H., Ooi, K.-B., Leong, L.-Y., & Lin, B. (2014). Predicting the drivers of behavioral intention to use mobile learning: A hybrid SEM-Neural Networks approach. *Computers in Human Behavior, 36*, 198–213.

Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data, 5*(1), 1–9.

Veronese, F., Branciforti, F., Zavattaro, E., Tarantino, V., Romano, V., Meiburger, K. M., … Savoia, P. (2021). The role in teledermoscopy of an inexpensive and easy-to-use

smartphone device for the classification of three types of skin lesions using convolutional neural networks. *Diagnostics, 11*(3), 451. https://doi.org/10.3390/diagnostics11030451

von Kries, J. (1970). *Chromatic adaptation, Sources of Color Vision*. Cambridge, Mass: MIT Press.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600–612.

Yuan, E., & Suh, J. (2018). Neural stain normalization and unsupervised classification of cell nuclei in histopathological breast cancer images. *ArXiv Preprint*. ArXiv: 1811.03815.

Zanjani, F. G., Zinger, S., Bejnordi, B. E., van der Laak, J. A. W. M., & de With, P. H. N. (2018). Stain normalization of histopathology images using generative adversarial networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 573–577.

Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2016). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging, 3*(1), 47–57.

Zheng, Y., Jiang, Z., Zhang, H., Xie, F., Shi, J., & Xue, C. (2019). Adaptive color deconvolution for histological WSI normalization. *Computer Methods and Programs in Biomedicine, 170*, 107–120.

Zhou, N., Cai, D., Han, X., & Yao, J. (2019). Enhanced cycle-consistent generative adversarial network for color normalization of H&E stained images. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 694–702.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2223–2232).