An Encoder-Decoder CNN for Hair Removal in Dermoscopic Images

Lidia Talavera-Martínez\textsuperscript{a,b,*}, Pedro Bibiloni\textsuperscript{a,b}, Manuel González-Hidalgo\textsuperscript{a,b}

\textsuperscript{a}SCOPIA research group. University of the Balearic Islands. Dpt. of Mathematics and Computer Science. Crta. Valldemossa, km 7.5, E-07122 Palma, Spain
\textsuperscript{b}Health Research Institute of the Balearic Islands (IdISBa), E-07010 Palma, Spain

Abstract

The process of removing occluding hair has a relevant role in the early and accurate diagnosis of skin cancer. It consists on detecting hairs and restore the texture below them, which is sporadically occluded. In this work, we present a model based on convolutional neural networks for hair removal in dermoscopic images. During the network’s training, we use a combined loss function to improve the restoration ability of the proposed model. In order to train the CNN and to quantitatively validate their performance, we simulate the presence of skin hair in hairless images extracted from publicly known datasets such as the PH2, dermquest, dermis, EDRA2002, and the ISIC Data Archive. As far as we know, there is no other hair removal method based on deep learning. Thus, we compare our results with six state-of-the-art algorithms based on traditional computer vision techniques by means of similarity measures that compare the reference hairless image and the one with hair simulated. Finally, a statistical test is used to compare the methods. Both qualitative and quantitative results demonstrate the effectiveness of our network.

Keywords: Hair removal, Dermoscopy, Deep Neural Networks, Skin Lesion, Image Processing, Inpainting

1. Introduction

Melanoma is the most aggressive, metastatic and deadliest type of skin cancers, turning this disease into a major problem for public health. In Europe, it accounts for 1–2% of all malignant tumors \cite{2}, and its estimated mortality in 2018 was 3.8 per 100,000 men and women per year \cite{1}. Although melanoma is still incurable, its early diagnosis is of great importance. Its early detection can prevent malignancy and increase the survival rate and the effectiveness of the treatment. Nowadays, practitioners rely on the dermoscopic evaluation for completing the clinical analysis and the diagnosis of melanoma. This practice improves the diagnostic accuracy up to 10–30\% \cite{23} compared to simple clinical observation. This in-vivo, non-invasive skin imaging technique enables the visualization of specific subsurface structures, forms and colors that are not discernible by a simple visual inspection. The diagnosis of skin lesions is mainly based on their morphological characteristics, such as an irregular shape, asymmetry and a variety...
of colors, along with a history of changes in size, shape, color and/or texture. However, their evaluation might be altered by the individual judgment of the observers, which depends on their experience and subjectivity [6]. Thus, in order to help physicians to obtain an early, objective, and reproducible diagnosis of skin lesions, sophisticated Computer-Aided Diagnosis (CAD) software are developed. These computational tools are designed based on clinical protocols [4] [18] [25] [32] and focus mainly on image acquisition, artifact removal (hairs, bubbles, etc.), segmentation of the lesion, extraction and selection of features, and final classification of the lesion.

In the pre-processing stage during the analysis of the lesions, hair removal is one of the key steps. The presence of hair in dermoscopic images usually occludes significant patterns reducing the accuracy of the system. Once the hairs are detected and removed, the next step is to estimate and restore the underlying information (i.e., color and texture patterns) of the skin pixels underneath the hair regions. Extensive previous research has been done addressing the hair removal process in dermoscopic images. To the best of our knowledge, previous works presented approaches based on traditional computer vision techniques. Thus, tackling the problem with classical generative and discriminative models, which rely on hand-crafted features. However, hand-crafted features needed to be defined and their impact is typically tested with small datasets [28]. In recent years, deep learning has shown to be a powerful tool for image analysis. More specifically, deep learning techniques have achieved a higher performance with respect to traditional approaches for the majority of applications within the medical field [20]. Deep Convolutional Neural Networks (CNNs) allow to automatically learn features of different complexity directly from data through a set of trainable filters. Moreover, it has shown to be a powerful tool when working with large datasets.

As we have mentioned before, CAD systems may include a preprocessing step in which hair removal stands out as one of the most useful and used methods. However, traditional approaches are still used, nowadays, for this task in more advanced systems in which the main model uses deep learning techniques [8] [30]. Thus, we face the task of developing a deep learning model for the detection and removal of hairs. Such model could be integrated into a complete CAD system based entirely on deep learning. Our objective is threefold. First, to design a novel model based on deep CNNs for the removal of skin hair in dermoscopic images and the subsequent restoration of the affected pixels. Second, to qualitative and quantitatively assess its performance. Third, to compare the results with other hair removal strategies using the same database, which would provide an objective comparison of the strengths and weaknesses, and therefore of the quality of the method presented.

In summary, we make the following contributions:

- To the best of our knowledge, we are the first to propose the use of a CNN model for hair removal in dermoscopic images.

- We introduce a loss function for the detection and posterior restoration of hair’s pixels based on the combination of well-known loss functions. We also present an ablation study of the different loss terms.

- We extend the dataset presented in [33], through the creation of more synthetic hair on dermoscopic images from public datasets. This allows us to carry out a quantitative evaluation of hair removal approaches.
The rest of the document is structured as follows. First, in Section 2, we review the related work on hair removal methods in dermoscopic images. Then, in Section 3, we present our method based on CNNs and detail the loss function that we have used to train the network. Next, in Section 4, we describe the database and the implementation details, as well as presenting the results obtained, and performing an ablation study of the loss terms and some architecture’s aspects. Finally, in Section 5, we discuss the previous results and study the strengths and limitations of this study.

2. Related Work

This section briefly describes previous works that addressed the task of hair removal.

Several traditional computer vision approaches have been used for hair removal in dermoscopic images. Here we highlight six state-of-the-art algorithms, based on their availability and wide use in the literature. These are the ones proposed by Lee et al. [19], Xie et al. [41], Abbas et al. [3], Huang et al. [15], Toossi et al. [35] and Bibiloni et al. [10]. A summary of the approaches considered by each of them can be seen in Table 1.

| Year | Method | Hair segmentation | Inpainting method | Color space |
|------|--------|-------------------|-------------------|-------------|
| 1997 | Lee et al. [19] | Grayscale closing | Bilinear interpolation | RGB |
| 2009 | Xie et al. [41] | Grayscale top-hat | Non-linear PDE | RGB |
| 2011 | Abbas et al. [3] | Derivative of Gaussians | Coherence transport | CIELab |
| 2013 | Huang et al. [15] | Conventional matched filters | Region growing algorithms and linear discriminant analysis | RGB |
| 2013 | Toossi et al. [35] | Canny edge detector | Coherence transport | RGB |
| 2017 | Bibiloni et al. [10] | Color top-hat | Morphological inpainting | CIELab |

Table 1: Detection and inpainting techniques employed in the literature to remove hair from dermoscopic images.

Deep learning techniques have been used to face a large number of tasks in computer vision. Specifically, deep learning-based image restoration techniques have been used in other fields for image inpainting [42], image deblurring and image denoising [34], among others. These methods learn the parameters of the network to reconstruct images directly from training data, that is composed by pairs of clean and corrupted images. This is usually more effective in real-world images. For instance, Xie et al. [42] proposed an approach for image denoising and blind inpainting that combines sparse coding with pre-trained deep networks, achieving good results in both tasks. Vincent et al. [36] presented a stack of denoising autoencoders for image denoising that is applied not only to the input, but also recursively to intermediate representations, to initialize the deep neural network. Also, Cui et al. [12] proposed a cascade of multiple stacked collaborative local autoencoders for image super-resolution. Their method searches in each layer non-local self-similarity to enhance high frequency texture details from the image patches to suppress the noise and combine the overlapping patches. In [22], Mao et al. proposed an encoding-decoding framework for image denoising and super-resolution combining convolutions and deconvolution layers linked symmetrically by skip connections, which helps improving the training process and the network’s performance. Finally, Jain et al. [16] and
Dong et al. [13] proposed a fully convolutional CNN for image denoising and image super-resolution, respectively. The authors shown that their methods achieve comparable results to traditional computer vision techniques.

Despite the good performance of the above mentioned works, and to the best of our knowledge, no previous research has been done on the field of hair removal in dermoscopic images relying on deep learning. Although Attia et al. works with this technique to simulate hair in dermoscopic images. This work aims to fill that gap and serve as a baseline for future works.

3. Methodology

In this section, we describe our model for hair removal in dermoscopic images. We specifically focus on our proposed reconstruction loss function.

We designed a convolutional encoder-decoder architecture for the task of hair removal in dermoscopic images. The model, with 12 layers, is described in Figure 1. The input of our model is the pair formed by the reference image without hair, used only when computing the loss function, and its corresponding image with simulated hair. On one hand, the encoder, with an input of size $512 \times 512 \times 3$, extracts a hidden representation of high-level features from the image. When looking for such features, the encoder tends to ignore noise. In our case, we expect the network to output hairless images, therefore hair becomes the noise that can be ignored by the encoder. On the other hand, the decoder aims to recover the missing information from the high-level feature representation. Its output is a $512 \times 512 \times 3$ cleaned version, without skin hair, of the input image. Both the encoder and the decoder have two blocks. Each block of the encoder consists of one $3 \times 3$ convolution, of 128 filters in the first block and 256 filters in the second one, followed by a down-sampling operation, which is applied by a two-strided $3 \times 3$ convolution to reduce the spatial resolution. On the other side, each block of the decoder consists of an up-sampling of the feature map by a deconvolution of $3 \times 3$ with strides of two in each dimension. A skip connection follows, which concatenates the up-sampled output with the corresponding feature map from the layer of equal resolution of the encoder. This enables the decoder to recover image details, and therefore improves the restoration performance. Next, a $3 \times 3$ convolution is applied over the merged data. Finally, in the last block, an additional $3 \times 3$ convolution is added to reduce the feature map to the number of output channels.

Reconstruction Loss Function

The loss function guides how the network learns by serving as a criterion that numerically reflects the errors of the model. It is computed between the network output and its corresponding hairless reference image, also known as Ground Truth (GT). There are several loss functions that have been used in image restoration tasks. Some widespread losses are the Mean Square Error (MSE) or the Mean average Error (MAE). These measurements exclusively depend on the difference between the corresponding pixels of the two images. Therefore, the results might have poor quality in terms of human perception, since the noise of a pixel should not be considered independently of the error of
Figure 1: Architecture of our proposed network. The pairs of reference hairless images and its corrupted (hair simulated) images are passed through the encoder to extract complex features. The decoder, connected to the encoder with skip-connections, reconstructs the image.
its neighbouring pixels. To overcome these limitations, other loss metrics have been proposed, such as the Structural
Similarity Metric (SSIM) or the Multiscale Structural Similarity Metric (MSSSIM), which depend on local luminance,
contrast and structure [43].

To achieve results appealing to a human observer, and inspired by the results obtained by Liu et al. in [21], we
propose to capture the best characteristics of the loss functions that measure statistical features locally along with
other per-pixel losses.

Thus, our reconstruction loss is defined as follows:

$$L_{\text{rec}} = \alpha L_{\text{foreground}} + \beta L_{\text{background}} + \gamma L_{\text{composed}} + \delta L_{\text{SSIM}} + \lambda L_{\text{tv}},$$

(1)

where $\alpha, \beta, \gamma, \delta$ and $\lambda$ are the weights of each term of the linear combination that define the reconstruction loss function.

We opted to perform a random hyperparameter search as there are many parameters of which we have to find the
optimal value, and a grid search would require a higher computational cost. Specifically, we performed 10 runs of
our model, assigning in each case a random value between 0 and 10 to each of the weights. Afterwards, a statistical
test indicates which are the best set of weights according to the measurements explained in Section 4.4. The term
$L_{\text{foreground}}$ is the $L_1$ distance between the GT and the prediction of the network only between those pixels belonging
to the hair areas. Next, $L_{\text{background}}$ estimates the $L_1$ distance between the GT and the network’s prediction only among
the background pixels. Then, $L_{\text{composed}}$ computes the $L_2$ restricted to the hair, but normalizing over all pixels rather
than the amount of hair pixels. The term $L_{\text{SSIM}}$ calculates the loss function based on the SSIM metric over the whole
image. Finally, we use a total variation loss, $L_{\text{tv}}$, as a regularizer to smooth the transition of the predicted values for
the regions corresponding to hair, according to their surrounding context. A more detailed description of this term can
be found in [29].

4. Experimental framework and Results

In this section, we first establish the experimental framework by describing the database used and the implemen-
tation details of our method. Then, we analyze the results obtained by our method and compare them, qualitative and
quantitatively, to the six traditional hair removal methods presented in Section 2. To obtain the numerical results, we
rely on several performance measures. We determine which method outperforms the rest by means of a statistical test.
Finally, we conducted an ablation study of the loss terms, as well as of some aspects of the model’s architecture.

4.1. Dataset description

In order to train the CNN and to quantitatively validate, in an effective way, the performance of our method,
we need a dataset. It must contain pairs of images: images with hair, used as the algorithm input, along with their
corresponding “clean” version, in this case the same image without hair.

Finding this type of data is a challenging task, the same dermoscopic image can not be captured with and without
hair. To tackle this problem, we decided to simulate the presence of skin hair in hairless images extracted from five
publicly available datasets, i.e. PH2 [24], dermquest[1] dermis[2] EDRA2002 [5] and from the ISIC Data Archive[3].

We have avoided selecting images with other artifacts (e.g. ruler, bandages etc) that are not hairs. Three different hair simulation methods have been used. The first is the one, presented by Attia et al. [7], is based on generative adversarial networks. The second one was implemented by Mirzaalian et al. [27], whose software “HairSim” is publicly available in [26]. Finally, the last approach we have used involved extracting hair masks by an automated method, proposed by Xie et al. [41], and superimpose them on hairless images.

We constructed a dataset with 618 images, which consists of 322 images from the EDRA2002 dataset, 239 images from the PH2 dataset, 46 images from the ISIC Data Archive, 6 images from the dermis dataset, and 5 from the dermquest dataset. During the experimentation we divide it into 70% for training and 30% for the test, which gives us 433 and 185 images, respectively. It is composed of images with diverse hair that present a variety of hair thickness, density and color, ranging from coarse to more realistic. This diversity guarantees that we consider hairs with very different characteristics when training the network. We also introduced images without hair to teach the network that some images must not be modified and their textures must be maintained. In Figure 2, we show examples of original hairless images from the dataset, together with the simulations obtained using each of the hair simulation approaches.

As we can see, with the first and third method we achieve a much more natural morphology, quantity and distribution that resembles real hair compared to the second. We consider a reduced number of testing images to maximize the training samples and, thus, to help the network learn and generalize well.

4.2. Experimental setup

We implemented the proposed architecture using Keras [11]. The network was trained from scratch with randomly initialized weights and using the Adam [17] optimizer with a learning rate experimentally set to $10^{-4}$. The coefficients for the different terms of the reconstruction loss function with which the network has been trained were experimentally found to be: $\alpha = 2.626, \beta = 3.892, \gamma = 0.309, \delta = 0.398$ and $\lambda = 0.597$. The network was trained on a single NVIDIA GeForce GTX 1070 with a batch size of 4.

In Figure 3, we can see how our model trained over almost 25 epochs (reaching an early stopping policy based on monitoring the validation loss), and how the Peak Signal-to-Noise Ratio (PSNR) metric evolves satisfactory during the training.

4.3. Qualitative results

We conducted a qualitative and quantitative analysis of the results obtained. In Figure 4, we can see that our proposed method attains visually appealing results. In addition, we present a visual comparison of our results with the ones obtained by other hair removal methods presented in Section 2. Let us remark that we compare with these methods as they are the ones that have been applied to dermoscopic images.

[1] Was deactivated on December 31, 2019
[2] www.dermis.net
[3] www.isic-archive.com
As can be seen in Figure 5, not all the methods are successful in both the hair removal and the subsequent process of inpainting. For instance, Abbas et al.’s and Toossi et al.’s methods are not capable of segment the hairs as it seems they are not able to detect them properly. In contrast, Huang et al.’s and Xie et al.’s methods are capable of detecting much of the hair. However, their inpainting process seems to leave a trail of them. Finally, our results and the ones of Bibiloni et al.’s and Lee et al.’s methods seem to adjust to the reference image at first sight. Although, the last two leave traces, while the new method does not. However, it may be the case that some of them introduce some alterations, to a greater or lesser extent, that blur the lesion’s features, such as streaks or reticular textures.

Both in Figure 4 and 5, we have seen that our method reaches good visual results when evaluated on synthetic
images. In Figure 6, we show its effectiveness and its generalization ability in dermoscopic images with real hair. We show images from the 5 databases, to demonstrate that although the data is not balanced, the network has not suffered database-specific overfitting.

4.4. Quantitative results

It is worth noting that once the hairs are removed there are many possible solutions as to what is the expected inpainting result, always with the aim of preserving the texture of the area involved. Therefore, a qualitative evaluation

| Reference image | Hair simulation | Our method |
|-----------------|-----------------|------------|
| ![Reference image](image1) | ![Hair simulation](image2) | ![Our method](image3) |
| ![Reference image](image4) | ![Hair simulation](image5) | ![Our method](image6) |
| ![Reference image](image7) | ![Hair simulation](image8) | ![Our method](image9) |
| ![Reference image](image10) | ![Hair simulation](image11) | ![Our method](image12) |
| ![Reference image](image13) | ![Hair simulation](image14) | ![Our method](image15) |

Figure 4: Example of the hair removal results obtained by our method.
is not enough to evaluate the quality of the different methods introduced. In the following, we introduce an automatic, objective and comparable performance evaluation system.

We used a set of nine objective error metrics to quantitatively assess the quality of the results obtained by the proposed CNN-based hair removal approach, with respect to the original hairless image. We cluster these measures into three different groups. The first one are the Mean Squared Error (MSE) \[38\], the Peak Signal-to-Noise Ratio (PSNR) \[39\], the Root Mean Squared Error (RMSE) \[9\], and the Structural Similarity Index (SSIM) \[39\], which are per-pixel metrics. Within the second group we consider the Multi-Scale Structural Similarity Index (MSSSIM) \[40\] and the Universal Quality Image Index (UQI) \[37\], which measure statistical features locally and then combine them. Finally, the Visual Information Fidelity (VIF) \[31\], the PSNR-HVS-M \[14\] and PSNR-HVS \[14\], conforming the third group, have been designed to obtain more similar results to those perceived by the Human Visual System (HVS).
This set of metrics constitutes a representative selection of the state-of-the-art performance metrics for restoration quality. We must recall that largest values of PSNR, SSIM, MSSSIM, UQI, VIF, PSNR-HVS-M, and PSNR-HVS are indicators of a better quality of reconstructed images. On the other hand, lowest values of MSE and RMSE are
indicators of higher similarity.

In Table 2, we show the mean and standard deviation of the results obtained for the 185 images of the test set, and for each of the nine performance measures. In addition, we make a comparison of our results against the ones obtained by applying the six state-of-the-art hair removal methods, detailed in Section 2, to the same 185 images.

|          | Our method | Abbas | Bibiloni | Huang | Lee | Toossi | Xie |
|----------|------------|-------|----------|-------|-----|--------|-----|
| MSE      |            |       |          |       |     |        |     |
| \(\mu\)  | 27.847     | 258.347 | 103.903  | 404.366 | 175.303 | 221.346 | 55.311 |
| \(\sigma\) | 35.261     | 302.243 | 108.762  | 589.341 | 219.633 | 256.319 | 103.553 |
| SSIM     |            |       |          |       |     |        |     |
| \(\mu\)  | 0.926      | 0.867  | 0.885    | 0.851  | 0.890 | 0.864  | 0.921 |
| \(\sigma\) | 0.026      | 0.051  | 0.048    | 0.050  | 0.075 | 0.053  | 0.041 |
| PSNR     |            |       |          |       |     |        |     |
| \(\mu\)  | 35.137     | 26.080 | 29.785   | 27.001 | 29.158 | 26.570 | 33.096 |
| \(\sigma\) | 3.006      | 4.079  | 4.073    | 6.603  | 5.783 | 3.852  | 3.705 |
| RMSE     |            |       |          |       |     |        |     |
| \(\mu\)  | 4.790      | 14.226 | 9.207    | 15.485 | 11.032 | 13.287 | 6.318 |
| \(\sigma\) | 2.220      | 7.501  | 4.386    | 12.864 | 7.340 | 6.711  | 3.935 |
| VIF      |            |       |          |       |     |        |     |
| \(\mu\)  | 0.525      | 0.526  | 0.499    | 0.402  | 0.531 | 0.509  | 0.592 |
| \(\sigma\) | 0.099      | 0.180  | 0.176    | 0.130  | 0.186 | 0.178  | 0.186 |
| UQI      |            |       |          |       |     |        |     |
| \(\mu\)  | 0.997      | 0.991  | 0.996    | 0.990  | 0.995 | 0.992  | 0.997 |
| \(\sigma\) | 0.004      | 0.009  | 0.005    | 0.013  | 0.006 | 0.008  | 0.004 |
| MSSSIM   |            |       |          |       |     |        |     |
| \(\mu\)  | 0.978      | 0.870  | 0.934    | 0.917  | 0.945 | 0.875  | 0.955 |
| \(\sigma\) | 0.011      | 0.073  | 0.039    | 0.042  | 0.051 | 0.069  | 0.037 |
| PSNR-HVS-M|            |       |          |       |     |        |     |
| \(\mu\)  | 36.802     | 25.078 | 29.404   | 26.248 | 28.445 | 25.519 | 33.005 |
| \(\sigma\) | 3.703      | 3.861  | 4.245    | 6.854  | 6.247 | 3.680  | 4.106 |
| PSNR-HVS  |            |       |          |       |     |        |     |
| \(\mu\)  | 35.168     | 24.628 | 28.681   | 25.738 | 27.826 | 25.065 | 32.186 |
| \(\sigma\) | 3.465      | 3.823  | 4.158    | 6.634  | 5.991 | 3.639  | 3.896 |

Table 2: Mean and standard deviation of the similarity measures obtained to compare our method with six state-of-the-art hair removal algorithms.

The next step in our work, is to study if one algorithm outperforms another one significantly. Given a fixed similarity measure, we decided to use a statistical test to contrast the means of all pairs of algorithms. Specifically, we have used the t-test if the samples pass the Shapiro-Wilk normality test, or the Wilcoxon signed-rank test, otherwise, both considering a significance level of 0.05. According to the statistical test we can determine which method surpasses the others. Table 3 summarizes the results obtained. In it, rows represent all the pairs of algorithms in which the statistical test was applied, and the columns correspond to the measures of similarity. As an example, let us interpret the test comparing Abbas vs Huang. According to the SSIM and VIF performance measures, Abbas’ algorithm significantly outperforms Huang’s algorithm. While according to the PSNR, MSSIM, PSNR-HVS-M and PSNR-HVS measures, Huang’s algorithm significantly outperforms Abbas’ algorithm. For the rest of measures, both methods obtain statistically comparable results. Let us remark that Abbas’ algorithm is non-statistically superior in all of them.

As it can be seen in Table 3, taking into account all the considered performance measures, the proposed method out-stands according to the majority of similarity measures. It is only significantly outperformed on the VIF perfor-
Table 3: Classification of algorithms according to objective similarity measures. The results are as follows: ✓ ✓ if the population mean of the first algorithm is better than that of the second algorithm; ✓ if the mean of the first algorithm is better but statistically comparable to that of the second algorithm; ✗ if the mean of the first algorithm is worse but statistically comparable to that of the second algorithm; ✗ ✗ if the population mean of the first algorithm is worse than that of the second algorithm.

| Algorithm 1 vs. Algorithm 2 | p-value | Statistical test |
|-----------------------------|---------|-----------------|
| MSE                         | SSIM    | PSNR            | RMSE | VIF | UQI | MSSSIM | PSNR-HVS-M | PSNR-HVS |
| Our method vs. Abbas        | 2.62e-28 | 1.51e-29 | 1.49e-30 | 3.89e-29 | 0.452 | 1.09e-28 | 4.49e-32 | 1.87e-31 | 3.33e-31 |
| Our method vs. Bibiloni     | 4.44e-30 | 1.29e-25 | 3.38e-31 | 1.27e-30 | 4.07e-04 | 2.05e-10 | 2.68e-30 | 1.06e-31 | 1.49e-31 |
| Our method vs. Lee          | 1.21e-30 | 2.23e-17 | 4.66e-30 | 1.57e-30 | 0.160 | 8.52e-13 | 3.55e-27 | 1.81e-31 | 3.85e-31 |
| Our method vs. Huang        | 5.54e-30 | 2.18e-30 | 1.75e-31 | 9.41e-31 | 4.66e-23 | 1.51e-29 | 1.87e-31 | 7.19e-32 | 9.63e-32 |
| Our method vs. Toossi       | 1.49e-27 | 3.95e-29 | 5.12e-30 | 1.60e-28 | 4.90e-03 | 7.83e-27 | 5.28e-32 | 2.79e-31 | 5.83e-31 |
| Abbas vs. Bibiloni          | 1.86e-14 | 9.59e-04 | 1.14e-14 | 1.32e-14 | 2.31e-10 | 4.05e-05 | 3.07e-27 | 1.36e-25 | 1.72e-19 | 7.22e-18 |
| Abbas vs. Lee               | 5.24e-23 | 1.27e-12 | 1.40e-23 | 2.42e-23 | 0.012 | 8.04e-29 | 3.07e-27 | 1.36e-25 | 1.72e-19 | 7.22e-18 |
| Abbas vs. Huang             | 1.36e-11 | 4.70e-16 | 2.48e-15 | 7.87e-13 | 1.03e-10 | 5.69e-23 | 3.96e-26 | 1.63e-15 | 1.22e-15 |
| Abbas vs. Toossi            | 0.839    | 1.22e-03 | 5.1e-02  | 0.488  | 3.95e-21 | 0.118  | 1.49e-17 | 1.61e-03 | 1.56e-03 |
| Abbas vs. Xie               | 9.08e-28 | 3.52e-07 | 2.93e-27 | 1.27e-27 | 4.14e-32 | 4.76e-25 | 3.27e-08 | 5.66e-25 | 4.13e-25 |
| Bibiloni vs. Lee            | 4.56e-32 | 4.21e-32 | 4.64e-32 | 4.56e-32 | 4.64e-32 | 8.74e-32 | 5.64e-32 | 4.87e-32 | 4.87e-32 |
| Bibiloni vs. Huang          | 5.07e-05 | 6.25e-13 | 2.51e-02 | 3.71e-04 | 1.34e-17 | 2.05e-10 | 1.36e-15 | 2.42e-03 | 2.36e-03 |
| Bibiloni vs. Toossi         | 5.13e-13 | 9.88e-24 | 9.07e-15 | 7.50e-14 | 4.51e-17 | 1.72e-17 | 4.58e-14 | 2.30e-16 | 3.38e-16 |
| Bibiloni vs. Xie            | 7.98e-21 | 1.59e-14 | 2.37e-21 | 3.29e-21 | 0.298 | 1.56e-27 | 4.32e-27 | 5.77e-23 | 1.22e-22 |
| Lee vs. Huang               | 1.03e-18 | 4.40e-29 | 2.24e-20 | 1.66e-19 | 4.21e-32 | 5.35e-12 | 8.77e-14 | 1.74e-18 | 1.68e-19 |
| Lee vs. Toossi              | 1.24e-11 | 6.62e-17 | 1.61e-15 | 9.23e-13 | 4.76e-20 | 2.92e-21 | 1.44e-17 | 9.01e-16 | 1.92e-15 |
| Lee vs. Xie                 | 4.13e-07 | 1.75e-16 | 9.88e-12 | 7.17e-09 | 2.16e-14 | 1.32e-20 | 4.08e-26 | 4.36e-12 | 3.29e-12 |
| Huang vs. Toossi            | 1.11e-16 | 2.09e-10 | 1.15e-15 | 1.63e-16 | 6.04e-28 | 5.12e-06 | 0.869  | 2.22e-16 | 8.65e-17 |
| Huang vs. Xie               | 0.496    | 7.83e-03 | 0.133   | 0.871  | 1.70e-17 | 0.594  | 5.44e-16 | 0.050  | 5.38e-02 |
| Toossi vs. Xie              | 6.91e-25 | 7.80e-32 | 9.36e-27 | 8.70e-26 | 2.82e-28 | 2.40e-24 | 1.08e-25 | 7.17e-27 | 3.89e-27 |

Table 3: Classification of algorithms according to objective similarity measures. The results are as follows: ✓ ✓ if the population mean of the first algorithm is better than that of the second algorithm; ✓ if the mean of the first algorithm is better but statistically comparable to that of the second algorithm; ✗ if the mean of the first algorithm is worse but statistically comparable to that of the second algorithm; ✗ ✗ if the population mean of the first algorithm is worse than that of the second algorithm.
mance measure compared to Abbas et al.’ and Lee et al.’s algorithms. Among the rest of methods, we can see that Lee et al.’s algorithm surpasses statistically Huang et al.’ and Toossi et al.’s algorithms. However, when comparing Lee et al.’s algorithm with Bibiloni et al.’s, the former is statistically better in very specific settings, namely the SSIM, VIF and MSSSIM measures. It is Xie et al.’s algorithm that outperforms Lee et al.’s in all performance measures. In the comparison between the algorithms of Bibiloni with Huang and Toossi, it is the first that outperforms statistically the other two in the majority of measures. Finally, the Toossi et al.’s algorithm is statistically superior to the Abbas et al.’s, except in the SSIM and VIF performance measures. These two algorithms are the ones that provide statistically worse results compared to the rest of the algorithms.

4.5. Ablation study

Some works [21, 22] defend the fact that using skip connections, or convolutions/deconvolutions instead of pooling/unpooling layers may decrease the amount of detail loss and deteriorate the restoration performance. We study how these layers can affect the learning of our model by replacing the pooling layers with convolutions, and introducing skip-connection layers. In Table 4, we can see that the introduction of skip-connections does improve the results numerically, in terms of the similarity measures previously presented. However, these do not vary significantly when pooling layers are used instead of convolutional ones in our model. In this case, we can visualize its effects in Figure 8, where we compare the results of using or not skip-connections. As can be seen, the network is able to create a more detailed prediction with them, especially when it comes to dermoscopic structures such as streaks or globules.

| Loss with no $L_{\text{foreground}}$ term | 41.197 | 0.881 | 32.719 | 6.138 | 0.437 | 0.997 | 0.954 | 33.946 | 32.717 |
| Loss with no $L_{\text{background}}$ term | 26.398 | 0.928 | 35.306 | 4.689 | 0.521 | 0.997 | 0.979 | 36.862 | 35.209 |
| Loss with no $L_{\text{SSIM}}$ term | 31.220 | 0.926 | 34.488 | 5.119 | 0.523 | 0.997 | 0.979 | 36.802 | 35.168 |
| Loss with no $L_{\text{tv}}$ term | 30.881 | 0.926 | 34.463 | 5.113 | 0.523 | 0.997 | 0.978 | 36.744 | 35.143 |
| Loss with no $L_{\text{comp}}$ term | 30.621 | 0.923 | 34.673 | 5.039 | 0.522 | 0.997 | 0.977 | 36.599 | 34.985 |

Table 4: Mean of the similarity measures obtained on the test set for the ablation study. The second and third row correspond to the skip-connections and pooling layers’ ablation study. Then, from the fourth to the last row correspond to the ablation study of the loss terms.

Another study that we believe is of great importance is the evaluation of the relevance of each term of the loss function. As in the previous case, we show in Table 4 and Figure 7 the quantitative and qualitative results, respectively, of our model trained by removing in each case one of the terms that compose the loss function. As can be seen, most of the performance measures and resulting images are worsened by deleting some of the terms. This is not the case when
we stop computing the $L_1$ distance between the GT and the network’s prediction only among the background pixels ($L_{1}^{\text{background}}$). By comparing Figures 7c and 7e, we can see that when we do not use this term, the structures tend to be more blurrier. Such blurrier regions may not be penalized as much when calculating performance measures.

Figure 7: (a) Original hairless image, (b) Input image with simulated hair, (c) result of our model trained with the complete loss, (d)-(h) results of our model trained by removing (d) $L_{1}^{\text{foreground}}$, (e) $L_{1}^{\text{background}}$, (f) $L_{2}^{\text{composed}}$, (g) $L_{\text{SSIM}}$, and (h) $L_{\text{tv}}$. 

15
5. Discussion and Conclusions

In this work, we have presented a novel CNN-based method to the task of hair removal in dermoscopic images. We have built an encoder-decoder architecture, which has shown good results in reconstruction tasks like the one at hand. We highlight an architectural aspect of the network: the use of skip connections helps to retrieve details. The benefits of its use have been demonstrated with an ablation study. In addition, we have analyzed the performance of our method and compared it with six state-of-the-art approaches. To carry out the experiments we created a dataset using different hair simulating strategies over images from publicly available dermoscopic datasets. For the validation of the algorithms, we calculated nine measures of similarity between the hairless reference images and their corresponding image with simulated hair. Finally, we performed a statistical test to objectively study and compare their performance.

The results obtained by means of the statistical tests applied to these measures lead to the conclusion that for eight of the performance measures, our method is statistically the best algorithm. Except for the VIF measure and when we compare it with Abbas’ et al. and Lee et al.’s methods. As reflected in Figure 5 and in Table 3, Abbas’ and Toossi’s algorithms produce the least suitable results. This bad behavior may be due to the fact that these algorithms do not seem to distinguish well hairs of greater thickness or dark colors.

It is worth mentioning that we have evaluated our model in dermoscopic images with real hair, obtaining good visual results and demonstrating, thus, its effectiveness.

As future work, we aim to use our approach on a more complete skin lesion analysis system, leveraging the
knowledge to extract other characteristics. Also, increasing the number of images used in the dataset to train the network might enhance the network’s generalization capabilities.

Acknowledgments

This work was partially supported by the project TIN 2016-75404-P AEI/FEDER, UE. Lidia Talavera-Martínez also benefited from the fellowship BES-2017-081264 conceded by the Ministry of Economy, Industry and Competitiveness under a program co-financed by the European Social Fund. We thank Dr. Mohamed Attia from the Institute For Innovation and Research, Deakin University, Australia, for providing the GAN-based simulated hair images.

References

[1] European Cancer Information System. [https://ecis.jrc.ec.europa.eu/index.php](https://ecis.jrc.ec.europa.eu/index.php) Accessed on 4th April 2020.
[2] Melanoma Molecular Map Project. [http://www.mmmp.org/MMMP/welcome.mmmp](http://www.mmmp.org/MMMP/welcome.mmmp) Accessed on 4th April 2020.
[3] Abbas, Q., Celebi, M.E., García, I.F., 2011. Hair removal methods: a comparative study for dermoscopy images. Biomedical Signal Processing and Control 6, 395–404.
[4] Argenziano, G., Fabbrocini, G., Carli, P., De Giorgi, V., Sammarco, E., Delfino, M., 1998. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. Archives of dermatology 134, 1563–1570.
[5] Argenziano, G., Soyer, H., De Giorgi, V., Piccolo, D., Carli, P., Delfino, M., 2000. Interactive atlas of dermoscopy (Book and CD-ROM). EDRA Medical Publishing & New media.
[6] Argenziano, G., Soyer, H.P., Chimenti, S., Talamini, R., Corona, R., Sera, F., Binder, M., Cerroni, L., De Rosa, G., Ferrara, G., et al., 2003. Dermoscopy of pigmented skin lesions: results of a consensus meeting via the internet. Journal of the American Academy of Dermatology 48, 679–693.
[7] Attia, M., Hossny, M., Zhou, H., Nahavandi, S., Asadi, H., Yazdabadi, A., 2020. Realistic hair simulator for skin lesion images: A novel benchmarking tool. Artificial Intelligence in Medicine 108, 101933.
[8] Bakkouri, I., Afdel, K., 2020. Computer-aided diagnosis (cad) system based on multi-layer feature fusion network for skin lesion recognition in dermoscopy images. Multimedia Tools and Applications 79, 20483–20518.
[9] Barnston, A.G., 1992. Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. Weather and Forecasting 7, 699–709.
[10] Bibiloni, P., González-Hidalgo, M., Massanet, S., 2017. Skin hair removal in dermoscopic images using soft color morphology, in: Conference on Artificial Intelligence in Medicine in Europe, Springer. pp. 322–326.
[11] Chollet, F., et al., 2018. Keras: The python deep learning library. Astrophysics Source Code Library .
[12] Cui, Z., Chang, H., Shan, S., Zhong, B., Chen, X., 2014. Deep network cascade for image super-resolution, in: European Conference on Computer Vision, Springer. pp. 49–64.
[13] Dong, C., Loy, C.C., He, K., Tang, X., 2015. Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence 38, 295–307.
[14] Egiazarian, K., Astola, J., Ponomarenko, N., Lukin, V., Battisti, F., Carli, M., 2006. New full-reference quality metrics based on hvs, in: Proceedings of the Second International Workshop on Video Processing and Quality Metrics.
[15] Huang, A., Kwan, S.Y., Chang, W.Y., Liu, M.Y., Chi, M.H., Chen, G.S., 2013. A robust hair segmentation and removal approach for clinical images of skin lesions, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 3315–3318.
[16] Jain, V., Seung, S., 2009. Natural image denoising with convolutional networks, in: Advances in neural information processing systems, pp. 769–776.

[17] Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[18] Kittler, H., 2007. Dermatoscopy: introduction of a new algorithmic method based on pattern analysis for diagnosis of pigmented skin lesions. Dermatopathology: Practical & Conceptual 13, 3.

[19] Lee, T., Ng, V., Gallagher, R., Coldman, A., McLean, D., 1997. Dullrazor®, A software approach to hair removal from images. Computers in biology and medicine 27, 533–543.

[20] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88.

[21] Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B., 2018. Image inpainting for irregular holes using partial convolutions, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 85–100.

[22] Mao, X., Shen, C., Yang, Y.B., 2016. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections, in: Advances in neural information processing systems, pp. 2802–2810.

[23] Mayer, J., et al., 1997. Systematic review of the diagnostic accuracy of dermatoscopy in detecting malignant melanoma. Medical journal of Australia 167, 206–210.

[24] Menzies, S.W., Ingvar, C., Crotty, K.A., McCarthy, W.H., 1996. Frequency and morphologic characteristics of invasive melanomas lacking specific surface microscopic features. Archives of Dermatology 132, 1178–1182.

[25] Menzies, S.W., Ingvar, C., Crotty, K.A., McCarthy, W.H., 1996. Frequency and morphologic characteristics of invasive melanomas lacking specific surface microscopic features. Archives of Dermatology 132, 1178–1182.

[26] Mirzaalian, H., Lee, T.K., Hamarneh, G., 2014. Hair enhancement in dermoscopic images using dual-channel quaternion tubularness filters and MRF-based multilabel optimization. IEEE Transactions on Image Processing 23, 5486–5496.

[27] Mirzaalian, H., Lee, T.K., Hamarneh, G., 2014. Hair enhancement in dermoscopic images using dual-channel quaternion tubularness filters and MRF-based multilabel optimization. IEEE Transactions on Image Processing 23, 5486–5496.

[28] O’Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G.V., Krpalkova, L., Riordan, D., Walsh, J., 2019. Deep learning vs. traditional computer vision, in: Science and Information Conference, Springer. pp. 128–144.

[29] O’ Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G.V., Krpalkova, L., Riordan, D., Walsh, J., 2019. Deep learning vs. traditional computer vision, in: Science and Information Conference, Springer. pp. 128–144.

[30] Poess, J.-A.A., Ruiz, C., 2018. Using deep learning to detect melanoma in dermoscopy images. International Journal of Machine Learning and Computing 8.

[31] Poess, J.-A.A., Ruiz, C., 2018. Using deep learning to detect melanoma in dermoscopy images. International Journal of Machine Learning and Computing 8.

[32] Sheikh, H.R., Bovik, A.C., 2004. Image information and visual quality, in: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, pp. iii–709.

[33] Sheikh, H.R., Bovik, A.C., 2004. Image information and visual quality, in: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, pp. iii–709.

[34] Sheik, H.R., Bovik, A.C., 2004. Image information and visual quality, in: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, pp. iii–709.

[35] Stolz, W., 1994. ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma. Eur. J. Dermatol. 4, 521–527.

[36] Talavera-Martínez, L., Bibiloni, P., González-Hidalgo, M., 2019. Comparative study of dermoscopic hair removal methods, in: ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing, Springer. pp. 12–21.

[37] Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., Lin, C.W., 2019. Deep learning on image denoising: An overview. arXiv preprint arXiv:1912.13171.

[38] Toossi, M.T.B., Pourreza, H.R., Zare, H., Sigari, M.H., Layegh, P., Azimi, A., 2013. An effective hair removal algorithm for dermoscopy images. Skin Research and Technology 19, 230–235.

[39] Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A., 2008. Extracting and composing robust features with denoising autoencoders, in: Proceedings of the 25th international conference on Machine learning, pp. 1096–1103.

[40] Wang, Z., Bovik, A.C., 2002. A universal image quality index. IEEE signal processing letters 9, 81–84.

[41] Wang, Z., Bovik, A.C., 2002. A universal image quality index. IEEE signal processing letters 9, 81–84.
Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al., 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13, 600–612.

Wang, Z., Simoncelli, E.P., Bovik, A.C., 2003. Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. Ieee. pp. 1398–1402.

Xie, F.Y., Qin, S.Y., Jiang, Z.G., Meng, R.S., 2009. PDE-based unsupervised repair of hair-occluded information in dermoscopy images of melanoma. Computerized Medical Imaging and Graphics 33, 275–282.

Xie, J., Xu, L., Chen, E., 2012. Image denoising and inpainting with deep neural networks, in: Advances in neural information processing systems, pp. 341–349.

Zhao, H., Gallo, O., Frosio, I., Kautz, J., 2016. Loss functions for image restoration with neural networks. IEEE Transactions on computational imaging 3, 47–57.