Multiscale skin analysis and parametrization

Multiscale skin analysis and parametrization

Alessio Gallucci

MULTISCALE SKIN ANALYSIS AND PARAMETRIZATION

Alessio Gallucci

Alessio Gallucci
Eindhoven University of Technology
Department of Mathematics and Computer Science

E-mail: alessio.gallucci@gmail.com
Printed by Gildeprint, The Netherlands.
A catalogue record is available from the Eindhoven University of Technology Library.
ISBN: 978-94-6419-595-8

The cover was made by nskvsky.

MULTISCALE SKIN ANALYSIS AND PARAMETRIZATION

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven,
op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens,
voor een commissie aangewezen door het College voor Promoties, in het
openbaar te verdedigen op vrijdag 7 oktober 2022 om 11:00 uur

door

Alessio Gallucci

geboren te Moncalieri, Italië

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof. dr. Edwin van den Heuvel
1e promotor:	prof. dr. Milan Petkovic
copromotor(en):	dr. Dmitry Znamenskiy (Philips Research)
leden:	prof. dr. Josien Pluim
	prof. dr. ir. Peter H.N. de With
	prof. dr. Elmar Eisemann (Technische Universiteit Delft)
	prof. Alberto Signoroni (University of Brescia)
adviseur(s):	dr. Nicola Pezzotti

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

CONTENTS

1. INTRODUCTION

The *skin* is the largest organ of the human body and performs essential biological functions such as protection, regulation, and sensation. It protects from UV rays and harnesses nutrients from the sun while also receiving its heat. The outer layer of the skin, shown in Figure 1.1, is called the *epidermis* and, together with skin *hair* and *lesions*, is the focus of this dissertation. Since the epidermis is the external part of the body, we can easily collect data such as photos and scans from it by means of external devices such as cameras and scanners. This raw skin data can then be analyzed by looking at the *local* distribution of skin artefacts to deduce information and relationships that can guide the decision-making process and facilitate the work of, for example, a dermatologist. On the other hand, skin data can be used to parametrize the *global* geometry and texture of the skin surface. Parametrization and analysis can enable other downstream applications such as body part surface estimation, body size and shape estimation, and animation of virtual characters.

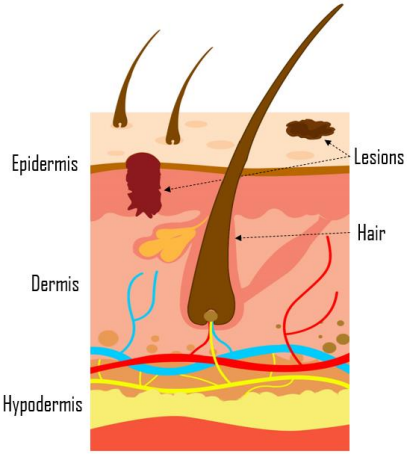


Figure 1.1. Example of the human skin. It is composed of three layers called epidermis, dermis, and hypodermis. Skin artefacts like hair and lesions are often visible by naked human eye in the epidermis.

Historically, these applications were often driven by human labor and expertise, for instance, body measures taken by a tailor for custom garments or doctors inspecting a suspicious skin lesion. Nowadays, the advancement in *Artificial Intelligence* and *Computer Vision* enables telemedicine applications where human observations are complemented and enhanced by technologies such as deep *Artificial neural networks* (ANNs) – a digital learning system inspired by the way biological neurons function and communicate. Convolutional neural networks (CNNs) – a popular version of ANNs for processing images – were first introduced in 1989 by Yan LeCun to recognize handwritten digits collected from the U.S. postal service [1]. In 2012, CNNs reached human and super-human performance in various vision tasks [2] such as classifying and detecting objects in natural images. Recently, in 2019, CNNs reached dermatologists' performance in classifying images of suspicious skin lesions [3]. Similarly, the rise of modern 3D scanning technologies allowed the accurate acquisition of the skin surface over the complete human body used to build high resolution data-driven models of the human body and face outer shape [4], [5]. Thanks to these new advances it is now possible to automate and enrich the collection, analysis, and parametrization of the skin reducing the need for repetitive and tedious human work while enabling new technologies. This can potentially improve millions of people's lives by impacting various industries including the medical and personal care ones.

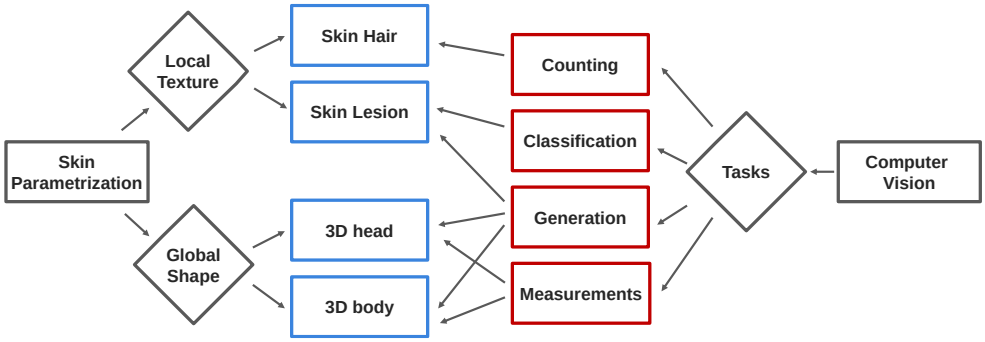


Figure 1.2 presents the concepts explored in this work of research, mapping the computer vision research field to the parametrization and analysis of the skin. From left to center, the skin is analyzed and parametrized as local texture when considering relatively small 2D skin patched, or global shape, when considering the appearance and distribution of a whole 3D body part. From right to center, the computer vision is divided into possible tasks, which are mapped and applied to the relative skin surface or patch.

Consider applications such as shaver trimming the complex beard style, a surgical arm performing a skin cut or a photo epilation device removing skin hairs on a body part. In the above examples, the smart functioning of the device requires its localization while displacing it over body parts and skin spots. When addressing the skin localization, one can first consider global body shape parameters, while the features classification and recognition would be carried on by models operating mostly on local skin texture. A common framework to describe this problem is the Simultaneous Localization and Mapping (SLAM) of a device navigating an unseen environment [6]. This typical scenario involves a robot that, deployed in a new environment, needs to figure out his relative position with respect to the map which also needs to be constructed by him while exploring the environment (hence called simultaneous). While not directly covered in this work, we consider of interests SLAM since it is a natural way to combine the different scales and tasks. Similarly, shape and texture can be combined in the same model to analyze how particular skin artefacts, like skin lesions, evolve by mapping, localizing, and retrieving them [7]. However, SLAM applications are only an ideal end goal. The current research presented in this work focuses more on downstream applications, such as detecting local skin features like hair and lesions or global analysis of skin surface areas and body measurements. In the following section, we introduce the relevant background and challenges in computer vision applied to the skin.

1.1. Background and Challenges

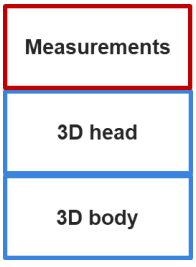
The localization of a device operating on top of the skin mentioned in the previous section is one major motivation underlying the need to capture multiscale skin variations – skin data (signals, photos, scans, …) under various settings (light variations, perspective distortions, resolution, …). This skin data is used to explore the recent developments in deep learning and shape modeling to advance the analysis and parametrization of the outer body skin. Figure 1.2 maps the main areas touched by this work by applying *computer vision tasks* (right) to *skin shape and*

Figure 1.3. Sample skin patch from the author left hand in different locations and dates.

texture (left). The main computer vision tasks considered are counting, classification, generation and measurements. We applied counting, classification and generation to local skin patches presenting skin hairs and skin lesions. The underlying data are images when referring to local skin patches without incorporating the skin surface's geometry and shape. Instead, we consider methods for estimating measurements and generative models on the shape and global geometry of the face and full body. In the following sections, we introduce the four primary computer vision archetypal tasks: *measurements*, *counting*, *classification*, and *generation*.

While presenting the tasks, we highlight some common challenges in applying the computer vision tasks to the skin domain. Skin images statistics and distribution are different from natural images. Their textures offer challenges with the 3D structure which reflect light in many directions causing difficulties in various areas of computer vision: it is more challenging for a camera to auto-focus or for an AI algorithm to collect skin features. An example of skin patches collected from the left hand of the author is shown in Figure 1.3: skin appearance varies significantly by collecting images of the same patch in different dates, locations, or camera orientations. The patches show some parts of the images that are out of focus where one can see hair blurred is some versions of the patch. Similarly, the light conditions affect the skin color, and the perspective distortions affect the scale of the skin features of the skin patch. In the following sections, we will introduce all computer vision tasks using the same elements and blocks presented in Figure 1.2. We will start with the global skin shape and its measurements and then local skin patches analysis using the counting, classification, and generation tasks.

MEASUREMENTS



In several applications it is often required to estimate and parametrize skin surfaces. For example, this is motivated by the fact that collecting body shapes and size measurements is often a complex task and prone to errors. It is challenging to define a proper set of instructions to follow by a person trying to measure body size accurately. While it can be easy to collect a person's height with accuracy already, other measurements like the torso circumference could present obstacles – what is the torso, and

Figure 1.4 shows in the top row present 3 raw scans acquired with a 3D scanner. The bottom row shows a common template mesh registered to the raw scans.

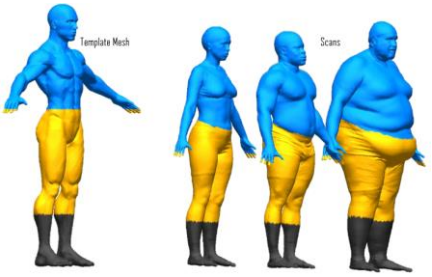


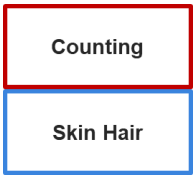
Figure 1.5. Example of the template mesh with a texture pattern on the left. On the right, three registered scans with the texture propagated from the template.

what anatomical reference can we use to define the height of the circumference? What if we want to measure more complex areas such as the calf area to estimate the total number of skin hair or the dosage of a particular local anesthetic?

One possible solution lies in inferring complex sizes from easy ones by building a statistical model. One can, for example, regress the torso circumference from age, weight, and height input variables. Wang et al. [8] find the optimal formula to regress Body Surface Area from weight and height. However, such simple models are difficult to generalize to the surface areas of particular body parts or “shape measurements”, when a specific (medical) product is interfacing with the body and therefore the product shape should be complementary to the body shape. Shape modeling is a methodology aimed at converting a group of similar instances into a common representation to have a better way of computing shape statistics and understanding the variability and characteristic of the instances – which, in this case, are 3D heads and bodies. A general procedure consists of first collecting a raw representation of the instances. The raw representation can be an image, a 3D scan or any other signal representing the instance shape and position in space. Then, a single template mesh geometry is used to represent and register the raw scans reaching a more manageable and common representation. The raw signal may contain artifacts or missing data such as spikes and holes or other imperfections which are not native to the instances but are usually caused by the acquisition method, obstructions or poses as shown in the top row of Figure 1.4. A linear space of bodies’ shapes is obtained by registration of a common template mesh to all subjects such that each template vertex is registered to a specific location on the body therefore creating a meaningful correspondence across all subjects. We can now, for example, apply a pattern of colors in the template and propagate it instantly to all instances – an example is shown in Figure 1.5 where a three colored texture pattern is propagated from the template on the left to the three scans on the right. Next, a relevant path to follow is to encode all vertex positions of registered template in a few parameters by reducing the dimensionality of the linear space using techniques like Principal Component Analysis. With the parameters, we will then be able to control meaningful measurements like height and circumference of the body.

The are many challenges in modeling the distribution of shapes representing the same instance, e.g., skin shape of a human. When the input class shape is complex the challenges rank up exponentially. Similarly, for our own case of human's skin surface areas' measurements, body regions like hands and finger, nostrils, or ears require particular care. Another important aspect of consideration is the difficulty in acquiring high resolution data. Acquiring complex and large shapes may be costly and not always possible, imagine the case of a patient lying in bed. Hence a common research issue is to build proper datasets or make use of the power of skin statistical shape modelling without the need to collect expensive 3D scans. Another related complexity is the naturally soft and flexible tissue which deform and bend – different body poses result in different body shapes for the same subject identity. It is often the case that one wants a measurement that is not dependent on body pose. Similarly, another relevant task is identifying a subject no matter the body pose and its tissue deformation.

COUNTING



The main objective of the counting task is to estimate the total number of an instance or object template in an image. Various real-world applications require robust automated counting schemes: estimating the number of people in crowds for surveillance and safety reasons [9], counting cells in a skin patches or biopsies for supporting medical diagnosis and detecting animals in their natural habitat for wildlife preservation [10].

Note that contrary to the measurements task or the generation task, in the counting task one does not need to model the full instance of the template, but it might be enough to recognize a single discriminating feature. In fact, counting requires mapping an input image to the number of instances, and this can be done without learning the full ranges or distributions of the instances. When designing a people's counter scheme, one may build an algorithm that recognizes only one identifying feature (for example the face proportions or a body part detector) while not understanding what a "person" in the whole means. This can result in a perfect people counter while not being able to recognize the full body but only a body part. Similarly, a classifier can discriminate people versus animals just by looking at their limps' geometry.

Counting is often coupled with other features detection algorithms: in crowd estimation, for example, often shape models are used as a building block for the detection and tracking of bodies. In the past, general algorithms have been implemented to detect any object instance robustly when presented in different orientation or resolution. They learn to recognize unique features at different scales like the seminal Scale Invariant Feature Transform [11], [12].

However, the instances' appearance can hide many challenges: they can be partially hidden or confused in the background, intersect other instances, present complex color variations depending on the environment or, by nature, be represented by a multitude of colors. For example, it might not always be easy to spot whether two or many bodies intersect making the feature detection prone to errors and counting task

challenging. In our work, we count instances of skin hair that, while intuitively simple to detect since low-degree splines can approximate them, they present challenges due to their lack of distinct and unique features. Even when the skin patch picture presents good settings, like in the example of Figure 1.6. where there are little variations in light conditions and perspective distortions, distinguishing and counting hair in crowded skin patches is complex for humans as well for traditional computer vision approaches.

Figure 1.6. Example of a crowded skin patch where it is hard to discriminate and count hair despite the good light conditions and the absence of perspective distortion.

CLASSIFICATION

In computer vision, classification has the objective of producing a label representing the belonging of an object to a group, class, or label. A classifier is a function that maps an object to a label. For example, a melanoma classifier will produce a score representing the probability of a melanoma's presence in the skin image or an email spam detector will produce a (binary) label categorizing the email into spam or inbox. Usually, classification tasks in computer vision are driven by a dataset in which we have images coupled with ground truth labels. The labels are used to steer the AI algorithm by tweaking its parameters according to the known label. A common algorithm in neural networks to update the parameters is the backpropagation, also called backprop [13].

The classification is non-binary when multiple classes are possible. An example presented in this thesis is the International Skin Imaging Collaboration (ISIC) Classification Challenge Dataset [14]. The input images contain a skin lesion as the main subject and a label representing the associated diagnosis, as presented in Figure 1.7. Classic classification datasets usually address common objects in natural images. Medical and personal care datasets of images often present additional hurdles. A major one is the lack of big data sets, since collecting samples from patients is more difficult than natural images that can be for example downloaded from social networks or the web. In fact, data connected to the skin might be personal

Figure 1.7. Examples of skin lesions and their labels from the ISIC 2020 challenge. From left to rights examples of lesions and their ground truth labels.

or connected to health sensitive data. The pictures might present unique identifying features like tattoos or birthmarks while 3D faces are by nature connected to the subject's identity. Another related hurdle is the underrepresentation or unbalances of the dataset: diseases or lesions are often underrepresented due to their low prevalence or importance in the healthcare community, or, due to biases such as skin color.

A possible solution to the imbalance of the dataset could be to enrich the former with realistic augmentations or to create synthetic samples for the classification task. However, it is crucial to understand which augmentations are possible and whether the presence of noise in the image reduces the performance of the classifier. Classification algorithms are also important for other computer vision tasks. For example, in the generation task presented in the following section, a classification algorithm may be crucial to benchmark its performances in the generation of synthetic instances.

GENERATION

Generative models are used for various exciting applications including super-resolution [15], neural style transfer [16], audio-video facial synthesis [17], and text to speech synthesis [18]. The common aim is to generate novel data that are also statistically indistinguishable from the original input dataset to mimic. It is usually more challenging than counting and classification also from an intuitive human perspective as introduced in the previous section on counting. In fact, it is not enough or interesting if the synthesis replicates the input data since the novelty and usefulness are missing. Hence a requirement is that generated samples are "sufficiently" different from the originals. The opposite also must hold that the new sample, while different from originals, must also belong to the same instance class. This, being difficult to define also heuristically, makes it always challenging to measure appropriately with a quantitative measure. Several research works have been proposing and analyzing different metrics for generative models on images and 3D data while not yet producing a definitive and stable one. In 2015 an analysis of several metrics is presented in Theis et al. [19], outlaying the need to evaluate generative models on the intended downstream applications. For example, when creating novel skin lesions to augment a dataset used for classification, one might want to evaluate the classifier performance using the synthetic data (e.g., in the application the data was created for) rather than evaluating the new data quality itself directly. These new ideas are captured in the Classification Accuracy Score metric presented in Suman et al. [20], which aimed at improving common and popular generative metrics such as the Inception Score [21] or Fréchet Inception Distance [22].

Particular attention must be taken in preprocessing the data and selecting the appropriate generative when considering the skin domain. Apart from the similar challenges presented in counting and classification here, the problem is usually amplified by the nature of the complex task. For example, while counting under

different lighting conditions may be difficult, the generation may fail without first normalizing the different conditions.

1.2. RESEARCH QUESTIONS AND CONTRIBUTIONS

In this work, we consider the use of *3D human body and face shape models* to enable and enhance skin measurement applications. Skin shape modeling is relevant for several fields such as healthcare, cognitive science [23], [24], online shopping [25], [26], clothing [27] and virtual reality [28], [29]. For example in healthcare, the knowledge of 3D body shape measurements can help in the assessment of the Psoriasis Area and Severity Index (PASI) [30], dosing chemotherapy according to the Body Surface Area (BSA) [31] or estimating a burned body part [32]. However, many applications of such technology lack precision in estimation [33]–[35] and accurate body shape prediction would help the dosage of a particular drug. The prediction of (less accurate) 3D models from anthropometric measurements – body measures which are easy to collect and not invasive like age, gender, waist circumference, height etc. – can be considered as a lower cost alternative to full body 3D scanning and processing involving the recognition and processing of different body parts.

One of the obstacles in the efficient processing of full body 3D models is the high volume of data. The cost and volume of the data required can be drastically reduced by learning a statistical representation of the human shape space. Only sparse data, combined with the learned space, are needed to reconstruct a full body scan instead of a dense representation. Hence the first research question we considered is

> How accurately can one estimate skin shape
> from anthropometric measurements?

and our contribution to define a methodology to systematically answer such questions for any body part. The foundation for our research is the seminal work of A**.** Blanz and Vetter [4]. They defined how to learn a statistical shape space of the face and then used measurements and semantic descriptors to modify the face appearance. Successively, Allan *et al.* [5], replicated the work for full bodies. Many other works improved such models by increasing the number of parameters in the model, including dynamic models and learning soft tissue deformation. We follow current state-of-the-art to build two parametric models for a full body and 3D heads considering a dataset containing more than 4000 high-resolution scans from various nationalities [36], [37]. Then we construct predictive statistical models to retrieve estimations of skin surfaces given a set of anthropometric measurements. The lowest full body surface estimation reaches 13mm when the input are 12 standard measurements including age, gender, height and weight and face shape.

As the main application, we used this methodology to estimate the number of laser flashes when epilating with a Laser Hair Removal (LHR) device (see Figure 1.8) which depends on the surface area of a body part. The LHR technological base has

been built by Anderson and Parrish's principle of selective photo thermolysis [38], [39]. In [40], Grossman et al., were the first to use it for photo-epilation and, since its approval in the Food and Drug Administration (FDA), has been growing its popularity due to its safe, fast and effective use for hair removal [41], [42].

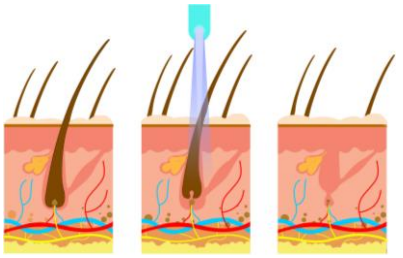


Figure 1.8. Laser Hair Removal Process from left to right. The hair bulb is burnt with light emitted from the personal care device.

Body hair is one of the main features of the skin. Their function is to facilitate heat regulation and protection against rash, but nowadays, people are much more interested in removing them for beauty reasons. Many commercial solutions to the hair removal problem are available. However, the use of LHR is not restricted to qualified personnel, e.g., a dermatologist, and anyone can benefit from it in the comfort of a private home. Our goal is to make hair removal devices, like shavers or LHR, more efficient by automating and enhancing them with smart algorithms. The effectiveness of the LHR device needs to be proven for many reasons such as the FDA approval of new model versions, getting market shares or improving an older model. Classical experimental variables range from device settings like light temperature, device orientation, type of device heads to subject conditions like skin color, skin area surface and user engagement. A primary evaluation metric common to many different experiments, is the number of hairs removed by the device. The hair counting is usually done by visual inspection and visual hair count, which is a long and tedious manual work. Hence, we narrow down the research question:

Can a computer vision based automatic hair counter replace the human annotator?

Deep learning is state-of-the-art for object counting in various applications such as agriculture [43]; microbiology [44], [45]; security [46], [47]; and wild life conservation [48]. Concerning healthcare object counting is fundamental in microbiology when considering the count of total number of cells [44], [45]. However, most of the previous literature on hair count does not rely on machine learning [49] [50], [51] [52] and, while being good in a controlled image acquisition environment, they suffer from different light conditions, perspective distortions and a slight change in the acquiring device implies the need for re-tuning the numerous parameters of the computer vision algorithm.

Hence, our analysis focuses on *counting* hair in an end-to-end fashion to streamline and reduce the costs of the numerous studies carried on the personal care device since it is motivated by the lack of fully automated systems to count skin hair. The contribution of our work is as follows: we collected a dataset of skin patches, including more than 4000 images from more than 100 volunteers; propose to adopt a deep learning based automated hair counting algorithm, and; we investigate several deep learning architectures to perform end-to-end hair counting. Finally, we show that the adaptation of a segmentation network outperforms other architectures and shows the emergence of a segmentation behavior by only regressing the total

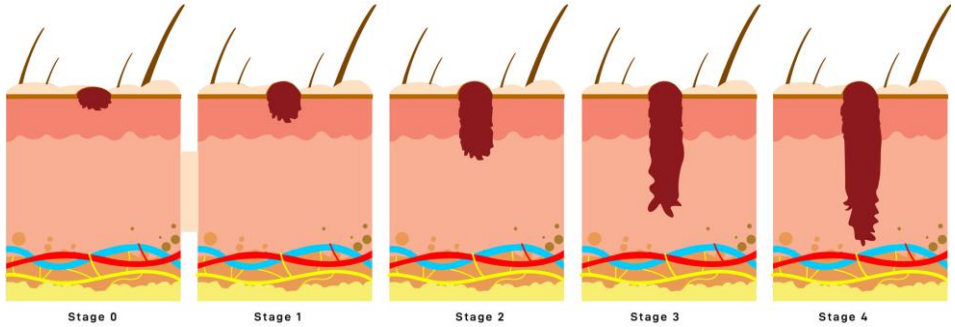


Figure 1.9. Example of Melanoma evolution over time from left to right. Early diagnosis is important at early stages to prevent the spread of it to other regions of the body and provide the right and timely treatment options

hair count. Our best end-to-end network can count skin hair with relatively low error of ~6 hair out of the ~55 in the images, hence, the expert human annotator is still to be preferred. Nevertheless, the automatic method is sufficient for the automatic evaluation of the LHR.

Building on top of the learned knowledge in the skin hair domain, we consider how the presence of skin hair may affect the diagnosis of *skin lesions*. There are various lesion types, and many recent research efforts try to collect huge datasets and classify them with Deep Learning. Comprehensive microscopic databases include around 30 thousand samples, limiting the richness of patterns that can be presented to machine learning. An example is the International Skin Imaging Collaboration (ISIC) challenge where the number of images submitted to the participants has increased steadily over the past 5 years. Their goal is to reduce mortality for skin cancer, see Figure 1.9, since it affects more than three million people only in the US. Usually, the screening and diagnosis of melanomas are primarily carried by clinical visual inspection and biopsy if necessary. However, there is an urge to deploy such results in the real-world scale to automate and facilitate this procedure with image-based screening [53] to lessen the burden of Dermatologists and satisfy the unmet request for screening of younger population. In fact, the prevalence of this type of cancer in the population is the highest compared to other types. To support an early diagnosis, effective and computationally efficient models can be deployed on smartphone devices to enable a first level of patient-driven screening. In this setting, the models have to deal with much less controlled conditions than in a laboratory environment. One objective is to understand and provide suggestion on whether to shave or not prior to acquiring skin pictures. Hence, we investigate the effect that a varying amount of skin hair have on the classification accuracy of deep learning classifiers and define the following research question:

Does and how hair affect skin lesion diagnosis in deep learning classifiers?

Our contribution is twofold: we first present an approach for the segmentation of hairs in existing skin image patches and second, we analyze how the segmented hair

influence deep learning skin lesion classifiers. The segmentation approach gives us realistic hair patterns that test for the robustness of skin lesions classifiers. Our second contribution is a set of image augmentation strategies, based on our first contribution, that form a testing pipeline for the robustness of skin lesions classifiers on skin hair presence. The results show that the presence of hair in skin images has little effect on the prediction of skin lesions. Quantitatively, the trained neural network with and without hair presence reached the same accuracy when classifying skin cancer versus nevi.

One clear challenge to achieve great prediction accuracy for the research is the lack of big data since collecting real samples is time-consuming and difficult, for example considering rare diseases. Another challenge is the fairness of the AI systems concerning underrepresented populations. Hence, we address the following connected research question

Can we model the distribution of skin lesion images and generate realistic looking synthetic examples with deep learning?

Another prominent area of deep learning research investigates the use of *generative models*. In the past years, we witnessed fast and significant improvement of generative models. The growth in popularity is related to the visually appealing results and the continually increasing computational power commonly available. The state-of-the-art models are called Generative Adversarial Networks (GANs) [54], and they rely on two competing architectures, one generating images starting from noise, the other trying to distinguish synthetic from real samples. The synthetic samples are difficult to spot compared to the original data even from a careful human inspector [55]. Recently, a different approach has proven similar results to GANs on image generation tasks. This approach use an autoencoder Vector Quantized Variational Autoencoder (VQ-VAE-2) [56] in combination with an autoregressive model called PixelSNAIL [57]. Key advantages of this approach are the explainability of the features and the potential for integrating advanced augmentation techniques.

Our contribution is to apply this novel approach in the field of skin lesions generation to augment and increase the number of images in the lesions' datasets. This offers benefits like the possibility to directly modify a certain local part of an input lesion without affecting its global structure. The quantitative results are promising but still show that the synthetic data is not good enough to improve on downstream tasks like classifying the diagnosis of a skin lesion.

We have considered various powerful Deep learning Models to analyze and synthetize skin texture. However, returning to our starting point, we are also interested in the synthetic full 3D skin. In fact, one major research challenge is the lack of high-resolution datasets of 3D scans due to the cost of acquisition and the lack of freely available big databases. Therefore, we wonder how to generate 3D realistic scans. The current state-of-the-art offers various directions to tackle such complex problems. Traditionally the generation of random 3D meshes is achieved by sampling using PCA decomposition. In fact, it is sufficient to sample the PCA scores as independent random variables assuming the orthogonality of the PCA

basis vectors. While the PCA analysis assumes that the marginal coefficient distributions are close to Gaussian, it is more reasonable to follow a data-driven approach and sample from the empirical distributions of each coefficient. While this approach is easy to implement and fast not all combinations of PCA scores result in a natural human shape. Hence, we tackle the following research question

> Can we generate realistic 3D skin shape
> exploiting the power of deep generative models?

Modern approaches rely on generative methods which could work with original high-dimensional vertex data. The current state-of-the-art advances in the field of geometric deep learning [58] try to use the power of CNNs by adapting them to work on meshes [59], [60]. Graph convolutions, however, restrict the resolution, and therefore, the accuracy of the 3D template.

Another prominent research direction is the idea of storing the 3D information into a 2D regular representation. One common definition of such idea uses UV maps, and provides bijective mapping from the 3D mesh triangles to their images on the texture image. For example, the UV maps for the face model can be created by warping of the 3D templates with a regular grid of the facial surface, see for example Booth *et al.* [61] for a list of possible optimal implementations. The body unwrapping is more complex and often requires tailored solution including multiple mappings for different body parts.

In our work, we borrow this alternative approach. We first consider a simpler case of the human facial surface using a 3D template with a CNN-friendly mesh and then a general case of the complete human body that cannot be naturally unwrapped into a 2D. Once we have 2D representations of 3D template, we can apply CNNs to generate new 3D shapes. We then rely on the same technique used for skin lesion generation applied to address the previous research question. Our main contributions are: (i) a new non-bijective way of creating the 2D representation of 3D template by using multi-view projections, and (ii) the generation of realistic high-resolution 3D scans by reducing the problem to 2D representations. Moreover, we demonstrate that our approach outperforms PCA-based sampling via quantitative and qualitative analysis of the synthetic scans by demonstrating that the synthetic scans are closer to the empirical distributions of the real test scans.

1.3. THESIS OUTLINE

In summary, the dissertation applies and innovate computer vision algorithms to parametrize global skin shape and local skin texture. It is organized as follows:

- In Chapter 2 we build 3D parametric body models and infer surface areas given anthropometric measurements like arm length, waist circumference, or age. The statistical models show that we can infer large surfaces (e.g., upper body one) with less than 1cm error with relatively few measurements.
- As shown in Chapter 3, artificial neural networks are used to automatically count skin hair in tiny skin patches. While the count is not optimal – human

annotator is still to be preferred – the average error of 6 skin hair is sufficient for many applications.
- In the following Chapter 4, we investigate how skin hair influences the prediction of skin lesions diagnosis made by deep learning classifiers. We show that epilation prior to image acquisition is unnecessary since hair presence does not hamper lesion classification performances.
- Then, in Chapter 5, we exploit recent advances in generative models using variational autoencoders to model the distribution of skin lesion images and generate synthetic ones.
- In Chapter 6 and 7, we move back to 3D shapes by merging the two computer vision field to reach high resolution, deep generative shape models, for faces and bodies.
- Finally, in Chapter 8 we discuss future research directions and conclude the body of the dissertation by elaborating on the results achieved.

1.4. RESEARCH OUTPUT

Our research led to the following publications:

1. Gallucci, A., Znamenskiy, D. & Petkovic, M. Prediction of 3D Body Parts from Face Shape and Anthropometric Measurements. *J. Image Graph.* 8, (2020).
2. Gallucci, A., Znamenskiy, D., Pezzotti, N. & Petkovic, M. Hair counting with deep learning. in *2020 International Conference on Biomedical Innovations and Applications (BIA)* 5–9 (2020).
3. Gallucci, A., Znamenskiy, D., Pezzotti, N. & Petkovic, M. Don't Tear Your Hair Out: Analysis of the Impact of Skin Hair on the Diagnosis of Microscopic Skin Lesions. in *Artificial Intelligence for Healthcare Applications. Lecture Notes in Computer Science Series (LNCS)* (2021).
4. Gallucci, A., Pezzotti, N., Znamenskiy, D. & Petkovic, M. A latent space exploration for microscopic skin lesion augmentations with VQ-VAE-2 and PixelSNAIL. in *SPIE Medical Imaging Proceedings* (2021).
5. Gallucci, A., Znamenskiy, D., Pezzotti, N. & Petkovic, M. Generating High-Resolution 3D Faces Using VQ-VAE-2 with PixelSNAIL Networks. in *International Conference on Image Analysis and Processing* 228–239 (2022).
6. Gallucci, A., Znamenskiy, D., Long, Y., Pezzotti, N. & Petkovic, M. Generating high-resolution 3D faces and bodies using VQ-VAE-2 with PixelSNAIL networks on 2D representations. Journal of MDPI Sensors Special Issue on Computer Vision in Human Analysis: From Face and Body to Clothes (currently under submission).

Other publications not directly presented in the current work are:

1. Valev, H., Gallucci, A., Leufkens, T., Westerink, J. & Sas, C. Applying Delaunay Triangulation Augmentation for Deep Learning Facial Expression Generation and Recognition. in *Pattern Recognition. ICPR International*

Workshops and Challenges 730–740 (Springer International Publishing, 2021).
2. Zicari, R. V *et al.* Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier. *Front. Hum. Dyn.* **3**, 40 (2021).
3. Zicari, R. V *et al.* On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. *Front. Hum. Dyn.* 3, 30 (2021).
4. Zicari, R. V *et al.* How to Assess Trustworthy AI in Practice. (2022) doi:10.48550/ARXIV.2206.09887.
5. Allahabadi, H. *et al.* Assessing Trustworthy AI in times of COVID-19. Deep Learning for predicting a multi-regional score conveying the degree of lung compromise in COVID-19 patients. *IEEE Trans. Technol. Soc.* (2022).

Another patent application not directly presented in the current work is:

1. Znamenskiy, D., Heinrich, A., Gallucci, A., Ciuhu, C., Zeitouny, M., Kooijman, G.. Estimating A Surface Area And/or Volume Of A Body Or A Body Part Of A Subject (Issue WO 2020/234339 A1; Issue EP 3742397 A1).

2. SKIN PARAMETRIZATION AND PREDICTION

“When we are no longer able to change a situation,
we are challenged to change ourselves.”
— Viktor E. Frankl, Man's Search for Meaning

While 3D body models have been vastly studied in the last decade, acquiring accurate models from the sparse information about the subject and few computational resources is still a main open challenge. In this chapter, we propose a methodology for finding the most relevant anthropometric measurements and facial shape features for the prediction of the shape of an arbitrary segmented body part. For the evaluation, we selected 12 features that are easy to obtain or measure including age, gender, weight and height; and augmented them with shape parameters extracted from 3D facial scans. For each subset of features, with and without facial parameters, we predicted the shape of 5 segmented body parts using linear and non-linear regression models. The results show that the modeling approach is effective and giving sub cm reconstruction accuracy. Moreover, adding face shape features always significantly improves the prediction.

This chapter is based on the paper *Prediction of 3D Body Parts from Face Shape and Anthropometric Measurements* authored jointly with Dmitry Znamenskiy, and Milan Petkovic, which was published in Journal of Image and Graphics in September 2020.

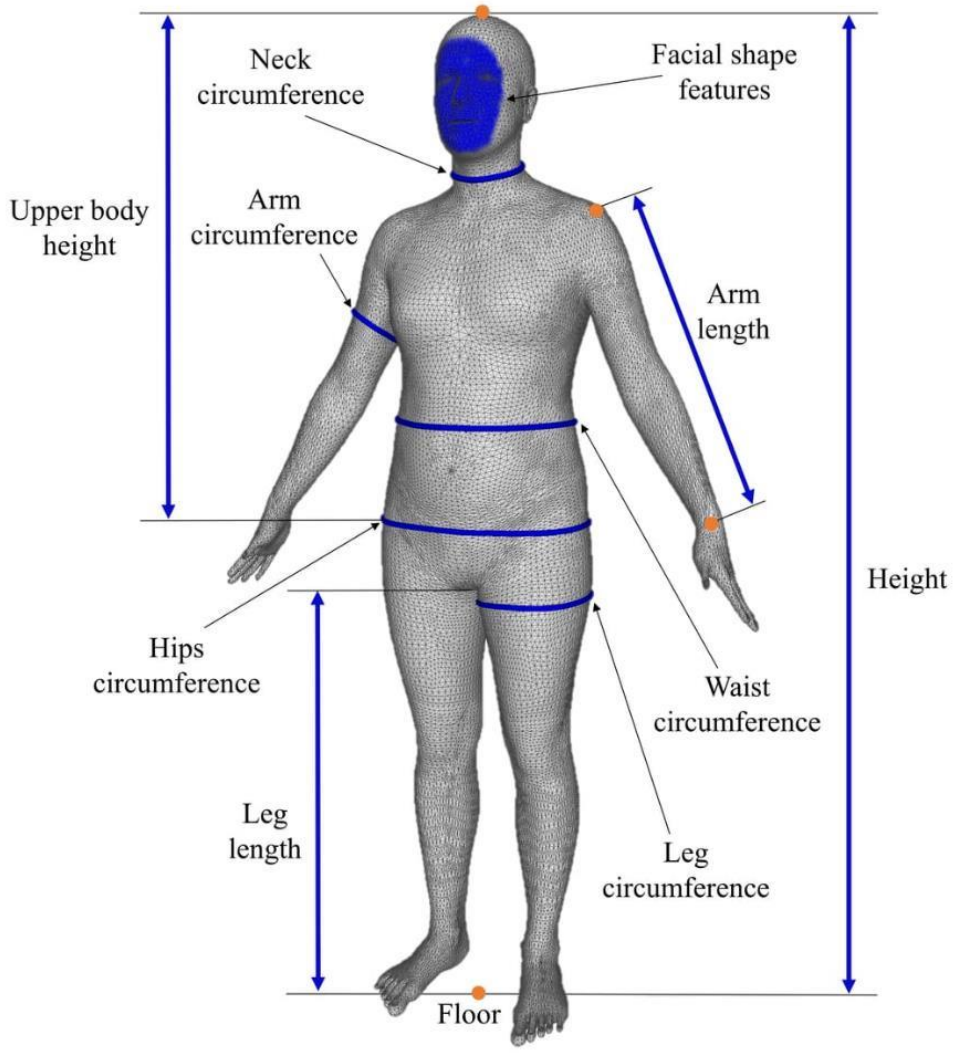


Figure 2.1. Facial shape features and measurements computed from the registered body meshes: height, upper body height, leg and arm length, the perimeters for waist, hips, arm, leg quadriceps and neck.

2.1. INTRODUCTION

The use of 3D human body shape has the potential of changing the way we interact with the world in a wide variety of ways. Applications of this technology have been proved helpful in several fields such as healthcare, cognitive science [23], [24], online shopping [25], [26],clothing [27] and virtual reality [28], [29]. For example, in the healthcare domain, the knowledge of 3D body shape can help in the assessment of the Psoriasis Area and Severity Index (PASI) [30], dosing chemotherapy according to the Body Surface Area (BSA) [31] or estimating a burned body part [32].

All this application lack precision in estimation [33]–[35] and accurate body shape prediction would help the dosage of a particular drug. The prediction of (a less accurate) 3D models from available metadata and body measurements can be considered as a lower cost alternative to full body 3D scanning and processing involving the recognition and processing of different body parts. Note that different practical applications can require 3D shapes and measurements with different precisions. Moreover, the obtaining of less accurate 3D body models computed from the available measurements can be used as a pre-processing tool for accurate registration of 3D models to raw scans.

One of the obstacles in the efficient processing of full body 3D models is the high volume of data. Cost and volume of the data required can be drastically reduced by learning a statistical representation of the human shape space, as described in [62], [63]. Only sparse data, combined with the learned space, are needed to reconstruct a full body scan instead of a dense representation. In the next section, we give an overview considering the prior art which relates and predicts the representation of the 3D body in the shape applications can require the prediction of 3D body shape using the least possible number of metadata and low-cost body measurements. Thus, in this chapter, we address the following points we consider novel: first, we evaluate the predictive power of different combinations of features and study how the error drops when their number increases. Second, we consider facial 3D scan as a lower-cost and less-obtrusive alternative to a full body 3D scan. We analyze the improvement of the body shape prediction when the metadata and the measurements are augmented with features extracted from the facial 3D scans. Third, we apply the above analysis to body parts, which can be arbitrarily segmented on the body.

In our analysis, we considered 12 features that can be relatively easily and reliably collected from a subject: gender, age, weight, and nine measurements shown in Figure 2.1 (most of the figures were generated using MeshLab [64]). Note that we have considered these features as an example, and others can be taken into account depending on the application and the data available. We apply our methodology for the prediction of five example body shapes, shown as segmentation masks in Figure 2.2. We selected those five due to possible applications in healthcare and personal care. For each body part, we assessed how well it can be predicted given each possible subset of the measurements. For each subset of features, we considered how much the prediction accuracy improves when adding to the feature set the features describing the facial geometry, i.e. coefficients in the facial shape space introduced in [4].

The rest of the chapter is organized as follows. Section 2.2 introduce the literature related to body shape analysis and modeling. Subsequently, in Section 2.3, we describe our approach: the registration of the database's population adopting a common template model, the encoding into a parametric shape space using Principal Component Analysis (PCA) and the prediction model used to link face and body shapes. In the last part of the section, we introduce the error measure used for the evaluation of the prediction. The results, presented in Section 2.4, demonstrate

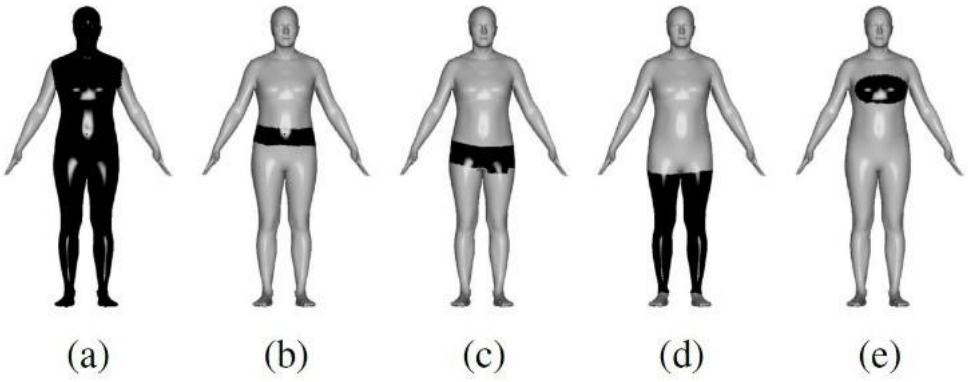


Figure 2.2. The body parts, also called segmentation masks, with segmented area highlighted in black. We refer to them as a) full body mask without arms, b) waistband, c) hips band, d) legs mask and e) breasts mask. The fraction of the segmented vertices with respect to the whole body are 0.78, 0.05, 0.08, 0.26, 0.02 respectively.

that face shape has a positive correlation with different body parts including hips, waist, breast and legs. In the conclusion section, we summarize our view on the 3D body shape prediction from sparse meta-data and the face, and we elaborate on future applications and directions of our work.

2.2. RELATED WORK

Many works have created models that correlate a body statistical shape space to other features, descriptors or meta-information but none of them define a strategy to find the optimal features and none include in the prediction another shape space (in our case the face).

Blanz and Vetter [4] defined how to learn a statistical shape space of the face and then used measurements and semantic descriptors to modify the face appearance. In [62], [63], Allen *et al.* were the first to employ the paradigm explained by [4] on 3D body scans. The authors paved the way for the application of this new method in exploring and studying human shape space. They first registered 250 scans, from dataset [36], solving an optimization problem that minimizes sparse markers' distance, vertices' distance and smoothness of the transformation. Then they learned a linear function mapping anthropometric measurements to the shape coefficient. In [65], [66], Seo *et al.* defined a model that can be modified or generated using only anthropometric measurements. They used radial basis interpolation to reconstruct the relationship between sizing parameters to shape space. Hasler *et al.* [67] includes in the registration phase the high level semantic parameters allowing the generation of realistic body meshes. Wuhrer and Shu [68] generated realist body shape fitting anthropometric measurements using non-linear optimization. Tsoli *et al.* [69] built a model to predict measurements from 3D scans. More recently in [70], [71], Hill *et al.* defined a linguistic space using common body words like fat, rounded or skinny. They first used Amazon Mechanical Turk to link descriptors and body shape by rating photographs. Streuber *et al.* [72] similarly used crowdsourcing to

define verbal descriptors and to demonstrate that they are sufficient for retrieving a realistic 3D scans. While previous works find a relationship between body measurements/characteristics and body shapes, they do not define a strategy to find the optimal subset of them for a specific body part. Moreover, they do not use facial features and/or another shape space as predictor.

Other works explored the correlation between face shape and textures to body parameters: Windhager *et al.* [73] linked facial features of young Caucasian females to body fat proportion using geometric morphometrics. Similarly, Mayer *et al.* [74] retrieved high-resolution face images and registered them using geometric morphometric. However, their experiments do not use the parametric modeling of the human body shape but they predicted a positive correlation between body mass index and waist-to-hip ratio with facial shape and texture. A similar approach to our work is presented in [75] where the authors model the difference between real and virtual measurements and fit a more advanced model with kinematic skeleton. However, they use a linear model for the mapping between features and body shape relying on very specific and not very accurate body measurements. They used VR controllers for collection adding the weight, probably because is a very strong predictor. Moreover, they selected the features based on their acquisition accuracy rather than their predictive power as presented in our work.

Multiple techniques are available to retrieve a 3d representation of a person from different sources (images, depth cameras, sparse markers, silhouettes, etc…). For example, Balan *et al.* [76] reconstructed the parametric shape model [77] using multiple images while more recent works [78]–[81] leverages only a single image and convolutional neural networks.

2.3. METHOD

We developed two parametric models, following the method explained in [4], [62], one for the body and one for the face shape. The face model was derived using more than 3000 3D scans, including the Size China Dataset [82][1]. The parametric body model was derived using more than 4000 full body scans standing in a frontal tree position as shown in Figure 2.3a. The scans were taken mainly from the CAESAR dataset [36].

In the following subsections, we describe the registration of the template meshes into 3D scans, the encoding of the registered models into the selected parameters. Then, we introduce the non-linear prediction model used to find the best subset of features for each segmented body part and, finally, the error measure used to evaluate the experiments.

[1]All other scans were collected at Philips

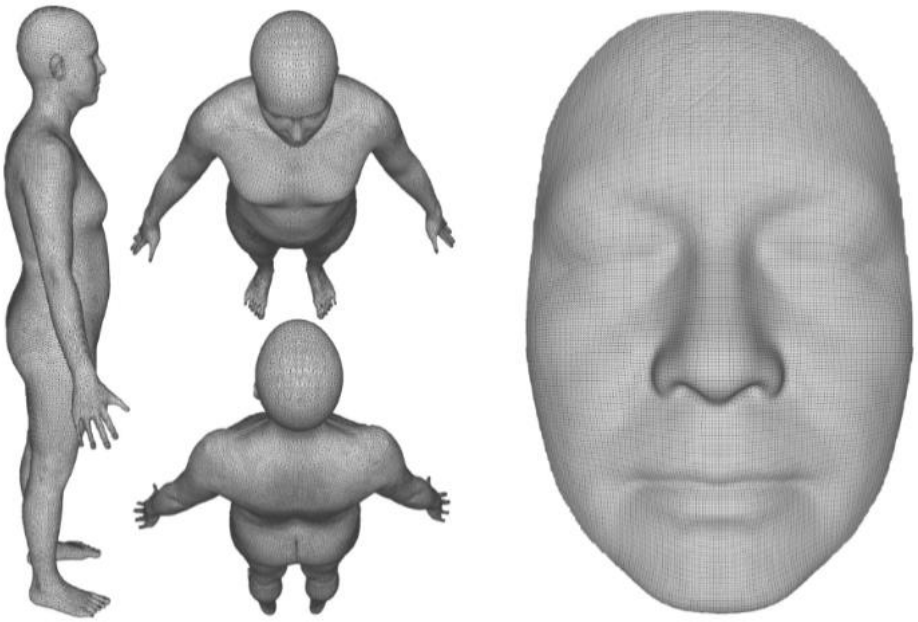


Figure 2.3. (a) Body template mesh standing in tree position and containing $N_P \approx 53000$ vertices. (b) Face template mesh containing $N_O \approx 23000$ vertices.

REGISTRATION

In order to register every face and full body mesh, we employed state-of-the-art non-rigid registration techniques [83]–[85]. We used a template mesh with about $N_P \approx 53000$ vertices for the body, see Figure 2.3a, and another template mesh with about $N_Q \approx 23000$ vertices for the face, see Figure 2.3b. Both template models were then used to register the full body scans dataset. We assessed the quality of the registration via visual inspection and other measures outlined in the survey [83]. For about $N \approx 3750$ full body scans both registrations have shown low fit error (below 0.5mm Root Mean Squared Error (RMSE), as surfaces distance, for the registration of the facial mesh and below 1.0mm RMSE for the full body).

Registration led to the following representation of each participant as the two morphed template meshes. Let $v_{i,j} \in \mathbb{R}^3$be the full body morphed coordinates of vertex $j \in N_P$ at participant $i \in N$. Furthermore, we can write the morphed coordinates of all vertices of scan $i \in N$ as a single flattened vector, stacking all vertices' coordinates together, as

$$\boldsymbol{p}_i^r = \left(\boldsymbol{v}_{i,1}^r, \boldsymbol{v}_{i,2}^r, \dots, \boldsymbol{v}_{i,N_P}^r\right) \in \mathbb{R}^{3N_P}, \qquad (1)$$

and collecting all participants into a rectangular matrix we have

$$P_r = (\boldsymbol{p}_1^r; \boldsymbol{p}_2^r; \dots; \boldsymbol{p}_N^r)' \in \mathbb{R}^{N \times 3N_P} \qquad (2)$$

In the same manner the definition of the face representation is

$$Q_r = (\boldsymbol{q}_1^r; \boldsymbol{q}_2^r; \dots; \boldsymbol{q}_N^r)' \in \mathbb{R}^{N \times 3N_Q}.$$

PARAMETRIC SPACES

The registered meshes were parametrized with Principal Component Analysis (PCA) transformation, using 200 eigenvectors for the body and 180 eigenvectors for the face. The PCA transformation can be written in matrix form as

$$P_r = \bar{P}_r + YD' + E_r \qquad (3)$$

where $\bar{P}_r \in \mathbb{R}^{N \times 3N_P}$ is the matrix of N times repeated average mesh coordinates

$$\bar{\boldsymbol{p}} = \left(\bar{p}_{1_x}, \bar{p}_{1_y}, \dots, \bar{p}_{N_{P_z}}\right) \in \mathbb{R}^{3N},$$
$$\bar{p}_{j_x} = \frac{\sum_i P_r(i, j_x)}{N_P}, \qquad (4)$$

$D \in \mathbb{R}^{3N_P \times 200}$ is the reduced eigenvectors matrix, composed of the 200 `principal` eigenvectors (*i.e.* eigenvectors with highest eigenvalues) of the covariance matrix $(P_r - \bar{P}_r)'(P_r - \bar{P}_r), Y \in \mathbb{R}^{N \times 200}$ is the reduced matrix of PCA coefficients, and $E_r \in \mathbb{R}^{3N_P}$ is the residual error, *i.e.*

$$P_r \approx P = \bar{P}_r + YD' \qquad (5)$$

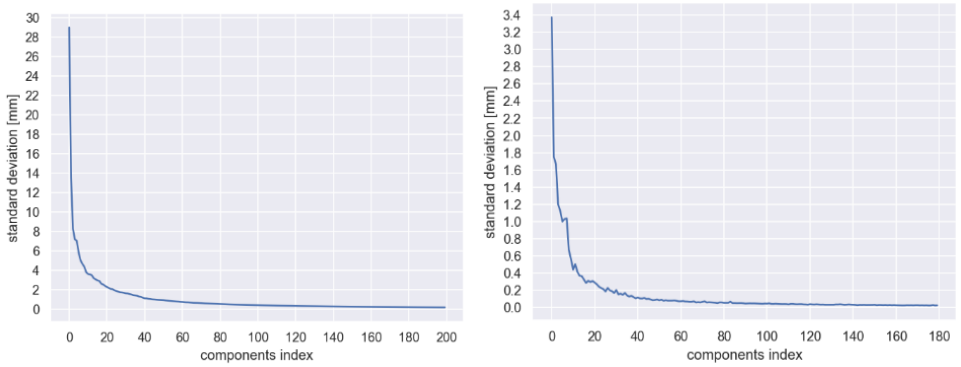
The transformation (5) gives a compact representation of 53000×3-dimensional vectors of vertex coordinates P_r with the 200-dimensional PCA coefficient vectors Y. In the same way, we apply the PCA transformation to the registered facial meshes:

$$Q_r \approx Q = \overline{Q_r} + X_Q D_Q' \qquad (6)$$

where $\bar{Q}_r \in \mathbb{R}^{N \times 3N_Q}$ is the matrix of N times repeated average mesh coordinates, $D_Q$consists of the 180 `principal` eigenvectors of the covariance matrix $(Q_r - \bar{Q}_r)'(Q_r - \bar{Q}_r)$, and $X_Q \in \mathbb{R}^{N \times 200}$ are the facial PCA coefficients. The results of the encoding for both models is shown in Figure 2.4. The residual error between P_r and P, computed using equation (14) and explained in Section Fitness Measures, is less than 2.5 mm. Similarly, the residual error for the face is less than 0.3 mm.

PREDICTION MODEL

In this section, we describe how the body shape coefficients Y are predicted using the subject's features, denoted as $X_F \in \mathbb{R}^{N \times (N_F + 1)}$, (where `+1` corresponds to free term in the regression model) and the face shape space X_Q. As subject features, we have considered *reported weight, age, gender*, and body measurements extracted from the registered meshes such as *body height, arm length, waist circumference*. This set was augmented by including their interactions up to $d = 3$-rd degree. Thus,



(a) Body shape coefficients (b) Face shape coefficients

Figure 2.4. The significance of the encoding, i.e., the standard deviation of the PCA coefficients for the body (a) and for the face (b). The decision to use 200 principal components for the body and 180 for the face was a heuristic decision seeking a compromise between the requirements to represent all shape spaces adequately and to not encode noise. The standard deviation of the last PCA body shape component is 0.18mm and for the face it is 0.025mm.

considering in total N_F personal features, the expanded set corresponds to the terms of the polynomial with degree d build from them. This holds for all features except the ones with lower interactions allowed, like *gender*. In the following, we denote the augmented set of features by $X_G \in \mathbb{R}^{N\times(N_G+1)}$, where the reader can derive the general formula for N_G using basic combinatorial techniques [86] as

$$N_G = \binom{N_F + d}{d} - 1 \tag{7}$$

which, in the case when the (binary) gender feature is included, becomes

$$N_G = \binom{N_F + d}{d} - 1 - (N_F + 1) \tag{8}$$

Equations (7)(8) are given for completeness but are not needed to understand the rest of the chapter or run algorithms which can simply count the combinations. To facilitate the notation, we include the constant term in both X_F and X_G, but it is not counted in N_F and N_G.

Then, we performed multi-linear regression for the body coefficients Y

$$Y = XB + \varepsilon \tag{9}$$

with four settings of the independent variable X, with and without interactions and with and without face coefficients:

$$\begin{aligned} &(a) \quad X = X_F \in \mathbb{R}^{N\times(N_F+1)} \\ &(b) \quad X = X_G \in \mathbb{R}^{N\times(N_G+1)} \\ &(c) \quad X = [X_F, X_Q] \in \mathbb{R}^{N\times(N_F+1+N_Q)} \\ &(d) \quad X = [X_G, X_Q] \in \mathbb{R}^{N\times(N_G+1+N_Q)} \end{aligned} \tag{10}$$

Next, we evaluated the predictions of specific body parts, using the segmentation masks shown in Figure 2.2. The arms were excluded from the segmentation masks

deliberately since subjects had visible variability in the arm positions and for lack of a pose model. To improve the prediction for each body part, instead of solving the basic regression (9), we solved the weighted versions as shown below. Let $I_m \in \mathbb{R}^{3N_P \times 3N_P}$ be the diagonal matrix of mask m, where $I_m(j,j) = 1$ if and only if the vertex is part of the segmentation mask. Recall $P = \bar{P}_r + YD'$ (equation (5) and note that for each body part m we want to have $I_m P$ accurately predicted. Then, assuming the regression model $Y = XB$, we get

$$\begin{aligned} \bar{P}_r I_m + YD'I_m &= \bar{P}_r I_m + XBD'I_m + \varepsilon D'I_m \\ YD'I_m &= XBD'I_m + \varepsilon D'I_m \\ YD'I_m D &= XBD'I_m D + \varepsilon D'I_m D \\ Y\Sigma_m &= XB\Sigma_m + \varepsilon\Sigma_m \end{aligned} \tag{11}$$

where $\Sigma_m = D'I_m D \in \mathbb{R}^{200\times 200}$. The least mean square estimate of B in the above equation is

$$\hat{B}_m = ((X'X)^{-1}X'Y\Sigma_m)\Sigma_m^{-1} \tag{12}$$

for each mask m.

FITNESS MEASURES

For each model and mask, we performed a leave-one-out cross validation on the N participants. In other words, the estimation of $\hat{B}$ has been carried out every time, leaving out the participant to predict. Once computed the predicted body coefficients $\hat{Y}$ we need to convert back, decode, using the PCA transformation (5) to reach the predicted vertices $\hat{P}$ as

$$\hat{P} = \bar{P}_r + \hat{Y}D' = \bar{P}_r + X\hat{B}D' \tag{13}$$

To evaluate the prediction, we first aligned the predicted $\hat{P}(i,:)$ to the original coordinates $\forall i \in [1,N]$with weighted Procrustes [87], and then we computed the vertex-wise RMSE over all participants for each vertex v_{ij} versus its predicted position $\hat{v}_{ij}$

$$E_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\boldsymbol{v}}_{i,j} - \boldsymbol{v}_{i,j}\right\|_2^2} \tag{14}$$

Since comparing the distribution of the vertices errors on the surface is beyond the scope of the research, as a final measure of fitness for the masks, we used the mean absolute error for all vertices:

$$E = \frac{1}{N_P}\sum_{j=1}^{N_P}|E_j| \tag{15}$$

Unlike other works, which used mainly point-to-surface distance, the above error measure also penalizes misplacement of the body part points on the surface and therefore can be considered more accurate.

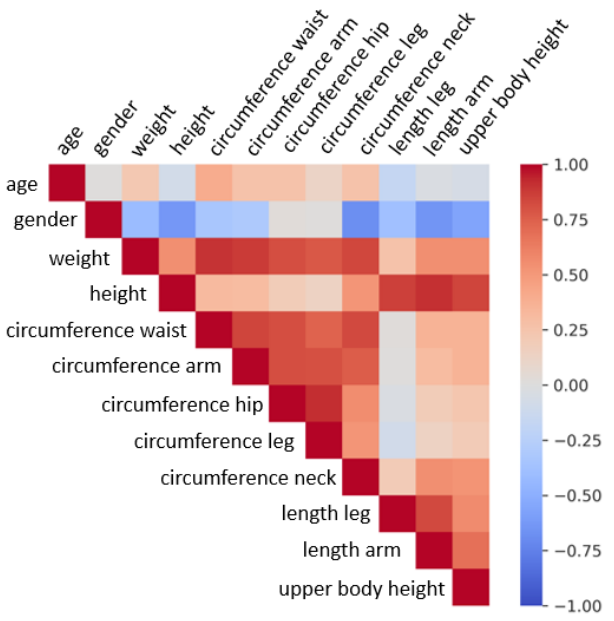
2.4. RESULTS

We evaluated 2 groups of features, listed in Table 2.1, with 12 features in total. The first group is composed of reported gender, age and weight (without clothes), all acquired in [11]. The second group includes parametric measurements that were computed from the registered body meshes: the height computed as head to floor; upper body height as head to the highest touchable point of the pelvis; arm length as the distance between acromion (shoulder) to the distal end of the middle finger; leg length from crotch to floor; the perimeters for waist as the midpoint between the lower margin of the last palpable rib and the top of the iliac crest; hips circumference it is performed at the most prominent point, on the major trochanters, and at the level of the maximum relief of the gluteal muscles; arm circumference taken from the midpoint of the total length of the arm, between acromion and olecranon; leg quadriceps circumference taken from the midpoint of the total length of the thigh; neck circumference taken from the midpoint of the total length of the neck. The covariance matrix of all the features is presented in Table 2.2.

Table 2.1. Features' definition where CQ stands for CAESAR Questionnaire and PM for Parametric Measurement

Name	Type	Mean	Source
Gender	Male[1] or Female[2]	1.53 ± 0.5	CQ
Age	Years	38.00 ± 12.59	CQ
Weight	Kg	74.56 ± 18.09	CQ
Height	Y-length [mm]	1701.38 ± 100.78	PM
Waist	Circumference [mm]	889.83 ± 150.45	PM
Arm	Circumference [mm]	306.00 ± 43.45	PM
Hip	Circumference [mm]	1037.99 ± 106.07	PM
Leg	Circumference [mm]	614.94 ± 67.91	PM
Neck	Circumference [mm]	364.09 ± 43.33	PM
Leg	Y-distance [mm]	764.41 ± 55.67	PM
Arm	distance [mm]	557.64 ± 40.59	PM
Upper Body	Y-distance [mm]	750.72 ± 42.90	PM

Table 2.2. Features' correlation matrix



We assessed the importance of each feature by performing a search over all possible combinations of the set X_F resulting in $2^{12} = 4096$ possible subset of features. We consider the empty subset as the error compared to the average of

Table 2.3. Error E for full body without arms using X_G best features

N_F	N_G	X_F	$[X_F, X_Q]$	X_G	$[X_G, X_Q]$	Features
0	0	36.70 ± 10.69	36.70 ± 10.69	36.70 ± 10.69	36.70 ± 10.69	Avg distance
1	3	20.92 ± 5.36	17.46 ± 3.52	20.89 ± 5.35	17.45 ± 3.52	Height
2	9	18.11 ± 3.20	16.29 ± 2.64	17.93 ± 3.14	16.21 ± 2.60	Weight, Height
3	19	17.00 ± 3.20	15.28 ± 2.70	16.75 ± 3.13	15.15 ± 2.64	Weight, Height, LegL
4	34	16.31 ± 2.80	14.95 ± 2.47	16.00 ± 2.69	14.73 ± 2.38	Height, WaistC, HipC, LegL
5	55	15.91 ± 2.66	14.65 ± 2.35	15.56 ± 2.57	14.37± 2.28	Height, WaistC, HipC, LegL, UBodyH
6	83	15.69 ± 2.67	14.51 ± 2.37	15.22 ± 2.55	14.13 ± 2.29	Height, WaistC, HipC, LegL, UBodyH, LegC
7	119	15.56 ± 2.61	14.46 ± 2.34	15.02 ± 2.51	13.98 ± 2.25	Row 6 + Age
8	164	15.50 ± 2.60	14.42 ± 2.34	14.80 ± 2.47	13.81 ± 2.23	Row 7 + Weight
12	**441**	**15.35 ± 2.54**	**14.32 ± 2.30**	**13.92 ± 2.28**	**13.00 ± 2.07**	**All 12 features**
4	29	17.41 ± 2.91	16.15 ± 2.65	17.11 ± 2.79	15.91 ± 2.54	Age, Gender, Weight, Height

the population. For each subset, we compared four different feature designs $X_F, X_G, [X_F, X_Q]$, $[X_G, X_Q]$. The maximum number of features reached by models without the face is N_G minus all combinations of the gender. In our example, the maximum number of features is $N_F = 12$ and all combinations of gender from second order are $N_F + 1$ hence using equation (8) we have that the maximum number of regressors, when using interactions is $N_G = 441$. Considering instead the example with age, gender, weight and height, where $N_F = 4$ we have $N_G = 29$. In the following, we present the errors for the full body mask without arms and next the errors for all the remaining four body parts.

FULL BODY MASK WITHOUT ARMS

Table 2.3 shows for each cardinality of X_F the best model for $X = X_G$ and the error using the other three input matrices. The most accurate single feature is the height with $E = 20.89$mm, because it is a major indicator of body size. The only feature that actually outperforms height is the body volume. However, we excluded it from the analysis since, it is a very uncommon measurement to be taken in a real-life scenario. Note that height alone is giving a smaller error than using weight and face shape combined. The best combination of two features is the height and the weight, resulting in 17.93mm residual error. The minimal error is achieved using all 12 features, and it is 13.92mm for X_G and 13.00mm for $[X_G, X_Q]$. The average error reduction for those 12 best models is 1.33mm, and the average effect of adding the face coefficients X_Q is the drop of the error with 8.12%. We think that adding new measurements will only affect the error minimally because of the small variability of different subjects' pose. In fact, the pose cannot be predicted from the measurements, and we believe that the addition of pose normalization methods would result in lower errors.

In order to evaluate the significance of adding the face shape, we considered the model with $X = [X_Q, X_G]$ where X_G is augmented from $X_F =$ [age, gender, weight, height]. This model has an error of 15.91mm, which is better than the error of the model with $N_F = 4$ best predictors without face. Hence the face shape can replace

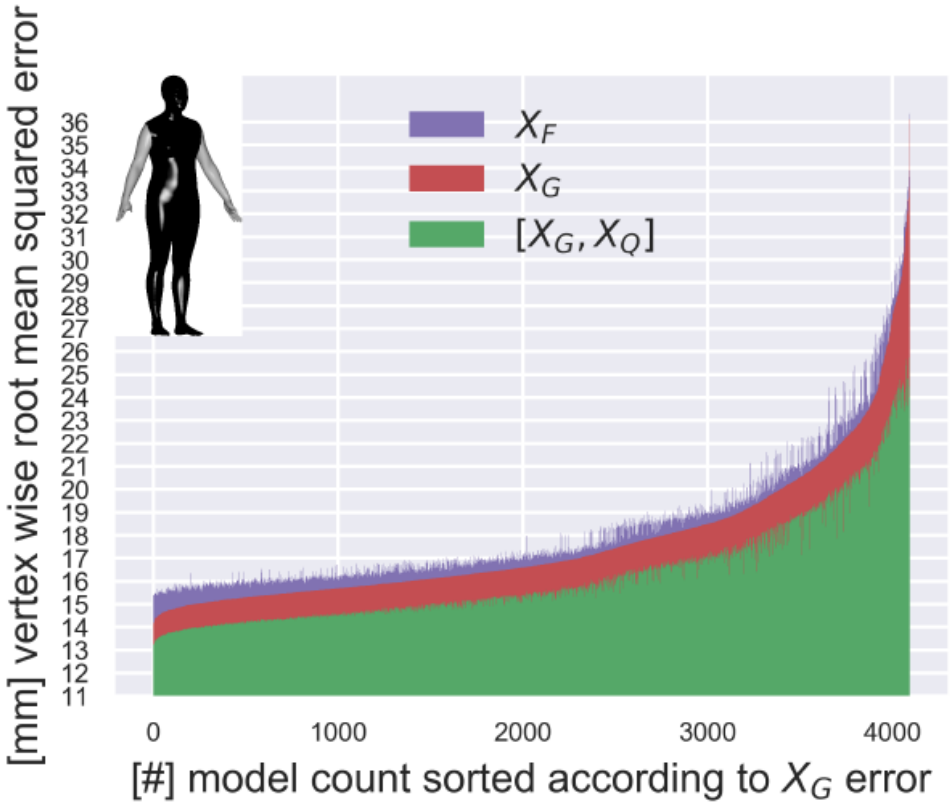


Figure 2.5. Full body mask with all 4095 models sorted according to the error of X = XG. The minimum error of 13mm is achieved using all 12 features plus the face coefficients.

detailed parametric measurements. Thus, for example, the face coefficients combined with age, gender and weight features give lower error than the prediction using waist, hip circumference and leg length features. In Figure 2.5 all possible subsets, excluding the empty one, visually demonstrate that the face has a significant positive contribution to the prediction. In fact, the average error drop, when extending X_G to $[X_G, X_Q]$, is 0.98mm or 9.72%. On opposite, the more features are considered the bigger the effect of adding interactions between them adding of interactions, as one can see in the Table 2.4 to Table 2.7, when comparing columns X_F to X_G.

Additional Four Masks

An interesting observation is that, while height is coming first for the body, it is not the case for hips and waist band prediction, where weight gives a better accuracy among the single feature predictors. As expected, the circumferences are now playing a much more significant role in the specific masks compared to the full body mask. This is shown in Table 2.4 to Table 2.7.

Table 2.4. Error E for waistband using X_G best features

N_F	N_G	X_F	$[X_F, X_Q]$	X_G	$[X_G, X_Q]$	Features
0	0	22.64 ± 3.66	22.64 ± 3.66	22.64 ± 3.66	22.64 ± 3.66	Avg distance
1	3	12.68 ± 2.34	10.51 ± 1.53	12.59 ± 2.32	10.44 ± 1.51	HipC
2	9	11.21 ± 1.08	9.59 ± 0.96	11.00 ± 1.05	9.47 ± 0.94	WaistC, HipC
3	15	10.63 ± 1.20	9.48 ± 0.92	10.37 ± 1.19	9.29 ± 0.97	Gender, WaistC, HipC
4	29	10.41 ± 1.11	9.38 ± 0.95	10.04 ± 1.08	9.11 ± 0.93	Gender, WaistC, HipC, LegC
5	49	10.29 ± 1.08	9.29 ± 0.92	9.87 ± 1.06	8.96 ± 0.90	Gender, WaistC, HipC, LegC, UBodyH
6	83	10.21 ± 1.08	9.20 ± 0.87	9.65 ± 0.98	8.78 ± 0.86	WaistC, HipC, LegC, UBodyH, Height, LegL
7	119	10.06 ± 1.05	9.13 ± 0.89	9.44 ± 0.98	8.65 ± 0.85	Row 6 + Age
8	155	10.00 ± 1.05	9.11 ± 0.88	9.25 ± 1.01	8.50 ± 0.87	Row 7 + Gender
12	**441**	**9.92 ± 1.02**	**9.05 ± 0.87**	**8.59 ± 0.94**	**7.93 ± 0.82**	**All 12 features**
4	29	12.06 ± 1.30	10.96 ± 1.13	11.54 ± 1.32	10.47 ± 1.14	Age, Gender, Weight, Height

Table 2.5. Error E for hips band using X_G best features

N_F	N_G	X_F	$[X_F, X_Q]$	X_G	$[X_G, X_Q]$	Features
0	0	20.42 ± 3.01	20.42 ± 3.01	20.42 ± 3.01	20.42 ± 3.01	Avg distance
1	3	11.76 ± 2.67	9.59 ± 1.63	11.70 ± 2.66	9.51 ± 1.61	HipC
2	9	10.76 ± 1.88	9.22 ± 1.52	10.58 ± 1.79	9.09 ± 1.47	HipC, NeckC
3	19	10.29 ± 1.80	8.92 ± 1.46	9.94 ± 1.78	8.69 ± 1.40	WaistC, HipC, LegC
4	29	9.88 ± 1.62	8.80 ± 1.42	9.48 ± 1.52	8.50 ± 1.34	Gender, WaistC, HipC, LegC
5	55	9.70 ± 1.56	8.69 ± 1.37	9.25 ± 1.47	8.34 ± 1.29	Height, WaistC, HipC, LegC, LegL
6	83	9.44 ± 1.47	8.50 ± 1.27	8.88 ± 1.36	8.06 ± 1.20	Height, WaistC, HipC, LegC, LegL, UBodyH
7	119	9.32 ± 1.44	8.44 ± 1.26	8.73 ± 1.34	7.95 ± 1.18	Row 6 + Age
8	164	9.28 ± 1.44	8.42 ± 1.25	8.59 ± 1.31	7.85 ± 1.16	Row 7 + Weight
12	**441**	**9.19 ± 1.41**	**8.37 ± 1.24**	**8.00 ± 1.21**	**7.33 ± 1.09**	**All 12 features**
4	29	11.66 ± 1.55	10.59 ± 1.36	11.17 ± 1.52	10.12 ± 1.33	Age, Gender, Weight, Height

For both the waist and hips masks, the best performing feature is the hip circumference, registering an error of 12.59mm and 11.70mm respectively. The lowest error reached using all features for the waist mask is 8.59mm whereas the hip mask achieved a minimum error of 8.00mm. For the breast mask, the best single feature is the waist circumference that reaches an error of 9.30mm, and as foreseen, gender plays an important role as well. For this mask the lowest error, achieved using all features, is 6.50mm. Finally, analyzing the error registered in the leg mask, it can be noticed that the leg length plays the most crucial role, reaching an error of 13.94mm. It is followed by the leg circumference and the height. The minimum error achieved in this mask, using all the features, is 10.12mm.

Overall, the face improves the most the hips band where the reduction for the best 12 models is 10.45% (0.99mm). For the waist mask the average reduction is 9.71% (0.98mm) and for the full body, described in Section A, the drop is 8.12% (1.33mm). Finally, the reduction for the legs is 7.32% (0.84mm) and the face achieves the least

Table 2.6. Error E for breasts using X_G best features

N_F	N_G	X_F	X_F, X_Q	X_G	X_G, X_Q	Features
0	0	13.77 ± 4.24	13.77 ± 4.24	13.77 ± 4.24	13.77 ± 4.24	Avg distance
1	3	9.38 ± 2.54	7.45 ± 1.70	9.30 ± 2.52	7.44 ± 1.69	WaistC
2	6	7.73 ± 1.72	7.15 ± 1.58	7.64 ± 1.66	7.07 ± 1.53	Gender, WaistC
3	15	7.55 ± 1.66	7.03 ± 1.55	7.42 ± 1.60	6.93 ± 1.49	Gender, Weight, WaistC
4	29	7.47 ± 1.64	6.96 ± 1.53	7.27 ± 1.55	6.81 ± 1.46	Gender, Weight, WaistC, UBodyH
5	49	7.42 ± 1.64	6.94 ± 1.52	7.18 ± 1.53	6.73 ± 1.45	Age, Gender, Weight, WaistC, UBodyH
6	76	7.39 ± 1.64	6.91 ± 1.52	7.10 ± 1.52	6.68 ± 1.43	Gender, Weight, Height, WaistC, LegC, LegL
7	111	7.36 ± 1.63	6.90 ± 1.51	7.01 ± 1.50	6.61 ± 1.42	Row 6 + Age
8	155	7.34 ± 1.63	6.89 ± 1.51	6.92 ± 1.48	6.54 ± 1.40	Row 7+ UBodyH
12	**441**	$\mathbf{7.29 \pm 1.60}$	$\mathbf{6.85 \pm 1.50}$	$\mathbf{6.50 \pm 1.38}$	$\mathbf{6.15 \pm 1.30}$	**All 12 features**
4	29	7.83 ± 1.81	7.27 ± 1.67	7.57 ± 1.69	7.07 ± 1.58	Age, Gender, Weight, Height

Table 2.7. Error E for legs using X_G best features

N_F	N_G	X_F	X_F, X_Q	X_G	X_G, X_Q	Features
0	0	19.52 ± 4.96	19.52 ± 4.96	19.52 ± 4.96	19.52 ± 4.96	Avg distance
1	3	13.99 ± 2.27	11.91 ± 1.78	13.94 ± 2.25	11.89 ± 1.78	LegL
2	9	12.14 ± 1.46	10.89 ± 1.26	12.07 ± 1.45	10.83 ± 1.25	LegC, LegL
3	19	11.65 ± 1.39	10.78 ± 1.25	11.54 ± 1.37	10.68 ± 1.22	Height, LegC, LegL
4	34	11.51 ± 1.35	10.69 ± 1.22	11.37 ± 1.31	10.55 ± 1.20	Height, HipC, LegC, LegL
5	55	11.40 ± 1.29	10.61 ± 1.20	11.15 ± 1.26	10.42 ± 1.18	Height, WaistC, HipC, LegC, LegL
6	83	11.31 ± 1.28	10.57 ± 1.19	11.00 ± 1.24	10.31 ± 1.16	Height, WaistC, HipC, LegC, LegL, UBodyH
7	119	11.25 ± 1.27	10.54 ± 1.19	10.86 ± 1.24	10.20 ± 1.16	Row 6 + Weight
8	164	11.21 ± 1.27	10.52 ± 1.20	10.72 ± 1.22	10.08 ± 1.15	Row 6 + Age, ArmC
12	**441**	$\mathbf{11.13 \pm 1.27}$	$\mathbf{10.47 \pm 1.19}$	$\mathbf{10.12 \pm 1.15}$	$\mathbf{9.53 \pm 1.09}$	**All 12 features**
4	29	12.88 ± 1.60	11.99 ± 1.44	12.70 ± 1.55	11.84 ± 1.40	Age, Gender, Weight, Height

reduction in the breasts area with 7.14% (0.54mm). Thus, we can deduce that the face is more relevant in predicting hips and waists compared to the legs and breasts.

2.5. DISCUSSION AND CONCLUSION

In this chapter, we showed how to couple two high-resolution parametric spaces of body and face with metadata and low-cost measurements. Initially, we predicted the body shape parameters using anthropometric measurements. In addition, we included the face shape parameters to our predictive model leading to the conclusion that they always improve the prediction. Moreover, we focused our analysis only on surface body parts as related to our applications; However, the same methodology and results can be presented for surface-area-to-volume estimations.

As far as our analysis is concerned, additional research could lead to the increase of the accuracy of our predictions. In the future, it would be helpful to include a skeleton model to factor out the pose, as shown in [88]. This, in turn, will prevent

information loss in the PCA encoding due to factors affecting the pose (e.g. the position of arms and legs). The regression model can be enhanced using regularizing techniques. We believe that Lasso [89] is the best to set the tail components of the face to zero when needed. We avoided presenting those regularizations since it is beyond the scope of this chapter. A further direction of research is the prediction of the face components out of images of the face. This way, one could predict the body shape coefficients using pictures instead of the 3D shape. A possible path to follow is extracting landmarks after aligning and then extrapolating 3D shapes. Suitable techniques to follow this approach are explained in [90], [91].

An interesting application of the procedure described in this chapter could lie in correlating other body parts to one another. In principle, any body part could be registered and encoded via PCA. As an example, the investigation of the relationship of foot features on the back has been studied via Geometric Morphometric in [92], [93]. They correlated foot shape with anthropometric measurements like height, Body Mass Index (BMI) and gender. Their work can be enriched by registering belly, hips or back areas and then by studying the effect on the back with our approach. Although, several studies have been conducted using BMI as a base factor, body shape coefficients have the potential of conveying more information and thus improving the prediction.

3. HAIR COUNTING WITH DEEP LEARNING

"Ignoring isn't the same as ignorance,
you have to work at it."
—Margaret Atwood, The Handmaid's Tale

We present a set of deep learning models aimed at solving the hair counting problem in human skin images. All the models are end-to-end, providing a mapping from the input image to a single scalar corresponding to the number of hair. The list of models corresponds to the most common deep learning architectures that worked over-time in various applications, where some of the networks were adapted to output the hair count. Results show that autoencoder architectures with skip connections work best for such end-to-end counting task, hinting at increased performance when multi-task learning is used. With the results presented, we speculate on the possibility to remove human annotator from the tedious task of manual counting of skin hair.

This chapter is based on the paper *Hair counting with deep learning* authored jointly with Dmitry Znamenskiy, Nicola Pezzotti and Milan Petkovic, which was published in International Conference on Biomedical Innovations and Applications Proceedings 2020.

Figure 3.1. Example of hairs to count. The image is captured using a DSLR camera with a special attachment controlling the illumination and the distance to the skin patch.

3.1. INTRODUCTION

Anderson and Parrish's principle of selective photo thermolysis [38], [39], build the technological base underlying the LHR (laser hair removal). In [40], Grossman *et al.*, were the first to use it for photo-epilation and, since its approval, in the Food and Drug Administration (FDA) has been increasingly growing its popularity due to safe, fast and effective method for hair removal [41], [42]. The LHR devices effectiveness needs to be proven for many reasons such as, mention above, the FDA approval, to get market shares or improve an older model. Classical experimental variables are ranging from device settings like light temperature, device orientation, type of device heads to subject conditions like skin color, skin area surface and user engagement.

A primary evaluation metric is the number of hair removed by the device. The hair counting is usually done by visual inspection and visual hair count, which is a long and tedious manual work. Object counting is relevant also when counting hair for applications related to skin beauty, dermatology and trichology [94]. In this chapter, we present several solutions for the replacement of the manual counting, with an automated deep learning counting system, and we review the most common end-to-end architectures adapted for counting. Deep learning is state-of-the-art for object counting in various applications such as agriculture [43]; microbiology [44], [45]; security [46], [47]; and wild life conservation [48].

In Philips, we collected a dataset of skin patches, including more than 4000 images from more than 100 volunteers. Each image, skin patch, is captured in various

sessions, usually four, using a DSLR camera with a special attachment controlling illumination and distance. A sample image is illustrated on Figure 3.1. Note that, since the effect of the photo epilation is not immediate, for better monitoring the experiment, the area of interest on the skin is shaved to optimize device performance [40], [95]. Therefore, a fraction of the collected pictures contains trimmed hair. A sizeable investment in time and resources was spent to manually count the skin hair in the above dataset. We believe that, by adopting an automatic system for statistical evaluation of hair removal devices, the FDA approval will be facilitated and streamlined.

To summarize, the contribution of our work is as follows: we propose to adopt a deep learning based automated hair counting algorithm, and we investigate several deep learning architectures to perform end-to-end hair counting. Finally, we show that the adaptation of a segmentation network does not only outperforms other architectures but shows the emergence of a segmentation behavior by only regressing the total hair count, hinting at the potential of adopting a multi-task learning approach.

3.2. RELATED WORK

The overview below describes prior-art object counting methods, which we will group, based on the type of used annotations: a) annotation free and unsupervised learning methods, b) dense annotation by segmenting each hair, c) point annotations, d) weak image annotation with the total count of hair.

Most of the earlier literature on hair counting and segmentation does not rely on machine learning. Thus Vallotton and Thomas [49] developed an approach for measuring body hair. They iteratively merge small line segments to account for curly variations. They only addressed cases in which the hair is darker than the skin, though their algorithm can be easily modified to predict the opposite. In [50], [51], Shih and Lin proposed an annotation free approach to count hair. The authors mainly used traditional computer vision techniques to detect lines. This approach, while good in a controlled image acquisition environment, suffers from different light conditions and perspective distortions. Besides, a slight change in the acquiring device implies the need for re-tuning the numerous parameters of the computer vision algorithm. Lim *et al.* [52], developed an automatic hair counting system to evaluate laser hair removal. They also validated their performance in clinical trials. They collected images from the thighs of five volunteers with Fitzpatrick skin type III-IV. Their percentage error was <5% in each subject.

The second group of the prior art use the finest image annotations where each hair is segmented in the image. While this is the most labor-consuming type of annotation, the segmentation of hair does not directly imply the count. Multiple hair, or object, could intersect and multiple instances of them rely on only one region, and for example, active contour will fail to count correctly. While this is not true when using instance segmentation. Multiple ways were presented with one very popular being the Mask R-CNN instance detector [96].

The third group of prior art apply supervised learning algorithm to the point annotation, recording a location $d \in \mathbb{R}^2$ in the image for each object to be counted

[97], [98]. Thus in [99], the authors propose an interactive annotation system to count object in which the user is requested to place a dot on top of each object. A big sub-group leverage the point annotations generating a continuous density [44]. In these methods, a density map (also referred as a proximity map, or heatmap) around each dot is generated using gaussian with center the dot or a ball with linear decay from each dot [100]. This will rephrase the problem to estimate the density map, and then retrieve the count integrating the density. Those last categories allow multitask learning where counting and segmentation can be combined. For example, in [48], Arteta *et al.* uses an FCN to count dot-annotated penguins in combination to segment the foreground, estimate a density map and uncertainty map. In [45], Falk *et al.* developed a tool to count, segment and measures cells with U-net.

Finally, we consider the prior art, which utilizes the weak image annotation where for each image x only the total count of hair y is recorded, which further reduces the annotation burden in comparison to the point annotation. A classical approach of using the above annotation in a supervised learning algorithm would be the detection of image features, for example, using a Gaussian pyramid, followed by the regression of the count y from the feature values. Recently, the feature detection part is replaced with an end-to-end artificial neural network, often a fully convolution one [101]. In this chapter, we consider the application of the different neural networks that transform the input image into the total count. Methods and architectures, used to regress the total count y, are presented in Section 3.3, and the experimental setup is described in Section 3.4.

3.3. METHODS

We selected five widely known models architectures to count hair with four of them which primary task is object classification: lenet-5 [102], vgg16 [103], resnet50 [104] and densenet121 [105]; and one, U-net, introduced to solve biomedical segmentation tasks [106].

We selected those models because of their success and popularity across different domains and applications. We adapted the classification to predict one scalar count—one class instead of e.g. 1000 class for the image net challenge—with a final Rectified Linear Unit activation (to keep the prediction strictly positive). We adapted the segmentation architecture using a global sum final [107]. We did not fine-tune the models, but every model was trained from scratch as specified in the optimizer section. Note that the models developed for classification are often used for regressing a single value, while the segmentation networks are hardly used for this task.

The Poisson Negative Log Likelihood is the loss most commonly applied used for object counting tasks:

$$L(y,\hat{y}) = \frac{1}{N}\sum_{p=0}^{N}(\widehat{y_p} - y_p \log\widehat{y_p})$$

where $y,\ \hat{y}$ are respectively the ground truth and predicted count. A relatively early example using such loss and relative neural network can be found in Fallah *et al.* [108]. Moreover, we used mean absolute error

$$MAE(y,\hat{y}) = \frac{1}{N}\sum_{p=0}^{N} |\widehat{y_p} - y_p|$$

and the Median and Mean Percentage Error

$$MeanAPE(y,\hat{y}) = \frac{1}{N}\sum_{p=0}^{N} \frac{\left|\widehat{y_p} - y_p\right|}{y_p}$$

as the metrics used to benchmark the experiments.

Adam [109] was used with: fixed $\beta_1 = 0.9$, $\beta_2 = 0.99$ and $\epsilon = 10^{-8}$; three different learning rates lr were $0.01, 0.001, 0.0001$; 1000 epochs and minibatches of size 16. We did not use early stopping or weight decay. All the networks were trained on a GPU Nvidia Tesla v100 with 16GB. As a framework, we opted for PyTorch [110] due to its simplicity to find and tailor model architectures.

We augmented the input data in different ways: A) no augmentations, B) random vertical and horizontal flips, C) random color jitter using PyTorch function ColorJitter(brightness=0.1, contrast=0.1, saturation=0.1, hue=0.1)), D) rotation by and random degrees, E) random scaling between 0.8 and 1.2, F) all the previous in the cascade. All the augmentations are performed during training, with the input image being transformed before the forward propagation. Hence, each input image is augmented a number of times equal to the number of epochs.

3.4. EXPERIMENTAL SETUP

We have analyzed a Philips dataset consisting of 4358 skin images which were collected from the different body parts of subjects varying in skin color and skin type. Various human annotators produced a single count $y \in \mathbb{N}^+$ for each image $x \in \mathbb{R}^p$ using the original resolution (4288, 2842) pixels. Figure 3.2 shows the distribution of the label value over the number of images present in the dataset. The labels y, representing the number of hair in the picture, where collected at Philips during multiple years and studies. Every annotator was first trained and evaluated by an expert supervisor. The annotator was required to annotate a test set—composed of n=30 examples—and accepted to perform further annotations only if her results were acceptable (<5% difference of what the supervisor counted). The annotator was provided with a special predefined hair-counting setup (image resolution, desk setting, etc…) in

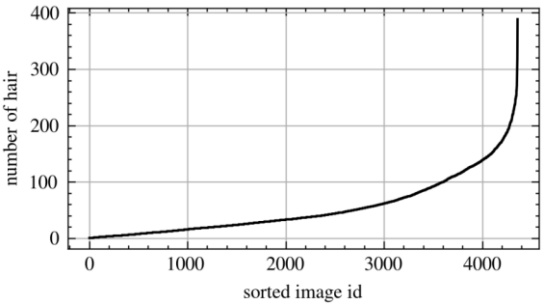


Figure 3.2. Distribution of hair count. The average is 54.5581 +- 51.2338 with a maximum of 389 hair and zero as minimum.

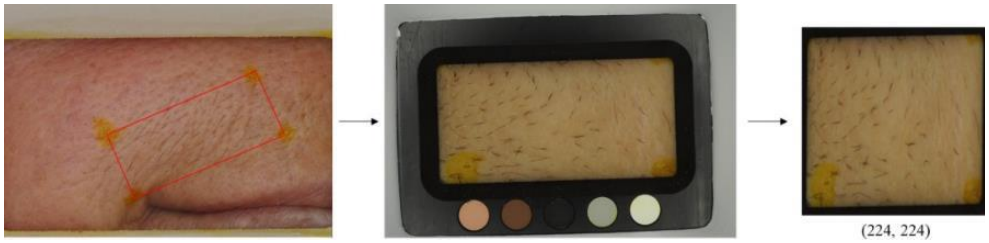


Figure 3.3. Data preprocessing pipeline. The skin patch is collected with a reference frame that is later removed before feeding it to the training loop.

order to have a similar look and feel across different sessions. The areas considered in the study are the upper lip, the armpit, the bikini area and the leg. These are common anatomical areas of interest for a skin beauty device such as the LHR.

For 85 images, the counting time t was recorded. By applying a simple regression, we observe that the time follows this rule

$$t = 1.03 + 0.02c$$

which we used to compute an estimate counting time of 9412 minutes for the whole dataset. Given that an annotator could not concentrate more than 2 hours per day on the counting, 9412 minutes would turn into 78 annotation days which could be potentially saved with an automatic counting solution.

Experimentally we found that removing of the image frame, occupying 52% of the image area as shown in Figure 3.3, improves the results. Therefore, we decided to crop out the image frame using the following pre-processing algorithm. We used the Resnet18 architecture to regress the corner positions where the training data was obtained using traditional computer vision techniques. Thus we first did identify the corners $B = [b_1; b_2; b_3; b_4] \in R^{4,2}$ using method [111] and then removed the outliers detected with unsupervised clustering, i.e. using method [112] with four centroids. In this way, we achieve a well-defined training image without the reference frame. Having the frame corners detected, we aligned the images on the corner with Kabsch [113] towards the mean of the corners and cropped the area containing the skin hair.

After the frame removal we reduced the resolution of the image to (224, 224) and normalized the image channels with means $[0.4725,\ 0.4375, 0.3773]$ and standard deviations $[0.2593, 0.2407, 0.2187]$ for the deep learning training and inference.

3.5. RESULTS

In this section, we present the results obtained with the experimental setup described above. We divided the dataset into train, validation and test set with relative proportions of 90%, 5% and 5%. To guide the model selection for each architecture, we did an exhaustive search for the best learning

Table 3.1. Best result for each architecture

Model	MAE	Mean APE	Median APE
U-net	6.42	22%	12%
Vgg16	7.70	26%	15%
Resnet50	7.72	29%	15%
Denset121	8.51	31%	15%
Lenet5	10.89	51%	21%

rate and augmentation strategy (A, B, C, D, F) according to the lowest validation error. We then report the results for the test set which we found similar to the validation one.

No matter which architecture the best strategy is always F, all the augmentation in cascade, thus probably a wider range of distortions can be applied to increase performance. Another hint in this direction is the fact that any degree rotation (D), apart from being the second best, outperform (B) in every run. On another side, the color augmentation (C) performed worse than no augmentation (A) in almost every run and therefore must be excluded in further experiments.

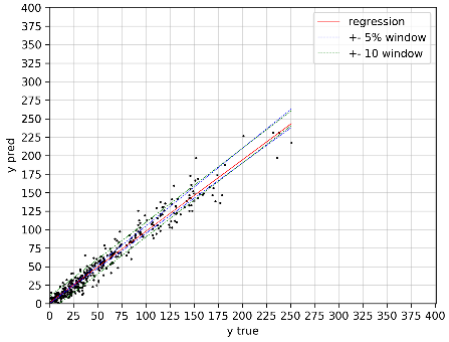


Figure 3.4. Scatter plot of predicted versus true count for the best model. In red, the regression line with coefficient of determination R2 = 0.9578.

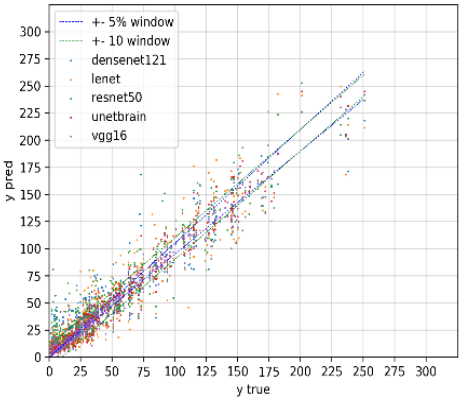


Figure 3.5. Scatter plot for each architecture. The ensemble does not improve the model and there is no apparent correlation between human and multiple machines.

While U-net always converged with the worst test MAE being 10.87 notably also densent121 always did, with a worst MAE 14.13, compared to the other “classification” architectures lenet5, vgg16 and resnet50. The best learning rate oscillates between 0.001 and 0.0001 without a big results difference. Table 3.1 below shows the result for each architecture corresponding the best learning rate (which is either 0.001 or 0.0001) and the best strategy (which is always F). See also that the best performing architecture is U-net with the hyperparameters, according to the validation set, $lr = 0.001$ and strategy F. Figure 3.4 shows the predicted versus actual count for the best model as well as the regression between them. The MAE is 6.99, and the MAPE mean and median are 24% and 12% respectively, which is worse compared to U-net.

3.6. DISCUSSION AND CONCLUSION

Figure 3.5 shows that none of considered methods can replace the qualified human annotator, (which would require the counting Mean APE errors of less than 5%), mainly due to errors on the images with low hair count. Note however, that for the practical task that we are proposing, i.e., streamlining the evaluation process of hair removal devices, the error is close to what is achieved by a human, while greatly reducing the cost by automating the process. Moreover, as noted in the results section, the best outcome is achieved by implementing all augmentations in cascade (e.g., strategy F). Hence, we believe that further augmentation optimization might increase performance and, hopefully, narrow the gap compared to human performance. Further, optimization and analysis of the color augmentation ranges,

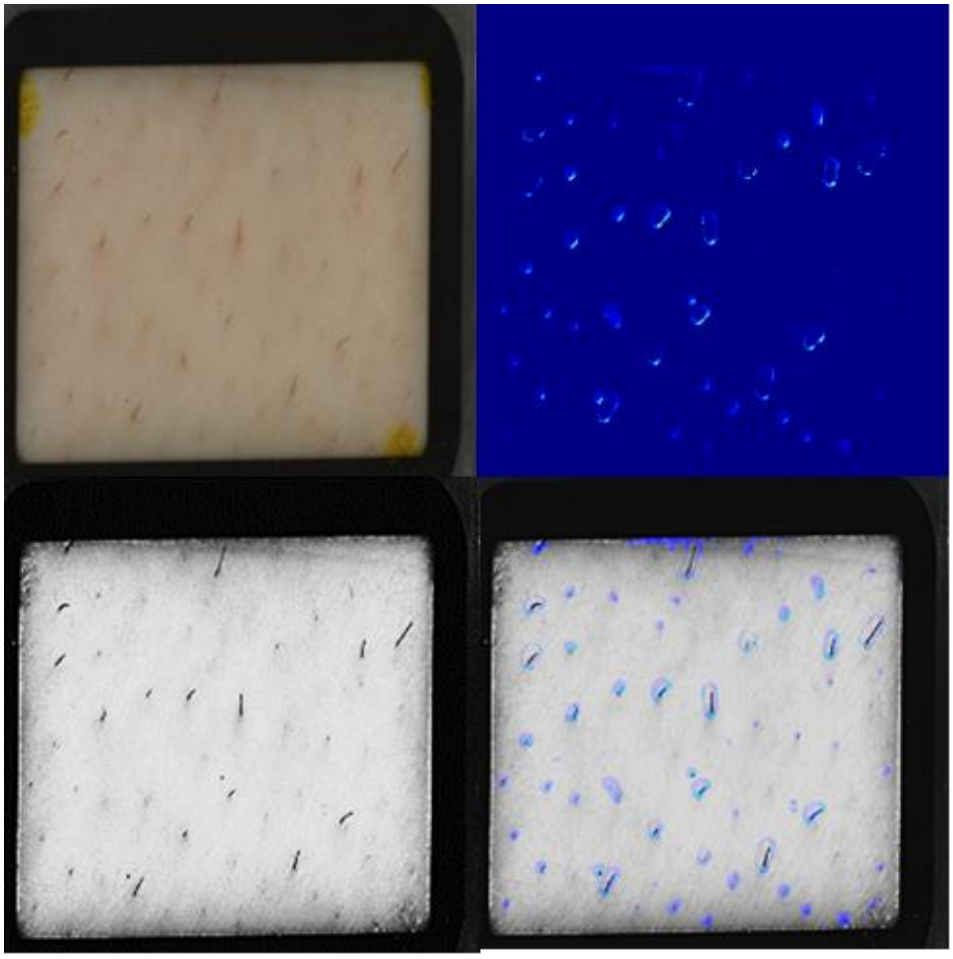


Figure 3.6. Left top: input image; left bottom: the luminance channel of the input image with enhanced visibility of hair; right top: the activation of the last u-net layer before summation; right bottom: the visualization of the activation superimposed as color overlay over the enhanced luminance channel of the input image.

for example by removing or reducing hue variations, might be beneficial since the current one is even lower than the baseline without augmentations.

A new research direction is found by investigating the heat maps generated by the last U-net layer, see Figure 3.6, which in our architecture has only one feature map. Observe that while the network was trained with the total hair count, the heat map shows that the network identifies and contours individual hair. We observe that the neuron activation on the contour of each hair fluctuates and does not sum up to one, as it should be for if an accurate hair count is to be achieved in the summation layer. In recent years, multi-task learning provided increased performance in several applications by training for different objectives. Therefore, we believe it is worth to investigate how to combine different tasks, e.g., hair detection, with our counting problem.

3.7. ACKNOWLEDGMENTS

The authors would like to thank Caroline Peters Yan Liu Lucja Bartula, Tim Tielemans, Evelyn Simons, Eva-Maria Rebernig, Birgit Krall and Jetten, Nadine for their contribution in the construction of the database and the collection of the annotations and Andrea Capra for the insightful discussions. A particular thanks for Binyam Gebre for the help in the design of the experiments and for the initial crucial support.

4. HAIR IMPACT ON SKIN LESIONS DIAGNOSIS

“Intelligence and education that hasn't been
tempered by human affection
isn't worth a damn.”
— Daniel Keyes, Flowers for Algernon

Recent work on the classification of microscopic skin lesions does not consider how the presence of skin hair may affect diagnosis. In this work, we investigate how deep-learning models can handle a varying amount of skin hair during their predictions. We present an automated processing pipeline that tests the performance of the classification model. We conclude that, under realistic conditions, modern day classification models are robust to the presence of skin hair and we investigate three architectural choices (Resnet50, InceptionV3, Densenet121) that make them so.

This chapter is based on the paper *Don’t Tear Your Hair Out: Analysis of the Impact of Skin Hair on the Diagnosis of Microscopic Skin Lesions.* authored jointly with Dmitry Znamenskiy, Nicola Pezzotti and Milan Petkovic, which was published in ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Science, vol 12661.

4.1. INTRODUCTION

Skin cancer is the most common cancer in the U.S. [114], and, the number of treated adults has increased over time, from 3.4 million in the 2002-2006 period to 4.9 million in the 2007-2011 period [115]. Usually, screening and diagnosis are primarily carried by clinical visual inspection and, if necessary, by biopsy. There is an urge to automate and facilitate this procedure with image-based screening since early detection is crucial for treatment options [53]. Deep convolutional neural networks (CNN) have demonstrated great potential for solving various vision tasks [2] and recently reached dermatologists performance in suspicious skin lesions classification [3]. To support early diagnosis, effective and computationally efficient models can be deployed on smartphone devices to enable a first level of patient-driven screening. In this setting, the models have to deal with much less controlled conditions than in a laboratory environment. In this chapter we investigate the effect that a varying amount of skin hair have on the classification accuracy of deep learning classifiers. We present an automated testing pipeline that, while currently focused on testing for the impact of skin hair, we plan to extend to other type of unforeseen circumstances, e.g., different camera models, light conditions and resolutions.

To perform such analysis, we first segment skin hair in images depicting small skin patches. These patches often contain skin lesions and we evaluate whether the learned segmentation helps to improve the robustness of the skin lesion classification. This work is motivated by the need to create and validate a screening approach that does not require removing skin hair. If feasible, such an approach has several

benefits in a professional healthcare setting, it can improve patient comfort, save time in the screening procedure, and improve diagnosis since the presence of hair can help differentiate skin lesions [116]. Moreover, it can allow the development of a first round of user-driven screening on, for example, smartphone devices.

The contribution of this chapter is twofold; we first present an approach for the segmentation of hairs in existing skin image patches. This segmentation approach gives us realistic hair patterns that can be used to test for the robustness of skin lesions classifiers. Our second contribution is a set of image augmentation strategies, based on our first contribution, that form a testing pipeline for the robustness of skin lesions classifiers. More specifically, for the first contribution we have developed an algorithm for the hair segmentation based on the state-of-the-art architecture for biomedical segmentation U-NET [106]. Since we needed data for training, which was difficult to find in open source datasets, we contributed to the enrichment of the public benchmark dataset HAM10000 [117] by annotating skin hair. Figure 4.1 shows eight examples out of 75 annotated images, now published

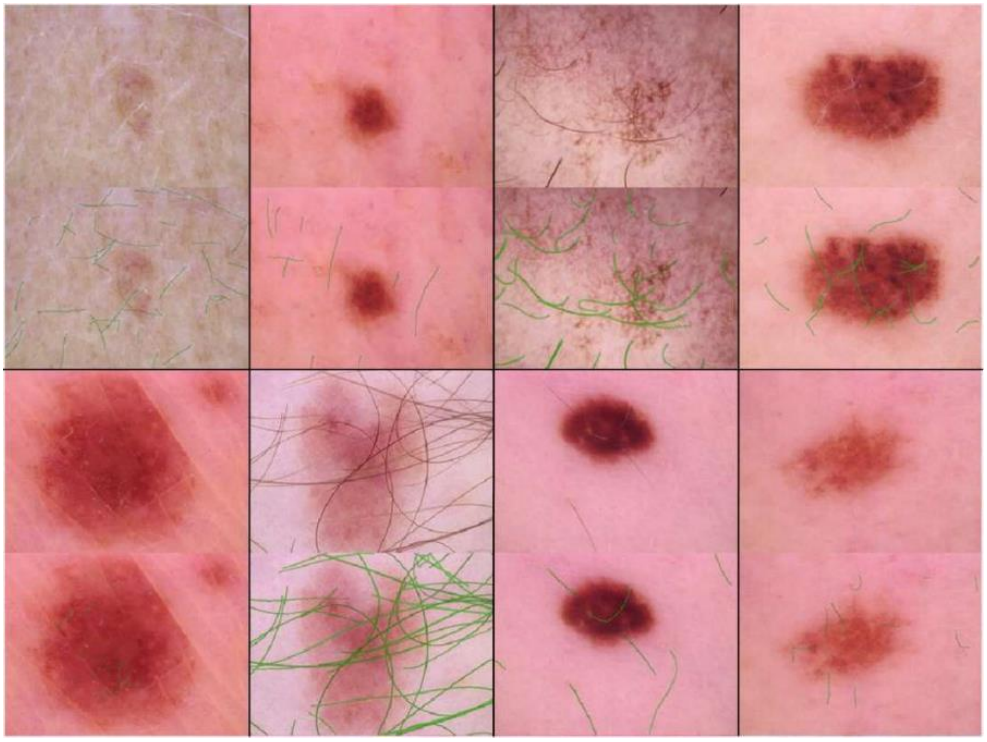


Figure 4.1. Eight skin hair annotation samples, each showing the original image (upper half) and the annotated version (lower half).

at https://doi.org/10.4121/uuid:9ed94e25-8b74-4807-b84a-2c54ec9d96f0. Note that the resolution of each image is of 600x450 pixels, highlighting the challenge of such a manual segmentation and the benefit that our segmentation network can provide to the research community.

As second contribution, we evaluate a new testing pipeline, relying on realistic image augmentations strategies, see Figure 4.2, for the skin condition classification task. We first consider a basic augmentation strategy with small rotation, sheer and scaling. Our main contribution, however, is to build on top of our segmentation approach, by adding realistic hair obtained from different skin images. In this approach, hereinafter referred to as “virtual hair transplantation”, we tested the addition of hair patterns in different positions, orientations and color. We test our method on the binary classification task of nevi versus melanoma with respectively 4522 and 12875 images taken from the datasets HAM10000 [117], BCN20000 [118] and MSK [119].

4.2. RELATED WORK FOR HAIR DETECTION

Multiple techniques are available to check and find hair on the human body. Most of them are relatively old, relying on the manual counting and traditional image processing techniques. The manual counting by visual inspection with a naked eye

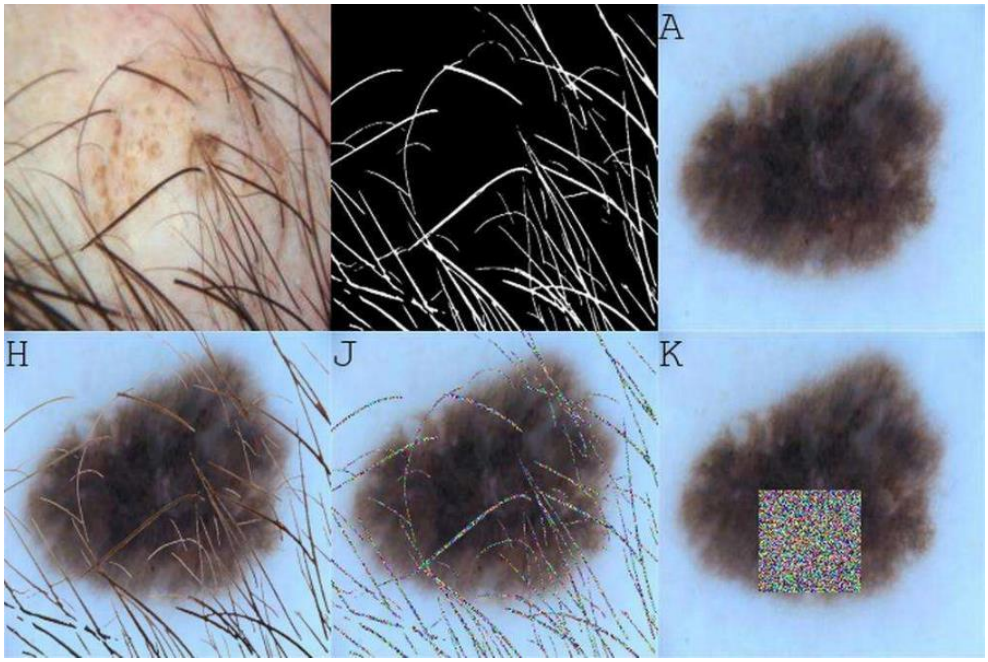


Figure 4.2. Top left is the input hair image. Top center the segmentation of hair. Top right an input image to be classified as nevi or melanoma. The bottom row shows the hair with their original color, transplanting hair with random pattern color and finally masking flat squares of equal area. We excluded scaling, sheer and rotation for visualization purposes.

or lens is the oldest, but still in use in many professional practices. The accuracy of this technique is naturally susceptible to loss of the instant local attention by the expert, which is influenced by tiredness, random gaze trajectory, and other psychophysical factors. Second, there are automatic systems present on the market like Chowis [120]. While Chowis mention the use of AI, they do not disclose the actual methods used in the products.

Other approaches relying on the traditional image processing have been used in the prior art to provide hair segmentation. Hoffmann [121] presents an automated system to detect hair loss and hair thinning conditions. Vallotton and Thomas [49] developed an approach for measuring body hair. They iteratively merge small line segments to account for curly variations. They only addressed cases in which the hairs are darker than the skin, though their algorithm can be easily modified to predict the opposite. In [50], [51], Shih and Lin proposed an unsupervised approach to count hair. The authors mainly used traditional computer vision techniques to detect lines. This approach, while good in a controlled image acquisition environment, suffers from different light conditions and perspective distortions. Besides, a slight change in the acquiring device implies the need for re-tuning the numerous parameters of the computer vision algorithm. Lim *et al.* [52], developed an automatic hair counting system to evaluate laser hair removal. They also validated their performance in

clinical trials. They collected images from the thighs of five volunteers with Fitzpatrick skin type III-IV. Their percentage error was <5% in each subject.

All the mentioned techniques may work on a dataset, but they are not easily generalizable and not easy to replicate, for example, when considering different skin types or hair colour. Since we were not aware of prior art on the use of neural networks for hair detection, we experimented with U-NET architecture which is widely used for biomedical image segmentation [122]. We did not try any other segmentation techniques or architecture since U-NET provided good hair segmentation sufficient for the virtual "hair transplantation" augmentation. Note that several hair segmentation techniques based on deep learning exist but are not focused on segmenting the individual hair but rather on the segmentation and color detection from frontal face pictures.

4.3. METHODS

In this section, we define the methodologies for the hair segmentation problem and for the skin lesion augmentations used to solve the binary lesion classification task.

HAIR SEGMENTATION

We have randomly selected 75 images to annotate from dataset [117] where we have first deleted all duplicates lesions. The implementation and definition of U-NET are taken from [123]. For the training of the network, we used the popular dice loss [124] (which is equal to 1 – dice coefficient) where we included a 'smooth' normalization parameter S = 0.0001 in the definition of the dice coefficient DSC, as presented in [125]:

$$\text{dice loss}\,(\mathrm{A},\mathrm{B}) = 1 - DSC = 1 - \frac{2|A \cap B| + S}{|A| + |B| + S},$$

where A is the binary ground truth annotation and B is the binarized prediction. We used the Jaccard index as the primary measure to evaluate the performance of similarity between binary label A and predicted segmentation map B:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

For each training strategy, the last transformation in the data generator is a spatial reduction with random crops of 128x128, which also removes, in the case of rotated images, the black margins. The input resolution of each image for this task is 600x450. After training, we generate the hair segmentation mask y for all skin lesions x in the combined datasets D.

AUGMENTATIONS

Using the hair segmentation mask, the hair can be copy-pasted from one image to another. Later, we make use of this mask in our testing pipeline to add hair from another image in the classification mask. Given an image x in the dataset D, we defined four different augmentation strategies to test for the skin lesion classification task.

The first strategy A is simply the identity or no augmentation, where the input image remains the same. The second is the basic augmentation strategy, denoted in the following by F where the input image x is randomly rotated by an angle $\theta \in [-20, +20]$, scaled by a factor between 0.8 and 1.2 and undergo a shear transform with parameter (0.05, 0.05). The three later strategies always add to F. Strategy H adds the transform that transplants hairs from image x_H with the corresponding hair mask y_H, where the image uniformly sampled from the dataset D. We define the strategy J as replacing x_H with x_R an image of the same size containing only random pixels. Finally, we define strategy K which adds, after augmentation F, a square containing random pixels with the size $e = sqrt(\sum y_H)$ where y_H is the hair segmentation mask of randomly chosen images from dataset D, so that the area of the square equals the total area of the segmentation mask. The above transformations can be notated as:

$$A: x = x$$
$$F: x = F(x)$$
$$H: x = F(x(1 - y_H) + x_H y_H)$$
$$J: x = F(x(1 - y_H) + x_R y_H)$$
$$K: x = F(x(1 - y_K) + x_R y_K)$$

where x_R is the image consisting of the random pixel values of the same size as x, and all multiplications are considered pixel-wise. We have to note that before applying the transplantation, we rotate the input image $x_H(x_R)$ by random $\theta \in [-180, +180]$ to get the rotation augmentation. One can see that the hair augmentation strategy 'H', when compared to J and K, is a smooth and soft way to perform a natural augmentation of the skin background in the skin lesion dataset. Moreover, this makes the prediction more robust to hair presence, as we will see below. The average area covered by hair is low and that less than 10% of images are covered by more than 5% of hair.

SKIN LESION CLASSIFICATION

As network architecture, we selected the state of the art Resnet50 [104] (achieves dermatologist level performance in [126]), InceptionV3 [127] (achieves dermatologist level performance in [6]), and Densenet [105] (best performing according to [128]). All the network are pre-trained on ImageNet [129] and fine-tuned for the skin lesion classification task. For the training with strategy H, in the augmentation phase as a hair source, we considered all input images. Figure 4.3 shows the distribution of the relative hair density in the test set. The hot pixels sum per image is computed as the sum of the segmentation before applying the binary threshold. We can

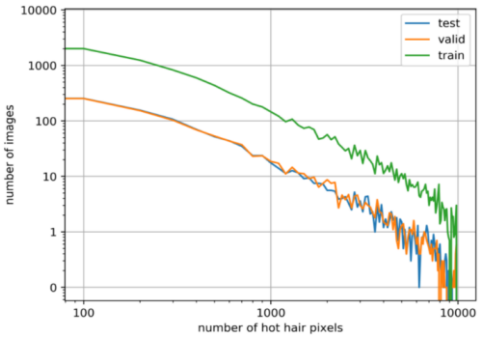


Figure 4.3. Average distribution of hot pixels highlighting hair in the ten-fold train, validation, and test set (images with resolution 224x224). On the y-axis the number of images and on the x-axis the number of hot pixels. Both axes are in log scale.

see that the average area covered by hair is low and that less than 10% of images are covered by more than 5% of hair.

We train all the networks for 100 epochs using the Adam optimizer [109] and learning rate 1e-4. For each combination of strategy and architecture, we randomly split D in train, validation and test set with each containing respectively 80%, 10%, and 10% of the images. We repeat all procedure 10 times to reach the final test set error.

4.4. RESULTS

In this section, we present the results for the two described problems: the preliminary task of hair segmentation and the evaluation of skin classification when applying the different augmentation techniques.

HAIR SEGMENTATION

Due to the limited dataset size, 75 annotated images, the hair segmentation task was relatively fast to accomplish but still resulting in a sufficient basis for the added augmentations. The parameters used for training are learning rate 0.01, epochs 500, batch size 8. The test set includes 12 images, and the resolution for testing is the original one. The trained U-NET achieves Jaccard value 0.51 with a discrete recall of 0.66 and accuracy 0.98. Therefore, the visual inspection, see Figure 4.4, offers promising results as a starting point to carry on the following skin classification analysis using hair augmentation. Cross-validation and a bigger dataset would be a possible next step to consolidate the hair segmentation results.

SKIN LESION CLASSIFICATION

The results presented in Table 4.1, overall shows the hair augmentation H does not essentially improve the performance of the models. Table 4.1 presents the average accuracy of ten repetitions of the experiment as introduced in the relative method section. In particular, H is not underperforming compared to the base F and the other color and shape pattern J, K. But when considering Densenet121 we see that H, J and K have positive regularization effect since they improve the accuracy compared to the baseline F. Apart from this difference, all three models show similar accuracies, with Resnet50 behind by only 1%.

Table 4.1. Test set accuracy metrics for the three different architectures and the four different augmentations plus no augmentation A. The standard deviation is the result of 10 iterations of the experiment. The total number of images in the set is 1740 with 516 average hot hair pixels.

	Resnet50	InceptionV3	Densenet121
A	0.886 ± 0.006	0.902 ± 0.009	0.905 ± 0.008
F	0.916 ± 0.010	0.926 ± 0.006	0.918 ± 0.008
H	0.916 ± 0.009	0.923 ± 0.007	0.922 ± 0.007
J	0.915 ± 0.009	0.922 ± 0.007	0.924 ± 0.008
K	0.917 ± 0.006	0.922 ± 0.005	0.923 ± 0.009

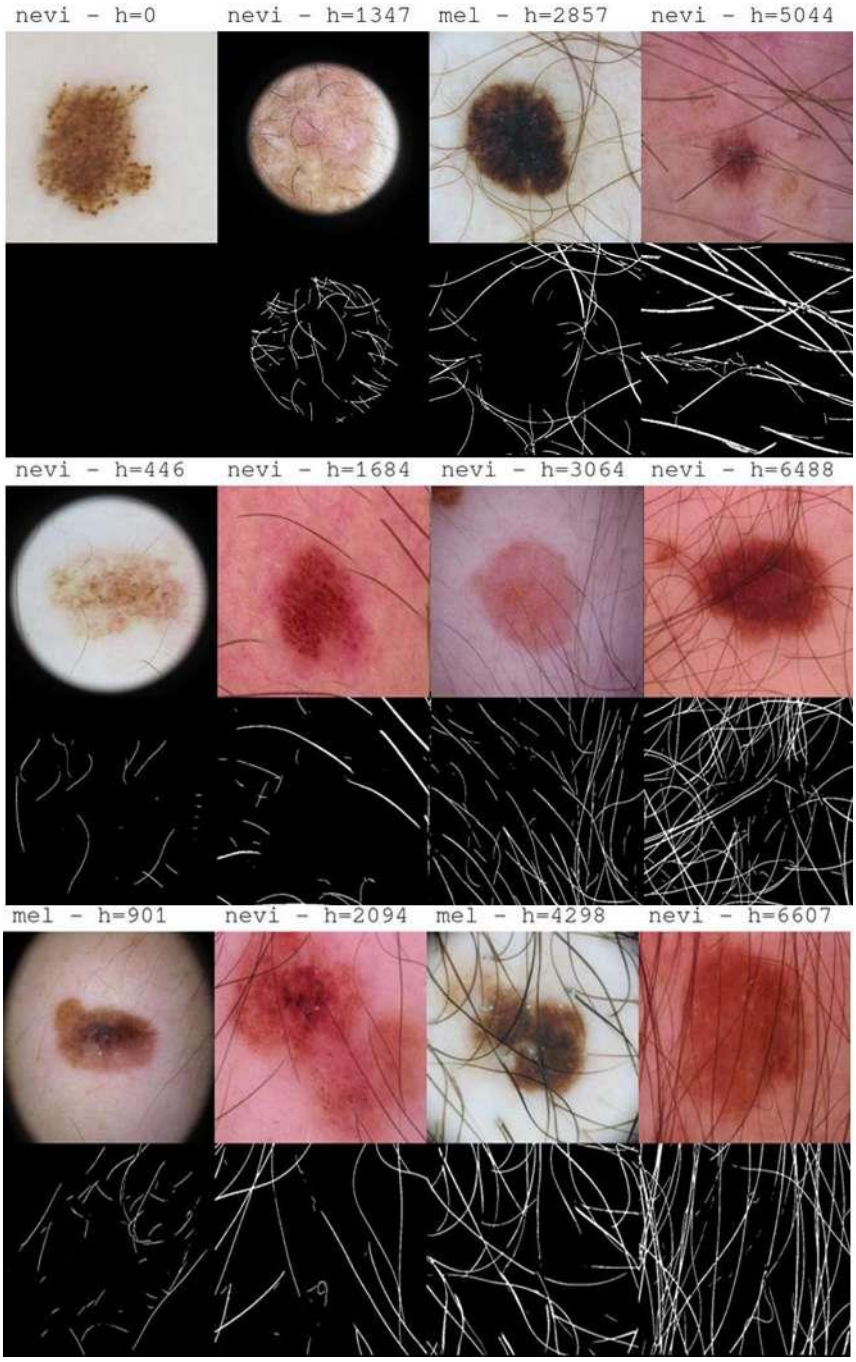


Figure 4.4. Examples of images with relative predicted hair. On top the label and the number of hot pixels in the 224x224 resized version of the input image.

Table 4.2. Same as Table 4.1 but considering only moderate to densely crowded hair images. Only images containing minimum 1500 hot hair pixels (average 3412 $\pm$ 79) with 153 images in total.

	Resnet50	InceptionV3	Densenet121
A	0.895 $\pm$ 0.018	0.902 $\pm$ 0.016	0.916 $\pm$ 0.016
F	0.924 $\pm$ 0.009	0.932 $\pm$ 0.011	0.933 $\pm$ 0.019
H	0.927 $\pm$ 0.009	0.927 $\pm$ 0.012	0.930 $\pm$ 0.019
J	0.922 $\pm$ 0.017	0.926 $\pm$ 0.024	0.933 $\pm$ 0.015
K	0.930 $\pm$ 0.014	0.930 $\pm$ 0.017	0.935 $\pm$ 0.011

Table 4.3. Same as Table 4.1 but considering only densely crowded hair images. Only images containing minimum 4000 hot hair pixels (average is 6011 $\pm$ 240) with 42 images in total.

	Resnet50	InceptionV3	Densenet121
A	0.920 $\pm$ 0.026	0.905 $\pm$ 0.024	0.932 $\pm$ 0.028
F	0.937 $\pm$ 0.035	0.956 $\pm$ 0.031	0.941 $\pm$ 0.027
H	0.951 $\pm$ 0.029	0.936 $\pm$ 0.030	0.940 $\pm$ 0.033
J	0.931 $\pm$ 0.028	0.923 $\pm$ 0.040	0.939 $\pm$ 0.030
K	0.941 $\pm$ 0.026	0.931 $\pm$ 0.035	0.936 $\pm$ 0.038

In Table 4.2, Table 4.3, we present the accuracy when considering only images with moderate and densely crowded skin hair presence. We believe that the increase in overall accuracy across compared to Table 4.1 is random. Even though the standard deviation increases, due to the lower number of images selected, the overall accuracy does not decrease when augmenting the images.

4.5. CONCLUSIONS AND FUTURE WORK

The results show that the presence of hair in skin images has little effect on the prediction of skin lesions. A practical consequence of this discovery can lead towards improving patient comfort and efficiency of screening, while opening the door to the investigation of a first-level screening performed by the user on mobile devices. We also show that the problem of hair segmentation on the skin images can be solved easily and robustly with deep learning.

For future work, we suggest to consider for the virtual ‘hair transplantation’ only images with sufficient hair density and consider improving the quality of the hair inpainting with skin color normalization which can reduce the visibility of the generated artefacts, or consider inpainting techniques based on the Generative Adversarial Networks, see [16], [130]. Finally, the proposed augmentation strategy of virtual ‘hair transplantation’ can be evaluated versus a more straightforward strategy where the real hairs are replaced with the random line patterns of the same area.

5. SKIN LESION GENERATION

"Every person must choose
how much truth he can stand."
—Yalom, I. (2005). When Nietzsche Wept.

Skin cancer affects more than 3 million people only in the US. Comprehensive microscopic databases include around 30 thousand samples, limiting the richness of patterns that can be presented to machine learning. To this end, generative models such as GANs have been proposed for creating realistic synthetic images but, despite their popularity, they are often difficult to train and control. Recently an autoregressive approach based on a quantized autoencoder showed state of the art performances while being simple to train and provide synthetic data generation opportunities. In the first part of this chapter, we evaluate the training of VQ-VAE-2 with different latent space configuration. In the second part, we show how to use a learned prior over the latent space with PixelSNAIL to generate and modify skin lesions. We show how this process can be used for powerful data augmentation and visualization for skin health, evaluating it on a downstream application that classifies malignant lesions.

This chapter is based on the paper *A latent space exploration for microscopic skin lesion augmentations with VQ-VAE-2 and PixelSNAIL* authored jointly with Nicola Pezzotti, Dmitry Znamenskiy and Milan Petkovic, which was published in SPIE Medical Imaging Proceedings 2021.

Figure 5.1. Synthetic skin lesions that do not exist in the real world. The lesions are generated in the latent space of a two-layer VQ-VAE-2 autoencoder. The codes are sampled from top left to bottom right using an autoregressive model. First, the top codes are generated, then the bottom ones conditioned on the top. Then, VQ-VAE-2 decoder reconstructs the image from the latent codes, which represent embedding quantizing vectors.

5.1. INTRODUCTION

Skin cancer is the most common cancer in the U.S. [114], and, the number of treated adults has increased over time, from 3.4 million in the 2002-2006 period to 4.9 million in the 2007-2011 period [115]. Usually, screening and diagnosis are primarily carried by clinical visual inspection and biopsy if necessary. There is an urge to automate and facilitate this procedure with image-based screening since early detection is crucial for treatment options [53]. Many state of the art Convolutional Neural Network (CNN) models performed on par, or better, compare to dermatologists: in 2017, Esteva *et al.* [126], trained a CNN that performed on par with 21 dermatologists and in 2019, Brinker *et al.* [3], [131], shows that Resnet50 [104] outperformed 136 of 157 dermatologists. However, core research is still needed to reach a fully automated diagnostic system which addresses the ethical and technical hurdles of deploying such models in the real world. A challenge is the lack of data since collecting real samples is time-consuming and difficult, for example, considering rare diseases. Another challenge is the fairness of the AI systems concerning underrepresented populations. Two common biases are the skin type—skin lesions datasets are skewed towards light skin types—and body location acquisition—most lesions come from easy to collect places such as limbs or torso and few from less conformable areas. While there are many reasons for the nature of such biases, some of them difficult to solve by simply collecting more data, it is possible to hamper their effect by creating synthetic examples covering the underrepresented class.

In the past years, generative models have improved very fast and significantly. The growth in popularity is related to the visually appealing results and the continually increasing computational power commonly available. The synthetic samples are difficult to spot compared to the original data even from a careful human inspector [55]. The state of the art models are called Generative Adversarial Networks (GANs) [54], and they rely on two competing architectures, one generating images starting from noise, the other trying to distinguish synthetic from real samples. The mapping from a noise vector to the sample coupled with the adversarial training produces very realistic images [132], but it is often challenging to train and control afterwards.

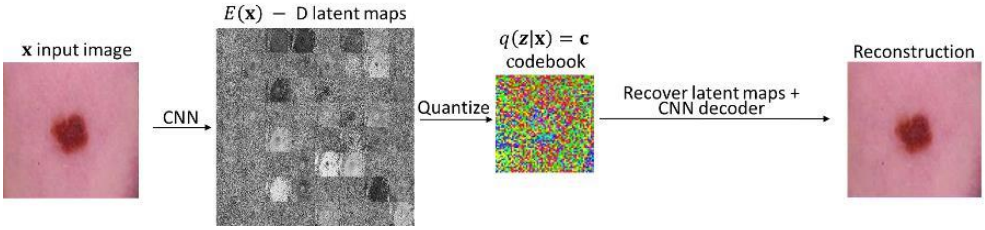


Figure 5.2. Encoding and quantizing a single image with a single layer vector quantizing autoencoder. The image is first encoded with a fully convolutional encoder into D latent maps and then quantized using K quantizing vectors.

Recently, a different approach has proven similar results to GANs on image generation tasks. This approach, use an autoencoder Vector Quantized Variational Autoencoder (VQ-VAE-2) [56] in combination with an autoregressive model called PixelSNAIL [57], and it is able to generate high resolution realistic pictures as shown in Figure 5.1. The autoencoder finds a compact quantized representation of the input data using stacked VQ-VAE [133] and represents an input image with a lower resolution map of codes. Since this representation is given at lower resolution, these codes represent different image patterns and, in the case of the hierarchical implementation, patterns at different scales. VQ-VAE-2 relies on feedforward networks, thus easy to train and, according to [56], it does not suffer that much from the common pitfalls of GANs, such as the mode collapse or lacking full support over the input distribution [134]. Moreover, it easily allows for the explainability of the generative model, as each code in the latent space can be seen as a self-supervised label extracted by the model. An example of a single layer autoencoder is presented in Figure 5.2, where the input image is encoded and quantized into discrete codes. Once learned the quantized latent representation, the PixelSNAIL autoregressive model is used to learn the prior distribution over the discrete latent codes and to generate new realistic images. Key advantages of this approach are the explainability of the features and the potential for integrating advanced augmentation techniques. Figure 5.3 shows the input to the VQVAE-2 model, and the corresponding codes in the two hierarchical levels (here color-coded with a randomized palette).

In this chapter, we apply the VQ-VAE-2 architecture to generate novel skin lesions, with and without the autoregressive model, to augment and increase the number of images in the lesions' datasets. While VQ-VAE-2 model has been proven to perform well in different domains like ImageNet [135], as shown in the ML community by A. D'Amour *et al.* [136] it is often a challenge to replicate ML pipelines in different domains, especially for medical images. This is one reason we present an exploration of the latent space generated and the behavior and hurdles we faced selecting the right latent space dimensions. For example, once trained with the original double-layer configuration, we report the top latent space's collapse leading to a one-layer autoencoder.

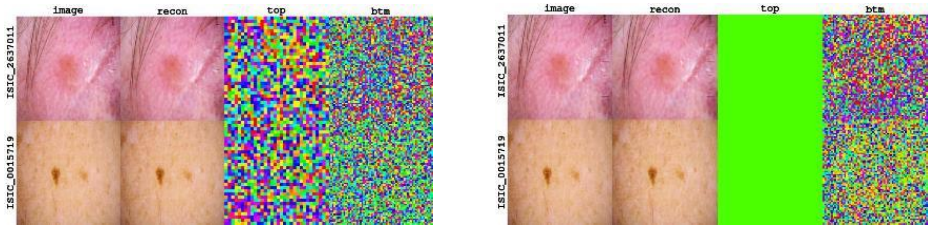


Figure 5.3. Left a model with latent space dimension $(K = 256, D = 8)$; Right a model with latent space dimension $(K = 512, D = 64)$, and the corresponding discrete embeddings, for 2 different images. Each row is a different image. The first column is the input image, the second the VQ-VAE-2 reconstruction. The third column presents the top encoding and the fourth the bottom encoding. Each colour in the third and fourth column represents a different integer code. The encoder quantizes each pixel, of $D = 8$ dimensional feature map, into one out of $K = 256$ codewords. The model on the right presents a single color in the top space since the top hierarchy is collapsed.

5.2. RELATED WORK

Several papers investigated the use of generative models in the context of skin lesions applications. In 2019, Ghorbani *et al.* [137] used Pix2Pix GAN to synthetize skin condition. Style transfer generative was used to enhance lesion segmentation [138] and other versions of GANs, such as LAPGAN, DDGAN, PPGAN has been tried in [139][140]. While all these examples provide visually appealing results, they suffer from the classical problems of GANs mentioned above. Another similar report of data augmentation [141] tried to hamper this using Self-Attention and PPGAN. In this chapter we investigate applications of autoregressive models, in combination with quantization based autoencoders. We moreover investigate how promising is to use such methods to augment skin lesion datasets.

5.3. VQ-VAE-2 AND PIXELSNAIL APPROACH

This section introduces another generative approach used in the following three chapters and not based on GANs. We present a basis for the approach by showing the VQ-VAE and VQ-VAE-2 setups and then we introduce the autoregressive models focusing on PixelSNAIL. In the last part of this section, we present other works which employ generative models, such as GANs, to create novel skin lesions.

VECTOR QUANTIZED VARIATIONAL AUTOENCODER

The VQ-VAE model is introduced by van den Oord *et al.* [133] and it builds on top of the Variational AutoEncoder (VAE) [142], [143] by generalizing ideas from classical image compression methods like jpeg. Given a dataset of observations $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}\}$, the goal of a VAE is to learn, without supervision, a lower dimensional representation in terms of latent variables $\mathbf{z}$. It is composed by an encoder E, which map the input image into latent variables, and a decoder, which reconstruct the image from the compressed representation. In other words, the

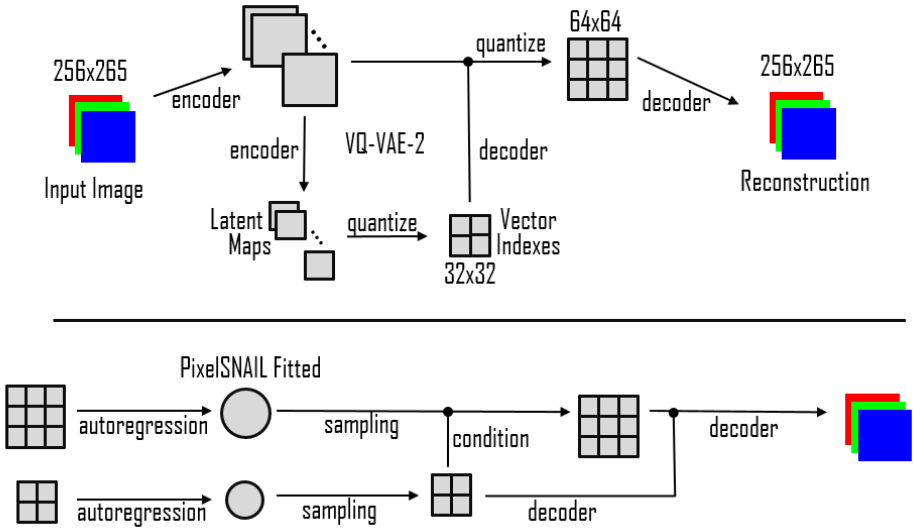


Figure 5.4. Method's flowchart. The input image is feed the VQ-VAE-2 autoencoder. The PixelSNAIL later learns a prior over the latent space, sample novel synthetics codes, decode them into the new geometric images, and, subsequently, to 3D scans.

decoder network models the joint distribution $p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$ while the encoder models the posterior distribution $q(\mathbf{z}|\mathbf{x})$.

In the VQ-VAE framework the prior distribution is based on K prototype latent vectors $\{\boldsymbol{e}^{(1)}, \boldsymbol{e}^{(2)}, \ldots, \boldsymbol{e}^{(K)}\}$ of dimension D which quantize the latent maps $E(\mathbf{x})$, generated by the encoder. There are exactly K different latent vectors to choose from, so each pixel on the latent maps is represented with the nearest quantizing vector, as shown in Figure 5.2. In Razavi *et al.* [56] the VQ-VAE-2 is presented, which is the upgrade of the VQ-VAE to include multiple hierarchical layers which provide different quantize codebooks at different hierarchies. An example with two-level is presented in the top part of Figure 5.4 where the processing flow of the autoencoder is depicted in a flowchart style from left to right. The image is encoded in quantized latent maps for the top and bottom levels. The decoder then reconstructs the image using the latent maps conditioning the higher levels, which have smaller resolution, to the bottom ones. In the original setup the input 24 bit image with resolution 256x256 was reduced to 64x64 bottom map and 32x32 top map with $K = 512 = 2^9$ different quantizing vectors of dimension $D = 64$. In [56] the authors present two layer network trained on ImageNet [129], and three layer network trained on FFHQ [144], for generating high-resolution photo-realistic facial images. In order to solve large scale dependencies, which are usually difficult to capture by the autoregressive decoder, Fauw *et al.* [145] successfully explored the possibility to use many layers encoder. While in another research work, Williams *et al.* [146] used Hierarchical Quantized Autoencoders for image compression purposes.

In our experiments we focus on a two layers hierarchy with input resolution of 256×256 and relative latent maps of dimension 64×64 and 32×32. We also call the relative quantized vector indices, or codebook, as $\boldsymbol{c}_B \in \{0, K\}^{64 \times 64}$ and $\boldsymbol{c}_T \in \{0, K\}^{32 \times 32}$ as shown in Figure 5.3. We first answer the question of which K, D are better in our small dataset and then we perform a data augmentation and test the impact on the classification task of predicting malignant melanomas.

AUTOREGRESSIVE MODELS AND PIXELSNAIL

Besides the analysis of the resulting latent space, we investigated the generative capabilities of the autoregressive model, PixelCNN [147] with self-attention [148], called PixelSNAIL [57]. In this setup the autoregressive model can efficiently model the prior distribution of the latent codes, creating photo-realistic synthetic images. The idea behind the PixelCNN model is to learn the conditional distribution of the given sequence of random variables. When applied on the latent space, the latent codes of the whole image are sorted from top left to bottom right to predict the next code value, which is a discrete probability distribution over the K codes, in an autoregressive fashion. In our example, the autoregressive model learns the join distributions of the latent codes on the top layer and then the distribution of the bottom codes conditioned on the top codes. There are different options to generate new samples once the two models are trained. The main approach is to perform a sampling of the top space $\boldsymbol{c}_T$ trained on specific image class label, and then sample the bottom space trained on the same label, while conditioning it on the sampled top codes. Figure 5.4 in the bottom part shows that two models are fitted, one for the top and one for the bottom space, for later sampling and decoding new synthetic images. An example of this approach is shown in Figure 5.1 where the two autoregressive models are trained on K=256, D=8 with all other hyperparameters equals to the original implementation in [11].

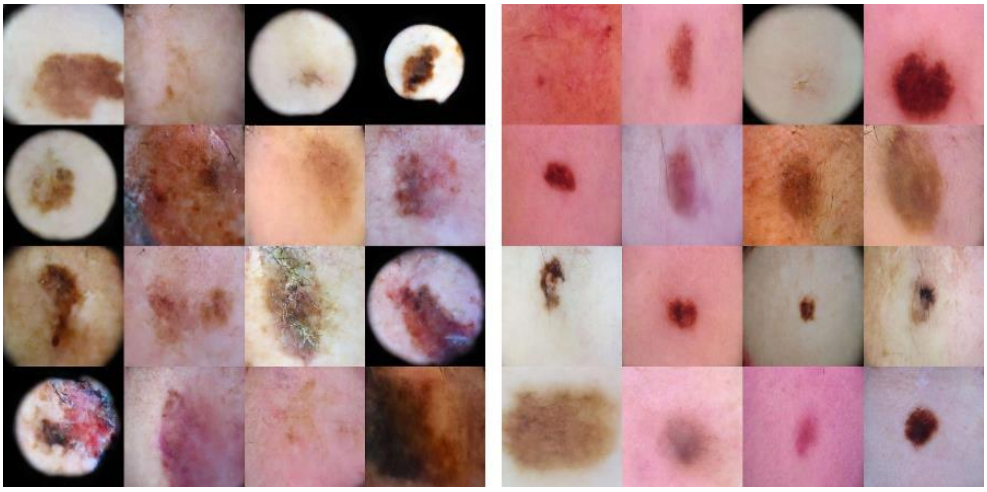


Figure 5.5. synthetic images generated by training the autoregressive model only on 4922 melanomas on the left and training only on 17685 nevi on the right.

5.4. DATASET, MATERIAL AND METHODS

In this chapter, we investigate the behavior of VQ-VAE-2 in combination with PixelSNAIL, applied to the skin lesion datasets 2020SIIM-ISIC [14], HAM10000 [117], BCN20000 [118] and MSK [119]. The analysis includes an extensive exploration of the impact of the model hyperparameters, as well as experiments to adopt the model as a way to augment the data for downstream tasks. For our experiments, we merge all datasets and then remove all duplicates images by using an EfficientNet pretrained model[2]. The total number of images includes 52302 benign and 4922 malign images (57224 in total). We split the dataset, stratified according to patient, into train (45718) and validation (11506) set to evaluate the reconstruction of images unseen by the network during training. We derive our PyTorch [110] implementation of the VQ-VAE-2 and PixelSNAIL from an openly available project[3]. Since we were not able to find or implement the class conditional sampling suggested in [11] we simply train one autoregressive model for nevi and another for melanoma, while keeping the "true class conditional sampling", with one PixelSNAIL model, for a future work. We present some examples of generated samples in Figure 5.5, where on the left we used the model trained only on melanomas while on the right it was trained only on nevi. The training images were obtained by cropping the center squared region and by scaling them to fit the 256x256 resolution used in our model.

Table 5.1. Report of three different experiments with $K = 512$. Each row is a different model trained with the same exact hyperparameters. The two rows are the latent space dimensions K and D. $|\mathrm{unique}(\boldsymbol{c}_T)|$, $|\mathrm{unique}(\boldsymbol{c}_B)|$ | represent the number of, top and bottom, codes used for encoding the whole dataset while $|\mathrm{unique}(\boldsymbol{c}_T, \boldsymbol{c}_B)|$ are the cooccurrences of used codes. The metrics are the mean squared error for validation set.

K	D	$\lvert\boldsymbol{c}_T\rvert$	$\lvert\boldsymbol{c}_B\rvert$	$\lvert(\boldsymbol{c}_T, \boldsymbol{c}_B)\rvert$	MSE
512	32	1	512	512	0.0025
512	64	8	512	4095	0.0030
512	64	1	512	140	0.0040

PRIOR LATENT SPACE DIMENSION

In the original setup, the hyperparameters $K = 512$ and $D = 64$ were used for both hierarchical layers trained on ImageNet [129], and the three layer model, trained on FFHQ [144]. Our first research question is what the best configuration of (K, D) is for the latent space, given that we now consider a much smaller dataset with respect to the two mentioned above. To understand the impact of the hyperparameters on the dataset, we did not use data augmentation techniques and worked solely on training data as-is. In our simulations we experimented in reducing the number of vectors K without decreasing the quality of the reconstructed image, in particular, considering that the autoregressive approach is noticeably slow when sampling new images [149], making it hardly scalable for real world applications. This analysis is motivated

[2] https://www.kaggle.com/shonenkov/merge-external-data; https://www.kaggle.com/shonenkov/dbscan-clustering-check-marking

[3] https://github.com/rosinality/vq-vae-2-pytorch

by the observation that some instances of the network, trained on the natural skin lesion images, resulted in a collapse to a single layer autoencoder. This is shown in the right part of Figure 5.3, where the reader can see that the top quantized map degenerates to a single code when we used the original configuration $K = 512$ and $D = 64$. We believe that, when reducing K, we also reduce the risk of the code collapse at the top layer which in turn allows for a higher number of code combinations on the top and bottom layers and, therefore, a better sampling and performance of the autoregressive models. This behavior is also confirmed by Table 5.1 where each row is a trained model and the columns are: the number K of latent vectors, the dimension D of each vector, the number of effectively used vectors in the top hierarchy $|\mathrm{unique}(\boldsymbol{c}_T)|$, the number of effectively used vectors in the bottom hierarchy $|\mathrm{unique}(\boldsymbol{c}_B)|$, the number of cooccurrences of vectors in the top and bottom space $|\mathrm{unique}(\boldsymbol{c}_T, \boldsymbol{c}_B)|$, and lastly the mean squared error for the validation set. It can be seen that, even if the top layer collapses in the model of the first row, it has a competitive MSE for the reconstructed image. We believe this is due to the tradeoff between a large latent space and the relatively small dataset compared to ImageNet. In other words, with some probability the method encodes images without taking advantage of the hierarchical model. Once this is established, we can refrain to use high values of K and D together and gain training and inference speed, which we need for the computationally demanding PixelSNAIL. A benefit for the community using this approach would be to find the optimal latent space dimension, according to their required use, in order to increase throughput speed when coming to the generation part.

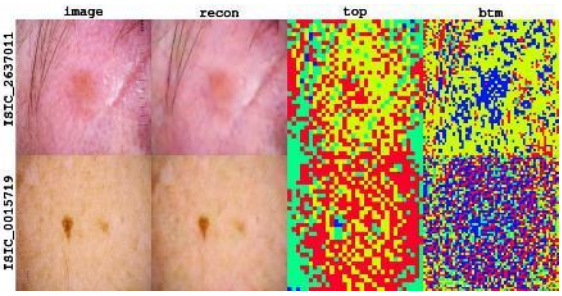


Figure 5.6. $(K = 4, D = 8)$ — Discrete embeddings for 2 different images. Each row is a different image. The first column is the input image, the second the autoencoder reconstruction. The third column presents the top encoding and the fourth the bottom encoding. Each colour in the third and fourth column represents a different integer code.

To understand better the models, it is relevant to visualize its latent space expression. For example, one can directly spot the richness of patterns (Figure 5.3 – left), or vice-versa the collapse of one hierarchy, see (Figure 5.3 – right). It is observed that, sometimes, the model extracts semantic regions, acting as a loosely defined segmentation model. This effect would be particularly interesting when clustering similar codes pattern after training or even during the training phase applying mask to objects of interest. In the following section, we explore the potential of code replacement in the latent representation for the generation of augmented, but realistic, training samples. While the top latent space, which undergo 8x compression of the image into D latent maps tends to encode low frequency information there is not always a clear boundary between it and the bottom space. In fact, as presented in the previous section, sometimes the information is encoded only in the bottom space and other instead use only a part of the top space. A

possible way to visualize the codes is by color-coding the input space, but with 256 colors is very difficult that any pattern emerges simply looking at those. A different situation arises when considering very small K. For this reason, we trained also a very small set of VQ-VAE-2 with $K = 4$. While such models do not provide high resolution reconstruction, they are still deeply helpful to test the model behaviors. In Figure 5.3 and Figure 5.6 respectively the latent space of three models is visualized. When using a small number of quantizing vectors, it is easier to visualize directly emerging patterns in the code space. Direct manipulation of the codes in the latent space as a way for creating and modifying lesions is possible. However, it is not the desired approach since complex relationships that arise in the pixel domain due to the interaction of the codes in the latent maps. Intuitively, the codes have an overlapping receptive field in the underlying images. Therefore, the code resulting image patches are not given solely by the corresponding latent codes, but by the interaction between the neighboring ones. In the next section, we provide evidence of this behavior, we present several examples of manipulations of the latent codes and demonstrate that, by using specific manipulations of the codes based on the autoregressive models, it can be used for data augmentations.

DATA AUGMENTATION

In this section, we present various techniques to augment the dataset using the VQ-VAE-2 and manipulations of its latent space. The power of having multiple layers in the VQ-VAE-2 architecture, is that one can modify an image by manipulating the latent codes only in one layer, for example the bottom one, retaining the top layer which usually encodes global structure information. Without learning any prior over the latent space one can already compose novel images by "mixing" codes from different images at the cost of losing spatial consistencies. This idea is similar to the pseudo-labeling [150], where to generate new labels and augment the dataset multiple images part are mixed together. An example of mixing skin lesion is presented by Perez *et al.* [128], where after learning a segmentation model, they mix foreground and background of different lesions. Here, instead of mixing codes based on regions, as shown in [151], we mix them by hierarchy of the learned VQ-VAE-2. This procedure can be achieved in many ways even without the autoregressive model. Given one input image $(\mathbf{c}_T, \boldsymbol{c}_B)$ we replace the bottom codes before reconstruction given another image with the same label, i.e. malignant melanomas or benign lesions. In Figure 5.7 an example of mixing top and bottom codes is

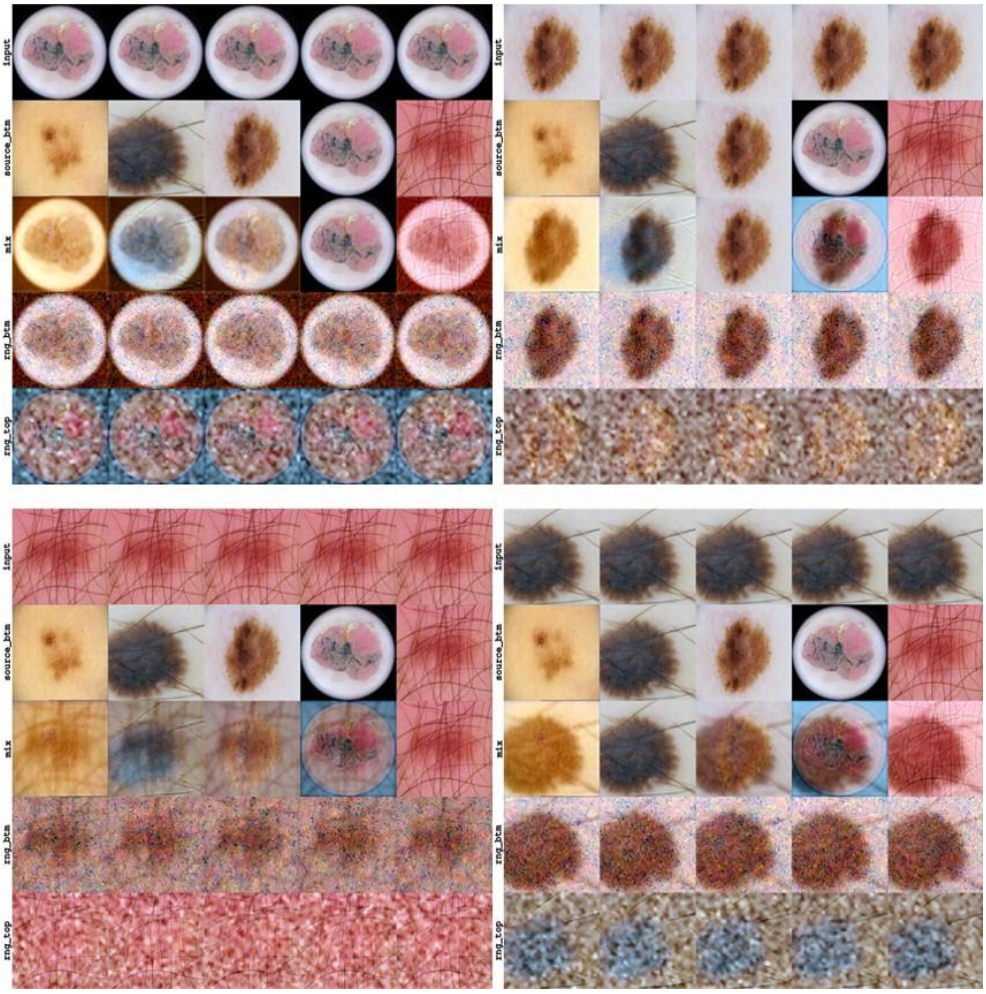


Figure 5.7. Examples of augmentations by mixing latent codes for the $K = 256, D = 8$ model. The top row represents the input image used to generate the top codes, while the second presents the one used for the bottom code. The third row presents the reconstruction obtained when the two codes are mixed. The bottom row is used for comparison and is created by randomly sampling the bottom codes, showing that a random replacement of the codes is not a viable solution and highlighting the relevance of the interplay between neighboring codes.

presented without considering the actual ground truth label. In the first row, the top code image is selected to be mixed with the second row, which is the source for the bottom codes. In the third row the mix of the first two rows shows that the global structure is mostly retained from the top source for example looking at the black circular box or the lesion geometry. On the contrary, the high frequencies patterns, like skin hair, are taken from the bottom source image. We highlight such behavior in the second to last bottom row (rng_btm) of the figure where the bottom codes are resampled randomly creating a noisy pattern but maintaining the global structure of the input source top image. On the other side, when the top codes are resampled

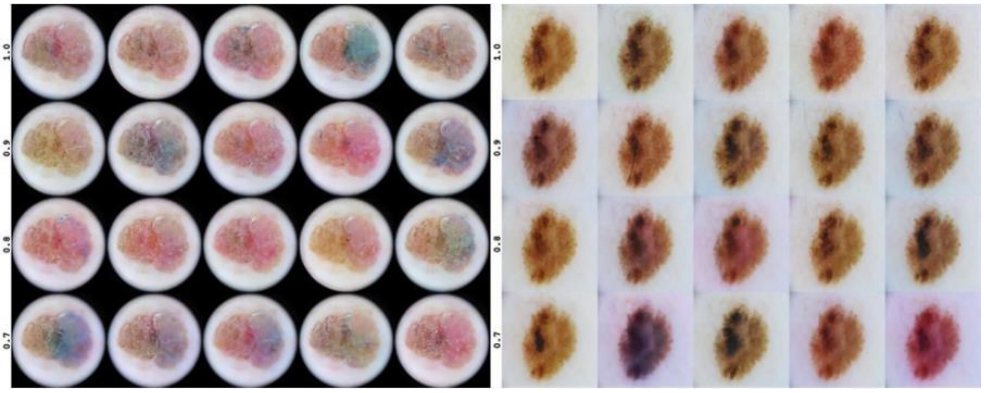


Figure 5.8. The same image, which original version can be seen in Figure 5.7, is modified by resampling the bottom codes. From the top row to the bottom row the temperature parameter goes from 1.0 to 0.7. The temperature flattens out the probability distribution before sampling according to a multinomial distribution. A lower temperature allows for more out of distribution, less realistic, samples.

randomly, as presented in the last row (rng_top) of the figure, the global structure is partially lost while it is easy to see the hair like patterns. It is clear by looking at the bottom two rows that this approach destroys the information and makes the resulting image not useful for any realistic task, but it provides a baseline behavior to compare other augmentations and to understand how much information is lost and which is most relevant for example when classifying lesions.

Another approach is to resample the bottom space by using the autoregressive model. Given an encoded image one can use the real top $\mathbf{c}_{\mathrm{T}}$ while sampling a new bottom encoding $\boldsymbol{c}_B^{new}$ and decoding using the original decoder network. This will lead to the same global low frequency structure while updating high frequencies encoded in the bottom space. An example of such behavior is presented in Figure 5.8, where two images augmented by resampling multiple times the bottom latent codes without conditioning on the diagnosis. While we expected a more unstable behavior, the change in skin tone does not ruin the quality of the image from an inexperienced human observer perspective. Also, to be observed is that the skin tone is consistent over the whole image for each resampling. This means that the model can represent very long relationships between pixels as also pointed out by

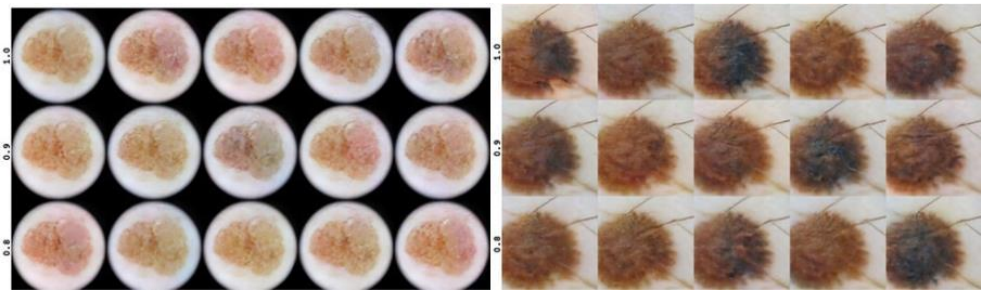


Figure 5.9. The images, which original version can be seen in Figure 5.7, are modified by resampling the bottom codes $\boldsymbol{c}_B$ using an autoregressive model trained only on Nevi.

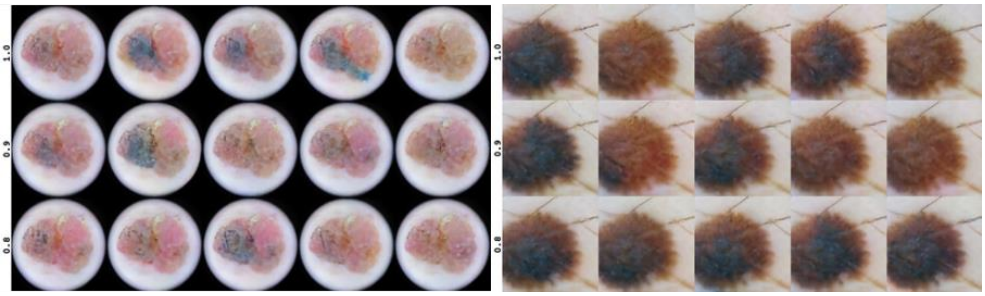


Figure 5.10. The images, which original version can be seen in Figure 5.7, are modified by resampling the bottom codes $\boldsymbol{c}_B$ using an autoregressive model trained only on Melanomas.

the original authors of the VQ-VAE-2 paper. The difference in each row lies in the temperature parameter, which flatten out the estimated probability distribution before sampling dividing the PixelSNAIL decoder's output by it before applying Softmax and sampling according to a multinomial distribution. While in Figure 5.8 we considered the diagnosis, and the pictures seems to have wide variety on the chrominance, when using models trained only on one particular diagnosis the pattern changes considerably. For example, when using the ABCD rule [152], a commonly used but easy to learn tool to judge lesions according to Asymmetry, Border, Colour and Diameter, multiple changing colors are usually features of Melanomas while a uniform color is characteristic of Nevi. In fact, in Figure 5.9 the model train only on nevi try to resample the bottom codes to match nevi-like global appearance by giving a uniform skin lesion color. On the contrary, in Figure 5.10 the model, trained only on Melanomas, prefers darker and changing colors as characteristic of Melanomas. While we appreciate this model behavior, we don't expect that it is what dermatologists want to see but simply observations of a particular dataset coupled with an augmentation technique. Finally, the completely synthetic images when training autoregressive model on a subset of the input data, 4922 melanoma and 17685 nevi, are shown respectively in Figure 5.5 left and right. We can see that the nevi seem to be sharper in details, and this can be due to the differences in the limited number of samples. Future work can increase the number of images or augment them prior to fitting the autoregressive models.

In the next section, we present preliminary results in which we use augmented data to enrich the training for a downstream task like the classification of skin lesions malignancy.

5.5. EXPERIMENTS AND QUANTITATIVE EVALUATION

In this section we report quantitative experiments for the respective methods presented above. We first present several results obtained by training various VQ-VAE-2 by changing the number of quantizing vectors and the number of latent codes. We begin by showing that *K* cannot drop much lower than 256 without impacting the MSE metric for the reconstruction. At the same time, it is difficult to spot big differences by human eyes even when considering only 64 latent quantizing vectors. We then show the results when the augmentations proposed in the previous section are used to train a downstream classification task. The experiments show that the performance of the model is impacted by the use of the synthetic images, hinting at the limitations of the approach. However, by showing that the impact on the performance is limited, we demonstrate the potential in the limited-data regime, where synthetically labeled data can be beneficial.

Table 5.2. Report of different experiments K and D. Each row is a different model trained with the same exact hyperparameters but K and D. $|\mathrm{unique}(\boldsymbol{c}_T)|$, $|\mathrm{unique}(\boldsymbol{c}_B)|$ | represent the number of, top and bottom, codes used for encoding the whole dataset while $|\mathrm{unique}(\boldsymbol{c}_T, \boldsymbol{c}_B)|$ are the cooccurrences of used codes. The metrics are the MSE for validation set.

K	D	$\lvert\boldsymbol{c}_T\rvert$	$\lvert\boldsymbol{c}_B\rvert$	$\lvert(\boldsymbol{c}_T, \boldsymbol{c}_B)\rvert$	MSE
256	8	256	256	64605	0.0024
512	32	1	512	512	0.0028
256	64	256	256	64687	0.0029
512	8	1	512	512	0.0029
128	8	128	128	16370	0.0030
512	64	8	512	4095	0.0030
256	32	47	256	10664	0.0031
256	64	1	256	256	0.0034
512	64	1	140	140	0.0040
128	64	1	128	128	0.0042
128	32	13	96	1248	0.0047
128	64	9	128	1152	0.0055
4	8	4	4	16	0.0137
4	32	2	4	8	0.0174
4	64	4	4	16	0.0175
4	64	2	4	8	0.0198

PRIOR LATENT SPACE DIMENSION

First, we explore the training of VQ-VAE-2. A first relevant question is which configuration of the latent space to select according to target dataset and its use. It is not clear a priori if one wants to use a single layer, two or even three hierarchical layers and what are the pros and cons of such a choice. Moreover, one can increase and decrease the latent dimensions by fixing the number of layers or the number of filters. The objective is to find the optimal values of K and D such that the model is not yet collapsing and giving a visible improvement in MSE. A similar exploration with smaller K, 2 or 4 codework, but with different objectives is carried on in [22] where the authors benchmark different lossy compression proposing a new scheme of Hierarchical Quantized Autoencoders.

The results of this training phase are presented in Table 5.2. Note that we trained twice the same configuration with 64 latent maps and in one model there was full collapse of the top space while in the other full use. Another relevant result is that, while there is no clear order between D, it seems easier for the model to learn with a small D $(= 2, 8)$. The model trained with $K = 128$ were performing worst in terms of metrics and visual inspection compared to $K = 256$, reaching the maximum valid

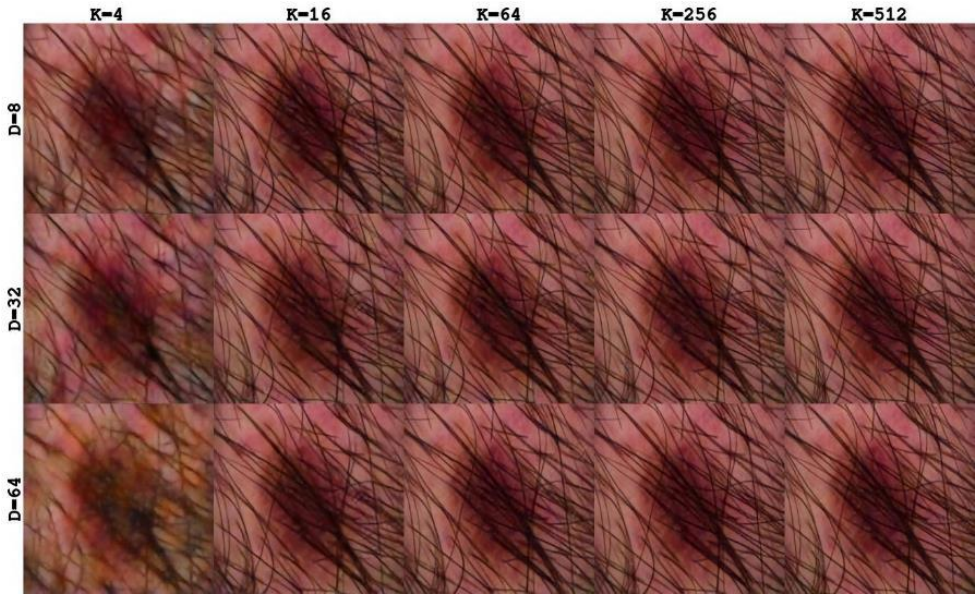


Figure 5.11. Comparison of VQVAE trained with different of latent codes K. All the rows show the reconstructed image ISIC_0068279.jpg. The visual quality of the reconstruction degrades very slowly, and big artefacts are clear only when using very small K.

score of 0.0026. We report here also very small configurations since we will display such models for visualization purposes and, therefore, understanding the behavior of the model in this extreme case. To compute the co-occurrences in column $|\mathrm{unique}(\boldsymbol{c}_T, \boldsymbol{c}_B)|$ we simply up-sampled $\boldsymbol{c}_T$ to have the same resolution as $\boldsymbol{c}_B$, using a nearest neighbor interpolation.

By inspecting the reconstruction in Figure 5.11, we observe that also when using small number of quantizing vectors, e.g. $K = 16$, it is difficult to observe big artifacts and only when reducing markedly the number of quantizing vectors, e.g. $K = 4$, the quality is distorted for this 256×256 resolution. On the contrary, the model is able to find the relevant features even when using few latent maps. Both the visual inspection and the results support this hypothesis; hence, we resolve to select $D = 8$ latent maps and $K = 256$ quantizing vectors as a latent space to fit the autoregressive model.

DATA AUGMENTATIONS

For the following section we only consider the model with K=256, D=8, we investigate how the model can be used for augmenting and manipulating the data. The model was chosen as a good tradeoff between training stability and richness of representation.

In order to evaluate the results, we trained a fast, yet powerful, EfficientNet [153] model pretrained on ImageNet to use it as a scoring function for the augmentations. Similarly to Classification Accuracy Score (CAS) [20] we train the classifier replacing the real dataset with the generated samples and compare the results with the

reconstructed version, passing through the autoencoder. The difference between real and reconstructed is what is lost in the lossy compression. Using such approach for scoring proves the quality of the samples directly in a relevant task avoiding the perplexities generated by other metrics. To facilitate and speed up the analysis, particularly when considering the cost of training PixelSNAIL for the generating of new images, we consider only Nevi (17685) and Melanoma (4922). We train the model for 50 epochs with ADAM [109] and learning rate 0.001. Prior to the training phase we split the dataset into train (80%) and test (20%) stratified according to patient. We use, as validation metric, the Area Under the Curve instead of the accuracy as it is also used as the primary metric in the ISIC2020 challenge.

Table 5.3. Metrics when replacing input dataset with Reconstruction of autoencoder, mixing images with same diagnosis, random bottom, synthetic novel images with autoregressive decoder matching diagnosis, resampled bottom codes according to diagnosis, random top codes.

Training set	AUC
real	0.934
reconstruction	0.920
mixing	0.893
synthetic	0.798
resample($\boldsymbol{c}_B$)	0.759
rand($\boldsymbol{c}_B$)	0.752
rand($\boldsymbol{c}_T$)	0.698

The results, presented in Table 5.3, shows that there is a loss in performance when no real data is used. More specifically, we observe a small difference of 0.14 points between “real” data and “reconstruction”, same images passed through the autoencoder. This shows that, although some reduction in performance is unavoidable due to the introduced domain shift, this is minimal and can open up new applications where data is synthetically generated.

When the top and bottom codes are mixed, we see a further drop in performance compared to the reconstructed and real images. The performances of the autoregressive models are lower, and completely “synthetic” images reaches 0.798 AUC, while resampling only low frequencies “resample($\boldsymbol{c}_B$)” achieves 0.759 AUC. The latter, while for human eye seems to produce realistic samples, it performs poorly since it is similar to randomly replacing the bottom codes. The worst result occurs when replacing top codes with random ones hinting that the model takes decisions more on high-frequency details rather than global structures.

5.6. Discussion and Conclusions

There are still many hurdles in generating and controlling high resolution medical images. Overcoming these issues can provide several benefits for training machine learning models, especially when limited labeled training data is available. In this chapter we presented VQ-VAE-2 as an alternative to GANs in the context of skin lesion analysis. We provided an exploration of the hyperparameter settings, as well as novel ways to augment the data based on the VQ-VAE-2 model and the manipulation of the latent space, including the use of an autoregressive model. We also showed that the generated images are not competitive on a downstream task, hinting a limitation of the methods driven by the inability to capture fine grained structures in the images and an introduced domain shift.

Several further research directions can be investigated in the future. We believe one interesting direction is to exploit the hierarchical structure of the autoencoder directly to separate patterns. By this, we suggest constraining a quantization only on a particular region of the image, for example, using a segmentation network. In this way the augmentation process will be streamlined and facilitated by selecting and sampling only relevant features for each layer. Moreover, an interesting direction is to modify the loss function used for training, for example by encoding the downstream task directly into the autoencoder training. Finally, we would like the integration of semantic information directly into the training of the model, for example by using unsupervised segmentation models.

To conclude, we explored the use of VQ-VAE-2 to generate skin lesions, performing a detailed analysis of the resulting latent space and performing an extensive hyperparameter analysis. Then, we investigated how an autoregressive model, called PixelSNAIL, can be used to generate synthetics lesions. We presented several possibilities to create these synthetic skin lesions and we evaluated them by training a classifier on real data. The qualitative results prove the methods effective for microscopic skin lesions generation while being relatively easy to train and control. The quantitative results prove the work is promising but cannot outperform, or perform similarly, to classifiers training on real data. However, we believe that our investigation provides relevant information to devise methods to train machine learning model for skin lesions in the low-data regime.

6. 3D FACES GENERATION

“The attempt to escape from pain,
is what creates more pain.”
— Gabor Maté

The realistic generation of synthetic 3D faces is an open challenge due to the complexity of the geometry and the lack of large and diverse publicly available datasets. Generative models based on convolutional neural networks (CNNs) have recently demonstrated great ability to produce novel synthetic high-resolution images indistinguishable from the original pictures by an expert human observer. However, applying them to non-grid-like data like 3D meshes presents many challenges. In our work, we overcome the challenges by first reducing the face mesh to a 2D regular image representation and then exploiting one prominent state-of-the-art generative approach. The approach uses a Vector Quantized Variational Autoencoder VQ-VAE-2 to learn a latent discrete representation of the 2D images. Then, the 3D synthesis is achieved by fitting the latent space and sampling it with an autoregressive model, PixelSNAIL. The quantitative and qualitative evaluation demonstrate that synthetic faces generated with our method are statistically closer to the real faces when compared to a classical synthesis approach based on Principal Component Analysis (PCA).

This chapter is based on the paper Generating High-Resolution 3D Faces Using VQ-VAE-2 with PixelSNAIL Networks authored jointly with Dmitry Znamenskiy, Nicola Pezzotti and Milan Petkovic, which was published in Image Analysis and Processing. ICIAP 2022 Workshops 228–239.

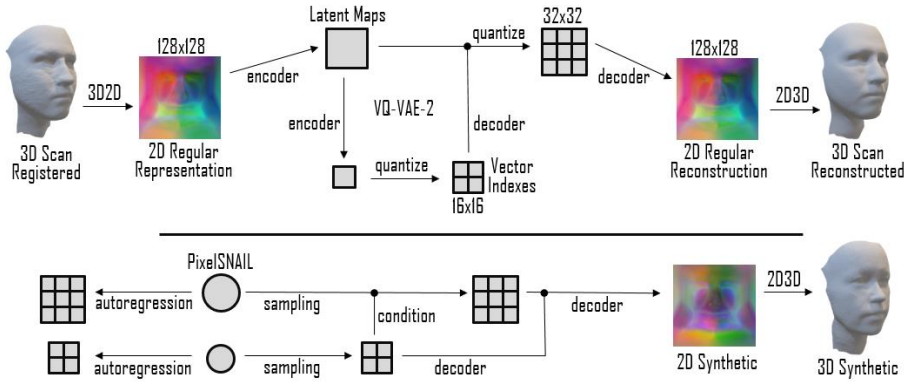


Figure 6.1. Method's flowchart. The registered 3D scans are first converted into a regular 2D image to feed the VQ-VAE-2 autoencoder. The PixelSNAIL later learns a prior over the latent space, sample novel synthetics codes, decode them into the new geometric images, and, subsequently, to 3D scans.

6.1. INTRODUCTION

In the last two decades, the applications of virtual 3D models have risen exponentially. Two factors behind the rise are the growth in computational power and the economic benefit derived by simulating physical phenomena employing 3D models. Today, 3D face models serve many fields including animation of faces [154]–[156], recognition of expression [157], and face recognition [158]. For example, realistic faces generation is important for crowd generation in virtual reality environments [159]. However, the applications are often limited since face data contains privacy and sensitive information that reduces or blocks data sharing and aggregation from multiple sources. This poses a limitation to the generation of realistic 3D faces which can be used in several contexts.

To overcome such limitations, we propose to replace the original dataset with a synthetic replica and present a compelling solution to the generation of synthetic 3D faces using a machine learning approach. As a first step, we follow the seminal work of Blanz and Vetter [4] and register a common reference 3D template into every scan bringing all raw scans in full correspondence with a common parametrization. The parametrization is insufficient to generate synthetic scans since the registered template has thousands of highly correlated vertices. Generation methods should consider this correlation by finding a low-dimensional decorrelated surface representation. Thus, Blanz and Vetter [4] reduced the vertex coordinates to a small number of decorrelated scores with a data-driven approach using Principal Component Analysis (PCA). Sampling new scans with PCA is then straightforward; however, interpolating in the reduced PCA subspace will not always result in natural human shapes due to the linear nature of the method. To overcome the drawbacks of PCA and similar linear methods, deep generative models based on Convolutional Neural Networks (CNNs) are employed to capture more complex non-linear interactions in the data. The current state of the state of the art advances in the field

of geometric deep learning [58] leverage the power of CNNs by adapting them to work on meshes [59], [60]. Graph convolutions, however, restrict the resolution, and therefore, the accuracy and amount of surface details of the 3D template.

By choosing the 3D mesh template with vertices connected as a 2D grid, our approach processes 3D meshes as 2D images. This makes the 3D2D mapping straightforward, enables 2D image synthesis methods, and avoids the challenges of graph convolutions [160]. Figure 6.1 shows schematically the solution we adopted, which is defined by two phases: in the first phase, see the top part of the figure, the method learns a discrete latent image representation given by a two layer quantized variational autoencoder VQ-VAE-2 [56]; then, in a later stage shown in the bottom of the figure, a powerful autoregressive network PixelCNN [147], [161] with self-attention [148], called as PixelSNAIL[162] is used to fit the latent space and sample from it. By employing such a novel pipeline, we empirically found that our method gives more natural 3D shapes compared to the PCA-based one. Due to the high variability of plausible 3D human shapes, the subjective evaluation is not enough to properly assess the quality and diversity of the generated face. A major challenge is defining a proper surrogate measure that evaluate how “human” is a 3D scan. In the literature, two metrics are commonly used to evaluate 3D scans: specificity [163] measures how close a scan is to the (training) set and diversity [164] measures the difference between a pairs of scans. In our work, instead of reporting a single number generated by such metrics, we compare their empirical distributions of the synthesized scans versus a test set of real faces. This allows us to measure the realism and diversity of the generated faces in terms of a previously unseen test dataset. The remaining chapter content is structured as follows: the next section gives an overview of the prior art, section 6.3 describes the approach for the synthetic generation, in the fourth section, we present the experiments and the relative quantitative evaluation. Lastly, we conclude and give acknowledgments.

6.2. RELATED WORKS

In the following, we present various approaches to synthetic head generation. Many works of research still rely on linear models [165] or multilinear models [166] due to their simplicity and due to the expansion of 3D Morphable Models [167]. Tran *et al.* [168] proposed a robust CNN-based approach to regress the PCA scores from pictures for face recognition and discrimination. In another work, the multilinear models are used to transfer facial expression and have the ability to animate faces [166]. While being simple and easy to train, they do not consider the input geometry. Additionally, a review of current methods regressing and sampling PCA scores is beyond the scope of chapter.

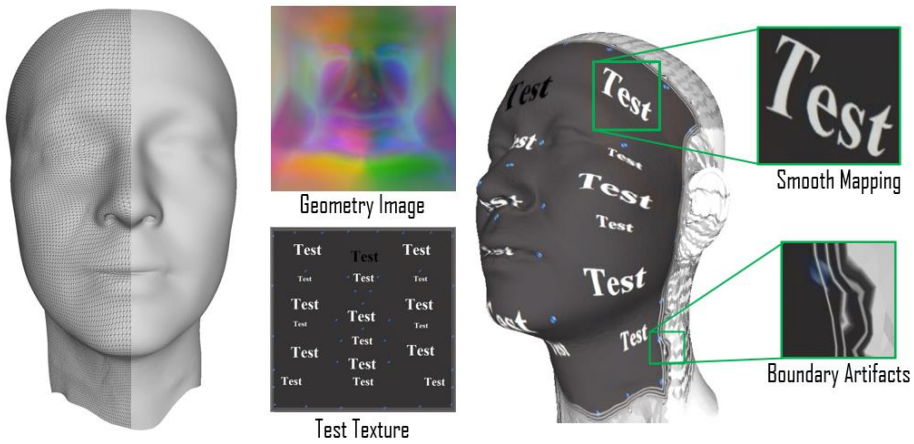


Figure 6.2. Template. The face template on the left has a regular (triangular) mesh grid – for visualization purposes the template is half rendered surfaces and half mesh. On the middle top column, the so-called 128×128 geometry image for the facial template, which is naturally derived from the regular structures of the template mesh, and a test texture to visualize the smoothness of the UV map. On the right the test visualization with example of boundary artifacts of the UV map.

3D TO 2D REPRESENTATIONS

Many 2D representation methods originate from the solution of the rendering problem which relies on so called UV maps to map 2D texture image on a 3D object. The UV maps, by definition, provide a bijective mapping from the 3D mesh triangles to their images on the texture image. In 2002, Gu *et al.* [169] showed how to optimally cut a surface and sample it over a regular 2D square grid generating the so-called geometry image. The problem might be more straightforward for a face geometry since the UV maps can be created by warping of the 3D templates with a regular grid. Booth *et al.* [61] presented a list of possible optimal implementations. Figure 6.2 shows the selected regular template geometry for this paper on the left, the geometry image derived from the grid and a test texture on the middle, and the texture rendered on the template on the right. As shown in the picture, the main drawbacks of such methods are the artifacts around the cuts or borders. However, in this example, such artifacts do not conflict with our requirements for an accurate face model and not a full-head one.

3D FACE GENERATION WITH GANS

The most common generative models employ Generative Adversarial Networks (GANs) [130]. Abrevaya *et al.* [164], investigated the use of Wasserstein GAN [170] to generate novel 3D faces with the ability to control and modify their expression. However, in our work, we directly map the input surface into a geometry image since we believe their fully connected generator cannot efficiently handle the complexity of the shapes. Slossber *et al.* [171] and the extension work in Shamai *et al.* [172], similarly to our work, converted the 3D into a 2D regular representation through non-rigid registration techniques. In Moschoglou *et al.* [173], the template was mapped using a cylindrical unwrapping as introduced by Booth and Zafeiriou [61]. While using

similar concepts, our work does not use adversarial training, a key difference that makes our method easier to train and less affected by the so-called *mode collapse* which affects GAN architectures.

3D FACE GENERATION WITH AUTOENCODERS

Apart from generative GANs models, recently, many works have overcome the linear modeling limitations by using VAEs [142]. For example, Bagautdinov *et al.* [174] modeled the face using a multiscale approach for different frequencies of details. Fernandez Abrevaya *et al.* [175] exploited the power of CNN-based encoder by coupling it with a multilinear decoder. In Li, K. *et al.* [176], a multi-column graph convolutional networks is designed to synthesize 3D surfaces. They first applied a spectral decomposition of the meshes and then trained multiple columns of graph convolutional networks. While these methods are similar to our approach, they also differ as no one uses the quantized autoencoder with an autoregressive network. Moreover, they do not convert the data into 2D geometry images but directly feed the registered 3D scans.

6.3. METHODS

The definition and registration of the face template are out of the scope of the current chapter. Conceptually, we have followed the method explained in Blanz and Vetter [4] [6] and have morphed all scans by means of non-rigid registration methods [83], [177]. A more detailed description of our parametric models is reported in Chapter 2. Since the face template already has a grid structure of 128×128 vertexes, we apply a vertex-based normalization to map the range of input values into interval $[0, 1]$, and therefore, to facilitate the follow up processing with the neural networks. The mapping to $[0, 1]$ also facilitates the normalized xyz facial data visualization of the as rgb (geometry-) images, see Figure 6.2 for an example. The range parameters for each grid vertex were retained for denormalizing the synthetic images into 3D shapes. The neural network approach VQ-VAE-2 with PixelSNAIL used in the current chapter is presented in the previous chapter, section 5.3.

METRICS FOR QUANTITATIVE EVALUATION

Our goal is to provide always realistic synthetic samples, and to achieve it, we visually inspected the generated scans and selected suitable metrics to prove this statement. The main idea is to prove that synthetic scans are statistically indistinguishable from a test set of original scans excluded from training. Before computing the metrics, the scans need to have identical parametrization corresponding to the 3D template. The identical parametrization enables a simple distance metric between a pair of scans, defined as the Root Mean Squared Distance between the corresponding pairs of vertices, after the rigid alignment of one scan to another [113].

We have employed two derivative metrics used in the literature to evaluate synthetic scans. The first metric is called diversity and has been introduced by Abrevaya *et al.* [164] with the aim to produce a single number measuring the heterogeneity of a set of scans. The diversity is defined as a distance between a random pair of synthetic

scans. In our work, we compare the empirical distribution of the diversity of 250 generated scans with the empirical distribution of diversity in 250 original scans from the test dataset. The second selected metric is called specificity and is defined in Davies *et al.* [163] for a scan as the minimal distance to the scans in the training dataset. Similar to the diversity distribution, we evaluate the empirical distribution of the specificity over 250 synthetic scans and compare it with the empirical distribution of specificity in 250 original scans from the test dataset.

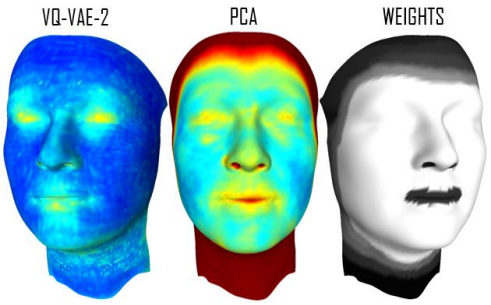


Figure 6.3. Test set reconstruction mean absolute weighted error on the left using VQ-VAE-2 and on the middle using PCA. The meshes jet color-code ranges from blue 0.0mm error map to red 1.0mm. On the right the used vertex weights are shown with color-code that ranges from 0.0 black to 1.0 white.

6.4. EXPERIMENTS

Within this work, we have considered two datasets of registered scans already available at Philips Research the *SizeChina* dataset of 3D head scans [82], and the *CAESAR* of full body scans [36], [178] where only the head was extracted. The above data gave us more than 5000 registered 3D face templates. We augment the face dataset by performing a symmetric reflection over the y-axis. While in this application we assume that asymmetries are normally distributed on the left and on the right of face we do not know if this is true. Nonetheless, we still believe this augmentation does not hamper the results of the approach. The dataset was split stratified according to participant id into train 90%, validation 5%, and test set 5%. We tested and computed the metrics only on the test set without considering the augmentations. For the sake of experimental reproducibility, we did not perform any other augmentation neither in training nor in test time. However, we believe further realistic augmentations would impact and consolidate the results. Nevertheless, we also notice that the current set of scans is enough to achieve the desired outcome of statistical indistinguishability from the test set.

In our experiments we focus on a two layers VQ-VAE hierarchy with input grid resolution of 128×128 and relative latent maps of dimension 32×32 and 16×16. We follow the approach described in chapter 5 to find the best combination of $K = [64, 128, 256\ 512], D = [2, 4,\ 8, 16, 32, 64]$ and found out that, according to reconstruction error, $K = 512$ is always better than smaller values. Vice versa for big enough K we notice that smaller dimension of D provides the best outcomes. Hence, we used $D = 2$ for our final VQ-VAE-2 model. We also reduced the batch size to 32 compared to the original implementation for both the autoencoder and the autoregressive model. The reconstruction root mean squared weighted error per participant is 0.29mm compared to 0.97mm for PCA and is mostly accumulated in the areas with higher curvature or with low weights such as eyes, mouth, nostril, and neck, as shown in Figure 6.3. We use vertex weights to improve the results in three different situations: to facilitate the registration of the raw scans, to maximize the

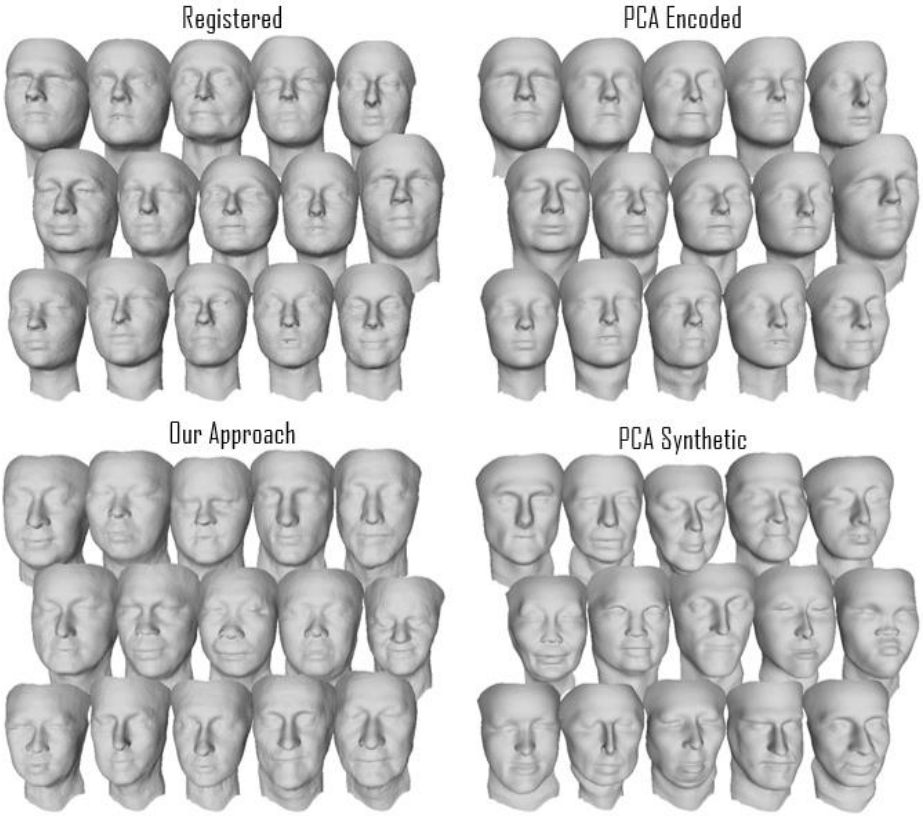


Figure 6.4. Example of facial scans (without selection). A batch of registered scans (top left), same scans encoded (top right), synthetic scans generated with our approach (bottom left), and PCA synthetic scans (bottom right).

PCA encoding energy in the face area of interest, and within the quantitative metrics, reported in the next section, to focus the attention of the metrics on more important facial areas of the model. Figure 6.3 shows the errors maps per vertex across the test set population: as expected the neural network reconstruction outperforms the smaller and linear PCA model – where we encoded the registered scans in "only" 200 principal components as done in chapter 2. To sample a new scan with PCA, we decode a 3D scan from the random PCA scores, where each score was sampled independently from the respective marginal empirical Cumulative Density Function (eCDF) computed over all PCA encoded scans. The above procedure guarantees that the synthetic PCA scans inherit the eCDF from the original data. Concerning the PixelSNAIL autoregressive model, we used the original configuration for ImageNet apart from the batch size, 32 in our example, and total number of epochs, 420 for both top and bottom hierarchy. The autoregressive models' validation accuracies in predicting the latent codes after 420 epochs are 0.87 for the top space and 0.91 for the bottom one. All the models were trained on PyTorch [110] with the same

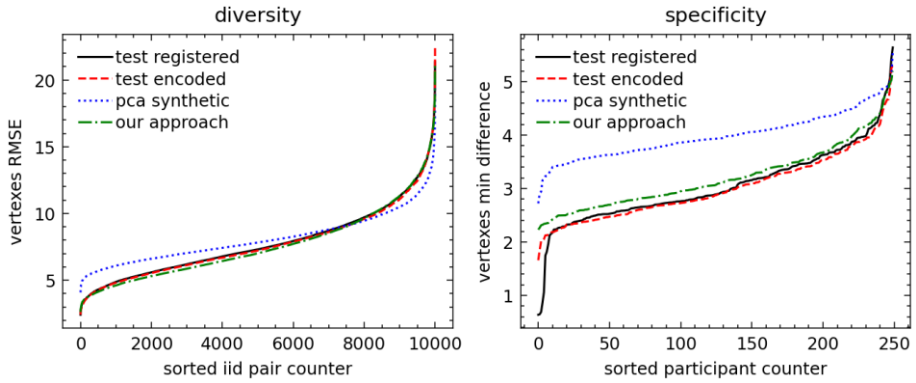


Figure 6.5. Quantitative metrics. On the left, the diversity is plotted for each i.i.d. pair and shows that the PCA distribution is "flatter" as excepted by the linear method. On the right, the specificity (the minimum distance versus the training set is kept) shows that our approach is much closer to the test set. Moreover, the specificity also proves that we do not replicate the input training scans since the minimum distance is markedly above 0mm – with our approach above 2mm.

hyperparameters as in the original implementation (excluding the one explicitly mentioned above).

Comparing the scans by visual inspection is not a trivial task and often not an objective metric. However, we believe that it is possible to spot some differences in the shape distribution between the different sets by looking at the overall scan's appearance. We present some scans randomly selected in Figure 6.4: the top right shows registered scans with the relative PCA encoded version on the top left; the bottom scans are synthetic and generated with our approach on the left and with PCA on the right. The PCA ones present more variability, or, in other words, more shapes differences compared to the other sets. The synthetic scans generated with our approach, from a visual inspection, present similar shape variability to the original scans compared to the PCA synthetic. Nevertheless, this is not enough since our approach may simply replicate or clone the original training data. We test these hypotheses in the following quantitative analysis proving that the synthetic scans are novel and different from the original training ones.

QUANTITATIVE EVALUATION

We have analyzed 2D representations for registered raw versus PCA-encoded vertices and computed empirical distributions for specificity and diversity metrics. Given V the vertices of a synthetic scan its specificity S is defined has

$$S(V) = \min_{t \in T} \left[\frac{\sum_{i=1}^{N} w_i \| v_i - v_i^t \|_2^2}{\sum_{i=1}^{N} w_i} \right]^{\frac{1}{2}}$$

where t is the index of the training set, $N = 128 \times 128 = 16384$ the total number of vertices $v_i \in V, w_i \in W$ are the weights for the i^{th} vertex as shown on the right of Figure 6.3. The diversity D of a pair of scans with vertexes $v_i^1 \in V^1, v_i^2 \in V^2$ is defined as

$$D(V^1, V^2) = \left[\frac{\sum_{i=1}^{N} w_i \|v_i^1 - v_i^2\|_2^2}{\sum_{i=1}^{N} w_i}\right]^{\frac{1}{2}}.$$

The empirical distributions presented in Figure 6.5 show that our approach results in synthetic faces which are statistically close to the original scans in the test set, unlike the PCA-based method which shows a flattened diversity distribution and higher specificity. The figure shows that our approach closely follows the distribution of the test-registered and -encoded scans, both in terms of diversity and specificity distributions. Moreover, since the specificity is computed against the training scans, we demonstrate that the faces generated by our approach are diverse from the training set since they do not collapse to zero and have a minimum distance above 2mm. The higher specificity of the PCA-synthetic scans also confirms the qualitative evaluation, see example scans in Figure 6.4, that PCA-based faces have more extreme characteristics.

6.5. Discussion and Conclusion

We presented a novel approach that can generate high-resolution synthetic 3D scans that combine traditional 3D parameterization approaches with the recent VQ-VAE-2 and PixelSNAIL deep learning based generative models. Our approach does not require the parametrization of the 3D face model and can be directly applied to registered templates, hence, allowing for a richer generation domain since synthetic scans can be outside the PCA linear sub-space. However, the major contribution of our work is that our method strictly outperforms the linear PCA classical approach and generates realistic high-resolution scans. We consider this only a first step in proving the validity of this approach – future work will perform a benchmark versus other state-of-the-art generative models. One main challenge is the lack of a clear quantitative metric to judge whether a scan belongs to the “real” class since the proposed diversity and specificity metrics may not be enough to capture all the relevant shape information. Additionally, while we believe the two selected metrics are suited for the current evaluation of realistic human faces, different metrics can be developed in the future. A natural extension of our approach that can partially solve the metrics problem could combine the 3D shape synthesis with the photo-realistic texture synthesis adding the rgb to the xyz channels within the 2D representation.

6.6. Acknowledgments

We thank Philips Research for providing access to the datasets of facial scans and software resources to manage the high-resolution parametric models.

7. 3D BODY GENERATION

“Vulnerability is the birthplace of innovation, creativity and change.”
— Brene Brown

Modelling and representing 3D shapes of the human body is a prominent field due to its applications in healthcare, clothes design, virtual fit rooms, and the movie industry. The realistic generation of synthetic 3D shapes is an open challenge due to the complexity of the geometry and the lack of large and diverse publicly available datasets. While generative models based on convolutional neural networks (CNNs) have shown the ability to produce novel synthetic high-resolution images indistinguishable from the original pictures by an expert human observer, applying them to non-grid-like data like 3D meshes presents many challenges.

In our work, we tackle the problem of 3D body synthesis by reducing 3D meshes to 2D image representations. We show previously that the face can naturally be modelled on 2D grid, while for the more challenging 3D body geometry we propose a novel non-bijective 3D-2D conversion method representing the 3D body mesh as a plurality of rendered projections on the 2D grid. Then we train a state-of-the-art Vector Quantized Variational Autoencoder VQ-VAE-2 to learn a latent representation of 2D images. The 3D synthesis is achieved by training on the latent space and sampling it with a powerful PixelSNAIL autoregressive model.

We evaluate our method of 3D shape synthesis versus a classical one based on Principal Component Analysis (PCA), where the synthetic shapes are obtained by sampling from the empirical cumulative distribution of the PCA scores. We use the empirical distributions of two commonly used metrics, called specificity and diversity, to demonstrate that synthetic bodies generated with our method are statistically closer to the real faces when compared to the PCA ones. This experiment on 3D body geometry is a preliminary work and requires further research to match the test set statistics but shows promising results.

This section is a part of the paper Generating High-Resolution 3D Shapes using VQ-VAE-2 with PixelSNAIL Networks authored jointly with Dmitry Znamenskiy, Long, Yuxuan, Nicola Pezzotti and Milan Petkovic, which is currently under review in Special Issue "Computer Vision in Human Analysis: From Face and Body to Clothes". A special issue of Sensors (ISSN 1424-8220). This special issue belongs to the section "Sensing and Imaging".

7.1. INTRODUCTION

In the last two decades, the use and applications of virtual 3D models in the real world have risen exponentially. There are many reasons behind this, from the growth in computational power to the economic benefit of using a parametric model to simulate physical phenomena. Today, 3D models serve many fields including animation of character [67] and faces [154]–[156], recognition of expression [157], face recognition [158] and inferring body shapes and measurement to be used, for example, in the clothing industry for virtual try-on [179] or in the medical field to estimate fat distribution.

The analysis of personal 3D data is subject to privacy constraints, limiting the 3D data sharing. Modern 3D scanners operate with sub-millimetre accuracy so that a person can be identified from its 3D scan. At the same time, the use of obscuring and decimation methods on 3D data can conflict with the modelling objectives and accuracy requirements of body or face measurements for a particular healthcare device. The sharing of randomly generated 3D subjects, using generative models trained on original 3D data, could enable 3D analysis when the sharing of the original data is constrained for example when data protection laws forbid sharing them between hospitals.

The analysis of a dataset of 3D models requires that all scans are brought in full correspondence. This can be achieved by registering a common reference 3D template into every scan, so that all morphed templates, representing individual scans, will have a common parametrization, see the seminal work of Blanz and Vetter [4] for faces and a Allen *et al* [5] similar work on full-bodies. Once having a common representation, we tackle the problem of generating synthetic and realistic shapes. The registered template has thousands of vertices, but their positions are highly correlated. Generation methods should take this correlation into account by finding a low-dimensional de-correlated representation of the surface. Thus Blanz and Vetter [4] used Principal Component Analysis (PCA) as a data driven approach, to reduce a stacked array of thousands of vertex coordinates to a small number of PCA scores. Analogously in [5], [180] Allen *et al.* applies the same approach to learn a data-driven human shape model.

The generation of synthetic 3D meshes by sampling using PCA decomposition is straightforward: assuming the orthogonality of the PCA basis vectors it is sufficient to sample the PCA scores as independent random variables. While the PCA analysis assumes that the marginal coefficient distributions are close to Gaussian, it is more reasonable to follow a data-driven approach and sample from the empirical distributions of each coefficient. This approach brings multiple benefits: It is easy to implement, it is fast, and the sampled shapes will have the marginal coefficient distributions statistically indistinguishable from the original data. However, there are also disadvantages. In fact, not all combinations of PCA scores result in a natural human shape as shown in the previous experiment for the face geometry.

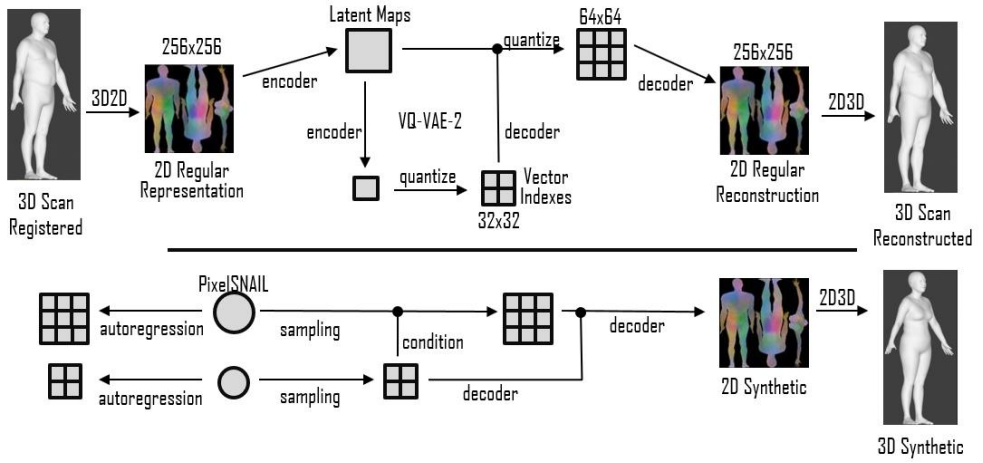


Figure 7.1. Flowchart of Full body. Definition of the components for our approach on synthetic body generation. The registered 3D scans are first converted into a regular 2D image to feed the VQ-VAE-2 autoencoder. The PixelSNAIL later learns a prior over the discrete codes and then it is used to sample novel synthetics codes, which are then decoded into the new geometric images and 3D scans.

Stepping back from PCA, we explore generative methods which could work with original high-dimensional vertex data. The recent advances in Convolutional Neural Networks (CNNs) where the deep generative modelling can capture more complex interactions in the data, thus potentially improving over linear methods. The current state of the state of the art advances in the field of geometric deep learning [58] uses the power of CNNs by adapting them to work on meshes [59], [60]. Graph convolutions, however, restrict the resolution, and therefore, the accuracy of the 3D template.

In our work, we decided to use an alternative approach (A schematic representation of our approach is presented in Figure 7.1): we work with the representation of the registered 3D template on a regular 2D grid to fully employ CNNs. Hence, a part of this work is devoted to the conversion from 3D shapes to 2D grids. We first considered a simpler case of the human facial surface using a 3D template with a CNN-friendly 2D grid structure as shown in the previous chapter. The face can be represented as a 2D image where a single piece UV map is easy to generate. Once can simply register starting with a regular grid the face scans. Here, we consider a general case of the complete human body that cannot be naturally unwrapped into a 2D grid. It is in fact hard to cut and unwrap the mesh considering the complex manifold of the full body. Therefore, we propose a novel non-bijective method when the 3D shape is represented as a sequence of projections on the 2D grid, as illustrated in the right of Figure 7.2, and explained in detail in the next sections. The advantage of this method is the simple creation of 2D body template by means of rendering, which works with any parametrization of the 3D grid. The inverse 2D to 3D conversion is achieved by aggregating and regularizing multiple body projections.

Once we have 2D representations of 3D template, we can apply CNNs to generate new 3D shapes. For our experiments, we use the relatively novel 2D image synthesis

method employing a latent image representation given by a quantized variational autoencoder VQ-VAE-2 [56], which is sampled using the autoregressive network PixelCNN [147] with attention, called as PixelSNAIL[162].

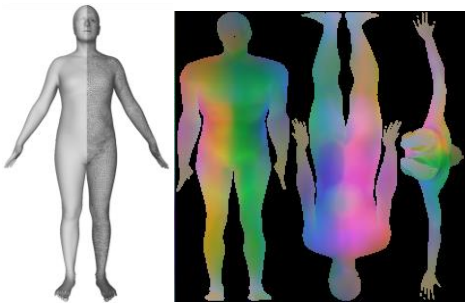


Figure 7.2. Template and UV Map. The average full-body template with half rendered surfaces and half mesh on the left and the relative regular representation on the right. The mesh is composed of more than 50.000 vertexes. The geometry images for the full body have resolution of 256×256 and is derived from 3 projections: frontal view, the middle from a rear view and the right from the bottom.

The remaining chapter content is structured as follows. In the next section we will introduce related works for full body representations. In Section 7.3, we introduce our method for generating 2D regular representation. The fourth section describes the input data and the experiments performed, and the fifth section contains the results of objective evaluation. Finally, we discuss the results, conclude the chapter, and give acknowledgements.

7.2. Related Works

Many 2D representation methods originate from the solution of the rendering problem which relies on so called UV maps for the mapping of 2D texture image on 3D object. The UV maps, by definition, provides bijective mapping from the 3D mesh triangles to their images on the texture image. The UV maps for the face model can be created by warping of the 3D templates with a regular grid of the facial surface, see for example Booth *et al.* [61] for a list of possible optimal implementations. In the previous chapter we utilize a facial parametric model which 3D template has already a regular 2D grid. Instead, the full body UV maps can be coarsely divided into two classes: single connected piece and a patch work of multiple pieces. Regarding the single piece solution, in 2002, Gu *et al.* [169] showed how to cut a surface and sample it over a regular 2D square grid generating the so-called geometry image. While the work did minimize the artefacts along the cuts, it naturally cannot solve huge distortions when mapping complex 3D body shape with arms and legs into a square grid. The majority of the recent prior art follows UV map from the SMPL model [181] which has a patchwork of different body parts: head, two palms, two arms, torso, two feet, two legs, which are economically mapped each in their own place on the 2D grid, so that the total area of background pixels is minimized. While the patchwork has more control over the surface area, it has many cuts through the surface and every cut creates a challenge to get the continuity in the generated vertex positions on both sides of the cut. Thus, in a more recent work, Zeng *et al.* [182] revert to a single piece solution where the geometrical distortions are reduced for the cost of increased percent of background pixels present in the UV map.

Observe that the use of prior-art UV maps for the 3D-2D conversion is subject to licenses that constrain commercial applications and limit the impact this technology

can have in several real-world applications. Thus, in this work we propose a novel non-bijective method that maps 3D body into multiple 2D projections rendered from a set of camera views. This method minimizes the number of pieces and is generalizable to arbitrary 3D body template.

7.3. FROM 3D TO 2D REPRESENTATIONS

Within this work, we have considered two datasets of registered scans already available at Philips Research: one corresponding to a face template with a regular grid of 128×128 vertices introduced in the previous chapter, and a full body template with about 50K vertices randomly placed over the body surface shown on the right of Figure 7.2. Conceptually, we have followed the method explained in [4], [62] and have morphed all scans by means of non-rigid registration methods [83], [84]. A more detailed description of our parametric models is reported in Chapter 2.

Since the face template already has a grid structure, we have only applied vertex-based normalization to map the range of input values into interval $[0, 1]$, and therefore, to facilitate the follow up processing with the neural networks. The mapping to $[0, 1]$ also facilitates the visualization of the normalized *xyz* facial data as *rgb* images, as illustrated on Figure 6.2.

In contrast to the face, the body template's complex shape cannot be naturally unwrapped into a 2D grid due to its watertight geometry, where cutting and unfolding to a planar geometry without introducing extreme deformations is challenging. We propose a novel non-bijective method that represents the 3D shape as a sequence of projections on the 2D grid. The advantage of this method is the simple creation of 2D body template by means of rendering, which works with any parametrization of the 3D grid. In our experiments we have rendered three body views: from front, from back and from bottom, where in the first two views we have morphed the arms down to save space on the rendered geometry image. The use of multiple views ensures that almost all template vertices are visible in at least one view, see Figure 7.3 (left), where their displacement can be computed via bilinear interpolation of the 2D grid values.

The inverse conversion from 2D to 3D is achieved by aggregating bilinear interpolated vertex displacements from the views followed by regularization of the 3D shape using mesh Laplacian, as described in [183]. Thus, the vertexes positions V are found by minimizing quadratic cost function:

$$V = \operatorname{argmin}(\|G \cdot V - P\|_2 + \alpha \|L \cdot V - L \cdot V_0\|_2)$$

where G is the sparse registration matrix for 2D grid points on the body, P are the de-normalized xyz vertex positions at the 2D grid points, α = 0.001 is the regularization parameter (for units defined in mm), L is the sparse matrix corresponding to discrete Laplacian, and V_0 are the average body vertices. Due to the use of the quadratic norm, it is easy to derive a closed form solution for V which gives us a mean vertex error of 0.14 mm, distributed as illustrated on Figure 7.3 (right).

Once the 3D bodies are mapped to the 2D grid, we apply the same normalization as in the case of faces to map the range of input values into interval [0, 1]. The range parameters for each grid point were retained for use during the de-normalization and conversion of the synthetic images into 3D shapes.

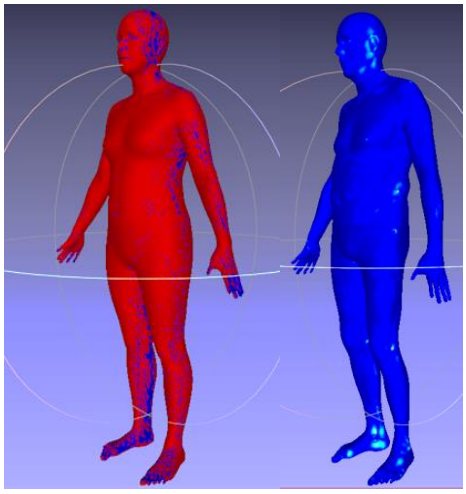


Figure 7.3. 2D-3D reconstruction accuracy. Left image shows visible (red) vs invisible(blue) vertexes in the template body. Right image shows distribution of the reconstruction error, according to 'jet' colormap (blue=0.0mm and red= 1.00mm), in a sample body.

7.4. EXPERIMENTS

In our experiments we have used 3D scans from the CAESAR dataset of full body scans [36], [178]. The above data gave us more than 4000 registered 3D body scans which we augmented to 50000 registered bodies using 'Age' and 'Weight' growth models. The dataset was split stratified according to participant id into train 90%, validation 5%, and test set 5%, both for face and full-body experiment. We tested and computed the metrics only on the test set without considering the augmentations. For the sake of experimental reproducibility, we did not perform any other augmentation neither in training nor in test time. However, we believe further realistic augmentations would impact and consolidate the results.

After registering the common templates, we encoded the information in 200 principal components (both for the head and full body). The description of our parametric model is presented in Chapter 2. We then computed the eCDF for all PCA encoded scans separately. To sample a new scan, we simply sample each score independently and then decode 3D scan from the scores.

In our experiments we focus on a two layers VQ-VAE hierarchy with input resolution of grid resolution of 256×256 (and relative latent maps of dimension 64×64 and 32×32) for the full body. We follow the approach described in Chapter 5 and Chapter 6 to fine tune hyperparameters such as the latent space dimensions. We also reduced the batch size to 32 compared to the original implementation for both the autoencoder and the autoregressive model. The reconstruction error average per vertex for the full body is presented in Figure 7.5. The figure shows an error bigger than 1.00mm on foot and hands as expected due to the complexities of fingers and the relatively few pixels devoted to them. However, we believe we should reduce the reconstruction error in the hip-waist area by improving the methods e.g., by providing and a better 2D representation.

Concerning the autoregressive model, we used the original configuration for ImageNet apart from the batch size, 32 in our example, and total number of epochs,

1000 for both top and bottom hierarchy. The autoregressive models' validation accuracies in predicting the latent codes after 1000 epochs reaches 0.45 top accuracy and 0.88 bottom accuracy. All the models were trained on PyTorch [110] with the same hyperparameters as in the original implementation (excluding the one explicitly mentioned above).

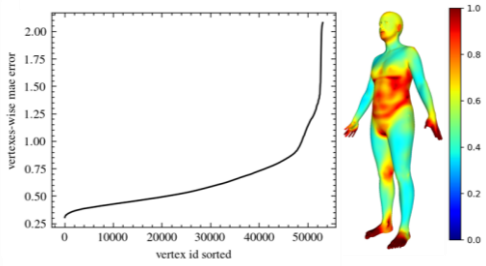


Figure 7.5. Reconstruction error vertex-wise mae for the best VQ-VAE-2 full body model. On the left the full-body error is first plotted and then displayed on the full body mesh surface. The meshes jet color-code ranges from blue 0.0mm error map and to red 1.0mm.

In Figure 7.4 one example of how PixelSNAIL is failing to produce correct geometry images is shown. This is happening for 0.5% of generated scans when using temperature 1.0. However, the failing generation can be easily detected by checking whether the background pixels are synthesized at the same places as in the original 2D template.

7.5. EVALUATION

We have analysed 2D representations for registered raw vs PCA-encoded vertices and computed empirical distributions for specificity and diversity metrics. The empirical distributions of metrics are shown in Figure 7.6. Contrary to the face metrics which show that our approach results in synthetic faces which are statistically close to the original scans in the test set, the full-body metrics do not confirm our subjective evaluation. The distribution of diversity and specificity metrics for the PCA-based method seems closer to the distribution in the test dataset selected from original scans. We can suggest the following reasons explaining this counterintuitive result. First, the full body, compared to face, has more challenging 3D-2D mapping with value discontinuity at the border of the projections. Quantitatively this is already seen in the autoencoder, and autoregressive worse performances compared to the

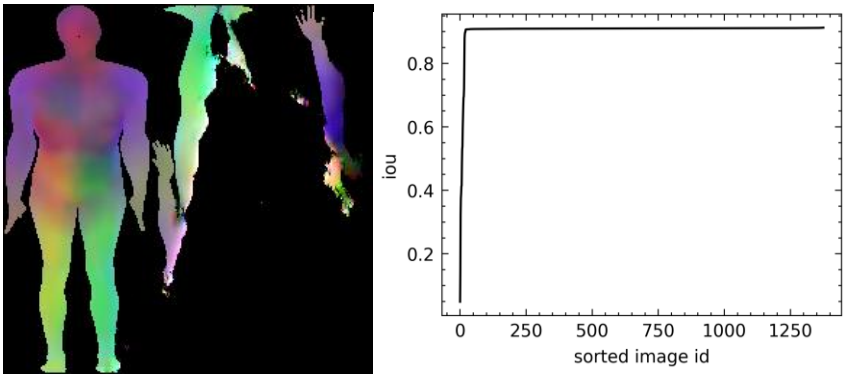


Figure 7.4. Corrupted synthetic 2D representations. An example in which the PixelSNAIL fails to produce a correct geometrical image on the left. The right is the distribution of Intersection Over Unions over the segmented UV body mask.

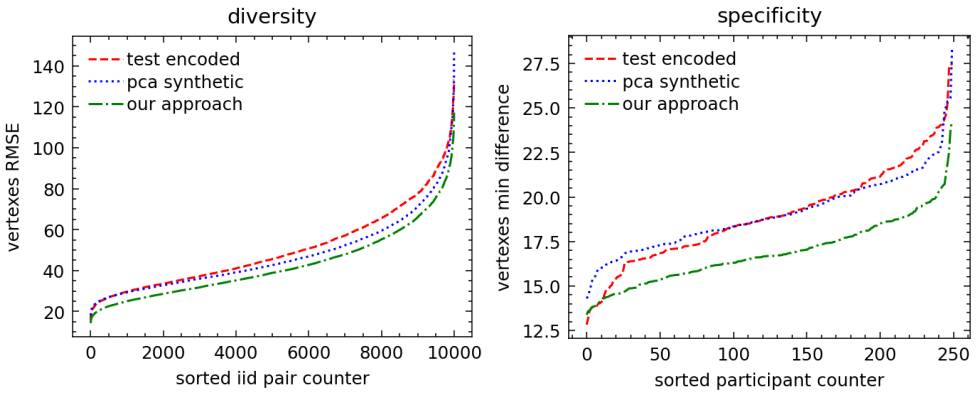


Figure 7.6. The full body metrics. The full body metrics do not show promising results as in the face geometry and further work are experiments are necessary to reach the desired quality of scans. Both diversity and specificity for our approach are lower than test set distributions allow us for example to calibrate the model and increase the variability of scans by raising the sampling temperature and therefore increasing entropy, diversity and specificity of the scans.

face experiment which we believe is the main reason for the lower specificity and diversity. In fact, we believe that a less accurate autoregressive model will lead to closer to average distributions of latent codes and 3D synthetic shapes. Another major point of discussion that will be analysed in future work is to benchmark the effectiveness and quality of our augmentations prior to model training.

Few example of synthetic bodies generated with our methods and with PCA are shown in Figure 7.7. The qualitative comparison seems promising but further experiments and quantitative evaluation is necessary.

7.6. Discussion And Conclusion

While the experiments successfully generate realistic high-resolution 3D faces and full bodies, we consider this only a first step in proving the validity of this approach.

One main challenge is the lack of a clear, quantitative, metric to judge whether a scan belongs to the “real” class (of faces or full bodies). These proposed diversity and specificity metrics may not be enough to capture all the relevant shape information. Moreover, the full-body experiment still requires further investigation in many directions. In fact, the purpose of generating full body 2D representation was required for different applications. We suggest that future research consider optimizing the number of projections and their position, taking particular interests in minimizing the background pixels to facilitate the NNs training.

Another potential issue is the geometric distortion introduced by the reconstruction from multiple projections. This can be explored are reduced in future research by using a deep learning approach to rendering for example as described in Kato *et al.* [184].

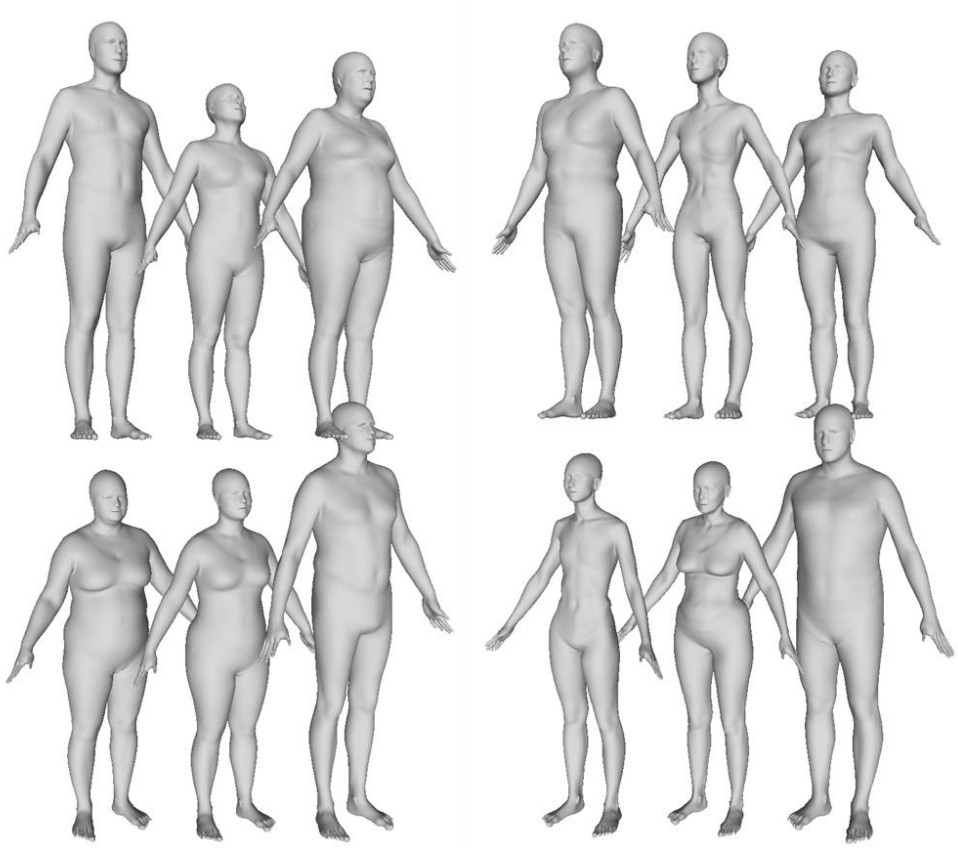


Figure 7.7: Full body scans. The three columns on the left are sampled using our approach, the three columns on the right are samples generated using the PCA based approach. Subjectively, our approach more often provides realistic scans compared to PCA-based approach where we observe extreme shape features and often bodies with mixed gender that are not representative of the input dataset.

In summary we presented a novel approach that is able to generate high resolution synthetic 3D faces in the previous chapter and full bodies. While tackling the more complex full-body geometry we presented a new non-bijective way of creating the 2D representation of 3D template by using multi-view projections. The advantage of this method is that it is agnostic to the shape of the 3D template and can be adapted to any 3D template: of foot, hand, torso, etc. The remaining challenge (and the subject of optimization) is finding of the best set of camera views while minimizing the overlap and the number of invisible vertices.

Another relevant strength of our approach is that it does not require the parametrization of 3D face/body model and can be directly applied to registered templates. This leads to the possibility to generate scans outside of the PCA linear sub-space. However, the major contribution of our work is that our method strictly outperforms the linear PCA classical approach and generates realistic high-resolutions scans. Moreover, our method also outperforms the PCA-based one when

training it over the PCA encoded scans for the facial scans. Further work is required to ensure that the same conclusions hold also for the full body surface.

To conclude we demonstrate that is possible to apply the state-of-the-art CNNs for the generation of realistic high-resolution 3D scans by reducing the problem to 2D representations. In particular we have shown that the combination of VQ-VAE-2 with PixelSNAIL, which was previously used for the generation of realistic facial images and skin lesions, it also applicable to 3D meshes when representing them as images.

7.7. ACKNOWLEDGMENTS

We thank Philips Research for providing access to the datasets full body scans and software resources to manage the high-resolution parametric models.

8. CONCLUSIONS

The analysis and parametrization of the outer body skin have been proven vital in many fields spanning from healthcare to the clothes and movie industry. This has not been possible before the advent of modern computer vision and the harness of artificial intelligence power. In this thesis, we explored recent advances in the modeling and parametrization of the body skin on different scales: from the global body measurements and shape to the face shape, and then to the local description and classification of skin features such as skin hair and lesions. While applying and advancing deep learning and shape modeling, we proposed novel approaches to the prediction of the skin surfaces and to the analysis of local skin patches effectively satisfying several real-life needs and enabling a few industrial applications.

8.1. OVERVIEW

Our research was driven by consumer and business needs and our aim from the beginning was to blend them with the research goals and research questions. We developed skin modeling and parametrization technologies to advance devices, solutions and digital tools for personal health and healthcare. Thus, a large part of the research is devoted to the Laser Hair Removal applications, where we first addressed the image-based hair counting for automatic assessment of the epilation efficiency. Second, we considered the skin lesion classification which can be treated differently during the epilation. Third, the epilation application motivated the analysis of 3D body surface shape and area, where we generated synthetic 3D data to avoid privacy-related issues. Another application that benefits from the synthetic 3D facial surfaces is the analysis of the facial geometry for making a perfectly fitting oxygen mask for patients suffering from sleep apnea.

We defined five research questions to advance the state of the art in the field of computer vision applied to the skin:

RQ1. How accurately can one estimate skin shape from anthropometric measurements?

RQ2. Can a computer vision based automatic hair counter replace the human annotator?

RQ3. Does hair affect skin lesion diagnosis in deep learning classifiers?

RQ4. Can we model the distribution of skin lesion images and generate realistic looking synthetic examples with deep learning?

RQ5. Can we generate realistic 3D skin exploiting the power of deep generative models?

And our main contributions answering these questions respectively are:

- a methodological approach suggesting the best combination of simple body measurements to acquire for the shape estimate of any selected body part demonstrating high accuracy in surface prediction (Chapter 2, RQ1)
- an end-to-end deep learning approach to automatically count skin hair in small skin patches which has the potential to replace the manual hair counting but is not yet mature enough to replace human work fully (Chapter 3, RQ2)
- a pipeline demonstrating that hair presence does not affect state-of-the-art skin lesion classification models (Chapter 4, RQ3)
- an effective approach to the controllable generation of artificial skin lesion images using deep generative models (Chapter 5, RQ4)
- and, finally, a successful adaptation of the above approach, which was initially developed for the generation of images, to the generation of synthetic 3D skin surfaces of faces and bodies (Chapter 6 and 7, RQ5).

8.2. Limitations and Future Work

Hence, all questions have been positively answered except for RQ2, where further work will be needed to create more robust hair counters to replace the human annotators. Despite the adoption of such solutions by the industry, many limitations are still present, and many open problems need future work. One major limitation is the ability to process 3D meshes efficiently and similarly to the 2D images defined on a regular pixel grid. Such a possibility was not available at the start of the research where now many directions in the deep learning field are rising like geometric deep learning. Another limitation posed to such meshes models is capturing and normalizing proper skin textures on top of the model. Concerning the skin patches analysis, we see a need for deep learning models to properly handle medical data that suffer from difficult light conditions and unbalance in the classes. Note that the controllable generation of synthetic skin lesions presented in Chapter 5 and similar related research work partially solves the problem, as we do not factorize the effect of illumination. Without factorization, the presented generative approach requires a significant amount of training data to be able to extrapolate unseen data or hypotheses. However, as presented in our work, the ability to manipulate only a salient part of the image allows for a richer data augmentation possibility, for example, in the case of a rare class representing a rare disease.

Over the past four years, many researchers have provided feedback on our work. Additionally, a collaboration called “z-inspection” [185], [186] with a group of external researchers aimed at addressing the trustworthy use of AI in healthcare raised a significant ethical awareness related to the research presented in the dissertation. In the following part of this sub-section the most important ethical considerations are discussed.

Artificial Intelligence versus Human Labor

The work on hair count was not possible without the work of many human annotators. More than 4000 pictures were annotated by hand with the total hair count. Nevertheless, the annotator’s work is the same work that might help replace himself

with the automatic end-to-end deep learning solution. We believe that particular care for such topics is crucial, and we recommend that AI researchers be well acquainted with these ethical tensions by taking the necessary steps to ensure their work is not misused to reduce overall population wellbeing. This can be achieved only by the cooperation between different institutions and stakeholders since no single person or entity can predict or understand how complex new technologies could evolve; however, the AI researcher should raise awareness as much as possible within its capabilities.

An ethical assessment attached to a submitted paper is one practical step that AI venues and journals may implement. Another possibility is the full ethical assessment by a third party or independent organization such as z-inspection to detect human concepts in skin lesions [187].

ACCESSIBILITY

Ideally, the benefit of the research should be accessible to everyone no matter their social status or geographic location. This is not always possible, but steps may be taken to ensure broader access. For example, one of the main limitations of the work presented in Chapter 2 is the need for constructing a full parametric model in the training phase. This implies that a retrain of the model including different dependent variables or dependent parametric scores (for example the calf to predict foot shape) would require building a new parametric model.

We hope that the synthetic generation of skin surfaces presented in Chapter 6 may improve the accessibility of such data. Similarly, the generative models may allow everyone to access expensive and rare skin images in the future.

PRIVACY AND DATA COLLECTION

Another tension that often emerged in the dissertation is the attention to privacy versus the need for big data collection: skin images and 3D human avatar has always the potential to leak privacy-sensitive information.

Ideally skin images should be cleaned to remove any birthmark or identifying tattoo. One suggestion for future research is to exploit the generative approach presented in Chapter 5 to inpaint and remove the area of the image which are potentially linked to the user identity.

DEVELOPMENT VERSUS DEPLOYMENT

After the major dissemination in a conference or journal, all models developed in the dissertation were deployed in production for various research and business applications. They were all sent to production and maintained by the engineering teams. This implies that the model must be robust and stable over time. These are qualities that are sometimes overlooked for research outputs. Future research will constrain the deep learning models, especially generative ones, to be deployment friendly for example by reducing their complexity.

ENERGY CONSUMPTION

A similar action might be taken to reduce energy usage. In fact, training deep neural networks requires great computational power and, therefore, great consumption of energy. In this work, we trained several neural networks models and in future research we will aim to reduce the energy costs in deep generative models. For example we are currently improving our DL pipelines by maximizing the use of pre-trained models and training efficient and smaller ones such as EfficientNet [153] as already seen in Chapter 5.

8.3. IN CONCLUSION

In conclusion, we presented several techniques to use modern computer vision to enhance skin-based applications. We believe that a combination of shape modeling and texture analysis using deep learning methods will play a crucial role in the future of skin mapping, navigation, and parametrization enabling game-changing technologies such as the full integration of robotic tools to help in surgical operations.

BIBLIOGRAPHY

[1] Y. LeCun *et al.*, “Backpropagation applied to handwritten zip code recognition,” *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “2012 AlexNet,” *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2012, doi: http://dx.doi.org/10.1016/j.protcy.2014.09.007.

[3] T. J. Brinker *et al.*, “A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task,” *Eur. J. Cancer*, vol. 111, pp. 148–154, 2019.

[4] V. Blanz and T. Vetter, “A morphable model for the synthesis of 3D faces,” *Proc. 26th Annu. Conf. Comput. Graph. Interact. Tech.*, pp. 187–194, 1999.

[5] B. Allen, B. Curless, and Z. Popović, “The space of human body shapes: reconstruction and parameterization from range scans,” *ACM Trans. Graph.*, 2003.

[6] H. Durrant-Whyte and T. Bailey, “Simultaneous localization and mapping: Part I,” *IEEE Robot. Autom. Mag.*, vol. 13, no. 2, pp. 99–108, 2006, doi: 10.1109/MRA.2006.1638022.

[7] F. Bogo, J. Romero, E. Peserico, and M. J. Black, “Automated Detection of New or Evolving Melanocytic Lesions Using a 3D Body Model,” pp. 593–600, 2014, Accessed: Nov. 29, 2017. [Online]. Available: http://files.is.tue.mpg.de/fbogo/papers/Bogo_MICCAI2014.pdf.

[8] Y. Wang, J. Moss, and R. Thisted, “Predictors of Body Surface Area,” *Butterworth-Heinemann J. Clin Anesth*, vol. 4, pp. 4–10, 1992.

[9] A. C. Davies, J. H. Yin, and S. A. Velastin, “Crowd monitoring using image processing,” *Electron. \& Commun. Eng. J.*, vol. 7, no. 1, pp. 37–47, 1995.

[10] A. S. Laliberte and W. J. Ripple, “Automated wildlife counts from remotely sensed imagery,” *Wildl. Soc. Bull.*, pp. 362–371, 2003.

[11] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the seventh IEEE international conference on Computer vision, 1999.*, 1999, vol. 2, pp. 1150–1157.

[12] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[14] V. Rotemberg *et al.*, “A Patient-Centric Dataset of Images and Metadata for Identifying Melanomas Using Clinical Context,” *arXiv Prepr. arXiv2008.07360*, 2020.

[15] C. Ledig *et al.*, “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[17] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, “Neural Voice Puppetry: Audio-driven Facial Reenactment,” *arXiv Prepr. arXiv1912.05566*, 2019.

[18] Y. Jia *et al.*, “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” in *Advances in neural information processing systems*, 2018, pp. 4480–4490.

[19] L. Theis, A. van den Oord, and M. Bethge, “A note on the evaluation of generative models,” *arXiv Prepr. arXiv1511.01844*, 2015.

[20] S. Ravuri and O. Vinyals, “Classification accuracy score for conditional generative models,” in *Advances in Neural Information Processing Systems*, 2019, pp. 12247–12258.

[21] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” *Adv. Neural Inf. Process. Syst.*, vol. 29, pp. 2234–2242, 2016.

[22] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local Nash equilibrium,” *Adv. Neural Inf. Process. Syst.*, pp. 6626–6637, 2017.

[23] K. K. Cornelissen, K. McCarty, P. L. Cornelissen, and M. J. Tovée, “Body size estimation in women with anorexia nervosa and healthy controls using 3D avatars,” *Sci. Rep.*, vol. 7, no. 1, p. 15773, 2017.

[24] S. C. Mölbert *et al.*, “Assessing body image in anorexia nervosa using biometric self-avatars in virtual reality: Attitudinal components rather than visual body size estimation are distorted,” *Psychol. Med.*, vol. 48, no. 4, pp. 642–653, 2018, doi: 10.1017/S0033291717002008.

[25] K. Kristensen *et al.*, “Towards a next generation universally accessible ‘online shopping-for-apparel’ system,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8006 LNCS, no. PART 3, pp. 418–427, 2013, doi: 10.1007/978-3-642-39265-8_47.

[26] L. G. Christiansen, A. L. Brooks, E. P. Brooks, and T. Rosenørn, “A Virtual Dressing Room for People with Asperger’s Syndrome,” in *International Conference on Universal Access in Human-Computer Interaction*, 2014, pp. 163–170.

[27] N. Magnenat-Thalmann, H. Seo, and F. Cordier, “Automatic modeling of virtual humans and body clothing,” *J. Comput. Sci. Technol.*, vol. 19, no. 5, pp. 575–584, 2004.

[28] C. Kuster *et al.*, “Towards next generation 3D teleconferencing systems,” in *2012 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, 2012, pp. 1–4.

[29] S. Orts-Escolano *et al.*, “Holoportation: Virtual 3d teleportation in real-time,” in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, 2016, pp. 741–754.

[30] T. Fredriksson and U. Pettersson, “Severe psoriasis — Oral therapy with a new retinoid,” *Dermatology*, vol. 157, no. 4, pp. 238–244, 1978, doi: 10.1159/000250839.

[31] D. Pinkel, “The use of body surface area as a criterion of drug dosage in cancer chemotherapy,” *Cancer Res.*, 1958.

[32] G. A. Knaysi, G. F. Crikelair, and B. Cosman, “The rule of nines: its history and accuracy,” *Plast. Reconstr. Surg.*, 1968.

[33] E. D. Dommasch, D. B. Shin, A. B. Troxel, D. J. Margolis, and J. M. Gelfand, “Reliability, validity and responsiveness to change of the Patient Report of Extent of Psoriasis Involvement (PREPI) for measuring body surface area affected by psoriasis,” *Br. J. Dermatol.*, 2010.

[34] R. J. Hunter, M. A. Navo, P. H. Thaker, D. C. Bodurka, J. K. Wolf, and J. A. Smith, “Dosing chemotherapy in obese patients: actual versus assigned body surface area (BSA),” *Cancer Treat. Rev.*, 2009.

[35] D. Parvizi *et al.*, “The potential impact of wrong TBSA estimations on fluid resuscitation in patients suffering from burns: things to keep in mind,” *Burns*, vol. 40, no. 2, pp. 241–245, 2014.

[36] K. M. Robinette, H. Daanen, and E. Paquet, “The CAESAR project: a 3-D surface anthropometry survey,” in *Second International Conference on 3-D Digital Imaging and Modeling (Cat. No.PR00062)*, 1999, pp. 380–386.

[37] R. M. Ball and J. F. M. Molenbroek, “Measuring Chinese heads and faces,” *Proc. 9th Int. Congr. Physiol. Anthropol. Hum. Divers. Des. life*, 2008.

[38] R. R. Anderson and J. A. Parrish, “Selective photothermolysis: precise microsurgery by selective absorption of pulsed radiation,” *Science (80-.).*, vol. 220, no. 4596, pp. 524–527, 1983.

[39] G. B. Altshuler, R. R. Anderson, D. Manstein, H. H. Zenzie, and M. Z. Smirnov, “Extended theory of selective photothermolysis,” *Lasers Surg. Med. Off. J. Am. Soc. Laser Med. Surg.*, vol. 29, no. 5, pp. 416–432, 2001.

[40] M. C. Grossman, C. Dierickx, W. Farinelli, T. Flotte, and R. R. Anderson, “Damage to hair follicles by normal-mode ruby laser pulses,” *J. Am. Acad. Dermatol.*, vol. 35, no. 6, pp. 889–894, 1996.

[41] K. Nouri, V. Vejjabhinanta, S. S. Patel, and A. Singh, “Photoepilation: a growing trend in laser-assisted cosmetic dermatology,” *J. Cosmet. Dermatol.*, vol. 7, no. 1, pp. 61–67, 2008.

[42] S. D. Gan and E. M. Graber, “Laser hair removal: a review,” *Dermatologic Surg.*, vol. 39, no. 6, pp. 823–838, 2013.

[43] M. Rahnemoonfar and C. Sheppard, “Deep count: fruit counting based on deep simulated learning,” *Sensors*, vol. 17, no. 4, p. 905, 2017.

[44] W. Xie, J. A. Noble, and A. Zisserman, “Microscopy cell counting and detection with fully convolutional regression networks,” *Comput. methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 6, no. 3, pp. 283–292, 2018.

[45] T. Falk *et al.*, “U-Net: deep learning for cell counting, detection, and morphometry,” *Nat. Methods*, vol. 16, no. 1, pp. 67–70, 2019.

[46] C. Zhang, H. Li, X. Wang, and X. Yang, “Cross-scene crowd counting via deep convolutional neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 833–841.

[47] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, “Multi-source multi-scale counting in extremely dense crowd images,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2547–2554.

[48] C. Arteta, V. Lempitsky, and A. Zisserman, “Counting in the wild,” in *European conference on computer vision*, 2016, pp. 483–498.

[49] P. Vallotton and N. Thomas, “Automated body hair counting and length measurement,” *Ski. Res. Technol.*, vol. 14, no. 4, pp. 493–497, 2008, doi: 10.1111/j.1600-0846.2008.00322.x.

[50] H.-C. Shih, “An unsupervised hair segmentation and counting system in microscopy images,” *IEEE Sens. J.*, vol. 15, no. 6, pp. 3565–3572, 2014.

[51] H.-C. Shih and B.-S. Lin, “Hair segmentation and counting algorithms in microscopy image,” in *2015 IEEE International Conference on Consumer Electronics (ICCE)*, 2015, pp. 612–613.

[52] H. Lim *et al.*, “Development of a Novel Automated Hair Counting System for the Quantitative Evaluation of Laser Hair Removal,” *Photomed. Laser Surg.*, vol. 35, no. 2, pp. 116–121, 2017, doi: 10.1089/pho.2016.4140.

[53] R. A. Smith *et al.*, “Cancer screening in the United States, 2018: a review of current American Cancer Society guidelines and current issues in cancer screening,” *CA. Cancer J. Clin.*, vol. 68, no. 4, pp. 297–316, 2018.

[54] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, “Generative Adversarial Networks,” *arXiv Prepr. arXiv …*, pp. 1–9, 2014, doi: 10.1017/CBO9781139058452.

[55] S. Zhou, M. Gordon, R. Krishna, A. Narcomey, L. F. Fei-Fei, and M. Bernstein, “Hype: A benchmark for human eye perceptual evaluation of generative models,” in *Advances in Neural Information Processing Systems*, 2019, pp. 3449–3461.

[56] A. Razavi, A. van den Oord, and O. Vinyals, “Generating diverse high-fidelity images with vq-vae-2,” in *Advances in Neural Information Processing Systems*, 2019, pp. 14837–14847.

[57] X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel, “PixelSNAIL: An improved autoregressive generative model,” *arXiv Prepr. arXiv1712.09763*, 2017.

[58] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, “Geometric deep learning: going beyond euclidean data,” *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, 2017.

[59] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, “Generating 3D faces using convolutional mesh autoencoders,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 704–720.

[60] P. De Haan, M. Weiler, T. Cohen, and M. Welling, “Gauge equivariant mesh cnns: Anisotropic convolutions on geometric graphs,” *arXiv Prepr. arXiv2003.05425*, 2020.

[61] J. Booth and S. Zafeiriou, “Optimal uv spaces for facial morphable model construction,” in *2014 IEEE international conference on image processing (ICIP)*, 2014, pp. 4672–4676.

[62] B. Allen, B. Curless, and Z. Popović, “The space of human body shapes: reconstruction and parameterization from range scans,” *ACM Trans. Graph.*, vol. 22, no. 3, p. 587, 2003, doi: 10.1145/882262.882311.

[63] B. Allen, B. Curless, and Z. Popovic, “Exploring the space of human body shapes : data-driven synthesis under anthropometric control,” *Digit. Hum. Model. Des. Eng. Symp.*, 2004.

[64] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia, “MeshLab: an Open-Source Mesh Processing Tool,” 2008.

[65] H. Seo and N. Magnenat-Thalmann, “An automatic modeling of human bodies from sizing parameters,” *Proc. 2003 Symp. Interact. 3D Graph. - SI3D ’03*, 2003.

[66] H. Seo, F. Cordier, and N. Magnenat-Thalmann, “Synthesizing animatable body models with parameterized shape modifications,” *Sca ’03*, pp. 120–125, 2003.

[67] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H. P. Seidel, “A statistical model of human pose and body shape,” *Comput. Graph. Forum*, vol. 28, no. 2, pp. 337–346, 2009.

[68] S. Wuhrer and C. Shu, “Estimating 3D human shapes from measurements,” *Mach. Vis. Appl.*, vol. 24, no. 6, pp. 1133–1147, 2013, doi: 10.1007/s00138-012-0472-y.

[69] A. Tsoli, M. Loper, and M. J. Black, “Model-based anthropometry: Predicting measurements from 3D human scans in multiple poses,” *2014 IEEE Winter Conf. Appl. Comput. Vis.*, 2014.

[70] M. Hill, S. Streuber, C. Hahn, M. Black, and A. O’Toole, “Exploring the relationship between body shapes and descriptions by linking similarity spaces,” *J. Vis.*, vol. 15, no. 12, p. 931, 2015, doi: 10.1167/15.12.931.

[71] M. Q. Hill, S. Streuber, C. A. Hahn, M. J. Black, and A. J. O’Toole, “Creating Body Shapes From Verbal Descriptions by Linking Similarity Spaces,” *Psychol. Sci.*, vol. 27, no. 11, pp. 1486–1497, 2016, doi:

10.1177/0956797616663878.

[72] S. Streuber *et al.*, “Body Talk: Crowdshaping Realistic 3D Avatars with Words,” *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–14, 2016, doi: 10.1145/2897824.2925981.

[73] S. Windhager, K. Patocka, and K. Schaefer, “Body fat and facial shape are correlated in female adolescents,” *Am. J. Hum. Biol.*, 2013.

[74] C. Mayer, S. Windhager, K. Schaefer, and P. Mitteroecker, “BMI and WHR are reflected in female facial shape and texture: A geometric morphometric image analysis,” *PLoS One*, 2017.

[75] S. Pujades *et al.*, “The Virtual Caliper: Rapid Creation of Metrically Accurate Avatars from 3D Measurements,” *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 5, pp. 1887–1897, 2019, doi: 10.1109/TVCG.2019.2898748.

[76] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker, “Detailed human shape and pose from images,” in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[77] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, “Scape: shape completion and animation of people,” in *ACM SIGGRAPH 2005 Papers*, vol. 24, no. 3, 2005, pp. 408–416.

[78] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, “Joint training of a convolutional network and a graphical model for human pose estimation,” in *Advances in neural information processing systems*, 2014, pp. 1799–1807.

[79] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, “Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image,” in *European Conference on Computer Vision*, 2016, pp. 561–578.

[80] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, “Coarse-to-fine volumetric prediction for single-image 3D human pose,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7025–7034.

[81] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, “End-to-end recovery of human shape and pose,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7122–7131.

[82] R. Ball and J. F. M. Molenbroek, “Measuring Chinese Heads and Faces,” *Proc. 9th Int. Congr. Physiol. Anthropol. , Hum. Divers. Des. life*, no. January 2008, pp. 150–155, 2008.

[83] G. K. L. L. Tam *et al.*, “Registration of 3d point clouds and meshes: A survey from rigid to Nonrigid,” *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 7, 2013.

[84] O. van Kaick, H. Zhang, G. Hamarneh, and D. Cohen-Or, “A survey on shape correspondence,” *Eurographics Symp. Geom. Process.*, vol. 30, no. 6, pp. 1681–1707, 2011, doi: 10.1111/j.1467-8659.2011.01884.x.

[85] X. Li and S. S. Iyengar, “On Computing Mapping of 3D Objects,” *ACM*

Comput. Surv., vol. 47, no. 2, pp. 1–45, 2014, doi: 10.1145/2668020.

[86] W. Feller, “An Introduction to Probability Theory and Its Applications: Volume One.” John Wiley & Sons, 1950.

[87] F. L. Bookstein, “Principal Warps: Thin-Plate Splines and the Decomposition of Deformations,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 6, pp. 567–585, 1989, doi: 10.1109/34.24792.

[88] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele, “Building statistical shape spaces for 3D human modeling,” *Pattern Recognit.*, vol. 67, pp. 276–286, 2017, doi: 10.1016/j.patcog.2017.02.018.

[89] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. R. Stat. Soc. Ser. B*, 1996.

[90] X. Xiong and F. la Torre, “Supervised descent method and its applications to face alignment,” 2013.

[91] J. F. Cohn, “Dense 3D Face Alignment from 2D Videos in Dense 3D Face Alignment from 2D Videos in Real-Time,” no. MAY, 2015, doi: 10.1109/FG.2015.7163142.

[92] J. Domjanic, M. Fieder, H. Seidler, and P. Mitteroecker, “Geometric morphometric footprint analysis of young women,” *J. Foot Ankle Res.*, 2013.

[93] J. Domjanic, H. Seidler, and P. Mitteroecker, “A combined morphometric analysis of foot form and its association with sex, stature, and body mass,” *Am. J. Phys. Anthropol.*, vol. 157, no. 4, pp. 582–591, 2015, doi: 10.1002/ajpa.22752.

[94] D. Ferriman and J. D. Gallwey, “Clinical assessment of body hair growth in women,” *J. Clin. Endocrinol. Metab.*, vol. 21, no. 11, pp. 1440–1447, 1961.

[95] C. C. Dierickx, M. C. Grossman, W. A. Farinelli, and R. R. Anderson, “Permanent hair removal by normal-mode ruby laser,” *Arch. Dermatol.*, vol. 134, no. 7, pp. 837–842, 1998.

[96] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[97] V. Lempitsky and A. Zisserman, “Learning to count objects in images,” in *Advances in neural information processing systems*, 2010, pp. 1324–1332.

[98] L. Fiaschi, U. Köthe, R. Nair, and F. A. Hamprecht, “Learning to count with regression forest and structured labels,” in *Pattern Recognition (ICPR), 2012 21st International Conference on*, 2012, pp. 2685–2688.

[99] C. Arteta, V. Lempitsky, J. A. Noble, and A. Zisserman, “Interactive object counting,” in *European Conference on Computer Vision*, 2014, pp. 504–518.

[100] Y. Xie, F. Xing, X. Kong, H. Su, and L. Yang, “Beyond classification: structured regression for robust cell detection using convolutional neural network,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 358–365.

[101] J. Long and E. Shelhamer, “Fully Convolutional Networks for Semantic Segmentation,” 2015.

[102] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[103] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv Prepr. arXiv1409.1556*, 2014.

[104] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[105] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[106] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

[107] S. Aich and I. Stavness, “Object Counting with Small Datasets of Large Images,” *arXiv Prepr. arXiv1805.11123*, 2018.

[108] N. Fallah, H. Gu, K. Mohammad, S. A. Seyyedsalehi, K. Nourijelyani, and M. R. Eshraghian, “Nonlinear Poisson regression using neural networks: a simulation study,” *Neural Comput. Appl.*, vol. 18, no. 8, pp. 939–943, 2009.

[109] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv Prepr. arXiv1412.6980*, 2014.

[110] A. Paszke *et al.*, “Automatic differentiation in PyTorch,” 2017.

[111] D. Bradley and G. Roth, “Adaptive thresholding using the integral image,” *J. Graph. tools*, vol. 12, no. 2, pp. 13–21, 2007.

[112] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

[113] W. Kabsch, “A solution for the best rotation to relate two sets of vectors,” *Acta Crystallogr. Sect. A Cryst. Physics, Diffraction, Theor. Gen. Crystallogr.*, vol. 32, no. 5, pp. 922–923, 1976.

[114] R. S. Stern, “Prevalence of a history of skin cancer in 2007: results of an incidence-based model,” *Arch. Dermatol.*, vol. 146, no. 3, pp. 279–282, 2010.

[115] G. P. Guy Jr, S. R. Machlin, D. U. Ekwueme, and K. R. Yabroff, “Prevalence and Costs of Skin Cancer Treatment in the US, 2002- 2006 and 2007- 2011,” *Am. J. Prev. Med.*, vol. 48, no. 2, pp. 183–187, 2015.

[116] A. Huang, S.-Y. Kwan, W.-Y. Chang, M.-Y. Liu, M.-H. Chi, and G.-S. Chen, “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)*, 2013, pp. 3315–3318.

[117] P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Sci. data*, vol. 5, p. 180161, 2018.

[118] M. Combalia *et al.*, “BCN20000: Dermoscopic lesions in the wild,” *arXiv Prepr. arXiv1908.02288*, 2019.

[119] N. Codella *et al.*, “Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC),” pp. 1–12, 2019, [Online]. Available: http://arxiv.org/abs/1902.03368.

[120] Chowis, “https://chowis.com/hair-ai-diagnostic-technology/,” 2019. .

[121] R. Hoffmann, “TrichoScan Ein neues Werkzeug fr die digitale Haarzhlung,” *Der Hautarzt*, vol. 12, no. 53, pp. 798–804, 2002.

[122] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[123] M. Buda, A. Saha, and M. A. Mazurowski, “Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm,” *Comput. Biol. Med.*, vol. 109, pp. 218–225, 2019.

[124] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, “Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations,” in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, Springer, 2017, pp. 240–248.

[125] V. Iglovikov, S. Mushinskiy, and V. Osin, “Satellite imagery feature detection using deep convolutional neural network: A kaggle competition,” *arXiv Prepr. arXiv1706.06169*, 2017.

[126] A. Esteva *et al.*, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, p. 115, 2017.

[127] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[128] F. Perez, C. Vasconcelos, S. Avila, and E. Valle, “Data augmentation for skin lesion analysis,” in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, Springer, 2018, pp. 303–311.

[129] O. Russakovsky *et al.*, “ImageNet Large Scale Visual Recognition Challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015, Accessed: Feb. 21, 2018. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.693.9561&rep=rep1&type=pdf.

[130] I. Goodfellow *et al.*, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[131] T. J. Brinker *et al.*, “Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task,” *Eur. J. Cancer*, vol. 113, pp. 47–54, 2019.

[132] A. Brock, J. Donahue, and K. Simonyan, “Large scale gan training for high fidelity natural image synthesis,” *arXiv Prepr. arXiv1809.11096*, 2018.

[133] A. van den Oord, O. Vinyals, and others, “Neural discrete representation learning,” in *Advances in Neural Information Processing Systems*, 2017, pp. 6306–6315.

[134] A. Grover, M. Dhar, and S. Ermon, “Flow-gan: Combining maximum likelihood and adversarial learning in generative models,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32, no. 1.

[135] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” *2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 248–255, 2009, doi: 10.1109/CVPRW.2009.5206848.

[136] A. D’Amour *et al.*, “Underspecification Presents Challenges for Credibility in Modern Machine Learning,” *arXiv Prepr. arXiv2011.03395*, 2020.

[137] A. Ghorbani, V. Natarajan, D. Coz, and Y. Liu, “DermGAN: Synthetic Generation of Clinical Skin Images with Pathology,” *arXiv Prepr. arXiv1911.08716*, 2019.

[138] Y. Chi, L. Bi, J. Kim, D. Feng, and A. Kumar, “Controlled synthesis of dermoscopic images via a new color labeled generative style transfer network to enhance melanoma segmentation,” in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 2591–2594.

[139] C. Baur, S. Albarqouni, and N. Navab, “MelanoGANs: high resolution skin lesion synthesis with GANs,” *arXiv Prepr. arXiv1804.04338*, 2018.

[140] A. Bissoto, F. Perez, E. Valle, and S. Avila, “Skin lesion synthesis with generative adversarial networks,” in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, Springer, 2018, pp. 294–302.

[141] I. S. A. Abdelhalim, M. F. Mohamed, and Y. B. Mahdy, “Data augmentation for skin lesion using self-attention based progressive generative adversarial network,” *Expert Syst. Appl.*, vol. 165, p. 113922, 2021.

[142] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *2nd Int. Conf. Learn. Represent. ICLR 2014 - Conf. Track Proc.*, no. Ml, pp. 1–14, 2014.

[143] D. P. Kingma and M. Welling, “An introduction to variational autoencoders,” *arXiv Prepr. arXiv1906.02691*, 2019.

[144] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[145] J. De Fauw, S. Dieleman, and K. Simonyan, “Hierarchical autoregressive image models with auxiliary decoders,” *arXiv Prepr. arXiv1903.04933*, 2019.

[146] W. Williams, S. Ringer, T. Ash, D. MacLeod, J. Dougherty, and J. Hughes, “Hierarchical Quantized Autoencoders,” *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020.

[147] A. den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, and others, “Conditional image generation with pixelcnn decoders,” in *Advances in neural information processing systems*, 2016, pp. 4790–4798.

[148] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[149] P. Ramachandran *et al.*, “Fast generation for convolutional autoregressive models,” *arXiv Prepr. arXiv1704.06001*, 2017.

[150] D.-H. Lee, “Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks,” in *Workshop on challenges in representation learning, ICML*, 2013, vol. 3, no. 2.

[151] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, “Classmix: Segmentation-based data augmentation for semi-supervised learning,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1369–1378.

[152] F. Nachbar *et al.*, “The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions,” *J. Am. Acad. Dermatol.*, vol. 30, no. 4, pp. 551–559, 1994.

[153] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International Conference on Machine Learning*, 2019, pp. 6105–6114.

[154] S.-L. Liu, Y. Liu, L.-F. Dong, and X. Tong, “RAS: A Data-Driven Rigidity-Aware Skinning Model For 3D Facial Animation,” in *Computer Graphics Forum*, 2020, vol. 39, no. 1, pp. 581–594.

[155] E. Carrigan, E. Zell, C. Guiard, and R. McDonnell, “Expression Packing: As-Few-As-Possible Training Expressions for Blendshape Transfer,” in *Computer Graphics Forum*, 2020, vol. 39, no. 2, pp. 219–233.

[156] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, “Learning a model of facial shape and expression from 4D scans.,” *ACM Trans. Graph.*, vol. 36, no. 6, pp. 191–194, 2017.

[157] H. Valev, A. Gallucci, T. Leufkens, J. Westerink, and C. Sas, “Applying Delaunay Triangulation Augmentation for Deep Learning Facial Expression Generation and Recognition,” in *Pattern Recognition. ICPR International Workshops and Challenges*, 2021, pp. 730–740.

[158] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–

1708.

[159] G. Varol *et al.*, “Learning from synthetic humans,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 109–117.

[160] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, “Graph convolutional networks: a comprehensive review,” *Comput. Soc. Networks*, vol. 6, no. 1, pp. 1–23, 2019.

[161] A. Van Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel recurrent neural networks,” in *International conference on machine learning*, 2016, pp. 1747–1756.

[162] X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel, “PixelSNAIL: An improved autoregressive generative model,” *35th Int. Conf. Mach. Learn. ICML 2018*, vol. 2, pp. 1364–1372, 2018.

[163] R. Davies, C. Twining, and C. Taylor, *Statistical models of shape: Optimisation and evaluation*. Springer Science \& Business Media, 2008.

[164] V. F. Abrevaya, A. Boukhayma, S. Wuhrer, and E. Boyer, “A decoupled 3D facial shape model by adversarial training,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9419–9428.

[165] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2face: Real-time face capture and reenactment of rgb videos,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.

[166] D. Vlasic, M. Brand, H. Pfister, and J. Popovic, “Face transfer with multilinear models,” in *ACM SIGGRAPH 2006 Courses*, 2006, pp. 24--es.

[167] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, “A 3d morphable model learnt from 10,000 faces,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5543–5552.

[168] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni, “Regressing robust and discriminative 3D morphable models with a very deep neural network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5163–5172.

[169] X. Gu, S. J. Gortler, and H. Hoppe, “Geometry images,” in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 2002, pp. 355–361.

[170] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *International conference on machine learning*, 2017, pp. 214–223.

[171] R. Slossberg, G. Shamai, and R. Kimmel, “High quality facial surface and texture synthesis via generative adversarial networks,” in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, p. 0.

[172] G. Shamai, R. Slossberg, and R. Kimmel, “Synthesizing facial photometries

and corresponding geometries using generative adversarial networks," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 15, no. 3s, pp. 1–24, 2019.

[173] S. Moschoglou, S. Ploumpis, M. A. Nicolaou, A. Papaioannou, and S. Zafeiriou, "3dfacegan: Adversarial nets for 3d face representation, generation, and translation," *Int. J. Comput. Vis.*, vol. 128, pp. 2534–2551, 2020.

[174] T. Bagautdinov, C. Wu, J. Saragih, P. Fua, and Y. Sheikh, "Modeling facial geometry using compositional vaes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3877–3886.

[175] V. F. Abrevaya, S. Wuhrer, and E. Boyer, "Multilinear autoencoder for 3d face model learning," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 1–9.

[176] K. Li, J. Liu, Y.-K. Lai, and J. Yang, "Generating 3D Faces using Multi-column Graph Convolutional Networks," in *Computer Graphics Forum*, 2019, vol. 38, no. 7, pp. 215–224.

[177] O. van Kaick, H. Zhang, G. Hamarneh, and D. Cohen-Or, "A survey on shape correspondence," *Eurographics Symp. Geom. Process.*, 2011.

[178] K. M. Robinette and H. Daanen, "Lessons Learned From Caesar: a 3-D Anthropometric Survey," no. May, p. 5, 2003, Accessed: Dec. 05, 2017. [Online]. Available: http://childergo.com/ASC031101.pdf.

[179] D. Siegmund, T. Samartzidis, N. Damer, A. Nouak, and C. Busch, "Virtual Fitting Pipeline: Body Dimension Recognition, Cloth Modeling, and On-Body Simulation.," *VRIPHYS*, vol. 14, pp. 99–107, 2014.

[180] B. Allen, B. Curless, and Z. Popovic, "Exploring the space of human body shapes : data-driven synthesis under anthropometric control," *Digit. Hum. Model. Des. Eng. Symp.*, no. c, pp. 1–4, 2004, doi: doi:10.4271/2004-01-2188.

[181] M. Loper, N. Mahmood, J. Romero, G. Pons-moll, and M. J. Black, "SMPL : A Skinned Multi-Person Linear Model," *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1--248:16, 2015, doi: 10.1145/2816795.2818013.

[182] W. Zeng, W. Ouyang, P. Luo, W. Liu, and X. Wang, "3d human mesh regression with dense correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7054–7063.

[183] O. Sorkine, "Laplacian mesh processing," in *Eurographics (State of the Art Reports)*, 2005, pp. 53–70.

[184] H. Kato, Y. Ushiku, and T. Harada, "Neural 3d mesh renderer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3907–3916.

[185] R. V Zicari *et al.*, "Z-Inspection®: A Process to Assess Trustworthy AI," *IEEE*

Trans. Technol. Soc., 2021.

[186] R. V Zicari *et al.*, “How to Assess Trustworthy AI in Practice.” arXiv, 2022, doi: 10.48550/ARXIV.2206.09887.

[187] R. V Zicari *et al.*, “Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier,” *Front. Hum. Dyn.*, vol. 3, p. 40, 2021, doi: 10.3389/fhumd.2021.688152.

SUMMARY

The skin is the largest human organ and performs essential biological functions like protection, regulation, and sensation. These functions are often impaired or changed due to natural conditions or human intervention and lifestyle. For example, human interventions applying products to preserve skin beauty are common when the skin is aging. Similarly, skin ageing also impairs the normal protection against DNA damage and, therefore, it can lead to skin cancer. Skin specialists can restore the functions and detect and prevent diseases by looking at and analyzing the skin surface and texture. Their work can be improved with the help of computer vision – a scientific field under artificial intelligence that studies how computers can observe, recognize, and interact with the world via visual inputs such as videos, images, and 3D data.

Within the computer vision fields, several archetypal tasks can be considered: classifying and labeling objects in the images, counting the total number of instances of an object, detecting a human constitution through the analysis of 3D scans, and health state evaluation through a sequence of images. By applying computer vision to parametrize and analyze the skin surface distribution and the local texture, one can improve personal care devices such as shavers, epilators, and skin care products and support dermatologists in their daily work. In this dissertation, we leverage and improve modern computer vision algorithms to enable various human skin applications. We use shape analysis and synthetic generation of skin surfaces to reduce the costs of acquiring and using precious 3D human scans. We use artificial neural networks to detect, classify and count skin features, such as potentially harmful skin lesions or skin hair.

We improve the analysis of 3D human skin surfaces by defining a novel methodology to infer the surface shape distribution from sparse measurements like weight, age, height, and arm length. Predictive statistical models can estimate skin surfaces given a set of anthropometric measurements with superior accuracy. Successively, we leverage deep generative models –a subfield of artificial neural networks aiming at generating synthetic data– to generate novel 3D scans. Our contribution is a novel approach that converts 3D representation into a 2D one, where a generative model is then applied, enabling the generation of surfaces instead of textures. The quantitative comparison of synthetic test shapes versus real ones demonstrates the strength of this approach in terms of diversity and fidelity.

Besides the analysis of 3D skin, the dissertation investigates the use of deep learning to detect local skin features to enable several applications. Usually, the screening and diagnosis of skin lesions like melanomas are primarily carried out by clinical visual inspection and biopsy if necessary. However, this process is slow and there is a pressing need to reduce the amount of work of dermatologists and skin specialists employing computer assistant telemedicine. We use state-of-the-art architectures to automatically count skin hair in a tiny skin patch, to understand wheatear skin hair presence influence skin lesions classification and generate novel skin lesions for training new classification models.

To count skin hair, we collected a dataset of 4000 skin patches from more than 100 volunteers. Based on this data, we trained a prediction model that can count hairs with an accuracy of ~6 hair out of the ~55 in the images. While not beating an expert human annotator, the automatic method is extremely fast and deemed sufficient for an industrial evaluation of epilation devices. Building on top of learned knowledge in the hair domain, we investigate how their presence may influence the diagnosis of skin lesions. We trained neural networks with and without hair, demonstrating that their presence does not hamper the ability to classify skin cancer correctly. Hence, we show that shaving prior to acquiring skin pictures is not required, thus facilitating an early diagnosis of potential skin cancer.

One clear challenge to achieving great prediction accuracy for skin lesions is the lack of a large amount of annotated data. Collecting real samples is time-consuming and difficult, for example, considering rare diseases. In this work we propose using a novel approach to generate synthetic skin lesions to augment and increase the number of images in the lesions' datasets using state-of-the-art neural networks. This offers benefits including directly modifying a local part of an input lesion without affecting its global structure. The quantitative results are promising but still show that the synthetic data are not good enough to improve downstream tasks such as the classification one.

To conclude, in this work we explored several multiscale computer vision methodologies and confirmed the potential of deep learning to improve the analysis and parametrization. While doing so, we presented new compelling applications that are currently used in the industry practice or that will soon be deployed and adopted.

CURRICULUM VITAE

Alessio Gallucci was born in August 1991 in Moncalieri, Turin, Italy. He showed a passion for mathematics and computers from early years, reason why he decided to combine them and pursuing a degree in Applied Mathematics, obtaining both his Bachelor's and Master's degree at Politecnico di Torino, Italy.

During the Bachelor he developed fundamental geometrical and algebraic skills while in the Master's he went on to build a solid understanding of statistics, algorithms, and artificial intelligence. During his Master's degree in 2015 he had a first working experience as a Data Scientist Intern at Elis Consulting in Rome, where he learnt industry practices and developed further understanding of artificial intelligence in the field of natural language processing. After this experience, he decided to combine his Master's Thesis with an internship and joined Philips Research, working on 3D modeling of human bodies.

He then decided to expand his knowledge on artificial intelligence and computer vision for personal and healthcare applications and started his PhD, which he is currently pursuing, at Eindhoven University of Technology, The Netherlands, in collaboration with Philips Research. During his PhD, Alessio was an active member of the Philips Interns and PhD communities where he developed is communication and organizational expertise. In 2022, he joined Philips Research as a data scientist where he is continuing his research.

The collaboration between academia and industry resulted in ten scientific publications, two patent application and several technical notes and inventions disclosures. His current research interests include deep learning, digital image processing, 3D registration and trustworthy AI.

STUDENT SUPERVISION

All the students were supervised jointly with the co-promotor Dmitry Znamenskiy following his leadership.

2022 Sun, Xinyu, 3D Body Estimation from 2D representations

2021 Yuxuan Long, EPFL, 3D Body Reconstruction from 2D images

2020 Florian Delberghe, EPFL, Skin Mapping and Navigation

2019 Jawahar Ponathipan, TU/e, 3D Head Reconstruction

REVIEWER SUPPORT

2022 Towards a Complete Analysis of People: From Face and Body to Clothes Workshop at ICPR

2021 International Journal of Advanced and Applied Sciences

2021 WSCG International Conferences in Central Europe on Computer Graphics, Visualization and Computer Vision

2021 ISAIC International Symposium on Automation, Information and Computing

AWARDS

2022 Towards a Complete Analysis of People ICIAP 2021, Best paper

2021 Philips, Grooming and Beauty Global Creation Event 2021, Winner

2019 ICBSP, Best Presentation in session one

ACKNOWLEDGMENTS

This PhD thesis marks the end of a great and arduous journey. Through a philosophical, spiritual, and ethical inner quest many identities have died, and new ones have been born. A small hint of it can be appreciated by reading the quotes presented in each chapter. Clearly, the main drivers of the quest are the people surrounding me and the emotions that, created, sparked and light constantly my fire.

First, thanks to my promotor Milan for giving me the opportunity to take on the challenges of the PhD and thanks for giving me the tools to overcome them. You kindly but timely pointed out my hidden biases showing me the right path. Thanks to my supervisor Dmitry for always being present (days and nights), for inspiring me and for giving me always the spirit and energy to produce great work. Thank you for the great life lessons you always implicitly and silently give with your behaviors. Thanks to my advisor Nicola for giving many great advice. Your introduction in the middle of the path contribution has been crucial for the success of the PhD. Thanks to Jim for the crucial support. Thanks to the committee members and thanks to my great other many other colleagues and advisors in Philips and TU/e. Thanks to Pauline, Odette and Mirjam for their constant present in times of bureaucratic needs. Thanks to Mounir who keeps teaching me new great lessons since day one. Thanks to Erik for being a great colleague and life coach. Thanks to Calina for our warm coffees. Thanks to Roman for our cold beers. Thanks again Dmitry for being not only a colleague but also a friend.

Thanks to Pepo for teaching me not to judge. For being a great hidden life teacher apart, of course, for teaching me chess. Thanks to Alessandro for our amazing connection and understanding; for being the one I want to have always in my team and much more. Thanks to Joshua for showing me the peace and for always making me feel free to express and be myself. Thank you being so close to me during so many years. It is great to see how we change and grow together. Thanks to Manuel for always being playful and funny; please never lose it. Thank you for being a life-friend no matter what. Thanks to Antonio for your equanimity and your belly. Thank you also for many other things, nights, and days. Thanks to Luca for spending time sharing on books and always giving me so much energy. Thanks to Simone for our long-standing friendship, for our chill and natural sharing of disparate topics and projects. Thanks to Ezio and Daniela for being so their big heart. Thanks to Renato for supporting me for many years.

Thanks to Julien and Rose for paving my sustainability path. Who knows if you will condition my future career choices. Thanks to my gosht friend Sandro whom I shared great part of my childhood and whom nature I've encountered. Thanks to him my skills and training in videogames, and, therefore in life, have born and rise. Thanks to the Hamsters. Thanks to Luis, Sauvik, and Andra for being present since day 1 of the journey and always supporting me.

Thanks to Andrea. Thank you for making my university life bearable and fun. Thanks for all the time spent together. Thanks for all our beers and all our talks and all our understanding. Thanks also for our misunderstanding. You know that I am never

afraid of them and remember that I always believe in you. Thanks to Niccolò for being so full of energy. Thanks for being always ready to show me new life forces. Thanks for the great vibes in this continent and the ones you always offer me in others which soon I hope to catch.

Thanks to Moussa for stepping up my game. For listening to my complaints and for your true understanding. For showing me a new world and new culture. Thanks for our ability to put aside strongholds in favors of sweet hills. Thanks to Alessandro and Daan for sharing great Tuesdays, office, swim and chess and more games.

Thanks to my great chess big family and thanks to the SST which provided me a second home and many lovely memories: to remember just one out the many I can refer to a milestone for this journey, the BSc degree party. Thanks to my soccer friends, in particular, ohhh Deportivo, La Scoccaccio, and friends in general. Thank you, guys, for letting my rage flow freely in the pitch. Thanks to the Philips Interns Community for all the activities done together, all the moment shared. Thanks to An, Melanie, Meghna and Boncho for making me feel at home in the Netherlands. Thanks to the Philips PhD Community for helping me become a responsible adult. Thanks to Steven for being always honest and making me grow like no one. Thanks to Marco, Shin, Marta, Olek, Maretha, Anastasia, Irene, Manolis, Vladislava, Nuoya, Rachna, Tom, Hakim, Laura, Shaniq, Siebren, Giuseppina, Tony, Simone, Gianluca, Vide, Umberto, Edoardo and Francesca. Thanks to Hristo for being a deeply needed beer partner and for our great talks and sharing. Thanks to Thanks to all the people I will not list here. Thanks to all the recent and old friends. Thanks to all my dearest friends the one written and the one missed because of lack of time in writing this section.

Thanks to my additional brother Francesco. You always give meaning to my life. Our time is never enough but also always also good enough to make me happy.

Thanks to my large family. To my aunts and uncles who have always been very close to me and provided me with additional places where to feel at home during my life. Thanks to my dear cousins which are like brothers to me. Thanks to Lucia, Roberto, Eleonora, Ruben, Alberto, Orsi for letting me be a son and a brother. Thanks to Barbara, Maurizio for being second parents. Thanks to Federico, Andrea, and Alessandro. Thanks, Carlo, for always making me laugh. Thanks to Daniela for our true understanding and all the fun we always have together. Thanks to Claudia, Valerio, Sara e Ilaria.

Thanks to my father Flavio for showing me that being or having are ephemerals attachment of the mind. Thank you. Thanks to Eliana for showing me the path to forgiveness. Thanks to Giorgia for the great philosophical talks. You are very wise.

Thanks to my sister Giulia. Thank for being my family, for showing me how to love, for giving a place to feel at home. Thanks for giving me the space to be myself without judgment always and forever. Thanks to Stefano for being always openminded and Buddhist with me. Thank you for your generosity. Thanks to Matilde and Tommaso. Thank you for bringing me two opposite forces; for diverging my spirit by relighting the child that is in me while also letting me become a responsible uncle.

Thanks to my mother Mariangela for being my first and long-standing supporter. For believing always in me and my potential and encouraging me to pursue my careers and goals. Thank you. Thanks to Andrea for being always there, always; and providing the very needed safe space to grow. Thanks to Davide for showing me and a new way into arts and emotions.

Thanks to Giliana and Adriana. You two both knows how much I love you. Last, but not least, thanks to all other supporters that are not anymore present physically but constantly drive my efforts. Thanks to Franca, Martino and Eduardo. Thank you for being not only my north star but also my quasar wonder.

Very last thanks to the most important person, whom, without, I might still be a lost soul in the nowhere country. The person who allowed me to carry all the meaningful work. The person who showed me the path to the heart. The path to compassion. The path to goodness. Thanks, Ilaria.

Utrecht, in a place close by the station
August 2022, near by the end of the month but in the present moment
Alessio