DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Face Relighting in the Wild

Juan Raúl Padrón Griffe

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Face Relighting in the Wild

Beleuchtungsbearbeitung von Gesichtern in Portraitbildern

Author:	Juan Raúl Padrón Griffe
Supervisor:	Matthias Nießner
Advisor:	Justus Thies
Submission Date:	15-07-2020

I confirm that this master's thesis in informatics is my own work and I have documented all sources and material used.

Munich, 15-07-2020 Juan Raúl Padrón Griffe

Acknowledgments

First of all, I would like to thank my supervisor Matthias Nießner for the unique opportunity to do a master thesis with an outstanding research group like the Visual Computing lab. I want to highlight his willingness to provide all the necessary resources and conditions to carry out the research project including the capture equipment, the training server and the strong background of the group. I am particularly grateful to my advisor Justus Thies for his guidance and feedback during the project. His incredible expertise and exceptional commitment were essential to overcome the difficulties of the topic. I want to mention that his 3D Scanning and Motion Capture lecture, co-organized with Angela Dai, inspired me to work on this exciting and fascinating field and broadened my horizons. Finally, I would also like to thank the fellow students at the lab for all the interesting discussions and practical tips.

On a personal level, I am very grateful for all the support of my close friends during my studies. In particular, I would like to express my sincere gratitude to the Tigges family for their help and kindness since the beginning of my studies. Finally, I must express my deepest gratitude to my family, specially my parents for their unconditional love and care. I dedicate this work to them, who motivate my interest and passion for science.

Abstract

Relighting plays an essential role in realistically transferring objects from a captured environment into another one. In particular, current applications like telepresence need to relit faces consistently with the illumination conditions of the target environment to offer an authentic immersive experience. Traditional physically-based methods for portrait relighting rely on an intrinsic image decomposition step, which requires to solve a challenging inverse rendering problem in order to obtain the underlying face geometry, reflectance material and lighting. Inaccurate estimation of these components usually leads to strong artifacts (e.g. artificial highlights or ghost effects) in the subsequent relighting step. In recent years, several deep learning architectures have been proposed to address this limitation. However, none of them are free from these artifacts.

In this thesis, we propose a general framework for automatic relighting enhancement using the StyleGAN generator as a photorealistic portrait prior. Specifically, we apply the ratio image-based face relighting to an artificial portrait dataset generated using the StyleGAN model. Next, we refine this dataset by projecting back the relit samples into the StyleGAN space. Then, we train an autoencoder network to relit portrait images from a source portrait image and a target spherical harmonic lighting. We evaluate the proposed method on our synthetic dataset, the Laval face and lighting dataset and the Multi-PIE dataset both qualitatively and quantitatively. Our experiments prove that this method can enhance the state of the art single portrait relighting algorithm for synthetic datasets. Unlike this algorithm, we achieve these results relying on a synthetic dataset five times smaller employing a traditional training scheme.

Contents

1. Introduction

Compositing virtual objects into photographs or videos is a fundamental technique in applications like visual effects, augmented reality, product visualization and telepresence. In particular, the illumination of the virtual objects should be consistent with respect to the lighting conditions of the surrounding environment. This task of rendering scenes under novel lighting conditions is known as *relighting* and it has been a long studied vision and graphics challenge [1, 2].

Reconstruction based relighting is a traditional approach used in the community to address the relighting problem. In this approach, an inverse rendering step is required to recover the current geometry, illumination and reflectance of the scene from the capture images. Once the image intrinsics are estimated, the scene can be rendered under any novel lighting condition. Barron and Malik [1] estimate simultaneously shape, illumination and reflectance from a single image relying on strong priors like Lambertian reflectance and low-dimensional shape spaces. These assumptions indeed impact significantly the realism of the final rendering and the imperfect 3D content leads to rendering artifacts.

Image based relighting introduced in the seminal work by Debevec et al. [2] is another classical approach, which avoid the explicit reconstruction by approximating directly the light transport function as a linear combination of the capture images under a fixed viewpoint. The results are photo-realistic and in fact the technique has been used to create virtual actors in films such as Spider-Man 2, The Avengers and Avatar. However, this strategy requires a large number of images to be captured using a complex acquisition equipment known as a light stage or light cage. In this calibrated multi-view setup, a human subject is recorded under various lighting conditions by using one light at the time from a large number of white LED lights surrounding the subject.

Recent works [3, 4, 5, 6, 7, 8] rely on Deep Learning advances to address the limitations of traditional approaches, specially on end-to-end autoencoder architectures. Sengupta et al. [3] decompose portrait images in the wild conditions into shape, reflectance and illuminance using an end-to-end framework, which learns from a mixture of labeled synthetic data (low frequency variations) and unlabeled real world images (high frequency details). Xu et al. [4] apply image based relighting on general scenes from only five images captured under specific directional lights by training a neural network that exploits the coherence of the light transport function across scenes. Meka et al. [7] perform human relighting using only a pair of images recorded under spherical gradient illumination.

Other works [5, 6] perform relighting and lighting estimation at the same time from a single image. Zhou et al. [5] use spherical harmonics for the target lighting representation and train an autoencoder to relight single portrait images on a synthetic dataset built with a traditional ratio image-based algorithm. An adversarial training is further applied to reduce the gap between the synthetic and real lighting. On the other hand, Sun et al. [6] train their network on captured reflectance field data to relit portrait images under arbitrary user-specified environment maps. The ground truth rendering consists on a weighted combination of the images in the light stage dataset according to the corresponding projections of the target low-resolution environment map onto the LED basis. In a similar line of work, Nestmeyer et al. [8] train an end-to-end architecture to both de-light and relight human faces under directional lighting. The architecture consists on a physics-based diffuse render and a residual correction network to handle non-diffuse effects (e.g. cast shadows or specular highlights).

A common practice across these works is the fact the training data is paired, .ie., the only difference between the original image and the target image is the illumination. To satisfy this property, previous works rely on light stage datasets [6, 7, 8] or synthetic datasets [3, 5]. There is a enormous gap between the quality of models trained on real data and models based synthetic illumination. For instance, the synthetic illumination proposed by Zhou et al. [5] produces artifacts like artificial highlights or ghost effects, especially notorious around the nose as can be seen in figure 1.1. Unfortunately, light stage datasets of human faces are not openly available.

Face image synthesis has achieved a tremendous success in the past few years due to the fast progress of Generative Adversarial Networks (GANs) [9]. State-of-the art GAN models, such as the recent StyleGAN [10], can generate high-quality virtual face images that are sometimes even hard to distinguish from real ones. Furthermore, the novel architecture enables some control over a variety of face attributes depending on the scale: coarse (geometry, pose), medium (expressions, facial hair) and fine (color scheme and microstructures). Despite of the promising results, the attributes cannot be fully controlled in a intuitive way as the traditional modeling and rendering tools. In addition, several semantic attributes are still entangled such as facial hair and illumination or heap pose and face identity.

In this work, we propose a face relighting method that combines the strengths of the high photo-realism of generative face models with the recent deep-learning based relighting methods. We employ a fixed and pretrained StyleGAN generator to automatically refine synthetic relighting datasets by projecting the relit images into the StyleGAN space. This projection can be interpreted as a photo-realistic constraint, where artifacts such as aliasing or wrong shadows would potentially be removed (figure 1.1). Our experiments show that this refinement can improve the results of the deep single image portrait relighting network [5], which is one the state of the art algorithms on portrait relighting. Finally, the proposed method can be seamlessly applied to different deep-learning based relighting system as the mentioned above.

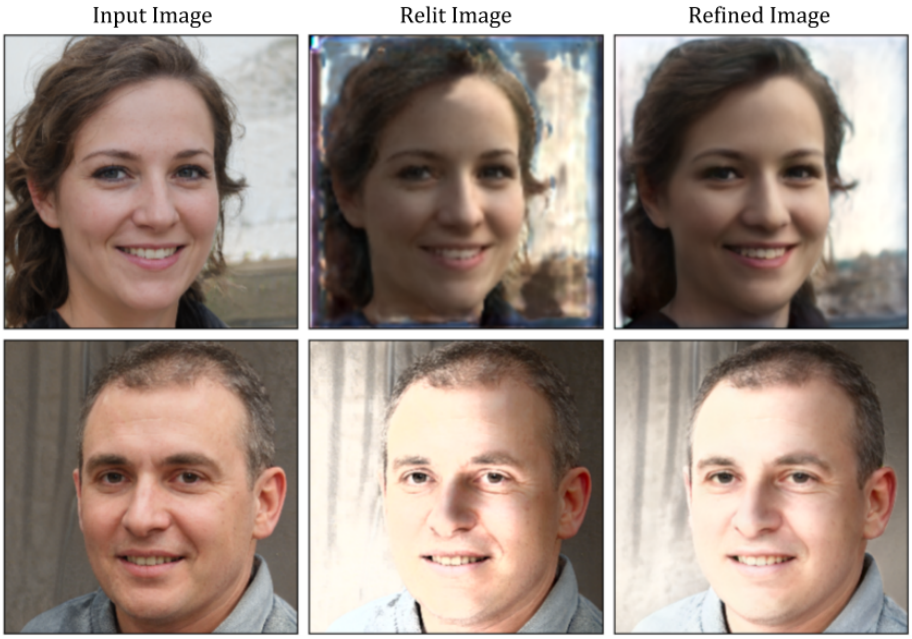

Figure 1.1.: StyleGAN refinement on synthetic illumination: SfSNet [3] (first row) and DPR [5] (second row).

2. Related Work

Portrait relighting can be seen as a special case of image relighting or a direct application of scene reconstruction. Style transfer is another potential approach as the lighting can be considered part of the style of headshot portraits. In this section, we will review briefly the most significant works in the field and discuss how they are related to the face relighting task. We will cover classical approaches from the computer graphics and computer vision communities as well as recent deep learning based approaches.

2.1. Inverse Rendering of Portrait Images

Decomposing an image into its intrinsic components (e.g. shading, reflectance, and shape) [11] is one of the fundamental challenges in computer vision, which is known as intrinsics image decomposition or inverse rendering. *Shape from Shading* [12] is a particular scenario, where the algorithms try to recover only the geometry assuming the materials and illumination are known or fixed. The formulation is still ill-posed even for this simplified case as different shapes could be inferred from the same image [13]. Barron and Malik [1] propose *SIRFS*, the first technique to simultaneously recover shape, illumination, reflectance and shading from a single image. The authors rely on extensive priors to guide the inverse rendering optimization procedure. Once the image is decomposed, the scene can be relighted by changing the illumination and keeping the other components fixed.

A 3D Morphable Model (3DMM) [14] is a common face representation, where faces are parameterized by the identity, expressions, skin reflectance and scene illumination. Expressions are commonly modeled using blend shapes and illumination is generally modeled via spherical harmonics parameters. The models are traditionally learned via Principal Component Analysis (PCA) based on scans of human faces [15, 16]. Recent works focus more on raw Internet footage because the 3D scan data is limited both in size and variety.

Shu et al. [17] integrate the image formation and shading computation as layers in an end-to-end network, which is trained on portraits in-the-wild by weak supervision on prior knowledge of the physical process. For instance, the Retinex Theory [18] indicates that the shading should be spatially smooth and the reflectance should be statistically sparse. Similarly, Sengupta et al. [3] introduce *SfSNet*, a end-to-end CNN architecture for inverse face rendering trained on a mixture of labeled synthetic and unlabeled real world images. Low-frequency variations are learned from synthetic 3DMM examples based on the traditional

diffuse rendering model, while simultaneously the high-frequency details are inferred from real data using shading cues through the photometric reconstruction loss. Tewari et al.[19] create *MOFA*, a self-supervised training scheme for 3D human face reconstruction from a single RGB image. The end-to-end approach combines a convolutional encoder network with a fixed differentiable parametric decoder designed to render the semantic vector in the bottleneck, which is interpreted as a 3DMM model.

The main limitation of the 3DMM models is the lack of realism. For instance, mouth interiors, hair, or eyes are often not included by such representations. Inverse rendering methods tend to be computationally expensive due to the complexity of the task. As a result, most of them can only process low-resolution images. For example, SfSNet can only work on (128×128) images. For a detailed survey of 3D morphable face models over the last years please refer to the following sources [20, 21].

2.2. Image based Relighting

Debevec et al. [2] establish for the first time that relighting can be formulated as the linear combination of many images of a subject from the same viewpoint under different known illuminations. The approach is known as image based relighting and the images are referred as the reflective field. Image based relighting has been essential specially for realistic virtual characters in blockbuster films and games [22]. Images are typically acquired in a calibrated environment called light stage, where a group of synchronized LED light sources surround the subject. Then, cameras in front of the subject record images for a short period of time where each light source flash is turned on one at a time. Unfortunately, the solution is subject dependent and it requires around 2000 images per subject.

The light transport function is known to be highly coherent [23] and this fact inspired subsequent works to apply image based relighting on a small number of images using low-dimensional functions. Malzbender et al. [24] propose the Polynomial Texture Maps (PTM), where the radiance values of each pixel are modelled as polynomial functions of the lighting directions. Ren et al. [25] use a shallow neural network instead of polynomials to approximate the relighting function. The network was able to produce impressive complex light transport effects (e.g. caustics or interreflections) using around 100-900 images. Xu et al. [4] present a novel scheme to relight any scene under novel illumination from only five images by exploiting the light transport correlations across multiple scenes. For this purpose, the authors train an end-to-end network to jointly learn the optimal input lighting directions and the relighting function. The entire system is trained on a large synthetic dataset of random shapes with complex SVBRDF rendered using Mitsuba [26]. Xu et al. [27] later address novel viewpoint rendering of a single object from only six wide baseline views.

Thies et al. [28] present Deferred Neural Rendering, a new paradigm for image synthesis that combines the traditional graphics pipeline with learnable components. First, the neural

textures are sampled using a classical computer graphics rasterizer and a valid uv-map parameterization (texture coordinates) of a given object. Then, the final output image is generated from the rendered feature maps using a small U-Net (deferred neural renderer), which is trained in conjunction with the neural textures. This new rendering pipeline enables photo-realistic (re-)renderings of imperfect 3D content obtained from photo-metric reconstructions with noisy and incomplete surface geometry, which was applied successfully in tasks such as novel view synthesis, scene editing or face reenactment. As a part of my guided research project, we demonstrated that this pipeline can be extended to perform relighting tasks as well by concatenating the light information to the neural textures. For instance, a directional light can be encoded using three feature maps with the same spatial resolution of the neural textures, where each feature map corresponds to one of the Cartesian coordinates of the light direction. One caveat of the method is the fact it is trained for a specific scene or object.

Meka et al. [7] come on with Deep Reflectance Fields, a novel approach to decompose full reflectance fields by training a convolutional neural network that maps two spherical gradient images to any of the light stage images. The Relightables system by Guo et al. [29] achieved an unprecedented level of photorealism for a volumetric capture pipeline based on the same principle. Despite of the impressive results, most image based rendering algorithms rely heavily on high quality data from complex calibrated setup. A summary of recent trends and applications of neural rendering approaches has been published by Tewari et al. [30].

2.3. Portrait Style Transfer

Style transfer consists of applying the style of a reference image (e.g. painting) to an input image (e.g. city), preserving the general content of the input image. The seminal work of Gatys et al. [31] show for the first time that a deep neural network can be used to do this job. The style is preserved by matching the second-order statistics captured by the Gram matrix between the feature activations of the generated image and the reference image. This method can enforce arbitrary styles, but it requires an expensive optimization process per pair of content and style images. Johnson et al. [32] suggest to train a feedforward neural network per style in order to perform style transfer with a single forward pass. Huang et al. [33] introduce an approach to apply arbitrary styles to any input image in real time by performing an Adaptive Instance Normalization (AdaIn) in the feature space.

The illumination of a scene can be treated as a part of the style of the image. As a result, the style transfer techniques can also be applied for portrait relighting. In this case, the style image is a non-occluded reference of a subject with the desired lighting information. Shih et al. [34] use a classical multiscale technique to transfer the local image statistics of a reference facial portrait onto a new one, matching properties such as local contrast and the overall lighting direction. Shu et al. [35] formulate the relighting task as a mass-transport problem

between the input and reference images. The optimization is designed as a geometry-aware color transfer, where authors fit a 3D morphable face model to both input and reference portraits in order to extract the color (RGB), position (2D) and normal (3D) per pixel. The results are compelling, but the algorithm fails to add or remove non-diffuse effects. Furthermore, the method shares the aforementioned drawbacks of the 3DMM-based approaches.

In general, portrait style transfer methods require a reference image. Thus, they do not allow general purpose relighting based on arbitrary illumination conditions, which restricts the possible application scenarios.

2.4. Lighting Estimation

Predicting the lighting of a particular scene from a single image is a fundamental problem in computer vision. Hold-Geoffroy et al. [36] model outdoor lighting with the parametric Hosek-Wilkie sky model [37] and estimate its parameters from an individual image. Later Hold-Geoffroy et al. [38] propose DeepSky, a deep autoencoder that can directly estimate an HDR environment map of the outdoor lighting without relying on analytical models. Gardner et al. [39] use an end-to-end learning approach to estimate directly the illumination of indoor scenes from a single RGB image. Calian et al. [40] create the Face2Light system, which learns the space of outdoor lighting using a deep autoencoder and combine this latent space with an inverse optimization framework to predict lighting from a face.

Nishino and Nayar [41] suggest the eyes as a light probe. They recover high frequency lighting from the reflections of the subject eyes. LeGendre et al. [42] introduce DeepLight, a general deep-learning based method to estimate the illumination of general scenes from a mobile phone camera with a limited field of view (FOV). The ground truth HDR lighting is inferred only from LDR images using three spheres with different reflections properties: specular, glossy and diffuse. The authors build their own capture system by arranging the phone and the spheres in a fix distance using a selfie stick tripod, where the mirror balls can be seen in the bottom portion of the camera's FOV. The capture system is indeed effective and remarkable simple in comparison with previous works.

Overall, the methods for lighting estimation enable the rendering of new objects in a particular scene under the recovered illumination conditions. However, they do not address the relighting of existing objects.

2.5. Portrait Relighting

Marschner and Greenberg [43] suggest the ratio of two images under different lighting conditions in order to relit the first one under the lighting conditions of the second one.

Shashua and Riklin-Raviv [44] present the first paper using the notion of a quotient image for image-based relighting of human faces. Stoschek [45] combines the technique with image morphing based on facial landmarks alignment in order to support arbitrary poses. Wen et al. [46] relight faces using the ratio of radiance environment maps represented as spherical harmonics coefficients. Peers et al. [47] present an image-based post-production relighting system where the performance capture and lighting design are decoupled, .i.e., the reflectance field of a reference actor is transferred onto the dynamic performance of the same actor or a subject with similar appearance.

Zhou et al. [5] propose a single-image portrait relighting solution inspired by these works. A large scale portrait relighting dataset is created by applying the ratio image based trick to the high resolution CelebA (CelebA-HQ) dataset. A convolutional autoencoder network is then trained on this dataset to take a portrait image and a target lighting as input and produce a relit version as the output. In addition, the illumination of the portrait image is predicted from the bottleneck feature using a small fully connected network. As a consequence, the system can perform both relighting and light transfer. Finally, an adversarial loss is further applied to remove the artifacts caused by the ratio-image relighting algorithm.

In a similar line of work, Sun et al. [6] train an autoencoder on a small database of subjects captured under different directional light sources in a controlled light state setup. In this paper, the target light is represented as an environment map, while Zhou et al. [5] report the usage of spherical harmonics coefficients. The training data is built as a weighted combination of the light stage images according to the projection of the target environment map. A confidence-weighted average predicts the illumination of the source image, .i.e, the network predict a complete RGB environment map as well as a confidence map associated with the location in the bottleneck. The confidence learning and a cycle consistency loss contribute to improve the performance on the relight task as well as the lighting prediction accuracy.

The training dataset of Sun et al. [6] is a smooth version of the original light stage captures. Thus, the system cannot handle hard shadows and sharp specularities. Nestmeyer et al. [8] propose an end-to-end deep learning architecture to learn the face relighting function under strong directional lighting by training directly on the light stage data. The architecture consists of the integration of two stages. The first one is a U-NET generator with an explicit implementation of the diffuse rendering process. The second stage is a refinement U-NET that predict a residual correction and a binary visibility map, which are required to model non-diffuse effects such as specularities and cast shadows respectively. The intrinsic predictions are guided by losses with respect to the ground truth maps of the photometric stereo reconstruction.

Recent approaches bypass the image decomposition by using deep learning models to regress directly to the relit image. As a result, they enable relighting applications on consumer cameras. For example, the Sun et al. [6] system can relight 640×400 portrait images on a

mobile phone approximately in 160 ms. The representation of the light source and the quality of the training data are fundamental aspects of these algorithms. On the one hand, models trained on synthetic data tend to be unrealistic due to the strong assumptions like Lambertian reflectance and low-dimensional spaces for the shapes. On the other hand, algorithms based on light state captures could not generalize well enough to real scenarios. Furthermore, light stage data is really tough to acquire and it is unfortunately not openly available.

2.6. Deep Generative Models

In 2012, Krizhevsky et al. [48] beat all the competitors in the Large Scale Visual Recognition Challenge (ILSVRC) [49] by a large margin (10.8%) using a convolutional neural network. Since then, deep learning based approaches have dominated many computer vision tasks. In particular, face image synthesis has advanced rapidly in the past few years since Goodfellow et al. [9] introduced the Generative Adversarial Networks (GANs) in 2014. This model consists of two neural networks competing with each other. A generator network synthesizes an image from a noise vector, while a discriminator network try to distinguish between the synthetic and real samples. Later, Radford et al. [50] create DCGAN, the first implementation of convolutional and convolutional-transpose layers in both the generator and the discriminator.

Karras et al. [51] show for the first time that GANs can be trained to generate high-resolution photorealistic images of human faces by using a progressive training strategy where both the generator and discriminator are built layer by layer. This strategy turns out to be very effective in stabilizing and speeding up the training phase. Karras et al. [52] introduce StyleGAN, a state-of-the art GAN model, which can synthesize high-fidelity face images while allowing for more control over the output. Later, the same authors identified some issues such as water-droplet and phase artifacts in the synthetic images. They fixed them by redesigning the original synthesis network architecture in the follow-up paper [10].

Several research groups have recently explored the latent space of GANs for image editing. Abdal et al. [53] proposed an efficient embedding algorithm to map images into the latent space of a pre-trained StyleGAN model. They concluded that embedding into the extended latent space works better and that any kind of image can be embedded. Härkönen et al. [54] use PCA to find disentangled linear directions in the StyleGAN and BigGAN [55] latent manifolds and then they define interpretable controllers for image synthesis (e.g. aging, lighting, emotions) by applying the directions to a specific set of layers.

Tewari et al. [56] recently proposed StyleRig, a novel approach to control a pretrained StyleGAN network via a 3DMM model. A rigger network (RigNet) is responsible to learn the mapping between the semantic parameters of the 3DMM representation and the extended latent space of the StyleGAN generator. The network is trained in a self-supervised manner based on the consistency of the predictions in the image space using a differentiable renderer.

In a similar direction, Deng et al. [57] employ Variational Autoencoders (VAEs) to translate the coefficients of the input latent space of a StyleGAN model to the parameters of a 3DMM representation. A VAE is defined per each independent semantic factor: identity, pose, expression, illumination and random noise. Factors of variation are well disentangled through a imitative-contrastive learning scheme based on the priors from the rendered 3D faces.

Previous publications on GANs were focused mostly on improving the quality of the image synthesis, while recent ones are focusing more on semantic control and disentanglement of the latent space. StyleRig is a promising work in this direction, but the control of real images has not well investigated yet.

3. Method

In this chapter, we present a general framework for automatic enhancement of learning-based face relighting methods. Our goal is removing potential artifacts of synthetic datasets using the StyleGAN model as a photorealistic regularizer. First, we provide an overview of our framework in Section 3.1 and then we explain in more detail the underlying components. We start with the physically based relighting algorithm (Section 3.2) responsible of generating the portrait relighting dataset. Next, we describe the relighting network architecture (Section 3.3) and the training strategy (Section 3.4). Finally, we introduce the StyleGAN refinement algorithm (Section 3.5) and conclude with the implementations details of the solution (Section 3.6).

3.1. Overview

Figure 3.1 illustrates our general framework to boost learning-based relighting methods using the StyleGAN space as a portrait image prior. First, we produce random portraits using a pretrained StyleGAN generator and store both the images and their corresponding latent codes into an initial portrait dataset. Next, we build the portrait relighting dataset using a relighting system given the portraits and a lighting prior database. For that we sample a pair (input face, target lighting) from the corresponding datasets and then we use the relight system to obtain the tuple (input face, source light, relit face, target light). Since the relit samples could contain artifacts or incorrect lighting, we project them back into the StyleGAN space (photorealistic portrait manifold) to obtain the pair (projected relit image, relit latent code) using a stochastic gradient descent algorithm to minimize the photometric error. We initialize the optimization with the original portrait code in order to reduce significantly the number of iterations. The projection itself acts as a regularization, but we could further refine the results by the truncation of the relit latent codes. Once we remove the potential artifacts of the samples, we train the relight network on the new version of the portrait relighting dataset to predict the light source and relit the input image. Finally, we could obtain new relight samples using the relighting system or directly the relight network and then repeat the process as many times as necessary.

In particular, we test the approach on the state of the art single image portrait relighting algorithm (DPR). We use their ratio-image based relighting (physically-based method) as the initial relighting system (Section 3.2) and their autoencoder network (learning-based method) as the relighting network (Section 3.3). Finally, we adapt the original StyleGAN projection algorithm to the refinement goal (Section 3.5).

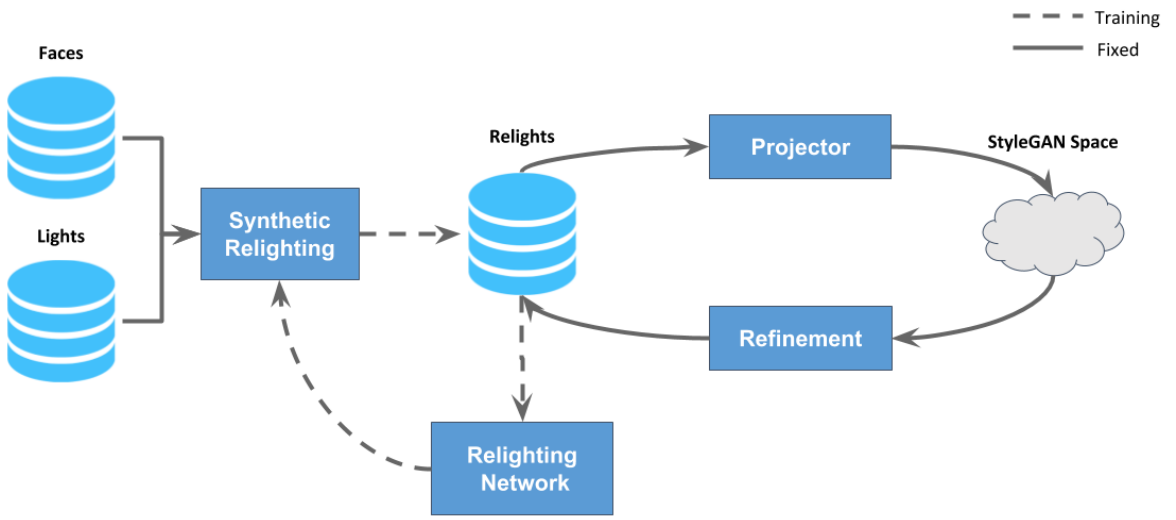

Figure 3.1.: Our general StyleGAN refinement method for face relighting. A synthetic relighting portrait dataset containing incorrect lights and artifacts is automatically enhanced by projection to the StyleGAN space, a real human face manifold. Once the refined samples are processed, the relighting network is trained on the update dataset.

3.2. Synthetic Relighting

We build the portrait relighting dataset using a ratio image-based rendering algorithm under five randomly selected lighting conditions from a lighting prior dataset [58]. The dataset is built on top of a synthetic portrait dataset generated set up by the StyleGAN generator, which contains 5000 face images and its corresponding latent codes. In total, the dataset consists of 25,000 relit pairs.

The ratio image-based relighting algorithm renders an image under a target lighting condition by multiplying the source image with the ratio of the target shading and source shading. An overview of this algorithm can be found in figure 3.2. The image formation under Lambertian reflectance can be represented by the following equation:

$$I = R \odot f(N, L) \tag{3.1}$$

where the face image I is the result of the element-wise product between the reflectance and Lambertian shading function f, which depends on the normal N and lighting L. Then, the ratio image trick [44] is applied by introducing information from the input image into the

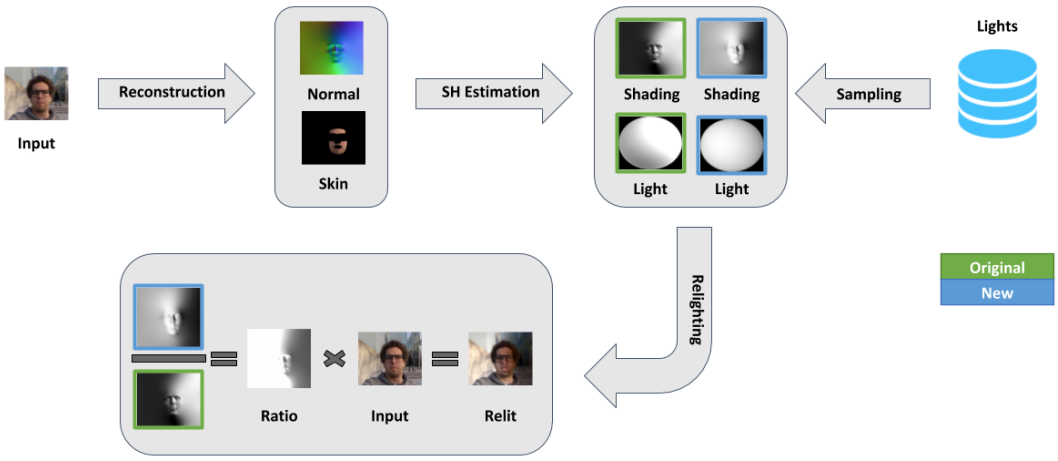

Figure 3.2.: Ratio image-based relighting algorithm. The relit image is computed by multiplying the source image with the ratio of the target shading (blue) and source shading (green). The source lighting must be previously inferred, while the target one is randomly sampled from the lighting prior dataset.

target image as follows:

$$\begin{aligned} I^* &= R \odot f(N, L^*) \\ &= R \odot f(N, L^*)(1) \\ &= R \odot f(N, L^*) \frac{R \odot f(N, L)}{R \odot f(N, L)} \\ &= \frac{R \odot f(N, L^*)}{R \odot f(N, L)} R \odot f(N, L) \\ I^* &= \frac{f(N, L^*)}{f(N, L)} I \end{aligned} \tag{3.2}$$

The shading function f using spherical harmonics lighting (SH lighting) can be defined as the product between spherical harmonics coefficients [59]:

$$S = f(N, L) = YL \tag{3.3}$$

where $L \in \mathbb{R}^{9\times1}$ is the spherical harmonics representation of the light and $Y \in \mathbb{R}^{N\times9}$ is the basis matrix, whose rows consist of the spherical harmonics basis derived by the normal of the corresponding pixel. Please refer to Section A.1 for more details about the spherical harmonics basis. The authors also assume the lighting is monochromatic and the relighting is only applied to the lighting channel of the LAB color space, .i.e, the final image is the combination of the original AB channels and the relit L channel. In this way, the algorithm does not have to deal with the ambiguity between the lighting and skin colors.

The face relighting can be computed given the portrait image I, its normal N, its lighting L and a specific target lighting L^*. The main steps of the reconstruction pipeline are illustrated in figure 3.3. We replicate the original pipeline with small modifications. The normal of the given portrait image is estimated by fitting a 3DMM model with additional refinements steps. In practice, we use the PyTorch implementation of the 3D Dense Face Alignment (3DDFA) work [60, 61] to recover the depth map from the portrait image. First, the 3DMM parameters (shape, pose and expressions) are regressed from the image via pretrained Cascaded CNNs. After that, the 3D vertices are inferred from the 3DMM parameters and the triangle mesh is built using a Delaunay triangulation. Finally, the albedo, normal, uv map and semantic labels are obtained by rendering the 3D model.

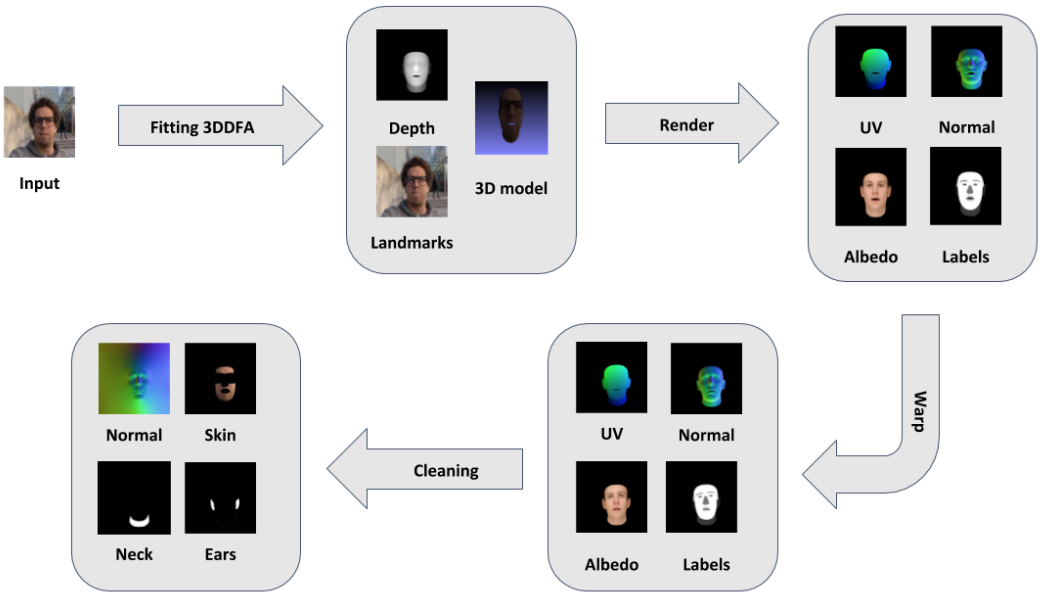

Figure 3.3.: 3D face reconstruction pipeline. First, the normal is estimated by 3DDFA [60, 61] and it is then aligned to the portrait image using an ARAP-based warping method. Finally, problematic regions of the 3DMM model are removed and then the full normal image is obtained by solving a Poisson equation.

In most cases, the resulting normal is unfortunately not accurately aligned with the portrait image due to the limitations of the 3DMM models regarding the variations of the face geometry and expressions. To address this issue, the authors propose an alignment algorithm based on the As-Rigid-As-Possible formulation (ARAP) [62]. The algorithm builds a triangle mesh for both the portrait image and the generic reflectance map of the 3DMM model using a Delaunay triangulation and 68 detected facial landmarks (Dlib) [63] as anchor points. A warp function is then estimated by imposing the ARAP deformation constraint, which is later applied to the 3DDFA normals to obtain the refined normals.

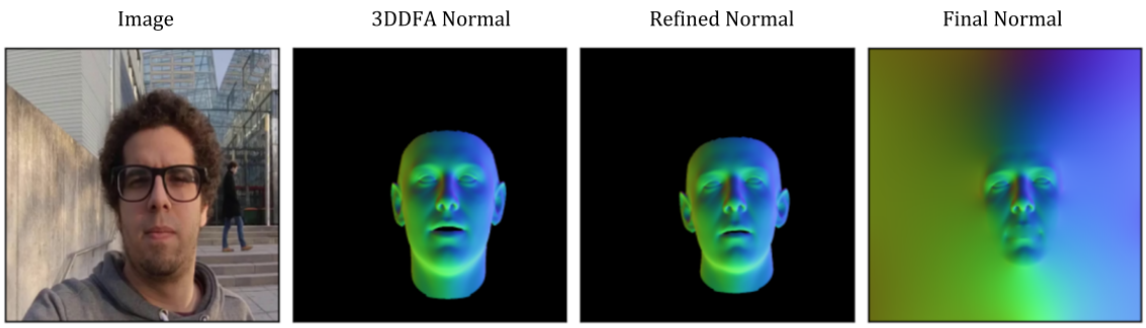

Figure 3.4.: Outcomes of the normal estimation through the 3D face reconstruction pipeline..

The new alignment is significantly better than the original one, especially at the eyes and mouth. Nevertheless, there is still a misalignment in some regions such as the ears and neck. For this reason, the authors remove these regions and fill the missing information (neck, ears and background) by solving a Poisson equation [17]. This equation arises from minimizing the difference between the gradient of the indicator function u (scalar function) and the face normals f (vector field) in the least square sense:

$$\begin{aligned} \min_u \|\nabla u - f\|_2^2 \\ \nabla(\nabla u) - \nabla f = 0 \\ \Delta u = \nabla^2 u = \nabla f \end{aligned} \tag{3.4}$$

where Δ denotes the Laplace operator and ∇f is the divergence of the given normals. In the implementation, the 2D Laplace operator is discretized via finite-difference and it is then represented as a compressed sparse matrix in order to save computation. Then, the large sparse linear system is solved using the conjugate gradients iteration of the scipy package under Dirichlet boundary conditions. Figure 3.4 shows the results of the face reconstruction and its further refinements.

Once the normal has been computed, the next step is estimating the lighting of the original portrait. There are several ways to estimate the spherical harmonics from faces. One way is using SfSNet [3], which predicts albedo, normal and lighting from a RGB image. Another way is solving a semi-definite programming problem, where the main challenge is enforcing the lighting to be non-negative. This task can be formulated as the following optimization:

$$\begin{aligned} \min_L \|I - R \odot YL\|_2^2 \\ \text{s.t. } CL \geq 0 \end{aligned} \tag{3.5}$$

where the Toeplix matrix $T(L) = CL$ corresponds to the linear combination of the precomputed Gaunt coefficient matrices C and the SH lighting L. R is the diffuse reflectance of the

face model, which is approximated as the average pixel intensities associated with the skin region. Shirdhonkar and Jacobs [64] prove that the Toeplitz matrix of a non-negative SH is positive semi-definite. In practice, the whole convex optimization problem is computed only in the face skin region using a binary mask and then it is solved using the CVXPY python package. The original pipeline derives the binary mask from the face labels inferred with the official MATLAB implementation of the multi-objective CNN proposed by Liu et al. [65], while we rely on a PyTorch implementation of a bilateral segmentation network trained on the CelebAMask-HQ dataset for the face segmentation task [66, 67, 68]. This network is originally designed to achieve the right balance between the speed and segmentation performance by including a spatial path and context path, which deal with the loss of spatial information and the shrinkage of the receptive field respectively.

Finally, a new relit image is created using the equation 3.2 given the final normal and a target SH lighting randomly sampled from a lighting prior dataset [58]. The target SH lighting is rotated using a random azimuthal angle and normalized by a random factor within range $[0.2, 0.7]$ in order to encourage more variety in the lighting conditions. Some results are shown in figure 3.5.

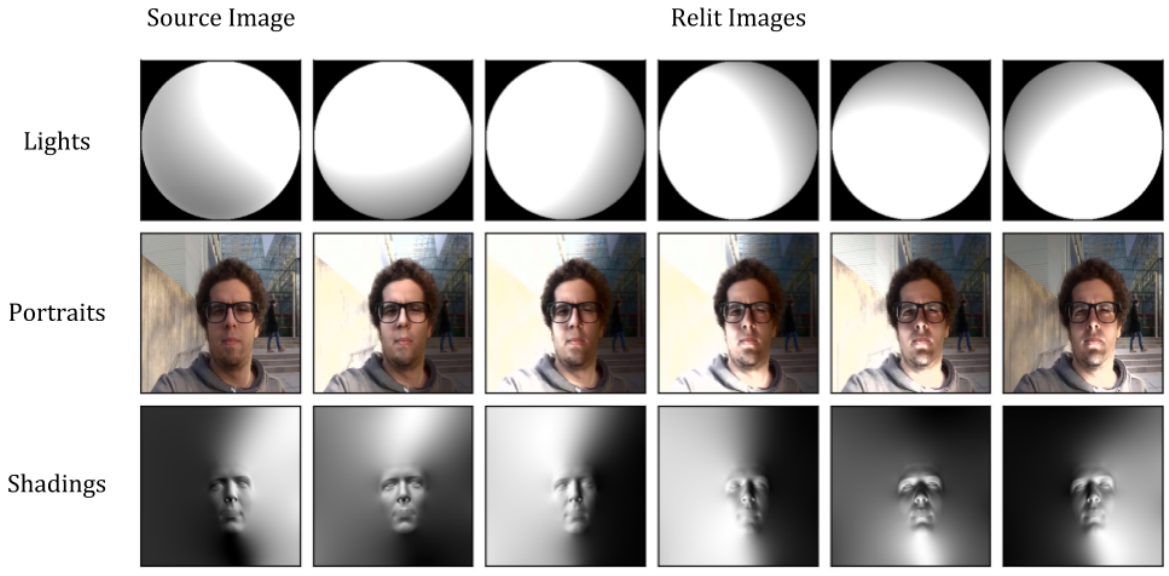

Figure 3.5.: Examples of the synthetic relighting algorithm on a given portrait image under several SH lighting conditions, including their corresponding shading maps.

3.3. Relighting Network Architecture

Once the relighting portrait dataset is built, we train the DPR hourglass network [5] to relit a single portrait image under the SH target illumination. Figure 3.6 shows the structure of the network. The encoder consists of downsample layers H_1, H_2, H_3 and H_4, while the decoder consists of upsample layers H_5, H_6, H_7, H_8. Each convolutional block in the encoder is connected to the corresponding decoder features by skip connections defined as residual blocks [69]. One convolutional block consists of one convolutional layer followed by a batch normalization layer and a ReLU activation function. We do not use the upsample and downsamples layers C_1 and C_2 because we decide to work on 256×256 and 512×512 images. Table 3.1 includes the details of these blocks, while table 3.2 presents the specifications of the lighting network.

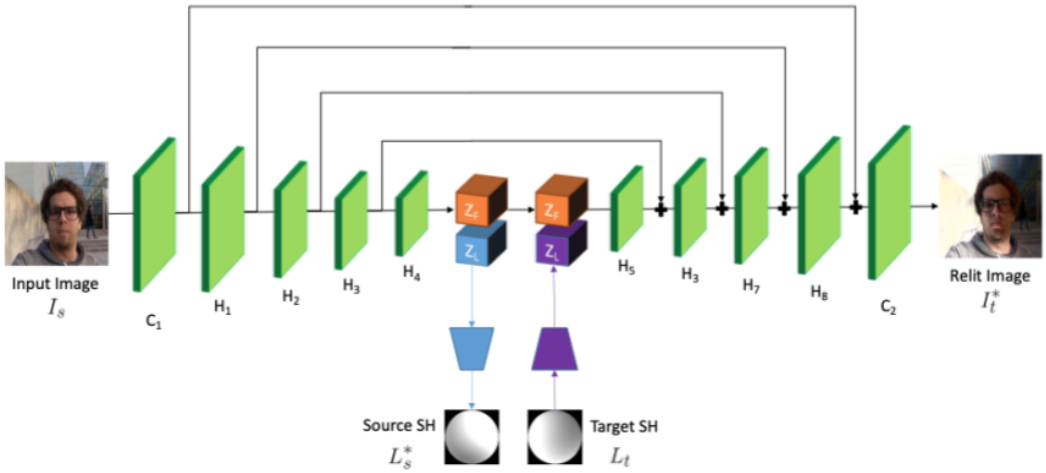

Figure 3.6.: Relighting network architecture. An encoder-decoder neural network that takes as input a single portrait image and a target illumination (injected in the bottleneck layer), and produces as output a relit version of the portrait image. Image adapted from [5].

The feature space Z is divided into two parts: a face feature Z_F (128 channels) and lighting feature Z_L (27 channels). Z_L is fed into a lighting prediction network (blue component), which predicts the spherical harmonics lighting of the source image. This network is defined as an average pooling layer followed by two fully connected layers of 128 and 9 channels respectively. The other network (purple component) maps the target spherical harmonics lighting into the lighting feature space by using two fully connected layers of 128 and 27 channels respectively. The resulting lighting feature (27 dimensional vector) is then repeated spatially to match the same spatial resolution as the face feature. In practice, we turn the fully connected layers into 1×1 convolutional layers in order to have a fully convolutional network.

Encoder	Decoder
H_1 $(3 \times 3, 16)$ Convolution	H_5 $(3 \times 3, 64)$ Upsampling
H_2 $(3 \times 3, 32)$ Convolution	H_6 $(3 \times 3, 32)$ Upsampling
H_3 $(3 \times 3, 64)$ Convolution	H_7 $(3 \times 3, 16)$ Upsampling
H_4 $(3 \times 3, 155)$ Convolution	H_8 $(3 \times 3, 16)$ Upsampling

Table 3.1.: Details of the U-NET network. H_1 $(3 \times 3, 16)$ Convolution denotes a convolutional layer with filter size 3×3 and 16 output channels. H_5 $(3 \times 3, 64)$ Upsampling denotes an upsampling layer (nearest neighbor) followed by a convolutional layer with filter size 3×3 and 64 output channels.

Regression	Encoding
Average pooling	Spatial repetition
128 Fully-connected	27 Fully-connected
9 Fully-connected	128 Fully-connected

Table 3.2.: Specifications of the lighting network. 128 Fully-connected denotes a dense layer with 128 channels.

3.4. Training Strategy and Loss Function

We first describe the original strategy for the training supervision of the relighting network and then we suggest some modifications. In the next chapter, we explore in detail the impact of both the original strategy and the proposed strategy.

The model is trained through the minimization of a weighted combination of four loss functions as follow:

$$\mathcal{L} = \mathcal{L}_I(I_t, I_t^*) + \mathcal{L}_L(L_s, L_s^*) + \lambda \mathcal{L}_F + \mathcal{L}_{GAN} \tag{3.6}$$

The photometric supervision is applied using a L1 loss between the ground truth relit image I_t and the predicted image I_t^*. The authors also add a L1 loss to the gradients to preserve edges and avoid blurring results:

$$\mathcal{L}_I = \frac{1}{N_I}(\|I_t - I_t^*\|_2 + \|\nabla I_t - \nabla I_t^*\|_2) \tag{3.7}$$

The lighting loss supervises the correct estimation of scene the illumination via L2 loss between the ground truth source lighting L_s and the predicted lighting L_s^*:

$$\mathcal{L}_L = (L_s - L_s^*)^2 \tag{3.8}$$

A feature matching loss enforces images of the same person under different lighting conditions to have the same face features using a L2 loss:

$$\mathcal{L}_F = \frac{1}{N_F}(Z_{f_{ori}} - Z_{f_i})^2 \tag{3.9}$$

where $Z_{f_{ori}}$ are the face features from the original image, while Z_{f_i} are the face features of one of the relit versions.

The relit images generated by the synthetic relighting algorithm are not real ground-truth images and in fact they look unrealistic especially under extreme lighting cases. Furthermore, they may contain artifacts due to inaccurate estimations during the data generation process (e.g. face normals or lighting predictions). Therefore, the authors use a patch GAN to constraint locally the relit portraits to be as similar as possible to the real images. The adversarial loss is formulated as a LS-GAN loss:

$$\mathcal{L}_{GAN} = \mathbb{E}_I(1 - D(I))^2 + \mathbb{E}_{I_s} D(G(I_s, I_t))^2 \tag{3.10}$$

where G and D stand for the generator and discriminator networks respectively. The generator network is the relighting network mentioned above. I is the real image and it is labeled as 1, while $G(I_s, I_t)$ produces the fake image labeled as 0.

The relighting network is trained end-to-end following a special training procedure denoted by the authors as skip training. First, the network is trained for five epochs without any skip connections. After that each skip connection is added progressively one per epoch in a similar fashion to this publication [51]. Next, the face feature loss is included after ten epochs. The main goal of this strategy is avoiding the facial information to be passed through the skip layers, which leads to noticeable artifacts especially around the nose [5].

We suggest to employ the perceptual loss LPIPS $\mathcal{L}_{LPIPS}$ introduced by Zhang et al. [70] instead of the original image loss $\mathcal{L}_I$. In this loss function, both the input and target images are fed into a pre-trained VGG16 network [71] and then the final error is measured by $L1$ loss at different layers of the network. Additionally, we prefer our refinement strategy instead of the adversarial training. Finally, we also propose to include a reconstruction loss $\mathcal{L}_I(I_s, I_s^*)$ in a similar way to Sun et al. cost function [6]. We replace the true target illumination L_t with the predicted one L_s^*. As a result, we can measure the reconstruction error of the model in a self-supervised way. To avoid generalization issues, the predicted source illumination is slightly jittered before is feeding back into the decoder.

3.5. StyleGAN Refinement

The StyleGAN model is able to synthesize realistic portrait images of faces, including the scene illumination. On the other hand, the performance of the relighting network depends

heavily on the quality of the training set. Therefore, we use the StyleGAN generator network to improve the quality of synthetic relighting datasets such as the DPR one. We first introduce the StyleGAN architecture with an emphasis on the projection operation. Then, we present our refinement approach, which is inspired in the self-supervised bootstrapping strategy of the InverseFaceNet [72].

StyleGAN is an extension of the progressive growing GAN network [51], where learned styles are applied per scale in order to provide more control over the image generation process. An overview of the architecture is shown in figure 3.7. The generator is divided into two components: the mapping network and the synthesis network. The synthesis network is a modified version of the original generator. The mapping network is comprised of eight fully connected layers, which map a random sampled point Z from the input latent space to an intermediate latent space W. Then, the resulting style vector is applied to each block of the synthesis network via a learned affine transformation and an Adaptive Instance Normalization (AdaIN) layer [33]. This normalization is an effective way to enforce the statistics of the activation maps to match the statistics of the given style. In practice, the operation consists of first transforming the output of the feature maps to a standard Gaussian and then the target distribution is obtained by scaling and translating the results using the style scale and bias respectively. Finally, the authors add a Gaussian noise image to each scale using learnable scaling factors, which allow the synthesis network to control the amount of noise to be injected into the corresponding feature map. The goal is providing the network with the capability to represent details such as skin pores or curly hair.

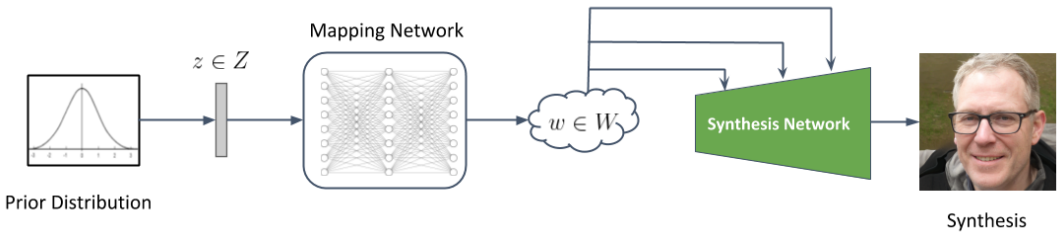

Figure 3.7.: Architecture of the StyleGAN generator [10]. A random vector $z \in Z$, which is sampled from a multivariate normal distribution, is first transformed into an intermediate latent space $w \in W$ via the mapping network. Then a realistic image is created by the synthesis network under the styles encoded in the intermediate latent vector w.

The proposed method relies on two key operations: image projection and truncation. Projecting a target image x into the StyleGAN space consists of finding the corresponding $w \in W$ and the per-layer noise maps n_i. In essence, we need to invert the synthesis network,

which can be formulated as an optimization task [10] with the following cost function:

$$\mathcal{L} = \mathcal{L}_{image}(x, x^*) + \alpha \mathcal{L}_{reg}(n_i) \tag{3.11}$$

where $x^* = g(w^*, n_i)$ is the image produced by the synthesis network g. This optimization is usually solved directly by running stochastic gradient descent methods for approximately 1000 iterations. The first component $\mathcal{L}_{image}$ is the data term and it is computed using the LPIPS distance [70]. Both images are downsampled to 256×256 resolution to reduce the computation effort of the LPIPS distance. The second component is a regularization term of the noise maps and it is performed on multiple resolutions. This expression constraints the optimization to not introduce image content into the noise maps. In practice, a style vector w is optimized for every layer of the synthesis network. The concatenation of the 18 different style vectors is known as the extended latent space W^+ and previous research [52, 53] suggests that leads to higher quality results.

Low probability density regions in z or w do not have enough training data to learn an accurate representation. As a result, the synthesis network may not be able to generate high quality images from them. One way to address this complication is using the truncation trick. In the original StyleGAN publication [52] is performed as follows:

$$w' = \overline{w} + \psi(w - \overline{w}) \tag{3.12}$$

where $\overline{w}$ is the mean of the W space and w is the original latent variable. The deviation scale parameter ψ controls the linear interpolation between the average face ($\psi = 0$) and the original face ($\psi = 1$). This trick can indeed reduce the artifacts of the synthetic images at the cost of the variation.

Figure 3.8 shows an overview of our refinement strategy. First, we sample 5000 latent codes z from a standard normal distribution and generate the corresponding photorealistic face images and their corresponding latent codes w using a pretrained StyleGAN network $I_w = StyleGAN(z)$. Then, we relit the faces using the ratio image based relighting algorithm described in Section 3.2. After that, we project back the relit portraits into the extended StyleGAN space W^+ in order to refine the quality of the images. Once we estimate the refined images, we use them to train the relighting network. It is important to point out that the network could be improved by applying this strategy many times.

The projection of a relit image into the extended latent space W^+ is slightly different from the original StyleGAN formulation. First, we initialize the candidate latent variable with the original intermediate latent variable $w^* = w_{img}$. In this way, we would require less iterations to converge to a good solution. Second, we include an extra regularization term to prevent the candidate latent variable to differ too much from the original one. The intention is finding a latent variable which captures the light signal from the relit image preserving the quality of

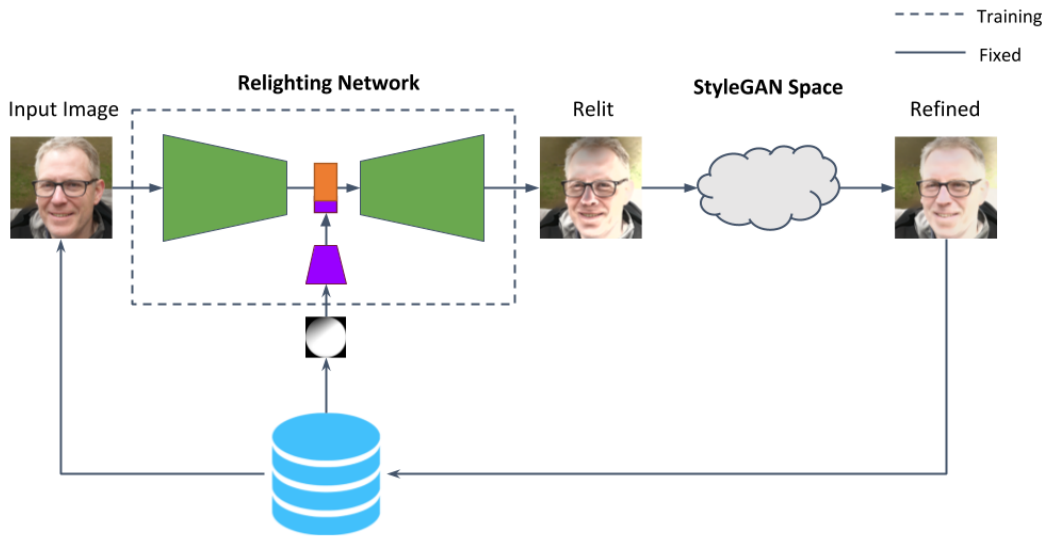

Figure 3.8.: StyleGAN refinement strategy. The relighting network takes an input image and a target ilumination and produces a relit image. After that a refined version is obtained by projecting the relit sample into the extended latent space W_+. The original sample is replaced with the refined one and then the network is retrain.

the original portrait. The final cost function is then designed as follows:

$$\mathcal{L} = LPIPS(x_{relit}, g(w^*, n_i)) + \alpha(w_{img} - w^*)^2 \tag{3.13}$$

where the noise maps n_i are fixed, .i.e, they are not trainable parameters. For our method, we decide to ignore the explicit optimization of the noise maps in order to simplify the optimization process. In addition, we truncate the resulting latent variable w^* using equation 3.12. Finally, we also consider the optimization of specific layers of the extended latent space associated with the lighting of the scene (e.g. 9, 9-18).

In next chapter, we study the accuracy of the design choices suggested above: the initialization, the target layers and the noise optimization. Furthermore, we evaluate the impact of the different parameters: the truncation parameter ψ, the number of steps and the regularization factor α.

3.6. Implementation

We train the relighting network using the Adam optimizer with default parameters [73] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$). The network was implemented in python 3.7.7 using the deep learning framework PyTorch 1.3.1 [74]. The experiments were performed on a NVIDIA GeForce GTX 1080 Ti device, which enables a maximum batch size of 16 and 8 for images

of size 256×256 and 512×512 respectively. Training the relighting network on the portrait relighting dataset of 25,000 relit images takes approximately three hours. For the inference time, the network requires around 223 milliseconds per image ignoring the time for loading the input image and the SH light vector.

For the StyleGAN refinement, we use a PyTorch implementation of the original StyleGAN 2 paper [10, 75]. We converted the original weights from the official TenforFlow implementation to PyTorch weights using the *convert_weight.py* script. In the case of the StyleGAN projection, we use the default parameters of the *projector.py* script, but we do not optimize the noise maps. For the rest of the setting, we use the same setup as in the relighting network. The projection of five images using 200 iterations takes around 40 seconds. The refinement of the complete portrait relighting dataset takes approximately 69 hours on the GeForce GTX 1080 Ti device.

4. Experiments

In this chapter, we evaluate our refinement approach both quantitatively and qualitatively. First, we describe the datasets and evaluation metrics used in the experiments. Second, we compare the method with previous state-of-the-art algorithms. After that we present the results of the ablation studies corresponding to the two main components of the proposed refinement technique: the relighting network and the StyleGAN refinement strategy. Finally, we show some results on direct StyleGAN relighting using a small light encoder.

4.1. Datasets

In this work, we rely on several datasets for different purposes. An overview is presented in table 4.1. We use ours synthetic dataset to train the relighting network and to perform the qualitative evaluation of the refinement strategy. We propose the Multi-PIE dataset [76] for the quantitative evaluation of the experiments. We consider the Laval face and lighting dataset [40] useful for the comparison of the proposed method against the baselines. Finally, we employ the Deep Portrait Relighting dataset [5] as the training dataset for the ablation studies of the relighting network.

Dataset	Size	Identities	Resolution	Domain
Ours	$25,000$	$5,000$	256×256	synthetic
Laval [40]	137	9	256×256	real (unpaired)
Multi-PIE [76]	$4,980$	249	128×128	real
DPR [5]	138,135	$27,627$	1024×1024	synthetic

Table 4.1.: Summary of the datasets used in this work. Size and Identities refer to the number of relit pairs and the number of subjects on the corresponding dataset respectively.

The Laval dataset consists of 137 face/lighting pairs from 9 subjects (8 males and 1 female) under 25 different lighting conditions. Subjects mostly follow a neutral expression protocol. The acquisition was performed in two steps: the illumination estimation and the portrait session. First, a HDR spherical environment map was computed by merging a sequence of images at different orientations using a Canon 5D Mark III camera mounted on a robotic tripod. Once the environment was captured, the tripod was removed and each subject was recorded at the same location. We show some example pairs in figure 4.1. We can predict the

SH lighting from a given face image in this dataset using the relighting network and then relit another arbitrary portrait image under this reference lighting.

Figure 4.1.: Examples of the Laval face and lighting dataset [40]. Each pair includes the face probe and the corresponding lighting conditions represented as a HDR spherical environment map. Image adapted from [40].

Deep Portrait Relighting (DPR) dataset is a large scale dataset built on top of the high resolution CelebA-HQ dataset published in the ProgressiveGAN work [51]. For the final version, authors discard images in which the landmarks extraction was not possible. The ratio image-based rendering algorithm is then applied to each of the resulting images under five randomly selected lighting conditions to obtain the face relit images (Section 3.2). In total, the dataset consists of 138,135 relit images. For our synthetic dataset, we first generate 5000 identities using the StyleGAN generator. Next, we use the synthetic relighting algorithm in a similar way as the DPR case to build a portrait relighting dataset of 25,000 images.

We cannot evaluate the accuracy of the relighting task using any of the datasets described so far. On the one hand, synthetic relit images should not really be considered ground-truth. On the other hand, the Laval dataset is actually designed for the lighting estimation task. For a given subject, the image content is not exactly the same under different illumination conditions. For example, there are variations of the locations and face postures. For this reason, we consider the Multi-PIE dataset [76] for quantitative evaluation because it contain real images of the same person under different lighting conditions.

The Multi-PIE dataset covers more than 750,000 images of 337 people under 15 view points and 20 lighting conditions (including two no flash images). Subjects were also instructed to display different facial expressions such as neutral, smile, disgust, surprise or squint. The capture system consists on 15 cameras and 18 flashes connected to a group of computers in order to record the different variations in a systematic way. The calibration setup and samples

of one session are illustrated in figure 4.2. For our experiments, we consider a subset of this dataset of cropped images around the face from 249 subjects under the frontal view and the neutral expression. For each person, we build the test pairs using the average of the relit images as input and each of these samples as an output. In total, the test set contains 4980 images.

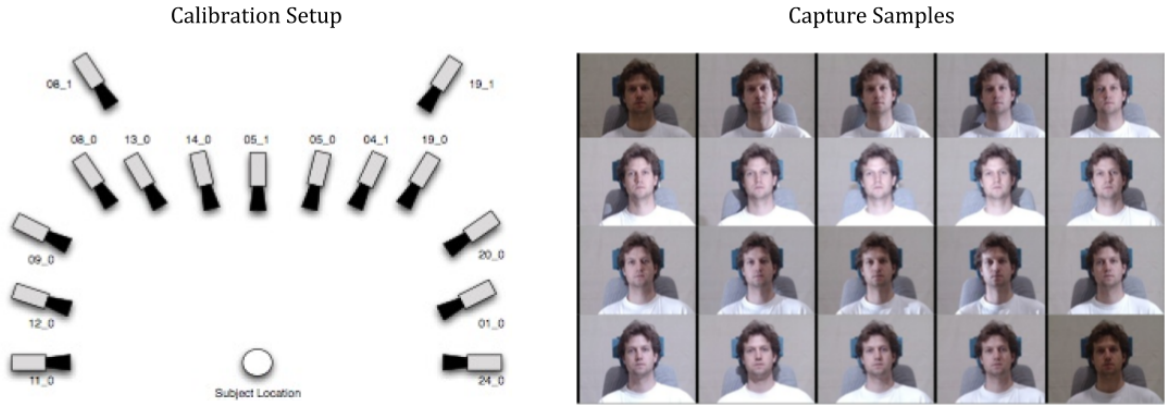

Figure 4.2.: Multi-PIE dataset [76] under different lighting conditions. The arrangement of the cameras and their corresponding 20 relit images for a particular subject in frontal view. Each sampled was obtained by firing only the flash of the corresponding camera. The first and the last images are without flash. Image adapted from [76].

4.2. Metrics

To evaluate the performance of the relighting algorithms we need to measure the similarity between the original image and the relit image. We consider the following three metrics: mean-square error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM). These metrics cover different degrees of invariance from none (MSE) to local and global scaling invariance (SSIM). A summary is presented in table 4.2.

MSE is one of the most popular error estimators in statistics. It is computed by the average squared difference between the predicted values and the actual values. The PSNR ratio is typically used as a quality assessment between an image and its compressed version. The higher the PSNR, the better the quality of the reconstruction. The PSNR is computed using the following equation:

$$PSNR(x,y) = 10\log_{10}\frac{max_{img}}{MSE(x,y)} \tag{4.1}$$

where max_{img} is the maximum possible intensity of the image. For example, 255 for 8 bits images. The estimation is expressed in terms of the logarithmic decibel scale in order

Metric	Invariance	Human judgement	Neural network
MSE	none	no	no
PSNR	global	no	no
SSIM [77]	local and global	yes	no
FID [78]	local and global	yes	yes

Table 4.2.: Overview of the metrics for quantitative evaluation. In this table, we compare the different properties of the metrics such as invariance and the correlation of the visual quality with respect to the human criteria. The first three metrics are traditional, while the last one (FID) is a deep-learning based score.

to deal with high dynamic range values. It is important to point out that MSE and PSNR are not considered robust estimators to measure similarity in terms of human perception. For instance, MSE would obtain the same value if we add a constant value or noise to the original image. The structural similarity index introduced by Wang et al. [77] overcomes these drawbacks by taking into account structural information such as luminosity (average brightness), contrast (texture) and structural similarity (correlation). The SSIM formula between the image x and y is defined as:

$$\begin{aligned} SSIM(x,y) &= l(x,y)c(x,y)s(x,y) \\ l(x,y) &= \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \\ c(x,y) &= \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \\ s(x,y) &= \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \end{aligned} \tag{4.2}$$

where $l(x,y)$, $c(x,y)$, $s(x,y)$ measure the similarity with respect to the luminosity, contrast and structural similarity respectively. The luminosity, contrast and structural similarity of the samples are captured using the average, standard deviation and covariance of the intensity values. The equation is designed to enforce each component to be between 0 and 1, where x and y have exactly the same corresponding attribute when the result is 1. c_1, c_2 and c_3 are constants introduced to avoid numerical issues, .i.e, division by 0.

The index is typically computed using a 11×11 Gaussian filter, which implies SSIM is invariant to local scaling. We compute SSIM on each RGB channel independently, which makes the metric invariant to color shifts too. The metrics discussed so far can be computed in practice using the scikit-image python package.

We also decide to consider the Fréchet Inception distance proposed by Heusel et al. [78], one of most popular metrics to evaluate the performance of GANs. The authors show that

lower FID scores correlate with human judgement of visual quality. The score is calculated measuring the Fréchet distance between two multivariate Gaussians:

$$FID = \|\mu_r - \mu g\|_2^2 + Tr(\sigma_r^2 + \sigma_g^2 - 2\sqrt{\sigma_r \sigma_g}) \tag{4.3}$$

where $X_r \sim N(\mu_r, \sigma_r)$ and $X_g \sim N(\mu_g, \sigma_g)$ correspond to the statistics of the 2048-dimensional activations of the pretrained Inception-v3 last pooling layer for real and generated samples respectively. *Tr* refers to the trace operation, i.e., the sum of the elements along the main diagonal of a square matrix.

4.3. Comparison with Baselines

We propose the state-of-the-art methods SfSNet [3] and DPR [5] as baselines since they provide the pretrained models and they are trained on synthetic data. Both methods can perform relighting in two different ways: sh-based lighting and reference-based lighting. The first approach generates directly the relit image from a source image and a target lighting encoded as a spherical harmonics vector. The second approach, first predicts the spherical harmonics lighting from a reference image and then it infers the relit image of another image using this predicted light, i.e., the model performs light transfer. This option is particularly helpful when the target lighting is not available (e.g. MultiPIE dataset, Laval dataset) or the coordinate system of the given spherical harmonics vector is different with respect to the method's coordinate system.

In figure 4.3, we present a comparison of the techniques on conventional portrait images. For this test, we transfer the illumination conditions of some Laval portrait images to our example images. We can see that all methods generate images under the correct lighting. However, the SfSNet cannot deal with the background correctly unless a binary mask is provided and it generates 128×128 images, which we consider are too small for portrait relighting applications. Figure 4.3 also reveals that DPR and our method disentangle correctly the lighting from albedo, which is clearly not the case for SfSNet. Finally, we think the relighting of the proposed method looks slightly better than DPR.

In table 4.3 we compare the performance of our method against the baselines for the single-image relighting task on the Multi-Pie test set. Since the ground truth lighting is not provided with the dataset, we use the data of the first subject as the reference for the scene illumination. We predict the spherical harmonics lighting for all the images of this identity and then we relit the rest of subjects using the predicted lights. We apply the same methodology for all methods in order to avoid unfair comparisons. The results demonstrate that the proposed method achieves the best results for SSIM and FID, which we believe are the most suitable metrics for the task.

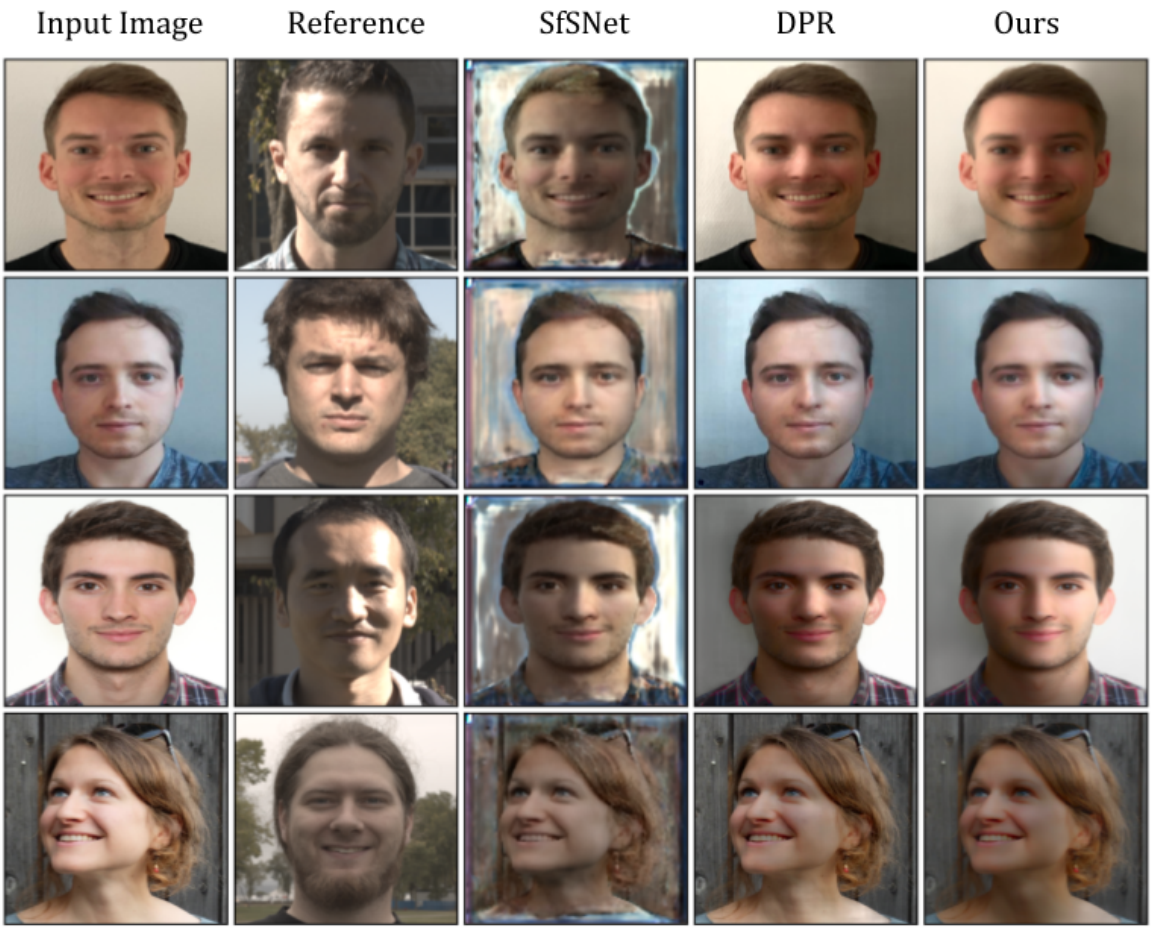

Figure 4.3.: Qualitative comparison of the proposed method with previous methods (SfSNet [3], DPR [5]) on our portrait images. We rely on some examples of Laval face and lighting HDR Dataset [40] as references for the illumination conditions. For each method, we first predict the SH lighting from the reference image and then we relit the input image under that lighting conditions.

Model	MSE ↓	PSNR ↑	SSIM ↑	FID ↓
SfSNet [3]	928.0101	18.8402	0.8449	280.1141
DPR [5]	**528.1843**	**21.4433**	0.9096	40.3443
Ours	594.4948	21.1055	**0.9183**	**28.0343**

Table 4.3.: Quantitative comparison between the proposed method and baselines on the Multi-PIE dataset. We use the captures of the first person as the reference images for the light transfer to other subjects.

In figure 4.4, we show some examples of the comparison. Each image of the Multi-PIE dataset was captured under a dominant point light source. We can notice that any of the methods fully captures the reference lighting in that sense. The main reason is the representation of the scene illumination as a spherical harmonics vector, which is more suited for diffuse lighting conditions. We attribute the blurry results of our predictions to the domain gap between the Multi-PIE and the synthetic portrait relighting datasets.

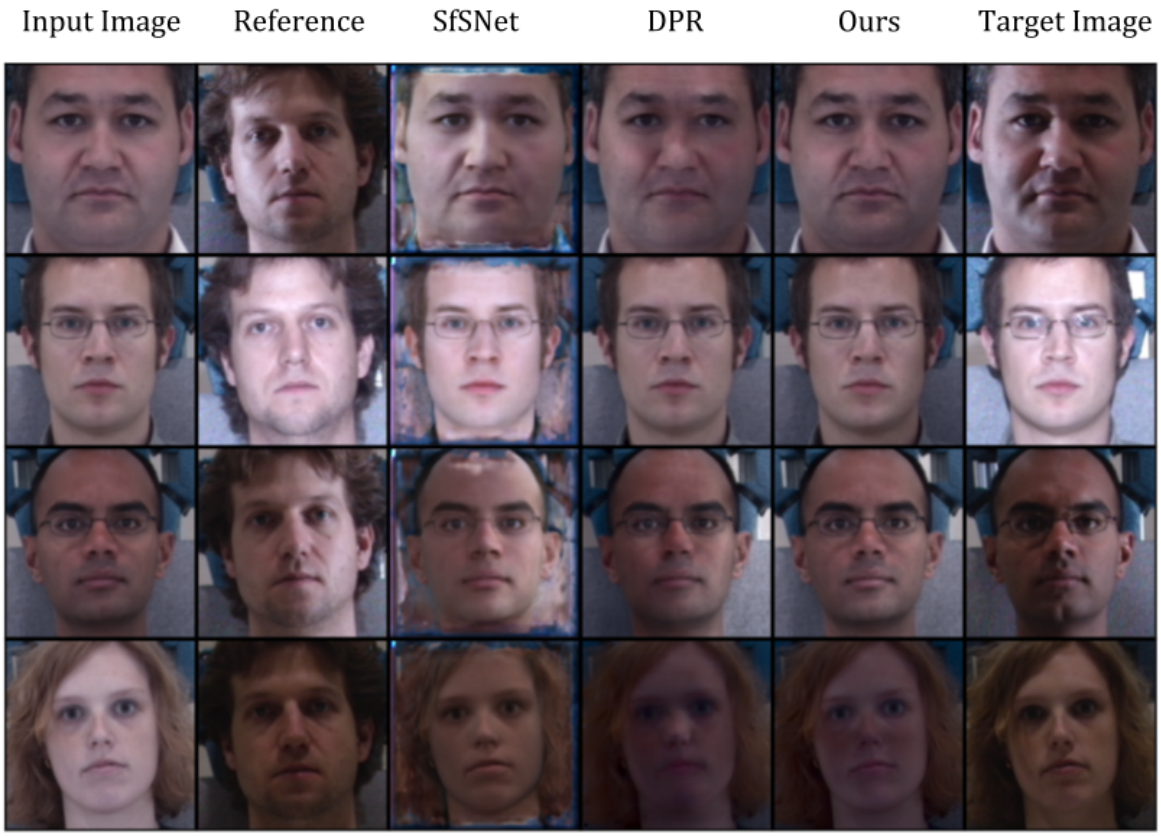

Figure 4.4.: Qualitative comparison of the proposed method with baselines (SfSNet [3], DPR [5]) on MultiPIE dataset [76]. For each method, we first extract the SH lighting from the reference image and then we relit the input image under that lighting conditions. The difference between the predictions of the methods and the ground-truth image is remarkable.

4.4. Ablation Studies for Relighting Network

In this section, we run several tests in order to evaluate the importance of different design choices on the relighting network performance. We assess the original formulation of the DPR

work [5] and then we contrast with our modifications. We train all the models on the original DPR dataset for a fair comparison. We consider the following factors: loss functions, dataset size, training strategy and color space. In the case of the loss functions, the $\mathcal{L}_L$ is essential for the relighting task since it provides the illumination supervision. Thus, the lighting loss is required in all experiments.

We train the relighting network on $25,000$ relit pairs using the loss functions described in Section 3.4. In table 4.4 and figure 4.5 we present the quantitative and qualitative results respectively. We notice that adding each loss function improves the performance with the exception of the cycle loss $\mathcal{L}_{Cyc}$, which we argue may distract the network to focus more on the reconstruction task instead of the relighting one. Our suggested loss function ($\mathcal{L}_{LPIPS} + \mathcal{L}_F$) obtains the best result for all metrics. Additionally, we can appreciate in figure 4.5 that this combination indeed improves the visual quality of the images. For instance, images are sharper and contain significantly less grid artifacts especially in comparison with the vanilla loss $\mathcal{L}_I$. It is relevant to point out that we decide to exclude the adversarial loss of the analysis because it makes the training process more unstable and we believe it would have a similar effect to the refinement step.

Loss	MSE ↓	PSNR ↑	SSIM ↑
$\mathcal{L}_I$	2779.3725	14.5333	0.8101
$\mathcal{L}_I + \mathcal{L}_{Grad}$	2306.6228	15.4522	0.8296
$\mathcal{L}_I + \mathcal{L}_{Grad} + \mathcal{L}_F$	1703.8868	16.3749	0.8395
$\mathcal{L}_I + \mathcal{L}_{Grad} + \mathcal{L}_F + \mathcal{L}_{Cyc}$	4644.6799	11.9719	0.7688
$\mathcal{L}_{LPIPS} + \mathcal{L}_F$	**1642.7077**	**16.5897**	**0.8569**

Table 4.4.: Ablation study of the loss functions on MultiPie dataset [76]. Our proposal loss $\mathcal{L}_{LPIPS} + \mathcal{L}_F$ is more compact and effective for all metrics.

We evaluate the impact of the dataset size and the training strategy using the proposed loss ($\mathcal{L}_{LPIPS}, \mathcal{L}_F$). Table 4.5 shows we can improve the performance in the relighting task by increasing the size of the training set. Nevertheless, the improvement is not so impressive versus the computational effort. In table 4.6, we compare the efficacy of the skip training against a vanilla strategy on $125,000$ training pairs. We can conclude based on the results that the skip training strategy does not improve the performance. In addition, this strategy did not remove the artifacts around the nose. We argue that the skip training strategy is tailored for the adversarial training, which coincides with previous works such as ProgressiveGAN [51]. We could ameliorate those artifacts by replacing the nearest neighbor upsampling with a bilinear interpolation.

In table 4.7 and figure 4.6, we report the outcomes of the quantitative and qualitative experiments related to the color space. The results confirm the importance of the LAB space.

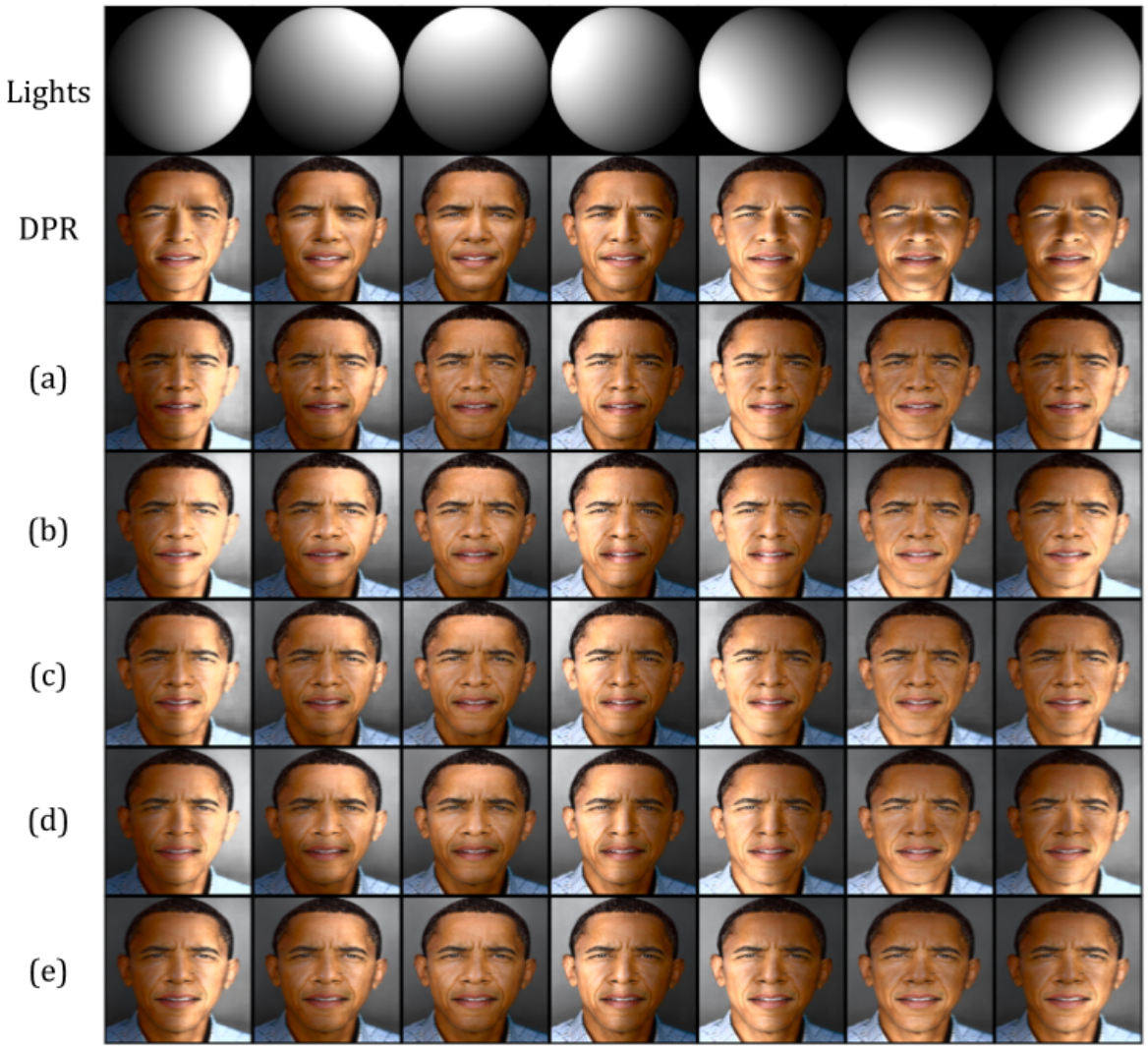

Figure 4.5.: Visual comparison of the effects of several loss functions on the results: (a) $\mathcal{L}_I$, (b) $\mathcal{L}_I + \mathcal{L}_{Grad}$, (c) $\mathcal{L}_I + \mathcal{L}_{Grad} + \mathcal{L}_F$, (d) $\mathcal{L}_I + \mathcal{L}_{Grad} + \mathcal{L}_F + \mathcal{L}_{Cyc}$, (e) $\mathcal{L}_{LPIPS} + \mathcal{L}_F$. DPR denotes the original pretrained model [5].

Dataset Size	MSE ↓	PSNR ↑	SSIM ↑
25000	1713.2614	16.5747	0.8541
75000	1551.1950	**16.9224**	**0.8614**
125000	**1528.3020**	16.8606	0.8584

Table 4.5.: Impact of the number of relit images on Multi-Pie evaluation using a vanilla training strategy.

Training Strategy	MSE ↓	PSNR ↑	SSIM ↑
Vanilla 256×256	1756.5665	16.7512	0.8580
Skip training 256×256	1578.3096	17.3136	0.8677
Vanilla 512×512	**1396.6921**	**17.5235**	**0.8788**
Skip training 512×512	1687.0097	16.9649	0.8585

Table 4.6.: Ablation study of the training strategy on Multi-Pie dataset for different image resolutions.

Optimizing only on the luminosity channel of this space not only reduces memory consumption but also leads to better results. RGB predictions look particularly unusual around the brighter parts of the relit skin.

Color Space	MSE ↓	PSNR ↑	SSIM ↑
LAB	**815.0624**	**19.8153**	**0.9067**
RGB	1547.8246	17.1339	0.8671

Table 4.7.: Quantitative evaluation study of the color space choice on Multi-Pie dataset. The relighting network clearly performs better on LAB space for all metrics.

In this Section, we found interesting insights about the relighting network. First, we demonstrate the effectiveness of the LPIPS distance in comparison to the original loss formulation. Second, we proved the skip training was not particularly superior to the vanilla training, especially regarding the stripe artifacts. We then suggested to remove these artifacts by replacing the nearest neighbor upsampling with a bilinear interpolation. Third, we realized the performance of the network does not improve significantly with the dataset size. Finally, we corroborate the relevance of the LAB space choice to deal with the ambiguity of the color between the light source and skin reflectance.

4.5. Ablation Studies for StyleGAN Refinement

In this section, we analyze the effect of the different hyperparameters involved in the refinement method through several qualitative comparisons. For these experiments, we generate five human portraits using the StyleGAN model and we relit them under different lighting conditions using the pretrained version of the DPR relighting network [5]. The hyperparameters to be studied are: the deviation scale of the truncation ψ, the regularization strenght α, the refinement layers and the number of optimization steps n. After that we compare the relighting portrait dataset with the refined version, where we emphasize the impact on relighting artifacts.

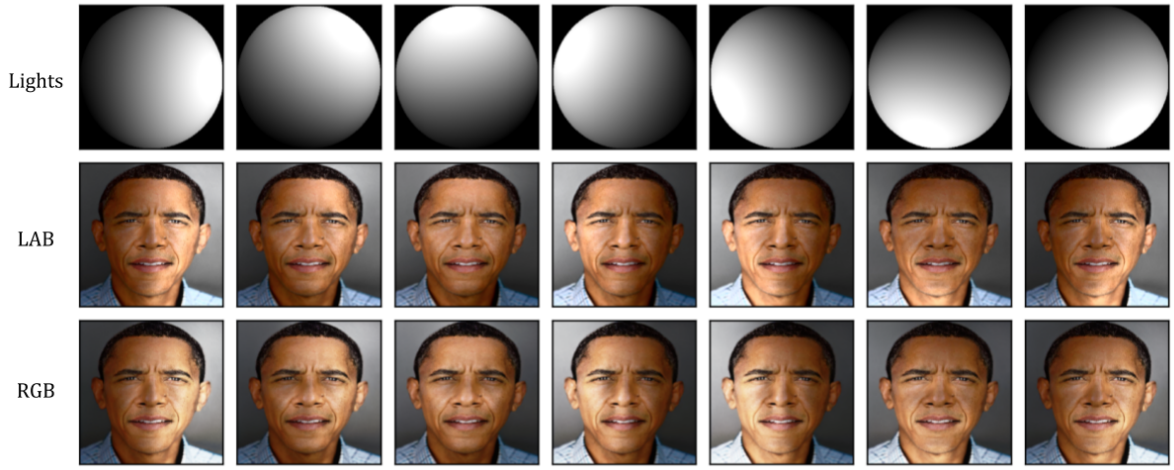

Figure 4.6.: Comparison of the relighting network predictions when it is trained on different color spaces. In particular, note the peculiar appearance of the RGB case for the lit regions of the skin.

All the light controllers discovered with the GANSpace technique [54] rely mostly on the layer 9 of the StyleGAN generator. For this reason, we decide to also study the impact of optimizing specific layers during the image projection. For this experimentation, we project the relit images to the extended StyleGAN space without truncation neither regularization. Figure 4.7 demonstrates that the optimization of specific layers like $9 - 18$ is indeed more effective than the original formulation when the optimization is run for only 50 iterations. However, the layer 9 seems to not be able to fully capture the illumination of the portraits.

The projection of arbitrary images into the StyleGAN space usually requires many iterations to converge to a good solution. Nevertheless, we believe that the latent code of the original portrait is a good guess to accelerate the projection of the relit version. For this experiment, we consider the optimization of all layers without regularization neither truncation trick. Figure 4.8 illustrates the impact of the number steps. Unfortunately, a low number of iterations like 50 does not capture well the scene illumination. On the other hand, a high number of steps like 1000, which takes 150 seconds per image (four times more), does not seem to incorporate significantly more information than the 200 case. Therefore, we think 200 iterations represents a good compromise between the quality of the results and the runtime performance.

For the truncation trick and regularization tests, we perform the projection of the relit images considering only the layers 9-18 using 200 optimizations steps. Both operations prevent the candidate latent vector to differ too much with respect to the latent code of the original image. The regularization is applied during the optimization, while the truncation is a post-projection operation. The results for the truncation case are shown in figure 4.9. As

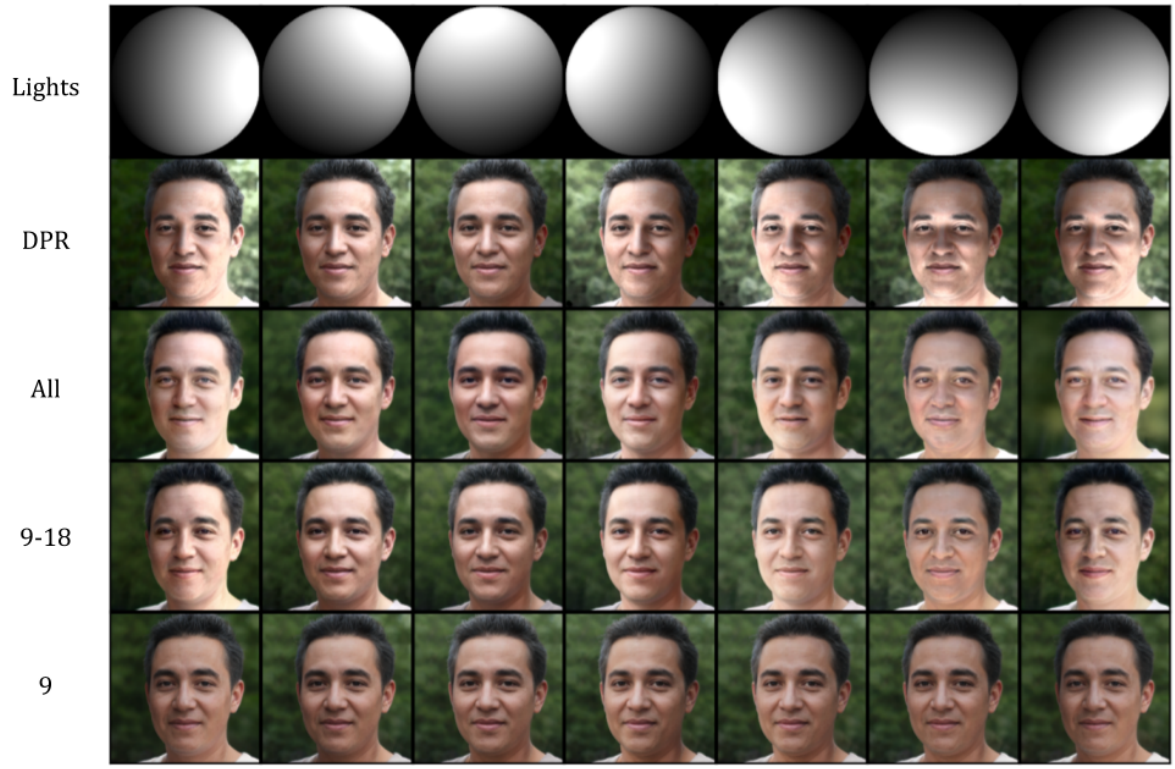

Figure 4.7.: Effect of specific layers on the projection of relit images. We perform the synthetic relighting using the pretrained DPR relighting network [5] under the given SH lightings. We run the optimization for 50 iterations without truncation neither regularization. The optimization on layers 9-18 seems to be more effective to capture the light information for the given number of steps.

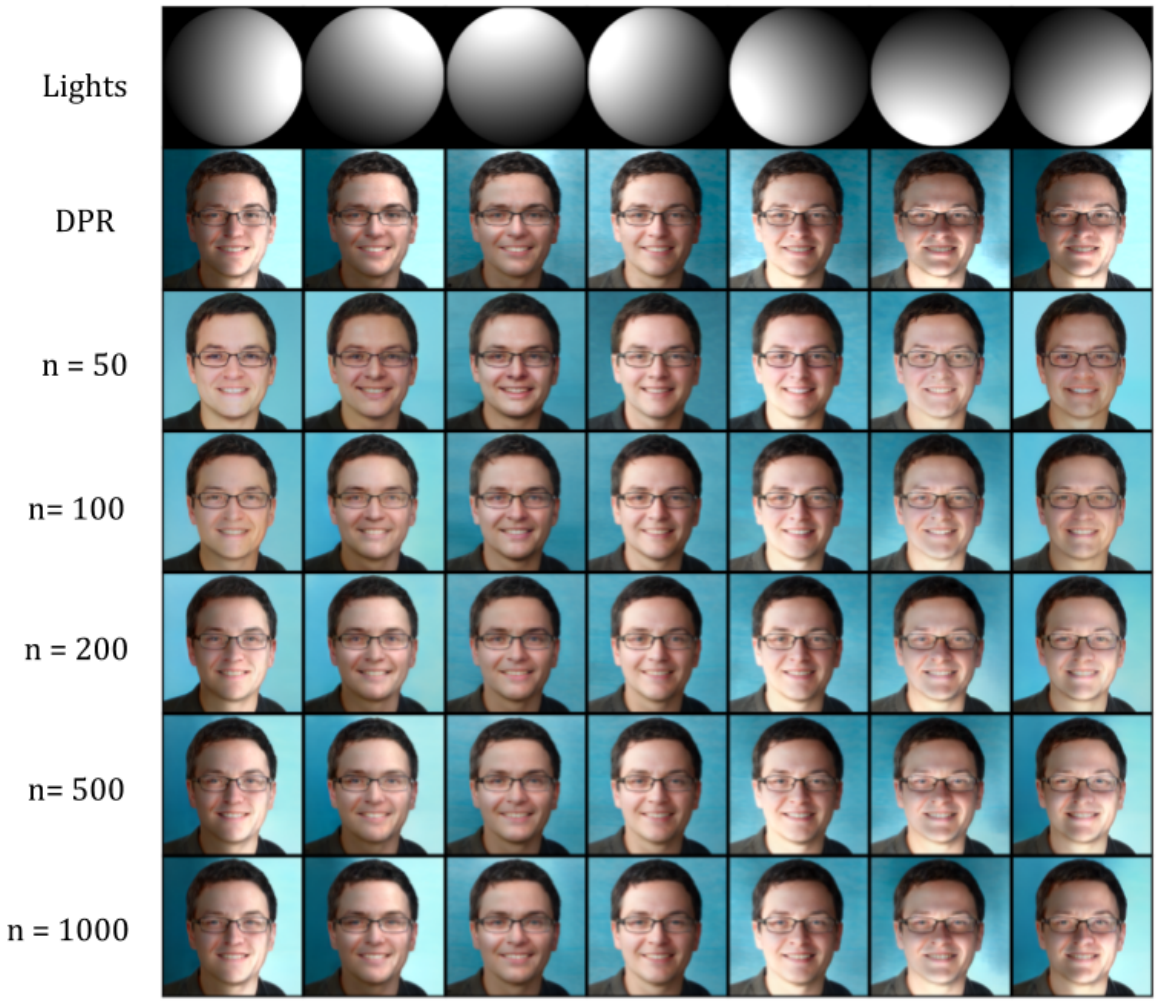

Figure 4.8.: Impact of the number iterations on the projection of relit samples. We perform the synthetic relighting using the pretrained DPR relighting network [5] under the target SH lightings. We run the optimization on all layers for different number of steps without truncation neither regularization. The main information seems to be captured around 200 iterations.

we explained already in Section 3.5, the lower the value of the deviation scale the stronger is the truncation to the original portrait. A deviation scale of $\psi = 0.7$ seems to be a reasonable value, while almost all lighting information is removed for $\psi = 0.3$. We present the results for the regularization experiment in figure 4.10. We can see that the regularization could be another effective technique to improve the quality of the relighting prediction, especially for strength values $1.0 \leq \alpha \leq 10$.

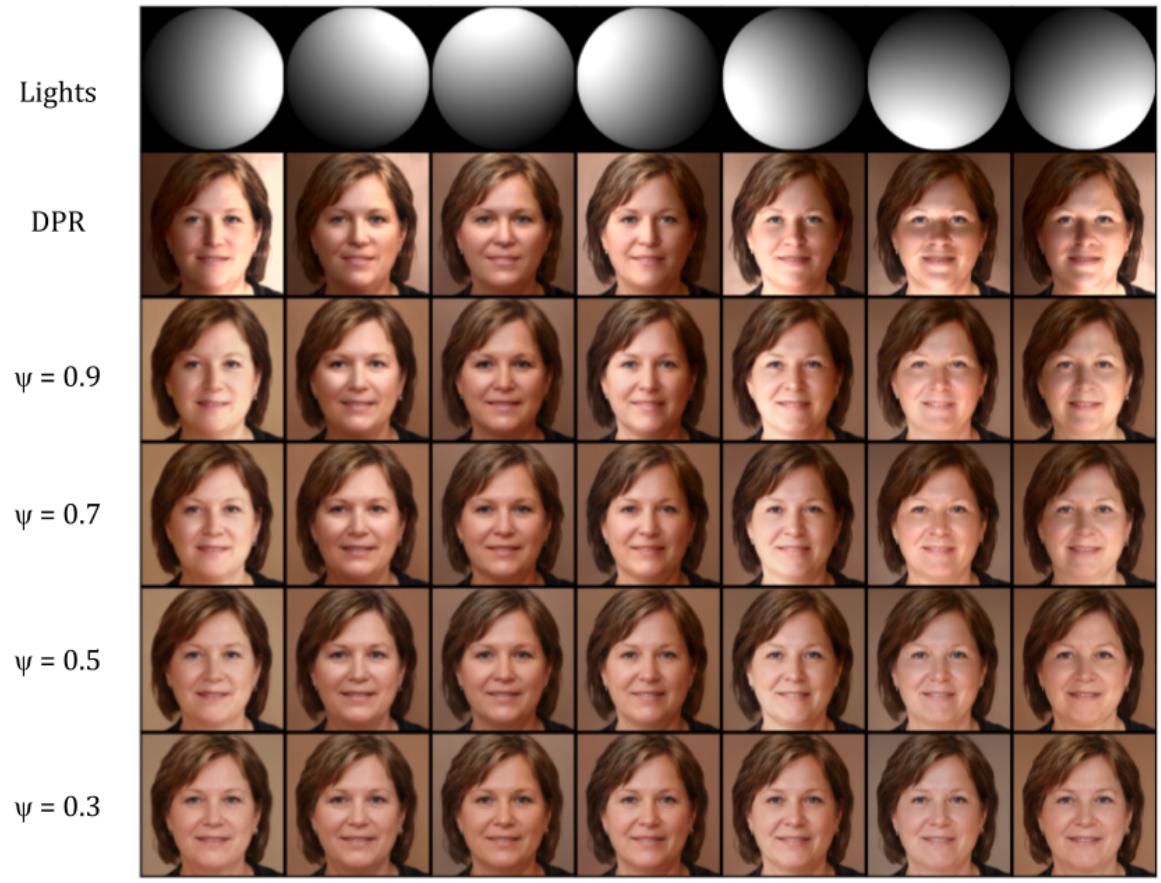

Figure 4.9.: Effect of the truncation trick on the projection of relit images for different deviation values ψ. We perform the synthetic relighting using the pretrained DPR relighting network [5] under the given SH lightings. We run the optimization only on the 9-18 layers for 200 iterations without regularization. Smaller values tend to ignore the light information.

Projection of the complete portrait relighting dataset is computationally expensive, which implies it is not possible to iterate fast on the experiments. For this reason, an exhaustive quantitative evaluation of the different hyperparameters mentioned above is unfeasible. For the final dataset, we choose the following configuration: all layers, $\psi = 0.9$, $\alpha = 0.0$ and $n = 200$. The distance between the relighting portrait dataset and the refine one is FID = 5.86,

Figure 4.10.: Impact of the regularization term on the projection of relit samples for different strength values α. We perform the synthetic relighting using the pretrained DPR relighting network [5] under the target SH lightings. We run the optimization only on the 9-18 layers for 200 steps without truncation. The main information seems to be captured around 200 iterations. Smaller values between $\alpha = 1.0$ and $\alpha = 10.0$ could be helpful, but larger values discard the light signals.

which suggests the refinement version is relative close to the original one. In figure 4.11, we show an example of this final dataset for comparison.

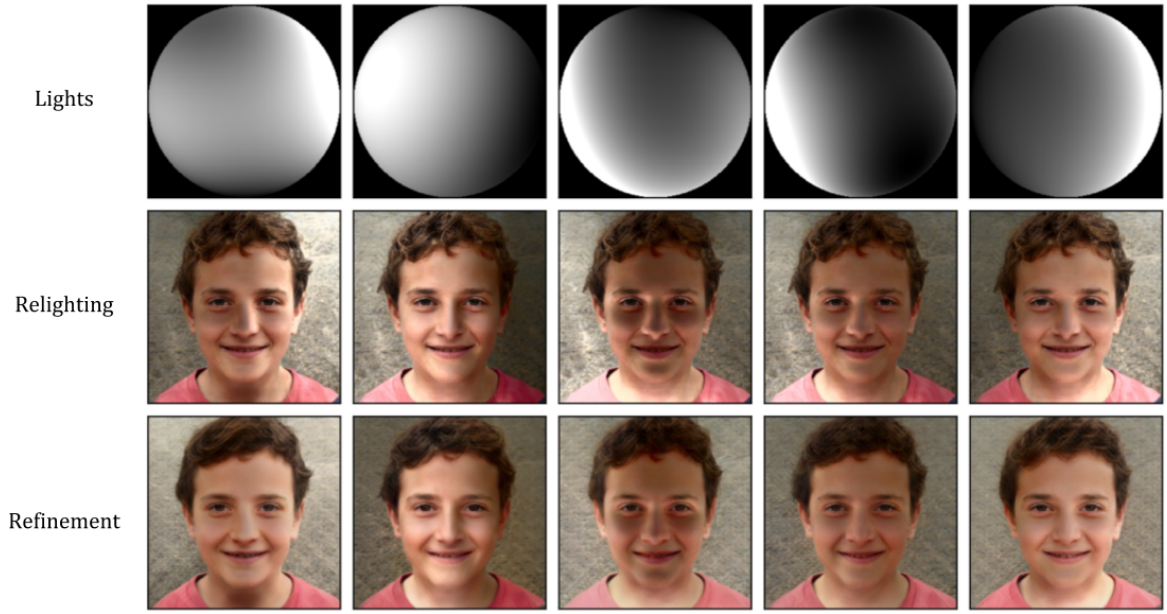

Figure 4.11.: Example of the refinement strategy for one of the synthetic portraits in our dataset. We sample five random target SH lights from a prior dataset. Then, we infer the relit images using the DPR relighting network. Finally, we obtain the refined images by projecting the relit ones into the StyleGAN space for 200 optimization steps.

In Section 3.2, we mentioned that the relighting algorithm is sensible to the imperfections of the geometry estimation. In particular, the face misalignment can cause ghost artifacts, especially around the nose. In figure 4.12, we can clearly notice the presence of these artifacts in the pretrained DPR relighting network predictions and we also demonstrate that the refinement strategy can effectively remove them.

4.6. StyleGAN Relighting

StyleGAN is unfortunately not fully controllable despite of the effort to make the intermediate space more disentangled. The GANSpace technique [54] can be used to discover interpretable light controllers, which can modify the scene illumination by traversing the latent space along the direction specified by the controller. For example, the "Sunlight in face" controller is defined by the PCA direction 10 with sigma -8 over the layer 8 of the mapping network. Nevertheless, we would like to control the scene illumination in a semantic and interpretable

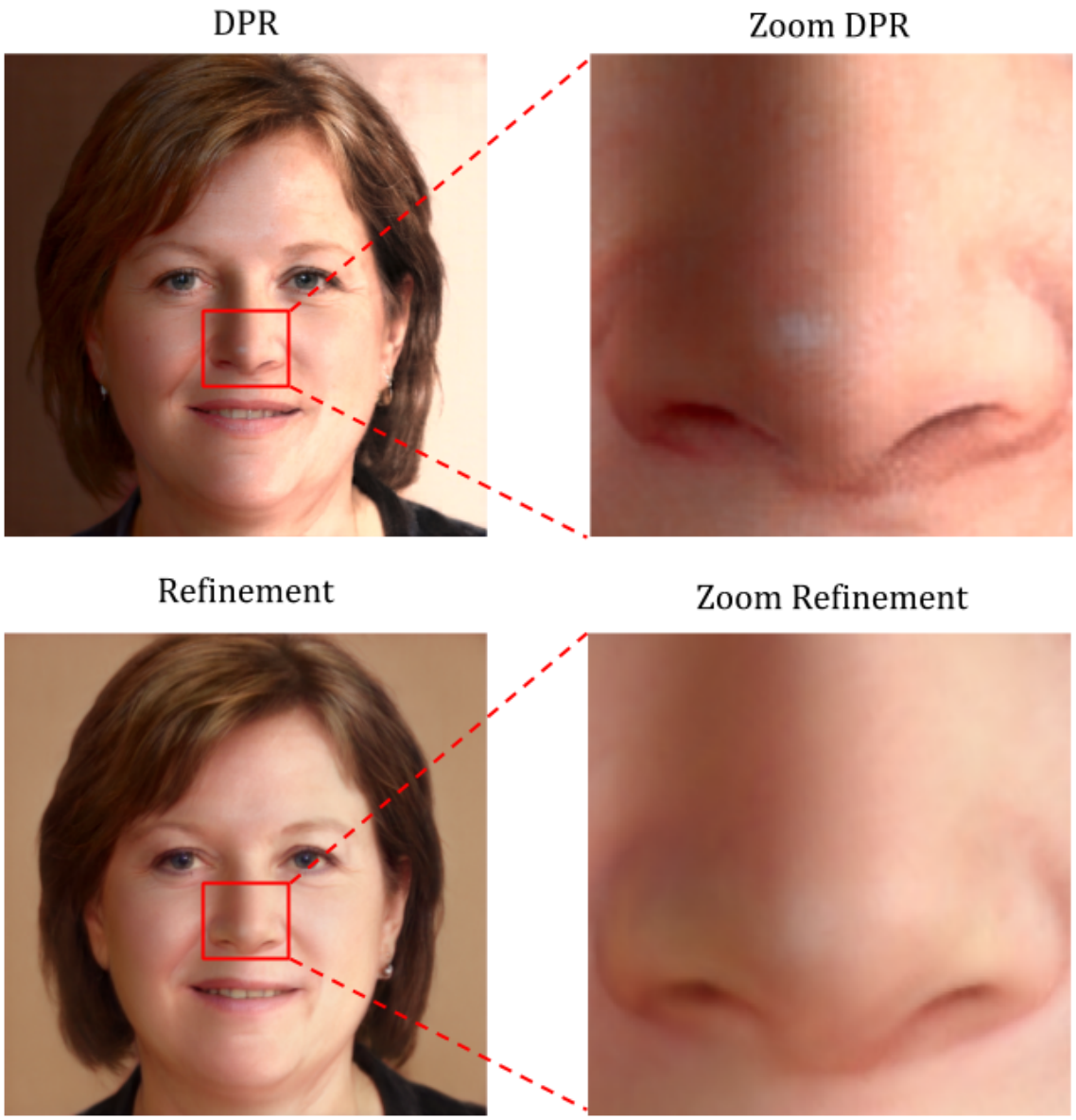

Figure 4.12.: Example of our refinement strategy effectiveness against relighting artifacts. The DPR relighting network produces an unpleasant shadow appearance around the nose of the subject. Note on the zoom images that the proposed refinement removes the grid artifacts and the artificial highlights.

way based on traditional light representations like spherical harmonics for smooth lighting environments or a 3D vector for directional light sources.

We propose explicit illumination control over the pretrained and fixed StyleGAN generator via the DPR relighting network. An illustration of the pipeline is shown in figure 4.13. Given an image I_w and their corresponding latent code w, we infer the relit images under different lighting conditions using a pretrained relighting network with fixed weights. Then, we can learn the function that outputs the relit latent codes conditioned in the target lighting $\hat{w} = LightNet(w, L_t)$ in a supervised fashion based on the error in the image domain between the relit image I_{relit} and the candidate relit image $\hat{I_w} = StyleGAN(\hat{w})$. The LightNet is a shallow neural network that learns the displacement from the original latent variable $\hat{w} = \nabla w + w$ instead of the direct mapping to the relit latent variable, which is an easier optimization problem.

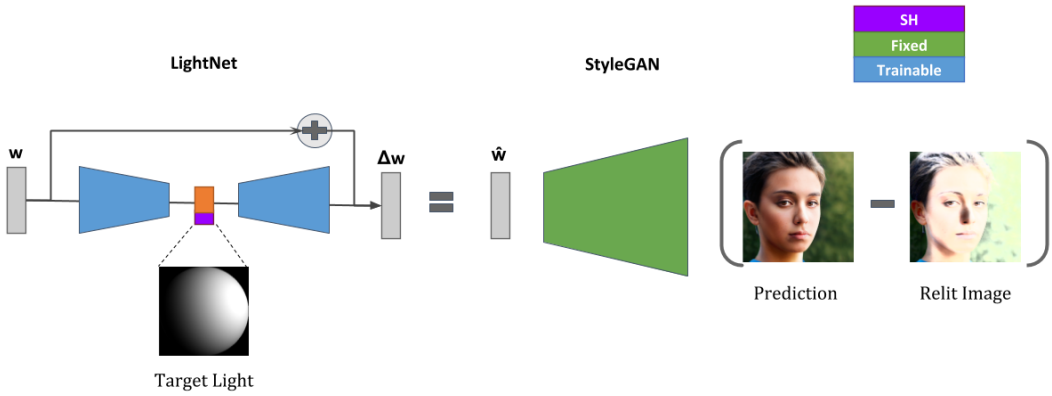

Figure 4.13.: StyleGAN relighting architecture. A shallow auto-encoder LightNet predicts the displacement of a given latent vector w under the target SH light (injected in the bottleneck layer) to the latent code $\hat{w}$ corresponding to the relit portrait image under the specified lighting conditions. We create relit images of the corresponding synthetic portrait I_w using a pretrained DPR relighting network [5] and then we use them to supervise the training of the LightNet in the image domain using a fixed and pretrained StyleGAN generator [10].

The LightNet is implemented as a two-layer perceptron with ReLU activations organized in an autoencoder fashion similar to the RigNet [56]. The first layer acts as an encoder by transforming the input latent vector of size 512 w into a lower dimensional vector l of size 32. Then, we concatenate this vector l with the target light vector L_t represented as a spherical harmonics vector and feed the resulting vector into the second layer, which acts as a decoder. Based on previous findings of the GANSpace technique, we decide to optimize only the displacement vector corresponding to the layer 8. This formulation reduces the number of parameters of the LightNet significantly, which is more efficient in terms of memory and computation.

We performed some initial experiments using this approach and we found the LightNet can indeed control the lighting conditions up to a certain degree. However, the light variations also affect another aspects of the portrait such as the eyes and hair colors. Furthermore, the StyleGAN model is big and we could only combine it with relative small models due to the hardware limitations.

5. Limitations and Future Work

The results of our StyleGAN refinement method are promising. However, we are aware of few limitations. First, we consider the limitations inherited directly from the underlying components of the proposed approach. Then, we examine the main disadvantage of the refinement method itself. Finally, we discuss another approach to combine the relighting network and the StyleGAN generator.

5.1. Hard Shadows and StyleGAN bias

We can observe in figure 5.1 that the strong highlights (first row) and the cast shadows caused by the glasses (second row) do not vary as the illumination conditions change. We attribute this behaviour to the underlying physically based relighting method for data generation. Smooth lighting approximations such as low-order spherical harmonics cannot produce hard shadows from dominant point light sources like the sun. Moreover, the ratio image-based relighting algorithm assumes the reflectance of human faces is Lambertian. In practice, human faces can also exhibit other complex reflectance properties such as specular lighting or subsurface scattering.

Ray tracing based methods could address these limitations, but they require an exact description of the geometry, material and light sources. We believe the network architecture is not suitable for the de-lighting process neither. Thus, we recommend the structured relighting architecture by Nestmeyer et al. [8], with an explicit diffuse-based image decomposition and a subsequent re-rendering to learn the non-diffuse residuals. The directional lighting representation, derived directly from the light stage data, can model any complex illumination as a sum of point lights.

On the other hand, the StyleRig paper [56] reports limitations (e.g. in plane rotations or asymmetrical expressions) on the StyleGAN model due to the bias inherited from the FFHQ dataset. We notice in our experiments that the generator tends to ignore the lighting conditions coming from the lower hemisphere, which we presume are not present in the training dataset.

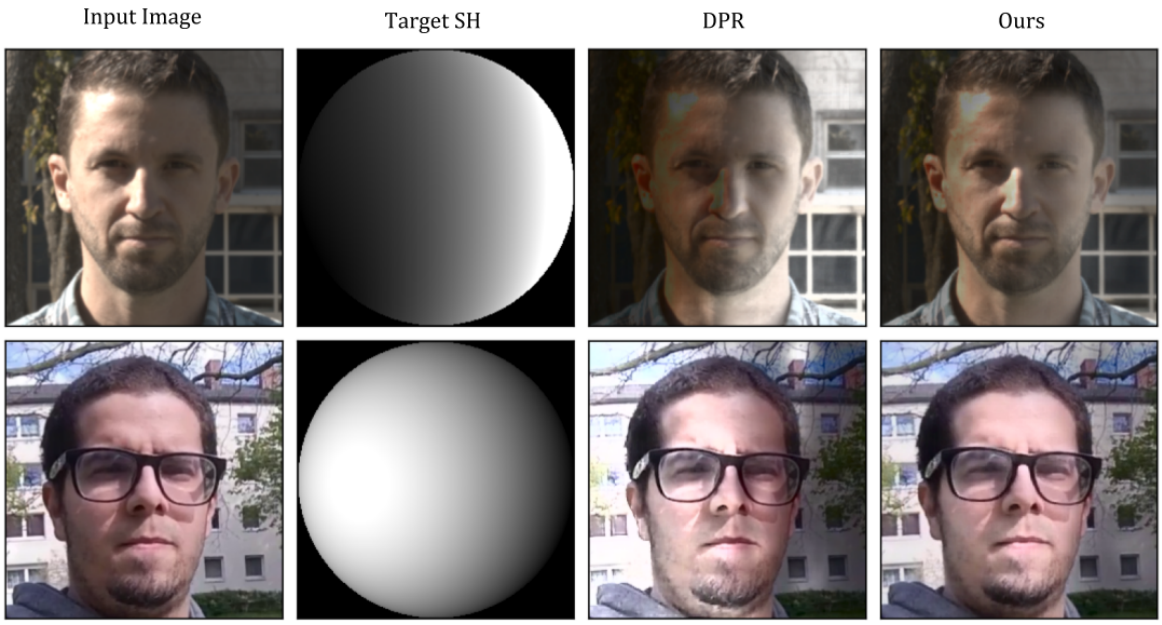

Figure 5.1.: Failure examples for the baseline DPR [5] and our method: strong highlights (first row) and glasses shadows (second row) persist on relit images.

5.2. Computational Complexity

The main drawback of our approach is the computational complexity of the refinement step in terms of both time complexity and memory requirements. The projection of each relit image into the StyleGAN space requires to solve an optimization problem using an ADAM optimizer, which is computationally expensive even for special cases with a relative low number of iterations. For instance, the refinement of 25,000 relit images using 200 iterations takes approximately 69 hours on a GeForce GTX 1080 Ti device. Additionally, we could project the images using a maximum batch size of only 4 due to the heavy memory requirements of the StyleGAN generator. For this reason, the relighting network cannot be combined with a fixed StyleGAN model in a end-to-end fashion neither.

A faster alternative could be training an image encoder (e.g. ResNet18 [69]) to predict the latent vector in the extended latent space. To facilitate the task, the encoder could only predict the offset to the mean latent code. However, it is not clear if it would generalize outside of the training dataset. An interesting direction for future work would be formulating the StyleGAN projection as a meta learning problem and leverage gradient-based meta-learning algorithms [79] to obtain the latent codes using significantly fewer iterations.

The official Python implementation of the ratio image-based relighting algorithm [5] is also expensive since several steps are executed in the CPU. For example, the cleaning normal step, which requires to solve a Poisson equation, takes approximately one day for a training set of

5000 images. In total, the complete sequential processing of this set lasts roughly 45 hours.

5.3. StyleGAN for Relighting

Another interesting direction for follow-up work is the explicit illumination control over the StyleGAN generator via the relighting network. We already showed some initial steps in this direction by training a small light encoder to predict the linear displacement from an original latent variable to its relit version. Despite of certain control, the lighting is still entangled to other factors like eyes and skin color. It would be interesting to try more advanced architectures similar to the recent StyleRig network [56] or Deng et al. work [57].

6. Conclusions

In this thesis, we have proposed a novel approach to automatically enhance learning-based face relighting methods by treating a fixed and pretrained StyleGAN model as a photorealistic portrait regularizer. The key idea is removing incorrect lighting and artifacts from a face relighting model by projecting the relit samples into the StyleGAN space and then retrain the model on the refined samples. In particular, we demonstrate the effectiveness of our approach on the DPR relighting network, a U-NET network trained to relit portrait images from a source portrait image and a target spherical harmonic lighting. This network can also perform light transfer since the relighting model predicts the lighting conditions of the input image too.

We show through both qualitative and quantitative experiments that the proposed refinement technique combined with the DPR relighting network outperforms the state-of-the-art synthetic relighting techniques in terms of the image quality and performance on the MultiPIE dataset. Furthermore, we achieve these results relying on a synthetic dataset five times smaller than the DPR dataset using a traditional training scheme guided by the LPIPS perceptual loss.

Our refinement strategy may also be helpful for other computer vision tasks such as denoising, super-resolution or inpainting. In addition, the face relighting model could also be a valuable data augmentation technique for domain adaptation on tasks such as facial recognition or 3D reconstruction. Finally, we believe that further research is needed on developing methods capable of learning from arbitrary real environments, where the illumination is directly unknown and uncontrollable. The fundamental challenge in this direction is how to leverage the relighting information from unpaired datasets, which as far as we know it has not been well investigated yet.

A. Appendix

A.1. Spherical Harmonics Lighting

We represent the lighting conditions as a second-order Spherical Harmonics (SH) as it is defined by Ramamoorthi and Hanrahan [80],.i.e., a 9 dimensional vector. Let the normal be n(p) = [x, y, z] at pixel p. Then, the 9 dimensional spherical harmonics basis $Y(p)$ in the Cartesian coordinates is formulated as follow:

$$Y = [Y_{00}, Y_{1-1}, Y_{10}, Y_{11}, Y_{2-2}, Y_{2-1}, Y_{20}, Y_{21}, Y_{22}] \tag{A.1}$$

where:

$$\begin{array}{ccccc} & & Y_{00} = 0.2821 & & \\ & Y_{1-1} = 0.4886y & Y_{10} = 0.4886z & Y_{11} = 0.4886x & \\ Y_{2-2} = 1.0925xy & Y_{2-1} = 1.0925yz & Y_{20} = 0.3154(3z^2 - 1) & Y_{21} = 1.0925xz & Y_{22} = 0.5463(x^2 - y^2) \end{array}$$

It is important to point out that the Basel face model [58] and SfSNet [3] use another coordinate systems for the spherical harmonics system. Zhou et al. [5] provide utils functions in their source code to transfer those coordinate system to their DPR coordinate system.

List of Figures

List of Tables

Bibliography

[1] J. T. Barron and J. Malik. "Shape, Illumination, and Reflectance from Shading." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 37.8 (2015), pp. 1670–1687. URL: http://dblp.uni-trier.de/db/journals/pami/pami37.html#BarronM15.

[2] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar. "Acquiring the Reflectance Field of a Human Face". In: *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '00. USA: ACM Press/Addison-Wesley Publishing Co., 2000, pp. 145–156. ISBN: 1581132085. DOI: 10.1145/344779.344855. URL: https://doi.org/10.1145/344779.344855.

[3] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. "SfSNet: Learning Shape, Refectance and Illuminance of Faces in the Wild". In: *Computer Vision and Pattern Recognition (CVPR)*. 2018.

[4] Z. Xu, K. Sunkavalli, S. Hadap, and R. Ramamoorthi. "Deep Image-based Relighting from Optimal Sparse Samples". In: *ACM Trans. Graph.* 37.4 (July 2018), 126:1–126:13. ISSN: 0730-0301. DOI: 10.1145/3197517.3201313. URL: http://doi.acm.org/10.1145/3197517.3201313.

[5] H. Zhou, S. Hadap, K. Sunkavalli, and D. W. Jacobs. "Deep Single-Image Portrait Relighting". In: *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2019.

[6] T. Sun, J. T. Barron, Y.-T. Tsai, Z. Xu, X. Yu, G. Fyffe, C. Rhemann, J. Busch, P. Debevec, and R. Ramamoorthi. "Single Image Portrait Relighting". In: *ACM Trans. Graph.* 38.4 (July 2019). ISSN: 0730-0301. DOI: 10.1145/3306346.3323008. URL: https://doi.org/10.1145/3306346.3323008.

[7] A. Meka, C. Haene, R. Pandey, M. Zollhoefer, S. Fanello, G. Fyffe, A. Kowdle, X. Yu, J. Busch, J. Dourgarian, P. Denny, S. Bouaziz, P. Lincoln, M. Whalen, G. Harvey, J. Taylor, S. Izadi, A. Tagliasacchi, P. Debevec, C. Theobalt, J. Valentin, and C. Rhemann. "Deep Reflectance Fields - High-Quality Facial Reflectance Field Inference From Color Gradient Illumination". In: vol. 38. 4. July 2019. DOI: 10.1145/3306346.3323027. URL: http://gvv.mpi-inf.mpg.de/projects/DeepReflectanceFields/.

[8] T. Nestmeyer, J.-F. Lalonde, I. Matthews, and A. M. Lehrmann. "Learning Physics-guided Face Relighting under Directional Light". In: *Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, June 2020.

[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative Adversarial Nets". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 2014, pp. 2672–2680.

[10] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. "Analyzing and Improving the Image Quality of StyleGAN". In: vol. abs/1912.04958. 2019.

[11] H. G. Barrow and J. M. Tenenbaum. *Recovering Intrinsic Scene Characteristics From Images*. Tech. rep. 157. 333 Ravenswood Ave., Menlo Park, CA 94025: AI Center, SRI International, Apr. 1978.

[12] B. K. Horn. *SHAPE FROM SHADING: A METHOD FOR OBTAINING THE SHAPE OF A SMOOTH OPAQUE OBJECT FROM ONE VIEW*. Tech. rep. USA, 1970.

[13] P. N. Belhumeur, D. J. Kriegman, and A. L. Yuille. "The Bas-Relief Ambiguity". In: *Int. J. Comput. Vision* 35.1 (Nov. 1999), pp. 33–44. ISSN: 0920-5691. DOI: 10.1023/A:1008154927611. URL: https://doi.org/10.1023/A:1008154927611.

[14] V. Blanz and T. Vetter. "A Morphable Model for the Synthesis of 3D Faces". In: *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '99. USA: ACM Press/Addison-Wesley Publishing Co., 1999, pp. 187–194. ISBN: 0201485605. DOI: 10.1145/311535.311556. URL: https://doi.org/10.1145/311535.311556.

[15] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. "Learning a model of facial shape and expression from 4D scans". In: *ACM Transactions on Graphics* 36.6 (Nov. 2017). Two first authors contributed equally, 194:1–194:17.

[16] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou. "Large Scale 3D Morphable Models". In: *Int. J. Comput. Vision* 126.2–4 (Apr. 2018), pp. 233–254. ISSN: 0920-5691. DOI: 10.1007/s11263-017-1009-7. URL: https://doi.org/10.1007/s11263-017-1009-7.

[17] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. "Neural Face Editing with Intrinsic Image Disentangling". In: July 2017, pp. 5444–5453. DOI: 10.1109/CVPR.2017.578.

[18] E. Land and J. McCann. "Lightness and Retinex Theory". In: *Journal of the Optical Society of America* 61 (Feb. 1971), pp. 1–11. DOI: 10.1364/JOSA.61.000001.

[19] A. Tewari, M. Zollöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and T. Christian. "MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction". In: *The IEEE International Conference on Computer Vision (ICCV)*. 2017.

[20] M. Zollhöfer, J. Thies, D. Bradley, P. Garrido, T. Beeler, P. Péerez, M. Stamminger, M. Nießner, and C. Theobalt. "State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications". In: (2018).

[21] B. Egger, W. A. P. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, C. Theobalt, V. Blanz, and T. Vetter. "3D Morphable Face Models - Past, Present and Future". In: *ACM Transactions on Graphics* (Sept. 2020).

[22] P. Debevec. "The Light Stages and Their Applications to Photoreal Digital Actors". In: *SIGGRAPH Asia*. Singapore, Nov. 2012. URL: http://ict.usc.edu/pubs/The%20Light%20Stages%20and%20Their%20Applications%20to%20Photoreal%20Digital%20Actors.pdf.

[23] D. Mahajan, I. K. Shlizerman, R. Ramamoorthi, and P. Belhumeur. "A Theory of Locally Low Dimensional Light Transport". In: *ACM Trans. Graph.* 26.3 (July 2007). ISSN: 0730-0301. DOI: 10.1145/1276377.1276454. URL: http://doi.acm.org/10.1145/1276377.1276454.

[24] T. Malzbender, D. Gelb, and H. Wolters. "Polynomial Texture Maps". In: *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '01. New York, NY, USA: ACM, 2001, pp. 519–528. ISBN: 1-58113-374-X. DOI: 10.1145/383259.383320. URL: http://doi.acm.org/10.1145/383259.383320.

[25] P. Ren, Y. Dong, S. Lin, X. Tong, and B. Guo. "Image Based Relighting Using Neural Networks". In: *ACM Trans. Graph.* 34.4 (July 2015), 111:1–111:12. ISSN: 0730-0301. DOI: 10.1145/2766899. URL: http://doi.acm.org/10.1145/2766899.

[26] W. Jakob. *Mitsuba renderer*. http://www.mitsuba-renderer.org. 2010.

[27] Z. Xu, S. Bi, K. Sunkavalli, S. Hadap, H. Su, and R. Ramamoorthi. "Deep View Synthesis from Sparse Photometric Images". In: *ACM Trans. Graph.* 38.4 (July 2019). ISSN: 0730-0301. DOI: 10.1145/3306346.3323007. URL: https://doi.org/10.1145/3306346.3323007.

[28] J. Thies, M. Zollhöfer, and M. Nießner. "Deferred Neural Rendering: Image Synthesis Using Neural Textures". In: *ACM Trans. Graph.* 38.4 (July 2019), 66:1–66:12. ISSN: 0730-0301. DOI: 10.1145/3306346.3323035. URL: http://doi.acm.org/10.1145/3306346.3323035.

[29] K. Guo, P. Lincoln, P. Davidson, J. Busch, X. Yu, M. Whalen, G. Harvey, S. Orts-Escolano, R. Pandey, J. Dourgarian, D. Tang, A. Tkach, A. Kowdle, E. Cooper, M. Dou, S. Fanello, G. Fyffe, C. Rhemann, J. Taylor, P. Debevec, and S. Izadi. "The Relightables: Volumetric Performance Capture of Humans with Realistic Relighting". In: *ACM Trans. Graph.* 38.6 (Nov. 2019). ISSN: 0730-0301. DOI: 10.1145/3355089.3356571. URL: https://doi.org/10.1145/3355089.3356571.

[30] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B. Goldman, and M. Zollhöfer. "State of the Art on Neural Rendering". In: *EG* (2020).

[31] L. A. Gatys, A. S. Ecker, and M. Bethge. "Image Style Transfer Using Convolutional Neural Networks". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[32] J. Johnson, A. Alahi, and L. Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution". In: *European Conference on Computer Vision*. 2016.

[33] X. Huang and S. J. Belongie. "Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization". In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 1510–1519.

[34] Y. Shih, S. Paris, C. Barnes, W. T. Freeman, and F. Durand. "Style Transfer for Headshot Portraits". In: *ACM Trans. Graph.* 33.4 (July 2014). ISSN: 0730-0301. DOI: 10.1145/2601097.2601137. URL: https://doi.org/10.1145/2601097.2601137.

[35] Z. Shu, S. Hadap, E. Shechtman, K. Sunkavalli, S. Paris, and D. Samaras. "Portrait Lighting Transfer Using a Mass Transport Approach". In: *ACM Trans. Graph.* 37.1 (Oct. 2017). ISSN: 0730-0301. DOI: 10.1145/3095816. URL: https://doi.org/10.1145/3095816.

[36] Y. Hold-Geoffroy, K. Sunkavalli, S. Hadap, E. Gambaretto, and J.-F. Lalonde. "Deep Outdoor Illumination Estimation". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017.

[37] L. Hosek and A. Wilkie. "An Analytic Model for Full Spectral Sky-Dome Radiance". In: *ACM Trans. Graph.* 31.4 (July 2012). ISSN: 0730-0301. DOI: 10.1145/2185520.2185591. URL: https://doi.org/10.1145/2185520.2185591.

[38] Y. Hold-Geoffroy, A. Athawale, and J.-F. Lalonde. "Deep Sky Modeling for Single Image Outdoor Lighting Estimation". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[39] M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J.-F. Lalonde. "Learning to Predict Indoor Illumination from a Single Image". In: *ACM Trans. Graph.* 36.6 (Nov. 2017). ISSN: 0730-0301. DOI: 10.1145/3130800.3130891. URL: https://doi.org/10.1145/3130800.3130891.

[40] D. A. Calian, J.-F. Lalonde, P. Gotardo, T. Simon, I. Matthews, and K. Mitchell. "From Faces to Outdoor Light Probes". In: *Computer Graphics Forum* (2018). ISSN: 1467-8659. DOI: 10.1111/cgf.13341.

[41] K. Nishino and S. Nayar. "Eyes for Relighting". In: *ACM Transactions on Graphics (also Proc. of SIGGRAPH)* 23.3 (July 2004), pp. 704–711.

[42] C. Legendre, W.-C. Ma, G. Fyffe, J. Flynn, L. Charbonnel, J. Busch, and P. Debevec. "DeepLight: Learning Illumination for Unconstrained Mobile Mixed Reality". In: June 2019, pp. 5911–5921. DOI: 10.1109/CVPR.2019.00607.

[43] S. Marschner and D. P. Greenberg. "Inverse Lighting for Photography". In: *Color Imaging Conference*. 1997.

[44] A. Shashua and T. Riklin-Raviv. "The Quotient Image: Class-Based Re-Rendering and Recognition with Varying Illuminations". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 23.2 (Feb. 2001), pp. 129–139. ISSN: 0162-8828. DOI: 10.1109/34.908964. URL: https://doi.org/10.1109/34.908964.

[45] A. Stoschek. "Image-based re-rendering of faces for continuous pose and illumination directions". In: vol. 1. Feb. 2000, 582–587 vol.1. ISBN: 0-7695-0662-3. DOI: 10.1109/CVPR.2000.855872.

[46] Z. Wen, Z. Liu, and T. Huang. "Face relighting with radiance environment maps". In: July 2003, pp. II–158. ISBN: 0-7695-1900-8. DOI: 10.1109/CVPR.2003.1211466.

[47] P. Peers, N. Tamura, W. Matusik, and P. Debevec. "Post-Production Facial Performance Relighting Using Reflectance Transfer". In: *ACM Trans. Graph.* 26.3 (July 2007), 52–es. ISSN: 0730-0301. DOI: 10.1145/1276377.1276442. URL: https://doi.org/10.1145/1276377.1276442.

[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[49] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[50] A. Radford, L. Metz, and S. Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2016. URL: http://arxiv.org/abs/1511.06434.

[51] T. Karras, T. Aila, S. Laine, and J. Lehtinen. "Progressive Growing of GANs for Improved Quality, Stability, and Variation". In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=Hk99zCeAb.

[52] T. Karras, S. Laine, and T. Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[53] R. Abdal, Y. Qin, and P. Wonka. "Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?" In: *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2019.

[54] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris. "GANSpace: Discovering Interpretable GAN Controls". In: *ArXiv* abs/2004.02546 (2020).

[55] A. Brock, J. Donahue, and K. Simonyan. "Large Scale GAN Training for High Fidelity Natural Image Synthesis". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=B1xsqj09Fm.

[56] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Pérez, M. Zöllhofer, and C. Theobalt. "StyleRig: Rigging StyleGAN for 3D Control over Portrait Images, CVPR 2020". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. June 2020.

[57] Y. Deng, J. Yang, D. Chen, F. Wen, and T. Xin. "Disentangled and Controllable Face Image Generation via 3D Imitative-Contrastive Learning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[58] B. Egger, S. Schönborn, A. Schneider, A. Kortylewski, A. Morel-Forster, C. Blumer, and T. Vetter. "Occlusion-Aware 3D Morphable Models and an Illumination Prior for Face Image Analysis". In: *Int. J. Comput. Vision* 126.12 (Dec. 2018), pp. 1269–1287. ISSN: 0920-5691. DOI: 10.1007/s11263-018-1064-8. URL: https://doi.org/10.1007/s11263-018-1064-8.

[59] R. Basri and D. W. Jacobs. "Lambertian Reflectance and Linear Subspaces". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 25.2 (Feb. 2003), pp. 218–233. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2003.1177153. URL: https://doi.org/10.1109/TPAMI.2003.1177153.

[60] X. Zhu, X. Liu, Z. Lei, and S. Z. Li. "Face alignment in full pose range: A 3d total solution". In: *IEEE transactions on pattern analysis and machine intelligence* (2017).

[61] J. Guo, X. Zhu, and Z. Lei. *3DDFA*. https://github.com/cleardusk/3DDFA. 2018.

[62] O. Sorkine and M. Alexa. "As-Rigid-as-Possible Surface Modeling". In: *Proceedings of the Fifth Eurographics Symposium on Geometry Processing*. SGP '07. Barcelona, Spain: Eurographics Association, 2007, pp. 109–116. ISBN: 9783905673463.

[63] V. Kazemi and J. Sullivan. "One Millisecond Face Alignment with an Ensemble of Regression Trees". In: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. CVPR '14. USA: IEEE Computer Society, 2014, pp. 1867–1874. ISBN: 9781479951185. DOI: 10.1109/CVPR.2014.241. URL: https://doi.org/10.1109/CVPR.2014.241.

[64] S. Shirdhonkar and D. W. Jacobs. "Non-negative lighting and specular object recognition". In: *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*. Vol. 2. Oct. 2005, 1323 -1330 Vol. 2 - 1323 –1330 Vol. 2. DOI: 10.1109/ICCV.2005.168.

[65] S. Liu, J. Yang, C. Huang, and M.-H. Yang. "Multi-Objective Convolutional Learning for Face Labeling". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.

[66] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. "BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation". In: *ECCV* (2018).

[67] zll. *Face parsing in PyTorch*. https://github.com/zllrunning/face-parsing.PyTorch. 2019.

[68] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. "MaskGAN: Towards Diverse and Interactive Facial Image Manipulation". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[69] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". In: June 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[70] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 586–595.

[71] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations*. 2015.

[72] H. Kim, M. Zollhöfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt. "InverseFaceNet: Deep Monocular Inverse Face Rendering". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[73] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2015. URL: http://arxiv.org/abs/1412.6980.

[74] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. "Automatic differentiation in PyTorch". In: *NIPS-W*. 2017.

[75] K. Seonghyeon. *Implementation of Analyzing and Improving the Image Quality of StyleGAN (StyleGAN 2) in PyTorch*. https://github.com/rosinality/stylegan2-pytorch. 2019.

[76] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. "Multi-PIE". In: *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE Computer Society, Sept. 2008. URL: https://www.microsoft.com/en-us/research/publication/multi-pie/.

[77] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity". In: *Trans. Img. Proc.* 13.4 (Apr. 2004), pp. 600–612. ISSN: 1057-7149. DOI: 10.1109/TIP.2003.819861. URL: https://doi.org/10.1109/TIP.2003.819861.

[78] 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". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6629–6640. ISBN: 9781510860964.

[79] C. Finn, P. Abbeel, and S. Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, 2017, pp. 1126–1135.

[80] R. Ramamoorthi and P. Hanrahan. "An Efficient Representation for Irradiance Environment Maps". In: *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '01. New York, NY, USA: Association for Computing Machinery, 2001, pp. 497–500. ISBN: 158113374X. DOI: 10.1145/383259.383317. URL: https://doi.org/10.1145/383259.383317.