PhotoApp: Photorealistic Appearance Editing of Head Portraits

MALLIKARJUN B R and AYUSH TEWARI, MPI for Informatics, SIC, Germany
ABDALLAH DIB, InterDigital R&I, France
TIM WEYRICH, University College London, UK
BERND BICKEL, IST Austria, Austria
HANS-PETER SEIDEL, MPI for Informatics, SIC, Germany
HANSPETER PFISTER, Harvard University, USA
WOJCIECH MATUSIK, MIT CSAIL, USA
LOUIS CHEVALLIER, InterDigital R&I, France
MOHAMED ELGHARIB and CHRISTIAN THEOBALT, MPI for Informatics, SIC, Germany

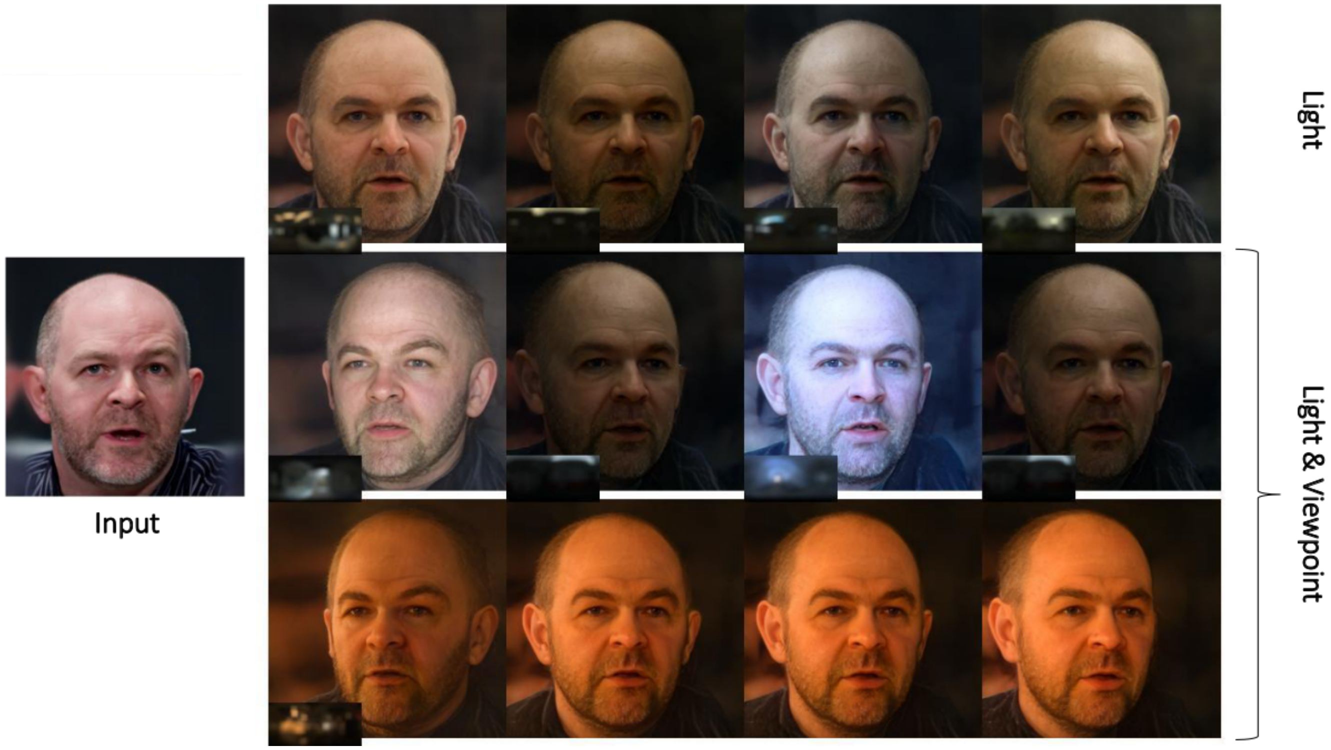

Fig. 1. We present a method for high-quality appearance editing of head portraits. Given an input image, our approach edits its appearance using a target environment map (see insets), and a target camera viewpoint. We achieve high-quality photorealistic results for in the wild images, capturing a wide variety of reflectance properties. Our method is trained on a light-stage dataset, using a combination of supervised learning and generative adversarial modeling which allows for accurate editing as well as generalisation outside the dataset.

Photorealistic editing of head portraits is a challenging task as humans are very sensitive to inconsistencies in faces. We present an approach for high-quality intuitive editing of the camera viewpoint and scene illumination (parameterised with an environment map) in a portrait image. This requires our method to capture and control the full reflectance field of the person in the image. Most editing approaches rely on supervised learning using training data captured with setups such as light and camera stages. Such datasets are expensive to acquire, not readily available and do not capture all the rich variations of in-the-wild portrait images. In addition, most supervised approaches only focus on relighting, and do not allow camera viewpoint editing. Thus, they only capture and control a subset of the reflectance field. Recently, portrait editing has been demonstrated by operating in the generative model space of StyleGAN. While such approaches do not require direct supervision, there is a significant loss of quality when compared to the supervised approaches. In this paper, we present a method which learns from limited supervised training data. The training images only include people in a fixed neutral expression with eyes closed, without much hair or background variations. Each person is captured under 150 one-light-at-a-time conditions and under 8 camera poses. Instead of training directly in the image space, we design a supervised problem which learns transformations in the latent space of StyleGAN. This combines the best of supervised learning and generative adversarial modeling. We show that the StyleGAN prior allows for generalisation to different expressions, hairstyles and backgrounds. This produces high-quality photorealistic results for in-the-wild images and significantly outperforms existing methods. Our approach can edit the illumination and pose simultaneously, and runs at interactive rates.

Additional Key Words and Phrases: Portrait Editing, Relighting, Pose Editing, Neural Rendering

1 INTRODUCTION

Portrait photos are among the most important photographic depictions of humans and their loved ones. Even though the quality of cameras and thus the photographs have improved dramatically, there arises many cases where people would like to change the scene illumination and camera pose after the image has been captured. Editing the appearance of the image after capture has applications in post-production, casual photography and virtual reality. Given a monocular portrait image and a target illumination and camera pose, we present a method for relighting the portrait and editing the camera pose in a photorealistic manner. This is a challenging task, as the appearance of the person in the image includes complex effects such as subsurface scattering and self-shadowing. Changing the camera requires reasoning about occluded surfaces. Humans are very sensitive to inconsistencies in portrait images, and a high level of photorealism is necessary for convincing editing. This requires our method to correctly reason about the interactions of the lights in the scene with the surface, and edit them at photorealistic quality. We are interested in editing in-the-wild images with a very wide range of illumination and pose conditions. We only rely on a single image of an identity unseen during training. These constraints make the problem very challenging.

Several methods have been proposed for editing portrait appearance in the literature. One category of methods [Debevec et al. 2000; Ghosh et al. 2011; Weyrich et al. 2006] address this problem by explicitly modelling the reflectance of the human face [Kajiya 1986]. While these approaches provide well-defined, semantically meaningful reflectance output, they require the person to be captured under multi-view and multi-lit configurations. They also do not edit the full portrait image, just the inner face region, missing out important portrait components such as hair and eyes. Recently, several deep learning-based methods have been proposed for appearance editing. These methods use large light-stage datasets which consist of a limited number of people illuminated by different light sources and captured from different camera viewpoints. A neural network is trained on such datasets which enables inference from a single image. Some methods [Lattas et al. 2020; Yamaguchi et al. 2018] regress the reflectance of the face from a monocular image in the form of diffuse and specular components. Neural representations for face reflectance fields have also been explored recently [B R et al. 2020]. While these methods can work with a single image, they still only model the inner face region, missing out on important details such as hair and eyes.

In contrast to the previous methods, several approaches only capture and edit a subset of the reflectance field. These approaches only allow for the editing of either scene illumination or camera pose. Most relighting methods directly learn a mapping from the input image to its relit version using a light-stage training dataset [Nestmeyer et al. 2020; Sun et al. 2019, 2020]. The controlled setting and limited variety of such datasets limits performance while generalising to in-the-wild images. Zhou et al. [2019] attempted to break out from the complexity of capturing light-stage datasets and from their limited variations. Instead, they proposed to use a synthetic dataset of in-the-wild images, synthesised with different illuminations. Illumination is modeled using spherical harmonics. The use of synthetic data impacts the photorealism of the results. All of these approaches do not allow for changing the camera pose. Several methods exist for only editing the camera pose and expressions [Averbuch-Elor et al. 2017; Geng et al. 2018; Kim et al. 2018; Nagano et al. 2018; Siarohin et al. 2019; Wiles et al. 2018]. These methods are commonly trained on videos. While person-specific methods [Kim et al. 2018; Thies et al. 2019] can obtain high-quality results, methods which generalise to unseen identities [Siarohin et al. 2019; Wiles et al. 2018] are limited in terms of photorealism. In addition, none of them can edit the scene illumination.

Recently, Tewari et al. [2020b] proposed Portrait Image Embedding (PIE), an approach for editing the illumination and camera pose in portrait images by leveraging the StyleGAN generative model [Karras et al. 2019]. PIE computes a StyleGAN embedding for the input image which allows for editing of various face semantics. As StyleGAN represents a manifold of photorealistic portraits, PIE can edit the full image with high quality. However, due to the absence of labelled data, the supervision for the method is defined using a 3D reconstruction of the face. This supervision is indirect and not over the complete image, leading to results that still lack sufficient accuracy and photorealism. It uses a low-dimensional representation of the scene illumination and can thus not synthesize results with higher-frequency lights. Furthermore, PIE solves a computationally expensive optimisation problem taking several minutes to compute the embedding.

We therefore propose a technique for high-quality intuitive editing of scene illumination and camera pose in a head portrait image. Our method combines the best of generative modeling and supervised learning approaches, and creates results of much higher quality compared to previous methods. We learn to transform the StyleGAN latent code of the input image into the latent code of the output. We perform this learning in a supervised manner by leveraging a light-stage dataset, containing multiple identities shot from different viewpoints and under several illumination conditions. Learning in the StyleGAN space allows us to synthesise photorealistic results for general person identities seen under in-the-wild conditions. Our method can handle properties such as shadows and other complex appearance, and can synthesise full portrait images including hair, upper body and background. We inherit the high photorealism and diversity of the StyleGAN portrait manifold in our solution, which allows us to outperform methods that only use light-stage training data [Sun et al. 2019]. Our method has analogies to self-supervised discriminative methods [Jing and Tian 2020]. We show that the StyleGAN latent representation allows for generalisation even with very little training data. We obtain high-quality results of our method even when trained on just 15 identities. Our formulation does not make any prior assumptions on the underlying surface reflectance or scene illumination (other than it being distant) and rather directly predicts the appearance as a function of the target environment map and camera pose. This leads to significantly more photorealistic results compared to methods that use spherical-harmonic illumination representations [Abdal et al. 2020; Tewari et al. 2020b; Zhou et al. 2019] which are limited to only modeling low-frequency illumination conditions. Furthermore, directly supervising our method using a multi-view and multi-lit light-stage dataset allows us to produce significantly more photorealistic results than PIE [Tewari et al. 2020b]. Our method can additionally edit at a faster speed, using just a single feedforward pass, and also edit both illumination and pose simultaneously, unlike PIE. Compared to traditional relighting approaches [Sun et al. 2020; Zhou et al. 2019], we obtain higher-quality results as well as allow for changing the camera pose. In summary we make the following contributions:

- We combine the strength of supervised learning and generative adversarial modeling in a new way to develop a technique for high-quality editing of scene illumination and camera pose in portrait images. Both properties can be edited simultaneously.
- Our novel formulation allows for generalisation to in-the-wild images with significantly higher quality results than related methods. It also allows for training with limited amount of supervision.

2 RELATED WORK

In this section we look at related works that can edit the scene parameters in a head portrait image. We refer the reader to the state-of-the-art report of Tewari et al. [2020c] for more details on neural rendering approaches.

The seminal work of Debevec et al. [2000] introduced a light-stage apparatus to capture the *reflectance field* of a human face, that is, its appearance under a multitude of different lighting directions. Through weighted superposition of images of the illumination conditions, their method recreates high-quality images of the face under any target illumination. By employing additional cameras and geometry reconstruction, and gathering data from the additional viewpoints, they further fit a simple bi-directional radiance distribution function (BRDF) allowing for novel-light *and* -view renderings of the face. Their method, however, is limited to reproducing the specific face that was captured. Weyrich et al. [2006] extend this concept using a setup with a much larger number of cameras (16) and a reconstruction pipeline that extracts geometry and a detailed spatially-varying BRDF (SVBRDF) of a face. By scanning hundreds of subjects that way, they extract generalisable statistical information on appearance traits depending on age, gender and ethnicity. The generative power of the extracted quantities, however, is heavily constrained, and examples of sematic appearance editing were limited to subjects from within their face database. In our work, we revisit their original dataset using a state-of-the-art learning framework.

Another category of methods tries to infer geometry and reflectance properties from single, unconstrained images. Shu et al. [2017] and Sengupta et al. [2018] decompose the image into simple intrinsic components, that is, normals, diffuse albedo and shading. With the assumption of Lambertian surface reflectance, these methods use spherical harmonics to model the scene illumination; however, the starkly simplified assumption ignores perceptually important reflectance properties which leads to limited photrealism. Others infer more general surface reflectance, with fewer assumptions about incident illumination [B R et al. 2020; Lattas et al. 2020; Yamaguchi et al. 2018]. While such techniques can capture rich reflectance properties, they do not synthesise the full portrait, missing out on important components such as hair, eyes and mouth.

Recently, several methods addressed the simplified problem of only relighting a head portrait in the fixed input pose [Meka et al. 2019; Nestmeyer et al. 2020; Sun et al. 2019; Wang et al. 2020; Zhang et al. 2020; Zhou et al. 2019]. Nestmeyer et al. [2020] used a light-stage dataset to train a model that explicitly regresses a diffuse reflectance, as well as a residual component which accounts for specularity and other effects. Similarly, Wang et al. [2020] used a light-stage dataset to compute the ground truth diffuse albedo, normal, specularity and shadow images. A network is trained to regress each of these components which are then used in another network to finally relight the portrait image. Instead of explicitly estimating the different reflectance components, methods such as Sun et al. [2019]; Zhou et al. [2019] directly regress the relighted version of the portrait given the imput image and target illumination. Here, the target illumination is parameterised either in the form of environment map [Sun et al. 2019] or spherical harmonics [Zhou et al. 2019]. While Sun et al. [2019] used light-stage data to obtain their ground truth for supervised learning, Zhou et al. [2019] used a ratio image-based approach to generate synthetic training data.

Recently, Zhang et al. [2020] proposed a method to remove harsh shadows from a monocular portrait image. They created a synthetic data from in-the-wild images by augmenting shadows and training a network to remove these shadows. Using a light-stage dataset, another network is trained to smooth the artifacts that could remain from the first network. While the methods of [Meka et al. 2019; Nestmeyer et al. 2020; Sun et al. 2019; Wang et al. 2020; Zhang et al. 2020; Zhou et al. 2019] can produce high-quality relighting results, they either focus on shadow removal [Zhang et al. 2020], are limited by spherical-harmonics illumination representation [Zhou et al. 2019]. In addition, methods trained on light-stage or synthetic datasets struggle to generalise to in-the-wild. They are also limited to only relighting, as they cannot change the camera viewpoint.

There are several methods for editing the head pose of portrait images [Averbuch-Elor et al. 2017; Geng et al. 2018; Kim et al. 2018; Nagano et al. 2018; Siarohin et al. 2019; Wiles et al. 2018]. While Kim et al. [2018] require a training video of the examined subject, the techniques of Averbuch-Elor et al. [2017]; Geng et al. [2018]; Nagano et al. [2018]; Siarohin et al. [2019]; Wiles et al. [2018] can directly operate on a single image. However, Nagano et al. [2018] does not synthesise the hair and the approaches of Siarohin et al. [2019]; Wiles et al. [2018] lack explicit 3D modeling and only allow for control using a driving video. The approaches of Averbuch-Elor et al. [2017]; Geng et al. [2018] rely on warping of the image guided by face mesh deformations, and are thus limited to very small edits in pose. Furthermore, these approaches can not change the scene illumination.

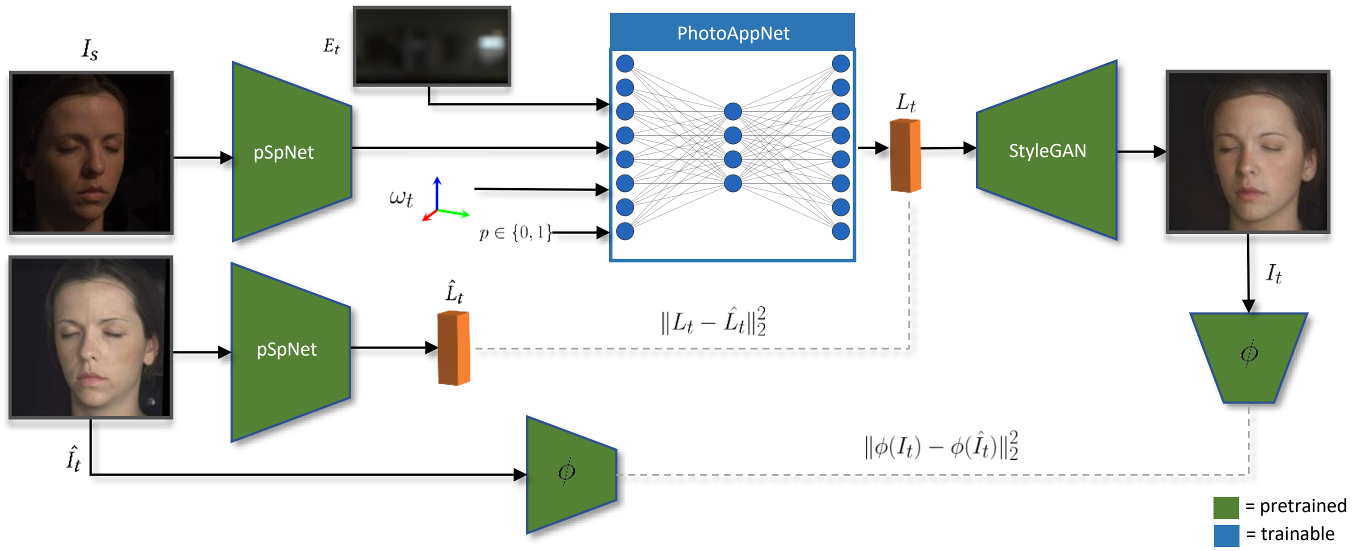

Fig. 2. Our method allows for editing the scene illumination E_t and camera pose ω_t in an input source image I_s. We learn to map the StyleGAN [Karras et al. 2020] latent code L_s of the source image, estimated using pSpNet [Richardson et al. 2020] to the latent code L_t of the output image. StyleGAN [Karras et al. 2020] is then used to synthesis the final output I_t. Our method is trained in a supervised manner using a light-stage dataset with multiple cameras and light sources. For training, we use a latent loss and a perceptual loss defined using a pretrained network ϕ. Supervised learning in the latent space of StyleGAN allows for high-quality editing which can generalise to in-the-wild images.

Recently, Tewari et al. [2020b] proposed PIE, a method which can relight, change expressions and synthesise novel views of the portrait image using a generative model. PIE is based on StyleRig [Tewari et al. 2020a] which maps the control space of a 3D morphable face model to the latent space of StyleGAN [Karras et al. 2019] in a self-supervised manner. It further imposes an identity perseverance loss to ensure the source identity is maintained during editing. Even though PIE inherits the high photorealism of the StyleGAN portrait manifold, its lack of direct supervision for appearance editing limits its performance and impacts the overall photorealism. The scene illumination is parameterised using spherical harmonics as it relies on a monocular 3D reconstruction approach to define its control space. Thus, it only allows for rendering using low-frequency scene illumination. In addition, PIE can not edit the illumination and pose simultaneously, but rather one at a time. PIE solves an expensive optimisation for the image which is time consuming, taking around 10 minutes per image on an NVIDIA V100 GPU. Concurrent to us, Abdal et al. [2020] also propose a method for semantic editing of portrait images using latent space transformations of StyleGAN. They use an invertible network based on continuous normalising flows to map semantic input parameters such as head pose and scene illumination into the StyleGAN latent vectors. The input parametrisation for the illumination is spherical harmonics like PIE, which limits its relighting capabilities. This method is also trained without explicit supervision, i.e., images of the same person with different scene parameters. This limits the quality of the results. While there are several other approaches which demonstrate transformations of StyleGAN latent vectors for semantic manipulation [Collins et al. 2020; Härkönen et al. 2020; Shen et al. 2020; Tewari et al. 2020a], these methods focus on StyleGAN generated images, and do not produce high-quality and high-resolution results for real existing images.

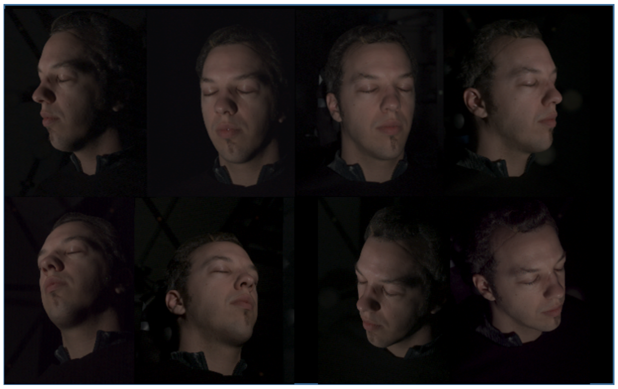

Fig. 3. Visualisation of the camera poses in the training dataset.

3 METHOD

Our method takes as input an in-the-wild portrait image, a target illumination and the target camera pose. The output is a portrait image of the same identity, synthesised with the target camera and lit by the target illumination. Given a light-stage dataset of multiple independent illumination sources and viewpoints, the naive approach could be to learn the transformations directly in image space. Instead, we propose to learn the mapping in the latent space of StyleGAN [Karras et al. 2020]. We show that learning using this latent representation helps in generalisation to in-the-wild images with high photorealism. We use StyleGAN2 in our implementation, referred to as StyleGAN for better comprehension.

3.1 Dataset

We make use of a light-stage [Weyrich et al. 2006] dataset for training our solution. This dataset contains 341 identities captured with 8 different cameras placed in the frontal hemisphere of the face. The camera poses available are shown in Fig. 3. The dataset also contains 150 light source evenly distributed on the sphere. Using this setup, each image is captured with one-light-at-a-time (OLAT) light. Given 150 OLAT images of a person with a specific camera pose, we can linearly combine them using an environment map to obtain relight portrait images [Debevec et al. 2000]. We use 205 HDR environment maps from the Naval Outdoor [Hold-Geoffroy et al. 2019] and 2233 from the Naval Indoor [Gardner et al. 2017] dataset for generating naturally lit images. Camera poses for the images are estimated using the approach of Yang et al. [2019]. Out of the 341 identities, we use 300 for training and the rest for testing. We synthesise 300 transformed images for each identity with randomly selected environment maps and camera viewpoints. Our training set consists of input-ground truth pairs of the same identity along with the target pose and environment map. The camera viewpoint of the ground truth is kept identical to the input for quarter of the training data. In the remaining, this camera viewpoint is randomly selected. The test set includes pairs from the test identities for quantitative evaluations, as well as in-the-wild images for qualitative evaluations, see Sec. 4.

3.2 Network Architecture

Fig. 2 shows an overview of our method. Our approach takes as input a source image I_s, target illumination E_t and camera pose ω_t, and a binary input p. The value of p is set to 0 when only relighting is performed, and 1 when we also want to edit the camera pose. This conditioning input helps in better preservation of the input camera pose for relighting results. The ground truth image for training is represented as $\hat{I_t}$. Camera pose is parameterised using Euler angles. We represent the illumination E_t as a 450 dimensional vectorised environment map. This corresponds to the 150 RGB discrete light sources. A core component of our approach is the PhotoAppNet neural network, which maps the input image to the edited output image in the latent space of StyleGAN (see Fig. 2). We first compute the latent representations of I_s and $\hat{I_t}$ as L_s and $\hat{L_t}$ using the pretrained network of Richardson et al. [2020] (pSpNet in Fig. 2). The latent space used is 18×512 dimensional, corresponding to the W+ space of StyleGAN. The output of PhotoAppNet is a displacement to the input in the StyleGAN latent space. This is then added to the input latent code to compute L_t, which is used by StyleGAN to generate the output image I_t. We only train PhotoAppNet, while pSpNet and StyleGAN are pretrained and fixed.

We use an MLP-based architecture with a single hidden layer of length 512. ReLU activation is used after the hidden layer. We use independent networks for each of the 18 latent vectors of length 512 corresponding to different resolutions. This is motivated by the design of the StyleGAN network where each 512 dimensional latent code controls a different frequency of image features. The output of each independent network is the output latent code corresponding to the same resolution.

3.3 Loss Function

We use multiple loss terms to train our network.

$$\mathcal{L}(I_t, L_t, \hat{I_t}, \hat{L_t}, \theta_\mathbf{n}) = \mathcal{L}_\mathrm{l}(L_t, \hat{L_t}, \theta_\mathbf{n}) + \mathcal{L}_\mathrm{p}(I_t, \hat{I_t}, \theta_\mathbf{n}) \ . \qquad (1)$$

Here, θ_n denotes the network parameters of PhotoAppNet. Both terms are weighed equally. The first term is a StyleGAN latent loss defined as

$$\mathcal{L}_\mathrm{l}(L_t, \hat{L_t}, \theta_\mathbf{n}) = \|L_t - \hat{L_t}\|_2^2 \ .$$

It enforces the StyleGAN latent code of the output image L_t to be close to the ground truth latent code $\hat{L_t}$. The second term is a perceptual loss defined as

$$\mathcal{L}_\mathrm{p}(I_t, \hat{I_t}, \theta_\mathbf{n}) = \|\phi(I_t) - \phi(\hat{I_t})\|_2^2 \ .$$

Here, we employ the learned perceptual similarity metric LPIPS [Zhang et al. 2018]. An ℓ_2 loss is used to compare the AlexNet [Krizhevsky et al. 2012] features $\phi()$ of the synthesised output and the ground truth images.

3.4 Network Training

We implement our method in PyTorch and optimise for the weights of PhotoAppNet by minimising the loss function in Eq. 1. We use Adam solver with a learning rate of 0.0001 and default hyperparameters. As mentioned earlier, the StyleGAN encoder (pSpNet) and generator [Karras et al. 2020; Richardson et al. 2020] are pretrained and fixed during training. We optimise over our training set samples using a batch size equal to 1. Since in-the-wild images are very different from the light-stage data, it is difficult to assess convergence using a light-stage validation dataset. As such, we train our networks using an in-the-wild validation set using qualitative evaluations. Our network take around 10 hours to train on a single NVIDIA Quadro RTX 8000 GPU.

3.5 Discussion

Existing image-based relighting approaches such as Sun et al. [2020]; Zhou et al. [2019] rely on much larger trainable networks with several loss functions, such as losses on the input environment map and adversarial losses. Approaches for pose editing such as Kim et al. [2018]; Siarohin et al. [2019]; Thies et al. [2019] rely on conditional generative networks trained with a combination of photometric and adversarial losses. Since we rely on a pretrained generator as our backend renderer, our training is much simpler than existing approaches. We do not need an adversarial loss as the pretrained generator already synthesises results at a high quality. As such, our training is more stable than approaches operating in image space. In addition. the StyleGAN latent representation allows for generalisation with high-quality, even when trained on a dataset with as little as 3 identities (Sec. 4.4). Many existing methods use specialised network architectures for editing the pose such as landmark-based warping of the features [Siarohin et al. 2019], or rendering of a coarse 3D face model [Kim et al. 2018; Thies et al. 2019]. Similarly, common relighting networks are designed in a task-specific manner where the illumination is predicted at the bottleneck

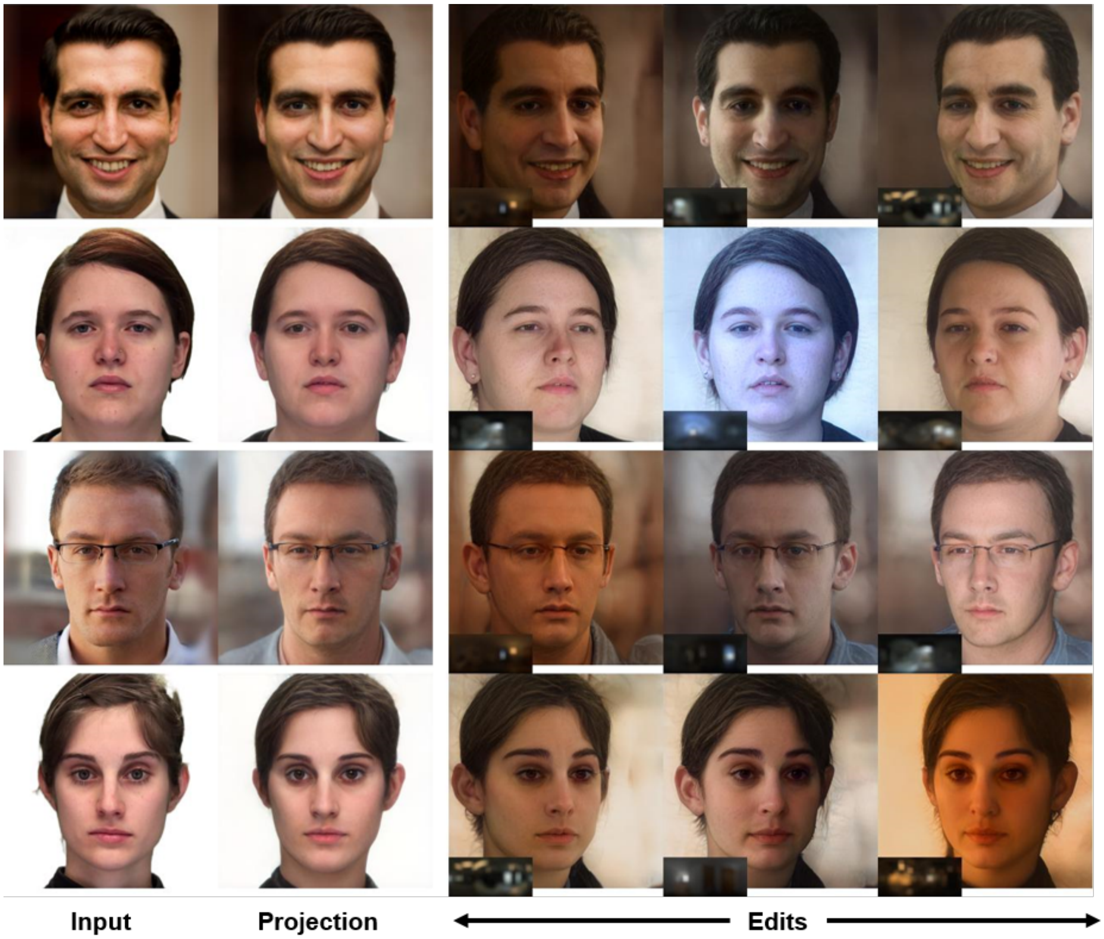

Fig. 4. Qualitative illumination and viewpoint editing results. The environment map of the target illumination is shown in the insets. We visualize the StyleGAN projection of the input image (second column). Our method produces photorealistic editing results even under challenging high-frequency light conditions.

of the architecture [Sun et al. 2020; Zhou et al. 2019]. Our design results in a compact and convenient to train PhotoAppNet network that does not require any sophisticated nor specialized network components. In addition, our method is also faster to train compared to these approaches, since the task we solve is only to transform the latent representation of images, unlike end-to-end approaches which also learn to synthesise high-quality images. When trained with 15 identities, our network only takes around 6 hours on a single RTX 8000 GPU to train. In contrast, the method of Sun et al. [2020] takes around 26 hours on 4 V100 GPUs for training at the same resolution.

4 RESULTS

We evaluate our technique both qualitatively and quantitatively on a large set of diverse images. The role of the different loss terms is studied in Sec. 4.2. We compare against several related techniques in Sec. 4.3 – the high-quality relighting approaches of Sun et al. [2019] and Zhou et al. [2019], as well as the recent StyleGAN-based image editing approaches of Tewari et al. [2020b] and Abdal et al. [2020] (the latter is concurrent to ours). Furthermore, we show that our method allows for learning from limited supervised training data by conducting extensive experiments in Sec. 4.4.

Data Preparation We evaluate our approach on portrait images captured in the wild [Karras et al. 2019; Shih et al. 2014]. All data in our work (including the training data) are cropped and preprocessed as described in Karras et al. [2019]. The images are resized to a resolution of 1024x1024. Since we need the ground truth images for quantitative evaluations, we use the test portion of our light-stage dataset composed of images of 41 identities unseen during training. We create two test sets, *Set1* has the input and ground truth pairs captured from the same viewpoint while *Set2* includes pairs captured from different viewpoints. The HDR environment maps, randomly sampled from the Naval Outdoor and Naval Indoor datasets [Gardner et al. 2017; Hold-Geoffroy et al. 2019] are used to synthesise the pairs with natural illumination conditions. Viewpoints are randomly sampled from the 8 cameras of the light-stage setup. The input and ground truth images are computed using the same environment map in *Set2* for evaluating the viewpoint editing

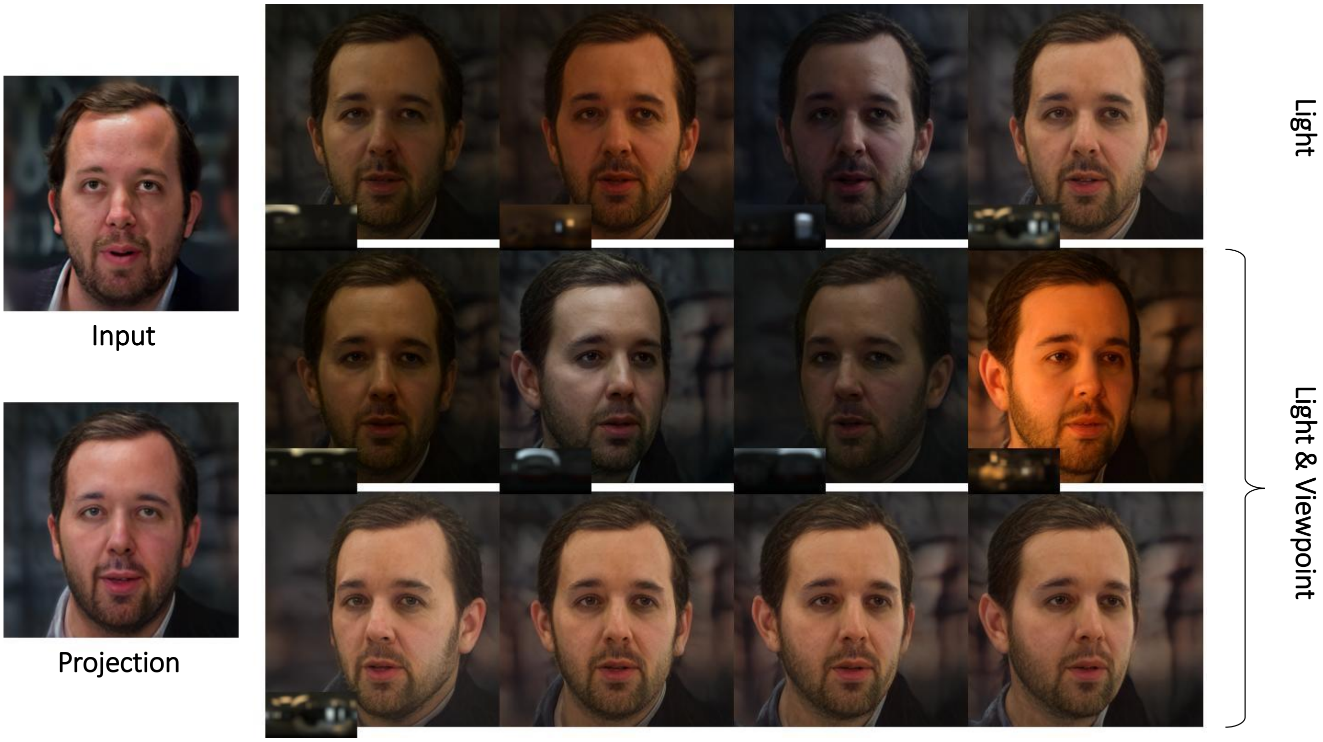

Fig. 5. Qualitative illumination and viewpoint editing results. In the first row, we show relighting results where the camera is fixed as in the input. The second column shows results where both illumination and camera pose is edited. The last row shows results with a moving camera under fixed scene illumination. Please note the local shading effects such as shadows, as well as view-dependent effects such as specularities in the image

results, while the pairs in *Set1* use different environment maps for relighting evaluations. *Set1* includes 883 and *Set2* include 792 image pairs after finding common set of images which works for all the methods. For each pair, we additionally provide a reference image, which is used by related methods to estimate the target illumination and pose in the representation they work with [Abdal et al. 2020; Sun et al. 2019; Tewari et al. 2020b; Zhou et al. 2019]. In *Set1*, the reference image is of an identity different from the input identity. The ground truth image is directly taken as the reference image for *Set2*, since there can be slight pose variations between different identities for the same camera.

4.1 High-Fidelty Appearance Editing

Figs. 4, 5, and 6 show simultaneous viewpoint and illumination editing results of our method for various subjects. We also show the StyleGAN projection of the input images estimated by Richardson et al. [2020]. Our approach produces high-quality photorealistic results and synthesises the full portrait, including hair, eyes, mouth, torso and the background, while preserving the identity, expression and other properties (such as facial hair). Additionally, the results show that our method can preserve a variety of reflectance properties, resulting in effects such as specularities and subsurface scattering. Please note the view-dependent effects such as specularities in the results(nose, forehead...). Our method can synthesise results even under high-frequency light conditions resulting in shadows, even though the StyleGAN network is trained on a dataset of natural images. In Figs. 5-6, we show more detailed editing results. As it can be noted, the relighting preserve the input pose and identity. Also, our method can change the viewpoint under a fixed environment map (third row for each subject).

4.2 Ablation Study

In this section, we evaluate the importance of the different loss terms of our objective function (Eq. 1). Results are shown in Fig. 7. The target illumination and viewpoint are visualised using a reference image (second column) with the same scene parameters. Removing the latent loss leads to clear distortions of the head geometry. Only using the perceptual loss leads to results with closed eye expressions, as our training data only consists of people captured with closed eyes. We found that the latent loss term helps in generalisation to unseen expressions. However, using only the latent loss is not sufficient for high-quality results. In such case, the facial identity and facial hair (see row 1) are not well preserved, and the relighting is not very accurate (see rows 1,2,6). A combination of both terms is essential for high-quality.

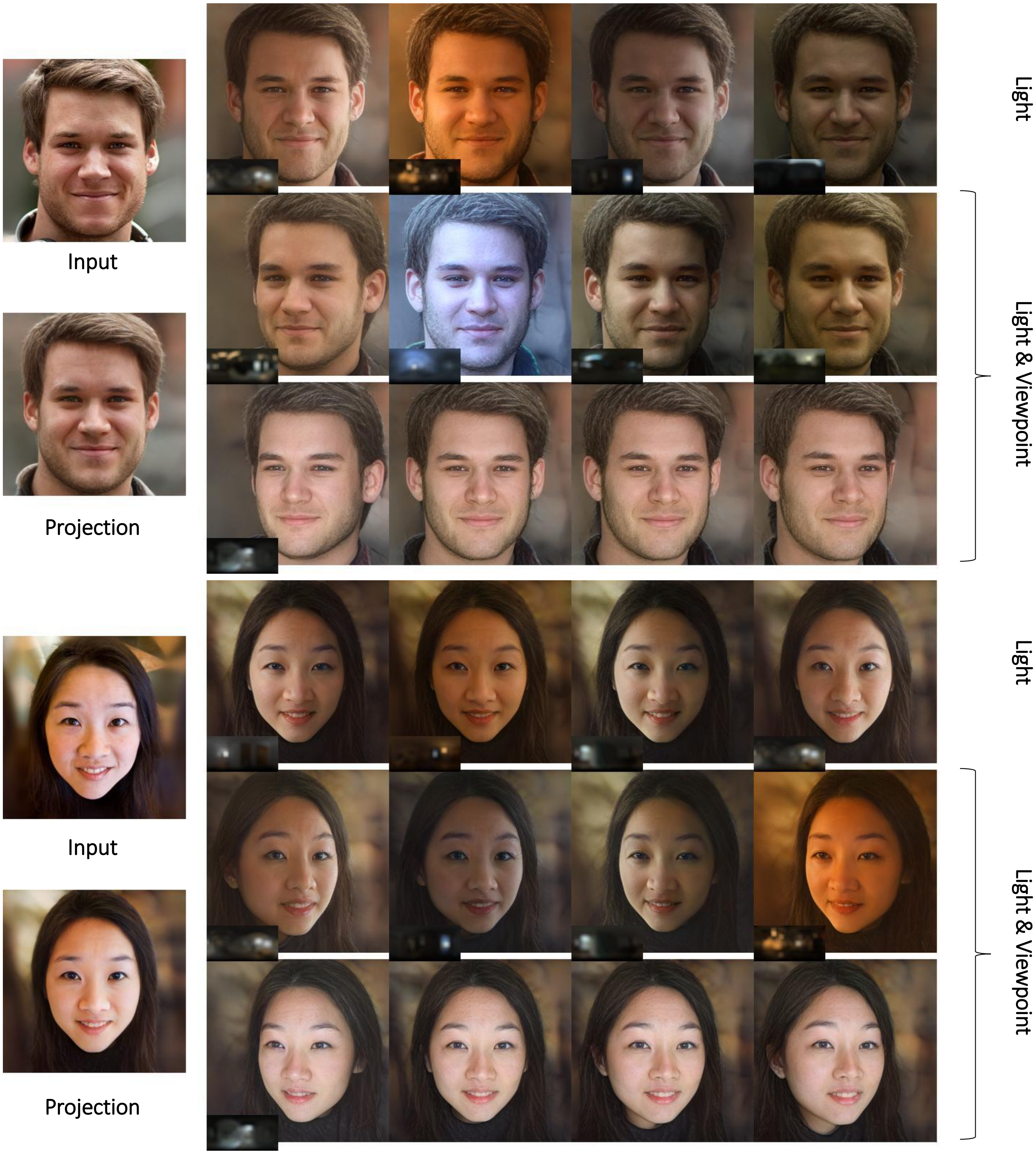

Fig. 6. Qualitative illumination and viewpoint editing results. In the first row, we show relighting results where the camera is fixed as in the input. The second column shows results where both illumination and camera pose is edited. The last row shows results with a moving camera under fixed scene illumination. Please note the local shading effects such as shadows, as well as view-dependent effects such as specularities in the image

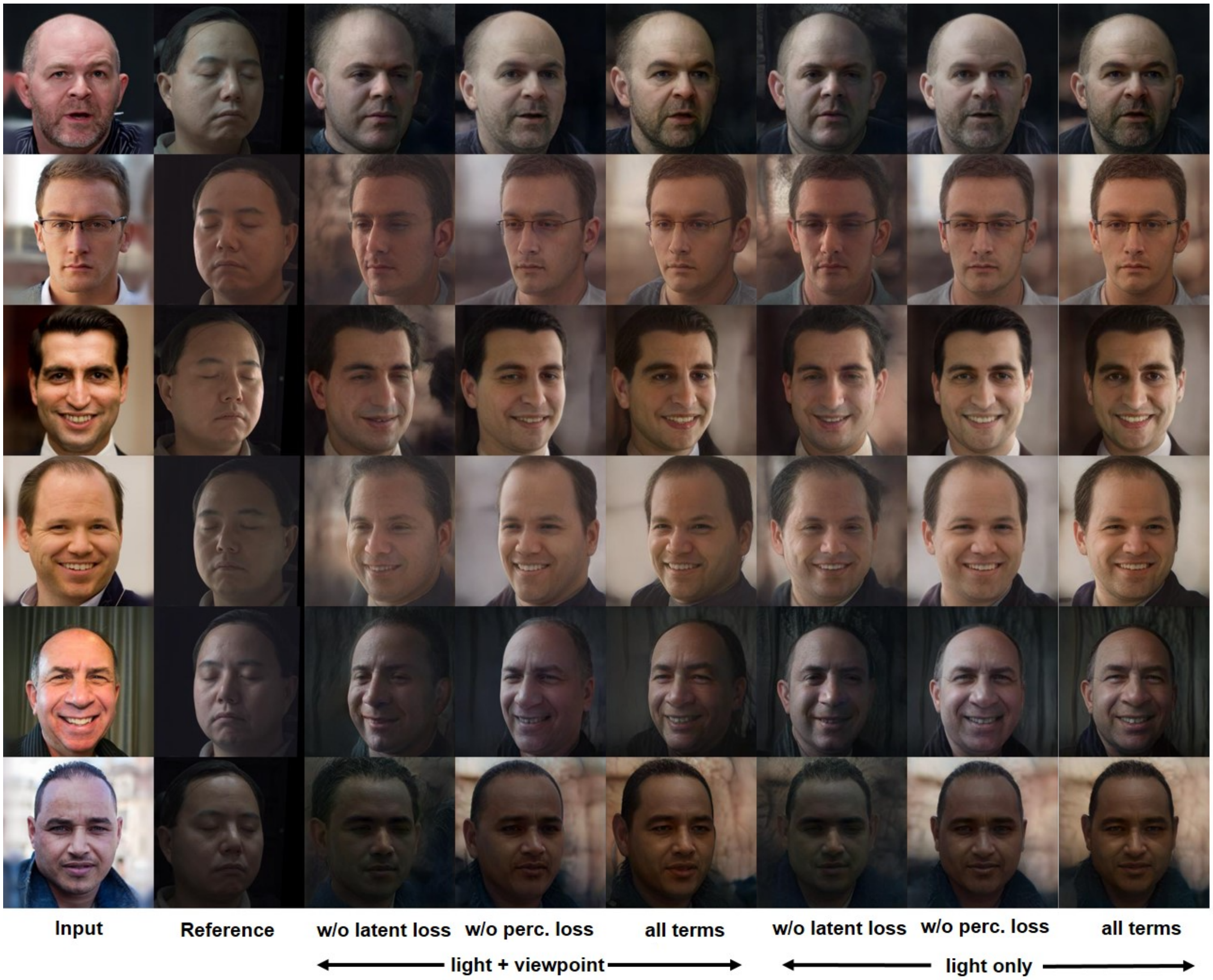

Fig. 7. Ablative study on the loss functions. The reference images visualise the target illumination and viewpoint. Removing the latent loss results in distortion of the head geometry and lower quality results. Removing the perceptual term leads to a loss of facial hair and identity preservation such as beards (for e.g., row 1, row 4,5 in light+viewpoint). It also often produces lower-quality results (e.g. row 1,2,6). Both terms are necessary for high-quality results.

4.3 Comparisons to Related Methods

We compare our method with several state of the art portrait editing approaches. We evaluate qualitatively on in the wild data, as well as quantitatively on the test set of the light-stage data. We compare with the following approaches:

- The relighting approach of Sun et al. [2019] which is a data-driven technique trained on a light-stage dataset. It can only edit the scene illumination.
- The relighting approach of Zhou et al. [2019] which is trained on synthetic data. It can also only edit the scene illumination.
- PIE [Tewari et al. 2020b] is a method which computes a StyleGAN embedding used to edit the image. It can edit the head pose and scene illumination sequentially (unlike ours, which can perform the edits simultaneously). It is trained without supervised image pairs.
- StyleFlow [Abdal et al. 2020], like PIE can edit images by projecting them onto the StyleGAN latent space. It is also trained without supervised image pairs. Please note that this paper is concurrent to us (not counted as prior art). However, we provide comparisons for completeness.

We show the relighting comparisons on in the wild data in Fig. 8. Here, the reference image in the second column is used to visualise the target illumination. Both the light-stage data-driven approach of Sun et al. [2019] and the synthetic data-driven approach of Zhou

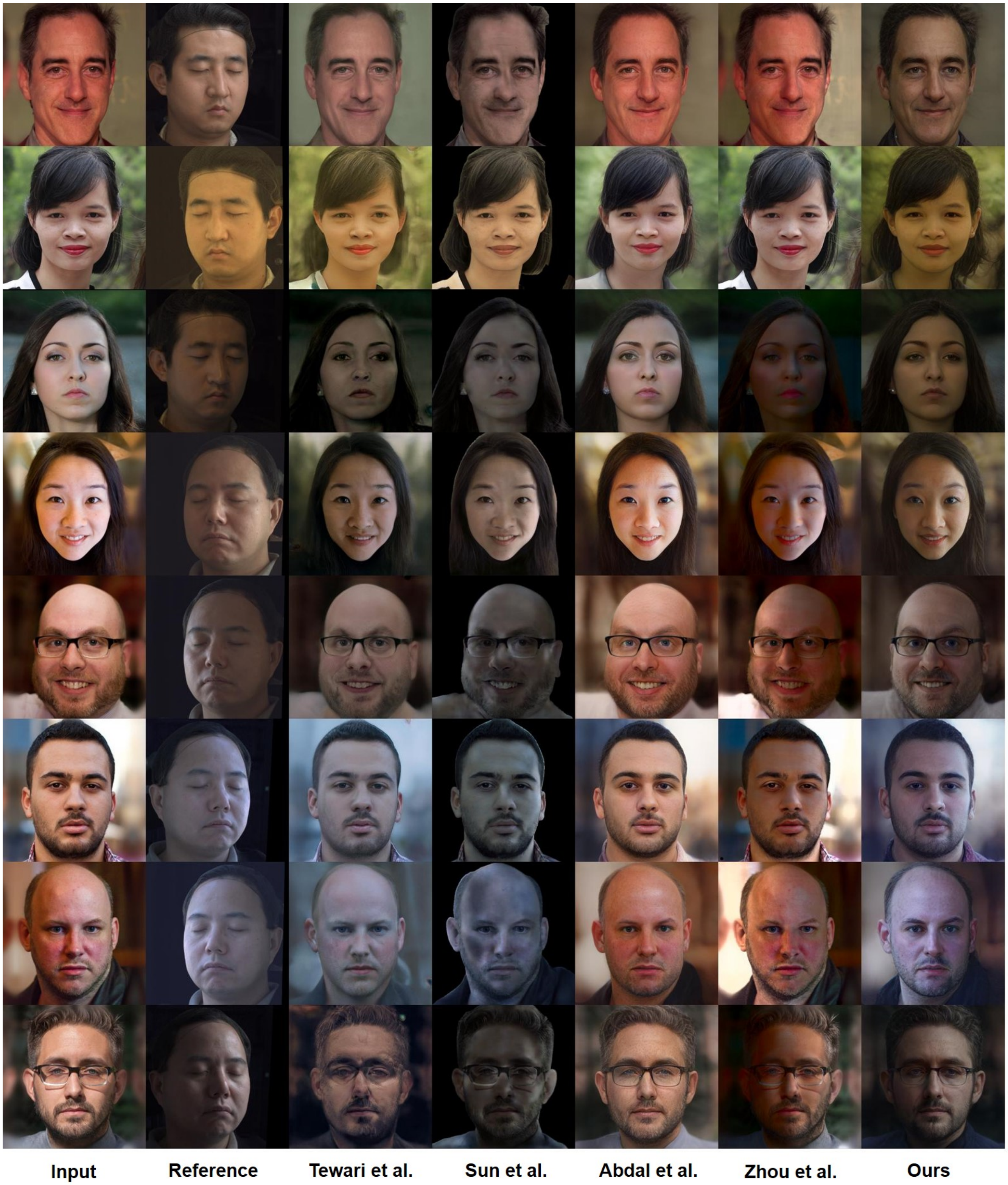

Fig. 8. Relighting comparisons. Target illumination is visualised using reference images. Our approach clearly outperforms all existing approaches.

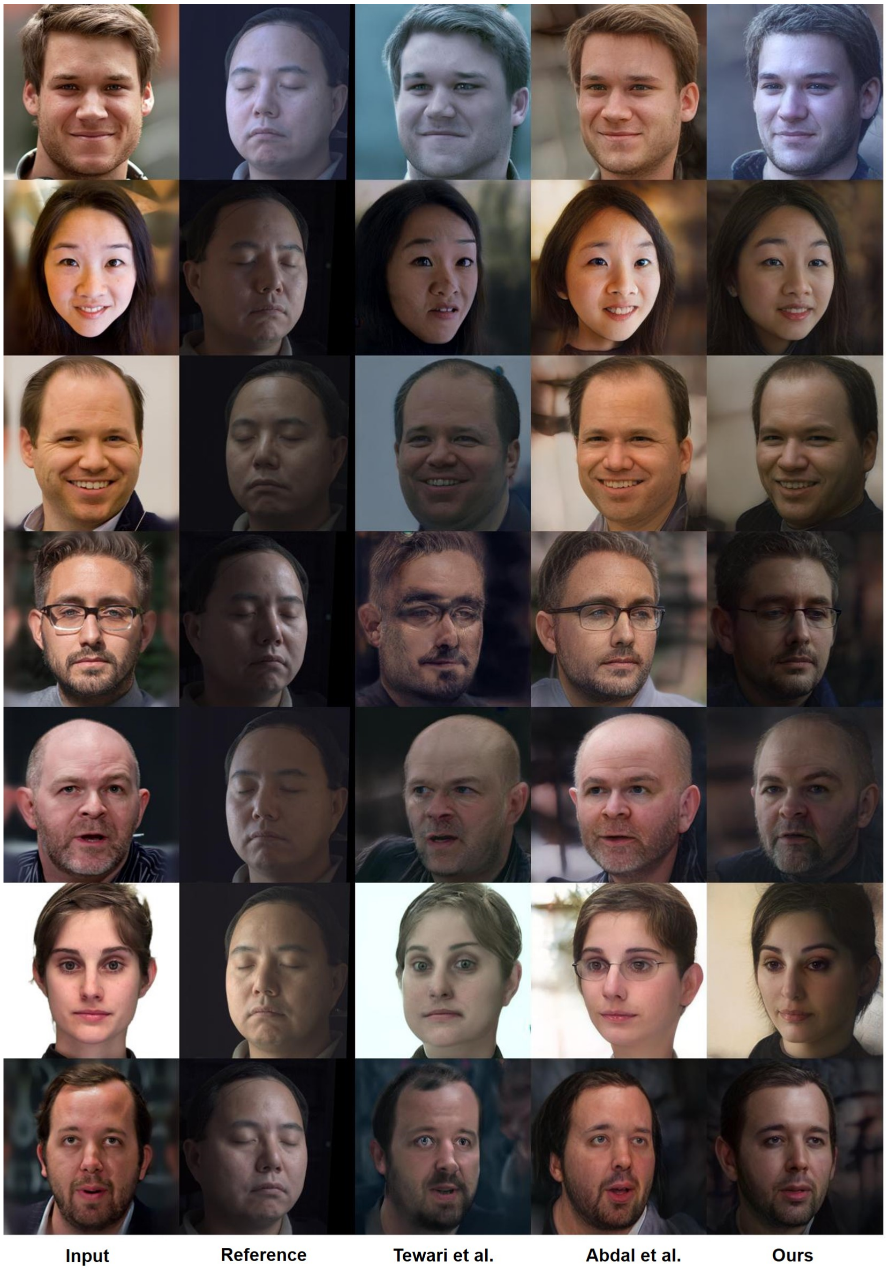

Fig. 9. Comparisons to PIE [Tewari et al. 2020b] and StyleFlow [Abdal et al. 2020]. The reference images visualise the target illumination and viewpoint. Our approach produces higher-quality results and clearly outperforms these methods.

et al. [2019] produce noticeable artifacts. The approach of Zhou et al. [2019] only uses single channel illumination as input and can thus not capture the overall color tone of the illumination. The StyleGAN-based approach of Abdal et al. [2020] produces less artifacts, however the quality of relighting is worse than other approaches as it mostly preserves the input lighting. In addition, similar to Zhou et al. [2019], this approach cannot capture the color tone of the environment map. PIE [Tewari et al. 2020b] produces better results but it does not capture local illumination effects like our approach (for eg., rows 5,6,7,8) and can produce significant artifacts in some cases (for eg., row 8). Our approach clearly outperforms all existing methods, demonstrating the effectiveness of a combination of supervised learning and generative modeling. It can capture the global color tone as well as local effects such as shadows and specularities. It can synthesise the image under harsh lighting (for e.g., rows 7,8)

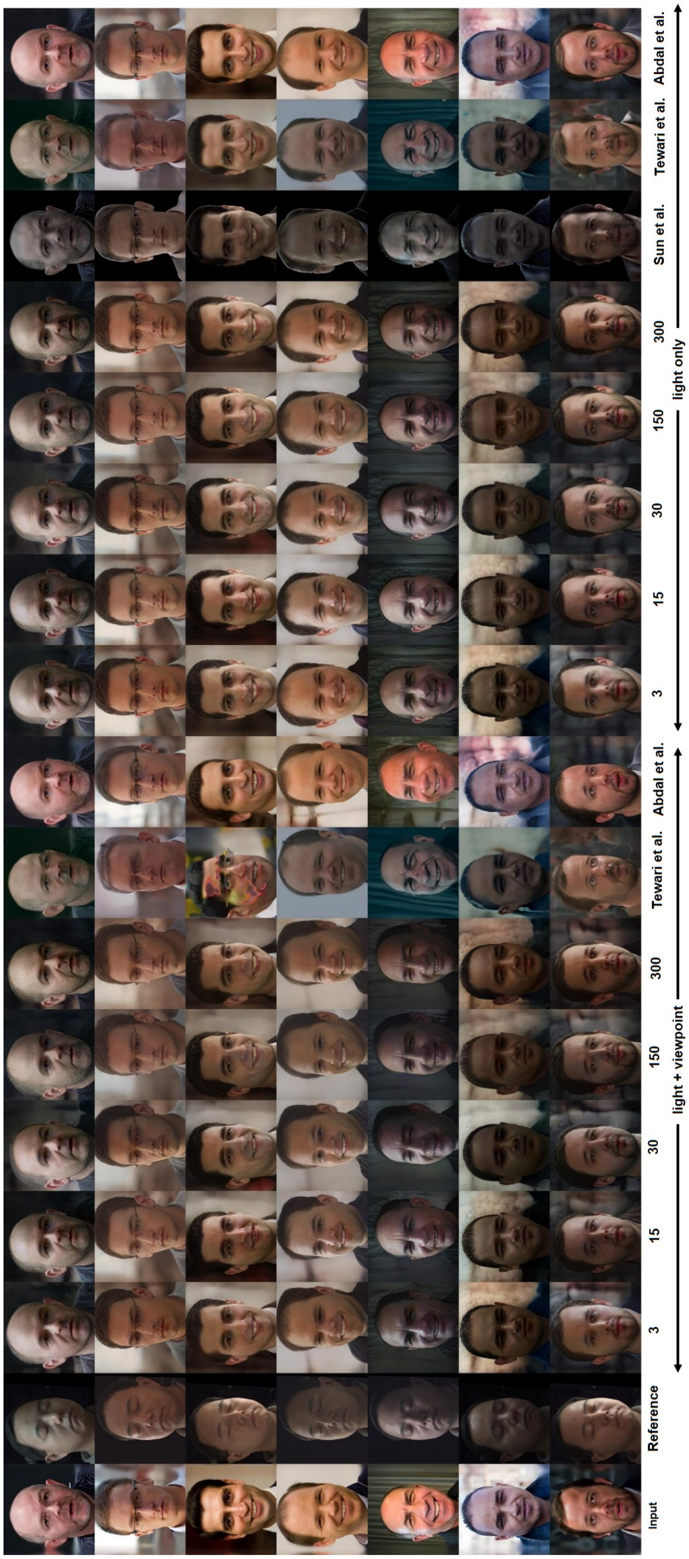

Fig. 10. Our method allows for training with very limited supervision. We show editing results when trained with 3,15,30,150 and 300 identities. Our approach produces photorealistic results, and outperforms existing methods even with limited training data.

and remove source-lighting related specularities on the glasses (for eg., row 5).

Tab. 1 shows the quantitative comparisons with these methods on the light-stage test set (*Set1*). We use the Scale invariant-MSE (Si-MSE) [Zhou et al. 2019] and SSIM [Zhou Wang et al. 2004] metrics. The Si-MSE metric does not penalize global scale offsets between the ground truth and results. Our method outperforms all methods using this metric. The method of Sun et al. [2019] outperforms other methods on SSIM. Since this method uses a U-Net architecture, it is easier to copy the details from the input image, and maintain the pixel correspondences. However, visual results clearly show that our approach outperforms all related methods, including that of Sun et al. [2019] (see Fig. 8).

Fig. 9 shows joint editing of the camera viewpoint and scene illumination for in the wild images. The target viewpoint and illumination are visualised using reference images (see second column). While PIE [Tewari et al. 2020b] can change the viewpoint, it often distorts the face in an unnatural way (e.g. row 1,2,7). It also does not capture local shading effects correctly (e.g. row 1,2,6) and can produce strong artifacts (e.g. row 4). PIE solves an optimisation problem to obtain the embedding for each image, which is slow, taking about 10 mins per image. In contrast, our method is interactive, only requiring 160ms to compute the embedding and edit it. StyleFlow [Abdal et al. 2020] can preserve the identity better than PIE, but results in less photorealistic results compared to our method. In addition, the relighting results of StyleFlow often fail to capture the input environment map. Our approach clearly outperforms both methods both in terms of photorealism as well as the quality of editing.

Tab. 2 quantitatively compares the joint editing of camera viewpoint and scene illumination. We use the Si-MSE and SSIM metrics and evaluate on the *Set2* of the light-stage test data. Our approach outperforms all methods here in both metrics.

4.4 Generalisation with Limited Supervision

The combination of generative modeling and supervised learning allows us to train from very limited supervised data. We show results of training with different number of identities in Fig. 10. Results of PIE [Tewari et al. 2020b], StyleFlow [Abdal et al. 2020] and Sun et al. [2020] are also demonstrated. Our relighting results outperform related methods both in terms of realism as well as quality of editing, even when trained with as little as 3 identities. We also consistently outperform PIE and StyleFlow when editing both viewpoint and illumination, even when trained with 30 identities. More identities during training help with better preservation of the facial identity during viewpoint editing. However, very small training data is sufficient for relighting.

We also quantitatively evaluate these results in Tables 1 and 2. In both tables, our method outperforms all related approaches using the Si-MSE metric, even when trained with just 3 identities. All versions of our approach perform similar in terms of SSIM. These evaluations show that while larger datasets lead to better results, only limited supervised data is required to outperform the state of the art. Finally, despite the limited expressivity of the training dataset (subjects in a single expression with eyes and mouth closed), our method is able to generalise to different expressions, as shown in our results (mouth and eyes open, smiling, etc.) (see Fig. 3).

Table 1. Quantitative comparison with relighting methods. Our approach achieves the lowest Si-MSE numbers. While Sun et al. [2019] achieves the highest SSIM score, qualitative results show that our method significantly outperforms all existing techniques on in the wild images (see Fig. 8).

	Si-MSE ↓	SSIM ↑
[Zhou et al. 2019]	0.0037	0.9197
	(σ= 0.0031)	(σ= 0.0744)
[Sun et al. 2019]	0.0026	**0.9591**
	(σ= 0.0024)	(σ= 0.0237)
[Tewari et al. 2020b]	0.0051	0.922
	(σ= 0.0036)	(σ= 0.029)
[Abdal et al. 2020]	0.0082	0.8909
	(σ=0.0056)	(σ= 0.04)
Ours	**0.002**	0.9199
	(σ=0.001)	(σ= 0.0297)
Ours	**0.002**	0.9192
(150 id.)	(σ=0.001)	(σ= 0.0351)
Ours	**0.002**	0.9188
(30 id.)	(σ=0.001)	(σ= 0.0300)
Ours	**0.002**	0.9191
(15 id.)	(σ=0.001)	(σ= 0.0306)
Ours	**0.002**	0.9193
(3 id.)	(σ=0.001)	(σ= 0.0293)

Table 2. Quantitative evaluation with illumination and pose editing methods using *Set2* of the light-stage test set. Our approach outperforms both competing methods, also clearly illustrated using qualitative results (see Fig. 9).

	Si-MSE ↓	SSIM ↑
[Tewari et al. 2020b]	0.0067	0.9005
	(σ= 0.0044)	(σ= 0.0363)
[Abdal et al. 2020]	0.0104	0.8812
	(σ=0.0071)	(σ= 0.0469)
Ours	**0.0039**	**0.9086**
	(σ=0.0029)	(σ=0.0307)
Ours	**0.0035**	0.9050
(150 id.)	(σ=0.0020)	(σ=0.0340)
Ours	0.0040	0.9021
(30 id.)	(σ=0.0033)	(σ=0.0312)
Ours	0.0046	0.8974
(15 id.)	(σ=0.0027)	(σ=0.0315)
Ours	0.0048	0.9000
(3 id.)	(σ=0.0031)	(σ=0.0316)

4.5 Supplemental Material

In the supplemental video, we show results on videos processed on a per-frame basis. We can synthesise the input video from different

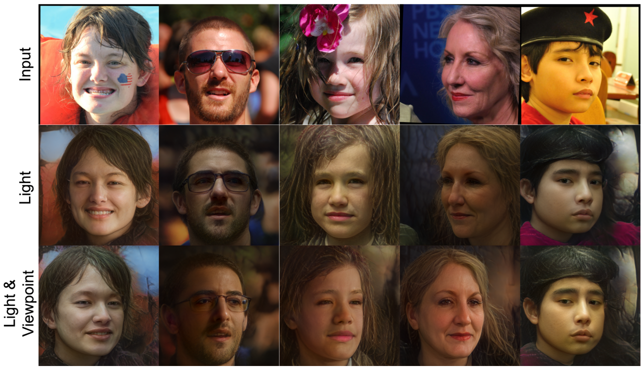

Fig. 11. Our method struggles in the presence of accessories such as caps and sunglasses, and background clutter. Extreme pose and illumination editing is also difficult for our method.

camera poses and under different scene illumination while preserving the expressions in the video. We also show additional results on a large number of images in the supplemental material.

5 LIMITATIONS

While we demonstrate high-quality results of our approach, several limitations exist, see Fig. 11. Our method can fail to preserve accessories such as caps and glasses in some cases. Background clutter can lead to a degradation of quality in the results. Our method struggles to preserve the facial identity under large edits for both camera pose and illumination. While we can preserve the input pose in the results, our method cannot edit the camera viewpoint without changing the illumination. In the future, we can use methods that estimate illumination from portrait images [LeGendre et al. 2020] to preserve input illumination when editing the viewpoint. While we show several high-quality results on video sequences, slight flicker and instability remain. A temporal treatment of videos could lead to smoother results.

6 CONCLUSION

We presented PhotoApp, a method for editing the scene illumination and camera pose in head portraits. Our method exploits the advantages of both supervised learning and generative adversarial modeling. By designing a supervised learning problem in the latent space of StyleGAN, we achieve high-quality editing results which generalise to in the wild images with significantly more diversity than the training data. Through extensive evaluations, we demonstrated that our method outperforms all related techniques, both in terms of realism and editing accuracy. We further demonstrated that our method can learn from very limited supervised data, achieving high-quality results when trained with as little as 3 identities captured in a single expression. While several limitations still exist, we hope that our contributions inspire future work on using generative representations for synthesis applications.

Acknowledgements. This work was supported by the ERC Consolidator Grant 4DReply (770784). We also acknowledge support from Technicolor and Interdigital.

REFERENCES

Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. 2020. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *arXiv e-prints* (2020), arXiv–2008.

Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F. Cohen. 2017. Bringing Portraits to Life. *ACM Trans. on Graph. (Proceedings of SIGGRAPH Asia)* 36, 6 (2017).

Mallikarjun B R, Ayush Tewari, Tae-Hyun Oh, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, and Christian Theobalt. 2020. Monocular Reconstruction of Neural Face Reflectance Fields. arXiv:2008.10247

Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. 2020. Editing in style: Uncovering the local semantics of gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 5771–5780.

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *Annual conference on Computer graphics and interactive techniques.*

Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to predict indoor illumination from a single image. *ACM Trans. on Graph. (Proceedings of SIGGRAPH Asia)* 36, 6, Article 176 (2017).

Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. 2018. Warp-Guided GANs for Single-Photo Facial Animation. *ACM Trans. on Graph. (Proceedings of SIGGRAPH Asia)* 37, 6 (2018).

Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. 2011. Multiview Face Capture Using Polarized Spherical Gradient Illumination. *ACM Trans. on Graph.* 30, 6 (Dec. 2011), 1–10.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546* (2020).

Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. 2019. Deep sky modeling for single image outdoor lighting estimation. In *Computer Vision and Pattern Recognition (CVPR).*

Longlong Jing and Yingli Tian. 2020. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

James T. Kajiya. 1986. The Rendering Equation. *SIGGRAPH Computer Graphics* 20, 4 (1986), 143–150. https://doi.org/10.1145/15886.15902

T. Karras, S. Laine, and T. Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Computer Vision and Pattern Recognition (CVPR).*

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Computer Vision and Pattern Recognition (CVPR).*

Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollöfer, and Christian Theobalt. 2018. Deep Video Portraits. *ACM Trans. on Graph. (Proceedings of SIGGRAPH)* 37, 4 (2018), 163.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.

Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. 2020. AvatarMe: Realistically Renderable 3D Facial Reconstruction "in-the-wild". In *Computer Vision and Pattern Recognition (CVPR).*

Chloe LeGendre, Wan-Chun Ma, Rohit Pandey, Sean Fanello, Christoph Rhemann, Jason Dourgarian, Jay Busch, and Paul Debevec. 2020. Learning Illumination from Diverse Portraits. In *SIGGRAPH Asia 2020 Technical Communications.* 1–4.

Abhimitra Meka, Christian Häne, Rohit Pandey, Michael Zollhöfer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, Peter Denny, Sofien Bouaziz, Peter Lincoln, Matt Whalen, Geoff Harvey, Jonathan Taylor, Shahram Izadi, Andrea Tagliasacchi, Paul Debevec, Christian Theobalt, Julien Valentin, and Christoph Rhemann. 2019. Deep Reflectance Fields: High-Quality Facial Reflectance Field Inference from Color Gradient Illumination. *ACM Trans. on Graph. (Proceedings of SIGGRAPH)* 38, 4 (2019).

Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. 2018. paGAN: real-time avatars using dynamic textures. In *ACM Trans. on Graph. (Proceedings of SIGGRAPH Asia).* 258.

Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas M Lehrmann. 2020. Learning Physics-guided Face Relighting under Directional Light. In *Computer Vision and Pattern Recognition (CVPR).*

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2020. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. *arXiv preprint arXiv:2008.00951* (2020).

Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. 2018. SfSNet: Learning Shape, Refectance and Illuminance of Faces in the Wild. In *Computer Vision and Pattern Regognition (CVPR).*

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the Latent Space of GANs for Semantic Face Editing. In *CVPR*.

YiChang Shih, Sylvain Paris, Connelly Barnes, William T. Freeman, and Frédo Durand. 2014. Style Transfer for Headshot Portraits. *ACM Trans. on Graph. (Proceedings of SIGGRAPH)* 33, 4, Article 148 (2014), 14 pages.

Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. 2017. Neural Face Editing with Intrinsic Image Disentangling. In *Computer Vision and Pattern Recognition (CVPR)*. 5444–5453.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. 2019. Single Image Portrait Relighting. 38, 4, Article 79 (July 2019).

Tiancheng Sun, Zexiang Xu, Xiuming Zhang, Sean Fanello, Christoph Rhemann, Paul Debevec, Yun-Ta Tsai, Jonathan T. Barron, and Ravi Ramamoorthi. 2020. Light Stage Super-Resolution: Continuous High-Frequency Relighting. In *ACM Trans. on Graph. (Proceedings of SIGGRAPH Asia)*.

Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. 2020a. StyleRig: Rigging StyleGAN for 3D Control over Portrait Images, CVPR 2020. In *Computer Vision and Pattern Recognition (CVPR)*.

Ayush Tewari, Mohamed Elgharib, Mallikarjun BR, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. 2020b. PIE: Portrait Image Embedding for Semantic Control. *ACM Trans. on Graph. (Proceedings SIGGRAPH Asia)* 39, 6.

Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. 2020c. State of the art on neural rendering. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 701–727.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–12.

Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. 2020. Single Image Portrait Relighting via Explicit Multiple Reflectance Channel Modeling. *ACM Trans. on Graph. (Proceedings of SIGGRAPH Asia)* 39, 6, Article 220 (2020).

Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, and Markus Gross. 2006. Analysis of Human Faces using a Measurement-Based Skin Reflectance Model. *ACM Trans. on Graphics (Proceedings of SIGGRAPH)* 25, 3 (2006), 1013–1024.

Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. 2018. X2Face: A network for controlling face generation using images, audio, and pose codes. In *European Conference on Computer Vision (ECCV)*.

Shuco Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. 2018. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Trans. on Graph. (Proceedings of SIGGRAPH)* 37, 4, Article 162 (2018).

Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. 2019. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1087–1096.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Xuaner Zhang, Jonathan T. Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E. Jacobs. 2020. Portrait Shadow Manipulation. *ACM Transactions on Graphics (TOG)* 39, 4.

Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. 2019. Deep Single-Image Portrait Relighting. In *International Conference on Computer Vision (ICCV)*.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.