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 generalization outside the dataset. Portrait images are from Shih et al. [2014] and environment maps are from [Gardner et al. 2017; Hold-Geoffroy et al. 2019].

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.

CCS Concepts:
- Computing methodologies → Reflectance modeling; Image representations; Image manipulation.

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

ACM Reference Format:

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 modelled 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 synthesise 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 multi-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 semantic 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 photorealism. Others infer more general surface reflectance, with fewer assumptions about incident illumination [BR 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 input 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], or 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.

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 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.

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, synthesized 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 generalization 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 Laval Outdoor [Hold-Geoffroy et al. 2019] and 2233 portrait images [Debevec et al. 2000]. We use 205 HDR environment maps from the Laval Outdoor [Hold-Geoffroy et al. 2019] and 2233 portrait images [Debevec et al. 2000]. We use 205 HDR environment maps from the Laval Outdoor [Hold-Geoffroy et al. 2019] and 2233 portrait images [Debevec et al. 2000].

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 \( \omega_t \), and a binary input \( p \). The value of \( p \) is set to 0 when the target pose is the same as that of the input, and 1 when they are different. This conditioning 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 \( 18 \times 512 \) 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.
\[
\mathcal{L}(I_s, L_s, \hat{I}_t, \hat{L}_t, \theta_h) = \mathcal{L}_1(L_s, L_t, \theta_h) + \mathcal{L}_p(I_s, \hat{I}_t, \theta_h) .
\]
(1)

Here, \( \theta_h \) denotes the network parameters of PhotoAppNet. Both terms are weighted equally. The first term is a StyleGAN latent loss defined as
\[
\mathcal{L}_1(L_s, L_t, \theta_h) = \| L_s - L_t \|_2^2 .
\]

It enforces the StyleGAN latent code of the output image \( L_t \) to be close to the ground truth latent code \( L_t \). The second term is a perceptual loss defined as
\[
\mathcal{L}_p(I_s, \hat{I}_t, \theta_h) = \| \phi(I_s) - \phi(\hat{I}_t) \|_2^2 .
\]

Here, we employ the learned perceptual similarity metric LPIPS [Zhang et al. 2018]. An \( \ell_2 \) 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 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 results, while the pairs in Set1 use different environment maps for relighting evaluations. Set1 includes 883 and Set2 includes 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-Fidelity 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). Our method works well on people of different races. 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 cases, 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. 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. Portrait images are from Shih et al. [2014] (first and third row) and from Livingstone and Russo [2018] (second and fourth row). Environment maps are from [Gardner et al. 2017; Hold-Geoffroy et al. 2019].

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 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. 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 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
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 row 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. Portrait images are from Shih et al. [2014] (first part) and Karras et al. [2019] (second part). Environment maps are from [Gardner et al. 2017; Hold-Geoffroy et al. 2019]
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 row 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. Portrait images are from Shih et al. [2014] and environment maps are from [Gardner et al. 2017; Hold-Geoffroy et al. 2019].
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. Images are from Shih et al. [2014] (first column) and Weyrich et al. [2006] (second column).

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...
Fig. 8. Relighting comparisons. Target illumination is visualised using reference images. Our approach clearly outperforms all existing approaches. Here, we compare our method with approaches of Tewari et al. [2020b], Sun et al. [2019], Abdal et al. [2020], Zhou et al. [2019]. Images are from Karras et al. [2019] (first column, row 1,5,8), Shih et al. [2014] (first column, row 2,3,4,6,7) and Weyrich et al. [2006] (second column).
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.
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. Here, we also compare with approaches of Sun et al. [2019], Tewari et al. [2020b] and Abdal et al. [2020]. Images are from Shih et al. [2014] (first column) and Weyrich et al. [2006] (second column).
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).

4.5 Preserving the Input Illumination

Our method can be easily extended for editing the viewpoint while preserving the input illumination, see Fig. 12. Here, we modify the network architecture in Fig. 2 by providing another binary input similar to \( p \), which is set to 0 when the target illumination is same as the input illumination, and 1 when they are different. This design helps in editing both viewpoint and illumination in isolation.

4.6 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

---

**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 the wild images (see Fig. 8).

<table>
<thead>
<tr>
<th>Method</th>
<th>Si-MSE ↓</th>
<th>SSIM ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Zhou et al. 2019]</td>
<td>0.0037</td>
<td>0.9197</td>
</tr>
<tr>
<td>[Sun et al. 2019]</td>
<td>0.0026</td>
<td>0.9591</td>
</tr>
<tr>
<td>[Tewari et al. 2020b]</td>
<td>0.0051</td>
<td>0.922</td>
</tr>
<tr>
<td>[Abdal et al. 2020]</td>
<td>0.0082</td>
<td>0.8909</td>
</tr>
<tr>
<td>Ours (3 id.)</td>
<td>0.002</td>
<td>0.9199</td>
</tr>
<tr>
<td>Ours (15 id.)</td>
<td>0.002</td>
<td>0.9192</td>
</tr>
</tbody>
</table>

**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).

<table>
<thead>
<tr>
<th>Method</th>
<th>Si-MSE ↓</th>
<th>SSIM ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Tewari et al. 2020b]</td>
<td>0.0067</td>
<td>0.9005</td>
</tr>
<tr>
<td>[Abdal et al. 2020]</td>
<td>0.0104</td>
<td>0.8812</td>
</tr>
<tr>
<td>Ours (15 id.)</td>
<td>0.0035</td>
<td>0.9050</td>
</tr>
<tr>
<td>Ours (30 id.)</td>
<td>0.0040</td>
<td>0.9021</td>
</tr>
<tr>
<td>Ours (15 id.)</td>
<td>0.0046</td>
<td>0.8974</td>
</tr>
<tr>
<td>Ours (3 id.)</td>
<td>0.0048</td>
<td>0.9000</td>
</tr>
</tbody>
</table>

ACM Trans. Graph., Vol. 40, No. 4, Article 44. Publication date: August 2021.
While we demonstrate high-quality results of our approach, several limitations remain. A temporal treatment of videos could lead to a degradation of quality in the results. 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. Images are from Karras et al. [2019].

5 LIMITATIONS

While we demonstrate high-quality results of our approach, several limitations exist, see Fig. 13. 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 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.

ACKNOWLEDGMENTS

This work was supported by the ERC Consolidator Grant 4DReply (770784). We also acknowledge support from Technicolor and InterDigital. We thank Tiancheng Sun for kindly helping us with the comparisons with Sun et al. [2019].

REFERENCES


