# PIE: Portrait Image Embedding for Semantic Control

AYUSH TEWARI, MOHAMED ELGHARIB, and MALLIKARJUN B R, Max Planck Institute for Informatics, SIC
FLORIAN BERNARD, Max Planck Institute for Informatics, SIC and Technical University of Munich
HANS-PETER SEIDEL, Max Planck Institute for Informatics, SIC
PATRICK PÉREZ, Valeo.ai
MICHAEL ZOLLHÖFER, Stanford University
CHRISTIAN THEOBALT, Max Planck Institute for Informatics, SIC

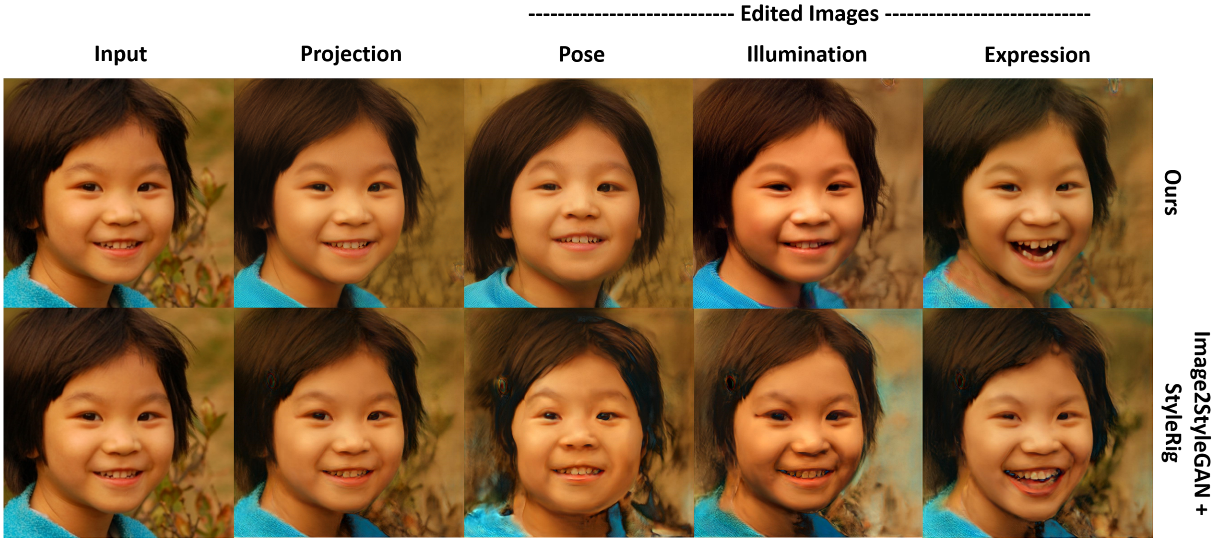

Fig. 1. We present an approach for embedding portrait images in the latent space of StyleGAN [Karras et al. 2019b] (visualized as "Projection") which allows for intuitive photo-real semantic editing of the head pose, facial expression, and scene illumination using StyleRig [Tewari et al. 2020]. Our optimization-based approach allows us to achieve higher quality editing results compared to the existing embedding method Image2StyleGAN [Abdal et al. 2019]. Image from Shen et al. [2016].

Editing of portrait images is a very popular and important research topic with a large variety of applications. For ease of use, control should be provided via a semantically meaningful parameterization that is akin to computer animation controls. The vast majority of existing techniques do not provide such intuitive and fine-grained control, or only enable coarse editing of a single isolated control parameter. Very recently, high-quality semantically controlled editing has been demonstrated, however only on synthetically created StyleGAN images. We present the first approach for embedding real portrait images in the latent space of StyleGAN, which allows for intuitive editing of the head pose, facial expression, and scene illumination in the image. Semantic editing in parameter space is achieved based on StyleRig, a pretrained neural network that maps the control space of a 3D morphable face model to the latent space of the GAN. We design a novel hierarchical non-linear optimization problem to obtain the embedding. An identity preservation energy term allows spatially coherent edits while maintaining facial integrity. Our approach runs at interactive frame rates and thus allows the user to explore the space of possible edits. We evaluate our approach on a wide set of portrait photos, compare it to the current state of the art, and validate the effectiveness of its components in an ablation study.

CCS Concepts: • **Computing methodologies** → **Image manipulation**.

Additional Key Words and Phrases: Portrait Editing, StyleGAN, StyleRig

**ACM Reference Format:**
Ayush Tewari, Mohamed Elgharib, Mallikarjun B R, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. 2020. PIE: Portrait Image Embedding for Semantic Control. *ACM Trans. Graph.* 39, 6, Article 223 (December 2020), 14 pages. https://doi.org/10.1145/3414685.3417803

## 1 INTRODUCTION

Portrait images, showing mainly the face and upper body of people, are among the most common and important photographic depictions. We look at them to emotionally connect with friends and family, we use them to best present ourselves in job applications and on social media, they remind us of memorable events with friends,

0730-0301/2020/12-ART223
https://doi.org/10.1145/3414685.3417803

and photographs of faces are omnipresent in advertising. Nowadays, tools to computationally edit and post-process photographs are widely available and heavily used. Professional and hobby photographers use them to bring out the best of portrait and social media photos, as well as of professional imagery used in advertising. Photos are often post-processed with the purpose to change the mood and lighting, to create a specific artistic look and feel, or to correct image defects or composition errors that only become apparent after the photo has been taken. Today, commercial software[1] or recent research software [Gatys et al. 2016; Luan et al. 2017] offers a variety of ways to edit the color or tonal characteristics of photos. Some tools even enable the change of visual style of photos to match certain color schemes [Luan et al. 2017; Shih et al. 2014], or to match a desired painterly and non-photo-realistic style [Gatys et al. 2016; Selim et al. 2016]. In many cases, however, edits to a portrait are needed that require more complex and high-level modifications e.g. modifying head posture, smile or scene illumination after the capture. Enabling such edits from a single photograph is an extremely challenging and underconstrained problem. This is because editing methods need to compute reliable estimates of 3D geometry of the person and lighting in the scene. Moreover, they need to photo-realistically synthesize modified images of the person and background in a perspectively correct parallax-respecting manner, while inpainting disoccluding regions.

For ease of use, editing methods should use semantically meaningful parameterizations, which for the rest of the paper means the following: Head pose, face expression and scene lighting should be expressed as clearly disentangled and intuitive variables akin to computer animation controls, such as coordinates and angles, blendshape weights, or environment map parameterizations. Existing methods to edit human portrait imagery at best achieve parts of these goals. Some model-based methods to realistically edit human expression [Thies et al. 2019, 2016] and head pose [Kim et al. 2018] fundamentally require video input and do not work on single images. Other editing approaches are image-based and cannot be controlled by intuitive parametric controls [Averbuch-Elor et al. 2017; Geng et al. 2018; Siarohin et al. 2019; Wang et al. 2019a; Zakharov et al. 2019], only enable editing of a single semantic parameter dimension, e.g., scene illumination [Meka et al. 2019; Sun et al. 2019; Zhou et al. 2019], or do not photo-realistically synthesize some important features such as hair [Nagano et al. 2018].

Recently, generative adversarial neural networks, such as StyleGAN [Karras et al. 2019b], were trained on community face image collections to learn a manifold of face images. They can be sampled to generate impressive photo-realistic face portraits, even of people not existing in reality. However, their learned parameterization entangles important face attributes (most notably identity, head pose, facial expression, and illumination), which thus cannot be independently and meaningfully controlled in the output. It therefore merely allows control on a coarse style-based level, e.g., to adapt or transfer face styles on certain frequency levels between images. To overcome this limitation, StyleRig [Tewari et al. 2020] describes a neural network that maps the parameters of a 3D morphable face model (3DMM) [Blanz and Vetter 1999] to a pretrained StyleGAN for face images. However, while their results show disentangled control of face images synthesized by a GAN, they do not allow for editing real portrait photos.

On the other hand, some approaches have tried to embed real images in the StyleGAN latent space. Abdal et al. [2019, 2020a] demonstrate high-quality embedding results, which are used to perform edits such as style or expression transfer between two images, latent space interpolation for morphing, or image inpainting. However, when these embeddings are used to edit the input images using StyleRig [Tewari et al. 2020], the visual quality is not preserved and the results often have artifacts. High-quality parametric control of expression, pose or illumination on real images has not yet been shown to be feasible.

We therefore present the first method for embedding real portrait images in the StyleGAN latent space which allows for photo-realistic editing that combines all the following features: It enables photo-real semantic editing of all these properties — head pose, facial expression, and scene illumination, given only a single in-the-wild portrait photo as input, see Fig. 1. Edits are coherent in the entire scene and not limited to certain face areas. Edits maintain perspectively correct parallax, photo-real occlusions and disocclusions, and illumination on the entire person, without warping artifacts in the unmodeled scene parts, such as hair. The embedding is estimated based on a novel non-linear optimization problem formulation. Semantic editing in parameter space is then achieved based on the pretrained neural network of Tewari et al. [2020], which maps the control space of a 3D morphable face model to the latent space of StyleGAN. These semantic edits are accessible through a simple user interface similar to established face animation control. We make the following contributions:

- We propose a hierarchical optimization approach that embeds a portrait image in the latent space of StyleGAN while ensuring high-fidelity as well as editability.
- Moreover, in addition to editability of the head pose, facial expression and scene illumination, we introduce an energy that enforces preservation of the facial identity.

## 2 RELATED WORK

We define face editing as the process of changing the head pose, facial expression, or incident illumination in a portrait image or video. Many recent editing techniques are learning-based. We distinguish between person-specific techniques that require a large corpus of images (or a long video) of the person, few-shot techniques that only require a small number of images, and single-shot techniques that only require a single image as input. Our Portrait Image Embedding (PIE) approach is part of the third category and enables intuitive editing of a portrait image by a set of semantic control parameters. In addition to these categories, we will also summarize existing works related to portrait relighting.

### 2.1 Person-specific Video Editing Techniques

There has been a lot of research on person-specific techniques [Bansal et al. 2018; Kim et al. 2019, 2018; Thies et al. 2019, 2016; Wiles et al. 2018] that require a large training corpus of the target person as input. These approaches can be classified into model-based [Kim

[1] For example: www.adobe.com/Photoshop

et al. 2019, 2018; Thies et al. 2019, 2016] and image-based [Bansal et al. 2018] techniques. Model-based techniques employ a parametric face model to represent the head pose, facial expression, and incident scene illumination. The semantic parameter space spanned by the model can be used to either perform intuitive edits or transfer parameters from a source to a target video. On the other end of the spectrum are image-based techniques that can transfer parameters, but do not provide intuitive semantic control.

*Model-based Video Editing Techniques.* Facial reenactment approaches [Thies et al. 2019, 2016] change the facial expressions in a target video to the expressions in a driving source video. These approaches achieve impressive results, but require a video of the target person as input and do not enable editing of the head pose and incident illumination. Kim et al. [2018] proposed the first full head reenactment approach that is able to edit the head pose as well as the facial expression. A conditional deep generative model is leveraged as a neural rendering engine. While these approaches [Kim et al. 2018; Thies et al. 2019, 2016] produce exciting results, they do not preserve the speaking style of the target. In Kim et al. [2019], an approach is proposed for editing the expressions of a target subject while maintaining his/her speaking style. This is made possible by a novel style translation network that learns a cycle-consistent mapping in blendshape space. In contrast to our approach, all these techniques require a long video of the target as input and cannot edit a single image of an arbitrary person.

*Image-based Video Editing Techniques.* Image-based techniques enable to control a target face through a driving video. The approach of Bansal et al. [2018] allows them to modify the target video while maintaining the speaking style. A novel recycle loss is defined in the spatio-temporal video domain. This approach obtains high-quality results for expressions and pose transfer. In contrast to our approach, image-based approaches do not provide intuitive control via a set of semantic control parameters and have to be trained in a person-specific manner. Thus, they cannot be employed to edit a single given image.

## 2.2 Few-shot Editing Techniques

Few-shot editing techniques [Wang et al. 2019a; Wiles et al. 2018; Zakharov et al. 2019] require only a small set of images of the target person as input. Given multiple frames showing a target person, X2Face [Wiles et al. 2018] drives a frontalized face embedding by a regressed warp field that is estimated by an encoder-decoder network. The approach can also drive faces based on audio. Wang et al. [2019a] presented a few-shot video editing approach and showed its application to driving a target face via a source video. A novel network weight generation module is proposed that is based on an attention mechanism. To animate faces, the network is trained to transfer image sketches to photo-realistic face images. The network is trained on a large multi-identity training corpus and can be applied to new unseen still images. Zakharov et al. [2019] presented a few-shot technique for animating faces. Their solution has three components: 1) a generator network that translates landmark positions to photo-realistic images, 2) an embedding network that learns an intermediate representation for conditioning the generator, and 3) a discriminator. The network is trained on a large corpus of face images across multiple identities and generalizes to new identities at test time. Impressive results are shown in animating images, including legacy photos and even paintings. The learned models of few-shot techniques [Wang et al. 2019a; Wiles et al. 2018; Zakharov et al. 2019] can be improved by fine-tuning on a few example images of the target person, e.g., images taken at different view-points or at different time instances. The learned models can also be applied directly to new still images without fine-tuning.

## 2.3 Single-shot Editing Techniques

Several works [Averbuch-Elor et al. 2017; Geng et al. 2018; Nagano et al. 2018] exist for controlling the expression and head pose given a single image as input. Nagano et al. [2018] presented *paGAN*, an approach for creating personalized avatars from just a single image of a person. However, the work does not synthesize photo-realistic hair. The approach of Averbuch-Elor et al. [2017] brings portrait images to life by animating their expression and pose. The target image is animated through a 2D warp that is computed from the movement in the source video. The mouth interior is copied from the source and blended into the warped target image. The approach of Geng et al. [2018] employs deep generative models to synthesize more realistic facial detail and a higher quality mouth interior. First, a dense spatial motion field is used to warp the target image. Afterwards, the first network corrects the warped target image and synthesizes important skin detail. Finally, the second network synthesizes the mouth interior, including realistic teeth. Siarohin et al. [2019] proposed a method for animating a single image based on a driving sequence. By detecting keypoints in both the target image and the driving frames, the method uses a neural network to compute a dense warping field, specifying how to translate the driving frames into the target image. Based on this information a second network produces high-quality output frames. Since keypoint extraction is also learned during training, the method is applicable for any category of input, and in particular works for face and full body images. While existing single-shot editing techniques can only be controlled via a driving video, our approach enables intuitive editing of the head pose, facial expression and incident illumination in a portrait image through intuitive parametric control, as well as through a driving video.

## 2.4 Portrait Relighting

Relighting approaches modify the incident illumination on the face [Meka et al. 2019; Peers et al. 2007; Shu et al. 2017; Sun et al. 2019; Zhou et al. 2019]. Earlier works [Peers et al. 2007; Shu et al. 2017] require an exemplar portrait image that has been taken under the target illumination conditions. More recent techniques use deep generative models [Meka et al. 2019; Sun et al. 2019; Zhou et al. 2019] and can relight images based on an environment map. Zhou et al. [2019] train a relighting technique based on a large corpus of synthetic images. Relighting is performed in the luminance channel, which simplifies the learning task. Sun et al. [2019] use light stage data to train their relighting approach. At test time, the network produces high quality relighting results, even for in-the-wild images. While training with light stage data leads to high-quality results, their scarcity and careful recording protocol can limit their adaptation. Meka et al. [2019] showed that the 4D reflectance field can be estimated from two color gradient images captured in a light stage. This provides more movement flexibility for the subject during recording, and hence takes an important step towards capturing relightable video.

### 2.5 Image Editing using SyleGAN

Several recent methods have been proposed to edit StyleGAN generated images. Most approaches linearly change the StyleGAN latent codes for editing [Härkönen et al. 2020; Shen et al. 2020; Tewari et al. 2020]. Non-linear editing has been shown in Abdal et al. [2020b]. Image2StyleGAN [Abdal et al. 2019, 2020a] is a popular approach for embedding real images into the StyleGAN latent space with very high fidelity. InterFaceGAN [Shen et al. 2020] and StyleFlow [Abdal et al. 2020b] demonstrate editing of real images using these embeddings. Very recently, Zhu et al. [2020] introduce a domain-guided embedding method which allows for higher-quality editing, compared to Image2StyleGAN. However, they do not demonstrate results at the highest resolution for StyleGAN. In this paper, we design an embedding method which allows for high-quality portrait editing using StyleRig [Tewari et al. 2020].

## 3 RIGGING STYLEGAN-GENERATED IMAGES

StyleGAN [Karras et al. 2019b] can synthesize human faces at an unprecedented level of photorealism. However, their edits are defined in terms of three main facial levels (coarse, medium and fine), with no semantic meaning attached to them. StyleRig [Tewari et al. 2020] attaches a semantic control for a StyleGAN embedding, allowing edits for head pose, illumination and expressions. The control is defined through a 3D Morphable Face Model (3DMM) [Blanz and Vetter 1999].

### 3.1 StyleRig in more detail

Faces are represented by a 3DMM model with $m = 257$ parameters

$$\theta = (\phi, \rho, \alpha, \delta, \beta, \gamma) \in \mathbb{R}^{257} \,. \tag{1}$$

Here, $(\phi, \rho) \in \mathbb{R}^6$ are the rotation and translation parameters of the head pose, where rotation is defined using Euler angles. The vector $\alpha \in \mathbb{R}^{80}$ represents the geometry of the facial identity, while $\beta \in \mathbb{R}^{64}$ are the expression parameters. Skin reflectance is defined by $\delta \in \mathbb{R}^{80}$ and the scene illumination by $\gamma \in \mathbb{R}^{27}$. The basis vectors of the geometry and reflectance models are learned from 200 facial 3D scans [Blanz and Vetter 1999]. The expression model is learned from FaceWarehouse [Cao et al. 2014] and the Digital Emily project [Alexander et al. 2010]. Principal Components Analysis (PCA) is used to compress the original over-complete blendshapes to a subspace of 64 parameters. Faces are assumed to be Lambertian, where illumination is modeled using second-order spherical harmonics (SH) [Ramamoorthi and Hanrahan 2001].

StyleRig [Tewari et al. 2020] allows one to semantically edit synthetic StyleGAN images. To this end, StyleRig trains a neural network, called *RigNet*, which can be understood as a function $\text{rignet}(\cdot, \cdot)$ that maps a pair of StyleGAN code $\mathbf{v}$ and subset of 3DMM parameters $\theta^\tau$ to a new StyleGAN code $\hat{\mathbf{v}}$, i.e. $\hat{\mathbf{v}} = \text{rignet}(\mathbf{v}, \theta^\tau)$. In practice, the 3DMM parameters are first transformed before being used in the network. Please refer to the supplemental document for details. With that, $\mathbf{I}_{\hat{\mathbf{v}}}$ shows the face of $\mathbf{I}_{\mathbf{v}}$ modified according to $\theta^\tau$ (i.e. with edited head pose, scene lighting, or facial expression), where $\mathbf{I}_{\mathbf{v}}$ is the StyleGAN image generated using the latent code $\mathbf{v}$. Thus, editing a synthetic image $\mathbf{I}_{\mathbf{v}}$ amounts to modifying the component $\tau$ in the parameter $\theta$, and then obtaining the edited image as $\mathbf{I}_{\hat{\mathbf{v}}} = \mathbf{I}(\text{rignet}(\mathbf{v}, \theta^\tau))$. Multiple RigNet models are trained, each to deal with just one mode of control (pose, expression, lighting). Although RigNet allows for editing of facial images, it has the major shortcoming that only *synthetic* images can be manipulated, rather than real images. This is in contrast to this work, where we are able to perform semantic editing of *real* images. Different from the original RigNet design where a differentiable face reconstruction network regresses the 3DMM parameters from a StyleGAN code, we use a model-based face autoencoder [Tewari et al. 2017] which takes an image as an input. This change is necessary, as we initially do not have the StyleGAN code for the real image we want to edit.

## 4 SEMANTIC EDITING OF REAL FACIAL IMAGES

We present an approach that allows for semantic editing of real facial images. The key of our approach is to embed a given facial image in the StyleGAN latent space [Karras et al. 2019b], where we pay particular attention to finding a latent encoding that is *suitable for editing the image.* This is crucial, since the parameter space of the StyleGAN architecture is generally under-constrained. For example, it has been shown that a StyleGAN trained for human faces is able to synthesize images that show completely different content with high fidelity, such as images of cat faces [Abdal et al. 2019] Our goal is to compute embeddings which can be edited using 3DMM parameters using StyleRig.

*Problem Statement.* We will refer to the image that we want to make editable as $\mathbf{I}$ (without any subscripts or arguments), which we assume to be a given input. Moreover, we will refer to the StyleGAN code that will make image $\mathbf{I}$ editable as $\mathbf{w}$, which is the desired output of our approach. As such, we will introduce an energy function $E(\mathbf{w})$, which is minimized by solving a numerical optimization problem. This energy function accounts for the high fidelity of the synthesized image based on $\mathbf{w}$ (explained in Sec. 4.1), for editing-suitability (described in Sec. 4.2), as well as for consistent face identity before and after the edit (Sec. 4.3). We emphasize that our approach is based on non-linear optimization techniques, and does not perform any learning of network weights, which in turn means that we do not require any ground truth data of edited facial images. In order to formulate the energy function we will make use of several existing neural networks, where all of them are pretrained and remain fixed throughout the optimization. We will now introduce some technical notations, which will allow us to have an additional layer of abstraction and thereby facilitate a more comprehensive description of the main concepts.

*Notation.* Throughout this paper we will use $\mathbf{w}$ exclusively to refer to the (unknown) StyleGAN embedding that we want to find, and we will use $\mathbf{v}$ (potentially with subscripts) to refer to general StyleGAN embeddings. We note that the StyleGAN embeddings $\mathbf{w}$

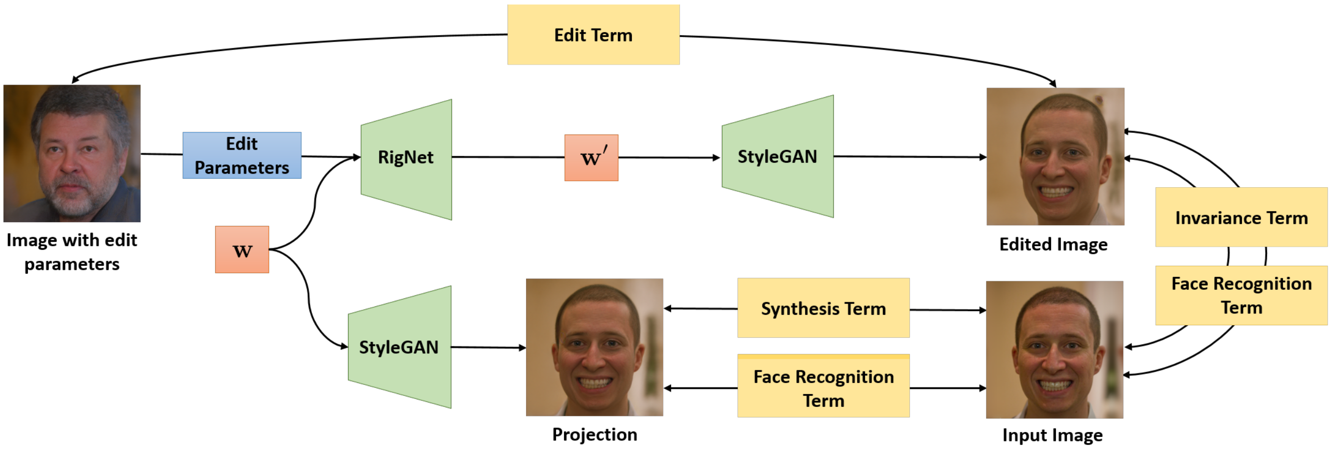

Fig. 2. Given a portrait input image, we optimize for a StyleGAN embedding which allows to faithfully reproduce the image (synthesis and facial recognition terms), editing the image based on semantic parameters such as head pose, expressions and scene illumination (edit and invariance terms), as well as preserving the facial identity during editing (facial recognition term). A novel hierarchical non-linear optimization strategy is used to compute the result. StyleGAN generated images (image with edit parameters) are used to extract the edit parameters during optimization. At "test time", i.e. for performing portrait image editing, the image with edit parameters is not needed. Note that the identity term is not visualized here. Images from Shih et al. [2014].

and $\mathbf{v}$ can have two different forms, where each form has a different dimension, which we will describe in detail in Sec. 4.4. StyleGAN can be understood as a function stylegan$(\cdot)$ that maps a given latent code to a portrait image. To simplify notation, we will use the function notation $\mathbf{I}(\mathbf{v}) := \text{stylegan}(\mathbf{v})$ in order to emphasize that we use the StyleGAN embedding $\mathbf{v}$ to generate the image $\mathbf{I}(\mathbf{v})$. Analogously, we overload $\mathbf{I}(\cdot)$, so that it can also take a 3DMM parameter $\theta$ as input. As such, $\mathbf{I}(\theta)$ refers to an image rendered using the face model that is parameterized by $\theta$ (Sec. 3.1), where differentiable rendering is employed [Tewari et al. 2017]. Note that this rendered image is only defined on foreground face pixels as opposed to StyleGAN images.

We use the variable $\tau \in \{\phi, \beta, \gamma\}$ to indicate the user-defined facial semantic variable that is to be edited, which in our case can be the head pose $\phi$, facial expression $\beta$, or illumination $\gamma$. Similarly, we use the complement notation $\overline{\tau} \subset \{\phi, \rho, \alpha, \delta, \beta, \gamma\}$, to indicate all other variables, i.e., the ones that shall not be modified. With that, we use the notation $\theta^{\tau}$ (or $\theta^{\overline{\tau}}$) to refer to the extraction of the $\tau$-component (or $\overline{\tau}$-components) of $\theta$. Since facial editing is implemented by modifying the $\tau$-component of the 3DMM parameter $\theta$, we write $\theta' = [\theta_1^{\overline{\tau}}, \theta_2^{\tau}]$ to indicate that the respective $\tau$-component of $\theta_1$ is replaced by the corresponding component in $\theta_2$. For example, for $\tau = \beta$,

$$\theta_1 = (\phi_1, \rho_1, \alpha_1, \delta_1, \beta_1, \gamma_1)\,, \quad \text{and} \tag{2}$$

$$\theta_2 = (\phi_2, \rho_2, \alpha_2, \delta_2, \beta_2, \gamma_2)\,, \quad \text{we have} \tag{3}$$

$$[\theta_1^{\overline{\tau}}, \theta_2^{\tau}] = (\phi_1, \rho_1, \alpha_1, \delta_1, \beta_2, \gamma_1)\,. \tag{4}$$

Moreover, we use the notation $\theta(\mathbf{v})$ to extract the 3DMM parameters from the StyleGAN embedding $\mathbf{v}$. In order to compute this, the embedding $\mathbf{v}$ is first used to synthesize the image $\mathbf{I}(\mathbf{v})$ (using StyleGAN), followed by performing a 3D reconstruction based on the pretrained *Model-based Face Autoencoder* (MoFA) network [Tewari et al. 2017]. Hence, for MoFA$(\cdot)$ being the function that performs 3D reconstruction for a given image by estimating the 3DMM parameters, we define

$$\theta(\mathbf{v}) = \text{MoFA}(\mathbf{I}(\mathbf{v}))\,. \tag{5}$$

For any image $\mathbf{I}'$, we use the short-hand notation $\theta(\mathbf{I}') = \text{MoFA}(\mathbf{I}')$. Similarly as above, we will use $\theta^{\tau}(\mathbf{v})$ and $\theta^{\tau}(\mathbf{I}')$ to extract only the $\tau$-component from the 3DMM parameters. Whenever arguments of $\theta(\cdot)$ or $\mathbf{I}(\cdot)$ are fixed, i.e., the arguments are not a variable, we use the short-hand notations $\theta_{\mathbf{v}} = \theta(\mathbf{v})$, $\theta_{\mathbf{I}'} = \theta(\mathbf{I}')$, or $\mathbf{I}_{\mathbf{v}} = \mathbf{I}(\mathbf{v})$. We summarize the most important part of our notations in Table 1.

Table 1. Summary of notation.

| Symbol | Meaning |
|---|---|
| $\mathbf{w}$ | StyleGAN embedding that we want to find |
| $\mathbf{v}$ | other StyleGAN embedding(s) |
| $\theta$ | 3DMM parameter |
| $\tau$ | component that is to be edited ($\tau \in \{\phi, \beta, \gamma\}$) |
| $\mathbf{I}$ | input image that we want to edit |
| $\mathbf{I}(\mathbf{v})$ | StyleGAN-synthesized image |
| $\mathbf{I}(\theta)$ | image of 3DMM rendering |
| $\theta^{\tau}$ | extraction of $\tau$-component of $\theta$ |
| $[\theta_1^{\overline{\tau}}, \theta_2^{\tau}]$ | combine $\overline{\tau}$-components in $\theta_1$ with $\tau$-component in $\theta_2$ |
| $\theta(\mathbf{v}), \theta_{\mathbf{v}}$ | 3D reconstruction of 3DMM parameters from $\mathbf{I}(\mathbf{v})$ |
| $\theta(\mathbf{I}'), \theta_{\mathbf{I}'}$ | 3D reconstruction of 3DMM parameters from $\mathbf{I}'$ |

*Objective function.* We solve for $\mathbf{w}$ by minimizing the energy function

$$E(\mathbf{w}) = E_{\text{syn}}(\mathbf{w}) + E_{\text{id}}(\mathbf{w}) + E_{\text{edit}}(\mathbf{w}) + E_{\text{inv}}(\mathbf{w}) + E_{\text{recog}}(\mathbf{w})\,. \tag{6}$$

$E_{\text{syn}}$ is a synthesis term enforcing the StyleGAN-synthesized image $\mathbf{I}(\mathbf{w})$ to be close to $\mathbf{I}$ (Sec. 4.1). $E_{\text{id}}$, $E_{\text{edit}}$, $E_{\text{inv}}$ are face modification terms (Sec. 4.2) enforcing edits to take place on the modified facial semantics while at the same time ensuring unmodified facial semantics to remain un-edited. $E_{\text{recog}}(\mathbf{w})$ is a face recognition term that will be introduced in Sec. 4.3. A conceptual illustration of the energy function and the overall pipeline is shown in Fig. 2. Next, we will discuss each term in more detail.

## 4.1 High-Fidelity Image Synthesis

Similarly to Image2StyleGAN [Abdal et al. 2019], we use the following energy term that accounts for the StyleGAN-synthesized image $\mathbf{I}(\mathbf{w})$ being close to $\mathbf{I}$:

$$E_{\text{syn}}(\mathbf{w}) = \lambda_{\ell_2} \|\mathbf{I} - \mathbf{I}(\mathbf{w})\|_2^2 + \lambda_{\text{p}} \|\mathbf{\Phi}(\mathbf{I}) - \mathbf{\Phi}(\mathbf{I}(\mathbf{w}))\|_2^2 . \tag{7}$$

The first term in the energy $E_{\text{syn}}$ penalizes the discrepancy between $\mathbf{I}$ and the synthesized image in terms of the (squared) $\ell_2$-norm, whereas the second term penalizes discrepancies based on the *perceptual loss* [Johnson et al. 2016]. The perceptual loss is estimated on images downsampled by a factor of 4, based on $\ell_2$-losses over VGG-16 layers `conv1_1`, `conv1_2`, `conv3_2` and `conv4_2` [Simonyan and Zisserman 2015]. The notation $\mathbf{\Phi}(\cdot)$ refers to the function that downsamples a given input image and extracts features. The scalars $\lambda_{\ell_2}$ and $\lambda_{\text{p}}$ are the relative weights of both terms.

In principle, we could minimize the energy $E_{\text{syn}}$ in (7) in order to obtain the StyleGAN code $\mathbf{w}$, as done in Abdal et al. [2019], and perform editing operations on $\mathbf{w}$. A so-obtained code vector $\mathbf{w}$ allows the use of StyleGAN to obtain a highly accurate synthetic version of the input face, which is even capable of reconstructing backgrounds with high accuracy. However, such a $\mathbf{w}$ is sub-optimal for performing *semantic face editing*, as we later demonstrate in Fig. 6.

## 4.2 Face Image Editing

We augment the synthesis term with an editing energy that is based on the StyleRig framework [Tewari et al. 2020], which allows us to obtain more accurate semantic editing while preserving the non-edited attributes. Here, the StyleGAN embedding $\mathbf{w}$ that is to be determined should have the following three properties in order to be suitable for semantic editing:

*Identity Property.* The identity property is phrased in terms of the $\ell_2$-norm of the difference of StyleGAN embeddings and is given by

$$E_{\text{id}}(\mathbf{w}) = \lambda_{\text{id}} \|\mathbf{w} - \text{rignet}(\mathbf{w}, \theta^{\tau}(\mathbf{w}))\|_2^2 . \tag{8}$$

As such, whenever the RigNet is used to modify $\mathbf{w}$ with $\theta^{\tau}(\mathbf{w})$, i.e., a component of the 3DMM parameter extracted from $\mathbf{w}$, the embedding $\mathbf{w}$ should not be modified.

*Edit Property.* In order to get around the obstacle of defining a suitable metric for 3DMM parameter vectors, whose components may be of significantly different scale, and the relative relevance of the individual components is not easily determined, we phrase the edit property in image space, as in StyleRig [Tewari et al. 2020]. As such, a facial edit is implicitly specified in image space via the StyleGAN embedding $\mathbf{v}$, where the $\tau$-component of the respective 3DMM parameters of $\mathbf{v}$, i.e. $\theta^{\tau}_{\mathbf{v}}$, specifies the edit operation. The image-space version of the edit property reads

$$\forall\, \mathbf{v}: \quad \mathbf{I}_{\mathbf{v}} = \mathbf{I}([\theta^{\overline{\tau}}_{\mathbf{v}}, \theta^{\tau}(\text{rignet}(\mathbf{w}, \theta^{\tau}_{\mathbf{v}}))]) . \tag{9}$$

Note that this true equality cannot hold in practice, since the two images are from different domains (real image and a mesh rendering). We are interested in minimimzing the difference between these terms. This equation is best fulfilled whenever the $\tau$-component of the edited 3DMM parameters $\theta^{\tau}(\text{rignet}(\mathbf{w}, \theta^{\tau}_{\mathbf{v}}))$ is equal to $\theta^{\tau}_{\mathbf{v}}$, i.e. the edit has been successfully applied. Since computationally we cannot evaluate all choices of $\mathbf{v}$, we sample StyleGAN embeddings $\mathbf{v}$ as done in Tewari et al. [2020], and then use the expected value as loss. For integrating this property into our optimization framework we use a combination of a photometric term and a landmark term, which is defined as

$$\ell(\mathbf{I}', \theta) = \lambda_{\text{ph}} \|\mathbf{I}' - \mathbf{I}(\theta)\|_{\odot}^2 + \lambda_{\text{lm}} \|\mathcal{L}_{\mathbf{I}'} - \mathcal{L}(\theta)\|_F^2 . \tag{10}$$

The norm $\|\cdot\|_{\odot}$ computes the $\ell_2$-norm of all *foreground* pixels (the facial part of the image), whereas $\|\cdot\|_F$ is the Frobenius norm. By $\mathcal{L}_{\mathbf{I}'} \in \mathbb{R}^{66\times2}$ we denote the matrix of 2D facial landmarks extracted from the image $\mathbf{I}$ (based on Saragih et al. [2011]), and $\mathcal{L}(\theta) \in \mathbb{R}^{66\times2}$ refers to the corresponding landmarks of the 3DMM after they have been projected onto the image plane. With that, the edit property energy reads

$$E_{\text{edit}}(\mathbf{w}) = \lambda_{\text{e}}\, \mathbb{E}_{\mathbf{v}}[\ell(\mathbf{I}_{\mathbf{v}}, [\theta^{\overline{\tau}}_{\mathbf{v}}, \theta^{\tau}(\text{rignet}(\mathbf{w}, \theta^{\tau}_{\mathbf{v}}))])] . \tag{11}$$

*Invariance Property.* Similarly as the edit property we phrase the invariance property also in image space as

$$\forall\, \mathbf{v}: \quad \mathbf{I} = \mathbf{I}([\theta^{\overline{\tau}}(\text{rignet}(\mathbf{w}, \theta^{\tau}_{\mathbf{v}})), \theta^{\tau}_{\mathbf{I}}]) . \tag{12}$$

While the edit property imposes that the $\tau$-component of the edited 3DMM parameter $\theta^{\tau}(\text{rignet}(\mathbf{w}, \theta^{\tau}_{\mathbf{v}}))$ is modified as desired, the invariance property takes care of all $\overline{\tau}$. It is fulfilled whenever it holds that $\theta^{\overline{\tau}}(\text{rignet}(\mathbf{w}, \theta^{\tau}_{\mathbf{v}})) = \theta^{\overline{\tau}}_{\mathbf{I}}$, i.e. the components $\overline{\tau}$ that are not to be edited are maintained from the input image $\mathbf{I}$.

Analogously to the edit property, we base the respective energy on the combination of a photometric term and a landmark term as implemented by $\ell(\cdot)$, so that we obtain

$$E_{\text{inv}}(\mathbf{w}) = \lambda_{\text{inv}}\, \mathbb{E}_{\mathbf{v}}[\ell(\mathbf{I}, [\theta^{\overline{\tau}}(\text{rignet}(\mathbf{w}, \theta^{\tau}_{\mathbf{v}})), \theta^{\tau}_{\mathbf{I}}])] . \tag{13}$$

## 4.3 Face Recognition Consistency

In addition to the synthesis and editing terms, we incorporate two face recognition consistency terms to preserve the facial integrity while editing. On the one hand, it is desirable that the synthesized image $\mathbf{I}(\mathbf{w})$ is recognized to depict the same person as shown in the given input image $\mathbf{I}$. On the other hand, the edited image, $\text{stylegan}(\text{rignet}(\mathbf{w}, \theta^{\tau}_{\mathbf{v}}))$ should also depict the same person as shown in the input $\mathbf{I}$.

In order to do so, we use VGG-Face [Parkhi et al. 2015] to extract *face recognition features*, where we use the notation $\mathbf{\Psi}(\cdot)$ to refer to the function that extracts such features from a given input image. We define the recognition loss

$$\ell_{\text{recog}}(\mathbf{I}', \mathbf{v}) = \|\mathbf{\Psi}(\mathbf{I}') - \mathbf{\Psi}(\mathbf{I}(\mathbf{v}))\|_F^2 , \tag{14}$$

which is then used to phrase the recognition energy term as

$$E_{\text{recog}}(\mathbf{w}) = \lambda_{\text{r}_{\mathbf{w}}}\, \ell_{\text{recog}}(\mathbf{I}, \mathbf{w}) + \lambda_{\text{r}_{\hat{\mathbf{w}}}}\, \mathbb{E}_{\mathbf{v}}[\ell_{\text{recog}}(\mathbf{I}, \text{rignet}(\mathbf{w}, \theta^{\tau}_{\mathbf{v}}))] . \tag{15}$$

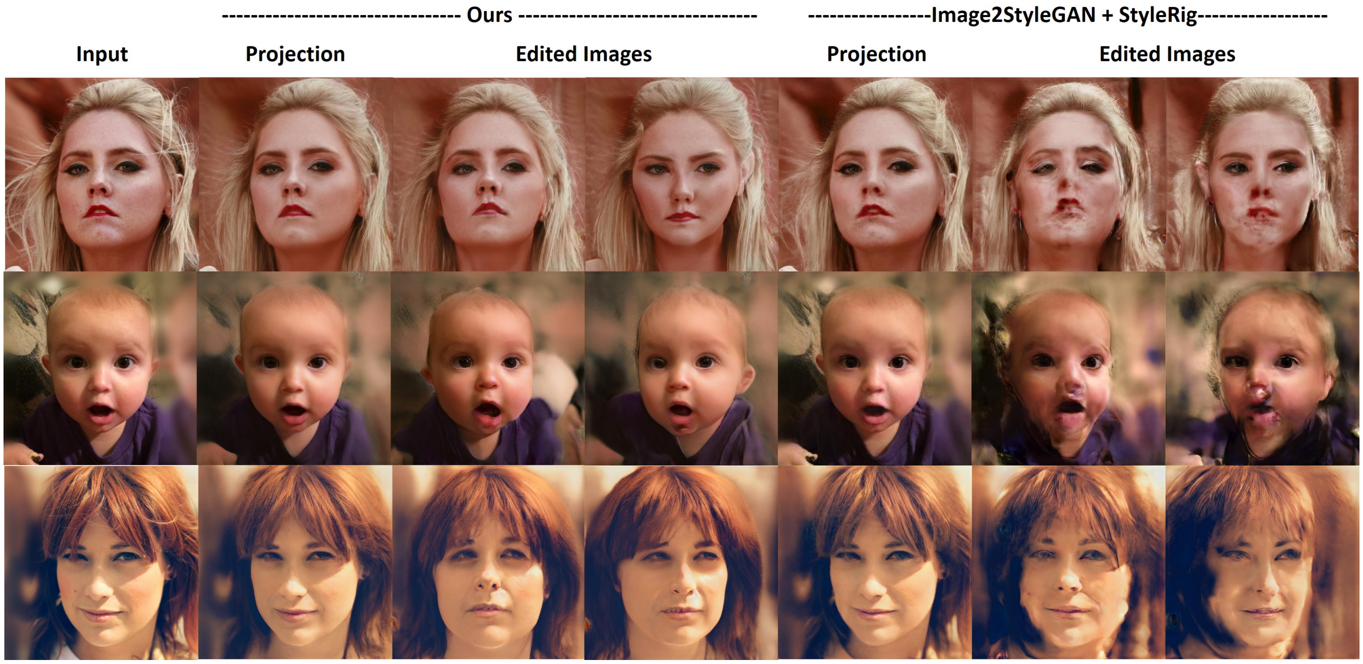
---------------------------------- Ours ---------------------------------
-----------------Image2StyleGAN + StyleRig------------------
Input
Projection
Edited Images
Projection
Edited Images

Fig. 3. Pose Editing. Our approach can handle a large variety of head pose modifications including out-of-plane rotations in a realistic manner. Image2StyleGAN [Abdal et al. 2019] embeddings often lead to artifacts when edited using StyleRig. Images from Shen et al. [2016].

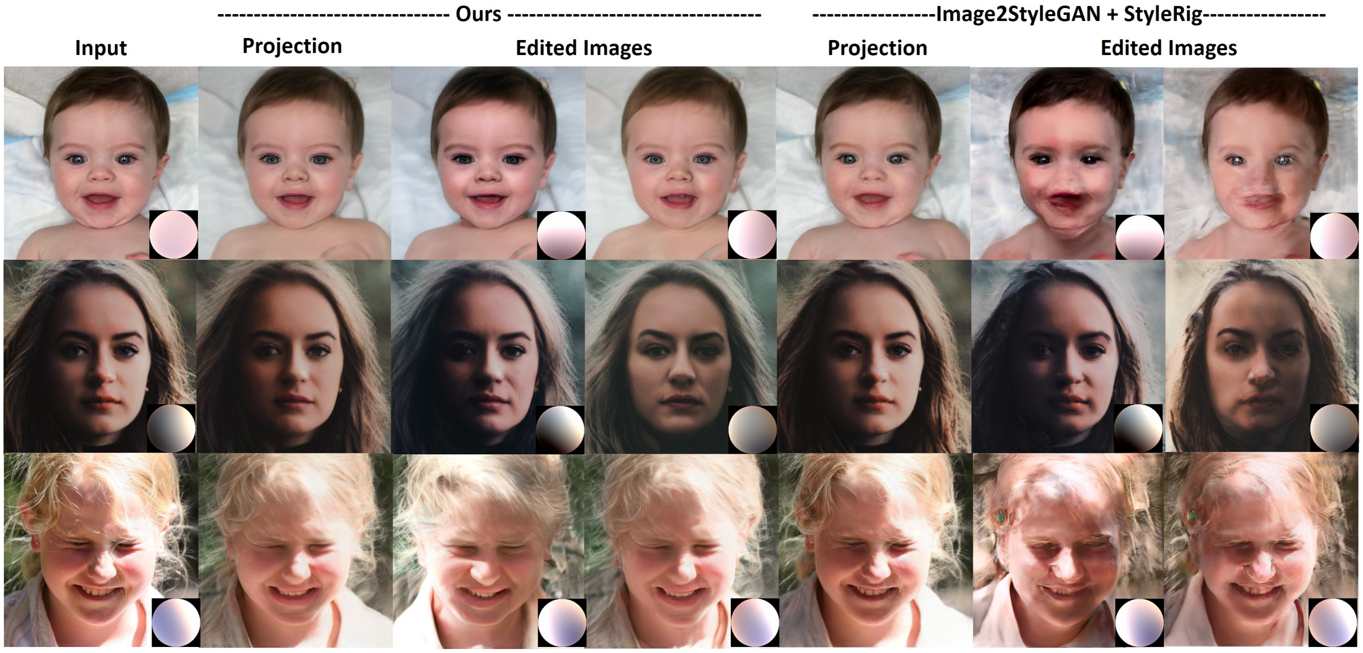
-------------------------------- Ours ----------------------------------
-----------------Image2StyleGAN + StyleRig-----------------
Input
Projection
Edited Images
Projection
Edited Images

Fig. 4. Illumination Editing. Our approach can realistically relight portrait images. Each edited image corresponds to changing a different Spherical Harmonics coefficient, while all other coefficients are kept fixed. The environment maps are visualized in the inset. Image2StyleGAN [Abdal et al. 2019] embeddings often lead to artifacts when edited using StyleRig. Images from Shen et al. [2016].

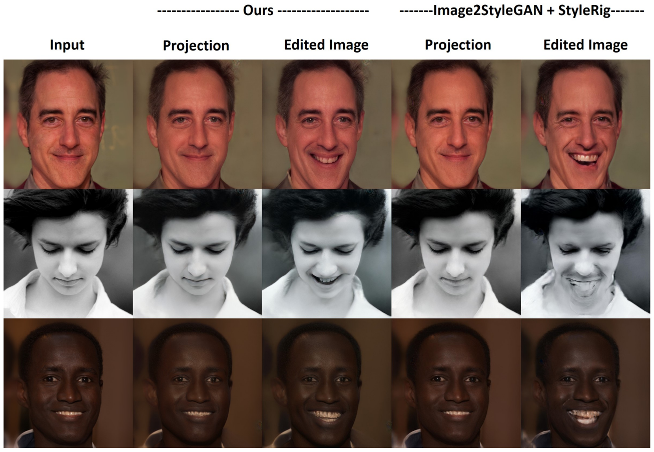
---------------- Ours ------------------
-------Image2StyleGAN + StyleRig-------
Input
Projection
Edited Image
Projection
Edited Image

Fig. 5. Expression Editing. Our approach can also be used to edit the facial expressions in a portrait image in a realistic manner. We obtain more plausible results, compared to Image2StyleGAN [Abdal et al. 2019] embeddings. Images from Shen et al. [2016] and Shih et al. [2014].

## 4.4 Optimization

Our energy function $E(\cdot)$ in (6) depends on a range of highly non-linear functions, such as stylegan$(\cdot)$, MoFA$(\cdot)$, $\Phi(\cdot)$ and $\Psi(\cdot)$, which are implemented in terms of (pretrained) neural networks. We implement our energy minimization within TensorFlow [Abadi et al. 2015] using ADADELTA optimization [Zeiler 2012]. In each iteration we stochastically sample a different $\mathbf{v}$. The optimization uses a hierarchical approach that we describe next.

*Hierarchical Optimization.* StyleGAN is based on a hierarchy of latent spaces, where a stage-one embedding $Z$ with $|Z| = 512$ is randomly sampled first. This is then fed into a mapping network that produces $W$ as output, where $|W| = 512$. Subsequently, $W$ is extended to $W^+$, where $|W^+| = 18 \times 512$, and used as input to 18 network layers. It has been shown that $W^+$ is the most expressive space for fitting to real images [Abdal et al. 2019]. However, we found that a direct optimization over this space leads to lower-quality editing results with severe artifacts. This is because the optimized variable can be far from the prior distribution of StyleGAN. To address this, we first optimize for the embedding in the $W$-space, meaning that in the first stage of our optimization the variable $\mathbf{w}$ is understood as an embedding in the $W$-space. We optimize in $W$-space for 2000 iterations. We then transfer the result to $W^+$-space, initialize the variable $\mathbf{w}$ respectively, and continue optimizing in the $W^+$-space for another 1000 iterations. Optimizing in this hierarchical way allows us to represent the coarse version of the image in the $W$-space, which is less expressive and thereby closer to the prior distribution. Finetuning on the $W^+$ space then allows us to fit the fine-scale details, while preserving editing quality.

## 5 RESULTS

In the following, we demonstrate the high-quality results of our method, analyze its different components, as well as compare to several state-of-the-art approaches for portrait image editing.

*Implementation Details.* We use the following weights for our energy terms: $\lambda_{\ell_2} = 10^{-6}$, $\lambda_{\mathrm{p}} = 10^{-6}$, $\lambda_{\mathrm{id}} = 1.0$, $\lambda_{\mathrm{ph}} = 0.001$, $\lambda_{\mathrm{lm}} = 0.2$, $\lambda_{\mathrm{e}} = 10.0$, $\lambda_{\mathrm{inv}} = 10.0$, $\lambda_{\mathrm{r_w}} = 0.1$, $\lambda_{\mathrm{r_{\hat{w}}}} = 0.1$. We use a starting step size of 50 when optimizing over embeddings in $W$ space, and 10 in $W^+$ space. The step size is then exponentially decayed by a factor of 0.1 every 2000 steps. Optimization takes approximately 10 minutes for 3000 iterations per image on an NVIDIA V100 GPU. Once the embedding is obtained, the portrait image can be edited at an interactive speed.

*Feedback.* We noticed that a simple feedback loop allows us to get more accurate editing results. We update the parameters used as input to RigNet in order to correct for the editing inaccuracies in the output. Given target 3DMM parameters $\theta$, we first obtain the embedding for the edited image, $\mathrm{rignet}(\mathbf{w}, \theta^\tau)$. We then estimate the 3DMM parameters from the edited embedding, $\theta_{\mathrm{est}} = \theta(\mathrm{rignet}(\mathbf{w}, \theta^\tau))$. The final embedding is computed as $\mathrm{rignet}(\mathbf{w}, \theta^\tau_{\mathrm{new}})$ with $\theta_{\mathrm{new}} = \theta + (\theta - \theta_{\mathrm{est}})$.

### 5.1 High-Fidelity Semantic Editing

We evaluate our approach on a large variety of portrait images taken from Shen et al. [2016] and Shih et al. [2014]. The images are preprocessed as in StyleGAN [Karras et al. 2019b]. Figs. 3, 4, 5 show results of controlling the head pose, scene illumination, and facial expressions, respectively. The projections onto the StyleGAN space are detailed, preserving the facial identity. Our approach also produces photo-realistic edits. Fig. 3 shows that our approach can handle a large variety of head pose modifications, including out-of-plane rotations. It also automatically inpaints uncovered background regions in a photo-realistic manner. Fig. 4 demonstrates our relighting results. Our approach can handle complex light material interactions, resulting in high photo-realism. The relighting effects are not restricted to just the face region, with hair and even eyes being relit. Our approach also allows for editing facial expressions, see Fig. 5. For smooth temporal editing results of portrait images, please refer to the supplementary video.

### 5.2 Ablation Studies

Here, we evaluate the importance of the different proposed loss functions, and also evaluate the hierarchical optimization strategy. Please refer to the supplemental document for the evaluation of the feedback strategy.

*Loss Functions.* Fig. 6 shows qualitative ablative analysis for the different loss functions. We group the edit, invariance and identity terms as *modification terms*. Adding face recognition consistency without the modification terms lead to incorrect editing in some cases. Adding the modification terms without face recognition consistency leads to the method being able to accurately change the specified semantic property, but the identity of the person in the image is not preserved. Using all terms together leads to results with photorealistic edits with preservation of identity. We do not evaluate the importance of the individual components of the modification terms, as it was already evaluated in Tewari et al. [2020].

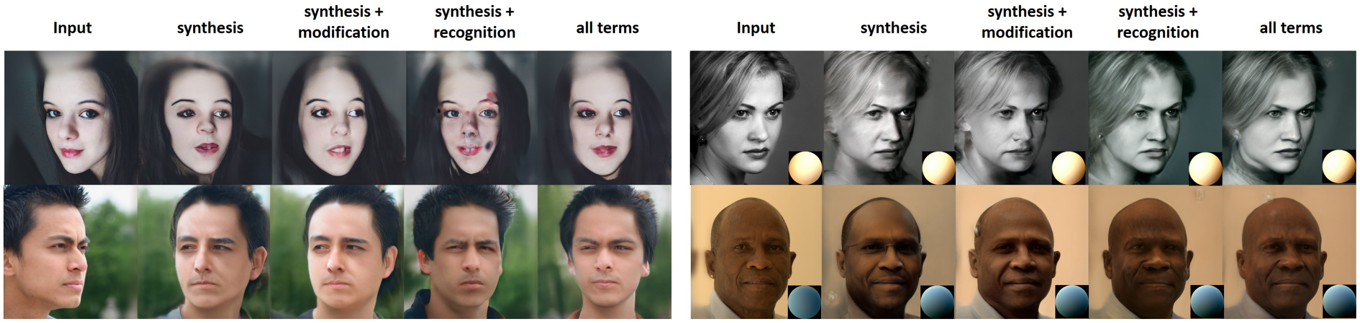

Fig. 6. Ablative analysis of the different loss functions. *Modification* refers to the edit, invariance and identity terms simultaneously. The left block shows results for editing the head pose and the right block shows results for editing scene illumination. All losses are required to obtain high-fidelity edits. Images from Shen et al. [2016].

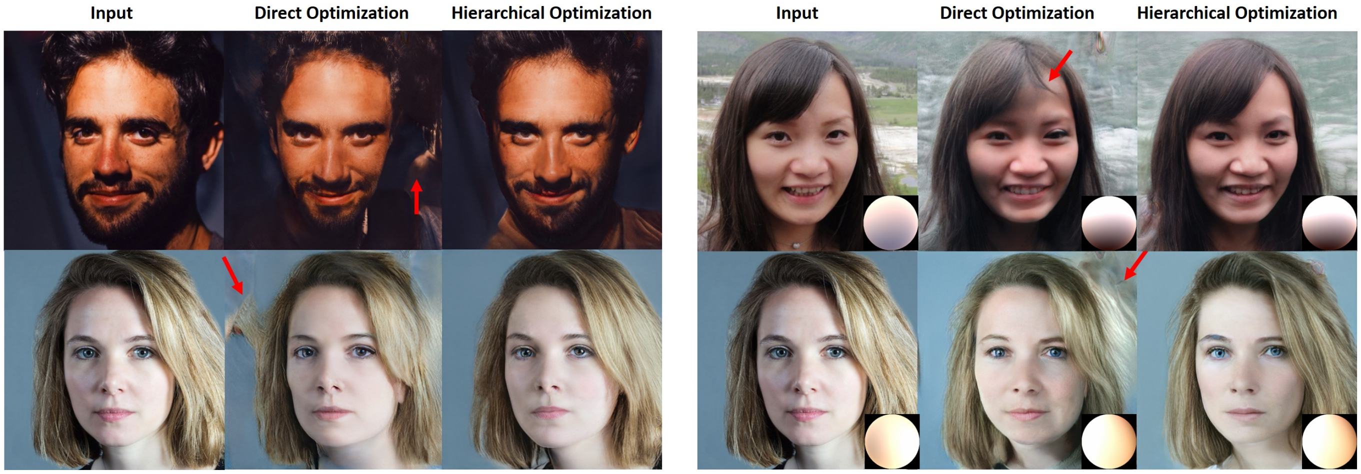

Fig. 7. Ablative analysis with and without hierarchical optimization. The left block shows the results for pose editing and the right block for illumination editing. Without our hierarchical optimization, the obtained embedding cannot be easily edited and artifacts appear in the modified images. Images from Shen et al. [2016].

Table 2. We compare different settings quantitatively using several metrics for pose editing. All numbers are averaged over more than 2500 pose editing results. We measure the quality of the fit by comparing them to the input image using PSNR and SSIM metrics. Editing error is measured as the angular difference between the desired and achieved face poses. Recognition error measures the value of the facial recognition error for the edited images. There is usually a trade-off between the quality and accuracy of editing, as lower recognition errors correspond to higher editing errors. We also compare to Image2StyleGAN [Abdal et al. 2019] embeddings using these metrics. While it achieves the highest quality fitting, the editing results do not preserve the facial identity well.

| | synthesis | synthesis + recognition | synthesis + modification | all terms (PIE) | all terms (direct opt.) | Image2StyleGAN |
|---|---|---|---|---|---|---|
| PSNR (dB) ↑ / SSIM ↑ | 30.15 / 0.70 | 29.84 / 0.69 | 30.15 / 0.70 | 29.96 / 0.70 | 29.76 / 0.69 | **31.21** / **0.75** |
| Editing Error (rad) ↓ | 0.06 | 0.11 | **0.036** | 0.08 | 0.037 | 0.07 |
| Recognition Error ↓ | 95.76 | 43.64 | 90.10 | **42.82** | 51.65 | 275.40 |

*Hierarchical Optimization.* Hierarchical optimization is an important component of our approach. Fig. 7 shows results with and without this component. Without hierarchical optimization, the method directly optimizes for $\mathbf{w} \in W^+$. While this leads to high-quality fits, the obtained embedding can be far from the training distribution of StyleRig. Thus, the quality of edits is poor. For example in Fig. 7 (top), the StyleGAN network interprets the ears as background, which leads to undesirable distortions. With hierarchical optimization, the results do not suffer from artifacts.

*Quantitative Analysis.* We also analyze the effect of different design choices quantitatively, see Tab. 2. We look at three properties, the quality of recostruction (measured using PSNR and SSIM between the projected image and the input), the accuracy of edits (measured as the angular distance between the desired and estimated head poses), and idenity preservation under edits (measured using the second term in Eq. 15) during editing. The numbers reported are averaged over more than 2500 pose editing results. We can see that removing the recognition term changes the identity of the face during editing, and removing the modification loss increases the editing and recognition error. Hierarchical optimization also leads to better facial identity preservation, compared to direct optimization. This is expected, since the results with direct optimization often have artifacts. Note that the artifacts outside of the face region (hair, ears) would not increase the recognition errors significantly. The recognition term introduces a clear trade-off between the quality of identity preservation under edits and the accuracy of edits. The modification terms allow for slight improvements in both identity preservation as well as the accuracy of the edits.

## 5.3 Comparison to the State of the Art

*5.3.1 Image2StyleGAN.* Image2StyleGAN [Abdal et al. 2019] also projects real images to the StyleGAN latent space, and is thus a closely related approach. The source code of Image2StyleGAN was kindly provided by the authors. We show editing results using Image2StyleGAN embeddings in Figs. 1, 3, 4 and 5. Since these embeddings are optimized only using the synthesis terms and without using hierarchical optimization, the results are often implausible, as is most evident when editing the head pose and scene illumination. However, Image2StyleGAN projections are more detailed than ours. We also quantitatively compare to Image2StyleGAN in Tab. 2. Image2StyleGAN obtains the highest quality projections in terms of PSNR and SSIM. When combined with StyleRig, it also leads to low editing errors. However, the recognition errors are very high due to the artifacts in the results, as shown in the qualitative results.

*5.3.2 Other Aproaches.* We also compare our approach to a number of related techniques, X2Face [Wiles et al. 2018], Geng et al. [2018] and Siarohin et al. [2019]. We compare our relighting capabilities to the single-image relighting approach of Zhou et al. [2019]. The source codes of these techniques are publicly available. For Geng et al. [2018], we estimated the landmarks using the dlib tracker [King 2009] as suggested by the authors. We also trained the few shot video-to-video translation method of Wang et al. [2019a] for portrait image editing. We trained on 700 videos from the FaceForensics dataset [Rössler et al. 2019]. Landmarks were extracted using the dlib tracker as recommended by the authors. The approaches of Geng et al. [2018], Wiles et al. [2018] , Wang et al. [2019a] and Siarohin et al. [2019] are trained on a video corpus. In contrast, our method does not use any direct supervision of the edited images. We compare to these methods in two different settings, self-reenactment and cross-identity reenactment.

*Self-Reenactment.* For self-reenactment, we capture several images of a person in different poses. We pick the first image and use the other images of the person as reference to edit the head pose. We captured 9 people in different poses, resulting in 31 images in the test set. Fig. 8 shows some qualitative results. Geng et al. [2018] use a warp-guided algorithm. While this enables expression changes and in-plane head motion, out-of-plane motion cannot be handled as shown in Fig. 8. We also compare to X2Face [Wiles et al. 2018], which samples a learned embedded face in order to synthesize portrait images with different poses and expressions. As such, it shares its limitations with Geng et al. [2018] and produces artifacts for strong pose changes. All approaches do not share the same cropping method, which makes it difficult to quantitatively evaluate the results. In addition, translation of the head during capture can lead to different illumination conditions. Thus, instead of directly computing errors in the image space, we first detect 66 facial landmarks [Saragih et al. 2009] on all results, as well as the reference images. We then compute the landmark alignment error, which is the averaged $\ell_2$-distance between the landmarks after 2D Procrustes alignment (including scale). The implementation of Geng et al. [2018] often fails to generate such large pose edits, so we do not consider this approach in the quantitative evaluation. Due to artifacts, the landmark detector fails on 29% images for the approach of Wiles et al. [2018] and on 23% for Wang et al. [2019a]. All our results, as well as those of Siarohin et al. [2019] pass through the detector. This can be considered as a pseudo-metric of realism, since the landmark detector is trained on real portrait images, implying that our results are better than those of Wiles et al. [2018] and Wang et al. [2019a], and on par with Siarohin et al. [2019]. Table 3 shows the errors for different methods. The low errors for Wiles et al. [2018] are possibly due to the landmark detector failing in challenging cases. We obtain only slightly worse results compared to Siarohin et al. [2019], even though our method does not have access to ground truth during training. Siarohin et al. [2019] train on videos allowing for supervised learning. In addition, their edits are at a lower resolution of 256 × 256, compared to our image resolutions of 1024 × 1024.

Table 3. Evaluation of pose edits: We measure landmark alignment errors for same-subject reenactment on 31 images, and facial recognition distances for cross-subject reenactment on 49 images. Existing landmark detection [Saragih et al. 2009] and facial recognition [King 2009] often fail on images from competing methods, implying higher realism of our results.

| | Landmark Alignment (number of images) | Recognition (number of images) |
|---|---|---|
| Wiles et al. [2018] | **10.9 (22)** | 0.52 (42) |
| Wang et al. [2019a] | 28.19 (24) | 0.49 (45) |
| Siarohin et al. [2019] | 11.97 (31) | 0.51 (46) |
| Ours | 20.12 (31) | **0.40 (49)** |

*Cross-identity Reenactment.* We also compare to others in cross-identity reenactment, which is closer to our setting of semantically disentangled editing. Here, the image being edited and the reference image have different identities. Fig. 8 shows some qualitative results. The implementation of Geng et al. [2018] does not support this setting. Wiles et al. [2018] and Wang et al. [2019a] result in similar artifacts as discussed before. Unlike other approaches, Siarohin et al. [2019] uses two driving images in order to edit the input image, where they use the deformations between the two images as input.

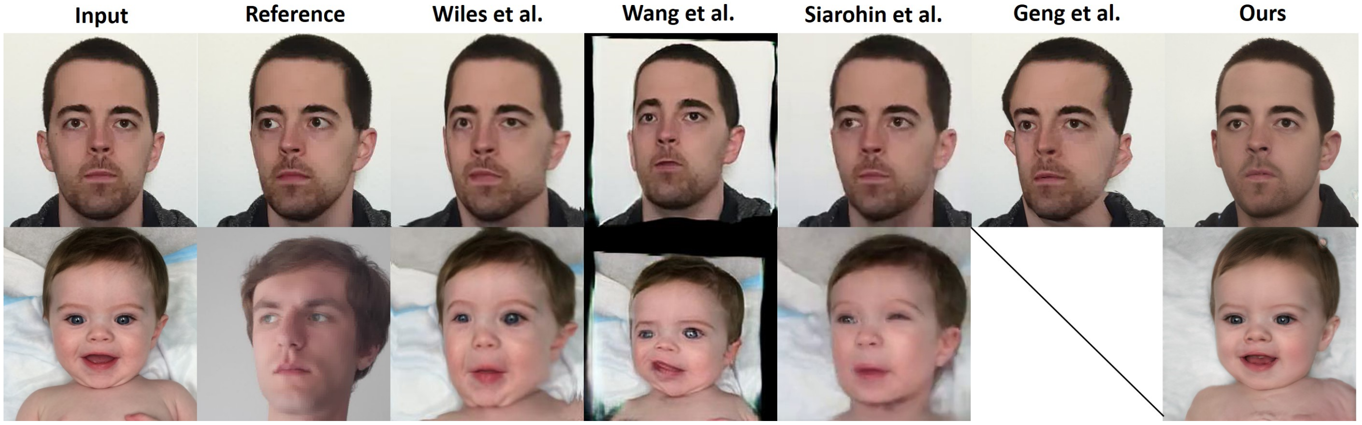

Fig. 8. Comparison of head pose editing for self-reenactment (first two rows) and cross-identity reenactment (last two rows). We compare our approach to Wiles et al. [2018], Wang et al. [2019b], Siarohin et al. [2019] and Geng et al. [2018]. The pose from the reference images is transferred to the input. Our approach obtains higher quality head pose editing results, specially in the case of cross-identity transfer. All approaches other than ours are incapable of *disentangled* edits, i.e., they cannot transfer the pose without also changing the expressions. The implementation of Geng et al. [2018] does not handle cross-identity reenactment. Note that while the three competing approaches require a reference image in order to generate the results, we allow for explicit control over the pose parameters. Image from Shen et al. [2016].

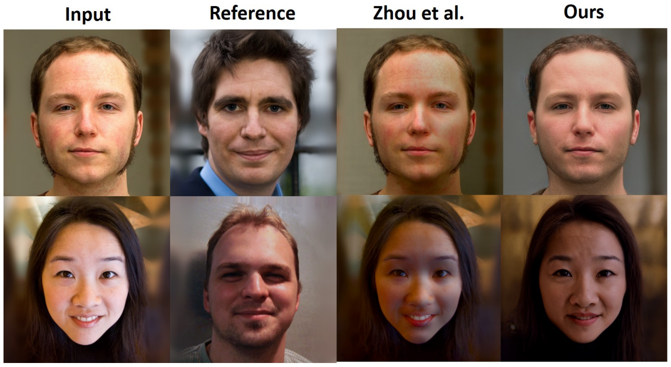

Fig. 9. Comparison of our relighting results with Zhou et al. [2019]. The illumination in the reference image is transferred to the input. Our results are more natural and achieve more accurate relighting. We can edit colored illumination while Zhou et al. [2019] can only edit monochrome light. In addition, we can also edit the head pose and facial expressions, while Zhou et al. [2019] is trained only for relighting. Images from Shih et al. [2014].

In the case of self-reenactment, we provide the input image as the first driving image. We do the same here, which leads to the two driving images with different identities. This significantly alters the facial identity in the output image. We also quantitatively evaluate the extent of identity preservation for different methods using a facial recognition tool [King 2009], see Table. 3. All methods other than ours do not support semantically disentangled editing. As can be seen in Fig. 8 (bottom), other methods simultaneously change the expressions in addition to the head pose.

*Interactive User Interface.* While all existing approaches need a driving image(s) for editing, we allow for explicit editing, using intuitive controls. We developed an interactive user interface to edit images, see supplemental video. The user can change the head pose using a trackball mouse interface. Spherical harmonic coefficients and blendshape coefficients are changed using keyboard controls. All editing results run at around 5fps on a TITAN X Pascal GPU.

*Relighting.* We compare our relighting results to the single-image relighting approach of Zhou et al. [2019], see Fig. 9. Our approach allows for colored illumination changes, as shown in Fig. 4. Our approach produces higher-quality and more realistic output images. We also quantitatively compare the relighting quality of these approaches in an illumination transfer setting, where the illumination in a reference image is transferred to a given input image. Since we do not have ground truth data available, we compare the results using a network which predicts the illumination from the reference and the relighted results. We use a model-based face autoencoder [Tewari et al. 2017], trained on the VoxCeleb dataset [Chung et al. 2018]. This network predicts a 27 dimensional spherical harmonics coefficients. We compare the predictions using a scale-invariant $\ell_2$-loss. We obtain higher quality (0.34), compared to Zhou et al. [2019] (0.36). The numbers are averaged over 100 relighting results. While the method of Zhou et al. [2019] is only trained for relighting, our method allows us to also edit the head pose and facial expressions.

## 5.4 Generality of the embeddings

*Sequential Editing.* Our method also allows for sequential editing of the different semantic parameters, see Fig. 10. Here, we optimize for the embedding using the pose RigNet network. After editing the pose, we can use the new embedding as input to the illumination and expression RigNets. Since all three versions of RigNet were trained on the same training data, this still produces plausible results.

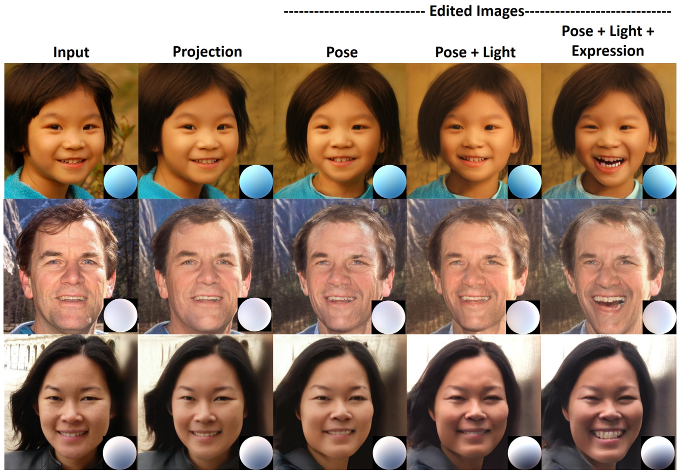
---------------------------- Edited Images----------------------------
Input
Projection
Pose
Pose + Light
Pose + Light + Expression

Fig. 10. PIE also allows for sequential editing. We optimize for the StyleGAN embedding using the pose RigNet. We can then use the edited pose results with the RigNets for other semantic components for sequential editing. Images from Shen et al. [2016].

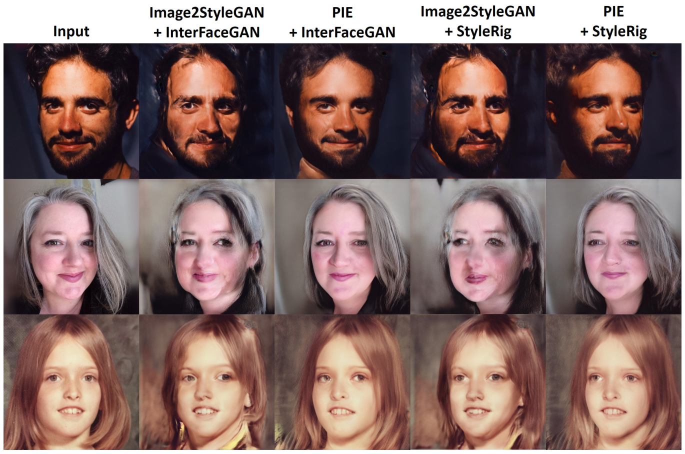

Fig. 11. Our embeddings obtain similar quality editing results with the InterFaceGAN [Shen et al. 2020] editing approach. We also notice similar improvements over Image2StyleGAN [Abdal et al. 2019] embeddings. Images from Shen et al. [2016].

*Other StyleGAN editing methods.* Our approach obtains a StyleGAN embedding which can be edited using StyleRig. In order to test the generality of these embeddings, we attempt to edit them using InterFaceGAN [Shen et al. 2020], see Fig. 11. Our improvements over Image2StyleGAN generalize to InterFaceGAN editings. We better preserve the facial identity and produce fewer artifacts. The editing results with InterFaceGAN are of a similar quality to those obtained using StyleRig. However, InterFaceGAN cannot change the scene illumination.

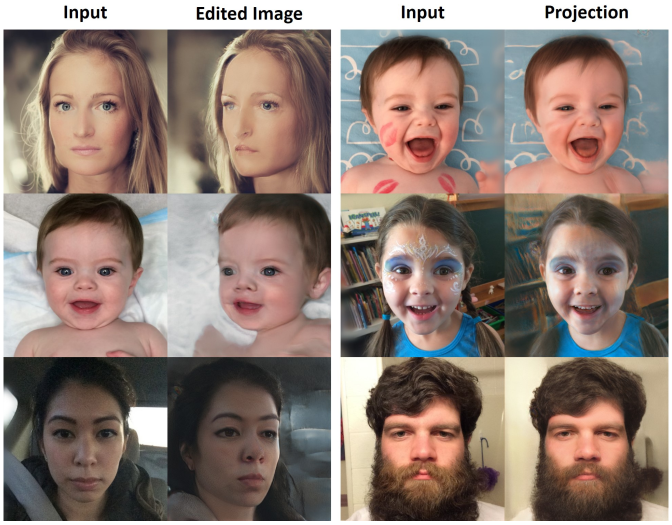

Fig. 12. Limitations: Large edits can lead to artifacts. High-frequency texture on the foreground or background is difficult to fit. Our method also cannot handle cluttered backgrounds or occlusions. Images from Shen et al. [2016].

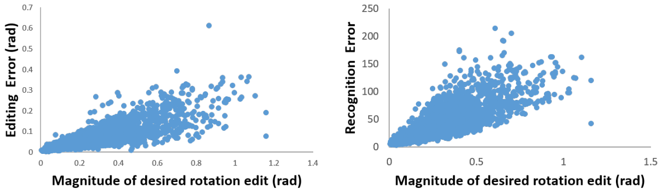

Fig. 13. Scatterplot of the editing (left) and recognition errors (right), with respect to the magnitude of the desired pose edits for over 2500 pose editing results. Larger edits lead to both higher editing and recognition errors.

## 6 LIMITATIONS

Even though we have demonstrated a large variety of compelling portrait image editing results, there is still room for further improvement of our approach: (1) At the moment, our approach has a limited expressivity, i.e., it does not allow the artifact-free exploration of the whole parameter space of the underlying 3D morphable face model. For example, we cannot change the in-plane rotation of the face or arbitrarily change the lighting conditions. The main limiting factor is the training corpus (FFHQ [Karras et al. 2019b]) that has been used to pretrain the StyleGAN-generator, since it does not contain such variations. Due to the same reason, our approach is also not yet suitable for video-based facial reenactment, since the variety of facial expressions in the training corpus is severely limited. This problem could be alleviated by pretraining the generator on a larger and less biased training corpus that covers all dimensions well. (2) Our method only allows for independent control over the semantic parameters, which is important for editing applications. While sequential control is possible, simultaneous control is a more challenging problem. (3) Our approach does not provide explicit control over the synthesized background. At the moment, the background changes during the edits and does not remain static as it should, since the network has learned correlations between the face and the background. This could potentially be alleviated by learning an explicit foreground-background segmentation and having a consistency loss on the static background region. (4) In challenging cases with large deformations, cluttered backgrounds or occlusions and high-frequency textures, our method can fail to faithfully fit to the input image and preserve editing properties at the same time, see Fig. 12. In addition, 3D face reconstruction also often fails under occlusions which would lead to incorrect data for our approach. (5) Larger edits generally correspond to worse results, and can often lead to artifacts, as shown in Fig. 12. This can also be seen in Fig. 13, where larger pose edits correlate with higher editing and facial recognition errors. (6) Similar to StyleGAN, our approach also sometimes shows droplet-like artifacts. This could be alleviated by switching to a higher quality generator architecture, such as StyleGAN2 [Karras et al. 2019a], which has been shown to solve this problem. (7) While we show results for people of different ethnicities, genders and ages, we did not extensively study the biases present in the method. Some of the components used, such as the 3DMM are known to have racial biases [Tewari et al. 2018]. (8) Our results are not guaranteed to be temporally consistent. While we show temporal editing results (in the supplemental video), our results could be made even more consistent by employing a temporal network architecture and space-time versions of our losses. Nevertheless, our approach, already now, enables the intuitive editing of portrait images at interactive frame rates.

## 7 CONCLUSION

We have presented the first approach for embedding portrait photos in the latent space of StyleGAN, which allows for intuitive editing of the head pose, facial expression, and scene illumination. To this end, we devised a hierarchical optimization scheme that embeds a real portrait image in the latent space of a generative adversarial network, while ensuring the editability of the recovered latent code. Semantic editing is achieved by mapping the control space of a 3D morphable face model to the latent space of the generator. In addition, a novel identity preservation loss enables to better preserve the facial identity.

Our approach is a first step towards intuitive and interactive editing of portrait images using a semantic control space akin to computer animation controls. In addition, our approach provides more insights into the inner workings of GANs, since it allows the intuitive and interactive exploration of the space of face images. This can shed light on the biases the model has learned from the employed training corpus. By using high-quality 3D face models, approaches such as StyleRig would produce better quality with more fine-grained control, and thus would further improve our results. Our paper brings the two different domains of 2D and 3D face models together, thus opening the road towards even more interesting edits.

## ACKNOWLEDGMENTS

We thank True-VisionSolutions Pty Ltd for providing the 2D face tracker. We thank Rameen Abdal for kindly providing the Image2StyleGAN code, Jalees Nehvi for helping us with the comparisons, and Gereon Fox for the video voiceover. This work was supported by the ERC Consolidator Grant 4DReply (770784), and the Max Planck Center for Visual Computing and Communications (MPC-VCC). We also acknowledge support from Technicolor and InterDigital.

## REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. http://tensorflow.org/ Software available from tensorflow.org.

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?. In *The IEEE International Conference on Computer Vision (ICCV)*.

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020a. Image2StyleGAN++: How to Edit the Embedded Images?. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. 2020b. StyleFlow: Attribute-conditioned Exploration of StyleGAN-Generated Images using Conditional Continuous Normalizing Flows. arXiv:2008.02401 [cs.CV]

Oleg Alexander, Mike Rogers, William Lambeth, Jen-Yuan Chiang, Wan-Chun Ma, Chuan-Chang Wang, and Paul Debevec. 2010. The digital emily project: Achieving a photorealistic digital actor. *IEEE Computer Graphics and Applications* 30, 4 (2010), 20–31.

Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F. Cohen. 2017. Bringing Portraits to Life. *ACM Transactions on Graphics (Proceeding of SIGGRAPH Asia 2017)* 36, 6 (2017), 196.

Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. 2018. Recycle-GAN: Unsupervised Video Retargeting. In *ECCV*.

Volker Blanz and Thomas Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*. 187–194.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2014. FaceWarehouse: A 3D Facial Expression Database for Visual Computing. *IEEE TVCG* 20, 3 (2014), 413–425.

J. S. Chung, A. Nagrani, and A. Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *INTERSPEECH*.

L. A. Gatys, A. S. Ecker, and M. Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2414–2423. https://doi.org/10.1109/CVPR.2016.265

Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. 2018. Warp-guided GANs for single-photo facial animation. *ACM Trans. Graph.* 37 (2018), 231:1–231:12.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. GANSpace: Discovering Interpretable GAN Controls. arXiv:2004.02546 [cs.CV]

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*.

Tero Karras, Samuli Laine, and Timo Aila. 2019b. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2019a. Analyzing and Improving the Image Quality of StyleGAN. *CoRR* abs/1912.04958 (2019).

H. Kim, M. Elgharib, M. Zollhöfer, H.-P. Seidel, T. Beeler, C. Richardt, and C. Theobalt. 2019. Neural Style-Preserving Visual Dubbing. *ACM Trans. on Graph. (Proceedings of SIGGRAPH-Asia)* (2019).

Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep Video Portraits. *ACM Trans. on Graph. (Proceedings of SIGGRAPH)* 37, 4 (July 2018), 163:1–163:14.

Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. 2017. Deep Photo Style Transfer. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 6997–7005.

Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhoefer, 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 Transactions on Graphics (Proceedings SIGGRAPH)* 38, 4. https://doi.org/10.1145/3306346.3323027

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. 1–12. https://doi.org/10.1145/3272127.3275075

Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition. In *British Machine Vision Conference*.

Pieter Peers, Naoki Tamura, Wojciech Matusik, and Paul Debevec. 2007. Post-production Facial Performance Relighting Using Reflectance Transfer. *ACM Trans. Graph.* 26, 3 (July 2007).

Ravi Ramamoorthi and Pat Hanrahan. 2001. An Efficient Representation for Irradiance Environment Maps. In *SIGGRAPH's Computer Graphics and Interactive Techniques*. 497–500. https://doi.org/10.1145/383259.383317

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. *arXiv* (2019).

Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. 2009. Face Alignment through Subspace Constrained Mean-Shifts. In *Proc. ICCV*. 1034–1041.

Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. 2011. Deformable Model Fitting by Regularized Landmark Mean-Shift. *IJCV* 91, 2 (2011).

Ahmed Selim, Mohamed Elgharib, and Linda Doyle. 2016. Painting Style Transfer for Head Portraits using Convolutional Neural Networks. (2016), 129:1–129:18.

Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. 2016. Deep automatic portrait matting. In *European conference on computer vision*. Springer, 92–107.

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. Graph.* 33, 4, Article 148 (July 2014), 14 pages. https://doi.org/10.1145/2601097.2601137

Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. 2017. Portrait Lighting Transfer Using a Mass Transport Approach. *ACM Trans. Graph* 36, 4 (July 2017).

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)*.

Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

Tiancheng Sun, Jonathan Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. 2019. Single Image Portrait Relighting. https://doi.org/10.1145/3306346.3323008

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

Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. 2018. Self-supervised Multi-level Face Model Learning for Monocular Reconstruction at over 250 Hz. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ayush Tewari, Michael Zollöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. 2017. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred Neural Rendering: Image Synthesis using Neural Textures. *ACM Trans. on Graph. (Proceedings of SIGGRAPH)* (2019).

Justus Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. 2016. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *CVPR*.

Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. 2019a. Few-shot Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2019b. Video-to-Video Synthesis. In *Proc. NeurIPS*.

O. Wiles, A.S. Koepke, and A. Zisserman. 2018. X2Face: A network for controlling face generation by using images, audio, and pose codes. In *European Conference on Computer Vision*.

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor S. Lempitsky. 2019. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. *CoRR* abs/1905.08233 (2019). arXiv:1905.08233 http://arxiv.org/abs/1905.08233

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* (2012).

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

Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. 2020. In-domain GAN Inversion for Real Image Editing. In *Proceedings of European Conference on Computer Vision (ECCV)*.