Rendering with Style: Combining Traditional and Neural Approaches for High-Quality Face Rendering

PRASHANTH CHANDRAN*, DisneyResearch|Studios, Switzerland and ETH Zurich, Switzerland SEBASTIAN WINBERG*, DisneyResearch|Studios, Switzerland GASPARD ZOSS, DisneyResearch|Studios, Switzerland and ETH Zurich, Switzerland JÉRÉMY RIVIERE, DisneyResearch|Studios, Switzerland MARKUS GROSS, DisneyResearch|Studios, Switzerland and ETH Zurich, Switzerland PAULO GOTARDO, DisneyResearch|Studios, Switzerland DEREK BRADLEY, DisneyResearch|Studios, Switzerland

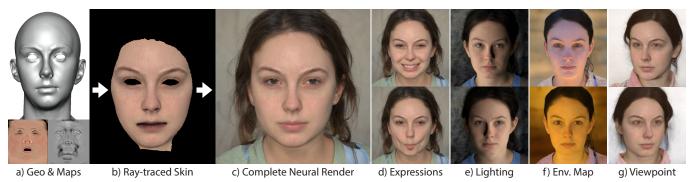


Fig. 1. Given a 3D face geometry and appearance maps (a), we perform traditional ray-tracing on skin pixels (b), and then project the result into a neural image generator network to inpaint non-skin pixels (c) resulting in a high quality render that matches the geometry and skin appearance. Our method can robustly render animations of facial expression (d), lighting changes (e), different environment maps (f) and viewpoint animations to some extent (g). Here, (d) through (g) were created independently and show different hairstyle, clothing and backgrounds from column to column, with consistency within each column.

For several decades, researchers have been advancing techniques for creating and rendering 3D digital faces, where a lot of the effort has gone into geometry and appearance capture, modeling and rendering techniques. This body of research work has largely focused on facial skin, with much less attention devoted to peripheral components like hair, eyes and the interior of the mouth. As a result, even with the best technology for facial capture and rendering, in most high-end productions a lot of artist time is still spent modeling the missing components and fine-tuning the rendering parameters to combine everything into photo-real digital renders. In this

*Indicates equal contributions by these authors.

Authors' addresses: Prashanth Chandran, DisneyResearch|Studios, Switzerland and ETH Zurich, Switzerland, prashanth.chandran@disneyresearch.com; Sebastian Winberg, DisneyResearch|Studios, Switzerland, sebastian.winberg.-nd@disneyresearch.com; Gaspard Zoss, DisneyResearch|Studios, Switzerland and ETH Zurich, Switzerland, gaspard.zoss@disneyresearch.com; Jérémy Riviere, DisneyResearch|Studios, Switzerland, jeremy.riviere12@gmail.com; Markus Gross, DisneyResearch|Studios, Switzerland and ETH Zurich, Switzerland, gross@disneyresearch.com; Paulo Gotardo, DisneyResearch|Studios, Switzerland, paulo.gotardo@disneyresearch.com; Derek Bradley, DisneyResearch|Studios, Switzerland, derek.bradley@disneyresearch.com

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM. 0730-0301/2021/12-ART223 \$15.00 https://doi.org/10.1145/3478513.3480509

work we propose to combine incomplete, high-quality renderings showing only facial skin with recent methods for neural rendering of faces, in order to automatically and seamlessly create photo-realistic full-head portrait renders from captured data without the need for artist intervention. Our method begins with traditional face rendering, where the skin is rendered with the desired appearance, expression, viewpoint, and illumination. These skin renders are then projected into the latent space of a pre-trained neural network that can generate arbitrary photo-real face images (StyleGAN2). The result is a sequence of realistic face images that match the identity and appearance of the 3D character at the skin level, but is completed naturally with synthesized hair, eyes, inner mouth and surroundings. Notably, we present the first method for multi-frame consistent projection into this latent space, allowing photo-realistic rendering and preservation of the identity of the digital human over an animated performance sequence, which can depict different expressions, lighting conditions and viewpoints. Our method can be used in new face rendering pipelines and, importantly, in other deep learning applications that require large amounts of realistic training data

Additional Key Words and Phrases: Face Rendering, Neural Rendering, GAN Inversion

with ground-truth 3D geometry, appearance maps, lighting, and viewpoint.

ACM Reference Format:

Prashanth Chandran, Sebastian Winberg, Gaspard Zoss, Jérémy Riviere, Markus Gross, Paulo Gotardo, and Derek Bradley. 2021. *Rendering with Style:* Combining Traditional and Neural Approaches for High-Quality Face Rendering. *ACM Trans. Graph.* 40, 6, Article 223 (December 2021), 14 pages. https://doi.org/10.1145/3478513.3480509

1 INTRODUCTION

The creation of digital humanoid characters continues to play a dominant role in film and video game productions. Today's advanced techniques greatly facilitate the creation and rendering of digital humans, increasing their popularity also in episodic content. With the advent of immersive virtual environments, AR/VR and the current push for virtual telepresence, there has never been a greater need for methods that can capture and render realistic digital humans efficiently, while requiring little artist effort.

Leveraging the latest advances in multi-view imaging and computational photography, an important body of research on facial scanning and performance capture has given rise to powerful methods for creating digital content that includes high-resolution 3D scans of an actor's face, with corresponding appearance textures for high-fidelity skin rendering. However, getting skin to render realistically is only part of the process, as other facial attributes such as eyes, the interior of the mouth, facial and scalp hair are all just as important to convey realism. Unfortunately, filling in the missing parts in 3D scans remains a tedious task that requires many work hours from skilled artists to obtain a high-quality digital human.

An attractive alternative to the traditional modeling and rendering pipeline is the recent advent of so-called neural rendering techniques [Tewari et al. 2020c], which can synthesize complete photo-realistic images without explicitly modeling the underlying scene properties nor its complex light transport process. Instead of using hand-crafted rendering models, neural rendering bypasses the traditional graphics pipeline and exploits deep learning techniques to encapsulate the complexity of the rendering task into a learned, data-driven rendering module. In particular, image generators based on Generative Adversarial Networks (GANs) [Goodfellow et al. 2014] have gained attention from the wider graphics and vision community, for their ability to learn the image formation model from large image datasets and then create new images that are indistinguishable from real ones. Along these lines, style-based image generators [Karras et al. 2019, 2020] have rapidly grown in popularity for their ability to synthesize a complete portrait of a human face with extremely high-fidelity. Follow up work has shown that these synthetic face images can be controlled by traversing the latent space of the network [Abdal et al. 2021; Härkönen et al. 2020; Shen et al. 2020], allowing semantic manipulation of attributes like head pose, illumination and facial expressions. It is even possible to project real face images into the latent space of these generators [Abdal et al. 2019, 2020; Xia et al. 2021; Zhu et al. 2020b], opening up a host of editing applications for portrait images. However, in comparison with traditional rendering, this neural approach offers significantly less control, in particular when the task is to render a specific identity in a desired expression with a given viewpoint and illumination conditions - the typical problem in computer graphics. The problem is further complicated when one wishes to render not just a single image but a consistent video of facial animation (e.g., performances or even simple camera movements) with facial attributes that change gradually and consistently across subsequent images, in a controllable manner. Despite the progress so far in both fields, there is still a large gap between the precise but time-consuming

controllability of traditional face rendering and the efficient but difficult-to-control neural approaches.

In this paper we take a step towards bridging this gap, as we propose a new neural rendering pipeline driven by traditional, highquality facial performance capture and skin appearance rendering. Our goal is to render multiple full-head portraits from realistic but incomplete facial scans, bypassing the burden of explicitly creating a complete 3D head asset with hair, eyes, and inner mouth. Our method starts by rendering the facial skin of the digital actor in high quality (without the complexity of modeled hair, eyes, and inner mouth) in the desired expressions and scene configurations. We then perform an optimization to simultaneously project these partial renders into the latent-space of a pre-trained image generator (StyleGAN2 [Karras et al. 2020]), obtaining full-head renders including the hair, eyes, and mouth interior. The neural rendered attributes are finally composited with the high-quality facial skin renders, yielding complete portrait images. Importantly, we take special care to optimize over all frames with a regularization approach that aims at consistently inpainting across a sequence of images, to preserve the face identity and scene properties over time. As in the traditional animation pipeline, artists have the familiar full level of control over skin appearance, facial expression, lighting and camera viewpoint, while the neural rendering step allows for realistically inpainting the missing areas of the rendered face mesh automatically. As a result, the input sequence of skin renders is transformed into a sequence of realistic, full-head portraits of a digital human (Fig. 1).

Going beyond face rendering in the field of entertainment, our work offers a second major benefit in the field of data-driven machine learning. Deep learning approaches for facial reconstruction [Feng et al. 2020; Lattas et al. 2020; Lin et al. 2020] and facial recognition [Wang and Deng 2021] rely on high quality labeled datasets of facial images. Obtaining such datasets of photo-real faces with ground truth labels (e.g. 3D geometry, appearance, pose, and lighting) is incredibly challenging. Our new neural rendering approach is ideally suited to help alleviate this problem. We demonstrate this ability by applying our rendering technique on a face model built from hundreds of state-of-the-art 3D face scans with high-quality appearance textures [Chandran et al. 2020]. This allows us to automatically create an unlimited number of photo-realistic portrait images with corresponding ground-truth skin geometry and appearance maps, besides known camera and lighting parameters, all of which can be used in training neural networks for downstream applications like face reconstruction or recognition.

2 RELATED WORK

We first review traditional methods for face modeling and rendering, and then discuss the recent explosion of neural rendering research.

2.1 Traditional Face Modeling And Rendering

Modeling a human face for photo-realistic rendering is a challenging task, as even the slightest inaccuracy could flag the entire render as uncanny. Over the last two decades, advances in computer vision have helped bootstrap face modeling, by providing artists with automatically generated, high-resolution 3D scans of faces, obtained

from multi-view imagery of an actor under passive illumination [Beeler et al. 2010, 2011; Bradley et al. 2010]. These approaches however are targeted at reconstructing the skin surface only and fail to properly reconstruct facial hair, eyes, teeth or the intrinsic appearance properties of the face, such as albedo, skin oiliness, roughness and translucency due to subsurface scattering. Some targeted methods were able to reconstruct eyes [Bérard et al. 2016, 2014], teeth [Velinov et al. 2018; Wu et al. 2016], and hair [Beeler et al. 2012; Hu et al. 2017, 2014] but unifying everything to a single complete digital human remains challenging, and computing photorealistic renders requires knowledge of the complex appearance properties of all the components. When it comes to the skin, a solid body of research has focused on facial appearance modeling and acquisition, following the seminal work of Debevec et al. [2000], which paved the way for a large body of techniques using active illumination in light stages [Fyffe et al. 2011; Ghosh et al. 2011, 2008], using flash photography [Fyffe et al. 2016], multiplexed illumination [Fyffe et al. 2016; Gotardo et al. 2015] or more recent passive methods [Gotardo et al. 2018; Riviere et al. 2020]. As with 3D scanning, these appearance capture techniques are limited to capturing skin and thus cannot accurately estimate geometry and light scattering properties of a complete human head, including inner mouth, eyes, or complex structures like hair. Nevertheless, our approach is to leverage these well-defined methods for high quality facial skin modeling and rendering, and we propose to build on top of these techniques by combining traditional rendering with recent advances in neural face rendering.

2.2 Neural Rendering

Even if traditional methods do exist for capturing, modeling and rendering all the components of a digital face, putting them all together for a photoreal digital human requires a lot of manual work by skilled artists. For this reason, researchers are turning to deep learning and using neural networks to circumvent the traditional rendering pipeline with its modeling requirements and complex simulations of light transport across multiple facial components.

Deep face rendering: An early source of inspiration comes from the fact that some complex light transport effects such as subsurface scattering can be modeled as simple filtering on the image plane [Jimenez et al. 2009]. Since convolutional neural networks (CNNs) are particularly well suited for estimating appropriate convolution kernels from training data, "deep shading" approaches began to emerge [Nalbach et al. 2017]. Another important step towards rendering realistic digital humans came from advances in image-toimage translation using U-Nets, which can be trained to translate a rendering of a human face or body into a more realistic image that closely matches reference real images [Martin-Brualla et al. 2018]. The deferred neural renderer of [Thies et al. 2019] further improved generalization over appearance properties and novel view synthesis by completely dropping the initial hand-crafted shaders and learning a texture of neural features, which is sampled onto the image plane of the desired camera view using a coarse 3D geometry estimate; the final step feeds the projected texture into the image-to-image translation CNN (the neural renderer). Other similar approaches have also been proposed that do not require a 3D

mesh as proxy geometry and, instead, learn neural features for point clouds [Aliev et al. 2020] or 3D grids of neural voxels [Lombardi et al. 2019]. Neural radiance fields (NeRFs) also leverage neural volume rendering by tracing a ray into a particular scene [Mildenhall et al. 2020]. These neural approaches can render full human heads (and even full bodies) with impressive realism. However, they are limited to rendering a particular person in a particular lighting environment, while their training requires an extensive set of images of that person across different views and body poses. A relightable neural renderer was proposed by Meka et al. [2020]; it was trained on an even larger image dataset, captured in a light stage under a single directional light at a time. An encoder was also trained that outputs a neural texture for a new person not seen during training; however, this encoder does not have the same fidelity of an optimized, personspecific neural texture and still requires input images and proxy geometry captured in a light stage, under controlled illumination.

Face image synthesis using GANs: Another line of neural rendering research explores Generative Adversarial Networks (GANs) [Goodfellow et al. 2014] that can be trained in an unsupervised way over very large face image datasets, without requiring any proxy 3D geometry. These networks can learn powerful implicit models that produce realistic human portraits with both male and female faces of different identities, ethnicities, ages, expressions, viewpoint, lighting, hair styles, accessories (glasses, ear rings) and backgrounds [Choi et al. 2018, 2020; Huang et al. 2017; Karras et al. 2018, 2019, 2020; Shen et al. 2018; Tang et al. 2018]. Some of these 2D image generators do have an explicit understanding of the underlying 3D scene [Chan et al. 2021; Nguyen-Phuoc et al. 2019; Schwarz et al. 2020], but do not yet provide the same visual quality of other image generators. Overall, as researchers soon realized, these pre-trained networks can be effectively turned into "neural morphable face models" when paired with a projection algorithm that optimizes for the network's input parameters as to approximate the appearance of a given real image [Abdal et al. 2019, 2020; Zhu et al. 2020b,a]. Among these pre-trained models, StyleGAN [Karras et al. 2019] and StyleGAN2 [Karras et al. 2021, 2020] have stood out not only due to their representative power and realism, but also due to the large degree of disentanglement in the different dimensions of the latent codes. Several recent papers have explored the latent space of StyleGAN and StyleGAN2 using facial attribute classifiers as to identify "editable" latent dimensions corresponding to semantically meaningful attributes such as facial pose, lighting, expressions, gaze, gender, age, glasses, hair and beard styles [Abdal et al. 2021; Härkönen et al. 2020; Shen et al. 2020; Shen and Zhou 2020; Wu et al. 2020]. Knowledge of such dimensions has then allowed for manipulating the latent code as to produce high-level semantic edits on both synthetic and (projected) real images with unprecedented ease and realism. In a non-face setting, Zhang et al. [2021] investigate the latent space of a pre-trained generative model by training an additional network with a differentiable renderer to identify latent dimensions that control viewpoints. Fine-grain control with such approaches is still lacking, as it is often difficult to edit a specific facial attribute without unintentional side-effects on other attributes such as facial identity. The StyleRig approach of Tewari

et al. [2020a] proposes controlling the portraits generated by Style-GAN by translating more intuitive edits applied on a 3D deformable face model (3DMM). However, only synthetic images generated by StyleGAN can be manipulated, and not real ones. In addition, the simple renderings of their 3DMM are devoid of skin detail and do not provide enough constraints on facial identity. Their subsequent work [Tewari et al. 2020b] describes an optimization approach for editing a real image with a pre-trained StyleRig network, using the real image to anchor the editing results via an identity preservation loss. Still, the demonstrated results show the expression edits are limited to adding a smile, pose remains nearly frontal, and lighting smooth. While we consider a different application and indeed find that StyleGAN2 limits viewpoint manipulation, we show results in a more varied range of expressions and lighting conditions with consistent identity. Kowalski et al. [2020] also approach the concept of semantically controlling real images by using intuitive parameters borrowed from computer graphics. They train their own GAN with two separate encoders (for real and synthetic data) with a shared latent space and a single image decoder. Garbin et al. [2020] project non-photorealistic face renders into the StyleGAN2 latent space to obtain more realistic full head portraints. Although their work is somewhat similar to ours, it has a different goal: they wish to approximately control the output of the generative network but allow the resulting face to deviate from the input non-realistic appearance; we provide final photo-realistic skin renders and only seek to consistently inpaint the missing components (hair, eyes, inner mouth, and background) from the StyleGAN2 output. Pernuš et al. [2021] recently explored a masked optimization scheme to obtaining more spatial control over the projection of individual frames into the StyleGAN2 latent space. Another stream of recent work explores training an encoder network that receives an input image and predicts the latent vector that approximates the input as closely as possible through a pre-trained generator. Such a StyleGAN2 encoder [Richardson et al. 2021] was extended by [Tov et al. 2021] to also allow for the semantic manipulation of the provided image. With such encoder, the recent *PhotoApp* method [Mallikarjun et al. 2021] achieves good consistency on viewpoint and lighting manipulations on a real face image, at the expense of requiring a large multiview, light stage training dataset for supervised learning. Alaluf et a. [2021] later identified that simple StyleGAN2 encoders are suboptimal in their ability to faithfully reconstruct the input image and proposed an iterative encoding scheme that improves reconstruction quality. Overall, previous work has focused on projecting and editing individual complete images. In contrast, our method focuses on inpainting the missing areas in skin renders and optimizes for simultaneous neural renderings under artist-controllable expression, camera viewpoint, and illumination. This is done while still matching the desired skin render and promoting a temporally consistent identity and configuration in the synthesized components.

3 RENDERING WITH STYLE

This section describes *Rendering with Style*, our hybrid face rendering approach that combines traditional, high-quality renderings

of incomplete facial scans and inpainting using a pre-trained neural face model. *Rendering with Style* allows for the generation of photorealistic sequences of full-head human portraits.

We consider a scenario in which high-quality 3D facial geometry and appearance maps are available a priori, either through capture methods using a state-of-the-art photogrammetry system [Gotardo et al. 2018; Riviere et al. 2020], or otherwise synthesized programmatically or artistically. These assets are typical results available when creating a digital human, but typically only represent the facial skin surface. We expect the skin geometry to be defined by a mesh with UV-parameterization defining the space of appearance maps such as diffuse and specular albedo, specular roughness, and high-frequency geometric detail from displacement maps. We also assume that a 3D facial blendshape model, another common asset used in facial animation, has been precomputed. Thus, using standard ray-tracing software, this 3D face skin mesh can already be rendered quite realistically in different facial expressions, viewing points, lighting conditions, and appearance parameters, with the familiar level of control. However, this model is still missing hair, ears, eyes, inner mouth (teeth, gums, tongue), all of which are important attributes that typically demand many more hours of work to generate a complete head of a realistic digital human asset.

Here, instead of continuing down the traditional digital human pipeline, we propose an alternative approach based on neural rendering. We leverage the fact that powerful image-based face models such as StyleGAN2 [Karras et al. 2020], pre-trained on very large face datasets, capture high-level semantics and correlations across the different elements of the human head. We thus leverage these correlations to generate full human head portraits from our high-quality 3D skin renders that, although incomplete, do encode rich information on facial identity, its many attributes, and the surrounding lighting environment. We highlight that our method can also be used without modification with the more recent Alias-Free Style-GAN2 [Karras et al. 2021], which is even better suited for animation. Our optimization energies are also largely generator agnostic, enabling our method to readily leverage future advances in GANs.

The following sections present a detailed description of the two main stages in our hybrid rendering pipeline: traditional rendering for the skin pixels (Section 3.1), and neural projection using Style-GAN2 as our face model for the rest of the image (Section 3.2). The final result will be a composite of the two steps, keeping the best from both approaches (Section 3.4).

3.1 Traditional Rendering

In the first step of our pipeline, we render high-fidelity facial skin geometry, consisting of high-resolution 3D meshes with 4K displacement maps that capture fine geometric detail down to pore level. The data also includes appearance maps (4K albedo and specular intensity) acquired using recent capture methods [Beeler et al. 2011; Riviere et al. 2020]. Following Riviere et al. [2020], we render skin as a two-layer model where the top layer is described by a Cook-Torrance [Cook and Torrance 1981] microfacet BRDF model covering a diffuse layer where we model subsurface scattering through diffusion, following the texture-space technique of d'Eon et al. [2007]. We further render a mask which covers only



Fig. 2. Our processing pipeline starts by generating high-quality skin renders via the traditional rendering pipeline with corresponding masks that indicate pixel targets for the rendering loss of the neural projection step.

parts of the face that correspond to skin and follow the approach described in Karras et al. [2019] to align the render to a 2D canonical space prior to projecting into the latent space of StyleGAN2. Fig. 2 shows an example input to our neural projection procedure, which we describe in the next section. Later, in Fig. 13, we also discuss how the quality of the data and the rendering itself can affect the visual quality of the final results.

Neural Projection

Given an input sequence with K high-quality rendered images I_k , depicting the skin surface of a particular person, our goal now is to generate a new sequence of neural projection images

$$P_k = \text{StyleGAN2}(\mathbf{x}_k) \approx I_k, \qquad k \in 1, 2, \dots, K,$$
 (1)

using a pre-trained StyleGAN2 generator as the image formation model of our full-head portraits P_k . Following the natural analysisby-synthesis approach for fitting morphable face models to images, we optimize the sets of StyleGAN2 input parameters \mathbf{x}_k (see below) as to generate facial images whose skin patches resemble those of our high-quality physically-based renderings. We thus explore the correlations learned by StyleGAN2 to plausibly and realistically inpaint the face elements that are missing in each I_k .

We focus on the scenario in which all the I_k depict the same person and, therefore, seek to preserve with high fidelity not only the facial identity, but also the (so far unconstrained) missing facial elements and surrounding features, which must be generated in a semantically consistent way over the output sequence P_k . These requirements rule out the straightforward naive approach of optimizing for each projection P_k independently, for two main factors: (1) even the optimization of a single StyleGAN2 latent code is a nontrivial, nonlinear optimization problem that can lead to different local optima, corresponding to very different inpainted areas (see Section 4.2 and Fig. 14 for an illustration); and (2) a naïve, greedy projection strategy inevitably overfits the highly-detailed skin renders in each I_k and, as a side-effect, small spurious correlations learned by StyleGAN2 introduce inconsistencies into the unconstrained inpainted areas.

To generate a sequence of full-head portraits with inpaintings that are semantically consistent and realistic, we therefore propose a novel optimization procedure that projects all the input renderings

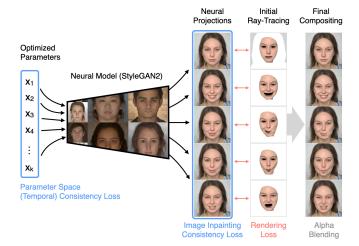


Fig. 3. The multi-frame neural projection step uses a pre-trained Style-GAN2 network as a morphable face model to realistically inpaint the missing attributes of an initial sequence of ray-traced images (with optional background cues). To avoid overfitting and facilitate consistent inpainting, projections are not required to exactly match the skin renders. The final compositing step embosses the full-detail raytraced skin appearance on top of the projection results.

 I_k simultaneously, while also enforcing additional constraints on the projections P_k and on their associated set of optimization parameters \mathbf{x}_k (Fig. 3). We formulate our search for the optimal set of image parameters X as an energy minimization problem over the entire input image sequence $I = \{I_1, I_2, \dots, I_K\},\$

$$\min_{\mathbf{X}} E_{rend}(\mathbf{X}, \mathbf{I}) + E_{cons}(\mathbf{X}), \quad \mathbf{X} = \{\mathbf{x}_k, \ k = 1, 2, \dots, K\}. \quad (2)$$

This problem comprises not only the usual data term with rendering constraints $E_{rend}(X, I)$, masked by the skin pixel mask, but also an inpainting consistency energy $E_{cons}(X)$ that largely operates in the nullspace of $E_{rend}(\mathbf{X}, \mathbf{I})$, as detailed in Section 3.2.2 and Section 3.2.3.

3.2.1 Parameterization

To derive an adequate parameterization for this non-trivial, nonlinear optimization problem, we first note that we do not require that the projections P_k exactly match the skin renders I_k , as the rendered skin appearance is restored by our final blending step (Section 3.4). Thus, our main goal here is the inpainting of missing portrait elements, which must look realistic and correlate well with the rendered skin patches. And to maintain realism, we must ensure that our solutions remain in a well-behaved location of the StyleGAN2 parameter space (although this generator has been pretrained on a large number of human images, it has been found to also generate unrealistic faces and even cat faces [Zhu et al. 2020b]). Further as we shown in Fig. 15, an unconstrained projection of partial renders into StyleGAN2 can result in unrealistic inpaintings.

For the aforementioned reasons, we model each parameter vector \mathbf{x}_k using convex linear combinations of N known, latent basis vectors \mathbf{b}_n that are randomly sampled in a well-behaved region of the StyleGAN2 latent space. Before the optimization, we sample basis vectors in the initial Z-space and feed each one through the

different MLPs within StyleGAN2, at each resolution level, to obtain basis vectors \mathbf{b}_n in the final S-space. As in Karras et al. [2020], we also apply truncation to remain in a well-behaved region near the origin. These pre-generated basis vectors are the N columns of a basis matrix B (similar to [Garbin et al. 2020]). Here, we further split matrix **B** (respectively, each \mathbf{b}_n) uniformly into many fixed-size segments \mathbf{B}_c of contiguous rows, $c \in \{1, ..., C\}$. We then model

$$\mathbf{x}_{k} = \begin{bmatrix} \mathbf{B}_{1} & & & \\ & \mathbf{B}_{2} & & \\ & & \ddots & \\ & & & \mathbf{B}_{C} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_{k1} \\ \boldsymbol{\alpha}_{k2} \\ \vdots \\ \boldsymbol{\alpha}_{kC} \end{bmatrix}, \text{ s.t. } \begin{cases} \alpha \geq 0, \ \forall \alpha \in \boldsymbol{\alpha}_{kc} \\ \|\boldsymbol{\alpha}_{kc}\|_{1} = 1, \ \forall kc \end{cases}, (3)$$

where each $\alpha_{kc} \in \mathbb{R}^N$ has weights of a convex linear combination that represents a segment of \mathbf{x}_k . Thus, solving for our sequence of projections X corresponds to optimizing for K weight vectors $\boldsymbol{\alpha}_k = [\boldsymbol{\alpha}_{k1}, \boldsymbol{\alpha}_{k2}, \dots, \boldsymbol{\alpha}_{kC}] \in \mathbb{R}^{NC}.$

This new representation with C partitions allows us to control the number of degrees of freedom in our parameterization α_k and its expressibility (more per-segment weights); it allows us to reach a balance between exploring parameter correlations within each block B_c , while also benefiting from the good semantic disentanglement across the different B_c , which define good building blocks of solutions with high realism. For our experiments in Section 4, we always sampled a set of N = 64 random basis vectors and partitioned them into 64 segments per resolution layer of StyleGAN2 (C = 1152 segments in total). Segment lengths do not change during optimization. Although future work could investigate better segmentation strategies, we empirically found this strategy to provide good expressiveness and also realistic solutions in well-behaved regions of the StyleGAN2 latent space. An ablation study on N and C is presented in Section 4.2.

During optimization, the weights α_{kc} of each segment are passed through a softmax function before applying them on their corresponding basis vectors. This ensures that the blended segment of \mathbf{x}_k is always within the convex hull of the basis segments. Although the total number of (softmax) weights in each vector α_k seems large, in practice very few of them are non-zero (roughly 6 per segment α_{kc}). In the following, for simplicity of notation, we define our optimization energies only in terms of X, x_k , P_k and I_k .

As mentioned, our basis vectors \mathbf{b}_n are defined in the S-space of StyleGAN2. Optimizing X in S-space is preferable due to its excellent level of feature disentanglement and fine control, as recently shown by Wu et al. [2020] for the manipulation of individual projections. In our case, computing the K projections in S-space allows some parameter segments to change per target I_k and better fit the individual skin renders, while other segments that represent inpainted areas can be consistently constrained across all K projections. In contrast to previous work, our projection method does not require hierarchical optimization over different spaces of StyleGAN2 to achieve good convergence [Abdal et al. 2020, 2021; Tewari et al. 2020b].

To introduce spatial variability and detail, StyleGAN2 adds random noise maps to its intermediary feature channels at different resolutions. These perturbations affect the generated images locally, for example changing a smooth hairstyle to a more frizzy one. In

our optimization, a single set of spatial noise maps (at different resolutions) is shared by all the generated projections P_k . In contrast to previous work, here we can simply sample these maps randomly and do not need to fit them to the input images. The one exception is for animations that contain camera viewpoint changes, where the fixed 2D detail generated by the noise maps would be inconsistent with the 3D projections. Thus, for these examples with camera motion we disable the noise component. Note that rendered skin areas still maintain their full level of detail due to our final, compositing step.

3.2.2 Rendering Energy

As already noted, we do not require StyleGAN2 to match with high fidelity the unique identity features seen in our high-quality input skin renders; optimization is thus focused on guiding the inpainting of the missing parts onto the neural projected image P_k , without overfitting the traditional renders. After optimization, a final compositing step is performed to emboss all the fine detail of I_k onto the resulting, complete portrait P_k (Section 3.4).

To guide inpainting in P_k using the information available in the ray-traced input RGB images I_k , we use a combination of the popular LPIPS perceptual loss [Zhang et al. 2018] and a face segmentation loss derived from [Yu et al. 2020],

$$E_{rend}(\mathbf{X}, \mathbf{I}) = \lambda_{rend} \sum_{k} \| M_k (\Phi(I_k) - \Phi(P_k)) \|_F^2 +$$

$$\lambda_{seg} \sum_{k} \| M_k (\Psi(I_k) - \Psi(P_k)) \|_F^2.$$
(5)

$$\lambda_{seg} \sum_{k} \| M_k \big(\Psi(I_k) - \Psi(P_k) \big) \|_F^2. \tag{5}$$

Here, $\Phi(\cdot)$ denotes the set of feature activations from layers *conv1-1*, conv1-2, conv2-2, conv3-3 of a pre-trained VGG-16 network [Simonyan and Zisserman 2015]; M_k denotes the masking of only those features corresponding to rendered skin patches in each I_k . To generate projections P_k with a better alignment of the contours of the eyes and mouth regions, we derive a loss term in Eq. 5, based on the activations of the final feature layer $\Psi(\cdot)$ of a face segmentation network, before its last softmax layer [Yu et al. 2020, 2018]. This additional term helps substantially improve the alignment of facial features especially the lips as we demonstrate in in Section 4.2. This additional segmentation term may also be used to control the spatial layout of the inpainted terms [Pernuš et al. 2021]. The weights λ_{rend} and λ_{seq} are used to balance the strength of these rendering energies relative to the consistency constrains below, to avoid overfitting.

Optionally, to provide a simple mechanism for controlling the inpainting of the background, we also allow the first rendering term in Eq. 4 to include small areas with render targets for background pixels. These per-image background constraints can come from simple scribble lines, or from parts of existing images (i.e., the lat-long environment map used to render the skin targets), thus providing additional information on background visibility and even lighting that is useful to guide the inpainting during the optimization (see Fig. 7 and Fig. 9 for examples).

3.2.3 Inpainting Consistency Energy

Since the rendering energy above still leaves some of the inpainted areas largely underconstrained and susceptible to spurious correlations in StyleGAN2, this section derives additional constraints that operate mainly in the nullspace of $E_{rend}(\cdot)$ and promote consistency across the inpainted areas of the sequence of projections P_k .

The inpainting consistency constraints comprise different terms in both the StyleGAN2 S-space and on the projected image plane,

$$E_{cons}(\mathbf{X}) = \lambda_{mean} \sum_{k} \|\mathbf{x}_{k} - \bar{\mathbf{x}}\|_{2}^{2} +$$

$$\lambda_{temp} \sum_{k} \|\mathbf{x}_{k} - \mathbf{x}_{k-1}\|_{2}^{2} +$$

$$\lambda_{inpt} \sum_{k} \|\widetilde{M}_{k} (\Phi(P_{k}) - \Phi(\bar{P}))\|_{F}^{2}.$$
(8)

$$\lambda_{temp} \sum_{k} \|\mathbf{x}_{k} - \mathbf{x}_{k-1}\|_{2}^{2} +$$
 (7)

$$\lambda_{inpt} \sum_{k} \| \widetilde{M}_{k} (\Phi(P_{k}) - \Phi(\bar{P})) \|_{F}^{2}. \tag{8}$$

The first loss term in Eq. 6 promotes consistency by minimizing the variance of each style parameter of the vectors \mathbf{x}_k , with $\bar{\mathbf{x}}$ denoting the mean vector. When the sequence of projections has a well-defined temporal ordering (i.e., animation), a non-zero weight λ_{temp} is used in Eq. 7 to penalize differences between temporal neighbors. This constraint enforces inpainting consistency by effectively minimizing the length of the path from the first through the last animation frames in latent space.

The third term in Eq. 8 specifically enforces consistency of the inpainted areas (hair, eyes, teeth, etc) and is applied on the image plane, where we can better specify the spatial extent of the consistency constraints. To better tolerate small in-plane motion, we apply the same LPIPS perceptual loss on the inpainted areas, as described for Eq. 4. This term is disabled ($\lambda_{inpt} = 0$) when the camera view (and head pose) changes significantly throughout the projections (when we expect large in-plane motion of inpainted areas) and when different per-frame background targets are provided in the rendering term in Eq. 4. The anchoring target \bar{P} for the inpainted areas is automatically generated by first computing a rough solution X with a single parameter segment (C = 1 in Eq. 3). From this first solution, we compute the mean projection \bar{P} , generated from the mean $\bar{\mathbf{x}}$ in latent space, and then penalize variations from this average inpainting on the image plane using a complement mask M_k , which does not affect the raytraced skin pixels.

Note that the parameterization in Section 3.2.1 already guarantees that our solutions remain in a well-behaved region of the latent parameter space (no other regularization term is needed to prevent drifting towards unrealistic face projections).

3.3 **Optimization Details**

Our multi-frame neural projection was implemented in pyTorch using an Adam optimizer. To optimize over arbitrarily long sequences of K projections, we store our global set of parameters **X** in a $K \times N \times C$ tensor, with N = 64 basis vectors and C = 1152 parameter segments. This tensor is optimized over multiple epochs just as when training a regular neural network. At each iteration in an epoch, we retrieve a small temporal window (batch) of consecutive projections \mathbf{x}_k to be feed forward through the StyleGAN2 generator, obtaining images that are subject to our rendering and consistency constraints. This strategy allows us to optimize X without constraining the sequence length K or running into GPU memory bottlenecks. For very long sequences, computation time is the main drawback. The experiments in Section 4 were all run on a single 1080Ti GPU, taking on average 60 seconds per input frame, with a pre-trained

StyleGAN2 model at resolution 1024 × 1024. Optimization was run for 200 iterations with a batch size of 2 projections and learning rate lr = 0.1. The different weights of our energy terms were set as: $\lambda_{rend} = 1, \lambda_{seq} = 0.01, \lambda_{mean} = 0.0001, \lambda_{temp} = 0.0001, \lambda_{inpt} = 0.1.$

3.4 Final Compositing

Although StyleGAN2 is a very powerful portrait generator, it still cannot faithfully reproduce the person-specific, high-frequency detail in the input skin renders I_k . For this reason, the optimization procedure described in Section 3.2 is designed to compute neural projections P_k that match each I_k only closely enough as to provide high-quality, realistic inpaintings for the areas that are missing on each I_k . After optimization, small discrepancies still exist between the desired skin appearance on I_k and the neural face render P_k .

To obtain final images I_k^* that combine the original skin renders and the inpainted areas, our method leverages the set of well-studied and well-defined methods for traditional face rendering, instead of trying to replace them. Thus, this final step in our rendering pipeline uses traditional compositing to blend the details from the original render I_k onto the neural projection P_k ,

$$I_{L}^{*} = (G * \hat{M}_{k})I_{k} + (1 - G * \hat{M}_{k})P_{k}.$$
(9)

Here, *G* denotes a Gaussian filter that is convolved on the skin mask \hat{M}_k to yield an alpha matte that has ones within the rendered skin area, zeros outside, and blended smoothly at the borders. For better blending, \hat{M}_k is the result of first applying morphological erosion (30 steps of 1 pixel) on the original skin mask M_k , illustrated in Fig. 2, before the Gaussian blur. We provide an evaluation of how close the neural projection result matches the target face render in Section 4.2, along with an illustration of the quality improvement achieved with this compositing step (Fig. 19).

RESULTS AND EVALUATION

Here and in the supplemental video, we showcase the results of Rendering with Style starting with different examples of neural projection rendering (Section 4.1), followed by an evaluation of the algorithm with ablations (Section 4.2). Finally we demonstrate how our method can be used to generate realistic, synthetic training data for deep learning applications (Section 4.3).

4.1 Neural Projection Rendering

We first demonstrate the power of our rendering approach and its ability to generate complete digital human renders for many faces in Fig. 4. Thanks to the generative power of StyleGAN2 combined with the compositing of rendered skin, the resulting faces match the identity, viewpoint and illumination of the traditional render, but are automatically completed with plausible hair, eyes and backgrounds.

Animating Expression. An important component of our approach is that we can render multiple frames in a consistent way. Fig. 5 illustrates several facial expressions, rendered with consistent identity including hair, eyes and background. Fig. 6 shows further examples of different subjects. Note that we include only one of the target renders for context, but all expression renders were used in our joint optimization across frames. We encourage the reader to view the



Fig. 4. Our method can automatically complete ray-traced faces (top) to create photo-realistic face renders that match diverse target skin identities (bottom).

supplemental video for more examples of consistent full portrait renderings under facial performance animations.

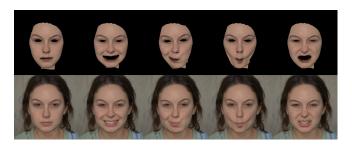


Fig. 5. Our multi-frame neural projection method yields consistent inpaintings across different expression frames. In addition, thanks to the final compositing step, the results also match the target high-detail skin renders.

Rendering with Style allows the typical levels of control available in traditional face rendering, for example, we can change the environment map to create a novel lighting condition, different from the captured studio lighting. Several results are shown in Fig. 7 for different people with animated expressions under different lighting environments. The figure also illustrates how the inpainting of the background is guided by parts of the image used as environment map when rendering the skin.

Animating Illumination. Now, consider the goal of rendering a consistent identity that matches the scene illumination, while the illumination changes. Fig. 8 shows three different illumination scenarios, with light coming from the left, front, and right. The results are photo-realistic renditions of the target digital human, consistently completed and plausibly lit from the desired light (a full animation is shown in the supplemental video). Pushing the method to more extreme scenarios, we can optimize over a sequence of renders showing more drastically varying illumination, created using wildly different environment maps. Fig. 9 shows that our method can maintain a consistent identity across frames, including all inpainted human body parts, despite the different per-frame lighting and background constraints.

Animating Viewpoint. Another parameter that is easy to control in the traditional graphics pipeline is camera viewpoint. In our

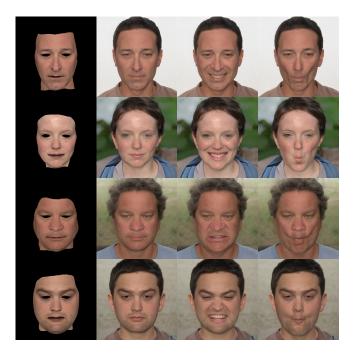


Fig. 6. Our face rendering method can produce consistent results across expression frames of different people. Here we show only the first input ray-traced render in the first column (to save space).

experiments, we have found that obtaining consistent facial inpaintings under free viewpoint variations is still very challenging with the current StyleGAN2 model. Reliably, our method can generate realistic facial renders within +/- 30 degrees. Fig. 10 shows an example of subjects rendered under different viewpoints. While the results look realistic, some temporal instability does occur (see accompanying video), and the method begins to degrade at more extreme viewpoints (around and beyond +/- 30 degrees), where the inpainted areas remain static, as visible in the video. However, there is already a considerable amount of work being done to improve current image generators and encoders [Mallikarjun et al. 2021] that can benefit our approach in the near future.

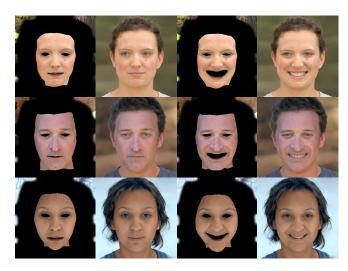


Fig. 7. Examples of consistent expression rendering under different lighting environment, for different subjects. Note that a small portion of the background is added as soft constraints to guide the neural rendering towards the image used as environment map for skin rendering.

Combined Animations. Finally, we show examples of animating multiple components in combination, for example animating the lighting or viewpoint during a performance, and simultaneously animating all three components. Some examples are shown in Fig. 11, and further illustrated in the supplemental video. It is clear that varying multiple scene properties at once does challenge the optimization, and some artifacts do start to appear, both in the form of under-fitting the desired skin renders as well as introducing minor temporal instabilities in the inpainted regions (e.g., the third row of Fig. 11). Figure 12 shows that better results are obtained when varying all scene components (expression, viewpoint and illumination), but only one at a time, still with a single optimization for a single identity. Note that for Fig. 11 and Fig. 12, the ray-traced skin renders have been omitted due to space limitations.

4.2 Evaluation

Recent work has shown that non-photorealistic renders of faces can be projected into the latent space of StyleGAN2 to approximately control the image generation process [Garbin et al. 2020]. One limitation of that approach is that the resulting face image does not closely match the rendered input geometry. Our method is designed to obtain a close match between the shape of the rendered face and the final face. Still, we require a more elaborate high-quality skin render as the input. Inspired by Garbin et al. [2020], here we first apply our method on simple OpenGL rendered faces, to determine the importance of high fidelity in the original renders. As Fig. 13 shows, the StyleGAN2 latent space is able to closely match even the non-photorealistic OpenGL renders, creating uncanny results. We conclude that our method works best with higher quality ray-traced face renders, as shown on the right of Fig. 13.

We also wish to highlight the importance of our main contribution - the ability to optimize the neural projection over all frames in

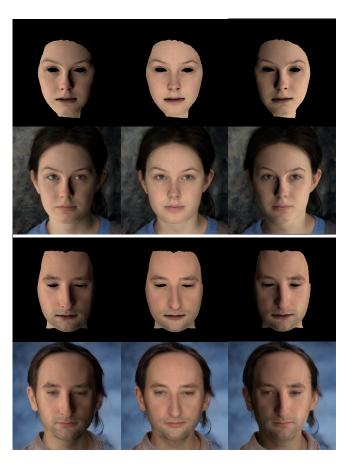


Fig. 8. Here we show that varying the lighting direction still produces a consistent identity with a realistic render, demonstrated on two subjects.



Fig. 9. We demonstrate consistent neural projection rendering under different (extreme) environment lighting.

a sequence consistently. Without this optimization, simply performing an independent projection for each frame in a sequence (e.g. following Image2StyleGAN++ [Abdal et al. 2020]) results in inconsistent renders as shown in Fig. 14. Clearly, the hair, background, clothing can all change from frame to frame, which is an unsuitable result for most applications.

Finally, we perform an ablation study to justify our design decisions and parameter value choices. As discussed in Section 3.2.1, we parameterize our problem using convex linear combinations of C

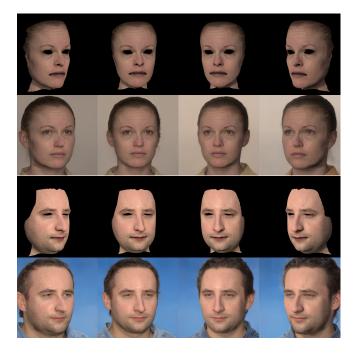


Fig. 10. Facial renders sequence with changing viewpoint for two subjects: under free viewpoint, obtaining consistent facial inpaintings is still very challenging with the current StyleGAN2 model. Our method can generate more consistent inpaintings within $\pm 10^{-2}$ degrees.

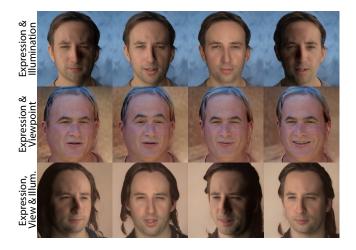


Fig. 11. Rendering with Style allows for varying multiple combinations of scene properties simultaneously in a single optimization, for example expression and illumination, expression and viewpoint, and expression, viewpoint and illumination all together. With increased variability, conflict between the energy terms also increases and results can degrade (bottom row).

segments from N known, latent basis vectors randomly sampled near the center of the StyleGAN2 latent space. One main reason for doing so is to ensure that the inpainted areas, such as eyes and teeth, remain realistic while optimizing to match the target skin pixels. In our first ablation, we compare this approach to general



Fig. 12. Simpler scenario where more consistent inpainting is obtained when varying all scene components (expression, viewpoint and illumination), but only one at a time, still with a single optimization for a single identity

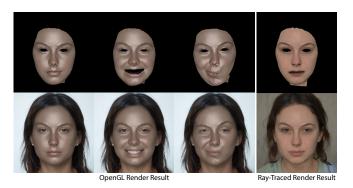


Fig. 13. Applying our method on non-photorealistic face renders yields uncanny results (first 3 columns), showing the importance of the high-quality ray-tracing component (compared in the 4th column).



Fig. 14. Optimizing each frame independently as done in related work (*e.g.*, Image2StyleGAN++) results in large inconsistencies across frames of the same sequence, whereas our method leads to consistent inpainting (Fig. 5).

unconstrained optimization in the latent space. As shown in Fig. 15, hierarchically optimizing in each of the StyleGAN2 parameter spaces (W, W+, and S) sacrifices inpainted regions in order to match the skin pixels, resulting in unrealistic eyes and inner-mouth regions. Iterating across these different spaces works well in the simpler case of projecting a *complete* image into StyleGAN2. However, when projecting incomplete renders instead of real images, the missing



Fig. 15. Unconstrained hierarchical optimization in StyleGAN2 latent spaces (W, W+, and S) results in unrealistic inpainting of the mouth and eye regions. Our convex optimization approach generates more plausible inpainting while still matching the target skin well.

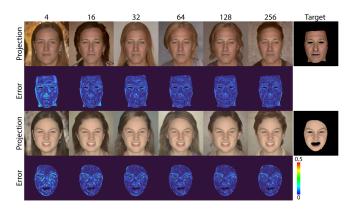


Fig. 16. We compare different basis sizes for our convex optimization approach. In practice, N=64 basis vectors produces good results with a manageable basis size. Error plots show the per-pixel norm of RGB errors, on a scale of 0-1.

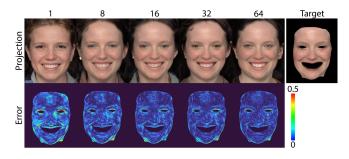


Fig. 17. Effect of splitting the parameter vector into different numbers of segments (per resolution layer): in practice, 64 segments per layer are enough to yield good fits to the rendered skin targets. Error plots show the per-pixel norm of RGB errors, on a scale of 0-1.

non-skin areas have no rendering target and remain largely unconstrained. In this case, hierarchical optimization occasionally drifts into regions of the latent space that produce unrealistic inpainting. The proposed convex optimization better leverages correlations between rendered and missing areas, providing better building blocks for more photorealistic solutions, particularly in the inpainted areas. Next we must determine the hyperparameters of this convex optimization, namely the number of basis vectors (N) to use, as

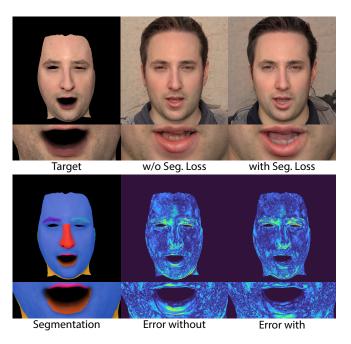


Fig. 18. Optimizing with a face segmentation loss helps the neural projection to match facial features like the mouth.

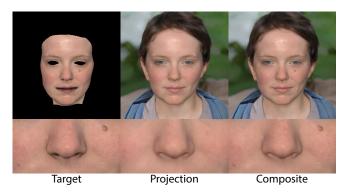


Fig. 19. Our neural projections closely match the target skin render but not exactly. Therefore, we composite the skin render with the projection in order to obtain the final face render.

well as the number of segments (*C*) to divide the parameter vector into. Fig. 16 shows the result of projecting 2 different subjects into StyleGAN2 with varying numbers of basis vectors. Given the overall quality of the projections and the skin error map, we found that N = 64 is a reasonable tradeoff of quality over basis size. Note that lower numbers of basis samples (e.g. 4, 16, 32) may predominantly contain samples from one gender, which can result in inpainting a male face with long hair and earrings as we see in the first row. Fig. 17 shows a similar evaluation of the number of segments to break the parameter vector into. We found that 64 segments per resolution layer (C = 1152) are enough to provide good fits to the skin renders, while more than 64 did not improve results. In addition, an important component of our optimization is the face segmentation

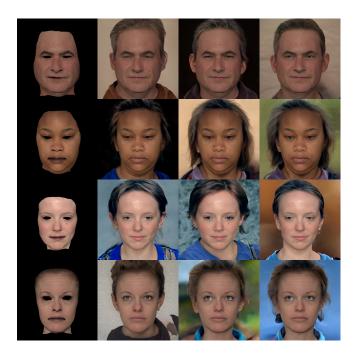


Fig. 20. Randomizing different components of the optimization process can result in different variations of neural face completion (columns 2, 3, 4) given a single ray-traced sample (column 1), which can be used as a tool for high quality synthetic data generation.

loss derived from [Yu et al. 2020] (Eq. 5). In Fig. 18 we show that optimizing with this loss helps the neural render to match facial features like the mouth region. Note that in all the ablation figures so far, we have shown pure projection results before our compositing step (Section 3.4). Despite our efforts to optimize within the StyleGAN2 space to match a desired face skin render, the goal is actually not to match it perfectly but rather generate plausible inpainting for the surrounding pixels. Fig. 19 shows how close the neural projection can match a target render, and additionally shows the final result obtained via our compositing approach, keeping the realistically inpainted pixels and the high quality traditional ray-traced pixels.

4.3 Dataset Generation

As we mentioned in Section 1, a second major application of our work is in support of large-scale dataset generation for deep learning. In particular, our method can be used to generate an unlimited number of photorealistic face images with corresponding ground-truth 3D geometry, appearance maps, viewpoint, and lighting. Such a dataset would have great value in the fields of monocular 3D face reconstruction and facial recognition under uncontrolled, in-the-wild conditions. Of course, one way to generate data samples is to vary the identity, expression, illumination and viewpoint as we have shown in Section 4.1. However, another very powerful approach is to vary the random seed used for sampling our latent basis vectors within the StyleGAN2 domain, which allows us to obtain different inpainting results even for the same identity, expression, lighting, and viewpoint combination. Specifically, with different seeds we can

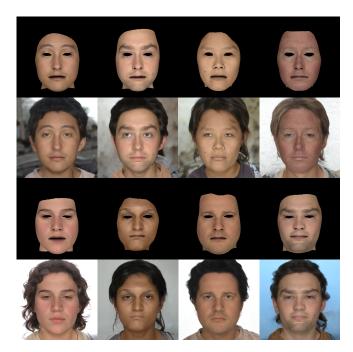


Fig. 21. Here we show that our method can render *fully synthetic*, controllable digital humans, created by a facial geometry variational autoencoder.

synthesize realistic renditions with different hairstyles, hair and eye colors, jewelry, clothing, and background, as illustrated in Fig. 20. Furthermore, our method is not limited to only captured data of real people, but can be applied to fully synthetic 3D face geometry (e.g. as generated by Chandran et al. [2020]). Fig. 21 demonstrates photorealistic renders of fully synthetic digital humans (with known ground-truth 3D geometry, appearance, viewpoint and lighting) that can help train downstream deep learning applications. As such, we believe our neural rendering method provides a valuable tool for many application domains.

5 CONCLUSION

This paper has presented *Rendering with Style*, a novel method for rendering high-quality, photorealistic digital humans, combining the high degree of controllability of traditional rendering with the representative power of a GAN. This new hybrid method leverages state-of-the-art techniques for the acquisition, modeling and rendering of skin appearance to render an incomplete face likeness in an arbitrary scene, and then project the skin renders into the latent space of a pre-trained image generator that plausibly synthesizes the missing parts. As a result, sequences of high-quality but incomplete ray-traced facial geometry are enriched with realistic hair, ears, eyes, and inner mouth areas that would otherwise require many hours of work from skilled artists to produce using traditional rendering alone.

Rendering with Style is the first method to leverage multiple, simultaneous neural projections with an optimization procedure especially designed to avoid overfitting, which if overlooked can lead to

unrealistic results with poor temporal coherency. To maintain realism, a novel parameterization is derived that provides new building blocks for estimating photorealistic solutions within a well-behaved convex subspace of the latent parameter domain. A novel rendering energy is also proposed to avoid overfitting and better align the neurally impainted and traditionally rendered face areas. Coherency across multiple neural renderings is promoted via a new inpainting consistency energy that acts on the missing image areas that are not constrained by the traditional renderings. The output is a complete, photorealistic image sequence that retains the identity of the person with artistic control on facial expression, lighting, and head pose.

Of course, Rendering with Style is not without its limitations. Like with most deep learning techniques, our approach can be limited by biases in the datasets used to pre-train the neural image model. These biases can result in sub-optimal performance for certain ethnicities, age groups, head poses, or facial expressions that are insufficiently captured by the generator's training data. More specifically, we currently rely on the StyleGAN2 model that was trained predominantly on portraits that are nearly frontal-facing. In addition, image generators such as StyleGAN2 capture 3D semantics only implicitly, making it difficult to inpaint some attributes consistently across different viewpoints (e.g., the complex occlusion and in-plane motion of long hair). We also found it challenging to consistently vary all scene components at the same time (expression, illumination and viewpoint). Regarding resolution, even though we can easily render high-resolution skin patches, the current resolution of our inpaintings is limited to the 1024×1024 spatial resolution of the last StyleGAN2 layer. Finally, our work has not yet focused on adding artistic control over the inpainted areas, such as the color of eyes, hair, and hair length, nor trying to match the inpainted regions to a photograph of a real individual. However, we believe that simple strategies can be used to alleviate this need, such as using hand-drawn scribble lines or rendered segmentation masks of simple 3D priors for regions such as hair [Hu et al. 2015], or even project a real image together with the traditional renderings to provide reference teeth, eyes, and hair style to guide the inpainting.

Nonetheless, given the current pace at which research is advancing in the fields of deep learning and neural rendering, we are confident that these limitations can be addressed in future work. For instance, StyleGAN2 has just been updated to remove aliasing artifacts and better generate animation [Karras et al. 2021], another step forward that nicely meets our goals. In addition, new image generators are being developed that do capture a more explicit understanding of the 3D scene [Chan et al. 2021]. Our approach is flexible and can readily incorporate these recent improvements.

We also expect that Rendering with Style will facilitate new advances in other applications of deep learning, such as monocular facial capture in-the-wild, by providing a means to generate a virtually unlimited amount of realistic training data with ground-truth 3D geometry, appearance, lighting, and viewpoint.

Another exciting topic for future work is to explore some of the dimensions of StyleGAN2 that have been recently identified as highly disentangled and capturing semantically meaningful facial attributes [Abdal et al. 2021]. These findings could be leveraged to further extend our hybrid rendering pipeline to allow for adding

beard, glasses and altering other facial attributes via simple edits to specific dimensions of our optimized parameter vectors.

REFERENCES

- Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2StyleGAN: How to embed images into the StyleGAN latent space?. In Proc. ICCV. IEEE, 4432-4441.
- Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. Image2StyleGAN++: How to edit the embedded images?. In Proc. CVPR. IEEE, 8296-8305.
- Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. 2021. StyleFlow: Attribute-Conditioned Exploration of StyleGAN-Generated Images Using Conditional Continuous Normalizing Flows. ACM Trans. Graphics 40, 3, Article 21 (May 2021), 21 pages. https://doi.org/10.1145/3447648
- Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. 2021. ReStyle: A Residual-Based StyleGAN Encoder via Iterative Refinement. In Proc. ICCV.
- Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. 2020. Neural Point-Based Graphics. In European Conference on Computer Vision (ECCV). Springer.
- Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. 2010. High-Quality Single-Shot Capture of Facial Geometry. ACM Trans. Graphics (Proc. SIGGRAPH) 29, 3 (2010), 40:1-40:9.
- Thabo Beeler, Bernd Bickel, Gioacchino Noris, Paul Beardsley, Steve Marschner, Robert W Sumner, and Markus Gross. 2012. Coupled 3D reconstruction of sparse facial hair and skin. ACM Trans. Graphics (Proc. SIGGRAPH) 31, 4 (2012), 1-10.
- Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W. Sumner, and Markus Gross. 2011. High-quality passive facial performance capture using anchor frames. ACM Trans. Graphics (Proc. SIGGRAPH) 30, Article 75 (August 2011), 10 pages, Issue 4.
- Pascal Bérard, Derek Bradley, Markus Gross, and Thabo Beeler. 2016. Lightweight Eye Capture Using a Parametric Model. ACM Trans. Graphics (Proc. SIGGRAPH) 35, 4. Article 117 (2016), 117:1-117:12 pages.
- Pascal Bérard, Derek Bradley, Maurizio Nitti, Thabo Beeler, and Markus H Gross. 2014. High-quality capture of eyes. ACM Trans. Graphics (Proc. SIGGRAPH) 33, 6 (2014),
- Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer, 2010, High resolution passive facial performance capture. ACM Trans. Graphics (Proc. SIGGRAPH) 29, 4 (2010), 41.
- Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. Pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis. In Proc. CVPR. 5799-5809.
- Prashanth Chandran, Derek Bradley, Markus Gross, and Thabo Beeler. 2020. Semantic Deep Face Models. In 2020 Intl. Conf. 3D Vision. 345-354.
- Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In Proc. CVPR. IEEE, 8789-8797.
- Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In Proc. CVPR. IEEE, 8188-8197.
- Robert Cook and Kenneth E. Torrance. 1981. A reflectance model for computer graphics. Computer Graphics (Proc. SIGGRAPH) 15, 3 (1981), 301-316.
- Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In ACM Trans. Graphics (Proc. SIGGRAPH). ACM Press/Addison-Wesley Publishing Co., ACM,
- Eugene d'Eon, David Luebke, and Eric Enderton. 2007. Efficient Rendering of Human Skin. In Proc. Eurographics Conf. on Rendering Techniques (EGSR'07). Eurographics Association, 147-157
- Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. 2020. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. (Dec. 2020). arXiv:2012.04012 [cs.CV]
- Graham Fyffe, Paull Graham, Borom Tunwattanapong, Abhijeet Ghosh, and Paul Debevec. 2016. Near-Instant Capture of High-Resolution Facial Geometry and Reflectance. Computer Graphics Forum 35, 2 (2016), 353–363.
- Graham Fyffe, Tim Hawkins, Chris Watts, Wan-Chun Ma, and Paul Debevec. 2011. Comprehensive Facial Performance Capture. Computer Graphics Forum 30, 2 (2011).
- Stephan J Garbin, Marek Kowalski, Matthew Johnson, and Jamie Shotton. 2020. High Resolution Zero-Shot Domain Adaptation of Synthetically Rendered Face Images. In Proc. ECCV. Springer, 220-236.
- Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. 2011. Multiview face capture using polarized spherical gradient illumination. In ACM Trans. Graphics (Proc. SIGGRAPH Asia). 1–10.
- Abhijeet Ghosh, Tim Hawkins, Pieter Peers, Sune Frederiksen, and Paul Debevec. 2008. Practical Modeling and Acquisition of Layered Facial Reflectance. ACM Trans. Graphics 27, 5 (Dec. 2008), 139:1-139:10.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. Advances in Neural Information Processing Systems 27 (2014), 2672-2680.

- Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2018. Practical Dynamic Facial Appearance Modeling and Acquisition. ACM Trans. Graphics (Proc. SIGGRAPH Asia) 37, 6 (2018), 232:1–232:13.
- Paulo Gotardo, Tomas Simon, Yaser Sheikh, and Iain Matthews. 2015. Photogeometric Scene Flow for High-Detail Dynamic 3D Reconstruction. In Proc. ICCV. IEEE, 846– 854.
- Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. GANSpace: Discovering Interpretable GAN Controls. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9841–9850.
- Liwen Hu, Derek Bradley, Hao Li, and Thabo Beeler. 2017. Simulation-ready hair capture. In Computer Graphics Forum, Vol. 36. Wiley Online Library, 281–294.
- Liwen Hu, Chongyang Ma, Linjie Luo, and Hao Li. 2014. Robust hair capture using simulated examples. ACM Trans. Graphics 33, 4 (2014), 1–10.
- Liwen Hu, Chongyang Ma, Linjie Luo, and Hao Li. 2015. Single-View Hair Modeling Using A Hairstyle Database. ACM Trans. Graphics (Proc. SIGGRAPH) 34, 4 (July 2015).
- Zhiwu Huang, Bernhard Kratzwald, Danda Pani Paudel, Jiqing Wu, and Luc Van Gool. 2017. Face Translation between Images and Videos using Identity-aware CycleGAN. arXiv:1712.00971 [cs.CV]
- Jorge Jimenez, Veronica Sundstedt, and Diego Gutierrez. 2009. Screen-space perceptual rendering of human skin. ACM Transactions on Applied Perception 6, 4 (2009), 23:1–23:15.
- Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In Intl. Conf. Learning Representations.
- Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-Free Generative Adversarial Networks. arXiv:2106.12423 [cs.CV]
- Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proc. CVPR*. 4401–4410.
- Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*. IEEE, 8110–8119.
- Marek Kowalski, Stephan J. Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. 2020. CONFIG: Controllable Neural Face Image Generation. In Proc. ECCV.
- Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. 2020. Avatarme: Realistically renderable 3d facial reconstruction 'in-The-wild'. In Proc. CVPR.
- Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. 2020. Towards high-fidelity 3D face reconstruction from in-the-wild images using graph convolutional networks. In Proc. CVPR.
- Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural Volumes: Learning Dynamic Renderable Volumes from Images. ACM Trans. Graphics (Proc. SIGGRAPH) 38, 4, Article 65 (July 2019), 14 pages
- B R Mallikarjun, Ayush Tewari, Abdallah Dib, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Louis Chevallier, Mohamed Elgharib, and Christian Theobalt. 2021. PhotoApp: Photorealistic Appearance Editing of Head Portraits. ACM Trans. Graph. (2021).
- Ricardo Martin-Brualla, Rohit Pandey, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Julien Valentin, Sameh Khamis, Philip Davidson, Anastasia Tkach, Peter Lincoln, Adarsh Kowdle, Christoph Rhemann, Dan B Goldman, Cem Keskin, Steve Seitz, Shahram Izadi, and Sean Fanello. 2018. LookinGood: Enhancing Performance Capture with Real-Time Neural Re-Rendering. ACM Trans. Graphics 37, 6, Article 255 (Dec. 2018), 14 pages.
- Abhimitra Meka, Rohit Pandey, Christian Haene, Sergio Orts-Escolano, Peter Barnum, Philip Davidson, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, Chloe Legendre, Wan-Chun Ma, Ryan Overbeck, Thabo Beeler, Paul Debevec, Shahram Izadi, Christian Theobalt, Christoph Rhemann, and Sean Fanello. 2020. Deep Relightable Textures - Volumetric Performance Capture with Neural Rendering. ACM Trans. Graphics (Proc. SIGGRAPH Asia) 39, 6.
- Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proc. ECCV*.
- O. Nalbach, É. Arabadzhiyska, D. Mehta, H.-P. Seidel, and T. Ritschel. 2017. Deep Shading: Convolutional Neural Networks for Screen Space Shading. Computer Graphics Forum 36, 4 (July 2017), 65–78.
- Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. 2019. HoloGAN: Unsupervised Learning of 3D Representations From Natural Images. In Proc. ICCV.
- Martin Pernuš, Vitomir Štruc, and Simon Dobrišek. 2021. High Resolution Face Editing with Masked GAN Latent Code Optimization. arXiv:2103.11135 [cs.CV]

- Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2021. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. In Proc. CVPR.
- Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2020. Single-shot high-quality facial geometry and skin appearance capture. ACM Trans. Graphics (Proc. SIGGRAPH) 39, 4 (2020), 81–1.
- Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2020. GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. In Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 20154–20166.
- Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. 2018. FaceID-GAN: Learning a Symmetry Three-Player GAN for Identity-Preserving Face Synthesis. In Proc. CVPR
- Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. 2020. InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs. IEEE Trans. PAMI PP (Oct. 2020).
- Yujun Shen and Bolei Zhou. 2020. Closed-form factorization of latent semantics in GANs. (July 2020). arXiv:2007.06600 [cs.CV]
- Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Intl. Conf. on Learning Representations.
- Xu Tang, Zongwei Wang, Weixin Luo, and Shenghua Gao. 2018. Face Aging with Identity-Preserved Conditional Generative Adversarial Networks. In Proc. CVPR. 7939–7947.
- Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. 2020a. StyleRig: Rigging StyleGAN for 3D Control over Portrait Images. In *Proc. CVPR*. 6142–6151.
- Ayush Tewari, Mohamed Elgharib, Mallikarjun B R., Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. 2020b. PIE: Portrait Image Embedding for Semantic Control. ACM Trans. Graphics (Proc. SIGGRAPH Asia) 39, 6 (2020).
- A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B Goldman, and M. Zollhöfer. 2020c. State of the Art on Neural Rendering. Computer Graphics Forum (EG STAR 2020) (2020).
- Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred Neural Rendering: Image Synthesis Using Neural Textures. ACM Trans. Graphics (Proc. SIGGRAPH) 38, 4, Article 66 (July 2019), 12 pages.
- Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. 2021.

 Designing an Encoder for StyleGAN Image Manipulation. arXiv preprint arXiv:2102.02766 (2021).
- Zdravko Velinov, Marios Papas, Derek Bradley, Paulo Gotardo, Parsa Mirdehghan, Steve Marschner, Jan Novák, and Thabo Beeler. 2018. Appearance capture and modeling of human teeth. ACM Trans. Graphics (Proc. SIGGRAPH Asia) 37, 6 (2018), 1–13.
- Mei Wang and Weihong Deng. 2021. Deep face recognition: A survey. Neurocomputing 429 (2021), 215–244.
- Chenglei Wu, Derek Bradley, Pablo Garrido, Michael Zollhöfer, Christian Theobalt, Markus H Gross, and Thabo Beeler. 2016. Model-based teeth reconstruction. ACM Trans. Graphics (Proc. SIGGRAPH) 35, 6 (2016), 220–1.
- Zongze Wu, Dani Lischinski, and Eli Shechtman. 2020. StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation. (Nov. 2020). arXiv:2011.12799 [cs.CV] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. 2021. GAN Inversion: A Survey. (2021). arXiv:2101.05278 [cs.CV]
- Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. 2020. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-time Semantic Segmentation. CoRR abs/2004.02147 (2020). arXiv:2004.02147
- Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proc. ECCV. 334–349.
- R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proc. CVPR*.
- Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, and Sanja Fidler. 2021. Image GANs meet Differentiable Rendering for Inverse Graphics and Interpretable 3D Neural Rendering. In Intl. Conf. on Learning Representations.
- Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. 2020b. In-Domain GAN Inversion for Real Image Editing. In *ECCV*.
- Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020a. Improved StyleGAN Embedding: Where are the Good Latents? (Dec. 2020). arXiv:2012.09036 [cs.CV]