Deep Face Normalization

KOKI NAGANO, Pinscreen
HUIWEN LUO, Pinscreen
ZEJIAN WANG, Pinscreen
JAEWOO SEO, Pinscreen
JUN XING, Mihoyo
LIWEN HU, Pinscreen
LINGYU WEI, Pinscreen
HAO LI, Pinscreen, University of Southern California, USC Institute for Creative Technologies

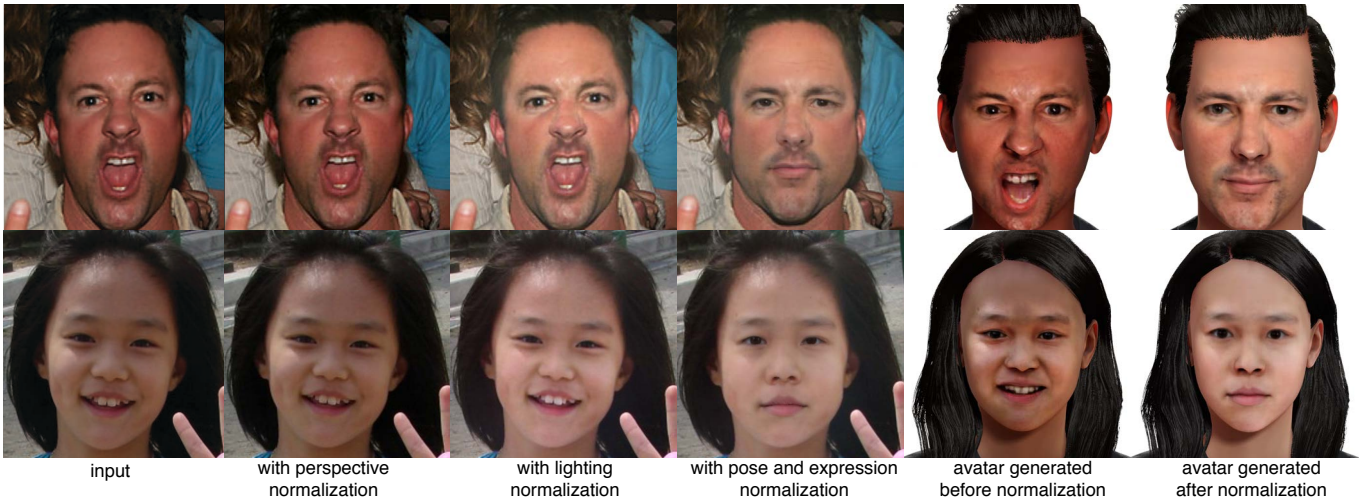

Fig. 1. We introduce a deep learning framework to normalize faces in unconstrained portraits w.r.t. perspective distortion, illumination, expression and pose. In addition to enhanced portrait manipulation capabilities, we can also generate more accurate and visually pleasing virtual 3D avatars. Original image courtesy of Bengt Nyman (top) and watchsmart/flickr (bottom).

From angling smiles to duck faces, all kinds of facial expressions can be seen in selfies, portraits, and Internet pictures. These photos are taken from various camera types, and under a vast range of angles and lighting conditions. We present a deep learning framework that can fully normalize unconstrained face images, i.e., remove perspective distortions, relight to an evenly lit environment, and predict a frontal and neutral face. Our method can produce a high resolution image while preserving important facial details and the likeness of the subject, along with the original background. We divide this ill-posed problem into three consecutive normalization steps, each using a different generative adversarial network that acts as an image generator. Perspective distortion removal is performed using a dense flow field predictor. A uniformly illuminated face is obtained using a lighting translation network, and the facial expression is neutralized using a generalized facial expression synthesis framework combined with a regression network based on deep features for facial recognition. We introduce new data representations for conditional inference, as well as training methods for supervised learning to ensure that different expressions of the same person can yield to not only a plausible but also a similar neutral face. We demonstrate our results on a wide range of challenging images collected in the wild. Key applications of our method range from robust image-based 3D avatar creation, portrait manipulation, to facial enhancement and reconstruction tasks for crime investigation. We also found through an extensive user study, that our normalization results can be hardly distinguished from ground truth ones if the person is not familiar.

CCS Concepts: • **Computing methodologies** → *Computer graphics; Image manipulation.*

Additional Key Words and Phrases: Virtual Avatar, Texture Synthesis, Deep Learning, Generative Adversarial Network, Image Processing.

Authors' addresses: Koki Nagano, Pinscreen; Huiwen Luo, Pinscreen; Zejian Wang, Pinscreen; Jaewoo Seo, Pinscreen; Jun Xing, Mihoyo; Liwen Hu, Pinscreen; Lingyu Wei, Pinscreen; Hao Li, Pinscreen, University of Southern California, USC Institute for Creative Technologies.

0730-0301/2019/11-ART183 $15.00
https://doi.org/10.1145/3355089.3356568

ACM Reference Format:
Koki Nagano, Huiwen Luo, Zejian Wang, Jaewoo Seo, Jun Xing, Liwen Hu, Lingyu Wei, and Hao Li. 2019. Deep Face Normalization. *ACM Trans. Graph.* 38, 6, Article 183 (November 2019), 16 pages. https://doi.org/10.1145/3355089.3356568

1 INTRODUCTION

A picture of a person's frontal face with blank expressions, captured in an evenly lit environment, and free from perspective distortion, is not only ideal for facial recognition, but also extremely useful for a wide range of graphics applications, ranging from portrait manipulation to image-based 3D avatar digitization. While billions of portraits and selfies are shared over the Internet, people tend to smile and express their emotions in front of the camera. Pictures are mostly taken under a vast range of challenging lighting conditions, and selfies generally cause noticeable facial distortions such as enlarged noses. In the context of counter-terrorism and law enforcement, images of suspects [Federal Bureau of Investigation 2019] are often limited and highly deteriorated.

Various solutions for image-based relighting and facial alterations exist, but they typically require accurate prior knowledge about the person's face or any available scene parameters. These algorithms work best if the faces are neutral and captured under well conditioned scene illuminations. Advanced camera effects for facial expression manipulation [Averbuch-Elor et al. 2017] are also difficult to apply on these unconstrained photos, since a neutral expression is often needed that is free from deformations. Furthermore, the ability to perform proper relighting on images with harsh illuminations is nearly impossible. In particular, when images with very different focal settings are used, 3D avatar creation apps (e.g., Pinscreen [2019], Loom.ai [2019], itSeez3D [2019]) tend to produce very different looking characters [Hu et al. 2017a; Nagano et al. 2018].

We introduce a deep learning-based framework, that can fully normalize a portrait taken in the wild into a canonical passport photo-like setting with a blank facial expression. While end-to-end face normalization systems exist [Cole et al. 2017], they can only produce low resolution images, which are not suitable for high-fidelity image-based 3D avatar creation or high-resolution portrait manipulation. Furthermore, individual normalization tasks for distortion, lighting, pose, and expressions are not possible. We propose a technique that does not have these limitations.

From an unconstrained picture, our method sequentially removes perspective distortion, re-illuminates the scene with an evenly lit diffuse illumination with proper exposure, and neutralizes the person's expression. For mild head rotations, our algorithm can successfully infer a frontal face with its nearby body and hair deformations. We first normalize a perspective distorted image into a near orthographic projection by predicting a dense flow image based on a variant of [Zhao et al. 2019], followed by a global warp, and inpainting operation. Next, we fit a 3D face model to the undistorted image and use this 3D geometry as a proxy to extract auxiliary information such as the spherical harmonics (SH) coefficients of the lighting, rigid pose parameters, and UV texture. Using the input picture and the estimated scene illumination, we use a generative adversarial network [Isola et al. 2017] to synthesize a high-quality image of a face lit under even lighting such that the true skin tone is reflected. We introduce an offset-based lighting representation in order to preserve high-frequency details such as facial hair and skin textures. The final step consists of frontalizing the face and neutralizing the facial expression. The resulting face must not only be plausible and faithful to the person's identity, but a consistent neutral face must be predicted from a wide range of expressions. While photorealistic facial expression synthesis networks have been introduced recently, they are only designed to produce expressions from a neutral face and cannot neutralize from arbitrary expressions. In particular, unwanted dynamic wrinkles caused generally persist after a neutralization attempt.

We propose to learn the mapping from a range of expressions to a single neutral one. We first decompose the problem into a task for facial geometry neutralization and one for texture neutralization. To predict a neutralized face geometry, we train a regressor that can infer identity parameters of a 3D morphable face model using deep features for facial recognition. Once the neutralized geometry is obtained, we generate a neutralized face texture using this as a condition. We then train a generalized version of the recently introduced photoreal avatar generative adversarial network (paGAN) [Nagano et al. 2018] using both, expression-to-neutral and neutral-to-expression data sets. Before inference, we adopt a similar data representation as paGAN using a combination of depth and normal maps, and incorporate a new technique that can inpaint occluded regions using symmetry-aware facial textures.

To facilitate the supervised learning, we also introduce a number of data augmentation techniques to generate synthetic faces, expressions, illumination, and perspective distortion variations for our normalization tasks. In particular, we simulate distortions using a variant of [Fried et al. 2016] and produce lighting variations based on a custom skin shader with soft shadows based on directional lighting, as well as additional image processing and grading operations. For facial expression neutralization, we use a recently proposed *StyleGAN* network to generate synthetic faces to augment identity variations and to train a generalized paGAN, we create renderings of synthetic faces by blending facial geometry with a UV texture, and use a pre-trained paGAN for expression synthesis.

We demonstrate the effectiveness of our method on a vast collection of pictures with varying expressions and lighting conditions. We can even show how different facial expressions of a same person can produce plausible and very similar neutral faces. Our user study shows that people who are not familiar with the input person, have difficulties distinguishing synthesized portraits from ground truth ones. In addition to an extensive set of evaluations and comparisons, we also showcase applications such as portrait manipulation, normalized face reconstruction, image-based 3D avatar creation, and improved 3D facial animation. We make the following contributions:

- We propose the first framework that can fully normalize focal length, lighting, facial expressions, and mild poses from a single unconstrained portrait, while preserving high-resolution facial details, the likeness of the person, as well as the original background.
- We introduce a highly effective facial expression neutralization technique, that can ensure a consistent mapping from a range of expressions to a single one. The method combines

a 3D neutral face regressor based on deep features for facial recognition and a generalized paGAN synthesis model.
- We present novel GANs with custom data representations for conditional inference that can preserve facial appearance details that are person specific, under varying lighting conditions, challenging facial expressions, and poses.
- We propose new techniques for augmenting the training data with synthetic faces, expressions, lighting, shading, and 3D facial perspective distortions for effective supervised training.

2 RELATED WORK

Lens distortion control, relighting, and facial expression manipulation have been extensively investigated as separate problems in the graphics and vision community. One can argue that estimating parameters for each of these tasks using existing methods can be used to undo these transformations in order to normalize an unconstrained portrait picture. Even if accurate scene and face parameters are recoverable, the ability to synthesize a plausible and photorealistic output is still challenging due to the complexity of facial deformations and appearance changes under intricate scene captures. Furthermore, the combination of these problems increase the difficulty of a proper disentanglement. For instance, a harsh lighting condition or perspective-distorted face can significantly deteriorate the ability to restore its neutral face from one with a smile. We will review the most relevant prior work for these three problems as well as recently proposed end-to-end approaches based on deep neural network feature inversion.

Perspective Distortion. As described in [Ward et al. 2018], facial shots from various distances, can cause distortive effects on the face and have a significant impact on the perceived nasal size. To this end, both, methods for estimating the camera-subject distance from a face photograph have been introduced, as well as algorithms for manipulating those as a post-effect. [Flores et al. 2013] are the first to propose a computational approach that can recover the camera distance from a face using an efficient Perspective-n-Point algorithm. The method relies on manual placements of facial landmarks. A fully automated approach for frontal faces was later introduced by [Burgos-Artizzu et al. 2014], which uses custom facial landmarks detected using a cascaded regression technique followed by a multivariate linear regressor for camera-subject distance estimation. Even if an accurate distance estimation is possible, we are interested in a way of undistorting the perspective of a given image.

[Bas and Smith 2018] recently presented a single-view 3D face modeling approach that can handle ambiguities with 2D constraints for both orthographic and perspective projections. In particular, if the camera-subject distance is known, a more accurate face shape can be estimated than without priors. While combining this modeling method with a reliable camera distance estimator could facilitate our task for face normalization, our approach consists of directly removing the perspective distortion in the input image using a deep neural network. A direct manipulation of perspective distortions on a portrait was introduced by [Fried et al. 2016]. Their technique relies on a successful 3D face model fitting, followed by a dense image warp, which is driven by the rendering of the facial distortion under varying camera-subject distances. If an initial camera distance cannot be provided and the input image exhibits significant distortion, the fitting of the 3D face model would fail, and a successful undistortion would not be possible as shown in [Zhao et al. 2019]. While we also predict a dense undistortion flow field for our output image, our proposed technique does not rely on fitting a 3D face model, and can therefore undistort an input picture without known camera distance parameters. Perspective distortion on wide-angle selfies has been explored by [Shih et al. 2019]. While their method can correct severe distortion caused by the wide field-of-view, their work does not address the effect of perspective distortion caused by subject distance.

Facial Relighting. Face relighting on photographs has been explored widely in the context of face recognition, which was first documented in [Adini et al. 1997]. A generative appearance-based technique for improved face recognition under different illuminations was later introduced in [Georghiades et al. 2001], where they successfully demonstrate renderings of input faces in novel lighting conditions using a sparse set of input images captured in a controlled setting. [Shashua and Riklin-Raviv 2001] employed a reference image with novel lighting condition to transfer its face illumination to an input picture using a ratio image-based formulation for face relighting. [Liu et al. 2001] extended this approach to handle facial expressions under illumination changes. Ratio image-based techniques have also been adopted to relight faces using radiance environment maps when the reference face shares the same albedo as the input one. Plausible delighting operations from a single input picture were demonstrated in the work of [Wang et al. 2009], in which a morphable face model is used in combination with spherical harmonics that model the illumination changes. While their method can handle extreme lighting conditions, the method only supports neutral expressions and facial details are lost since it is based on a linear face model and Lambertian shading under distant lighting.

A technique for transferring local contrasts and overall lighting directions from one portrait to another was introduced by [Shih et al. 2014]. While dramatic style transfers were shown, the method requires both input and reference pictures to have accurate 3D face models with compatible appearance attributes. In particular, beards and skin color have to look similar in order to work properly, as the algorithm would otherwise generate visible artifacts. A formulation that does not require an accurate modeling of the face was presented by [Shu et al. 2017] which solves a mass-transport problem to generate color remapping for more robust geometry-aware relighting. Both approaches are not suitable for delighting tasks since they cannot remove dark shadows, and brightening these regions would lead to significant artifacts. Further, unwanted facial details may carry over to the synthesized result.

For general photographs, [Barron 2015] introduced a highly effective approach that can auto white balance images with harsh lighting conditions. His convolutional color consistency method works by reformulating the problem as a 2D spatial localization task in a log-chrominance space using techniques from object detection and structure prediction. An improved approach using a network that can process patches with different confidence weights for color consistency estimation was later proposed by [Hu et al. 2017b]. While both approaches focus on removing illumination color casts

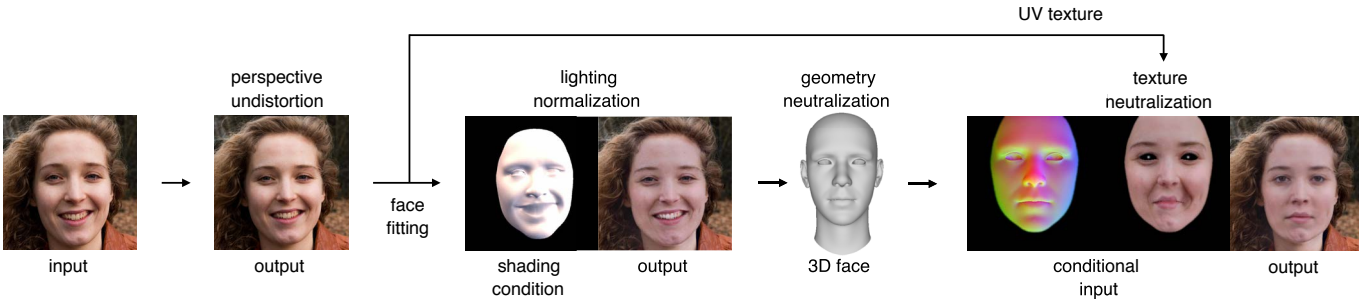

Fig. 2. System Overview. Given an unconstrained portrait, we remove the perspective distortion, predict an evenly lit face, reconstruct the neutralized geometry and facial texture, and combine all these output to generate a neutralized portrait. Original image is courtesy of Sterre Van Den Berge.

in images, they are not solving the lighting normalization problem in faces, and unwanted shadings can still remain. More recently [Sun et al. 2019] proposed a single image portrait relighting approach using a deep neural network. While their method addresses more general portrait relighting, the method requires a complex capture setup and is difficult to capture a large number of subjects for the model to be generalized.

Widely adopted facial appearance modeling techniques, such as morphable face models [Blanz and Vetter 1999] are linear and often adopt a simple Phong and Lambertian reflectance model for lighting estimation. Since the facial appearance cannot model high-frequency details and textures such as facial hair, relighting a face in unconstrained images can yield visible artifacts and unpleasant shadings. Despite the recent efforts in improving the separation and estimation of shape, reflectance, and illuminations in faces, it remains difficult to render and composite these faces on existing photographs without appearing uncanny. These intrinsic decomposition techniques include methods that use statistical reflectance priors [Li et al. 2014] as well as deep convolutional neural networks [Kim et al. 2018; Sengupta et al. 2018; Yamaguchi et al. 2018]. Our method first estimates the lighting condition using spherical harmonics, then uses an illumination-conditioned generative adversarial network to synthesize a face with normalized lighting conditions.

Facial Poses and Expressions. Morphable face models [Blanz and Vetter 1999] and many of their extensions with facial expressions, such as [Cao et al. 2014; Hu et al. 2017a; Thies et al. 2016] have enabled the modeling of fully textured 3D faces from a single input image. Frontalizing the face on a portrait can be achieved through smooth warp between the face region and the background as shown in [Cao et al. 2014] and [Averbuch-Elor et al. 2017]. For mild head rotations, the results are typically acceptable since facial textures only need to be generated in small occluded regions, while keeping background distortions minimal. [Hassner et al. 2015] shows that a simpler approach can be achieved by using a single 3D surface as opposed to a full 3D face model as approximation in order to improve face recognition capabilities. More recently, Huang et al. introduced TP-GAN [Huang et al. 2017], which can synthesize frontal views from extreme side views, using a generative adversarial network that takes advantage of symmetry and identity information. And extension of this work was presented by [Hu et al. 2018a] which can also synthesize arbitrary views. While we also adopt a GAN-based approach for frontalizing faces, we condition our generator to a dense 3D face geometry similar to [Nagano et al. 2018] which allows us to preserve high-resolution details.

Expressions can be neutralized by setting coefficients of a fitted linear facial expression model to zero [Cao et al. 2014; Hu et al. 2017a; Thies et al. 2016]. [Genova et al. 2018] propose a deep learning based approach that directly regresses morphable model parameters of neutral expressions from deep facial recognition features. However, if the appearance of a linear model is used, high-frequency facial details that are specific to the user will be lost. If we only deform the geometry of the input textures to the neutralized expression, unwanted dynamic wrinkles will persist. The deep learning-based photorealistic facial texture synthesis method of [Saito et al. 2017] could achieve more detailed results, but the generated textures may not match the original input, since they hallucinate plausible, instead of matching high-frequency details. Recently, several methods on synthesizing photorealistic facial expressions from photos have been introduced such as StarGAN [Choi et al. 2018], G2-GAN [Song et al. 2017], wg-GAN [Geng et al. 2018] and paGAN [Nagano et al. 2018]. Not only do most of these techniques require a successful initial face fitting, but they cannot ensure a plausible or consistent neutral face to be generated from a range of expressions.

Deep Neural Network Feature Inversion. The first deep learning framework that demonstrates face normalization capabilities was proposed by [Zhmoginov and Sandler 2016] in which they introduce an iterative and feed-forward technique for inverting a face recognition embedding which can be used to recover a frontal and neutral face of a person. While the likeness of the input subject is recognizable, the results are very blurry and noisy. An elegant end-to-end approach for full face normalization was recently proposed by [Cole et al. 2017], in which a frontal and neutralized facial expression is synthesized using facial identity features obtained from a facial recognition framework. The deep generative approach is extremely robust and successfully preserves the likeness of the input subject for extremely challenging scenarios. However, since it relies on globally learned identity features from a large face database, high-frequency details, such as facial hair and high-resolution skin appearances cannot be generated and the output is prone to noise and artifacts. Furthermore, the original background cannot be

preserved and individual controls for distortion, lighting, and facial expression manipulation are also not supported.

3 METHOD

Fig. 2 gives a high-level overview of our method. Due to the challenge of inferring fully normalized photos directly, we subdivide this challenging task into smaller ones, each addressed by a conditional generative adversarial network. These tasks consist of perspective undistortion, lighting normalization, pose frontalization, and expression neutralization. An additional benefit of breaking down the entire face normalization problem into sub-problems is that individual control is possible (e.g., expression neutralization without applying lighting normalization).

Given an unconstrained input image, we first transform its camera perspective into a near-orthographic one by predicting a flow field (Sec. 3.1). The flow is predicted on the face, then propagated to the background via inpainting, similar to [Fried et al. 2016]. Once the perspective distortion is removed, we perform 3D facial fitting to the input image using [Thies et al. 2016] in order to extract the lighting condition via spherical harmonics (SH) and the facial texture of the subject (Fig. 5 and Fig. 7). Notice that the fitted 3D face model from this stage is not used for the final 3D facial expression neutralization due to its sensitivity to extreme poses, occlusions and illuminations. Next, we predict an evenly lit face using a generative adversarial network (GAN) conditioned on the estimated SH shading image (Sec. 3.2) and use an offset representation to preserve high-frequency details. The light-normalized facial region is then composited back to the background using Poisson blending [Pérez et al. 2003].

A frontalized and expression neutralized 3D facial geometry is then obtained using a deep neural network for facial neutralization based on facial recognition features (Sec. 3.3). Finally, we synthesize a neutralized facial texture using another GAN generator, which is conditioned on the expression neutralized 3D facial geometry and a neutral-deformed expression texture (Sec. 3.3). This output can be used directly for our image-based 3D avatar digitization. However, if we wish to obtain a face normalized portrait, we composite the result to the background using the same warping technique used for perspective undistortion combined with Poisson blending.

Training Data. In this paper, we use the Chicago Face Dataset (CFD) [Ma et al. 2015] (597 subjects), the compound facial expressions (CFE) dataset [Du et al. 2014] (230 subjects), the Radbound Faces dataset [Langner et al. 2010] (67 subjects), and the Multi-PIE database (250 subjects) [Gross et al. 2008], which consists of a total of ~10K photographs and a wide variations of expressions and poses. Each stage uses a different set of data and we apply different data augmentation and simulation techniques for each network, which is explained in each sub-section.

3.1 Perspective Undistortion

Given a 2D input image $\mathbf{I}$, we first run landmark detection [Wei et al. 2016] to obtain 2D facial landmarks $\mathbf{L}$. Given $\mathbf{I}$ and $\mathbf{L}$, our perspective undistortion network G_{flow} predicts a dense 2D flow $\mathbf{F}$ to correct the distortion (Fig. 4):

$$\mathbf{F} = G_{flow}(\mathbf{I}, \mathbf{L}) \tag{1}$$

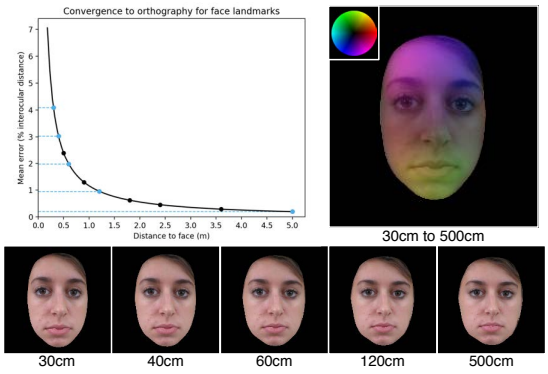

Fig. 3. Training data samples for perspective undistortion.

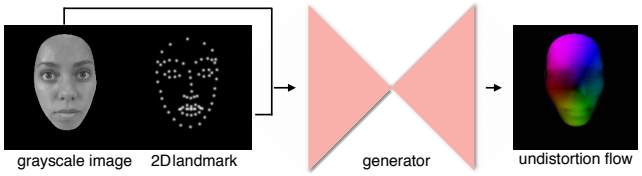

Fig. 4. Perspective undistortion pipeline.

We first convert the input image to grayscale and apply a facial mask to ignore the background. To train the network, we create dense flow fields by fitting 3D face models to input photographs with known focal length and camera-subject distance. We then simulate perspective distortion by rendering each subject with different camera-subject distances. As done in [Fried et al. 2016], the dense 2D flow field is derived by rasterizing the face model, before and after the distance manipulation given 3D mesh correspondence. Since we learn the flow field in 2D, it is more effective if we sample the training distance so that the 2D image space appearance changes evenly. To quantify the changes of the 2D warp field, we measure the mean 2D landmark distance between the perspective projections at a particular distance and the orthographic projection using a mean face of a 3D morphable face model [Blanz and Vetter 1999].

Perspective distortion is nonlinear to the camera-subject distance and focal length. In particular, it changes rapidly when the distance gets closer and/or the focal length becomes shorter. For this reason, we vary the sample rate along the distance to capture more changes in closer distances ($30cm$ to $1.2m$). In Fig. 3, we show that perspective distortion is roughly linear if the distances are sampled evenly in the vertical error scale (blue dots and the corresponding pictures in the bottom row). We sample 10 discrete distances (blue and black dots) for our synthetic training data. This procedure generates ~100K ground truth flow images for all subjects in the our training data. As seen in the graph, the perspective distortion converges nearly to an orthographic projection at $5m$. We chose this as our reference distance to warp all the input images, as if they were captured at $5m$ distance with a telephoto lens (approx. $350mm$ in $35mm$ camera).

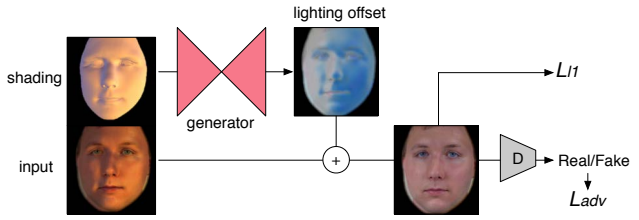

Fig. 5. Lighting normalization pipeline.

Fig. 6. Training data sample for lighting normalization.

An example of synthetic flow (from 30cm to 5m) is shown on the top right of Fig. 3.

To train the network, we employ a weighted L_2 pixel loss that measures the difference between the prediction from our U-net based generator [Isola et al. 2017] $G_{flow}(\mathbf{I}, \mathbf{L})$ and ground truth synthetic flow $\mathbf{F}_{gt}$:

$$L = \langle \mathbf{W}, \left\| \mathbf{F}_{gt} - G_{flow}(\mathbf{I}, \mathbf{L}) \right\|^2 \rangle \tag{2}$$

We accumulate the squared difference per pixel using a weight map $\mathbf{W}$, which is created by rasterizing the input 2D landmark image $\mathbf{L}$ (see Fig.4) to favor increased accuracy around the 2D facial features. We apply Gaussian blurring with a kernel size K ($K = 31$ in our experiment) to ensure smoothness of the output flow and use 10× higher weights around facial features. We also experimented with an adversarial loss as used in other normalization pipelines, but did not find it useful. In order to make the inference more robust against challenging input images, we added random brightness, contrast, and blurring during the training. Since our network is based on image-to-image translation, we found that the training is more efficient, if we estimate a flow that aligns with the input image pixels. A drawback of such forward warping is that a naive pixel-level mapping can cause holes in the target image. To properly warp all the pixels including the image background, we perform flow inpainting combined with Laplacian smoothing as done in [Fried et al. 2016]. Once the perspective distortion is removed, we perform 3D face fitting [Thies et al. 2016] to the corrected input image, and obtain a fitted 3D mesh, SH shading coefficients, and UV texture for subsequent steps.

3.2 Even Light Illumination

While spherical harmonics-based (SH) illumination models [Ramamoorthi and Hanrahan 2001] can represent real-world illuminations effectively, if the scene's reflectance is near Lambertian, the skin of human faces have generally more complex reflectance properties, such as specular reflections, subsurface scattering, cast shadows. As shown in Fig. 17, a naive shading decomposition can lead to significant artifacts. Nevertheless, we show that this SH-based shading information is sufficient in providing a coarse guide for the scene illumination when inferring an evenly lit face using a deep learning-based approach. Instead of directly using the estimated lighting condition to decouple the illumination, we perform this task using a conditional generative adversarial network for image synthesis by conditioning the inference using the estimated SH values, obtained from 3D face fitting (see Fig. 5).

To preserve important high-frequency facial details during the inference of high-resolution images, such as (pores, facial hair, specular reflections, etc.), we introduce an offset-based image representation, instead of inferring target pixel colors directly from a generator. Hence, we predict a lighting offset image $\mathbf{O}_{lit}$ by a generator G_{lit} that produces an evenly lit face, if it is added to the input image. More specifically, given a masked input image $\mathbf{I}$ and spherical harmonics shading image $\mathbf{S}$, the illumination normalized photograph $\mathbf{I}_\mathbf{l}$ is produced as

$$\mathbf{I}_{lit} = \mathbf{O}_{lit} + \mathbf{I} \tag{3}$$

where $\mathbf{O}_{lit} = G_{lit}(\mathbf{I}, \mathbf{S})$. Our experiments show that this approach is able to preserve significantly higher resolution details as shown in Fig. 18. To train our network, we produced a large volume of synthetic illumination data via portrait relighting. For each database picture in the training data, that is captured under uniformly lit white illumination, we fit a 3D morphable face model. We then use directional lighting and image-based lighting using custom OpenGL/GLSL shaders implementing soft shadows, microfacet specularities [Cook and Torrance 1982], and subsurface scattering to simulate a wide range of scene illuminations. We created 10 lighting variations (5 random directional lighting and 5 random HDR environments) per-subject, which leads to 100K training image samples in total. In order to relight eyes and teeth realistically, we created a billboard geometry for the eyes and mouth interiors, and perform inpainting on the shading image to relight partially visible hair on the face. Fig. 6 shows an example of our simulated training data for lighting augmentation. To further increase robustness, we also add random contrast and brightness perturbations to simulate poor quality input. Please refer to the supplemental material for more details.

For the training, we employ a multi-scale L1 pixel difference loss and an adversarial loss as follows:

$$L = L_{adv} + \lambda_{\ell_1} L_{\ell_1} \tag{4}$$

L_{ℓ_1} evaluates pixel differences at multiple scales to ensure globally consistent skin color estimation. Specifically,

$$L_{\ell_1} = \sum_{k=1}^{K} \left\| \mathbf{I}_{gt}^k - \mathbf{I}_{lit}^k \right\| \tag{5}$$

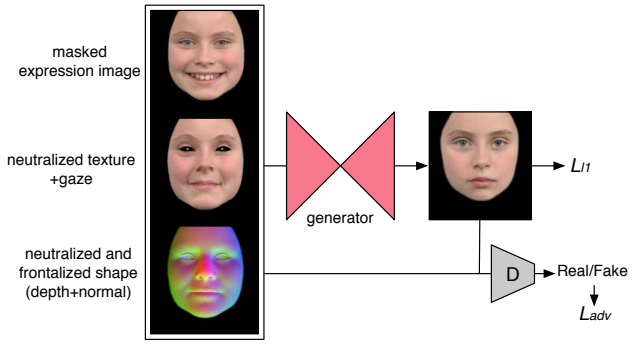

Fig. 7. Expression and pose normalization pipeline.

L_{adv} is a multi-scale adversarial loss adapted from [Wang et al. 2018].

$$L_{adv} = \mathop{\mathbb{E}}_{(\mathbf{I},\mathbf{S},\mathbf{I}_{gt})}\left[\log D_k\left(\mathbf{I},\mathbf{S},\mathbf{I}^k_{gt}\right)\right] + \mathop{\mathbb{E}}_{(\mathbf{I},\mathbf{S})}\left[\log\left(1 - D_k\left(\mathbf{I},\mathbf{S},\mathbf{I}^k_{lit}\right)\right)\right] \tag{6}$$

where $\{D_k\}_{k=1}^K$ are discriminators trained on different image scales to detect local and global artifacts. For both losses, we evaluate the error on an image pyramid [Wang et al. 2018] with $K = 2$ levels, where $\mathbf{I}^2_{\{gt,lit\}}$ are down-scaled to 1/4 width and height of the original images $\mathbf{I}^1_{\{gt,lit\}}$ (128 and 512 resolution in our experiment). We use $\lambda_{\ell_1} = 20$ for our experiments.

As our network predicts normalized appearances only inside the facial region, we perform Poisson image blending [Pérez et al. 2003] as a post-process to composite the normalized image seamlessly into the background.

3.3 Expression Neutralization

Our expression neutralization consists of geometry neutralization and facial texture neutralization, each of which is addressed by a dedicated deep neural network. After the facial expression is normalized, the face can be optionally composited to the background for portrait manipulation applications. We warp the background using 2D flow derived from 3D mesh correspondence before and after geometry normalization using the technique described in [Fried et al. 2016] and composite the face region to the background using Poisson blending [Pérez et al. 2003].

Geometry Neutralization. Despite the limited output resolution, Cole et al. [2017] introduced a highly robust method to synthesize a face with blank expressions using deep facial recognition features [Schroff et al. 2015]. While Cole et al. [2017] can infer 2D landmarks from the facial recognition features, our regressor infers neutralized full 3D models as similarly done in [Genova et al. 2018]. Let α be the identity coefficients of a linear 3D morphable model for the input $\mathbf{I}$. We train a regressor with multi-layer perceptron (MLP) layers $R(\mathbf{C})$ that takes the facial features $\mathbf{C}$ to predict α ($|\alpha| = 91$). For the facial recognition features $\mathbf{C}$, we used 1792-D vectors extracted from the last pooling layer using the Inception ResNet v1 [Szegedy et al. 2016] architecture, similar as in [Cole et al. 2017]. Training the network requires pairs of input facial recognition features and

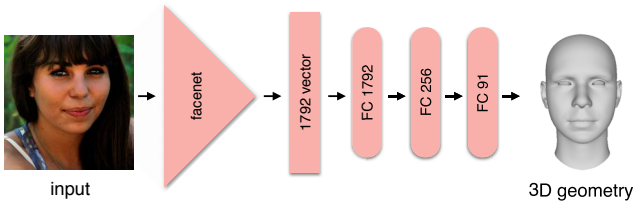

Fig. 8. Geometry neutralization pipeline. Original image is courtesy of Keniaxmargarita/flickr.

ground truth 3D geometry. We extract features from a pre-trained facial recognition network and use 3D face fitting results [Thies et al. 2016] with known camera and subject distances to produce the training data. The network is trained with the following loss:

$$L = \lambda_{pos}L_{pos} + \lambda_{land}L_{land} + \lambda_{prior}L_{prior} + \lambda_{sym}L_{sym} \tag{7}$$

L_{pos} presents the per-vertex position distance in 3D between the ground truth mesh $\mathbf{P}_{gt}$ and predicted mesh $\mathbf{P}$

$$L_{pos} = \left\|\mathbf{P}_{gt} - \mathbf{P}\right\|^2 \tag{8}$$

L_{land} is similar to L_{pos}, but measured on a subset of 68 vertex positions $\mathcal{L}$ corresponding to the facial features of [Kazemi and Sullivan 2014].

L_{sym} is a facial symmetry loss that minimizes the distortion by computing the difference of each corresponding pair of vertices $(l, r) \in \mathcal{L}$ on the left and right side of the face after flipping both to the same side.

$$L_{sym} = \sum_{(l,r)\in\mathcal{L}} \left\|\left|\mathbf{P}^l\right| - \left|\mathbf{P}^r\right|\right\|^2 \tag{9}$$

Lastly, L_{prior} accounts for the error between the predicted and ground truth blendshape coefficients:

$$L_{prior} = \left\|\alpha_{gt} - R(\mathbf{C})\right\|^2 \tag{10}$$

Our network employs three layers of MLP with Leaky ReLu nonlinearities with leakiness 0.2 (Fig. 8). In our experiments, we set $\lambda_{pos} = 2$, $\lambda_{land} = 0.01$, $\lambda_{prior} = 0.01$, and $\lambda_{sym} = 0.01$. Since our geometric loss formulation is generic, it is not limited to linear models, and more sophisticated ones can be used [Wu et al. 2018b]. While we have an immense amount of training samples, our training dataset only contains 1K unique facial identities, which can lead to overfitting during training. In order to augment the variation of unique facial identities, we introduce a way to synthesize novel identities by interpolating two identities continuously (described in the next paragraph), using features from the state-of-the-art *StyleGAN* network. Similar to [Cole et al. 2017], we only generated frontal faces of new identities with blank expressions for the data augmentation since the deep facial recognition network is robust to expressions and pose variations. We perform this identity augmentation on our training dataset and created 160K new subjects, resulting in 170K training data samples. We mixed our base training data and the augmented data with the ratio of 1 : 2 during the training (See Fig. 22 for the effect of the data augmentation).

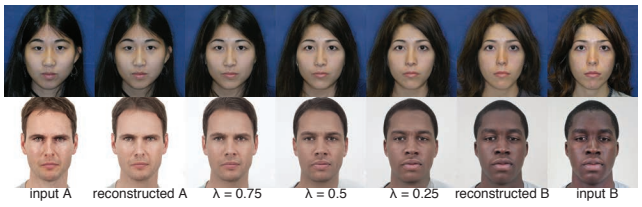

Fig. 9. StyleGAN embedding and interpolation. Real input images A and B are embeded into w_a and w_b, and their interpolations produce new synthetic faces.

Data Augmentation for Geometry Neutralization. [Karras et al. 2018] proposed an alternative generator architecture, *StyleGAN*, based on generative adversarial networks. We use *StyleGAN* to interpolate two neutral faces to synthesize new fake people in order to augment the number of neutral subjects in our dataset. We formulate this task as a latent space embedding problem. In particular, given an arbitrary image, we extract a representative latent vector w, which can be used as an input of *StyleGAN* and generate the same image. We embed two images into the latent space and obtain two latent vectors w_1 and w_2. Then, a linear function is used to interpolate the latent vector $w = \lambda w_1 + (1-\lambda)w_2$ and a new image can be generated using the new vector w. Given a real image $\mathbf{I}_r$, we first initialize a random latent vector w and $StyleGAN(w)$ to generate a random synthetic image $\mathbf{I}_f$. With a pre-trained model for perceptual loss [Johnson et al. 2016], our method minimizes the perceptual loss between $\mathbf{I}_r$ and $\mathbf{I}_f$ by freezing both generators and perceptual model weights, and optimizing w using gradient descent. In our implementation, we optimize the intermediate latent space (18 layers and each layer is a 512 vector) of *StyleGAN* and use the output layer `block4_conv2` of *VGG*-16 [Simonyan and Zisserman 2014] for the perceptual loss. We show the embedding and interpolation results in Fig. 9. In practice, we only add the mean interpolation results to the dataset and create 160K new subjects. We use *StyleGAN* augmentation to train the neutral geometry regressor.

Facial Texture Neutralization. We achieve pose and expression neutralization using a novel generalized variant of the photorealistic expression synthesis network of [Nagano et al. 2018]. Given a neutral picture of a source and target expressions, this work can synthesize arbitrary photorealistic facial expressions while maintaining person-specific identities. In our facial expression neutralization, we want the opposite effect, i.e., synthesize a photorealistic neutral expression from arbitrary facial expressions and pose of a person. Given the neutralized geometry inferred from the previous stage (Sec. 3.3), the rigid pose, and the UV expression texture from the initial face fitting process, we first frontalize the face by resetting the rotation component, and render the normal/depth image and the expression texture on the neutralized geometry to create images for conditioning the GAN (Fig. 7). While [Nagano et al. 2018] demonstrates photorealistic expression synthesis using a similar network, we found that naively providing pairs of input expressions and their corresponding neutral faces does not produce a high-quality result. Unlike facial expression synthesis from a neutral photograph, the neutralization training target needs to predict one exact neutral from a wide range of facial expressions. We conjecture that such many-to-one mapping is prone to overfitting, as perhaps the target lacks sufficient variations. Another concern is that since the input pictures exhibit a wide range of facial expressions, it is difficult for the network to extract consistent identity features.

From our experiments, we found that the network trained from a neutral input picture to a range of output facial expressions (i.e. paGAN by [Nagano et al. 2018]) is better at preserving person-specific identity features. Hence, we train a generalized version of the paGAN model by mixing both neutral-to-expression and expression-to-neutral datasets, and use only the expression-to-neutral dataset during test time. In this way, the network can learn the neutralization task (i.e. remove wrinkles, inpaint occluded areas, and synthesize plausible eyes) while better preserving the likeness of the person after inference. To train our generalized paGAN variant, we initialized the network using the original pre-trained paGAN model. In addition to the training strategy, we made an additional improvement on the conditional images. For side-facing training images, a naive facial texture computation with projection causes large visual artifacts in invisible or occluded areas. We address this by identifying invalid facial areas via facial symmetry, followed by Poisson blending and inpainting to recover from the artifacts. Our model is trained using the following loss function:

$$L = L_{adv} + \lambda_{\ell_1} L_{\ell_1} + \lambda_{Id} L_{Id} \tag{11}$$

where L_{adv} and L_{ℓ_1} are the multi-scale adversarial and pixel loss as in Sec. 3.2 and L_{Id}, an identity loss adapted from [Hu et al. 2018b] that minimizes features of the last pooling layer and fully connected layer of a pre-trained facial recognition network [Wu et al. 2018a]. In our experiment, we used $\lambda_{\ell_1} = 20$ and $\lambda_{Id} = 0.25$. Similarly, for the geometry neutralization training, in the previous section, we also perform data augmentation to train the generalized paGAN model. While the synthetic faces interpreted by *StyleGAN* are realistic with background and hair, it also provides high-frequency artifacts, which is not suitable when learning high-fidelity textures. Hence, we perform a different data augmentation to increase the identity variations by blending the 3D geometry and UV texture (see the next paragraph for details). Training a texture neutralization network requires pairs of a neutral and expression photos. To this end, we created synthetic expressions using a pre-trained paGAN model. We used 6 key expressions used in [Nagano et al. 2018]. In total this augmentation produces around 90K identities each with 6 expressions. During training, we mix synthetic faces and real photos from our training data with a ratio of 1 : 2 (See Fig. 23 for the effect of the data augmentation).

Data Augmentation for Texture Neutralization. We first synthesize fake frontal neutral faces which include both geometry and texture, and then we use paGAN to create expressions of these synthetic faces. Given a frontal face image I, we first fit a 3D morphable model to the image using the method of [Thies et al. 2016] to obtain the initial 3D mesh data $M_I = (\alpha_I, \beta_I, T_I)$, where α_I and β_I are the corresponding identity and expression coefficients. We then compute the face texture T_I which is unwrapped from I to UV-space. Given two face images A, B and their mesh data M_A, M_B, we interpolate the coefficients and textures of the two faces, independently. Given

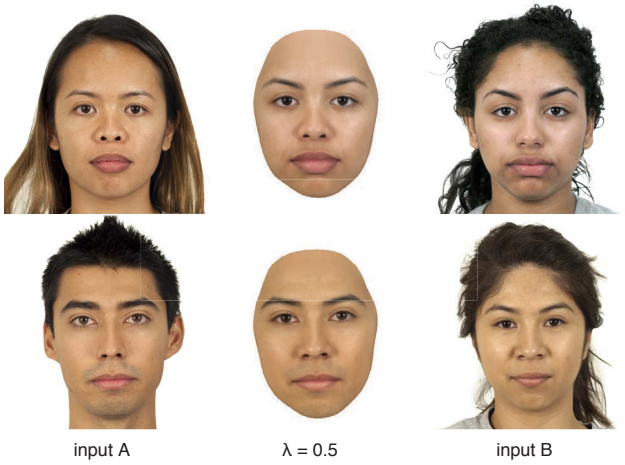

Fig. 10. Data augmentation by interpolating facial geometry and texture. A new face in the middle is generated by blending subjects A and B.

$M_A = (\alpha_A, \beta_A, T_A)$ and $M_B = (\alpha_B, \beta_B, T_B)$, a new face M_N is generated as $M_N = (\lambda\alpha_A + (1-\lambda)\alpha_B, \lambda\beta_A + (1-\lambda)\beta_B, \lambda T_A + (1-\lambda)T_B)$, with $\lambda \in [0, 1]$. Given a seed face A, we pick a target face by selecting one of the $k = 200$ nearest neighbors of A and interpolate them to obtain a new face with its 3D mesh. We use the same measurement for the distance between faces A and B as in [Cole et al. 2017]:

$$d(A, B) = \theta\|L_A - L_B\| + \|T_A - T_B\| \tag{12}$$

where L are matrices of 2D landmarks. We use $\theta = 10.0$ in our experiments. The blending results are shown in Fig. 10. We set $\lambda = 0.5$ to produce faces that are as different as possible as those we have, and ignore repetitive image pairs. While we could have used the second augmentation method for the geometry neutralization as well, we found that using the two different methods leads to better results as the second method can only synthesize the face region and cannot provide unique variations in the head shapes with background that are necessary for the face recognition network.

4 EXPERIMENTS AND RESULTS

We will showcase the results on three applications and highlight the effectiveness of our technique on a wide range of subjects of different skin tones and age, as well as extremely challenging lighting conditions, unknown camera properties, and even stylized input images. We further provide a comprehensive evaluation of individual algorithmic components, and compare our technique with the current state-of-the-art. We also refer to the accompanying video for additional results, including 3D avatar digitizations, progressive steps of perspective undistortion, and facial neutralization.

4.1 Applications

Portrait Manipulation. As our normalization techniques are modular, we can apply perspective undistortion, lighting normalization, and expression normalization individually to achieve portrait manipulation. In Figs. 1, 11 and 12, we demonstrate portrait manipulation results by sequentially applying an individual normalization component. After the lighting is normalized, the portrait can be re-lit with an arbitrary lighting condition using the proxy 3D geometry and texture obtained as part of the lighting normalization process (Fig. 12 fourth column). Here we demonstrate relighting using a directional light following the technique described in Sec. 3.2.

Single-View 3D Avatar Creation. Normalized portraits are highly suitable for image-based virtual avatar modeling tasks and are key for producing visually pleasing and high-fidelity results robustly. We produced all of our 3D avatar results using the method of [Nagano et al. 2018]. The fifth and sixth columns in Fig. 1 and Fig. 11 show 3D avatars created before and after the normalization. An undistorted input ensures accurate avatar geometry, normalized lighting produces a texture that can be re-lit with novel illuminations, and expression normalization enables correct facial animations, all of which are crucial for consumer accessible virtual avatars.

Face Reconstruction in Law Enforcement. In the context of crime investigation and counter-terrorism, there are often limited pictures of suspects or kidnapped persons, as shown for instance in the most wanted list maintained by the U.S. Federal Bureau of Investigation [Federal Bureau of Investigation 2019]. Graphical representations such as facial composites are often used to provide additional depictions on how these subjects may look like. In cases when the person is performing an expression (e.g., a smile) and the picture is taken in an uncontrolled lighting environment, we can show how a normalized face can provide additional information for identification and recognition, as shown in Fig. 13.

4.2 Performance

All of our networks are trained on a desktop machine with Intel i7-6800K CPU, 32 GB RAM and two NVIDIA GeForce GTX 1080 ti (12 GB RAM) GPUs using Pytorch. Our networks output cropped aligned face images with a resolution of 256x256 for perspective undistortion flows, and 512x512 for the other normalization outputs. Training takes 12 hours for the perspective undistortion network, 6 days for the light normalization network, 24 hours for the geometry neutralization network, and 2 days for the generalized paGAN model. All the results are produced on the same machine. The initial face fitting takes 0.5s. The prediction of undistortion flow takes 2ms and postprocess for perspective undistortion takes 5s. Light normalization inference takes 7ms and postprocess takes 2s. Expression geometry neutralization and texture inference take 1ms and 7ms, respectively. Final Poisson blending and inpainting take 5s. In total, our method takes 13s for a given portrait.

4.3 Perspective Undistortion

In Fig. 14, we show the generated avatars using the original images (left group) and the ones with perspective normalization (right group). Without perspective normalization, generated avatar shapes can exhibit a large variations, and using wrong focal lengths can produce avatars with wrong shapes (i.e. face becomes too wide or narrow). 3D face modeling can still produce correct avatar shapes using the correct focal length (indicated with black boxes), but usually it is unknown in advance for unconstrained portraits. On the

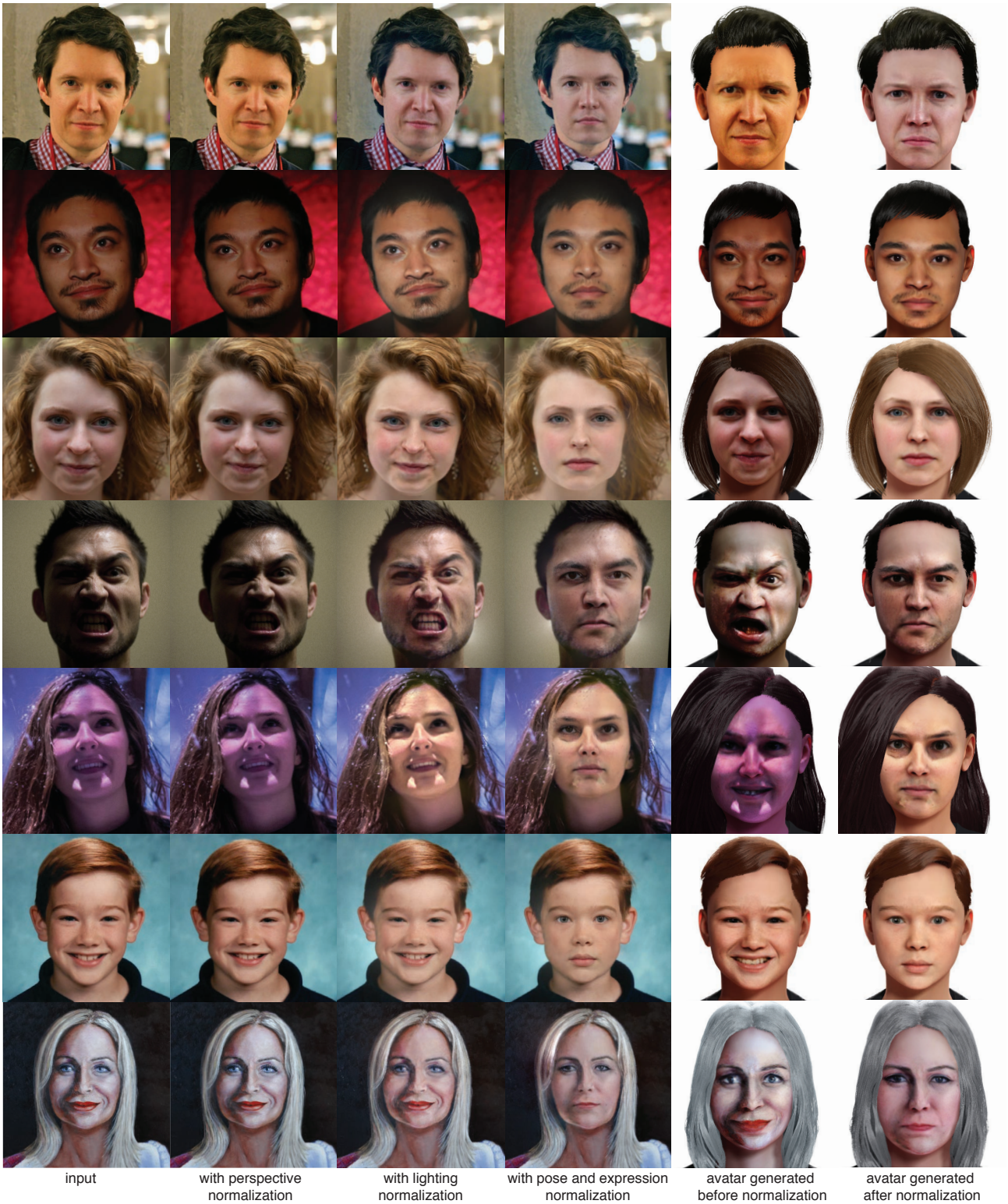

Fig. 11. First column: input photo. Second to fourth column: portrait with individual normalization component applied, i.e., perspective normalization, perspective+lighting normalization, and full normalization. The fifth and sixth columns show an avatar generated from an original input without normalization and from a fully normalized picture. From top to bottom, original images are courtesy of Daniel X. O'Neil, Michael Beserra, Benjamin Griffiths, Jens Karlsson, Remykennyl, Keith Parker, and Pamela Stone.

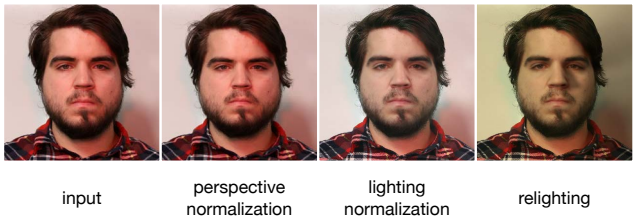

Fig. 12. Application: portrait manipulation is possible by individually applying each normalization component. After the lighting is normalized, the portrait can be relit with arbitrary lighting using an auxiliary 3D geometry and texture.

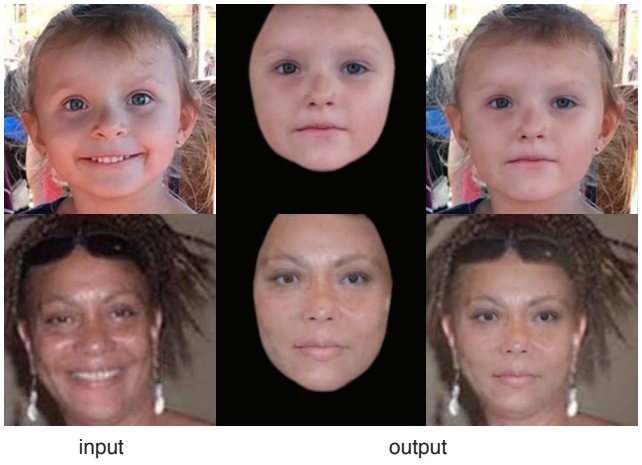

Fig. 13. Examples of neutral face reconstruction of a kidnapped child (top) and terrorist suspect (bottom) from the FBI Most Wanted database. Original images are courtesy of the Federal Bureau of Investigation.

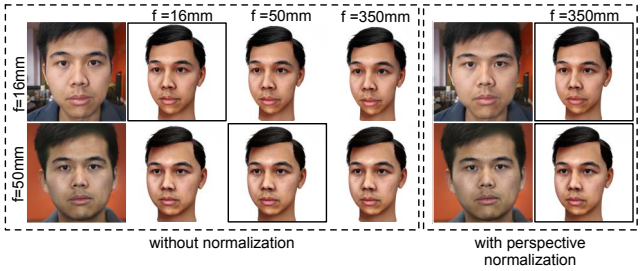

Fig. 14. Evaluation of perspective undistortion on 3D avatar modeling. Without normalization, 3D avatar modeling exhibits a large variations in facial shape depending on a focal length used in the facial fitting. Using the correct focal length can produce correct 3D shape (black boxes), but it is unknown in advance for unconstrained pictures. After perspective undistortion, face modeling can always produce correct shapes. Focal lengths (f) in 35mm sensor size.

right group, the normalization removes the perspective distortion of the input image and the facial modeling always produces plausible geometry with fixed focal length value (350mm in 35mm in our case), which closely matches to the one created with the original image and the correct focal length.

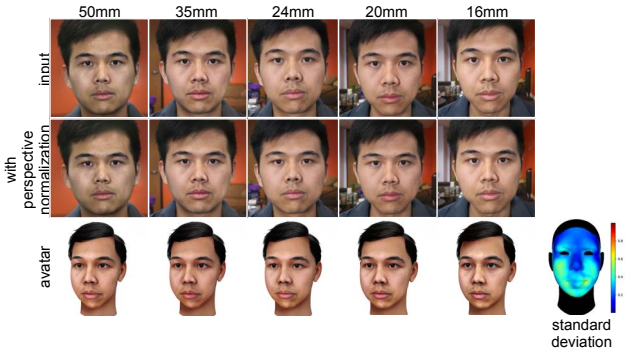

Fig. 15. Evaluation of perspective undistortion consistency. Photos taken from near to far distances exhibits significant variations in facial proportions (first row). After perspective undistortion, faces in normalized photos (second row) as well as the corresponding 3D avatar shapes (third row) are consistent. The heatmap on the lower right corner shows standard deviations for per-vertex Euclidean distance among 3D avatars.

We evaluate the robustness of our perspective undistortion method in Fig. 15 using a variety of distorted input images from near to far. Our algorithm can produce consistent portraits after perspective undistortion, as well as consistent avatar geometries.

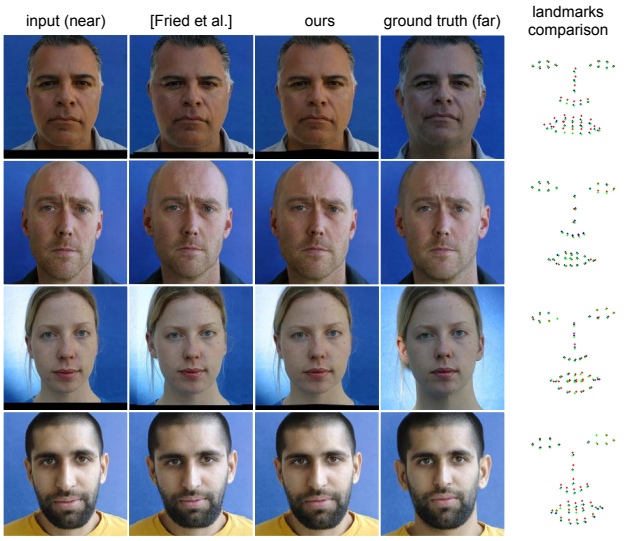

Fig. 16. Perspective undistortion comparison with [Fried et al. 2016]. The last column shows the 2D landmark alignments of our results (blue), [Fried et al. 2016] (green) and the ground truths (red).

In Fig. 16, we show a comparison with the previous work of [Fried et al. 2016] on the CMDP dataset [Burgos-Artizzu et al. 2014]. Our method can successfully remove the perspective distortions (especially in the nose area) for near range photos (first column),

Table 1. Ground truth Euclidean distance landmark error compared to [Fried et al. 2016]

Fried-GT(pix)	Ours-GT(pix)
4.557051	**3.948873**
1.192981	1.654887
2.379486	**2.262243**
3.185447	3.295902

producing visually similar facial proportion to the ground truth (fourth column), which are captured at a far distance. Table 1 shows the numerical difference computed from the aligned landmark to the ground truth. Our method produces comparable results to the previous work without knowledge of any camera parameters. Please see the supplemental material for additional numerical evaluations.

4.4 Even Light Illumination

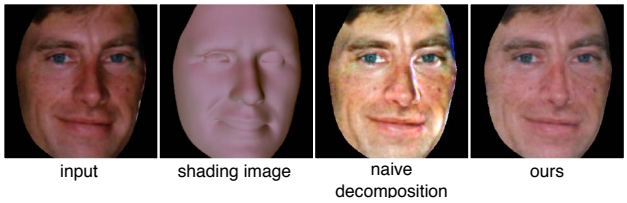

Fig. 17. From left to right: input image, shading estimation from 3D face fitting, naive decomposition by dividing the input image with the shading showing significant artifacts, and our result. Original image is courtesy of Keith Parker.

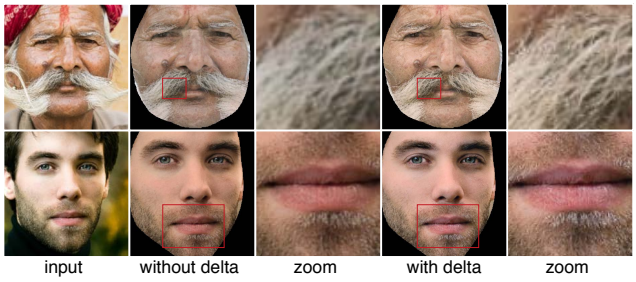

Fig. 18. Comparison of light normalization networks predicting the image $\mathbf{I}_{lit}$ (without delta) and predicting only the offset $\mathbf{O}_{lit}$ (with delta). Predicting $\mathbf{I}_{lit}$ directly yields loss in details. Original images are courtesy of Milena Martínez (top) and Trouni Tiet (bottom).

Fig. 17 compares our method with a naive lighting decomposition method using an estimated shading image. Due to the inaccuracies of the reconstructed 3d mesh and the limited capabilities of spherical harmonics, the naive decomposition (third column) exhibits significant artifacts, while our method ensures high-fidelity output.

We also compare our lighting normalization approach with a variant of the method in Fig. 18. By predicting the lighting offset image, we are able to preserve high-frequency intricate details (i.e.

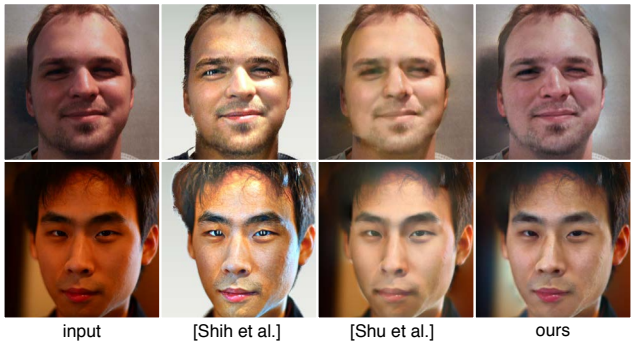

Fig. 19. Comparison with the portrait relighting methods [Shih et al. 2014] and [Shu et al. 2017]. Their results are produced using Fig.4 (a) in the supplemental material of [Shih et al. 2014] as a target lighting image. The previous work suffers from severe artifacts and cannot preserve the skin tone as well as ours. Original images are courtesy of Timothy Wood (top) and Toby Simkin (bottom).

facial hair) present in the input photograph (last two columns). The results tend to be blurred if the generator attempts to generate all the pixels from scratch (second and third columns).

In Fig. 19, we show how our method clearly outperforms state-of-the-art portrait relighting techniques for the task of delighting. In particular, only our approach can recover convincing skin tones of the subject that are appropriate for different ethnicity. Also our method does not rely on a lighting target reference.

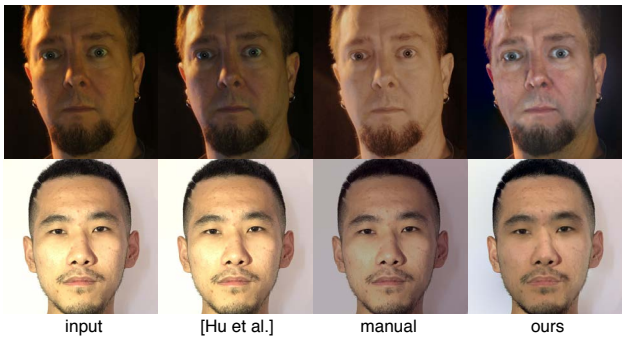

Fig. 20. Comparison with [Hu et al. 2017b] for portrait color correction, where our method is able to handle extreme directional lights and harsh shading on the face. Original images are courtesy of J. Davis Harvill (top).

Fig. 20 illustrates even more challenging lighting examples which include strong directional illuminations and rim lighting. Such harsh lighting conditions cannot be fixed using only color balance and exposure control, even when performed manually by a skilled digital artist (third column). On the other hand, our method can still recover plausible, though not perfect, facial appearances for both underexposed and overexposed areas. The robustness of our method is also demonstrated in Fig. 21. By illuminating the subject with a wide range of colored lighting conditions, our normalization method

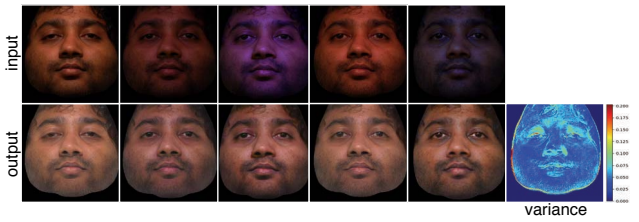

Fig. 21. Consistency of lighting normalization results for various input illumination colors. In the last column, we show the per-pixel variance computed on normalized pixel values, indicating the consistency of estimated result.

can reveal consistent facial skin tones of each person. On the last column we show the variance of estimated pixel colors, showing the consistency of our estimation.

4.5 Expression Neutralization

Fig. 22. Comparison with variants of geometry neutralization methods. Second row: naively resetting expression components to zero. Third row: our geometry neutralization without data augmentation. Our result (fourth row) shows the best results. Original images are courtesy of Michael Beserra, jcapaldi / flickr, Daniel X. O'Neil, and Sterre van den Berge (from left to right).

In Fig. 22, the second row shows the result using the geometry obtained by naively resetting all the expression parameters to zero and third row shows the results with neutralized geometry obtained from our method without synthetic data augmentation (Sec. 3.3). The identity is clearly better preserved with our full pipeline (fourth row) and also appears more natural using our geometry neutralization technique (Sec. 3.3).

Fig. 23 demonstrates the effect of data augmentation for facial texture neutralization in Sec. 3.3. Results without data augmentation lead to severe artifacts around the eyes.

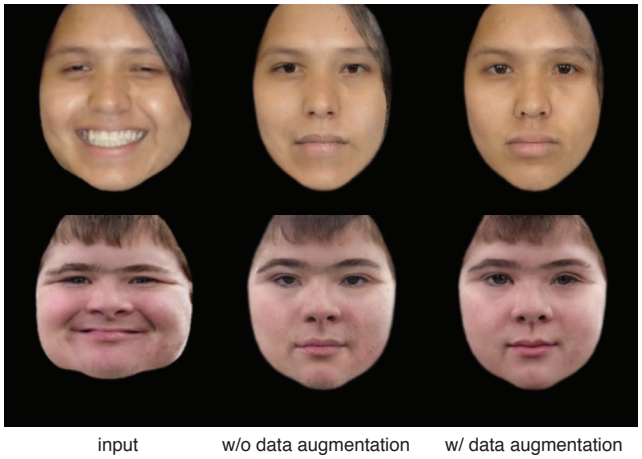

Fig. 23. Effects of data augmentation on the facial texture neutralization network. The second column shows the training results without data augmentation and the third column shows the results using synthetic faces. Original image are courtesy of NWABR/flickr (top) and Sheba_Also/flickr (bottom).

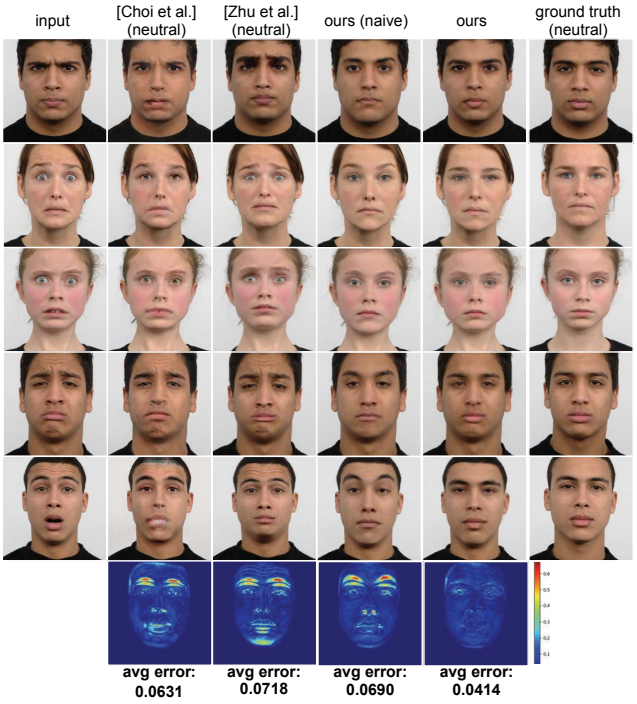

Fig. 24. Our expression normalization results compared with the state of the art methods and the ground truth. The bottom row visualizes the per-pixel error as heat maps as well as their average errors computed in normalized pixel values.

We compare our expression neutralization method with state-of-the-art GAN techniques of [Choi et al. 2018] and [Xiangyu Zhu et al. 2015] as well as a naïve approach by reseting all expressions to zero for the geometry (see Fig. 24). Even if intense facial wrinkles

are present in the expression, our method successfully removes them and produces a convincing neutral expression that reasonably matches the ground truth shown in the last column. Other methods however seem to struggle with removing strong expressions and yield artifacts especially in the mouth interior. In the bottom row of the figure, we show a quantitative comparison using pixel differences and average error for each method compared against the ground truth. The heatmap is computed after aligning each image to the ground truth using eye and nose landmarks.

Fig. 25. Consistency evaluation of expression neutralization. Our method produces a consistent neutral expression (bottom row) from various facial expressions in the input photos (top row).

We demonstrate the consistency of our expression neutralization method in Fig. 25. Although the input expressions shown in the first row exhibit a wide variety of facial deformations and wrinkles, our method can produce neutralized expressions that are reasonably consistent.

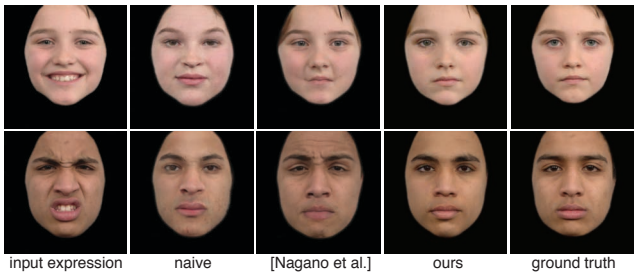

Fig. 26. From left to right: input images with expressions, naïve neutralization results trained with expression-to-neutral images, results from [Nagano et al. 2018], our result with generalized paGAN, and the ground truth neutral images.

Fig. 26 demonstrates the importance of having both neutral-to-expression and expression-to-neutral training data in order to ensure identity preservation and high-fidelity output. The facial identity tends to change after synthesis with an naive training (second column), which only consists of using expression-to-neutral samples. While an expression synthesis network [Nagano et al. 2018] can preserve facial identity well (third column), the network cannot fully remove unwanted wrinkles and expressions. The generalized paGAN model (fourth column) is most successful both in neutralizing expressions while keeping the identity.

We compare our normalization result with the state-of-the-art facial normalization approach [Cole et al. 2017] based on deep features for facial recognition (Fig. 27). While their method is robust

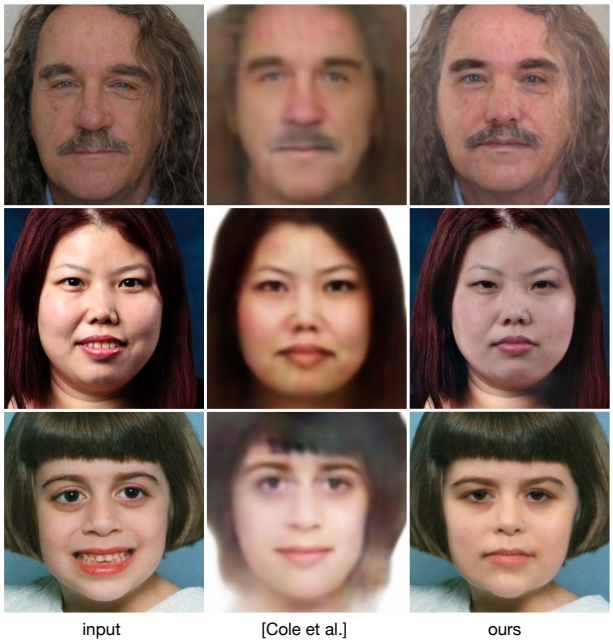

Fig. 27. Comparison with the end-to-end normalization method of [Cole et al. 2017]. Our results are much higher-fidelity and preserve high-frequency details from inputs. Original images are courtesy of Steven Damron, Tommy Low, and jcapaldi/flickr.

w.r.t. extremely challenging examples and can generate recognizable facial identities, our method can preserve high-frequency facial details and attributes, and produce higher resolution images.

4.6 User Study

We conducted a user study to assess the perceptual quality of our normalized faces. Over 400 testers from Amazon Mechanical Turk were asked to distinguish between our results and actual photos with proper portrait setting. Specifically, we provide 3 images of a subject in a row: (1) an input image with perspective distortion, extreme expression or unusual color lighting, (2) the ground truth image taken in a studio setting, and (3) the normalized result from our pipeline. Testers are required to pick the ground truth image from the latter two. 10 questions of different subjects are asked for each tester, with positions of answers randomly shuffled. Three questions of them have known answers to filter out random/malicious submission. Ideally, our pipeline should produce indistinguishable results from the ground truth, leading to a 50% fooling rate (testers cannot perform any better than random guessing). And a fooling rate of 0% would indicate that our results has noticeable difference from real photographs. As shown in Table 2 the fooling rate of each task is close to completely fooling users who are not familiar with the identities. Please refer to the supplemental material for more data including user study test samples.

Table 2. Fooling rate of user study. Our method performs very close to being able to fool the users absolutely (with a 0.5 rate).

Task	Fooling rate
perspective distortion	0.481
expression	0.438
lighting	0.446

4.7 Limitations

Our approach has several limitations (Fig. 28). Handling harsh lighting condition remains a challenge (top left). While our light normalization technique can produce high-resolution details (see Fig.18), the result could be slightly blurred or contain artifacts when composited to the background for portrait manipulation applications due to the Poisson blending. While our method can handle faces with mild poses, faces with strong side views are still challenging (top right). Also, our method does not handle frontalization of the body (bottom left). If the input has strong artistic filters such as Instagram filters, the skin color might not look convincing after lighting normalization (bottom right). While we show that neutralization from various expressions and lighting conditions can produce a reasonably consistent output, the results are still far from perfect.

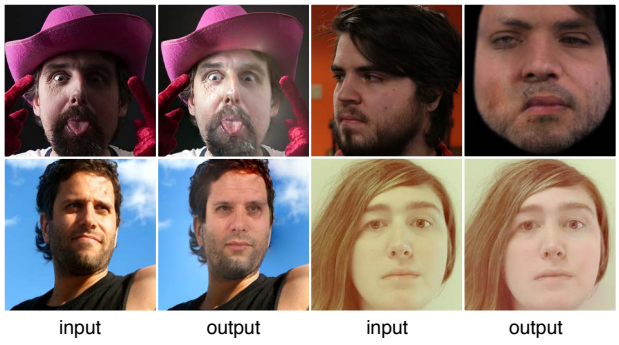

Fig. 28. Limitation: input images with extreme lighting conditions and large occlusions (top left), strong side-views (top right), non-frontal bodies (bottom left), and strong artistic filters e.g. Instagram selfies (bottom right). Original images are courtesy of Chris Roberts (top left) and Carlos Pacheco (bottom left)

5 DISCUSSION

We have shown that it is possible to fully normalize an unconstrained portrait while preserving facial details and resolution using a pipeline that performs perspective distortion removal, delighting, followed by facial frontalization and neutralization. Our perspecive normalization results are comparable to [Fried et al. 2016], but we do not require any camera parameter initialization (e.g. EXIF data) or 3D facial fitting. Our deep learning-based lighting normalization technique only approximates a real diffuse illumination conditions, and cannot be compared with ones that are obtained using a highly controlled capture environment [Ghosh et al. 2011]. Nevertheless, our delighting results are significantly superior than the current state-of-the-art and our produced skin tones are more consistent than end-to-end solutions such as [Cole et al. 2017] when different inputs are used. Furthermore, we show that our sequential normalization approach can produce significantly higher fidelity results than the end-to-end approach of [Cole et al. 2017], as high-frequency facial details are preserved. Furthermore, having intermediate steps in the pipeline gives us additional control for disentangled portrait manipulation. One of our core findings is that our combined approach of 3D neutral face regression based on deep features for face recognition combined with an inverted generalized paGAN can successfully invert highly complex facial expressions under extremely challenging capture conditions. As indicated by our user study, our normalization results are perceptually on par with real ones. However, if the person is appears familiar, it becomes harder to fool someone, whether the result is synthetic or not. Since portraits can now be normalized, more applications (e.g., portrait manipulation, 3D avatar creation) can benefit from these capabilities and generate more convincing results. We also believe that our solution can potentially impact law enforcement applications where high quality facial enhancement and reconstructions are needed due to limited available photographs of suspects or kidnapped persons.

Future Work. There are several directions we would like to pursue next. First, we would like to improve the robustness of our system, w.r.t., large head rotations and occlusions, especially by hair. Furthermore, we believe that providing additional input images or videos of the same person can be useful in determining a more accurate prediction of a normalized face, but handling additional unconstrained data may also increase the complexity of the problem. We would also like to extend our facial frontalization capabilities to other parts of the body, such as for hair, neck, and body regions. We believe that a more complete 3D human body and scene inference approach could help us produce plausible results from input data, where a large portion of a person's body is side facing. In the long term, we think that it would be possible to fully normalize high-resolution portraits, where the subject has more complex emotional states, such as crying, having tears in their eyes.

ACKNOWLEDGMENTS

We would like to thank Han-Wei Kung and Qingguo Xu for helping with generating results, Jiale Kuang, Erik Castellanos, Lain Goldwhite, Kyle San, and Aviral Agarwal for being our subject, Lain Goldwhite for proofreading, and Cole et al. for the comparisons.

REFERENCES

Y. Adini, Y. Moses, and S. Ullman. 1997. Face recognition: the problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, 7 (July 1997), 721–732.

Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F. Cohen. 2017. Bringing Portraits to Life. *ACM Trans. Graph.* 36, 4 (2017), to appear.

Jonathan T. Barron. 2015. Convolutional Color Constancy. In *IEEE ICCV (ICCV '15)*. IEEE Computer Society, Washington, DC, USA, 379–387. http://dx.doi.org/10.1109/ICCV.2015.51

Anil Bas and William A. P. Smith. 2018. Statistical transformer networks: learning shape and appearance models via self supervision. *CoRR* abs/1804.02541 (2018). arXiv:1804.02541 http://arxiv.org/abs/1804.02541

Volker Blanz and Thomas Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '99)*. 187–194.

Xavier P. Burgos-Artizzu, Matteo Ruggero Ronchi, and Pietro Perona. 2014. Distance Estimation of an Unknown Person from a Portrait. In *ECCV*. Springer International Publishing, Cham, 313–327.

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.

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 *IEEE CVPR*.

Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T. Freeman. 2017. Synthesizing Normalized Faces From Facial Identity Features. In *IEEE CVPR*.

R. L. Cook and K. E. Torrance. 1982. A Reflectance Model for Computer Graphics. *ACM Trans. Graph.* 1, 1 (Jan. 1982), 7–24.

Shichuan Du, Yong Tao, and Aleix M Martinez. 2014. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences* 111, 15 (2014), E1454–E1462.

Federal Bureau of Investigation. 2019. FBI Most Wanted. https://www.fbi.gov/wanted.

Arturo Flores, Eric Christiansen, David Kriegman, and Serge Belongie. 2013. Camera Distance from Face Images. In *Advances in Visual Computing*. Springer Berlin Heidelberg, Berlin, Heidelberg, 513–522.

Ohad Fried, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. 2016. Perspective-aware Manipulation of Portrait Photos. *ACM Trans. Graph.* (July 2016).

Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. 2018. Warp-guided GANs for Single-photo Facial Animation. *ACM Trans. Graph.* 37, 6, Article 231 (Dec. 2018), 12 pages.

Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. 2018. Unsupervised Training for 3D Morphable Model Regression. In *IEEE CVPR*.

A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. 2001. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 6 (June 2001), 643–660.

Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. 2011. Multiview Face Capture Using Polarized Spherical Gradient Illumination. *ACM Trans. Graph.* 30, 6, Article 129 (2011), 10 pages.

R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. 2008. Multi-PIE. In *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*. 1–8.

Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. 2015. Effective Face Frontalization in Unconstrained Images. In *IEEE CVPR*.

Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. 2017a. Avatar Digitization From a Single Image For Real-Time Rendering. *ACM Trans. Graph.* 36, 6 (2017).

Y. Hu, B. Wang, and S. Lin. 2017b. FC^4: Fully Convolutional Color Constancy with Confidence-Weighted Pooling. In *IEEE CVPR*. 330–339.

Yibo Hu, Xiang Wu, Bing Yu, Ran He, and Zhenan Sun. 2018a. Pose-Guided Photorealistic Face Rotation. In *IEEE CVPR*.

Yibo Hu, Xiang Wu, Bing Yu, Ran He, and Zhenan Sun. 2018b. Pose-Guided Photorealistic Face Rotation. In *IEEE CVPR*.

Rui Huang, Shu Zhang, Tianyu Li, and Ran He. 2017. Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis. In *IEEE ICCV*.

P. Isola, J. Zhu, T. Zhou, and A. A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *IEEE CVPR*. 5967–5976.

itSeez3D: Avatar SDK. 2019. https://avatarsdk.com.

Justin Johnson, Alexandre Alahi, and Fei-Fei Li. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. *CoRR* abs/1603.08155 (2016). http://arxiv.org/abs/1603.08155

Tero Karras, Samuli Laine, and Timo Aila. 2018. A Style-Based Generator Architecture for Generative Adversarial Networks. *CoRR* abs/1812.04948 (2018). http://arxiv.org/abs/1812.04948

Vahid Kazemi and Josephine Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *IEEE CVPR*. 1867–1874.

Hyeongwoo Kim, Pablo Carrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep Video Portraits. *ACM Trans. Graph.* 37, 4, Article 163 (July 2018), 14 pages.

Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, Skyler T Hawk, and AD Van Knippenberg. 2010. Presentation and validation of the Radboud Faces Database. *Cognition and emotion* 24, 8 (2010), 1377–1388.

Chen Li, Kun Zhou, and Stephen Lin. 2014. Intrinsic Face Image Decomposition with Human Face Priors. In *ECCV*. 218–233.

Ce Liu, Heung-Yeung Shum, and Chang-Shui Zhang. 2001. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *IEEE CVPR*, Vol. 1. I–I.

Loom.ai. 2019. http://www.loom.ai.

Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* 47, 4 (2015), 1122–1135.

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. *ACM Trans. Graph.* 37, 6, Article 258 (Dec. 2018), 12 pages.

Patrick Pérez, Michel Gangnet, and Andrew Blake. 2003. Poisson Image Editing. *ACM Trans. Graph.* 22, 3 (July 2003), 313–318.

Pinscreen. 2019. http://www.pinscreen.com.

Ravi Ramamoorthi and Pat Hanrahan. 2001. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 497–500.

Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. 2017. Photorealistic Facial Texture Inference Using Deep Neural Networks. In *IEEE CVPR*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *IEEE CVPR*.

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

Amnon Shashua and Tammy Riklin-Raviv. 2001. The Quotient Image: Class-Based Re-Rendering and Recognition with Varying Illuminations. *IEEE Trans. Pattern Anal. Mach. Intell.* 23, 2 (Feb. 2001), 129–139.

YiChang Shih, Wei-Sheng Lai, and Liang Chia-Kai. 2019. Distortion-Free Wide-Angle Portraits on Camera Phones. *ACM Trans. Graph.* 38, 4 (2019).

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.

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, Article 145a (Oct. 2017).

K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. 2017. Geometry Guided Adversarial Facial Expression Synthesis. *arXiv preprint arXiv:1712.03474* (2017).

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. *ACM Trans. Graph.* 38, 4 (2019).

Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *ICLR Workshop*.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE CVPR*. 2387–2395.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *IEEE CVPR*.

Y. Wang, L. Zhang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras. 2009. Face Relighting from a Single Image under Arbitrary Unknown Lighting Conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 11 (Nov 2009), 1968–1984.

Brittany Ward, Max Ward, Ohad Fried, and Boris Paskhover. 2018. Nasal distortion in short-distance photographs: The selfie effect. *JAMA Facial Plastic Surgery* 20, 4 (2018), 333–335. arXiv:/data/journals/faci/937383/jamafacial$_w ard_2 018_l d_1 80002.pdf$

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *IEEE CVPR*.

Chenglei Wu, Takaaki Shiratori, and Yaser Sheikh. 2018b. Deep Incremental Learning for Efficient High-fidelity Face Tracking. *ACM Trans. Graph.* 37, 6, Article 234 (Dec. 2018), 12 pages.

Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. 2018a. A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security* 13, 11 (2018), 2884–2896.

Xiangyu Zhu, Z. Lei, Junjie Yan, D. Yi, and S. Z. Li. 2015. High-fidelity Pose and Expression Normalization for face recognition in the wild. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 787–796. https://doi.org/10.1109/CVPR.2015.7298679

Shuco Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. 2018. High-fidelity Facial Reflectance and Geometry Inference from an Unconstrained Image. *ACM Trans. Graph.* 37, 4, Article 162 (July 2018), 14 pages.

Yajie Zhao, Zeng Huang, Tianye Li, Weikai Chen, Chloe LeGendre, Xinglei Ren, Jun Xing, Ari Shapiro, and Hao Li. 2019. Learning Perspective Undistortion of Portraits. *arXiv preprint arXiv:1905.07515* (2019).

Andrey Zhmoginov and Mark Sandler. 2016. Inverting Face Embeddings with Convolutional Neural Networks. https://arxiv.org/abs/1606.04189