3DFaceFill: An Analysis-By-Synthesis Approach to Face Completion

Rahul Dey   Vishnu Naresh Boddeti
Michigan State University
{deyrahul, vishnu}@msu.edu

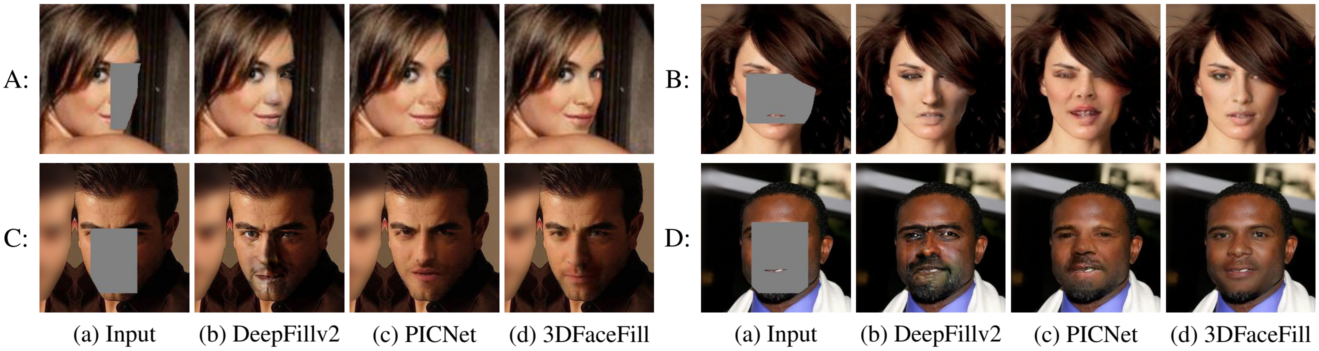


Figure 1: **Inpainting of face images under diverse conditions by 3DFaceFill and existing approaches**. By modeling the image formation process 3DFaceFill is able to generate more geometrically consistent and photorealistic completions across diverse scenarios such as non-frontal poses (A), light and dark complexions (B,D), non-uniform facial illumination (*e.g.* illumination is different on two sides of the nose in C) and in cases where the baselines tend to distort face components (*e.g.* nose in B).

Abstract

*Existing face completion solutions are primarily driven by end-to-end models that directly generate 2D completions of 2D masked faces. By having to implicitly account for geometric and photometric variations in facial shape and appearance, such approaches result in unrealistic completions, especially under large variations in pose, shape, illumination and mask sizes. To alleviate these limitations, we introduce 3DFaceFill, an analysis-by-synthesis approach for face completion that explicitly considers the image formation process. It comprises three components, (1) an encoder that disentangles the face into its constituent 3D mesh, 3D pose, illumination and albedo factors, (2) an autoencoder that inpaints the UV representation of facial albedo, and (3) a renderer that resynthesizes the completed face. By operating on the UV representation, 3DFaceFill affords the power of correspondence and allows us to naturally enforce geometrical priors (*e.g. *facial symmetry) more effectively. Quantitatively, 3DFaceFill improves the state-of-the-art by up to 4dB higher PSNR and 25% better LPIPS for large masks. And, qualitatively, it leads to demonstrably more photorealistic face completions over a range of masks and occlusions while preserving consistency in global and component-wise shape, pose, illumination and eye-gaze.*

1. Introduction

End-to-end image completion methods i.e., models that generate 2D completions directly from 2D masked images, have witnessed remarkable progress in recent years. These approaches rely primarily on architectural advances in neural network designs to implicitly account for photometric and geometric variations in image appearance. And even those that explicitly include scene geometry in their formulation do so largely in 2D. Consequently, object-based image completions from such methods often suffer from poor photorealism, especially under large variations in pose, shape, illumination of objects in the image and the inpainting mask. For example, in the context of faces, Fig. 1 shows face images having extreme poses (1.A), illumination variations across the face (1.C) and diverse appearances and shapes. Current state-of-the-art methods such as DeepFillv2 [41] and PICNet [46], both of which operate end-to-end on 2D image representations, often fail in preserving facial symmetry and the variations of the aforementioned factors (pose, illumination, texture, shape) while inpainting.

Several attempts have been made to customize generic image inpainting solutions for structured objects such as faces. General image inpainting approaches typically employ a CNN autoencoder as the inpainter and train it using a combination of photometric and adversarial losses

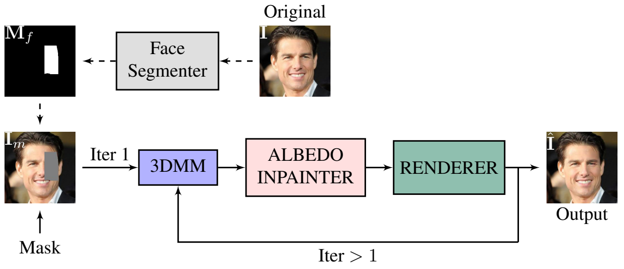


Figure 2: **Overview:** 3DFaceFill is an iterative inpainting approach where the masked face is disentangled into its 3D shape, pose, illumination and partial albedo by the 3DMM module, following which the partial albedo is inpainted and finally the completed image is rendered. During inference (only), the completed image is fed back through the whole pipeline in subsequent iterations, while using the initial mask for albedo inpainting. During training, a pre-trained model segments the image into face, hair and background for constraining the mask to lie only on the face. This segmentation is optionally used during inference if necessary.

[13, 23, 40, 46]. Face specific completion methods [19, 31] employ additional losses such as landmark loss, perceptual loss and face parsing loss. However, these approaches still do not account for all factors in the image formation process like illumination and pose variations and as such fail to effectively impose geometric priors such as facial symmetry. Moreover, the implicit enforcement of geometric priors is still done in 2D as opposed to in 3D. This is a significant limitation as faces are inherently symmetric 3D objects and their projections on 2D images are often affected by the aforementioned factors of pose, illumination, shape *etc*.

In contrast to the foregoing, this paper advocates for an analysis-by-synthesis approach for face completion that explicitly accounts for the 3D structure of faces i.e., shape and albedo, and image formation factors i.e., pose and illumination. The key insight of our solution is that performing face completion on the UV representation, as opposed to the 2D pixel representation, allows us to effectively leverage the power of correspondence and ultimately lead to geometrically and photometrically accurate face completion (see Fig.1). Our approach (see Fig. 2), dubbed 3DFaceFill, comprises of three components that are iteratively executed. First, the masked face image is disentangled into its constituent geometric and photometric factors. Second, an autoencoder performs inpainting on the UV representation of facial albedo. Lastly, the completed face is re-synthesized by a differentiable renderer. Our specific contributions are:

– We propose 3DFaceFill, a simple yet very effective face completion model that explicitly disentangles photometric and geometric factors and perform inpainting in the UV representation of facial albedo while preserving the associated facial shape, pose and illumination.

– We propose a 3D symmetry-aware network architecture and a symmetry loss for the inpainter to propagate albedo features from the visible to symmetric masked regions of the UV representation. Enforcing the symmetry prior in 3D, as opposed to 2D, allows 3DFaceFill to more effectively leverage and preserve facial *symmetry* while inpainting.

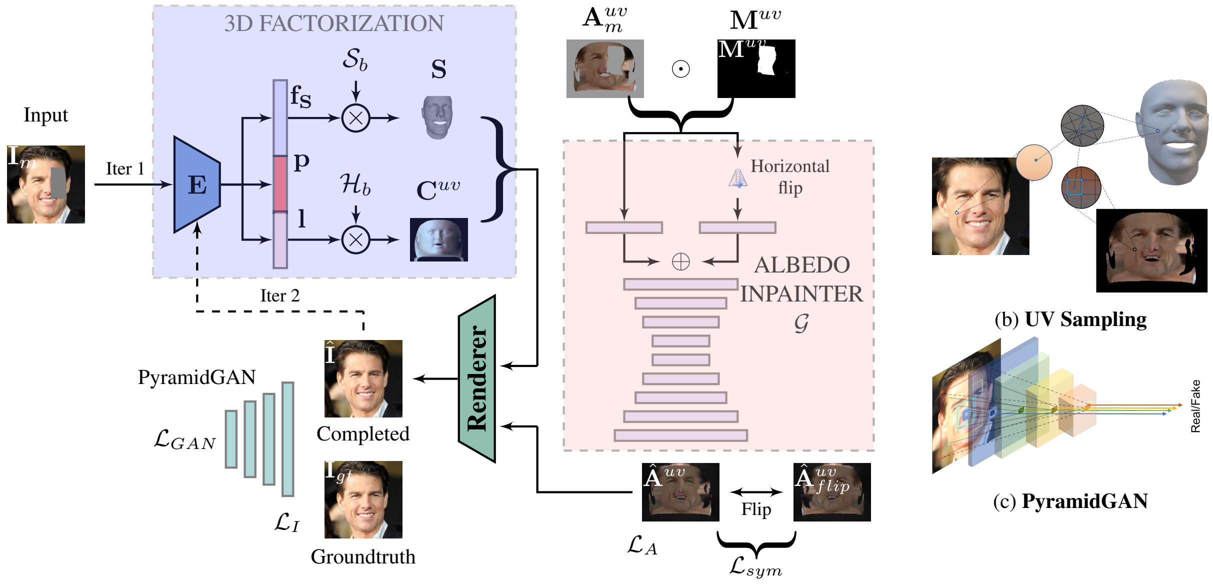
– Given our trained model, we propose a simple refinement process at inference by *iteratively* reprocessing the face completion through the model. This process enables us to address the “chicken-and-egg” problem of simultaneously inferring both the photometric and geometric factors and completion of the face from a masked image. The procedure is especially effective for heavily masked faces, improving the PSNR by up to 1dB.

– Extensive benchmarking on several datasets and unconstrained in-the-wild images results in 3DFaceFill producing photorealistic and geometrically consistent face completions over a range of masks and real occlusions, especially in terms of pose, lighting, and attributes such as eyegaze and shape of nose along with a quantitative improvement of upto 4dB PSNR and 25% in LPIPS [44].

2. Related Work

Image Inpainting: Earlier image inpainting approaches [1, 2, 6, 12] used diffusion or patch based methods to fill in the missing regions. This produced sharp results but often lacked semantic consistency. Recent techniques employ a CNN autoencoder along with a GAN loss to generate semantically consistent and realistic completions [13, 23, 39]. More recent methods focus on architectural enhancements to improve inpainting for variable and free form masks. These include a more refined discriminator in PatchGAN [14], contextual attention by DeepFill [40] and gated convolutions [21, 41]. In contrast, we adopt vanilla CNN architectures and instead rely on a more accurate analysis-by-synthesis modeling approach. Recently, Zheng *et al.* [46] generated multiple completions by sampling from a conditional distribution. Though this is a topic of interest, it is orthogonal to the goals of the current paper.

Face Completion: Face completion is a more challenging variant of image completion because of the complexity and diversity of faces. To address this, many approaches impose additional geometric and photometric priors in the form of face related losses [4, 19, 20, 31, 42, 45]. A recent approach called DSA [47] uses oracle-learned attention maps and component-wise discriminators to generate high-fidelity completions. While it often generates photorealistic completions in well-lit frontal faces, it still relies on implicitly learned priors which are insufficient to enforce correct geometry in challenging poses and illuminations. All these approaches rely on novel architectural advances and loss functions while 3DFaceFill focuses on more explicit and precise modeling of the image-formation process.



(a) **Architecture**

(b) **UV Sampling**

(c) **PyramidGAN**

Figure 3: **(a) Architecture:** Given a masked face $\mathbf{I}_m$, the 3DMM encoder extracts its shape $\mathbf{f_S}$, pose $\mathbf{p}$ and illumination $\mathbf{l}$ parameters, from which we obtain the full shape $\mathbf{S}$ and shade $\mathbf{C}^{uv}$ by linear combination of the corresponding bases. Then a partial albedo $\mathbf{A}^{uv}_m$ is obtained by first re-projecting the 3D mesh onto the masked image to obtain the UV-texture, as shown in **(b)** and then removing the shade from it $\mathbf{A}^{uv}_m = \mathbf{T}^{uv}_m \oslash \mathbf{C}^{uv}$. Finally, the albedo inpainter $\mathcal{G}$ completes the partial albedo as $\hat{\mathbf{A}}^{uv}$, conditioned on the UV-mask $\mathbf{M}^{uv}$, which is rendered along with the estimated shape, pose and shade to obtain the completed image $\hat{\mathbf{I}}$. To generate photorealistic completion, the completed and groundtruth images are evaluated by the proposed **(c) PyramidGAN** discriminator. **(b) UV Sampling:** 3D mesh is projected onto the face image to obtain per vertex RGB values $\mathbf{T}_\mathbf{v}(x, y, z)$. Each mesh face triangle $\mathbf{t} = (\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3})$ is mapped to a particular pixel in the UV space $\mathbf{T}^{v}_m(t) \rightarrow \mathbf{T}^{uv}_m(u, v)$ which allows us to sample the UV texture using barycentric interpolation.

Concurrently, Deng *et al.* [7] completed self-occluded UV texture to synthesize new face views. This assumes that the full face image and at least half of the UV texture is always visible. In contrast, we go beyond self-occlusion and instead, perform 3D factorization on the masked face and complete its *albedo* for *masked face completion*. Furthermore, since texture is not always symmetric due to illumination variations, [7] needs synthetically completed texture maps for training; whereas our model performs completion on albedo which is further disentangled from both geometry as well as illumination allowing us to effectively enforce symmetry prior, without needing synthetically completed UV-maps for training, as it bears out in our experiments. A few recent works have also attempted to leverage symmetry for face completion [18,43]. However, these approaches employ complex symmetry registration operations, which require huge computational resources; moreover these operations are often susceptible to large geometric variations.

3. Approach

In this section, we first present an overview of our proposed 3D face completion approach (dubbed 3DFaceFill) followed by the details of each component. As shown in Fig. 2, 3DFaceFill has three components: a 3DMM encoder, an albedo completion module and a renderer. Given a masked face, 3DFaceFill first resolves it into its constituent 3D shape, pose and illumination using the 3DMM encoder (Fig. 3). Then, we obtain the partial facial texture in the UV-domain by re-projecting the mesh onto the input image (Fig. 3b). We further remove the shading component to obtain an illumination-invariant partial albedo. The inpainter completes the partial albedo using symmetric and learned priors. Finally, the renderer combines the inpainted albedo with the estimated 3D factors to obtain the completed face. As a natural extension of the proposed approach, we use 3D factorization and completion in a complimentary way to further improve completion iteratively.

3.1. 3D Factorization

Existing face image completion approaches directly operate on 2D, which makes it non-trivial to enforce strong 3D geometric and photometric priors. This leads to poor face completion in challenging conditions of poses, geometry, lighting, *etc*. This motivates us to adopt explicit 3D factorization of face images to disentangle the appearance and geometric components, to enable robust completion.

Essentially, the 3D factorization module is an inverse renderer $\Phi : \mathbf{I} \rightarrow (\mathbf{S}, \mathbf{p}, \mathbf{l}, \mathbf{A})$ that resolves a 2D face $\mathbf{I}$

into its constituent shape $\mathbf{S} \in \mathbb{R}^3$, pose $\mathbf{p}$, illumination $\mathbf{l}$ and albedo $\mathbf{A}$. Various 3DMM approaches like [3, 8, 9] can be a natural fit for this. However, they are not real time, leaving learning based 3D reconstruction approaches [28, 30, 32–35, 37] as the obvious choices. While any of these approaches can potentially be used in our approach, for the purpose of this work, we adopt a simplified version of the nonlinear 3DMM presented by Tran *et al.* [34].

The 3D factorizaiton module consists of a 3DMM encoder and an albedo decoder (used only during training). The encoder $\mathcal{E}$ first resolves the image $\mathbf{I}$ in to its shape, albedo and illumination coefficients $(\mathbf{f_S}, \mathbf{f_A}, \mathbf{l})$ and pose $\mathbf{p} = (s, \mathbf{R}, \mathbf{t})$. Using the shape coefficients, we obtain the full shape $\mathbf{S}$ by linear combination with the Basel Face Model's (BFM) bases [24]. Similarly, we combine the illumination coefficients linearly with the spherical harmonics (SH) bases $\mathbf{H}_b$ [26] to obtain the surface shading $\mathbf{C}^{uv}$ (we assume *Lambertian* surface reflectance). The decoder $\mathcal{D}_\mathbf{A}$ maps the albedo coefficients into the full UV-albedo $\mathcal{D}_\mathbf{A} : \mathbf{f_A} \to \mathbf{A}^{uv}$, which is then multiplied with the shade to obtain the texture $\mathbf{T}^{uv} = \mathbf{A}^{uv} \odot \mathbf{C}^{uv}$. A differentiable renderer $\mathcal{R}$ [34] then reprojects the estimated 3D factors into image $\mathbf{I}_{ren}$ using the Z-buffer technique:

$$\mathbf{I}_{ren} = \mathcal{R}\left(\mathbf{S}, \mathbf{A}, \mathbf{p}, \mathbf{l}\right) \tag{1}$$

We train the module using masked images for robustness to partial inputs. For further details, refer the supplement.

3.2. Albedo Completion Module

Architecturally, our albedo completion module is similar to other adversarially trained image-completion autoencoders [19, 23, 40]. However, ours has the unique advantage of being solely focused on recovering the missing albedo, which has been disentangled from other variations in shape, pose and illumination through 3D factorization and is largely symmetric in its UV-representation. UVGAN [7] performs a similar completion of self-occluded UV-texture extracted from fully-visible face images. However, because of the entangled illumination, they don't use symmetry and need a synthetically completed texture map for supervision, whereas we use symmetry as self-supervision.

To this end, we discard the soft albedo obtained from the 3DMM albedo decoder and instead obtain the more realistic partial albedo from the input image in the UV space. This is done in two steps: first, we reproject the obtained 3D mesh onto the face image and use bilinear interpolation to sample the per-vertex texture (see Fig. 3b):

$$\mathbf{T}_m^{\mathbf{v}}(x, y, z) = \sum_{\substack{p \in \{\lfloor x \rfloor, \lceil x \rceil\} \\ q \in \{\lfloor y \rfloor, \lceil y \rceil\}}} \mathbf{I}_m^{p,q} (1 - |x - p|)(1 - |y - q|)$$

Then, we map the sampled partial texture $\mathbf{T}_m^{\mathbf{v}}$ onto the UV space using barycentric interpolation on the predefined mesh-to-uv mappings $\mathbf{T}_m^{\mathbf{v}}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) \to \mathbf{T}_m^{uv}(u, v)$. From the texture, we obtain the partial albedo by simply removing the estimated shade: $\mathbf{A}_m^{uv} = \mathbf{T}_m^{uv} \oslash \mathbf{C}^{uv}$, where $\oslash$ is the element-wise division operation. We perform similar operations to unwarp the mask $\mathbf{M}$ on-to the UV-space as $\mathbf{M}^{uv}$.

We use a U-Net [27] based autoencoder $\mathcal{G}$ to complete the partial albedo conditioned on the input mask, $\mathcal{G} : (\mathbf{A}_m^{uv}, \mathbf{M}^{uv}) \to (\hat{\mathbf{A}}^{uv}, \sigma^{uv})$, where $\hat{\mathbf{A}}^{uv}$ is the completed albedo and σ^{uv} is the uncertainty of completion. In order to leverage the bilateral symmetry of the UV facial albedo as an attention map, we modify the U-Net architecture (henceforth referred to as Sym-UNet). This is specially helpful since we do not have access to the full groundtruth albedo maps for training. To do so, we split the first convolution layer $f_{1:2c}$ into two parts: $f_{1,1:c}$ and $f_{2,c+1:2c}$ with equal number of output channels c (see Fig. 2). The first filter operates on the input albedo as such $\mathbf{h}_1 = f_1(\mathbf{A}_m^{uv})$. The second, instead, operates on the horizontally flipped albedo $\mathbf{h}_2 = f_2(hflip(\mathbf{A}_m^{uv}))$. We then concatenate the activations $\mathbf{h}_1$ and $\mathbf{h}_2$ from these two filters and pass it through the rest of the network. During training, the first filter learns to extract features from the visible parts of the albedo while the second filter learns to extract features corresponding to the symmetrically opposite visible parts to apply on the occluded regions (see Sec. 3.2 in the supplementary).

A naive approach of doing so, however, results in artifacts from the symmetrical counterparts to appear on the visible regions, making the network convergence difficult. Instead, we use gated convolutions [41] (in all but the final layer), to ensure that such symmetric features are only transferred to the masked regions and do not create artifacts on the visible regions. We use group normalization [38] and ELU activation [5] for all the feature layers and the final output is simply clipped between -1 and 1. We then render the completed albedo $\hat{\mathbf{A}}^{uv}$, along with the estimated shape, pose and illumination to obtain a completed image $\hat{\mathbf{I}}$ using eqn. 1. Finally, we simply blend the input and completed images to obtain the output image: $\mathbf{I}_{out} = \mathbf{I} \odot (1 - \mathbf{M}) + \hat{\mathbf{I}} \odot \mathbf{M}$.

PyramidGAN Discriminator: To generate sharp and semantically realistic completions, we use a multi-scale PatchGAN discriminator [29, 36], which we refer to as the *PyramidGAN*. The PyramidGAN evaluates the final output $\mathbf{I}_{out}$ at multiple locations and scales ranging from coarse and global to fine and local (refer to Fig. 3c). Features from each l-th downsampling layer of the PyramidGAN $\mathcal{D}_l$ are used to evaluate an average hinge loss [15, 41] for that layer. We then compute the average loss across all the layers as the total loss, thus giving equal weightage to each scale:

$$\begin{aligned} \mathcal{L}_\mathcal{G} &= -\,\mathbb{E}_{p(z)}\left[\mathbb{E}_{l \in L}\left[\mathcal{D}_l(\mathcal{G}(z))\right]\right] \qquad (2) \\ \mathcal{L}_\mathcal{D} &= \mathbb{E}_x\left[\mathbb{E}_{l \in L}[\mathbf{1} - \mathcal{D}_l(x)]_+\right] + \mathbb{E}_{p(z)}\left[\mathbb{E}_{l \in L}[\mathbf{1} + \mathcal{D}_l(\mathcal{G}(z))]_+\right], \end{aligned}$$

Training Losses: We train the albedo completion module with the following total loss:

$$\mathcal{L} = \lambda_1 \mathcal{L}_A + \lambda_2 \mathcal{L}_I + \lambda_3 \mathcal{L}_{sym} + \lambda_4 \mathcal{L}_{GAN} + \lambda_5 \mathcal{L}_{gp}, \quad (3)$$

where $\mathcal{L}_A = \mathcal{L}_\sigma(||\hat{\mathbf{A}}^{uv} - \hat{\mathbf{A}}^{uv}_{gt}||_1, \sigma^{uv})$ and $\mathcal{L}_I = \mathcal{L}_\sigma(||\hat{\mathbf{I}} - \mathbf{I}_{gt}||_1, \sigma)$ are the pixel losses for the albedo and the image, respectively, $\mathcal{L}_{sym}$ is the symmetry loss, $\mathcal{L}_{GAN}$ is the GAN loss given in eqn. 2 and $\mathcal{L}_{gp}$ is the WGAN-GP loss as described in [11]. The albedo symmetry loss is carefully applied on the masked regions whose symmetric counterparts are visible, to supplement as supervised attention:

$$\mathcal{L}_{sym} = \mathcal{L}_\sigma \left((\mathbf{1} - \mathbf{M}^{uv}) \mathbf{M}^{uv}_{flip} \odot ||\hat{\mathbf{A}}^{uv} - \hat{\mathbf{A}}^{uv}_{flip}||_1, \sigma^{uv} \right)$$

Here, $\mathcal{L}_\sigma(\mathbf{x}, \sigma) = \frac{1}{D} \sum_i \frac{1}{2} x_i exp(-\sigma_i) + \frac{\sigma_i}{2}$ is the aleatoric uncertainty loss [16]. The loss coefficients are set to have similar magnitude for all the loss components. In this paper, the goal is to show the efficacy of explicit 3D consideration on the geometric and photometric accuracy of face completion. So, *we withhold from using attention or face specific losses [19, 40, 41, 46, 47]* and leave them as future add-ons.

Iterative Refinement: 3D factorization is an important first step of our proposed approach, which itself leads to robust face completion in cases where 2D based methods fail. To make the 3D factorization itself robust to partial images, we train the 3DMM encoder on face images with randomly sized and randomly located masks. However, there is scope to further improve upon this and leverage the full power of our proposed two-step approach. To do this, we adopt a simple iterative refinement technique where face completion leads to improved 3D factorization and vice versa, as shown in Fig. 2. During inference, the masked face is used to distill the 3D factors in the first iteration; while in the next iteration, the completed face itself forms the input for 3D analysis. This leads to iteratively refined 3D analysis (*specially the 3D pose*) as well as face completion. Though one can repeat the iterative step many times, we experimentally found that two such iterations are usually sufficient.

4. Experimental Evaluation

Datasets: We evaluate the proposed 3DFaceFill on the CelebA [22] and CelebA-HQ [17] datasets. We use 80% split for training and 20% for evaluation. Further, to evaluate the robustness and generalization performance, we do a cross-dataset evaluation on the pose and illumination varying images from the MultiPIE [10] dataset and ∼50 in-the-wild face images downloaded from the internet[1].

Implementation Details: We train both the 3D factorization and the completion modules independently using the Adam optimizer with a learning rate of 10^{-4}. We first train the 3DMM module on the 300W-3D [48] and the CelebA [22] datasets. Once the 3DMM encoder is trained, we freeze it and use it to train the completion module on the CelebA [22] dataset for 30k iterations. We generate random rectangular masks of varying sizes and locations, and constrain them to lie in the segmented face region (Fig. 2). We provide further details on implementation and computational analysis in the supplementary.

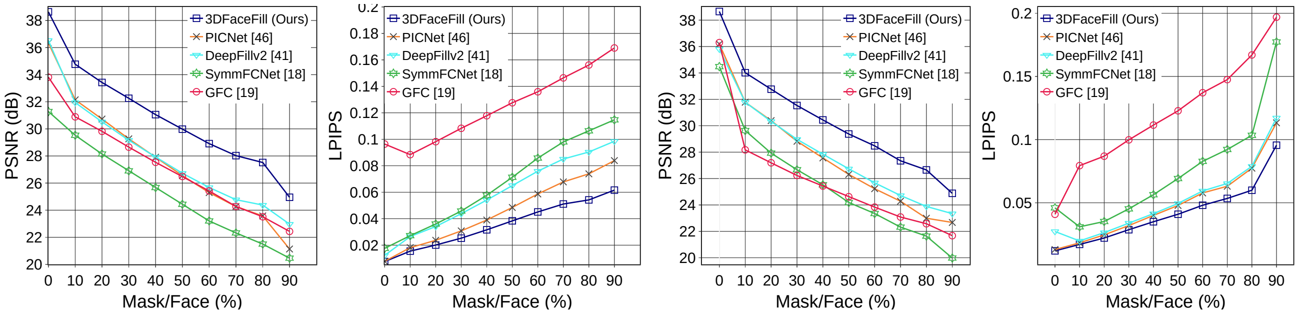
Baselines: To evaluate the efficacy of 3DFaceFill, we perform qualitative and quantitative comparison against baselines such as GFC [19], SymmFCNet [18], DeepFillv2 [40, 41] and PICNet[2] [46]. We use the publicly available pretrained face models for DeepFillv2 [41], PICNet [46] and SymmFCNet [18]. For GFC [19], the pretrained model was not trained on the same crop and alignment as ours, so we train it from scratch using their source code. Due to the absense of extensive results, we present additional evaluation against baselines that do not provide source codes or pre-trained models in the supplementary, using a small set of results obtained from the corresponding authors.

4.1. Results

Quantitative Evaluation: In addition to the typically used PSNR and SSIM metrics, we report LPIPS [44], which is more suitable for image completion. Table 4d reports the overall values of these metrics across all image-mask pairs for each dataset. Overall 3DFaceFill improves PSNR by 2dB-3dB and LPIPS by 5-10% over the closest baselines. In addition, for all the methods, we report PSNR and LPIPS as a function of mask to face area ratio ($\frac{\#MaskPixels}{\#FacePixels}$) in Fig. 4a, 4b and 4c for the CelebA, CelebA-HQ and Multi-PIE datasets, respectively. We make the following observations: (1) Across all the datasets, 3DFaceFill achieves significantly better PSNR and LPIPS across all mask ratios. (2) Among the baselines, PIC [46] and DeepFillV2 [41] perform comparably with the former being slightly better in terms of LPIPS. (3) The effectiveness of 3DFaceFill over the baselines is more apparent as larger parts of the face are to be completed i.e., as the mask ratio increases. (4) On the CelebA dataset [22], the improvement ranges from ∼2dB PSNR for 0-10% mask ratio to ∼4dB PSNR for 60-80% mask ratio. In terms of LPIPS, the improvement ranges from 5% for 0-10% mask ratio to 25% for 60-90% mask ratio. Similar trends are seen across the CelebA-HQ [17] and MultiPIE [10] datasets too. These results confirm our hypothesis that explicitly modeling the image formation process leads to significantly better face completion. We provide addtional quantitative comparisons against PConv [21], DSA [47] and UVGAN [7] in the supplementary since these results are based on a limited number of author-provided completions in the absense of source codes.

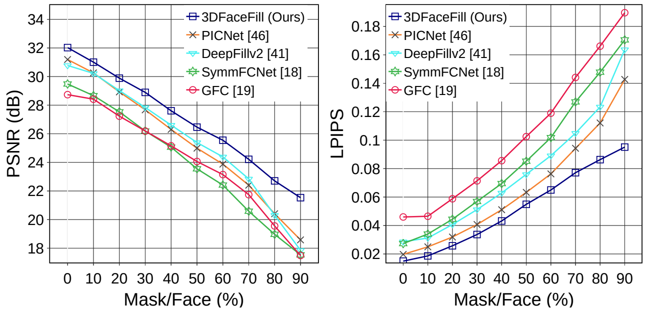
[1] Source: `https://unsplash.com/s/photos/face`

[2] Following author guidelines, we sample top 10 completions ranked by its discriminator and chose the one closest to the groudtruth for evaluation.



(a) CelebA dataset [22]

(b) CelebA-HQ dataset [17]



(c) MultiPIE dataset [10]

Dataset	**Metric**	GFC [19]	SymmFC [18]	DeepFill [41]	PIC [46]	3DFaceFill
	PSNR (↑)	27.0298	25.8817	28.2097	28.1262	**30.4917**
CelebA	**SSIM** (↑)	0.9257	0.9273	0.9356	0.9424	**0.9521**
	LPIPS (↓)	0.1134	0.0537	0.0499	0.0362	**0.0326**
	PSNR (↑)	25.5836	25.6203	27.9885	27.7020	**29.9398**
CelebAHQ	**SSIM** (↑)	0.8895	0.9232	0.9311	0.9380	**0.9492**
	LPIPS (↓)	0.1076	0.0535	0.0394	0.0376	**0.0365**
	PSNR (↑)	25.3805	25.1280	26.8225	26.5574	**28.7515**
MultiPIE	**SSIM** (↑)	0.9127	0.9266	0.9391	0.9397	**0.9553**
	LPIPS (↓)	0.0798	0.0645	0.0577	0.0472	**0.0436**

(d) Average metrics across all masks.

Figure 4: **Quantitative Evaluation:** We perform face completion over (a) CelebA [22], (b) CelebA-HQ [17] and (c) MultiPIE [10] datasets across a range (0-90%) of mask to face area ratios and evaluate the PSNR and LPIPS [44] metrics. In addition, we report the overall metrics across all mask-to-face are ratios in Table (d). 3DFaceFill consistently outperforms the baselines across all the datasets and mask ratios.

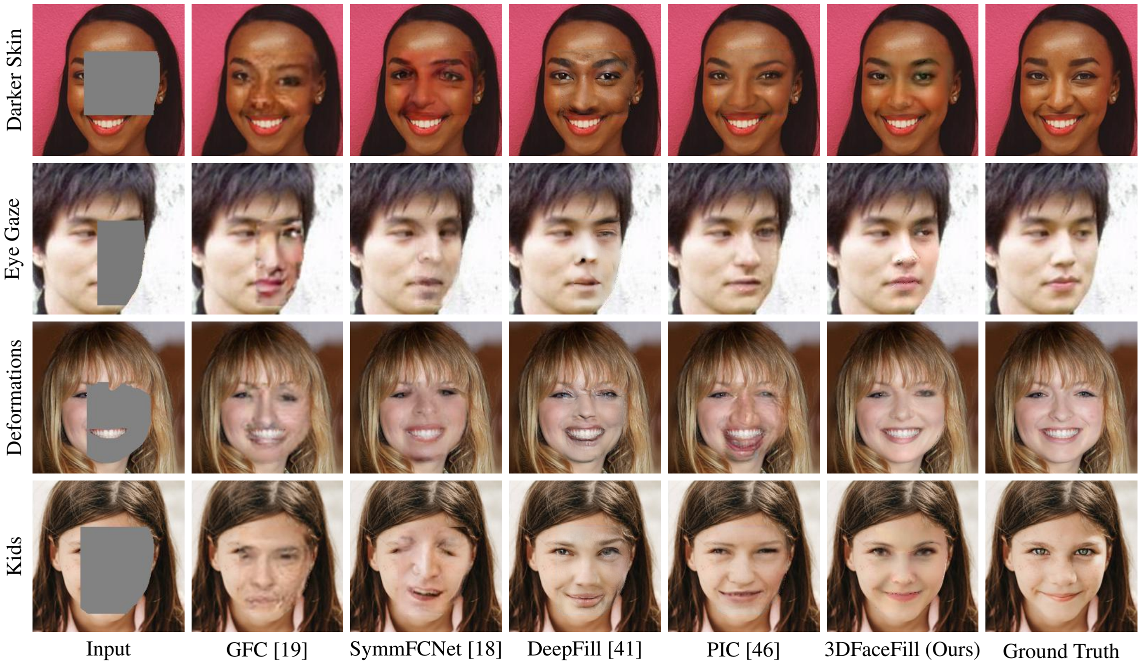


Figure 5: **Qualitative Evaluation:** Inpainting on faces from the CelebA [22] and CelebA-HQ [17] test sets (except last row downloaded from the internet). Across a variety of scenarios, the completions from baselines often have artifacts while those from 3DFaceFill are significantly more photorealistic due to explicit modelling of the image formation process. More examples can be found in the supplementary.

Qualitative Evaluation: Fig. 5 qualitatively compares face completion between 3DFaceFill and the baselines over a wide variety of challenging conditions. Completions by the baselines are less photorealistic and often contain artifacts in scenarios with dark complexion, tend to deform facial components (*e.g.* nose) and fail to preserve symmetry

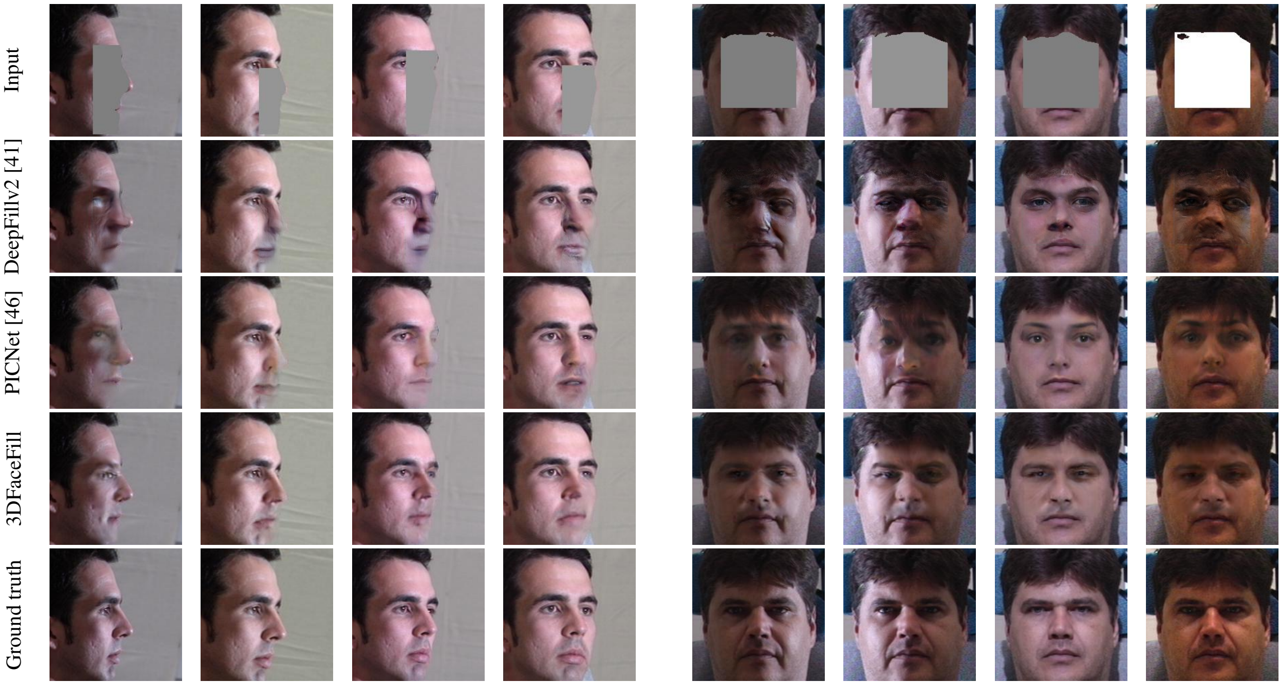


Figure 6: **Qualitative evaluation on the MultiPIE [10] dataset**. Compared to the baselines that generate deformed faces with artifacts in extreme poses and illumination, 3DFaceFill is more robust and generate geometrically accurate and illumination-preserving faces.

(*e.g*. eye-gaze or eye-brow shape). In addition, the baselines tend to deform the shape of small faces (*e.g*. children) since they are mostly trained on adult faces where the relative proportions of facial parts differs significantly. In contrast, 3DFaceFill generates more photorealistic completions in all these cases (diverse conditions and mask types) due to explicit 3D shape modeling, incorporating symmetry priors and disentanglement of pose and illumination.

Cross-Dataset Evaluation: To further demonstrate the improved generalization performance and robustness afforded by our method, we perform a cross-dataset comparison on the pose and illumination varying images from the MultiPIE [10] dataset, *using models that were trained on the CelebA dataset [22]*. Note that most baselines [19, 40, 46, 47] do not perform such an evaluation. Quantitative results are in the last rows of Table 4d, while Fig. 6 shows the qualitative results. Fig 6 (left) shows that the baselines generate fuzzy and deformed faces while 3DFaceFill generates consistently superior completion across all poses. Similarly, for the varying illumination case (Fig 6-right), 3DFaceFill not only generates superior completion but also preserves the illumination contrast across the face.

Real Occlusions: One of the potential applications of face completion is in de-occlusion. This is usually challenging when faces have large pose, illumination or shape variations. Fig. 7 shows a few real-world de-occlusion examples of faces in such conditions. Notice that, in cases of challenging pose, illumination, *etc*., the baselines tend to generate blurry and asymmetric face completions, whereas 3DFaceFill does more realistic de-occlusion.

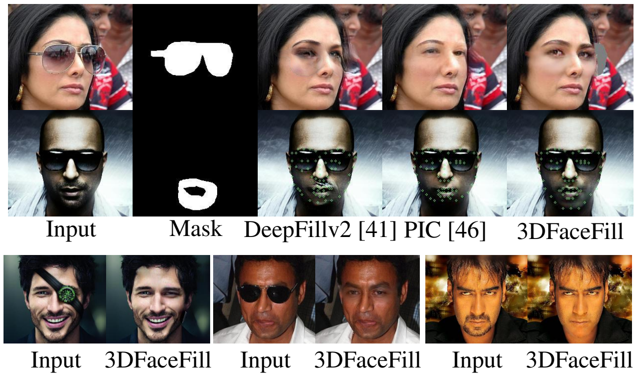


Figure 7: **FC on real occlusions**. Note the asymmetric eye-gaze (row 1) and blurry shape (row 2) by the baselines.

4.2. Ablation Studies

Iterative Refinement: To evaluate the effectiveness of iteratively refining face completion at inference, we compare the PSNR, SSIM and LPIPS [44] metrics on raw output images (before blending with the visible image) at each iteration. As reported in Table 1, iteration 2 significantly improves upon iteration 1 over all the metrics. After iteration 2, the metrics become more or less stable, with a slight dip in performance. We hypothesize that it is a result of

	Iter 1	Iter 2	Iter 3	Iter 4	Iter 5	Iter 6
PSNR (↑)	33.7587	**34.5347**	34.5018	34.4943	34.4428	34.4018
SSIM (↑)	0.9510	**0.9678**	0.9675	0.9670	0.9666	0.9652
LPIPS (↓)	0.0192	**0.0185**	0.0186	0.0187	0.0188	0.0188

Table 1: Quantitative evaluation of iterative refinement.

Metric	Full GAN	Patch GAN	NoSym	NoSym+Attn	Full Model
PSNR (↑)	31.7125	31.7552	31.6110	31.7969	**32.1950**
SSIM (↑)	0.9654	0.9658	0.9665	0.9667	**0.9678**
LPIPS (↓)	0.0462	0.0454	0.0446	0.0442	**0.0410**

Table 2: Quantitative evaluation between the different ablation models and our full model on masks blocking one-half of the face.

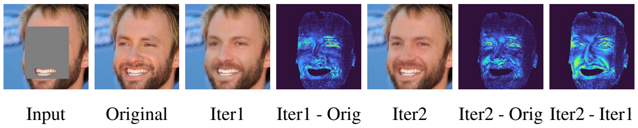


Figure 8: Visualization of raw completion (without blending) at iterations 1 and 2 along with the difference heatmaps.

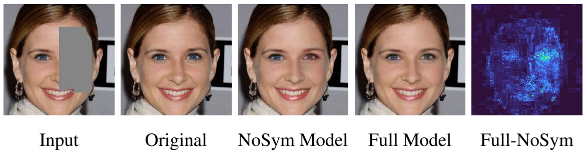


Figure 9: Visualizing the effect of symmetry for face inpainting. The full model includes Sym-UNet and symmetry loss (during training) and can copy symmetric features when available. The absolute difference heatmaps (Full-NoSym) shows that most difference is coming from components such as eyes, eye-brows, *etc*.

not training the model for iterative refinement and only performing it at inference. Further, we visualize the absolute difference heatmaps between the completed and the original image for both iterations 1 and 2 in Fig. 8 to understand which parts of the face benefit most from refinement. Observe that the largest differences are around the high-detail regions (eyes, beards, *etc*.), which we ascribe to more accurate 3D pose and shape estimation from the completed face after iteration 1 than from the partial face before.

Symmetry Constraint: To evaluate the effectiveness of Sym-UNet and the symmetry loss, we compare two variants of the full model (Sym-UNet + symmetry loss). These include, (1) **NoSym:** Sym-UNet replaced by standard UNet and with no symmetry loss, and (2) **NoSym+Attn:** NoSym model plus a self-attention layer after the 3rd upsampling layer in the UNet decoder. Attention layers are commonly employed by many inpainting models [40, 41, 46] for capturing long-range spatial dependencies, so this variant seeks to compare the utility of attention in lieu of symmetry priors for face inpainting. To best evaluate the benefit of symmetry constraints for faces, the above model variations are evaluated on face images masked on one side of the face as shown in Fig. 9.

The results in Table 2 indicate that the full model outperforms all the variants, with NoSym being the worst among them. Also the NoSym+Attn variant does perform slightly better than NoSym but is still far behind the full model. This indicates that, (i) though attention helps in the absence of any prior constraints, explicitly enforcing geometric priors associated with structured objects like faces is significantly more effective than implicitly learning them through attention, and (ii) symmetry is a more useful feature for face inpainting and behaves like an attention on the visible symmetric parts. As shown in Fig. 9, compared to the full model, the NoSym variant results in larger inpainting errors as indicated by the difference heatmaps. Therefore, unlike the full model the NoSym model tends to ignore the visible symmetric regions of the face leading to inconsistencies between the visible and inpainted regions.

4.3. Discussions

The above described experiments and ablation studies demonstrate the effectiveness of 3DFaceFill, along with the utility of each of its components in performing robust face completion in challenging cases of facial pose, shape, illumination, *etc*. However, the formulation of our proposed approach do impose a dependency on the fidelity of the underlying 3D model. Essentially, our approach cannot inpaint on regions which are not included in the underlying 3D model and the resolution of inpainting depends on the density of the 3D mesh. 3DFaceFill currently uses the BFM model [24], thanks to its widespread support. However, BFM [24] does not include the inner mouth, hairs and the upper head and has limited vertex density around the eyes, which restricts inpainting in these regions. However, these limitations of the underlying 3D model are not inherent to the proposed approach and do not invalidate the advantages of our model in improving the geometric and photometric consistency of completion. Furthermore, these limitations can potentially be mitigated by substituting BFM with a more detailed 3D face model, such as the Universal Head Model (UHM) [25], that includes the inner mouth and detailed eye-balls, along with other improvements.

5. Conclusion

In this paper, we proposed 3DFaceFill, a 3D-aware face completion method. Our solution was driven by the hypothesis that performing face completion on the UV representation, as opposed to 2D pixel representation, will allow us to effectively leverage the power of 3D correspondence and ultimately lead to face completions that are geometrically and photometrically more accurate. Experimental evaluation across multiple datasets and against multiple baselines show that face completions from 3DFaceFill are significantly better, both qualitatively and quantitatively, under large variations in pose, illumination, shape and appearance. These results validate our primary hypothesis.

References

[1] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009.

[2] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[4] Y.-A. Chen, W.-C. Chen, C.-P. Wei, and Y.-C. F. Wang. Occlusion-aware face inpainting via generative adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1202–1206. IEEE, 2017.

[5] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[6] A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212, 2004.

[7] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7093–7102, 2018.

[8] B. Egger, S. Schönborn, A. Schneider, A. Kortylewski, A. Morel-Forster, C. Blumer, and T. Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision*, 126(12):1269–1287, 2018.

[9] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[10] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[12] J. Hays and A. A. Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (ToG)*, 26(3):4–es, 2007.

[13] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[15] F. Juefei-Xu, R. Dey, V. N. Boddeti, and M. Savvides. Rankgan: a maximum margin ranking gan for generating faces. In *Asian Conference on Computer Vision*, pages 3–18. Springer, 2018.

[16] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

[17] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[18] X. Li, G. Hu, J. Zhu, W. Zuo, M. Wang, and L. Zhang. Learning symmetry consistent deep cnns for face completion. *IEEE Transactions on Image Processing*, 29:7641–7655, 2020.

[19] Y. Li, S. Liu, J. Yang, and M.-H. Yang. Generative face completion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3911–3919, 2017.

[20] Z. Li, Y. Hu, R. He, and Z. Sun. Learning disentangling and fusing networks for face completion under structured occlusions. *Pattern Recognition*, 99:107073, 2020.

[21] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[22] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[23] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[24] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. Ieee, 2009.

[25] S. Ploumpis, E. Ververas, E. O'Sullivan, S. Moschoglou, H. Wang, N. Pears, W. Smith, B. Gecer, and S. P. Zafeiriou. Towards a complete 3d morphable model of the human head. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[26] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 497–500, 2001.

[27] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[28] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6296–6305, 2018.

[29] A. Shocher, S. Bagon, P. Isola, and M. Irani. Ingan: Capturing and retargeting the" dna" of a natural image. In *Proceed-*

ings of the IEEE/CVF International Conference on Computer Vision, pages 4492–4501, 2019.

[30] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5541–5550, 2017.

[31] L. Song, J. Cao, L. Song, Y. Hu, and R. He. Geometry-aware face completion and editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2506–2513, 2019.

[32] A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017.

[33] L. Tran, F. Liu, and X. Liu. Towards high-fidelity nonlinear 3d face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1126–1135, 2019.

[34] L. Tran and X. Liu. On learning 3d face morphable model from in-the-wild images. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[35] A. Tuán Trán, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3935–3944, 2018.

[36] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[37] S. Wu, C. Rupprecht, and A. Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2020.

[38] Y. Wu and K. He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[39] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5485–5493, 2017.

[40] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[41] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4480, 2019.

[42] X. Yuan and I. K. Park. Face de-occlusion using 3d morphable model and generative adversarial network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10062–10071, 2019.

[43] J. Zhang, R. Zhan, D. Sun, and G. Pan. Symmetry-aware face completion with generative adversarial networks. In *Asian Conference on Computer Vision*, pages 289–304. Springer, 2018.

[44] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[45] S. Zhang, R. He, Z. Sun, and T. Tan. Demeshnet: Blind face inpainting for deep meshface verification. *IEEE Transactions on Information Forensics and Security*, 13(3):637–647, 2017.

[46] C. Zheng, T.-J. Cham, and J. Cai. Pluralistic image completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019.

[47] T. Zhou, C. Ding, S. Lin, X. Wang, and D. Tao. Learning oracle attention for high-fidelity face completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2020.

[48] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.