Generating Diverse 3D Reconstructions from a Single Occluded Face Image

Rahul Dey Vishnu Naresh Boddeti
Michigan State University
East Lansing, MI
deyrahul,vishnu@msu.edu

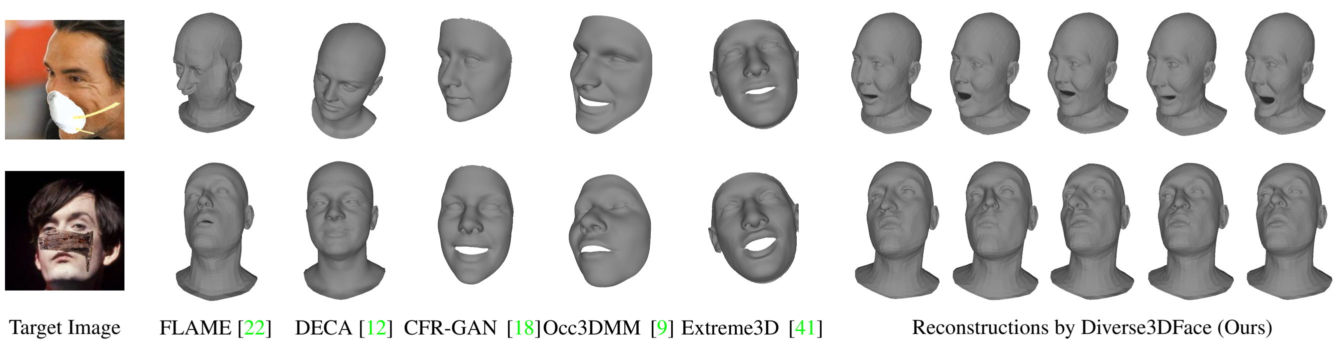


Figure 1. Diverse 3D reconstructions from a single occluded face image by Diverse3DFace *vs.* a singular solution by the baselines.

Abstract

Occlusions are a common occurrence in unconstrained face images. Single image 3D reconstruction from such face images often suffers from corruption due to the presence of occlusions. Furthermore, while a plurality of 3D reconstructions is plausible in the occluded regions, existing approaches are limited to generating only a single solution. To address both of these challenges, we present Diverse3DFace, which is specifically designed to simultaneously generate a diverse and realistic set of 3D reconstructions from a single occluded face image. It consists of three components: a global+local shape fitting process, a graph neural network-based mesh VAE, and a Determinantal Point Process based diversity promoting iterative optimization procedure. Quantitative and qualitative comparisons of 3D reconstruction on occluded faces show that Diverse3DFace can estimate 3D shapes that are consistent with the visible regions in the target image while exhibiting high, yet realistic, levels of diversity on the occluded regions. On face images occluded by masks, glasses, and other random objects, Diverse3DFace generates a distribution of 3D shapes having ∼50% higher diversity on the occluded regions compared to the baselines. Moreover, our closest sample to the ground truth has ∼40% lower MSE than the singular reconstructions by existing approaches.

1. Introduction

Single image-based 3D face reconstruction has improved significantly in recent years [10,53]. This includes advances in statistical models [3,22,27,28] as well as neural network-based models [12, 13, 32, 37–40, 44, 45]. However, facial occlusions remain a significant challenge to these tasks. In-the-wild face images often come with several forms of occlusions, and unless dealt with explicitly, often lead to erroneous 3D reconstruction in terms of shape, expression, pose, *etc*. [9, 10, 41].

3D reconstruction of partially occluded faces presents two main challenges. First, 3D reconstruction models need to selectively use features from the visible regions while ignoring those from the occluded parts. Failure to do so, either implicitly or explicitly, will lead to poor 3D reconstructions with an incorrect pose, expression, or both. Second, there could be a distribution of 3D reconstructions that are consistent with the visible parts in the image, yet diverse on the occluded parts. Failure to account for all such modes limits the utility of 3D reconstruction models. Addressing these two challenges is the primary goal of this paper.

Existing 3D face reconstruction solutions, however, are ill-equipped to overcome both of these challenges simultaneously. From a **reconstruction perspective**, a majority of the approaches that reconstruct 3D faces from a single image restrict themselves to fully-visible face images. And, even those that explicitly account for facial occlusions [9, 41], do so only in a holistic manner using a global model that implicitly uses features from the occluded regions as well. This form of global model-based fitting can lead to large errors (see Fig. 1) in the pose and expression of the 3D reconstruction, especially when large portions of the face are occluded. From a **diversity perspective**, existing approaches are, by design, limited to only generating a single plausible 3D reconstruction. However, in many practical applications, for a single occluded face image, it is desirable to generate multiple reconstructions that are consistent on the visible parts of the face, while spanning a diverse, yet realistic, set of reconstructions on the occluded parts (see Fig. 1). While the concept of generating diverse solutions has been explored in other contexts such as image generation [11], image completion [50], super-resolution [1] and trajectory forecasting [48], they have not been employed for monocular 3D face reconstruction of occluded faces.

In this paper, we propose Diverse3DFace which is designed to simultaneously yield a diverse, yet plausible set of 3D reconstructions from a single occluded face image. Diverse3DFace consists of three modules: a global + local shape fitting process, a graph neural network based variational autoencoder (Mesh-VAE), and a Determinantal Point Process (DPP) [21] based iterative optimization procedure. The global + local shape fitting process affords robustness against large occlusions by decoupling shape fitting on the visible regions from that of the occluded regions. The Mesh-VAE enables to learn a distribution over a compact latent space over the different factors of variation in the 3D shapes of faces. And, the DPP based iterative optimization procedure enables us to sample from the latent space of the Mesh-VAE and optimize them to generate a diverse set of reconstructions spanning the different modes of the latent space. Our specific contributions in this paper are:

– We propose Diverse3DFace, a simple yet effective diversity promoting 3D face reconstruction approach that generates multiple plausible 3D reconstructions corresponding to a single occluded face image.

– For robustness to occlusions, we propose a global + local PCA model based shape fitting that disentangles the fitting on each facial component from the others. The models are learned from a dataset of FLAME [22] registered 3D meshes. During inference, the local perturbations on various facial components are added on top of a coarse global fit to generate the final detailed fitting.

– We employ a DPP [21] based diversity loss in the context of generating diverse 3D reconstructions of faces. We define the quality and similarity terms in the DPP kernel so as to maximize diversity while remaining in the space of realistic 3D head shapes.

– We conduct extensive qualitative and quantitative experiments to show the efficacy of the proposed approach in generating 3D reconstructions that are faithful to the visible face, while simultaneously capturing multiple diverse modes on the occluded parts. The solution from Diverse3DFace that is closest to the ground truth is on average 30-50% better than the unique solutions of the baselines [12, 22] in terms of per-vertex L2-error.

2. Related Work

Single Image 3D Face Reconstruction: Blanz and Vetter [3] proposed the first 3DMM model of human faces. Since then, such models have grown to include complex pose and expression modalities in 3D faces [14,27]. Li *et al.* [22] proposed FLAME that models the full human head and allows non-linear control over joint poses to generate articulated expressive head instances. Many recent approaches adopted neural networks to model higher order complexities in the shape and expression spaces [12,19,29,31–33,37–40,45]. A few methods took a hybrid approach of fitting a non-linear neural network model to the target image to generate detailed 3D reconstructions [13, 47]. More recently, advances in graph neural networks [8, 20, 25, 42] have propagated using graph convolution operations to directly learn non-linear representation on a mesh surface while preserving the mesh topology [4, 30, 51]. Though these advances have significantly improved the modeling capabilities of 3D face reconstruction approaches, they are still limited when it comes to handling occlusions in the face images.

On the other hand, a few approaches are explicitly designed to handle such occlusions [9, 18, 41]. Tran *et al.* [41] trained a neural network to regress a robust foundation shape from a masked face image, over which a detailed bump map is later added. And, Egger *et al.* simultaneously optimized an occlusion mask and the model parameters from an occluded image. However, these approaches rely on a global model to account for the entire face, including the occluded parts, which is sub-optimal as the lack of information from such parts needs to be countered using strong regularization. Moreover, they are limited to reconstructing a singular 3D solution without considering the plurality of solutions that can explain the occluded regions. In contrast, we address the dual problems of robustness and lack of uniqueness through a multistage approach that disentangles fitting on the visible regions from diversity modeling on the occluded ones.

Diversity Promoting Generative Models: Diversity promoting algorithms have been employed in several areas in computer vision where a distribution of outcomes is more desirable than a singular solution. Conditioning [17, 46] and regularization [5, 15, 35, 36, 52] based techniques have been found to be useful to overcome mode-collapse and promote diveristy in GANs [16]. As ill-posed problems, diversity promoting algorithms are also particularly useful for image completion and image super-resolution. Zheng *et al.* [50] proposed a dual-pipeline C-VAE [34] that maintains groundtruth fidelity in one path, while allowing diversity on the other. While, Bahat *et al.* [1] generated diverse super-resolution explanations by only enforcing consistency in the low-resolution space. Compared to image based approaches that focus on diversity in the texture, 3D reconstruction requires modeling geometric diversity. As one of the most seminal works in this field, Kulesza and Taskar [21] introduced the framework of Determinantal Point Processes (DPPs) to model diversity in machine learning tasks such as inference, sampling, marginalization, *etc*. Yuan *et al.* [48, 49] adopted DPP to sample multi-modal latent vectors for diverse human trajectory forecasting. Elfeki *et al.* [11] devised a DPP based objective to train GANs and VAEs to emulate the diversity in real data. In this work, we adopt the idea of DPPs to generate diverse 3D reconstructions for an occluded face by discovering latent space representations that maximize plausible diversity on the occluded regions while remaining faithful to the visible parts.

3. Background

Statistical Models of 3D Face Reconstruction: Statistical 3D models such as BFM [3, 27] and FLAME [22] allow for generating new face instances. These models often consist of a *shape model* that explain geometric variations across identities, an *expression model* that accounts for variations due to different facial expressions, and additionally a *pose model* and an *appearance model* to account for variations in pose and appearance, respectively. Specifically, FLAME [22] defines a 3D shape as

$$S(\beta, \theta, \psi) = W(T(\beta, \theta, \psi), \mathbf{J}(\beta), \theta, \mathcal{W}), \tag{1}$$

where the parameters β, θ, ψ represent the shape, pose and expression parameters, respectively; $\mathbf{J} \in \mathbb{R}^{\mathbf{3K}}$ represents the locations of K face joints around which $T(\beta, \theta, \psi)$ is rotated, and finally smoothened by the blendweights $\mathcal{W}$. The un-aligned shape $T(\beta, \theta, \psi)$ is obtained by adding up the contributions of shape, expression and pose variations on top of a template shape $\bar{\mathbf{T}}$:

$$T(\beta, \theta, \psi) = \bar{\mathbf{T}} + B_S(\beta; \mathcal{S}) + B_P(\theta; \mathcal{P}) + B_E(\psi; \mathcal{E}) \tag{2}$$

The shape and expresssion variations are modeled by linear blendshapes $B_S(\beta; \mathcal{S}) = \mathcal{S}\beta$, $B_E(\psi; \mathcal{E}) = \mathcal{E}\psi$ where $\mathcal{S} \in \mathbb{R}^{3N \times |\beta|}$ and $\mathcal{E} \in \mathbb{R}^{3N \times |\psi|}$ are orthonormal shape and expression bases learned using PCA and N is the number of vertices. The pose blendshape function is defined as $B_P(\theta; \mathcal{P}) = (R(\theta) - R(\theta*))\,\mathcal{P}$, where $R(\theta)$ comprises of rotation matrices around the K joints and $\mathcal{P} \in \mathbb{R}^{3N \times 9K}$ are the pose blendshapes describing the vertex offsets from the rest pose activated by R.

Determinantal Point Processes: Determinantal Point Processes (DPPs) originated in quantum physics to model the negative correlations between quantum states of fermions [24]. DPPs were first introduced in machine learning by Kulesza and Tascar [21] as a probabilistic model of repulsion between points. A point process over a ground set $\mathcal{Y}$ describes the probability of all its $2^{\mathcal{Y}}$ subsets. A point process is determinantal when the probability of choosing a random subset $Y \subseteq \mathcal{Y}$ is given by the determinant of the sub-kernel matrix $\mathbf{L}_Y$ indexed by the elements of Y, *i.e.*, $P(Y \subseteq \mathcal{Y}) = det(\mathbf{L}_Y)$. Given a data matrix $B \in \mathbb{R}^{D \times N}$, we can compute the kernel as the Gram matrix $\mathbf{L} = B^T B$. In this case, the determinant of the sub-kernel matrix $det(\mathbf{L}_Y)$ is related to the volume spanned by the elements of B. Thus, conceptually, DPP assigns a higher probability to a subset whose elements tend to be orthogonal (diverse) to each other, thus spanning a larger volume.

4. Approach

Reconstructing diverse 3D fittings in a single stage, using a global model, is ineffective due to multiple reasons, as we show in our experiments (see Sec. 5.1). First, fitting a global model to a few visible sub-regions is sub-optimal, since the model has to strike a trade-off between robustness and local details. Secondly, diversification of the occluded regions will inadvertently affect the quality of fitting on the visible regions, and vice-versa. In view of this, we propose a three-step approach to generate diverse and realistic 3D reconstructions from an occluded face image. In step 1, we use an ensemble of disentangled global+local shape models to perform robust 3D reconstruction with respect to the visible parts in the face. In step 2, we employ a VAE to map the partial fit to a latent space from which multiple reconstructions can be drawn. In the third step, we iteratively optimize the latent embeddings to promote realistic geometric diversity on the occluded face regions while maintaining fidelity to the visible regions. We now describe the different components along with the full algorithm, below.

4.1. Global + Local Shape Model

A robust partial 3D reconstruction that fits accurately with the visible parts is a prerequisite for generating diverse solutions. Existing approaches of occlusion-robust 3D reconstruction, both fitting and regression based, typically exclude the occluded regions during inference [9, 41]. However, they still use a global model to do so. This is ineffective as a global model assumes the presence of full face for 3D reconstruction, lacking which, such models often introduce artifacts to the reconstructed shape (see Fig. 5). To mitigate this, such approaches impose high levels of regularization which leads to sub-optimal shape fitting. This observation motivated us to use an ensemble of global + local models as an effective approach to generate robust 3D reconstructions with respect to the visible parts. Note that, in this stage, we are not concerned about the reconstruction

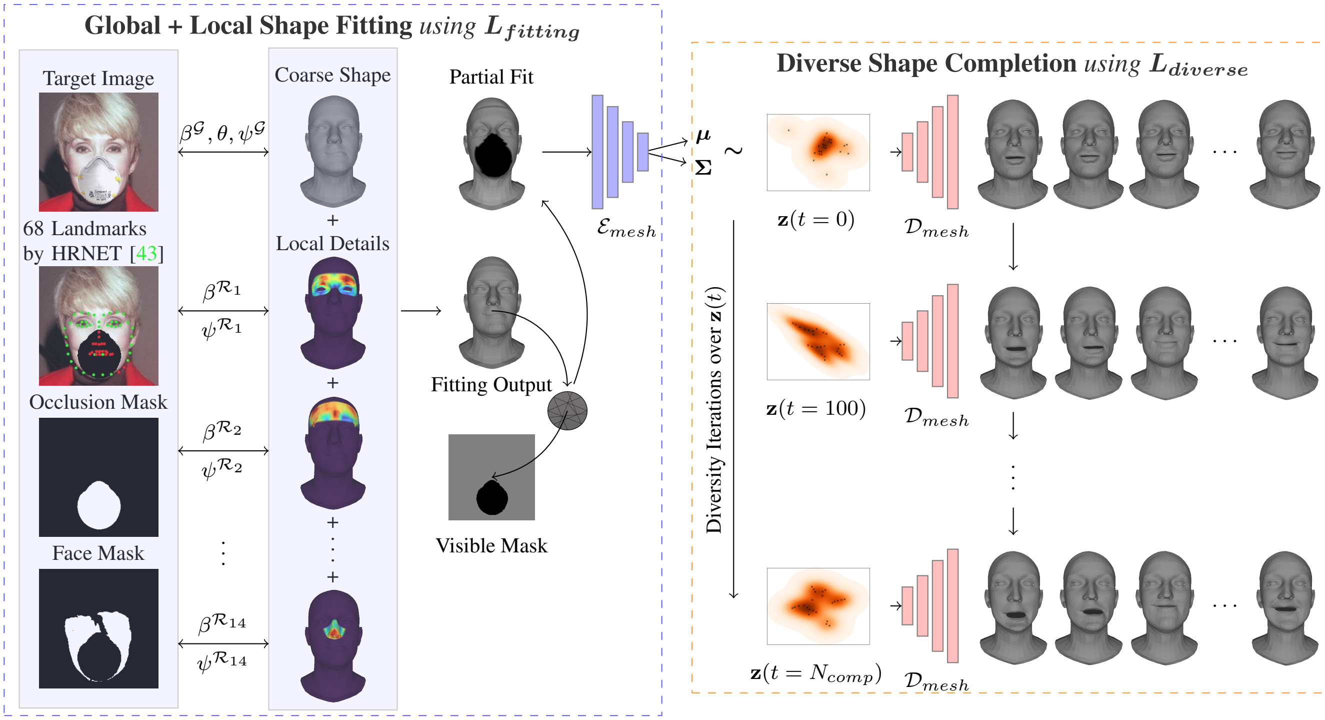


Figure 2. **Overview:** As input, we need the target image, the occlusion mask, facial landmarks, and optionally a face mask. We use the HRNET model [43] to obtain both the landmark locations and their confidence values, which we use to estimate the occlusion labels. Given these input, we first fit our proposed *global + local blendshape model* to obtain the coarse and local fittings as outlined in Algorithm 1, which we then add together to obtain the final fitting. We re-project the fitted shape onto the visible mask to obtain a partial fit, zeroed out on the occluded regions. We map the partial fit onto a latent space using the *Mesh-VAE* encoder $\mathcal{E}_{mesh}$ and sample N latent vectors $\mathbf{z}$. We then iteratively optimize the $\mathbf{z}$'s to capture diverse modes with respect to the occluded regions while remaining consistent with the visible regions as outlined in Algorithm 2 to obtain the final set of 3D reconstructions.

quality in the occluded regions. We now describe the details of our proposed global+local 3D head model.

Our global+local shape model is based on the FLAME mesh topology [22]. We use the FLAME registered D3DFACS [7] and CoMA [30] datasets to train the local PCA models. The FLAME [22] model comes with vertex masks corresponding to 14 parts on the human head. We trained individual PCA models corresponding to each of these parts to explain for the local variations. To do so, we first take FLAME-registered meshes and fit the full FLAME model [22] to these by optimizing the following fitting loss:

$$L_{fit} = \min_{\beta,\theta,\psi} ||S^{gt} - \tilde{S}(\beta,\theta,\psi)||, \tag{3}$$

Here $\tilde{S}(\beta,\theta,\psi)$ is obtained using Eqs. (1) and (2). We then *unpose* both the ground-truth and the fitted shapes by removing the variations due to pose θ as described in [22] and obtain S_0^{gt} and $\tilde{S}(\beta,0,\psi)$, respectively. The full FLAME model consists of $|\beta|= 300$ shapes and $|\psi|= 100$ expression bases to account for complete global variations. From this, we retain the top N_S shape and N_E expression bases (based on eigenvalues) and discard the rest to compute shape residuals $\tilde{S}^{res} = S_0^{gt} - \tilde{S}^{course}$, where

$$\tilde{S}^{course} = \bar{\mathbf{T}} + \sum_{n=1}^{N_S} \beta_n \mathcal{S}_n + \sum_{n=1}^{N_E} \psi_n \mathcal{E}_n \tag{4}$$

We then compute the region-wise shape and expression PCA models $(\mathcal{S}^{\mathcal{R}_i}, \mathcal{E}^{\mathcal{R}_i})$ using the region-wise residuals $M_{\mathcal{R}_i} \odot \tilde{S}^{res}$ (here $M_{\mathcal{R}_i}$ is the vertex-mask for the ith region). For computing the shape bases, we set $N_S = 10$ and $N_E = 100$ (removing all expression variations); while for the expression bases, we set $N_E = 10$ and $N_S = 300$ (removing all identity variations). The global + local model can then be represented as,

$$T(\beta^{\mathcal{G}}, \beta^{\mathcal{R}_i}, \theta, \psi^{\mathcal{G}}, \psi^{\mathcal{R}_i}) = T_{\mathcal{G}}(\beta^{\mathcal{G}}, \theta, \psi^{\mathcal{G}}) + T_{\mathcal{R}}(\beta^{\mathcal{R}_i}, \psi^{\mathcal{R}_i}), \tag{5}$$

where $T_{\mathcal{G}}(\beta^{\mathcal{G}}, \theta, \psi^{\mathcal{G}})$ is the course global shape given by the top N_S shape and N_E expression global blendshapes along with the pose blendshapes $\mathcal{P}$ (Eq. (2)); and $T_{\mathcal{R}}(\beta^{\mathcal{R}_i}, \psi^{\mathcal{R}_i})$ represent the local variations and is given by,

$$T_{\mathcal{R}}(\beta^{\mathcal{R}_i}, \psi^{\mathcal{R}_i}) = \sum_{\mathcal{R}_i} \left(\sum_{n=1}^{|\beta^{\mathcal{R}_i}|} \beta_n^{\mathcal{R}_i} \mathcal{S}_n^{\mathcal{R}_i} + \sum_{n=1}^{|\psi^{\mathcal{R}_i}|} \psi_n^{\mathcal{R}_i} \mathcal{E}_n^{\mathcal{R}_i} \right) \tag{6}$$

4.2. Shape Completion using Mesh-VAE

We use the global+local model to fit robust 3D reconstruction on the visible parts of the occluded face. But this does not guarantee robust and consistent reconstruction on the occluded parts, since the local PCA models have noisy (occluded) or no data to fit to. To address this, and to enable the generation of a distribution of plausible 3D reconstructions rather than a singular solution, which is one of our primary goals, we adopt a mesh based VAE (dubbed *Mesh-VAE*) as our shape completion model.

We assume that human head meshes can be mapped onto a continuous and regularized low-dimensional latent space $\mathcal{Z}$. Then, given a partial 3D mesh $\mathbf{S}_m$, the Mesh-VAE learns the conditional likelihood of mesh completions $\mathbf{S}_c$ and the corresponding latent embeddings $\mathbf{z}$:

$$p(\mathbf{S}_c, \mathbf{z}|\mathbf{S}_m) = p(\mathbf{z}|\mathbf{S}_m)p(\mathbf{S}_c|\mathbf{z}, \mathbf{S}_m), \tag{7}$$

4.3. DPP Driven Shape Diversification

Even though the Mesh-VAE can sample multiple shape completions from $p(\mathbf{S}_c|\mathbf{z}, \mathbf{S}_m)$, in practice, the generated samples from a VAE are not guaranteed to cover all the modes [48] (see Sec. 5.1). To guarantee diversity, we formulate a DPP on shape completions and develop a diversity loss to optimize their latent embeddings.

We adopt the quality-diversity based forumlation of the DPP kernel $\mathbf{L}$ [21] which soughts to balance the quality of samples with their diversity. Specifically, for elements i, j in a set, its kernel entry is given by $L_{i,j} = q_i S_{i,j} q_j$, where q_i denotes the quality of element i, and $S_{i,j}$ represents the similarity between i and j. Maximizing the determinant of such a kernel matrix implies maximizing the quality of each sample, while minimizing the similarity between distinct samples. For two shape completions $\mathbf{S}^i_c$ and $\mathbf{S}^j_c$, we define the similarity as

$$S_{i,j} = \exp(-k||\mathbf{S}^i_c - \mathbf{S}^j_c||_2), \tag{8}$$

where k is a scaling factor. To ensure that the completed samples look realistic, we relate the quality of a sample with the probability of its latent embedding $\mathbf{z}_i$ lying within 3σ of the prior $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and define it as:

$$q_i = \exp(-\max(0, \mathbf{z}_i^T \mathbf{z}_i - 3\sqrt{d})), \tag{9}$$

where d is the dimensionality of $\mathbf{z}_i$. For numerical stability [48], we adopt expected cardinality of $\mathbf{L}$ as the DPP loss:

$$L_{dpp} = -tr\left(\mathbf{I} - (\mathbf{L} + \mathbf{I})^{-1}\right) \tag{10}$$

4.4. Inference

Given an occluded face image $\mathbf{I}_m$, our goal is to generate a distribution of plausible 3D reconstructions $\mathbf{S}^1_c, ..., \mathbf{S}^M_c$. We do this in three steps which we describe below:

Step 1 *Partial Shape Fitting:* In this stage, we first fit our global + local PCA model on the visible parts of the face image $\mathbf{I}_m$ to obtain a partial reconstruction $\mathbf{S}_m$. We employ the following fitting loss:

$$L_{fitting} = \lambda^f_1 L_{lmk} + \lambda^f_2 L_{pho} + \lambda^f_3 L_{reg}, \tag{11}$$

where L_{lmk} is the landmark loss, L_{pho} is the photometric loss and L_{reg} applies $L2$-regularization over the model parameters. We use an off-the-shelf landmark detector HR-NET [43] to detect 68 landmarks on the face along with their confidence values. We mark those landmarks as visible whose confidence exceeds a threshold τ (set to 0.2), and apply the landmark loss on those points. To add local details, we apply a photometric loss between the input image and a rendered image $\mathbf{I}_{ren} = \mathcal{R}(\mathbf{S}_m, B_{tex}(\gamma, \mathcal{T}), c)$, where $B_{tex}(\gamma, \mathcal{T})$ is the estimated texture and c the estimated camera parameters. We restrict the photometric loss on the visible face regions using the face mask M_f and the occlusion mask M_o:

$$L_{pho} = ||(\mathbf{I}_m - \mathbf{I}_{ren}) \odot M_f \odot (\mathbf{1} - M_o)||_1 \tag{12}$$

Step 2 We use the encoder to map the partial fit $\mathbf{S}_m$ to a latent distribution from which we sample the latent embeddings $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma^2 I})$, where $\boldsymbol{\mu}, \boldsymbol{\sigma} = \mathcal{E}_{mesh}(\mathbf{S}_m)$.

Step 3 *Diversity Promoting Shape Completion:* In this stage, we perform a diversity promoting iterative shape completion routine which forces the latent embeddings towards diverse modes with respect to the occluded regions while remaining faithful to the visible regions. At each iteration, we obtain a distribution of shape completions using the decoder $\mathbf{S}^j_c = \mathcal{D}_{mesh}(\mathbf{z}_j), \forall j = 1...M$, and update the $\mathbf{z}$'s to minimize a diversity loss:

$$L_{diversity} = \lambda_1 L_S + \lambda_2 L_{pho} + \lambda_3 L_{dpp} \tag{13}$$

Here L_S is the L1 loss between $\mathbf{S}^j_c$ and $\mathbf{S}_m$ applied on the visible vertices, L_{pho} is the photometric loss (Eq. (12)) and L_{dpp} is the DPP loss (Eq. (10)).

We outline the full steps for partial fitting and shape completion diversification in Algorithm 1 and Algorithm 2, respectively.

5. Experimental Evaluation

Datasets: We use the FLAME [22] registered head meshes from the CoMA [30] and D3DFACS [7] datasets for training and evaluation. We split the two datasets into 80:10:10 train:val:test splits based on subject ID. We train the Mesh-VAE model using the combined training splits from the two datasets. During training, we augment the meshes with contiguous occlusion masks of random shapes and locations. To evaluate our approach, we use the test split of the CoMA dataset [30] consisting of subjects that were excluded from

Algorithm 1 Shape Fitting on the Visible Face Regions

Input: Image $\mathbf{I}_m$, Occlusion mask M_o, Face mask $\mathbf{M}_f$, Global models $\mathcal{S}, \mathcal{E}, \mathcal{P}$, Local models $\mathcal{S}^{\mathcal{R}_i}$, $\mathcal{E}^{\mathcal{R}_i}$ for $i =$ 1 to 14, Texture model $\mathcal{T}$, Landmarks detector $\mathcal{H}$
Parameters: $\beta, \theta, \psi, \gamma, c, \beta^{\mathcal{R}_i}, \psi^{\mathcal{R}_i}$ for $i = 1$ to 14
Hyperparameters: $\tau = 0.1, n_{iter}, \lambda_1^f, \lambda_2^f, \lambda_3^f, \eta$
Output: Partially fitted shape $\mathbf{S}_m$

Detect landmarks from image $\mathbf{L}_I, \mathbf{L}_{conf} \leftarrow \mathcal{H}(\mathbf{I}_m)$
Set $\mathbf{L}_{valid} \leftarrow 1$ when $\mathbf{L}_{conf} > \tau$ else 0
for $j = 1$ to n_{iter} **do**
  Obtain $\mathbf{S}_m$ using Eqs. (1), (2), (5) and (6)
  Select 68 landmarks from shape $\mathbf{L}_S \leftarrow M_{lmk}(\mathbf{S})$
  Obtain rendered image $\mathbf{I}_{ren} \leftarrow \mathcal{R}(\mathbf{S}, B_{tex}(\gamma, \mathcal{T}), c)$
  $L_{lmk}^f \leftarrow ||(\mathbf{L}_S - \mathbf{L}_I) \odot \mathbf{L}_{valid}||_1$
  $L_{pho}^f \leftarrow ||(\mathbf{I}_m - \mathbf{I}_{ren}) \odot \mathbf{M}_f \odot (\mathbf{1} - M_o)||_1$
  $L_{reg}^f \leftarrow$L2 regularization loss over all parameters
  $L_{fitting} = \lambda_1^f L_{lmk}^f + \lambda_2^f L_{pho}^f + \lambda_3^f L_{reg}^f$
  Update $p \leftarrow p - \eta \nabla_p L_{fitting}$ for $p \in \beta, \theta, \psi, \gamma, c, \beta^{\mathcal{R}_i}, \psi^{\mathcal{R}_i}$ for $i = 1$ to 14
end for

Algorithm 2 Diverse Shape Completions

Input: Mesh-VAE Encoder $\mathcal{E}_{mesh}$ and Decoder $\mathcal{D}_{mesh}$; From Algorithm 1: $\mathbf{I}_m, M_o, \mathbf{M}_f, \mathbf{L}_I, \mathbf{L}_{valid}, \theta, \gamma, c, \mathcal{T}$
Hyperparameters: $n_{comp}, \lambda_1, \lambda_2, \lambda_3, \eta$
Output: M Shape completions $\{\mathbf{S}_c^{j=1:M}\}$

Sample the vertex mask M_o^v by projecting $\mathbf{S}$ onto M_o
Obtain latent parameters $\boldsymbol{\mu}, \boldsymbol{\sigma} \leftarrow \mathcal{E}_{mesh}(\mathbf{S}_m \odot M_o^v)$
Sample M latent vectors $\mathbf{z}_1, ..., \mathbf{z}_M \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2 \boldsymbol{I})$
for $i = 1$ to n_{comp} **do**
  Obtain $\mathbf{S}_c^j \leftarrow \mathcal{D}_{mesh}(\mathbf{z}_j)$ for $j = 1...M$
  Obtain $\mathbf{I}_{ren,j} \leftarrow \mathcal{R}(\mathbf{S}_c^j, B_{tex}(\gamma, \mathcal{T}), c)$ for $j = 1...M$
  $L_S \leftarrow \sum_{j=1}^{M} ||(\mathbf{S}_c^j - \mathbf{S}_m) \odot (\mathbf{1} - M_o^v)||_1$
  $L_{pho} \leftarrow \sum_{j=1}^{M} ||(\mathbf{I}_m - \mathbf{I}_{ren,j}) \odot \mathbf{M}_f \odot (\mathbf{1} - M_o)||_1$
  $L_{dpp} \leftarrow \mathcal{L}_{dpp}(\mathbf{S}_c^{j=1:M} \odot M_o^v)$ using Eq. (10)
  $L_{diversity} = \lambda_1 L_S + \lambda_2 L_{pho} + \lambda_3 L_{dpp}$
  Update $\mathbf{z}_j \leftarrow \mathbf{z}_j - \eta \nabla_{\mathbf{z}_j} L_{diversity}$ for $j = 1$ to M
end for

training. We do further qualitative evaluation on the unannotated images from the CelebA dataset [23]. For both datasets, the test images are artificially augmented with occlusions such as masks, glasses and other random objects.

Implementation: We implement the Mesh-VAE as a fully convolutional graph neural network (GNN) based upon the MeshConv architecture presented in [51]. MeshConv [51] uses spatially varying convolution kernels to account for the irregularity of local mesh structures, and was shown to outperform fixed kernel based GNN approaches [4, 8, 20, 25, 30, 42] on reconstruction tasks. To train Mesh-VAE as a shape completion model, we augment the training meshes with random continuous masks covering 25-40% of the vertices. In practice, however, directly training the Mesh-VAE for inpainting is very challenging, especially with large degrees of occlusions. We adopt a curriculum learning [2] approach to overcome this challenge and progressively introduce larger occlusions during the training process, i.e., we start with easier shape completion tasks and progressively increase its difficulty. We use a combination of L1-reconstruction, L1-Laplacian and the KL-divergence losses to train the network.

Baselines: To evaluate the efficacy of Diverse3DFace in terms of diversity and robustness to occlusions, we perform evaluation against baselines such as FLAME [22], DECA [12], CFR-GAN [18], Occ3DMM [9] and Extreme3D [41] using their publicly available codes and pretrained models (whereever applicable). Due to the difficulty and unreliability in obtaining dense correspondence between FLAME and other mesh topologies, we perform quantitative comparison against only the FLAME [22] topology based methods. For the other baselines, we perform qualitative evaluation of 3D reconstruction from various occluded face images.

Metrics: The goal of this paper is to generate diverse, yet realistic 3D reconstructions on occluded face images. Such an approach should have three desired qualities: 1) the reconstructed shapes should fit as accurately as possible to the visible regions, 2) the occluded regions should be diverse from each other, and 3) at least one of the reconstructed shape should be very similar to the ground truth shape. There is no prior work on diverse 3D reconstruction, and as such there are no established metrics. So we define the following three metrics to evaluate the aforementioned qualities: (1) **Closest Sample Error (CSE)**: the per-vertex *L2*-error between the ground-truth shape and the closest reconstructed shape (lower is better), (2) **Average Self Distance-Visible (ASD-V)**: the per-vertex *L2*-distance on the visible regions between a 3D completion and its closest neighbor, averaged across all the samples (lower is better), and (3) **Average Self Distance-Occluded (ASD-O)**: ASD on occluded regions (higher is better). These metrics are inspired from those defined for diverse trajectory forecasting [48].

Occlusion	DECA [12]	FLAME [22]	Global+Local (Ours)
Glasses	57.83	47.89	**39.98**
Face-mask	61.18	30.37	**30.11**
Random	70.34	47.56	**38.27**
Overall	62.91	41.24	**35.85**

Table 1. Evaluation of 3D reconstruction accuracy in terms of mean shape error (MSE) $\times 10^{-3}$ on the CoMA dataset [30].

5.1. Quantitative Results

Tab. 1 reports the 3D reconstruction accuracy in terms of mean shape error (MSE) on artificially occluded test images from the CoMA dataset [30] for different approaches using

Occlusion Type	FLAME			Global+Local			Mesh-VAE			Diverse3DFace (Ours)		
	CSE (↓)	**ASD-V** (↓)	**ASD-O** (↑)	**CSE** (↓)	**ASD-V** (↓)	**ASD-O** (↑)	**CSE** (↓)	**ASD-V** (↓)	**ASD-O** (↑)	**CSE** (↓)	**ASD-V** (↓)	**ASD-O** (↑)
Glasses	41.26	3.83	3.26	38.17	2.25	3.11	**32.88**	1.01	1.38	36.30	**0.61**	**4.50**
Face-mask	28.14	3.07	4.58	28.06	2.30	3.57	**25.95**	0.89	1.79	27.58	**0.85**	**7.89**
Random	43.12	3.61	4.06	38.85	2.59	3.51	**36.58**	0.97	1.61	39.11	**0.72**	**5.62**
Overall	36.81	3.61	4.06	34.55	2.35	3.39	**31.18**	0.95	1.59	33.71	**0.73**	**6.05**

Table 2. Evaluation of diverse 3D reconstructions by the baselines *vs.* Diverse3DFace in terms of CSE, ASD-V and ASD-O (in order of 10^{-3}) on the CoMA dataset [30].

Occlusion Type	FLAME			Global+Local			Mesh-VAE			Diverse3DFace (Ours)		
	ASD-V (↓)	**ASD-O** (↑)	$\frac{\text{ASD-O}}{\text{ASD-V}}$(↑)	**ASD-V** (↓)	**ASD-O** (↑)	$\frac{\text{ASD-O}}{\text{ASD-V}}$(↑)	**ASD-V** (↓)	**ASD-O** (↑)	$\frac{\text{ASD-O}}{\text{ASD-V}}$(↑)	**ASD-V** (↓)	**ASD-O** (↑)	$\frac{ASD-O}{ASD-V}$(↑)
Glasses	3.44	2.98	0.866	2.15	2.99	1.391	0.81	1.17	1.444	**0.68**	**3.56**	**5.235**
Face-mask	3.45	4.93	1.429	2.85	3.99	1.400	**0.75**	1.62	2.160	1.03	**7.47**	**7.252**
Random	4.12	4.23	1.027	3.17	3.84	1.211	**0.79**	1.29	1.633	0.83	**4.30**	**5.181**
Overall	3.86	4.44	1.150	3.03	3.88	1.281	**0.78**	1.41	1.808	0.90	**5.41**	**6.011**

Table 3. Evaluation of diverse 3D reconstructions by the baselines *vs.* Diverse3DFace in terms of ASD-V and ASD-O (in order of 10^{-3}) and the ratio ASD-O/ASD-V on the CelebA dataset [23].

the FLAME [22] topology. Across all occlusion types, our proposed global+local model reports the lowest MSE values. The large gap between FLAME (fitting) [22], DECA [12] and our approach demonstrates the necessity of region-specific model fitting for occlusion robustness.

Due to the lack of existing diverse 3D reconstruction approaches, we formulate three baselines to evaluate the diversity performance of Diverse3DFace: 1) fitting FLAME on the visible parts with an additional DPP loss on the occluded parts, 2) fitting our global+local model in the same way as (1), and 3) shape completions by sampling from the conditional latent distribution $p(\boldsymbol{z}|\boldsymbol{S}_m)$ and decoding them through Mesh-VAE's decoder. We report the quantitative metrics on the CoMA dataset [30] in Tab. 2. Further, we report quantitative metrics on the CelebA dataset [23] in Tab. 3 without the *CSE* metric, due to the lack of groundtruth shape annotations. Across all occlusion types, the FLAME and the Global+Local based baselines report much higher *CSE* and *ASD-V,* and lower *ASD-O* than Diverse3DFace. Though Mesh-VAE obtains lower *CSE* than Diverse3DFace, it does so at the cost of reduced diversity in terms of *ASD-O*. On the other hand, Diverse3DFace reports the lowest *ASD-V*, the highest *ASD-O*, and the second lowest *CSE*, satisfying the three desired qualities mentioned earlier. This confirms our hypothesis that explicitly accounting for occlusions and explicitly optimizing for diversity can lead 3D reconstructions that are both more accurate (on visible regions) and more geometrically diverse (on occluded regions). Among the different occlusion types, we report the lowest *CSE* as well as the highest *ASD-O* for face-masks. This is consistent with the fact that human faces have higher variability in the mouth and nose regions, which

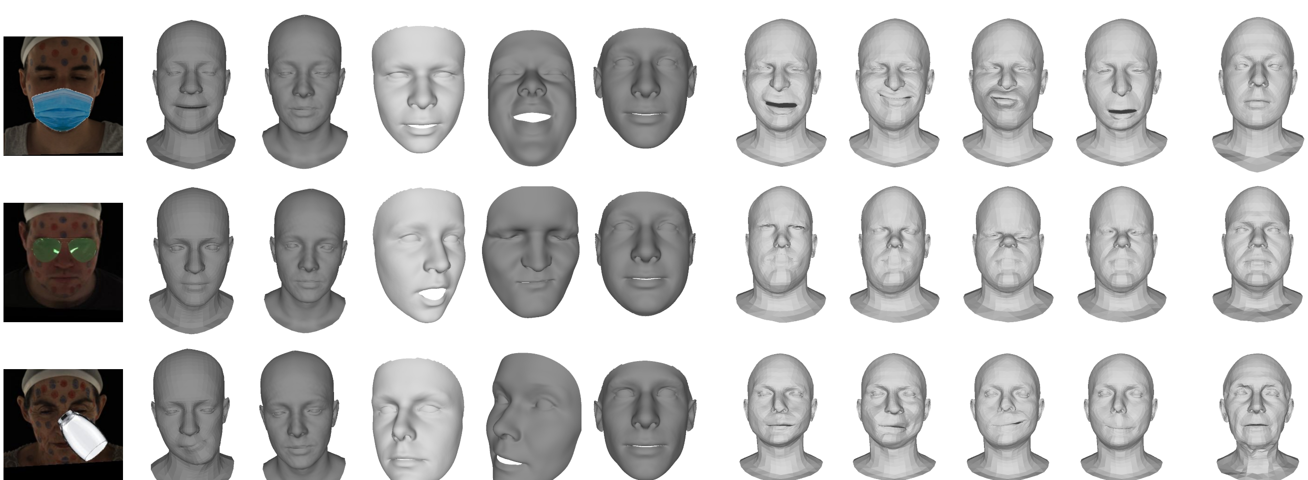


Figure 3. **Qualitative evaluation on the CoMA dataset [30]**: Reconstructed singular 3D meshes from the target image by the baselines *vs.* the diverse reconstructions from Diverse3DFace.

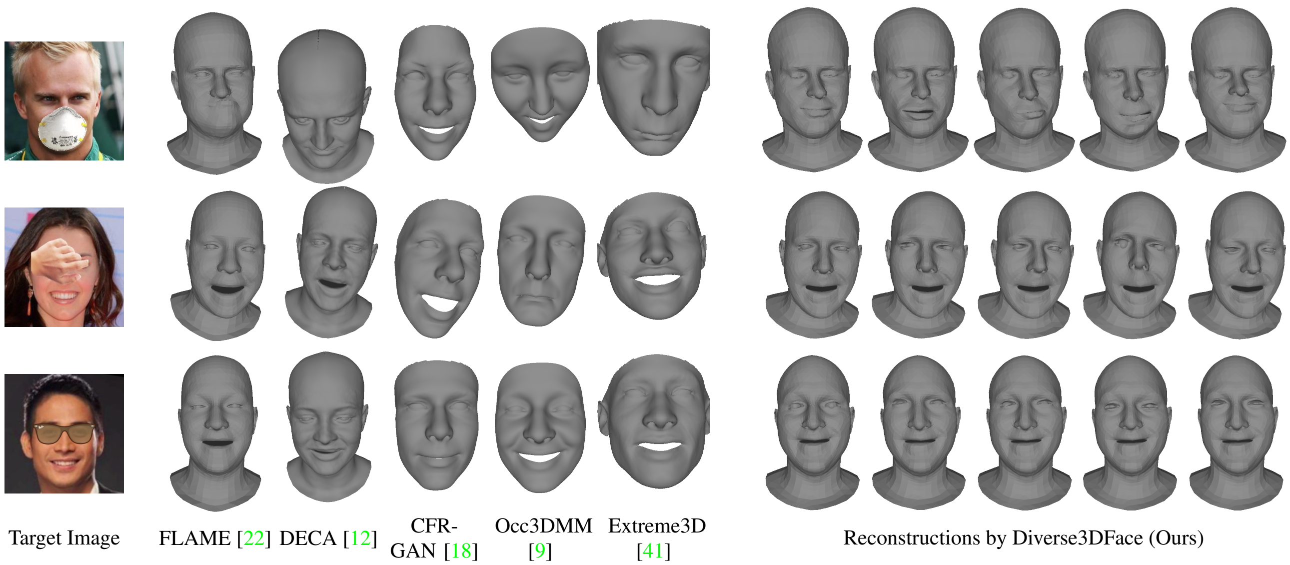


Figure 4. **Qualitative evaluation on the CelebA dataset [23]**: Reconstructed singular 3D meshes from the target image by the baselines *vs.* the diverse reconstructions from Diverse3DFace.

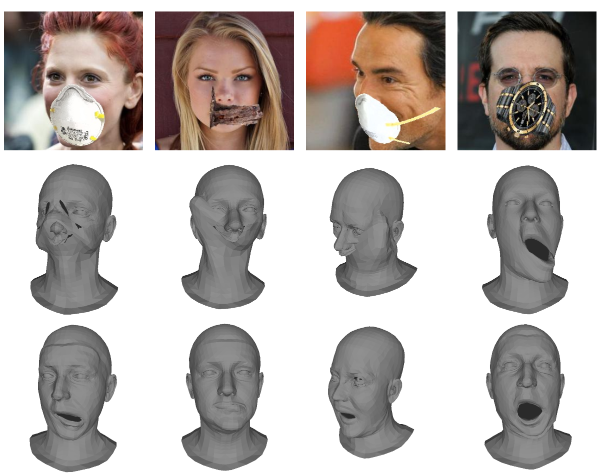


Figure 5. FLAME [22] based fitting (middle row) *vs.* our Global+Local fitting (last row) on occluded face images (top row).

our approach is able to learn and reproduce.

5.2. Qualitative Results

Fig. 3 shows qualitative results of 3D reconstruction on the artificially occluded CoMA [30] images. All the baselines can only generate a single 3D reconstruction with respect to the target image. We observe that, the reconstructions generated by Diverse3DFace look diverse, yet plausible, and visually more faithful to the ground truth on the visible regions. Compared to this, FLAME based fitting [22] and DECA [12] do not explicitly handle occlusions and generate soft and erroneous shapes. CFR-GAN [18] and Occ-3DMM [9] get the pose wrong in multiple instances. Extreme3D [41] generates visually better reconstructions with respect to the visible parts of the face, but gets the expression wrong in the second row. In Fig. 9, we show further visual comparisons on the occlusion-augmented images from the CelebA [23] dataset. Note that we do not have ground truth scans for these images. However, visual results suggest that the baselines, by virtue of being holistic models that do not explicitly exclude features from the occluded regions, often get incorrect pose and expression on these images. Meanwhile, the reconstructions from Diverse3DFace look diverse on the occluded regions yet consistent with respect to the visible parts of the face.

FLAME vs Global+Local PCA Model: In addition to the quantitative comparison done in Tab. 1, we qualitatively compare the occlusion robustness of the global FLAME [22] model *vs.* our global+local model. In Fig. 5, we show failure cases of the FLAME [22] based fitting on severely occluded images. Notice the severe deformations on the FLAME [22] fitted outputs, especially around the mouth. In contrast, the fittings by our global+local models look more faithful as well as detailed with respect to the visible parts. This further supports our claim that a global+local model based fitting performs better than a global model based fitting on occluded face images.

6. Conclusion

In this paper, we proposed Diverse3DFace, an approach to reconstruct diverse, yet plausible 3D reconstructions corresponding to a single occluded face image. Our approach

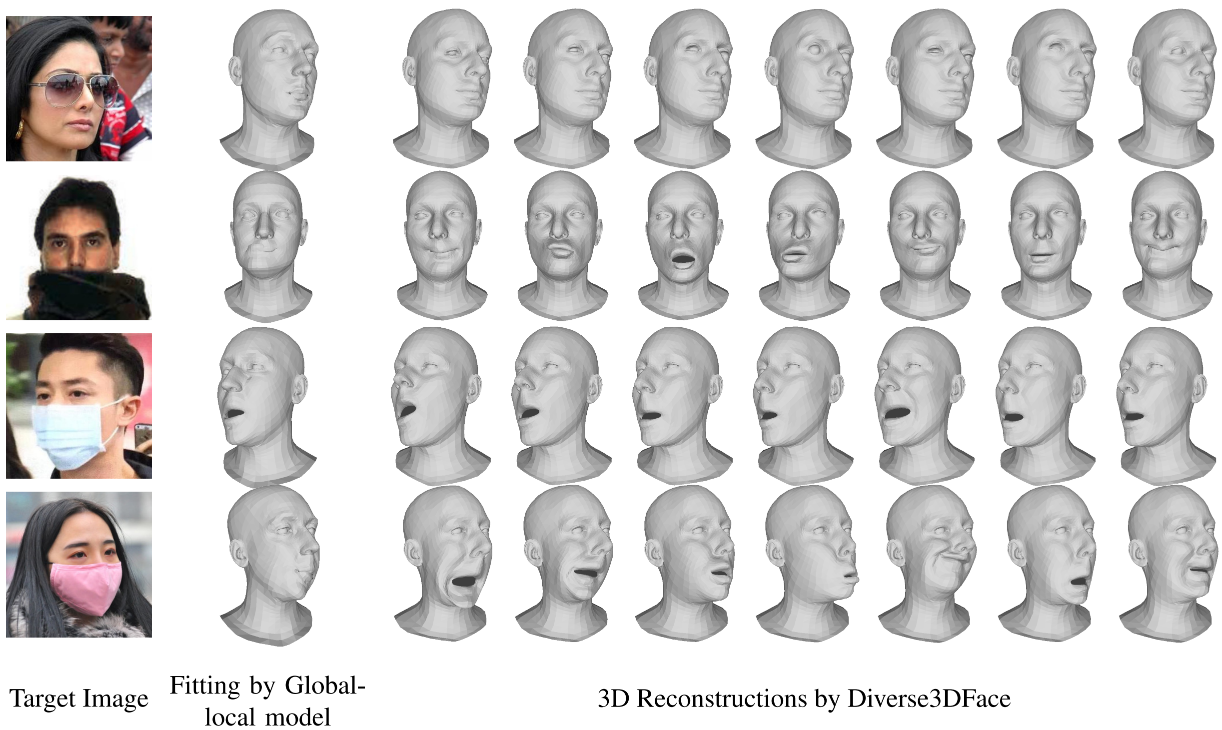


Figure 6. Set of 3D reconstructions by Diverse3DFace on real-world occluded face images.

was driven by the motivation that, in the presence of occlusions, a distribution of plausible 3D reconstruction is more desirable than a unique solution. We proposed a three step solution that first fits a robust partial shape using an ensemble of global+local PCA models, maps it to a latent space distribution and iterative optimizes the embeddings to promote diversity in the occluded parts while retaining fidelity with respect to the visible facial parts. Experimental evaluation across multiple occlusion types and datasets show the efficacy of Diverse3DFace, both in terms of robustness and diversity, as compared to multiple baselines. To the best of our knowledge, this is the first approach to generate diverse 3D reconstructions to explain an occluded face image. We foresee no potential negative impacts from the research presented in this paper.

A limitation of the proposed approach is its dependence on the robustness of the global+local fitting in the first step for further diverse completions. Although such a locally disentangled fitting demonstrably performs better than a global model fitting, it can still get affected in cases where the initial landmark or face-mask estimates are wrong.

Appendices

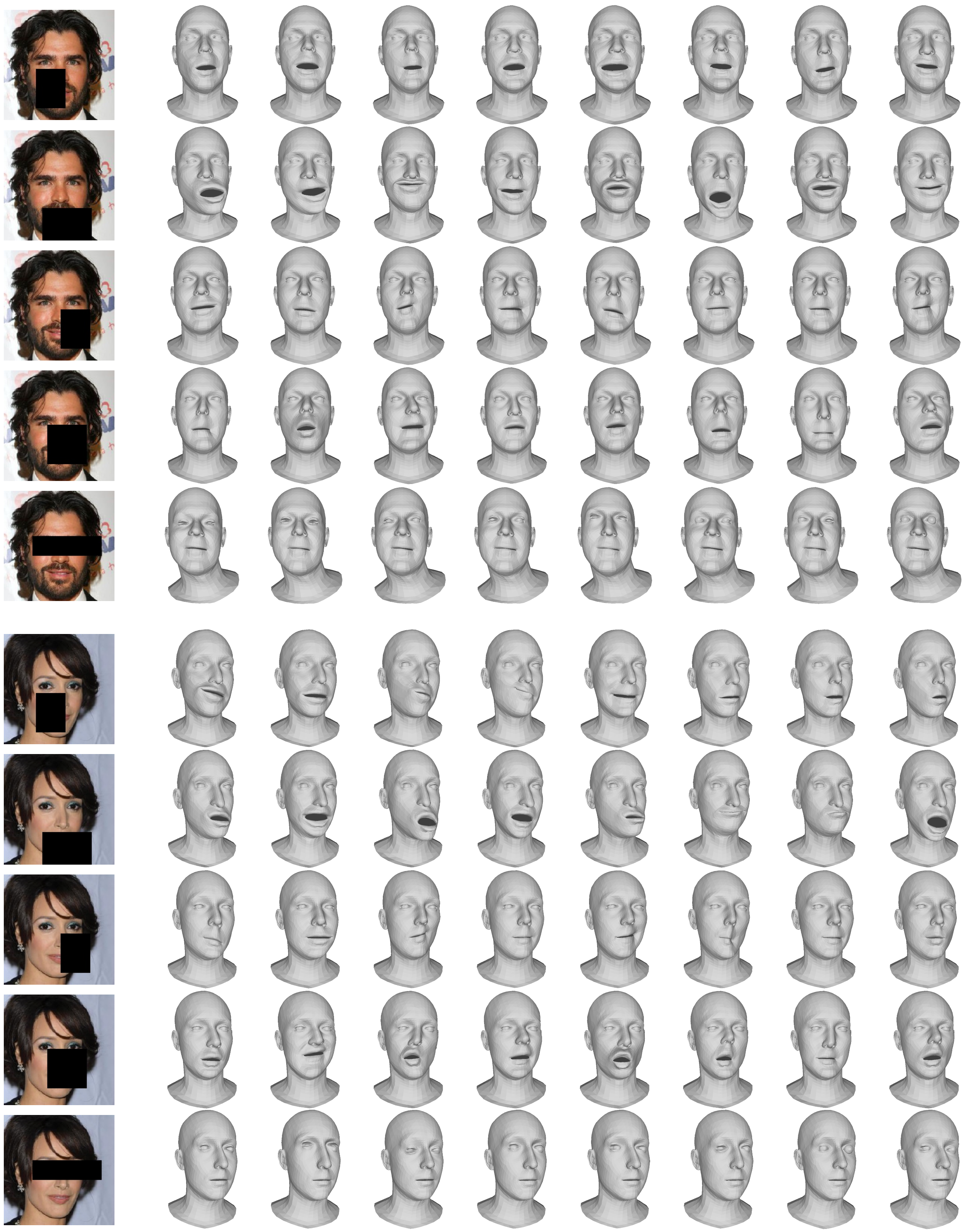
A. Further Experiments

A.1. Real-world Occlusions

We present examples of diverse 3D shape reconstructions by our approach on real-world occluded face images in Fig. 6. For these images, we inferred the occlusion mask using the face segmentation model by Nirkin *et al.* [26]. These results further demonstrate the efficacy of Diverse3DFace to generate diverse, yet plausible 3D reconstructions on real world occlusions ranging from glasses, scarf, facemasks, *etc*.

A.2. Moving the Occlusion Around the Face

In this section, we evaluate the diversity and robustness performance of Diverse3DFace to occlusions at different locations on the face. Fig. 7 shows the set of 3D reconstruction by Diverse3DFace when the occlusion moves around the face occupying the left cheek, mouth, the right cheek, center and the periocular (eye) regions of the face. Our method generates diverse, yet plausible set of 3D reconstructions for all the cases. We particularly note the high degree of diversity in expression that occurs when the mouth region is occluded, as is expected.



Target Image Diverse 3D Reconstructions by Diverse3DFace

Figure 7. Qualitative evaluation of the diversity and robustness performance of Diverse3DFace to occlusions at different facial locations.

Figure 8. **Controlled generated of diverse 3D reconstructions between two distinct modes.** Diverse3DFace can be used to generate controlled diversity on the occluded regions by performing interpolation between two distinct shapes in the latent space.

A.3. Diversity Interpolations

A potential application of Diverse3DFace is to perform controlled diversification around an occluded region during 3D reconstruction. To do this, we can first generate a set of diverse 3D reconstructions for an occluded target image and then allow the user to select two distinct samples to perform interpolation in-between. We perform interpolation in the latent space: $\mathbf{z}(\alpha) = \alpha\mathbf{z}_1 + (1-\alpha)\mathbf{z}_2$. This affords the user control over the extent and type of diversity. We present examples of such interpolations in Fig. 8.

A.4. Further Qualitative Results on CelebA Dataset

We show further qualitative results of diverse 3D reconstructions on occluded face images by Diverse3DFace, compared to the singular reconstruction by FLAME [22], DECA [12], CFR-GAN [18], Occ3DMM [9] and Extreme3D [41] in Fig. 9. While the baselines often get the pose, shape or expression wrong, Diverse3DFace generates 3D reconstructions that are consistent with the visible regions, yet plausibly diverse on the occluded regions.

B. Implementation Details

B.1. Mesh-VAE Network Architecture

The Mesh-VAE model is based on the fully convolutional mesh autoencoder (Meshconv) architecture proposed by Zhou *et al.* [51]. Meshconv [51] uses spatially varying convolutional kernels for different mesh vertices to account for the irregular structure of a 3D mesh. The spatially varying kernels are sampled from the span of a shared weight basis, using learned per-vertex coefficients. In addition, Meshconv defines pooling and unpooling operations on a 3D mesh by performing feature aggregation Monte Carlo sampling [51]. We detail the network architecture for the Mesh-VAE in Tabs. 4 and 5.

The abbreviated operators used are defined as follows:

- vcDownConv(in_c, out_c, s, r, M) + vcDownRes(s): Downward residual block (as defined in Meshconv [51]), with in_c input channels, out_c output channels, s stride, r kernel radius and M number of shared weight bases. The output is activated with ELU [6] activation.

- vcUpConv(in_c, out_c, s, r, M) + vcUpRes(s): Upward residual block (as defined in Meshconv [51]), with in_c input channels, out_c output channels, s stride, r kernel radius and M number of shared weight bases. The output is activated with ELU [6] activation.

References

[1] Yuval Bahat and Tomer Michaeli. Explorable super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2716–2725, 2020. 2, 3

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning*, 2009. 6

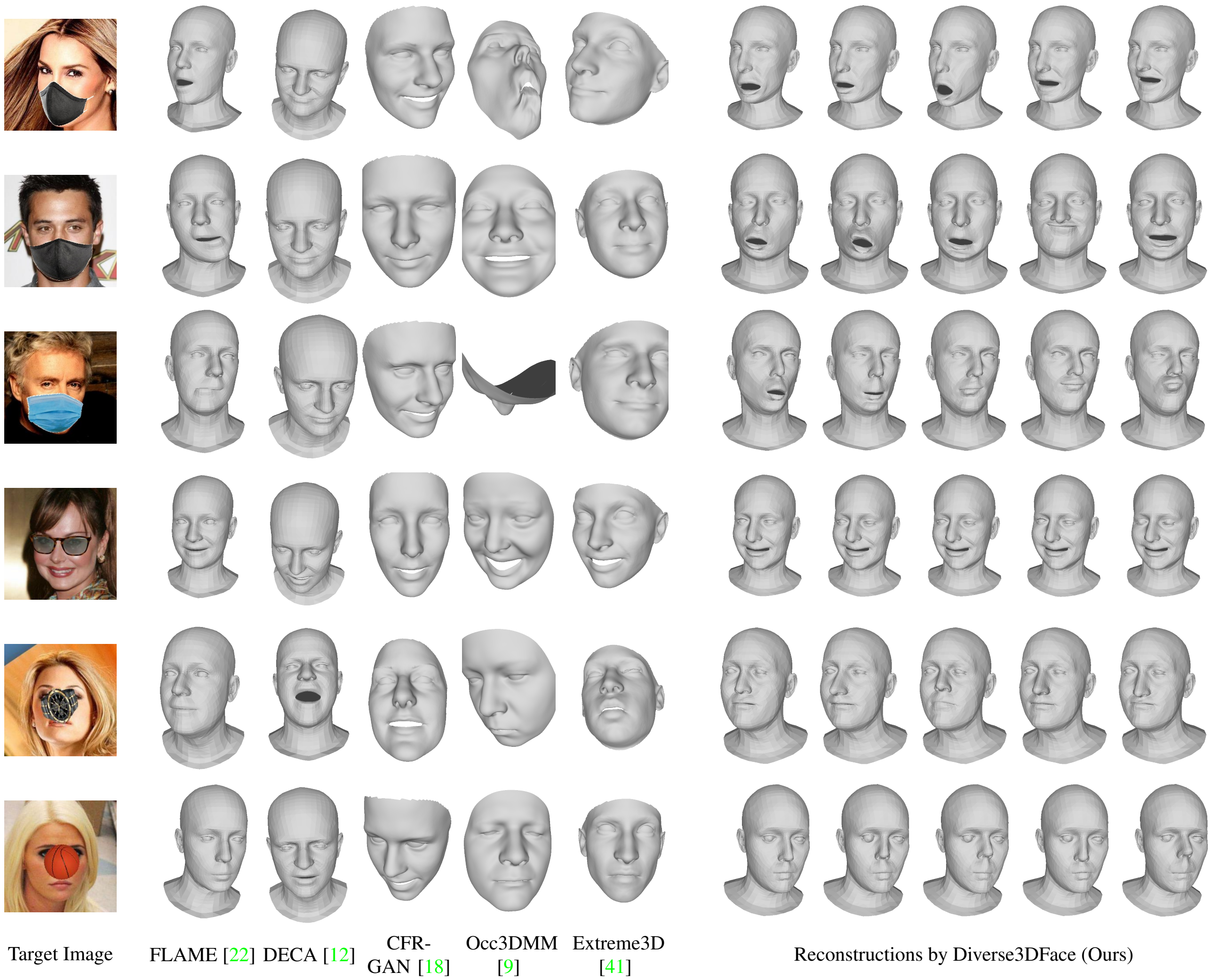


Figure 9. **More Qualitative evaluation on the CelebA dataset [23]**: Reconstructed singular 3D meshes from the target image by the baselines *vs.* the diverse reconstructions from Diverse3DFace.

Table 4. Network architecture of the Mesh-VAE Encoder $\mathcal{E}_{mesh}$.

Input	Layer	Output size	Output
5023×3 **Mesh**	$\rightarrow$ vcDownConv($in_c = 3, out_c = 32, s = 2, r = 43, M = 17$) + vcDownRes(2)	1367×32	
	vcDownConv($in_c = 32, out_c = 64, s = 1, r = 27, M = 17$) + vcDownRes(1)	1367×64	
	vcDownConv($in_c = 64, out_c = 128, s = 2, r = 54, M = 17$) + vcDownRes(2)	270×128	
	vcDownConv($in_c = 128, out_c = 256, s = 1, r = 25, M = 17$) + vcDownRes(1)	270×256	
	vcDownConv($in_c = 256, out_c = 512, s = 2, r = 81, M = 17$) + vcDownRes(2)	45×512	
	vcDownConv($in_c = 512, out_c = 1024, s = 1, r = 27, M = 17$) + vcDownRes(1)	45×1024	*feats*
feats	vcDownConv($in_c = 1024, out_c = 64, s = 2, r = 37, M = 17$) + vcDownRes(2)	10×64	$\boldsymbol{\mu}$
feats	vcDownConv($in_c = 1024, out_c = 64, s = 2, r = 37, M = 17$) + vcDownRes(2)	10×64	$\log \boldsymbol{\sigma}^2$
Model Complexity	9M		

[3] Volker Blanz and Thomas 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. 1, 2, 3

[4] Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, and Stefanos Zafeiriou. Neural 3d mor-

Table 5. Network architecture of the Mesh-VAE Decoder $\mathcal{D}_{mesh}$.

Input	Layer	Output size	Output
10×64 $\mathbf{z}$	vcUpConv($in_c = 64, out_c = 1024, s = 2, r = 8, M = 17$) + vcUpRes(2)	45×1024	
	vcUpConv($in_c = 1024, out_c = 512, s = 1, r = 27, M = 17$) + vcUpRes(1)	45×512	
	vcUpConv($in_c = 512, out_c = 256, s = 2, r = 16, M = 17$) + vcUpRes(2)	270×256	
	vcUpConv($in_c = 256, out_c = 128, s = 1, r = 25, M = 17$) + vcUpRes(1)	270×128	
	vcUpConv($in_c = 128, out_c = 64, s = 2, r = 12, M = 17$) + vcUpRes(2)	1367×64	
	vcUpConv($in_c = 64, out_c = 32, s = 1, r = 27, M = 17$) + vcUpRes(1)	1367×32	
	vcUpConv($in_c = 32, out_c = 3, s = 2, r = 24, M = 17$) + vcUpRes(2)	5023×3	**Output**
Model Complexity	8M		

phable models: Spiral convolutional networks for 3d shape representation learning and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7213–7222, 2019. 2, 6

[5] Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*, 2016. 2

[6] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015. 11

[7] Darren Cosker, Eva Krumhuber, and Adrian Hilton. A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In *2011 international conference on computer vision*, pages 2296–2303. IEEE, 2011. 4, 5

[8] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29:3844–3852, 2016. 2, 6

[9] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision*, 126(12):1269–1287, 2018. 1, 2, 3, 6, 7, 8, 11, 12

[10] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020. 1

[11] Mohamed Elfeki, Camille Couprie, Morgane Riviere, and Mohamed Elhoseiny. Gdpp: Learning diverse generations using determinantal point processes. In *International Conference on Machine Learning*, pages 1774–1783. PMLR, 2019. 2, 3

[12] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 1, 2, 6, 7, 8, 11, 12

[13] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos 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. 1, 2

[14] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018. 2

[15] Arnab Ghosh, Viveka Kulharia, Vinay P. Namboodiri, Philip H.S. Torr, and Puneet K. Dokania. Multi-agent diverse generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei 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. 2

[18] Yeong-Joon Ju, Gun-Hee Lee, Jung-Ho Hong, and Seong-Whan Lee. Complete face recovery gan: Unsupervised joint face rotation and de-occlusion from a single-view image. In *WACV*, 2022. 1, 2, 6, 7, 8, 11, 12

[19] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. Inversefacenet: Deep monocular inverse face rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4625–4634, 2018. 2

[20] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2, 6

[21] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012. 2, 3, 5

[22] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 1, 2, 3, 4, 5, 6, 7, 8, 11, 12

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 6, 7, 8, 12

[24] Odile Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122, 1975. 3

[25] Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4602–4609, 2019. 2, 6

[26] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 98–105. IEEE, 2018. 9

[27] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas 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. 1, 2, 3

[28] Stylianos Ploumpis, Evangelos Ververas, Eimear O'Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William Smith, Baris Gecer, and Stefanos P Zafeiriou. Towards a complete 3d morphable model of the human head. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[29] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5620–5629, 2021. 2

[30] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders. In *European Conference on Computer Vision (ECCV)*, pages 725–741, 2018. 2, 4, 5, 6, 7, 8

[31] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019. 2

[32] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David 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. 1, 2

[33] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5541–5550, 2017. 2

[34] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491, 2015. 2

[35] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3310–3320, 2017. 2

[36] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016. 2

[37] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian 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. 1, 2

[38] Luan Tran, Feng Liu, and Xiaoming 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. 1, 2

[39] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1, 2

[40] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5163–5172, 2017. 1, 2

[41] Anh Tuán Trán, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3935–3944, 2018. 1, 2, 3, 6, 7, 8, 11, 12

[42] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 2, 6

[43] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 4, 5

[44] Huawei Wei, Shuang Liang, and Yichen Wei. 3d dense face alignment via graph convolution networks. *arXiv preprint arXiv:1904.05562*, 2019. 1

[45] Shangzhe Wu, Christian Rupprecht, and Andrea 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. 1, 2

[46] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. *arXiv preprint arXiv:1901.09024*, 2019. 2

[47] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12803–12813, 2021. 2

[48] Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. *arXiv preprint arXiv:1907.04967*, 2019. 2, 3, 5, 6

[49] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pages 346–364. Springer, 2020. 3

[50] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019. 2

[51] Yi Zhou, Chenglei Wu, Zimo Li, Chen Cao, Yuting Ye, Jason Saragih, Hao Li, and Yaser Sheikh. Fully convolutional mesh autoencoder using efficient spatially varying kernels. *arXiv preprint arXiv:2006.04325*, 2020. 2, 6, 11

[52] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Multimodal image-to-image translation by enforcing bi-cycle consistency. In *Advances in neural information processing systems*, pages 465–476, 2017. 2

[53] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer Graphics Forum*, volume 37, pages 523–550. Wiley Online Library, 2018. 1