Learning an Animatable Detailed 3D Face Model from In-The-Wild Images

Yao Feng[1,2*] Haiwen Feng[1*] Michael J. Black[1] Timo Bolkart[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]Max Planck ETH Center for Learning System
{yfeng, hfeng, black, tbolkart}@tuebingen.mpg.de

Abstract

While current monocular 3D face reconstruction methods can recover fine geometric details, they suffer several limitations. Some methods produce faces that cannot be realistically animated because they do not model how wrinkles vary with expression. Other methods are trained on high-quality face scans and do not generalize well to in-the-wild images. We present the first approach to jointly learn a model with animatable detail and a detailed 3D face regressor from in-the-wild images that recovers shape details as well as their relationship to facial expressions. Our DECA (Detailed Expression Capture and Animation) model is trained to robustly produce a UV displacement map from a low-dimensional latent representation that consists of person-specific detail parameters and generic expression parameters, while a regressor is trained to predict detail, shape, albedo, expression, pose and illumination parameters from a single image. We introduce a novel detail-consistency loss to disentangle person-specific details and expression-dependent wrinkles. This disentanglement allows us to synthesize realistic person-specific wrinkles by controlling expression parameters while keeping person-specific details unchanged. DECA achieves state-of-the-art shape reconstruction accuracy on two benchmarks. Qualitative results on in-the-wild data demonstrate DECA's robustness and its ability to disentangle identity and expression dependent details enabling animation of reconstructed faces. The model and code are publicly available at https://github.com/YadiraF/DECA.

1. Introduction

Two decades have passed since the seminal work of Vetter and Blanz [76] that first showed how to reconstruct 3D facial geometry from a single image. Since then, 3D face reconstruction methods have rapidly advanced (for a comprehensive overview see [85]) enabling applications such as 3D avatar creation for VR/AR [32], video editing [71], face recognition [5, 54], virtual make-up [59], or speech-driven facial animation [17]. To make the problem tractable, most existing methods incorporate prior knowledge about geometry or appearance by leveraging pre-computed 3D face models [7, 20]. These models reconstruct the coarse face shape but are unable to capture geometric details such as expression-dependent wrinkles, which are essential for realism and for analysing human emotion.

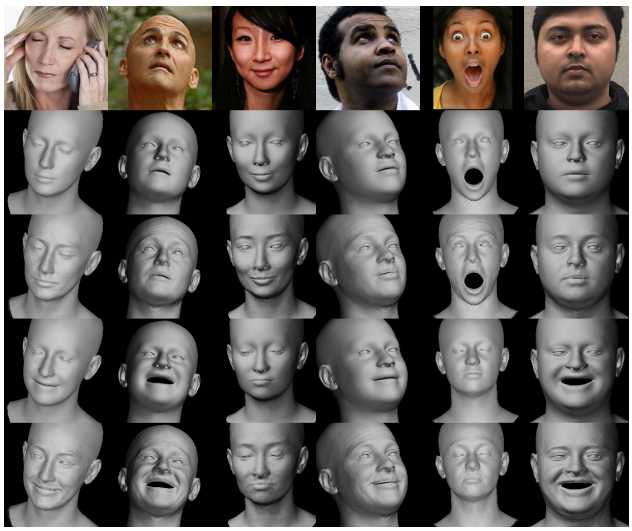


Figure 1: **DECA.** Example images (row 1), the regressed coarse shape (row 2), detail shape (row 3) and reposed coarse shape (row 4), and reposed with person-specific details (row 5). DECA is robust to occlusion and captures person-specific details as well as expression wrinkles that appear in regions like forehead and mouth. Our novelty is that this detail shape can be reposed (animated) such that the wrinkles are specific to the source shape and expression.

Several methods recover detailed facial geometry [1, 9, 15, 30, 53, 73, 74], however, they require high-quality training scans [9, 15] or lack robustness to occlusions [1, 30, 53]. None of these works explore how the recovered wrinkles

*equal contribution

change with varying expressions. Previous methods that learn expression-dependent detail models [14, 82] either use detailed 3D scans as training data and, hence, do not generalize to unconstrained images [82], or model expression-dependent details as part of the appearance map rather than the geometry [14], preventing realistic mesh relighting.

We introduce DECA (Detailed Expression Capture and Animation), which learns an *animatable* displacement model from in-the-wild images without 2D-to-3D supervision. In contrast to prior work, these *animatable expression-dependent wrinkles are specific to an individual* and are regressed from an image. Specifically, DECA jointly learns 1) a geometric detail model that generates a UV displacement map from a low-dimensional representation that consists of subject-specific detail parameters and expression parameters, and 2) a regressor that predicts subject-specific detail, albedo, shape, expression, pose, and lighting parameters from an image. The detail model builds upon FLAME's [42] coarse geometry, and we formulate the displacements as a function of subject-specific detail parameters and FLAME's jaw pose and expression parameters.

To gain control over expression-dependent wrinkles of the reconstructed face, while preserving person-specific details (i.e. moles, pores, eyebrows, and expression-independent wrinkles), the person-specific details and expression-dependent wrinkles must be disentangled. Our key contribution is a novel *detail consistency loss* that enforces this disentanglement. Given two images of the same person with different expressions, we observe that their 3D face shape and their person-specific details are the same in both images, but the expression and the intensity of the wrinkles differ with expression. During training, this observation is exploited by swapping the detail codes between different images of the same identity and enforcing the newly rendered results to look similar to the original input images. Once trained, DECA reconstructs a detailed 3D face from a single image (Fig. 1 third row) in real time (about 120fps on a Nvidia Quadro RTX 5000), and is able to animate the reconstruction with realistic adaptive expression wrinkles (Fig. 1 bottom).

In summary, our main contributions are: 1) The first approach to learn an animatable displacement model from in-the-wild images that can synthesize plausible geometric details by varying expression parameters. 2) A novel detail consistency loss that disentangles identity-dependent and expression-dependent facial details. 3) Reconstruction of geometric details that is, unlike most competing methods, robust to occlusions, poses, and illumination variation. This is enabled by our low-dimensional detail representation, the detail disentanglement, and training from a large dataset of in-the-wild images. 4) State-of-the-art shape reconstruction accuracy on two different benchmarks. 5) Code and model will be made publicly available for research purposes.

2. Related work

The reconstruction of 3D faces from visual input has received significant attention over the last decades after the pioneering work of Parke [47], the first method to reconstruct 3D faces from multi-view images. While a large body of related work aims to reconstruct 3D faces from various input modalities such as multi-view images [4, 11, 50], video data [23, 33, 35, 62, 66], RGB-D data [41, 70, 80] or subject-specific image collections [37, 56], our main focus is on methods that use only a single RGB image. For a more comprehensive overview, see Zollhöfer et al. [85].

Coarse reconstruction: Many monocular 3D face reconstruction methods follow Vetter and Blanz [76] by estimating coefficients of pre-computed statistical models in an analysis-by-synthesis fashion. Such methods can be categorized into optimization-based [2, 3, 5, 6, 26, 55, 71], or learning-based methods [13, 18, 25, 38, 52, 58, 69, 72, 75]. These methods estimate parameters of a statistical face model with a fixed linear shape space, which captures only low-frequency shape information. This results in overly-smooth reconstructions.

Several works are model-free and directly regress 3D faces (i.e. voxels [34] or meshes [19, 21, 28, 79]) and hence could capture more variation than the model-based methods. However, all these methods require explicit 3D supervision, which is provided either by an optimization-based model fitting [21, 28, 34, 79] or by synthetic data generated by sampling a statistical face model [19] and therefore also only capture coarse shape variations.

Instead of capturing high-frequency geometric details, some methods reconstruct coarse facial geometry along with high-fidelity textures [24, 57, 65, 81]. As this "bakes" shading details into the texture, lighting changes do not affect these details. To enable animation and relighting, DECA captures these details as part of the geometry.

Detail reconstruction: Another body of work aims to reconstruct faces with "mid-frequency" details. Common optimization-based methods fit a statistical face model to images to obtain a coarse shape estimate, followed by a shape from shading (SfS) method to reconstruct facial details from monocular images [36, 43] or videos [23, 66]. Unlike DECA, these approaches are slow, the results lack robustness to occlusions, and the coarse model fitting step requires facial landmarks, making them error-prone for large viewing angles and occlusions.

Most regression-based approaches [9, 15, 30, 53, 73] follow a similar approach by first reconstructing the parameters of a statistical face model to obtain a coarse shape, followed by a refinement step to capture localized details. Chen et al. [15] and Cao et al. [9] compute local wrinkle statistics from high-resolution scans and leverage these to constrain the fine-scale detail reconstruction from images [15] or videos [9]. Guo et al. [30] and Richardson et

al. [53] directly regress per-pixel displacement maps. All these methods only reconstruct fine-scale details in non-occluded regions, causing visible artifacts in the presence of occlusions. Tran et al. [73] gain robustness to occlusions by applying some face segmentation method [46] to determine occluded regions, and employ an example-based hole filling of the occluded regions. Further, model-free methods exist that directly reconstruct detailed meshes [60, 83] or surface normals that add detail to coarse reconstructions [1, 61].

Tran et al. [74] and Tewari et al. [67, 68] jointly learn a statistical face model and reconstruct 3D faces from images. While offering more flexibility than fixed statistical models, these methods capture limited geometric details compared to other detail reconstruction methods.

Unlike DECA, none of these detail reconstruction methods offer animatable details after reconstruction.

Animatable detail reconstruction: Most relevant to DECA are methods that reconstruct detailed faces while allowing animation of the result. Golovinski et al. [27], Shin et al. [63] and FaceScape [82] learn correlations between wrinkles and factors like age and gender [27] or expression [63, 82] from high-quality face scans. In contrast, DECA learns an animatable detail model solely from in-the-wild images without paired 3D training data. While FaceScape [82] predicts an animatable 3D face from a single image, the method is not robust to occlusions. This is due to a two step reconstruction process: first optimize the coarse shape, then predict a displacement map from the texture map extracted with the coarse reconstruction.

Chaudhuri et al. [14] learn identity and expression corrective blendshapes with dynamic (expression-dependent) albedo maps [45]. They model geometric details as part of the albedo map, and therefore, the shading of these details does not adapt with varying lighting. This results in unrealistic renderings. In contrast, DECA models details as geometric displacements, which look natural when re-lit.

3. Preliminaries

Geometry prior: FLAME [42] is a statistical 3D head model that combines separate linear identity shape and expression spaces with linear blend skinning (LBS) and pose-dependent corrective blendshapes to articulate the neck, jaw, and eyeballs. Given parameters of facial identity $\boldsymbol{\beta} \in \mathbb{R}^{|\boldsymbol{\beta}|}$, pose $\boldsymbol{\theta} \in \mathbb{R}^{3k+3}$ (with $k = 4$ joints for neck, jaw, and eyeballs), and expression $\boldsymbol{\psi} \in \mathbb{R}^{|\boldsymbol{\psi}|}$, FLAME outputs a mesh with $n = 5023$ vertices. The model is defined as

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}) = W(T_P(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}), \mathbf{J}(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathcal{W}), \quad (1)$$

with the blend skinning function $W(\mathbf{T}, \mathbf{J}, \boldsymbol{\theta}, \mathcal{W})$ that rotates the vertices in $\mathbf{T} \in \mathbb{R}^{3n}$ around joints $\mathbf{J} \in \mathbb{R}^{3k}$, linearly smoothed by blendweights $\mathcal{W} \in \mathbb{R}^{k \times n}$. The joint locations $\mathbf{J}$ are defined as a function of the identity $\boldsymbol{\beta}$. Further,

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

denotes the mean template $\mathbf{T}$ in "zero pose" with added shape blendshapes $B_S(\boldsymbol{\beta}; \mathcal{S}) : \mathbb{R}^{|\boldsymbol{\beta}|} \rightarrow \mathbb{R}^{3n}$, pose correctives $B_P(\boldsymbol{\theta}; \mathcal{P}) : \mathbb{R}^{3k+3} \rightarrow \mathbb{R}^{3n}$, and expression blendshapes $B_E(\boldsymbol{\psi}; \mathcal{E}) : \mathbb{R}^{|\boldsymbol{\psi}|} \rightarrow \mathbb{R}^{3n}$, with the learned identity, pose, and expression bases $\mathcal{S}, \mathcal{P}$ and $\mathcal{E}$. See [42] for details.

Appearance model: FLAME does not have an appearance model, hence we convert Basel Face Model's PCA albedo space [49] into the FLAME UV layout to make it compatible with FLAME. The appearance model outputs a UV albedo map $A(\boldsymbol{\alpha}) \in \mathbb{R}^{d \times d \times 3}$ for albedo parameters $\boldsymbol{\alpha} \in \mathbb{R}^{|\boldsymbol{\alpha}|}$.

Camera model: Photographs in existing in-the-wild face datasets are often taken from a distance. We, therefore, use an orthographic camera model $\mathbf{c}$ to project the 3D mesh into image space. Face vertices are projected into the image as $\mathbf{v} = s\Pi(M_i) + \mathbf{t}$, where $M_i \in \mathbb{R}^3$ is a vertex in M, $\Pi \in \mathbb{R}^{2 \times 3}$ is the orthographic 3D-2D projection matrix, and $s \in \mathbb{R}$ and $\mathbf{t} \in \mathbb{R}^2$ denote isotropic scale and 2D translation, respectively. The parameters s, and $\mathbf{t}$ are summarized as $\boldsymbol{c}$.

Illumination model: For face reconstruction, the most frequently-employed illumination model is based on Spherical Harmonics (SH) [44]. By assuming that the light source is distant and the face's surface reflectance is Lambertian, the shaded face image is computed as:

$$B(\boldsymbol{\alpha}, \mathbf{l}, N_{uv})_{i,j} = A(\boldsymbol{\alpha})_{i,j} \odot \sum_{k=1}^{9} \mathbf{l}_k H_k(N_{i,j}), \quad (3)$$

where the albedo, A, surface normals, N, and shaded texture, B, are represented in UV coordinates and where $B_{i,j} \in \mathbb{R}^3$, $A_{i,j} \in \mathbb{R}^3$, and $N_{i,j} \in \mathbb{R}^3$ denote pixel (i, j) in the UV coordinate system. The SH basis and coefficients are defined as $H_k : \mathbb{R}^3 \rightarrow \mathbb{R}$ and $\mathbf{l} = [\mathbf{l}_1^T, \cdots, \mathbf{l}_9^T]^T$, with $\mathbf{l}_k \in \mathbb{R}^3$, and $\odot$ denotes the Hadamard product.

Texture rendering: Once we have the geometry parameters $(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi})$, albedo $(\boldsymbol{\alpha})$, lighting $(\mathbf{l})$ and camera information $\boldsymbol{c}$, we can recover the 2D image I_r by rendering as $I_r = \mathcal{R}(M, B, \mathbf{c})$, where $\mathcal{R}$ denotes the rendering function.

FLAME is able to generate a face geometry with various poses, shapes and expressions from a low-dimensional latent space. However, the representational power of the model is limited by the low mesh resolution and therefore mid-frequency details are mostly missing in FLAME's surface. The next section introduces our expression-dependent displacement model that augments FLAME with mid-frequency details, and it demonstrates how to reconstruct detailed geometry from a single image and animate it.

4. Proposed method

The goal of DECA is to learn a parameterized face model with geometric detail solely from in-the-wild images (Fig. 2

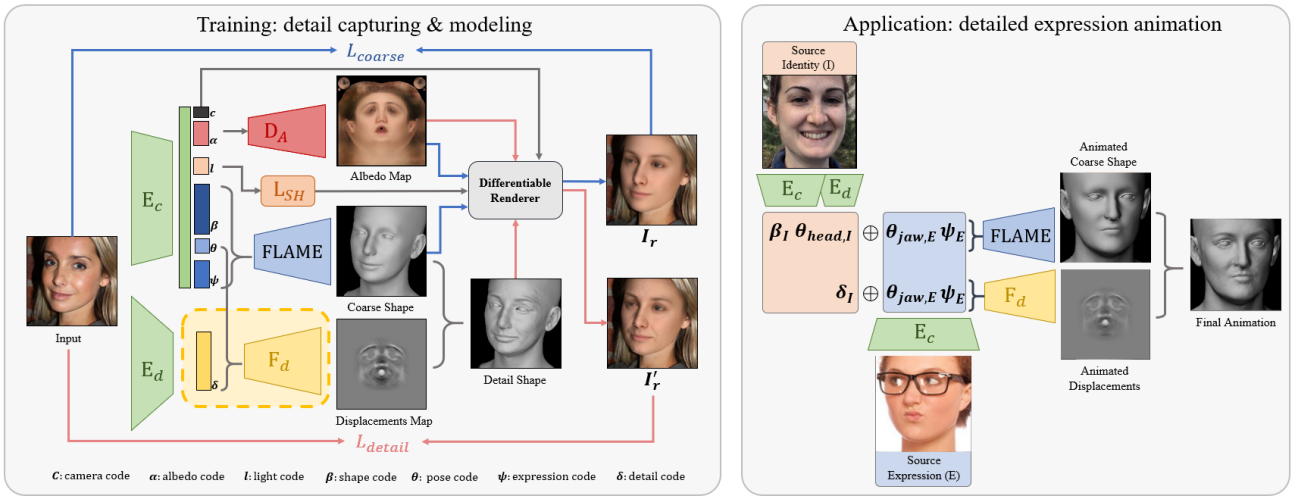


Figure 2: DECA training and animation. During training, DECA estimates parameters to reconstruct face shape for each image and, at the same time, learns an expression-conditioned displacement model by leveraging the shape and detail consistency information from multiple images of the same individual (see Sec. 4.3 for details, the yellow box region is further illustrated in Fig. 3). Once trained, DECA animates a face (right) by combining the reconstructed source identity's shape, head pose, and detail code, with the reconstructed source expression's jaw pose and expression parameters to obtain an animated coarse shape and an animated displacement map. Finally, DECA outputs an animated detail shape.

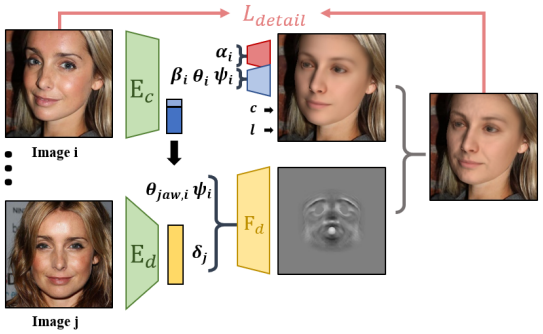


Figure 3: Detail consistency loss. See Sec. 4.3 for details.

left). Once trained, DECA reconstructs the 3D head with detailed face geometry from a single face image I. The learned parametrization of the reconstructed details enables us then to animate the detail reconstruction by controlling FLAME's expression and jaw pose parameters (Fig. 2 right). This synthesizes new wrinkles while keeping person-specific details unchanged.

Key idea: The key idea of DECA is grounded in the observation that an individual's face shows different details (i.e. wrinkles), depending on their facial expressions but that other properties of their shape remain unchanged. Consequently, facial details should be separated into static person-specific details and dynamic expression-dependent details such as wrinkles [40]. However, disentangling static and dynamic facial details is a non-trivial task. Static facial details are different across people, whereas dynamic expression dependent facial details even vary for the same person. Thus, DECA learns an expression-conditioned detail model to infer facial details from both the person-specific detail latent space and the expression space.

The main difficulty of learning a detail displacement model is the lack of training data. Prior work uses specialized camera systems to scan people in a controlled environment to obtain detailed facial geometry. However, this approach is expensive and impractical for capturing large numbers of identities with varying expressions and diversity in ethnicity and age. Therefore we propose an approach to learn detail geometry from in-the-wild images.

4.1. Coarse reconstruction

We first learn a coarse reconstruction (i.e. in FLAME's model space) in an analysis-by-synthesis way: given a 2D image I as input, we encode the image to a latent code, decode this to synthesize a 2D image I_r, and minimize the difference between the synthesized image and the input. As shown in Fig. 2, we train an encoder E_c, which consists of a ResNet50 [31] network followed by a fully connected layer, to regress a low-dimensional latent code. This latent code consists of FLAME parameters $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, $\boldsymbol{\theta}$ (i.e. representing the coarse geometry), albedo coefficients $\boldsymbol{\alpha}$, camera $\boldsymbol{c}$, and lighting parameters $\mathbf{l}$. More specifically, the coarse geometry uses the first 100 FLAME shape parameters ($\boldsymbol{\beta}$), 50 expression parameters ($\boldsymbol{\psi}$), and 50 albedo parameters ($\boldsymbol{\alpha}$). In total, E_c predicts a 236 dimensional latent code.

Given a dataset of $2D$ face images I_i with multiple images per subject, corresponding identity labels c_i, and

68 $2D$ keypoints $\mathbf{k}_i$ per image, the coarse reconstruction branch is trained by minimizing

$$L_{coarse} = L_{lmk} + L_{eye} + L_{pho} + L_{id} + L_{sc} + L_{reg}, \quad (4)$$

with landmark loss L_{lmk}, eye closure loss L_{eye}, photometric loss L_{pho}, identity loss L_{id}, shape consistency loss L_{sc} and regularization L_{reg}.

Landmark re-projection loss: The landmark loss measures the difference between ground-truth $2D$ face landmarks $\mathbf{k}_i$ and the corresponding landmarks in the FLAME's surface $M_i \in \mathbb{R}^3$, projected into the image by the estimated camera model. The landmark loss is defined as

$$L_{lmk} = \sum_{i=1}^{68} \|\mathbf{k}_i - s\Pi(M_i) + \mathbf{t}\|_1 . \quad (5)$$

Eye closure loss: The eye closure loss computes the relative offset of landmarks $\mathbf{k}_i$ and $\mathbf{k}_j$ on the upper and lower eyelid, and measures the difference to the offset of the corresponding landmarks in the FLAME's surface M_i and M_j projected into the image. Formally, the loss is given as

$$L_{eye} = \sum_{(i,j)\in E} \|\mathbf{k}_i - \mathbf{k}_j - s\Pi(M_i - M_j)\|_1 , \quad (6)$$

where E is the set of upper/lower eyelid landmark pairs.

Photometric loss: The photometric loss computes the error between the input image I and the rendering I_r as $L_{pho} = \|V_I \odot (I - I_r)\|_{1,1}$. Here, V_I is a face mask with value 1 in the face skin region, and value 0 elsewhere obtained by an existing face segmentation method [46], and $\odot$ denotes the Hadamard product. Computing the error in the face region only provides robustness to common occlusions by e.g. hair, clothes, sunglasses, etc.

Identity loss: Recent 3D face reconstruction methods demonstrate the effectiveness of utilizing an identity loss to produce more realistic face shapes [18, 24]. Motivated by this, we also use a pretrained face recognition network [10], to employ an identity loss during training.

The face recognition network f outputs feature embeddings of the rendered images and the input image, and the identity loss then measures the cosine similarity between the two embeddings. Formally, the loss is defined as

$$L_{id} = 1 - \frac{f(I)f(I_r)}{\|f(I)\|_2 \cdot \|f(I_r)\|_2}. \quad (7)$$

Shape consistency loss: Given two images I_i and I_j of the same subject (i.e. $c_i = c_j$), the coarse encoder E_c should output the same shape parameters (i.e. $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$). Previous work encourages shape consistency by enforcing the distance between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ to be smaller by a margin than the distance to the shape coefficients corresponding of a different subject [58]. However, choosing this fixed margin is challenging in practice. Instead, we propose a different strategy by replacing $\boldsymbol{\beta}_i$ with $\boldsymbol{\beta}_j$ while keeping all other parameters unchanged. Given that $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ represent the same subject, this new set of parameters must reconstruct I_i well. Formally, we minimize

$$L_{sc} = L_{coarse}(I_i, \mathcal{R}(M(\boldsymbol{\beta}_j, \boldsymbol{\theta}_i, \boldsymbol{\psi}_i), A(\boldsymbol{\alpha}_i), \mathbf{l}_i, \mathbf{c}_i)). \quad (8)$$

Regularization: L_{reg} regularizes shape $E_{\boldsymbol{\beta}} = \|\boldsymbol{\beta}\|_2^2$, expression $E_{\boldsymbol{\psi}} = \|\boldsymbol{\psi}\|_2^2$, and albedo $E_{\boldsymbol{\alpha}} = \|\boldsymbol{\alpha}\|_2^2$.

4.2. Detail reconstruction

The detail reconstruction aims at augmenting the coarse FLAME geometry with a detailed UV displacement map $D \in [-0.01, 0.01]^{d\times d}$ (see Fig. 2). Similar to the coarse reconstruction, we train an encoder E_d (with the same architecture as E_c) to encode I to a 128-dimensional latent code $\boldsymbol{\delta}$, representing subject-specific details. The latent code $\boldsymbol{\delta}$ is then concatenated with FLAME's expression $\boldsymbol{\psi}$ and jaw pose parameters $\boldsymbol{\theta}_{jaw}$, and decoded by F_d to D.

Detail decoder: The detail decoder is defined as

$$D = F_d(\boldsymbol{\delta}, \boldsymbol{\psi}, \boldsymbol{\theta}_{jaw}), \quad (9)$$

where the detail code $\boldsymbol{\delta} \in \mathbb{R}^{128}$ controls the static person-specific details. We leverage the expression $\boldsymbol{\psi} \in \mathbb{R}^{50}$ and jaw pose parameters $\boldsymbol{\theta}_{jaw} \in \mathbb{R}^3$ from the coarse reconstruction branch to capture the dynamic expression wrinkle details. For rendering, D is converted to a normal map.

Detail rendering: The detail displacement model allows us to generate images with fine-scale surface details. To reconstruct the detailed geometry M', we convert M and its surface normals N to UV space, denoted as $M_{uv} \in \mathbb{R}^{d\times d\times 3}$ and $N_{uv} \in \mathbb{R}^{d\times d\times 3}$, and combine them with D as

$$M'_{uv} = M_{uv} + D \odot N_{uv}. \quad (10)$$

By calculating normal N' from M', we obtain the detail rendering I'_r by rendering M with applied normal map as

$$I'_r = \mathcal{R}(M, B(\boldsymbol{\alpha}, \mathbf{l}, N'), \mathbf{c}). \quad (11)$$

The detail reconstruction is trained by minimizing

$$L_{detail} = L_{phoD} + L_{mrf} + L_{sym} + L_{dc} + L_{regD}, \quad (12)$$

with photometric detail loss L_{phoD}, ID-MRF loss L_{mrf}, soft symmetry loss L_{sym}, and detail regularization L_{regD}.

Detail photometric losses: With the applied detail displacement map, the rendered images I'_r contain some geometric details. Equivalent to the coarse rendering, we use a photometric loss $L_{phoD} = \|V_I \odot (I - I'_r)\|_{1,1}$, where, recall, V_I is a mask representing the visible skin pixels.

ID-MRF loss: We add an Implicit Diversified Markov Random Fields (ID-MRF) loss [78] to reconstruct geometric details. Given two images of the same person, the ID-MRF

loss extracts feature patches from different layers of a pre-trained network, and then minimizes the difference between corresponding nearest neighbor feature patches from both images. Following Wang et al. [78], the loss is computed on layers $conv3\_2$ and $conv4\_2$ of VGG19 [64] as

$$L_{mrf} = 2L_M(conv4\_2) + L_M(conv3\_2), \tag{13}$$

where $L_M(layer_{th})$ denotes the ID-MRF loss which is employed on the feature patches extracted from I'_r and I with layer $layer_{th}$ of VGG19. As for the photometric losses, we compute L_{mrf} only for the face skin region in UV space.

Soft symmetry loss: To add robustness to occlusions, we add a soft symmetry loss to regularize non-visible face parts. Specifically, we minimize

$$L_{sym} = \|V_{uv} \odot (D - flip(D))\|_{1,1}, \tag{14}$$

where V_{uv} denotes the face skin mask in UV space, and $flip$ is the horizontal flip operation.

Detail regularization: The detail displacements are regularized by $L_{regD} = \|D\|_{1,1}$ to reduce noise.

4.3. Detail disentanglement

Optimizing L_{detail} enables us to reconstruct faces with mid-frequency details. Making these detail reconstructions animatable however requires us to disentangle person specific details (i.e. moles, pores, eyebrows, and expression-independent wrinkles) controlled by $\boldsymbol{\delta}$ from expression-dependent wrinkles (i.e. wrinkles that change for varying facial expression) controlled by FLAME's expression and jaw pose parameters, $\boldsymbol{\psi}$ and $\boldsymbol{\theta}_{jaw}$. Our key observation is that the same person in two images should have both similar coarse geometry *and* personalized details. So for the rendered detail image, *exchanging the detail codes between two images of the same subject should have no effect on the rendered image.*

Detail consistency loss: Given two images I_i and I_j of the same subject (i.e. $c_i = c_j$), the loss is defined as

$$\begin{aligned} L_{dc} = L_{detail}(I_i, \mathcal{R}(M(\boldsymbol{\beta}_i, \boldsymbol{\theta}_i, \boldsymbol{\psi}_i), A(\boldsymbol{\alpha}_i), \\ F_d(\boldsymbol{\delta}_j, \boldsymbol{\psi}_i, \boldsymbol{\theta}_{jaw,i}), \mathbf{l}_i, \mathbf{c}_i)), \end{aligned} \tag{15}$$

where $\boldsymbol{\beta}_i$, $\boldsymbol{\theta}_i$, $\boldsymbol{\psi}_i$, $\boldsymbol{\theta}_{jaw,i}$, $\boldsymbol{\alpha}_i$, $\mathbf{l}_i$, and $\boldsymbol{c}_i$ are the parameters of I_i, while $\boldsymbol{\delta}_j$ is the detail code of I_j (see Fig. 3). We show the necessity and effectiveness of L_{dc} in Sec. 6.3.

5. Implementation Details

Data: We train DECA on three publicly available datasets: VGGFace2 [10], BUPT-Balancedface [77] and VoxCeleb2 [16]. VGGFace2 [10] contains images of over $8k$ subjects, with an average of more than 350 images per subject. BUPT-Balancedface [77] offers $7k$ subjects per ethnicity (i.e. Caucasian, Indian, Asian and African), and VoxCeleb2 [16] contains $145k$ videos of $6k$ subjects. In total, DECA is trained on 2 Million images.

All datasets provide an identity label for each image. We use FAN [8] to predict 68 2D landmarks $\mathbf{k}_i$ on each face. To improve the robustness of the predicted landmarks, we run FAN for each image twice with different face crops, and discard all images with non-matching landmarks. See Sup. Mat. for details on data selection and data cleaning.

Implementation details: DECA is implemented in PyTorch [48], using the differentiable rasterizer from Pytorch3D [51] for rendering. We use Adam [39] as optimizer with a learning rate of $1e-4$. The input image size is 224^2 and UV space size $d = 256$. See Sup. Mat. for details.

6. Evaluation

6.1. Qualitative evaluation

Reconstruction: Given a single face image, DECA reconstructs the 3D face shape with mid-frequency geometry details. The second row of Fig. 1 shows that the coarse shape (i.e. in FLAME space) well represents the overall face shape, and the learned DECA detail model reconstructs subject-specific details and wrinkles of the input identity (Fig. 1 row three), while being robust to partial occlusions.

Figure 4 qualitatively compares DECA results with state-of-the-art coarse face reconstruction methods, namely PRNet [21], RingNet [58], Deng et al. [18], FML [67] and 3DDFA-V2 [29]. Compared to these methods, DECA better reconstructs the overall face shape with details like the nasolabial fold (rows 1, 2, 3, 4, and 6) and forehead wrinkles (row 3). DECA better reconstructs the mouth shape and the eye region than all other methods. DECA further reconstructs a full head while PRNet [21], Deng et al. [18], FML [67] and 3DDFA-V2 [29] reconstruct tightly cropped faces. While RingNet [58], like DECA, is based on FLAME [42], DECA better reconstructs the face shape and the facial expression.

Figure 5 compares DECA visually to existing detail face reconstruction methods, namely Extreme3D [73], Cross-modal [1], and FaceScape [82]. Extreme3D [73] and Cross-modal [1] reconstruct more details than DECA but at the cost of being less robust to occlusions (rows 1, 2, 3). Unlike DECA, Extreme3D and Cross-modal only reconstruct static details. However, using static details instead of DECA's animatable details leads to visible artifacts when animating the face (see Fig. 6). While FaceScape [82] provides animatable details, unlike DECA, the method is trained on high-resolution scans while DECA is solely trained on in-the-wild images. Also, with occlusion, FaceScape produces artifacts (rows 1, 2) or effectively fails (row 3).

In summary, DECA produces high-quality reconstructions, outperforming previous work in terms of robustness, while enabling animation of the detailed reconstruc-

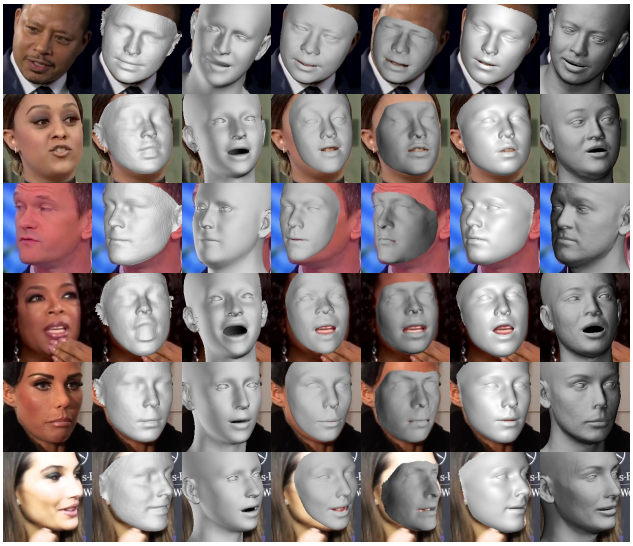


Figure 4: Comparison to other coarse reconstruction methods, from left to right: PRNet [21], RingNet [58] Deng et al. [18], FML [67], 3DDFA-V2 [29], DECA (ours).

tion. To demonstrate the quality of DECA and the robustness to variations in head pose, expression, occlusions, image resolution, lighting conditions, etc., we show results for 200 randomly selected ALFW2000 [84] images in the Sup. Mat. along with more qualitative coarse and detail reconstruction comparisons to the state-of-the-art.

Detail animation: DECA models detail displacements as a function of subject-specific detail parameters $\boldsymbol{\delta}$ and FLAME's jaw pose $\boldsymbol{\theta}_{jaw}$ and expression parameters $\boldsymbol{\psi}$. This formulation allows us to animate detailed facial geometry such that wrinkles are specific to the source shape and expression as shown in Fig. 1. Using static details instead of DECA's animatable details (i.e. by using the reconstructed details as a static displacement map) and animating only the coarse shape by changing the FLAME parameters results in visible artifacts as shown in Fig. 6 (top), while animatable details (middle) look similar to the reference shape (bottom) of the same identity. The Sup. Mat. shows more comparisons of animatable and static details.

6.2. Quantitative evaluation

We compare DECA with publicly available methods, namely 3DDFA-V2 [29], Deng et al. [18], RingNet [58], PRNet [21], 3DMM-CNN [72] and Extreme3D [73].

NoW benchmark: The NoW challenge [58] consists of 2054 face images of 100 subjects, split into a validation set (20 subjects) and a test set (80 subjects), with a reference 3D face scan per subject. The images consist of indoor and outdoor images, neutral expression and expressive face images, partially occluded faces, and varying viewing angles ranging from frontal view to profile view, and selfie images.

Figure 5: Comparison to other detail reconstruction methods, from left to right: Extreme3D [73], FaceScape [82], Cross-modal [1], DECA (ours).

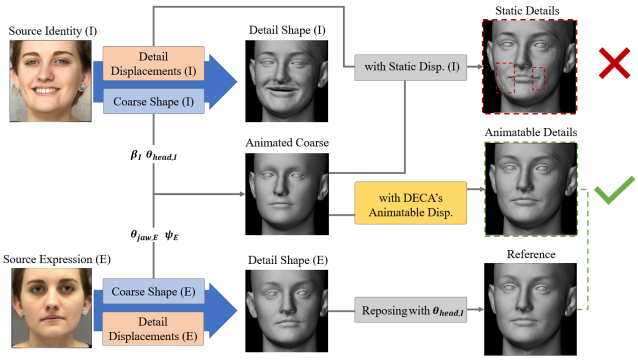


Figure 6: Effect of DECA's animatable details. Given images of source identity I and source expression E (left), DECA reconstructs the detail shapes (middle) and animates the detail shape of I with the expression of E (right, middle). This synthesized DECA expression appears identical to the reconstructed same subject's reference detail shape (right, bottom). Using the reconstructed details of I instead (i.e. static details) and animating the coarse shape only, results in visible artifacts (right, top). See Sec. 6.1 for details.

The challenge provides a standard evaluation protocol that measures the distance from all reference scan vertices to the closest point in the reconstructed mesh surface, after rigidly

Method	Median (mm)	Mean (mm)	Std (mm)
3DMM-CNN [72]	1.84	2.33	2.05
PRNet [21]	1.50	1.98	1.88
Deng et al.19 [18]	1.23	1.54	1.29
RingNet [58]	1.21	1.54	1.31
3DDFA-V2 [29]	1.23	1.57	1.39
DECA (ours)	**1.09**	**1.38**	**1.18**

Table 1: Reconstruction error on the NoW [58] benchmark.

Method	Median (mm)		Mean (mm)		Std (mm)	
	LQ	HQ	LQ	HQ	LQ	HQ
3DMM-CNN [72]	1.88	1.85	2.32	2.29	1.89	1.88
Extreme3D [73]	2.40	2.37	3.49	3.58	6.15	6.75
PRNet [21]	1.79	1.60	2.38	2.06	2.19	1.79
RingNet [58]	1.63	1.58	2.08	2.02	1.79	1.69
3DDFA-V2 [29]	1.62	1.49	2.10	1.91	1.87	**1.64**
DECA (ours)	**1.48**	**1.44**	**1.91**	**1.89**	**1.68**	1.66

Table 2: Feng et al. [22] benchmark performance.

aligning scans and reconstructions. For details, see [12].

We found that the tightly cropped face meshes predicted by Deng et al. [18] are smaller than the NoW reference scans, which would result in a high reconstruction error in the missing region. For a fair comparison to the method of Deng et al. [18], we use the Basel Face Model (BFM) [49] parameters they output, reconstruct the complete BFM mesh, and get the NoW evaluation for these complete meshes. As shown in Tab. 1 and the cumulative error plot in the Sup. Mat., DECA gives state-of-the-art results on NoW, providing the reconstruction error with the lowest mean, median, and standard deviation.

Feng et al. benchmark: The Feng et al. challenge [22] contains 2000 face images of 135 subject, and a reference 3D face scan for each subject. The benchmark consists of 1344 low-quality (LQ) images extracted from videos, and 656 high-quality (HQ) images taken in controlled scenarios. A protocol similar to NoW is used for evaluation that measures the distance between all reference scan vertices to the closest points on the reconstructed mesh surface, after rigidly aligning scan and reconstruction. As shown in Tab. 2 and the cumulative error plot in the Sup. Mat., DECA provides state-of-the-art performance.

6.3. Ablation experiment

Detail consistency loss: To evaluate the importance of our novel detail consistency loss L_{dc} (Eq. 15), we train DECA with and without L_{dc}. Figure 7 (left) shows the DECA details for detail code $\boldsymbol{\delta}_I$ from the source identity, and expression $\boldsymbol{\psi}_E$ and jaw pose parameters $\boldsymbol{\theta}_{jaw,E}$ from the source expression. For DECA trained with L_{dc} (top), wrinkles appear in the forehead as a result of the raised eyebrows of the source expression, while for DECA trained without L_{dc} (bottom), no such wrinkles appear. This indicates that without L_{dc}, person-specific details and expression-dependent wrinkles are not well disentangled. See Sup. Mat. for more disentanglement results.

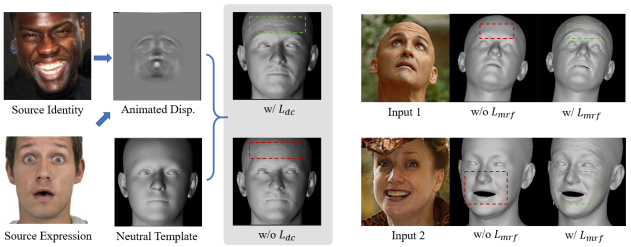


Figure 7: Ablation experiments. Left: Effects of L_{dc} on the animation of the source identity with the source expression visualized on a neutral expression template mesh. Without L_{dc}, no wrinkles appear in the forehead due to the surprise source expression. Right: Effect of L_{mrf} on the detail reconstruction. Without L_{mrf}, less details are reconstructed.

ID-MRF loss: Figure 7 (right) shows the effect of L_{mrf} on the detail reconstruction. Without L_{mrf} (middle), wrinkle details (e.g. in the forehead) are not reconstructed, resulting in an overly smooth result. With L_{mrf} (right), DECA captures the details.

7. Conclusion and discussion

We have presented DECA, which enables detailed expression capture and animation from single images by learning an animatable detail model from in-the-wild images. In total, DECA is trained from about 2M in-the-wild face images without 2D-to-3D supervision. DECA reaches state-of-the-art shape reconstruction performance enabled by a shape consistency loss. A novel detail consistency loss helps DECA to disentangle expression-dependent wrinkles from person-specific details. The low-dimensional detail latent space makes the fine-scale reconstruction robust to noise and occlusions, and the novel loss leads to disentanglement of identity and expression-dependent wrinkle details. This enables applications like animation, shape change, wrinkle transfer, etc. DECA is publicly available for research purposes. Due to the reconstruction accuracy, the reliability, and the speed, DECA is useful for applications like face reenactment or virtual avatar creation.

DECA opens the door for future work. First, our albedo model is dependent on the Basel face model, which lacks ethnic diversity and facial hair. This pushes skin tone into the lighting model and causes facial hair to be explained by shape deformations. We believe that we can learn a more diverse albedo model from in-the-wild images using our system. Second, we want to extend the model over time, both for tracking and to learn more personalized models of individuals from video where we could enforce continuity of intrinsic wrinkles over time. Third, while robust, our method

can still fail due to extreme head pose and lighting. This suggests the need for more diverse training data.

8. Acknowledgements

We thank S. Sanyal for providing us the RingNet PyTorch implementation, support with paper writing, and fruitful discussions, and M. Kocabas, N. Athanasiou, and V. Fernández Abrevaya for the helpful suggestions. We further thank all Perceiving Systems department members for the feedback. This work was partially supported by the Max Planck ETH Center for Learning Systems.

Disclosure: MJB has received research gift funds from Intel, Nvidia, Adobe, Facebook, and Amazon. While MJB is a part-time employee of Amazon, his research was performed solely at, and funded solely by, MPI. MJB has financial interests in Amazon and Meshcapade GmbH.

References

[1] Victoria Fernández Abrevaya, Adnane Boukhayma, Philip HS Torr, and Edmond Boyer. Cross-modal deep face normals with deactivable skip connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4979–4989, 2020. 1, 3, 6, 7, 15

[2] Oswald Aldrian and William AP Smith. Inverse rendering of faces with a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(5):1080–1093, 2013. 2

[3] Anil Bas, William A. P. Smith, Timo Bolkart, and Stefanie Wuhrer. Fitting a 3D morphable model to edges: A comparison between hard and soft correspondences. In *Asian Conference on Computer Vision Workshops*, pages 377–391, 2017. 2

[4] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-quality single-shot capture of facial geometry. *ACM Transactions on Graphics (TOG)*, 29(4):40, 2010. 2

[5] Volker Blanz, Sami Romdhani, and Thomas Vetter. Face identification across different poses and illuminations with a 3D morphable model. In *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 202–207, 2002. 1, 2

[6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, pages 187–194, 1999. 2

[7] Alan Brunton, Augusto Salazar, Timo Bolkart, and Stefanie Wuhrer. Review of statistical shape spaces for 3D data with comparative analysis for human faces. *Computer Vision and Image Understanding (CVIU)*, 128(0):1–17, 2014. 1

[8] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1021–1030, 2017. 6, 13

[9] Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (TOG)*, 34(4):1–9, 2015. 1, 2

[10] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 67–74, 2018. 5, 6, 13

[11] Xuan Cao, Zhang Chen, Anpei Chen, Xin Chen, Shiying Li, and Jingyi Yu. Sparse photometric 3D face reconstruction guided by morphable models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4635–4644, 2018. 2

[12] NoW challenge. https://ringnet.is.tue.mpg.de/challenge, 2019. 8

[13] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. Expnet: Landmark-free, deep, 3D facial expressions. In *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 122–129, 2018. 2

[14] Bindita Chaudhuri, Noranart Vesdapunt, Linda Shapiro, and Baoyuan Wang. Personalized face modeling for improved face reconstruction and motion retargeting. *arXiv preprint arXiv:2007.06759*, 2020. 2, 3

[15] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9429–9439, 2019. 1, 2

[16] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 6, 13

[17] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. Capture, learning, and synthesis of 3D speaking styles. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111, 2019. 1

[18] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In *Computer Vision and Pattern Recognition Workshops*, 2019. 2, 5, 6, 7, 8, 13, 15

[19] Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. End-to-end 3D face reconstruction with deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5908–5917, 2017. 2

[20] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhöfer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3D morphable face models - past, present and future. *ACM Transactions on Graphics (TOG)*, 2020. 1

[21] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3D face reconstruction and dense alignment with position map regression network. In *European Conference on Computer Vision (ECCV)*, pages 534–551, 2018. 2, 6, 7, 8, 13, 15

[22] Zhen-Hua Feng, Patrik Huber, Josef Kittler, Peter Hancock, Xiao-Jun Wu, Qijun Zhao, Paul Koppen, and Matthias

Rätsch. Evaluation of dense 3D reconstruction from 2D face images in the wild. In *International Conference on Automatic Face & Gesture Recognition (FG)*, 2018. 8, 13, 14

[23] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of personalized 3D face rigs from monocular video. *ACM Transactions on Graphics (TOG)*, 35(3):28, 2016. 2

[24] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1164, 2019. 2, 5

[25] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3D morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8377–8386, 2018. 2

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

[27] Aleksey Golovinskiy, Wojciech Matusik, Hanspeter Pfister, Szymon Rusinkiewicz, and Thomas A. Funkhouser. A statistical model for synthesis of detailed facial geometry. *ACM Transactions on Graphics (TOG)*, 25(3):1025–1034, 2006. 3

[28] Riza Alp Güler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. DenseReg: Fully convolutional dense shape regression in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6799–6808, 2017. 2

[29] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3D dense face alignment. In *European Conference on Computer Vision (ECCV)*, 2020. 6, 7, 8, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22

[30] Yudong Guo, Jianfei Cai, Boyi Jiang, Jianmin Zheng, et al. CNN-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 41(6):1294–1307, 2018. 1, 2

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[32] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (TOG)*, 36(6):195:1–195:14, 2017. 1

[33] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. Dynamic 3D avatar creation from hand-held video input. *ACM Transactions on Graphics (TOG)*, 34(4):45, 2015. 2

[34] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1031–1039, 2017. 2

[35] László A Jeni, Jeffrey F Cohn, and Takeo Kanade. Dense 3D face alignment from 2D videos in real-time. In *International Conference on Automatic Face & Gesture Recognition (FG)*, volume 1, pages 1–8, 2015. 2

[36] Luo Jiang, Juyong Zhang, Bailin Deng, Hao Li, and Ligang Liu. 3D face reconstruction with geometry details from a single image. *Transactions on Image Processing*, 27(10):4756–4770, 2018. 2

[37] Ira Kemelmacher-Shlizerman and Steven M Seitz. Face reconstruction in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1746–1753, 2011. 2

[38] 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 (CVPR)*, pages 4625–4634, 2018. 2

[39] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 6

[40] Hao Li, Bart Adams, Leonidas J. Guibas, and Mark Pauly. Robust single-view geometry and motion reconstruction. *ACM Transactions on Graphics (TOG)*, 28:175, 2009. 4

[41] Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. Realtime facial animation with on-the-fly correctives. *ACM Transactions on Graphics (TOG)*, 32(4):42–1, 2013. 2

[42] 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), 2017. 2, 3, 6

[43] Yue Li, Liqian Ma, Haoqiang Fan, and Kenny Mitchell. Feature-preserving detailed 3D face reconstruction from a single image. In *European Conference on Visual Media Production*, pages 1–9, 2018. 2

[44] Claus Müller. *Spherical Harmonics*. Springer Berlin Heidelberg, 1966. 3

[45] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. paGAN: real-time avatars using dynamic textures. *ACM Transactions on Graphics (TOG)*, 37(6):258:1–258:12, 2018. 3

[46] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 98–105, 2018. 3, 5

[47] Frederick Ira Parke. A parametric model for human faces. Technical report, University of Utah, 1974. 2

[48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 6

[49] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In *International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009. 3, 8

[50] Frédéric Pighin, Jamie Hecker, Dani Lischinski, Richard Szeliski, and David H. Salesin. Synthesizing realistic facial expressions from photographs. In *SIGGRAPH*, pages 75–84, 1998. 2

[51] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Pytorch3d. https://github.com/facebookresearch/pytorch3d, 2020. 6

[52] E. Richardson, M. Sela, and R. Kimmel. 3D face reconstruction by learning from synthetic data. In *3D Vision*, pages 460–469, 2016. 2

[53] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1259–1268, 2017. 1, 2, 3

[54] Sami Romdhani, Volker Blanz, and Thomas Vetter. Face identification by fitting a 3D morphable model using linear shape and texture error functions. In *European Conference on Computer Vision (ECCV)*, pages 3–19, 2002. 1

[55] S. Romdhani and T. Vetter. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 986–993, 2005. 2

[56] Joseph Roth, Yiying Tong, and Xiaoming Liu. Adaptive 3D face reconstruction from unconstrained photo collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4197–4206, 2016. 2

[57] Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. Photorealistic facial texture inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5144–5153, 2017. 2

[58] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7763–7772, 2019. 2, 5, 6, 7, 8, 13, 14

[59] Kristina Scherbaum, Tobias Ritschel, Matthias Hullin, Thorsten Thormählen, Volker Blanz, and Hans-Peter Seidel. Computer-suggested facial makeup. *Computer Graphics Forum*, 30(2):485–492, 2011. 1

[60] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1576–1585, 2017. 3

[61] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. SfSNet: learning shape, reflectance and illuminance of faces in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6296–6305, 2018. 3

[62] Fuhao Shi, Hsiang-Tao Wu, Xin Tong, and Jinxiang Chai. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Transactions on Graphics (TOG)*, 33(6):222, 2014. 2

[63] Il-Kyu Shin, A Cengiz Öztireli, Hyeon-Joong Kim, Thabo Beeler, Markus Gross, and Soo-Mi Choi. Extraction and transfer of facial expression wrinkles for facial performance enhancement. In *Pacific Conference on Computer Graphics and Applications*, pages 113–118, 2014. 3

[64] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 6

[65] Ron Slossberg, Gil Shamai, and Ron Kimmel. High quality facial surface and texture synthesis via generative adversarial networks. In *European Conference on Computer Vision Workshops (ECCV-W)*, 2018. 2

[66] Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M Seitz. Total moving face reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 796–812, 2014. 2

[67] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. FML: Face Model Learning from Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10812–10822, 2019. 3, 6, 7

[68] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 Hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2549–2559, 2018. 3

[69] Ayush Tewari, Michael Zollhöfer, 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 (ICCV)*, pages 1274–1283, 2017. 2

[70] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Transactions on Graphics (TOG)*, 34(6):183–1, 2015. 2

[71] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2Face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, 2016. 1, 2

[72] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard 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 (CVPR)*, pages 1599–1608, 2017. 2, 7, 8, 13

[73] Anh Tuan Tran, 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 (CVPR), pages 3935–3944, 2018. 1, 2, 3, 6, 7, 8, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22

[74] 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 (CVPR)*, pages 1126–1135, 2019. 1, 3

[75] Xiaoguang Tu, Jian Zhao, Zihang Jiang, Yao Luo, Mei Xie, Yang Zhao, Linxiao He, Zheng Ma, and Jiashi Feng. Joint 3D face reconstruction and dense face alignment from a single image with 2D-assisted self-supervised learning. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[76] Thomas Vetter and Volker Blanz. Estimating coloured 3D face models from single images: An example based approach. In *European Conference on Computer Vision (ECCV)*, pages 499–513, 1998. 1, 2

[77] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2019. 6, 13

[78] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 331–340, 2018. 5, 6

[79] Huawei Wei, Shuang Liang, and Yichen Wei. 3D dense face alignment via graph convolution networks. *arXiv preprint arXiv:1904.05562*, 2019. 2

[80] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. Realtime performance-based facial animation. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 30(4):77, 2011. 2

[81] Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)*, 37(4):162:1–162:14, 2018. 2

[82] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. FaceScape: a large-scale high quality 3D face dataset and detailed riggable 3D face prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 601–610, 2020. 2, 3, 6, 7, 14, 15, 16, 17, 18, 19, 20, 21, 22

[83] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. DF2Net: A dense-fine-finer network for detailed 3D face reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[84] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 787–796, 2015. 7, 14, 16, 17, 18, 19, 20, 21, 22

[85] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. State of the art on monocular 3D face reconstruction, tracking, and applications. *Computer Graphics Forum (Eurographics State of the Art Reports 2018)*, 37(2), 2018. 1, 2

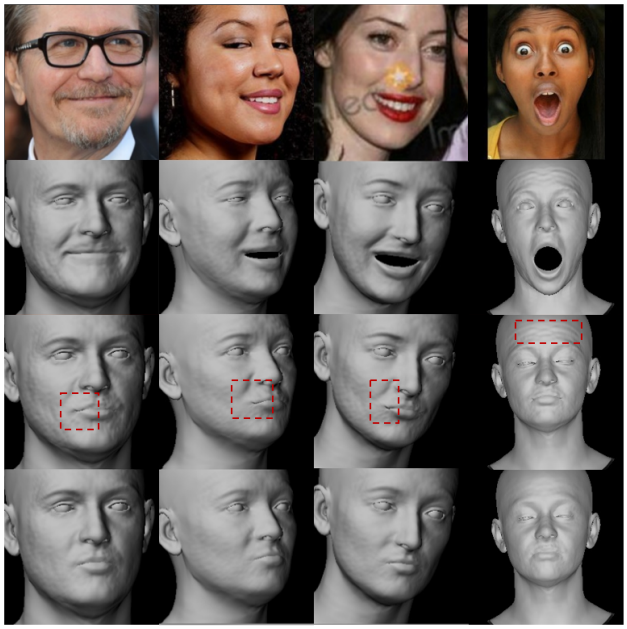


Figure 8: Effect of DECA's animatable details. Given a single image (top), DECA reconstructs a detailed mesh (second row). Using static details and animating the coarse FLAME shape only (third row) results in visible artifacts as highlighted by the red boxes. Instead, reposing with DECA's animatable details (bottom) results in a more realistic mesh with geometric details.

Appendices

A. Implementation Details

Data: DECA is trained on 2 Million images from VGGFace2 [10] and BUPT-Balancedface [77] and VoxCeleb2 [16]. From VGGFace2 [10], we randomly select $950k$ images such that $750K$ images are of resolution higher than 224×224, and $200K$ are of lower resolution. From BUPT-Balancedface [77] we randomly sample $550k$ with Asian or African ethnicity labels to reduce the ethnicity bias of VGGFace2. From VoxCeleb2 [16] we choose $500k$ frames, with multiple samples from the same video clip per subject to obtain data with variation only in the facial expression and head pose. We also sample $50k$ images from the VGGFace2 [10] test set for validation.

Data cleaning: We generate a different crop for the face image by shifting the provided bounding box by 5% to the bottom right (i.e. shift by $\boldsymbol{\epsilon} = \frac{1}{20}(b_w, b_h)^T$, where b_w and b_h denote the bounding box width and height). Then we expand the original and the shifted bounding boxes by 10% to the top, and by 20% to the left, right, and bottom. We run FAN [8], providing the expanded bounding boxes as input and discard all images with $\max_i \left\|\mathbf{D}(\mathbf{k}_i^2 - \boldsymbol{\epsilon} - \mathbf{k}_i^1)\right\| \geq 0.1$, where $\mathbf{k}_i^2$ and $\mathbf{k}_i^1$ are the ith landmarks for the original and the shifted bounding box, respectively, and $\mathbf{D}$ denote the normalization matrix $diag(b_w, b_h)^{-1}$.

Training details: We pre-train the coarse model (i.e. E_c) for two epochs with a batch size of 64 with $\lambda_{lmk} = 1e-4$, $\lambda_{eye} = 1.0$, $\lambda_{\boldsymbol{\beta}} = 1e-4$, and $\lambda_{\boldsymbol{\psi}} = 1e-4$. Then, we train the coarse model for 1.5 epochs with a batch size of 32, with 4 images per subject with $\lambda_{pho} = 2.0$, $\lambda_{id} = 0.2$, $\lambda_{sc} = 1.0$, $\lambda_{lmk} = 1.0$, $\lambda_{eye} = 1.0$, $\lambda_{\boldsymbol{\beta}} = 1e-4$, and $\lambda_{\boldsymbol{\psi}} = 1e-4$. The landmark loss uses different weights for individual landmarks, the mouth corners and the nose tip landmarks are weighted by a factor of 3, other mouth and nose landmarks with a factor of 1.5, and all remaining landmarks have a weight of 1.0. This is followed by training the detail model (i.e. E_d and F_d) on VGGFace2 and VoxCeleb2 with a batch size of 6, with 3 images per subject, and parameters $\lambda_{phoD} = 2.0$, $\lambda_{mrf} = 5e-2$, $\lambda_{sym} = 5e-3$, $\lambda_{dc} = 1.0$, and $\lambda_{regD} = 5e-3$. The coarse model is fixed while training the detail model.

B. Evaluation

B.1. Detail animation

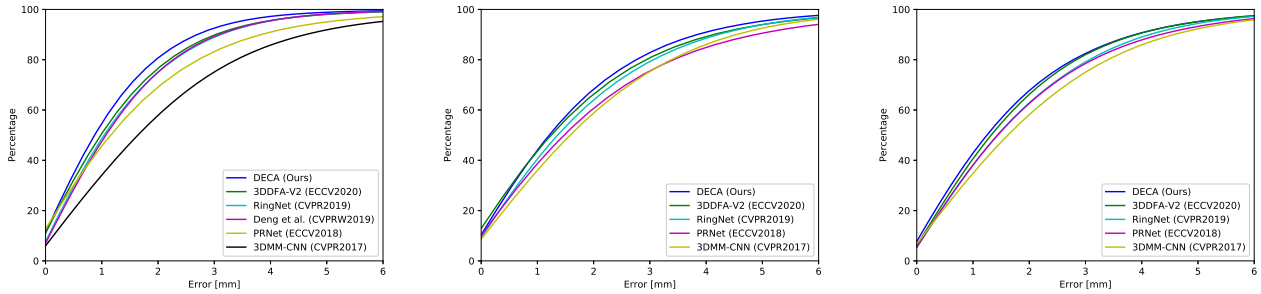
As described in Section 6.1 and shown in Figure 6 of the main paper, using a static displacement map to model geometric details instead of DECA's animatable details results in visible artifacts. Figure 8 shows more examples where using static details results in artifacts in the mouth corner or the forehead region, while DECA's animated results look plausible.

B.2. Quantitative evaluation

As described in Section 6.2 of the main paper, we quantitatively compare DECA with publicly available methods, namely 3DDFA-V2 [29], Deng et al. [18], RingNet [58], PRNet [21], 3DMM-CNN [72] and Extreme3D [73] on two existing 3D face reconstruction benchmarks, the NoW challenge [58] and the Feng et al. [22] benchmark. The left of Figure 9 shows the cumulative errors for Table 1 of the main paper, the middle and right of Figure 9 show the cumulative errors for Table 2 of the main paper. Note that in all cases, the DECA curve in dark blue is aboth that of the other methods. This demonstrates that DECA gives state-of-the-art reconstruction performance for both benchmarks.

B.3. Qualitative comparisons

Figure 10 shows additional qualitative comparisons to existing coarse and detail reconstruction methods. DECA better reconstructs the overall face shape than all existing methods, it reconstructs more details than existing coarse reconstruction methods (e.g. (b), (e), (f)), and it is more robust to occlusions compared to existing detail reconstruction methods (e.g. (c), (d), (g)).



NoW [58] Feng et al. [22] LQ Feng et al. [22] HQ

Figure 9: Quantitative comparison to state-of-the-art on two 3D face reconstruction benchmarks, namely the NoW [58] challenge (left) and the Feng et al. [22] benchmark for low-quality (middle) and high-quality (right) images.

As promised in the main paper (e.g. Section 6.1), we show results for more than 200 randomly selected ALFW2000 [84] samples in Figures 11, 12, 13, 14, 15, 16, and 17. For each sample, we compare the DECA's detail reconstruction (e) with the state-of-the-art coarse reconstruction method 3DDFA-V2 [29] (see (b)) and existing detail reconstruction methods, namely FaceScape [82] (see (c)), and Extreme3D [73] (see (e)). In total, DECA reconstructs more details then 3DDFA-V2, and it is more robust to occlusions than FaceScape and Extreme3D. Further, the DECA retargeting results appear realistic.

Figure 10: Comparison to previous work, from left to right: (a) Input image, (b) 3DDFA-V2 [29], (c) FaceScape [82], (d) Extreme3D [73], (e) PRNet [21], (f) Deng et al. [18], (g) Cross-modal [1], (h) DECA detail reconstruction, and (i) reposing (animation) of DECA's detail reconstruction to a common expression. The expression in (i) is from the source expression E in Figure 2 of the main paper. Blank entries indicate that the particular method did not return any reconstructed mesh.

Figure 11: Qualitative comparisons on random ALFW2000 [84] samples. a) Input image, b) 3DDFA-V2 [29], c) FaceScape [82], d) Extreme3D [73], e) DECA detail reconstruction, and f) reposing (animation) of DECA's detail reconstruction to a common expression. The expression in (i) is from the source expression E in Figure 2 of the main paper. Blank entries indicate that the particular method did not return any reconstructed mesh.

Figure 12: Qualitative comparisons on random ALFW2000 [84] samples. a) Input image, b) 3DDFA-V2 [29], c) FaceScape [82], d) Extreme3D [73], e) DECA detail reconstruction, and f) reposing (animation) of DECA's detail reconstruction to a common expression. The expression in (i) is from the source expression E in Figure 2 of the main paper. Blank entries indicate that the particular method did not return any reconstructed mesh.

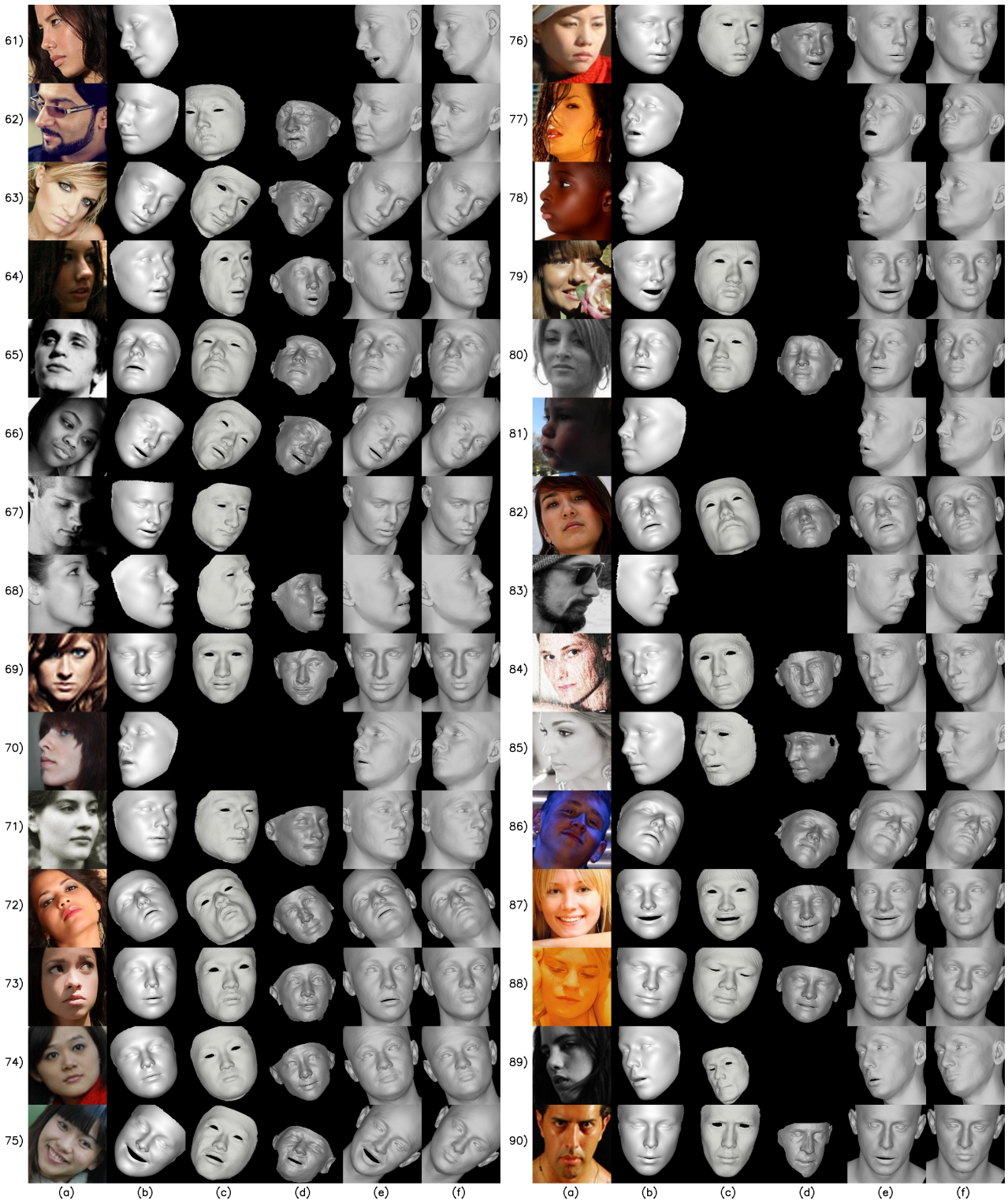


Figure 13: Qualitative comparisons on random ALFW2000 [84] samples. a) Input image, b) 3DDFA-V2 [29], c) FaceScape [82], d) Extreme3D [73], e) DECA detail reconstruction, and f) reposing (animation) of DECA's detail reconstruction to a common expression. The expression in (i) is from the source expression E in Figure 2 of the main paper. Blank entries indicate that the particular method did not return any reconstructed mesh.

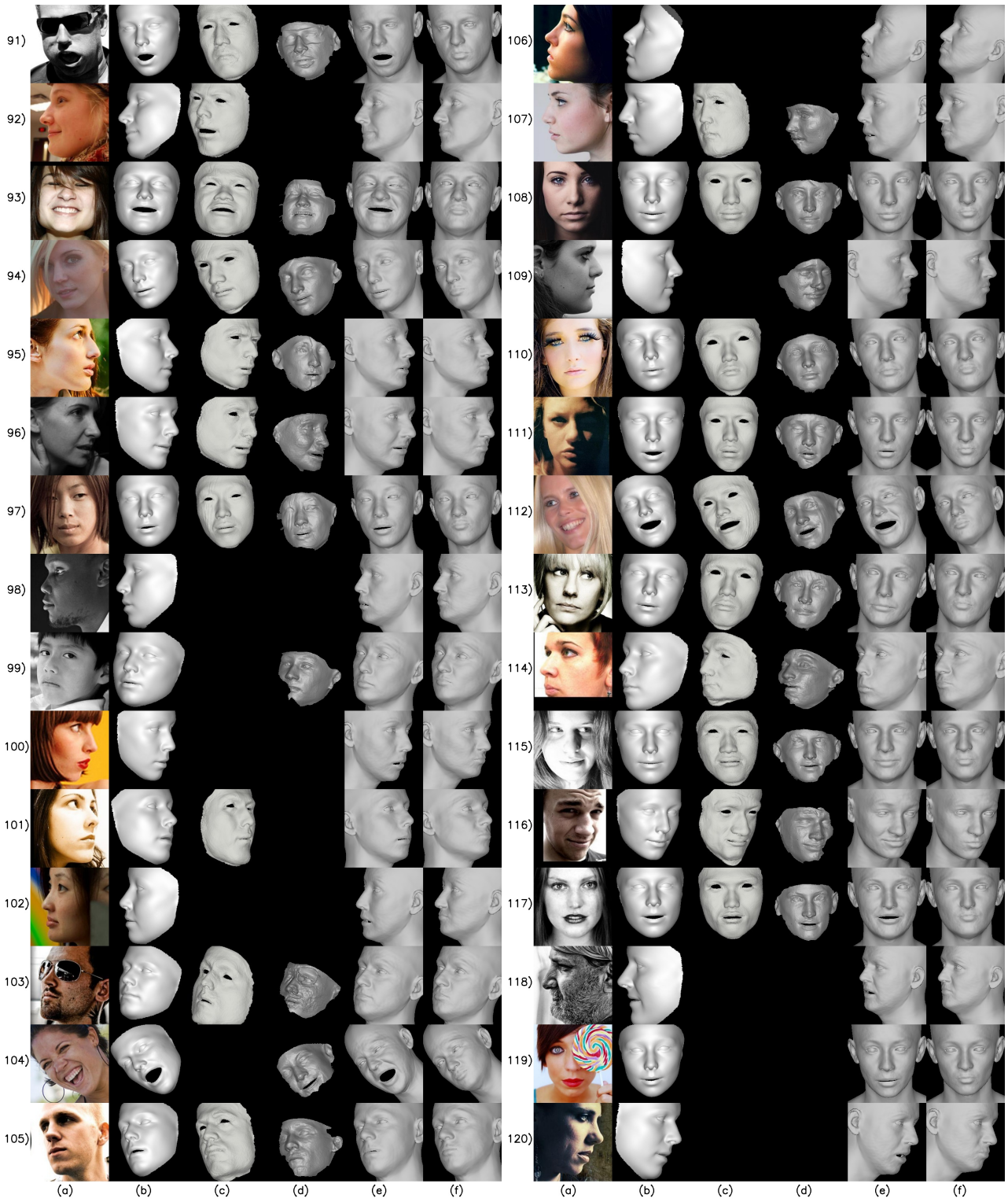


Figure 14: Qualitative comparisons on random ALFW2000 [84] samples. a) Input image, b) 3DDFA-V2 [29], c) FaceScape [82], d) Extreme3D [73], e) DECA detail reconstruction, and f) reposing (animation) of DECA's detail reconstruction to a common expression. The expression in (i) is from the source expression E in Figure 2 of the main paper. Blank entries indicate that the particular method did not return any reconstructed mesh.

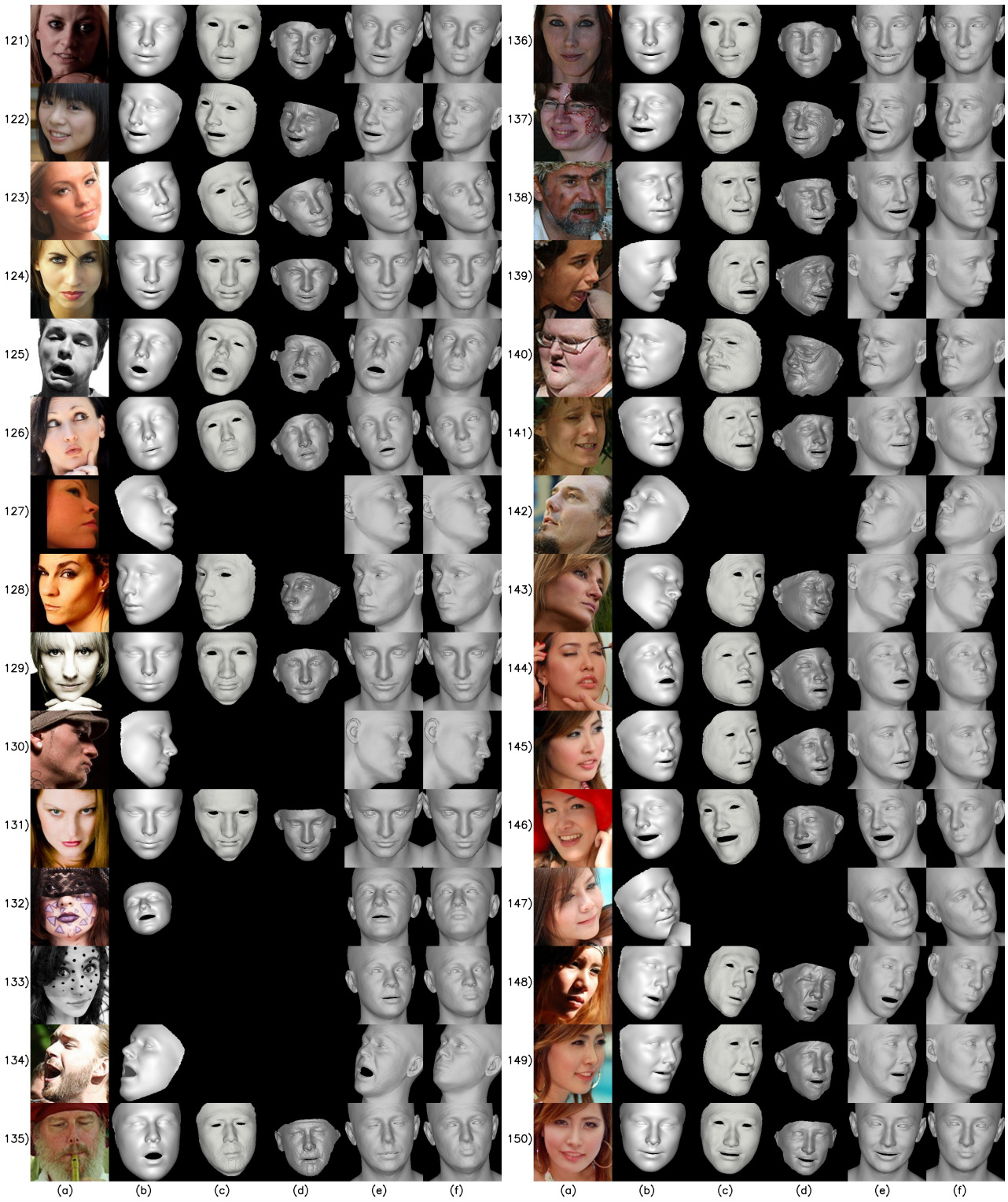


Figure 15: Qualitative comparisons on random ALFW2000 [84] samples. a) Input image, b) 3DDFA-V2 [29], c) FaceScape [82], d) Extreme3D [73], e) DECA detail reconstruction, and f) reposing (animation) of DECA's detail reconstruction to a common expression. The expression in (i) is from the source expression E in Figure 2 of the main paper. Blank entries indicate that the particular method did not return any reconstructed mesh.

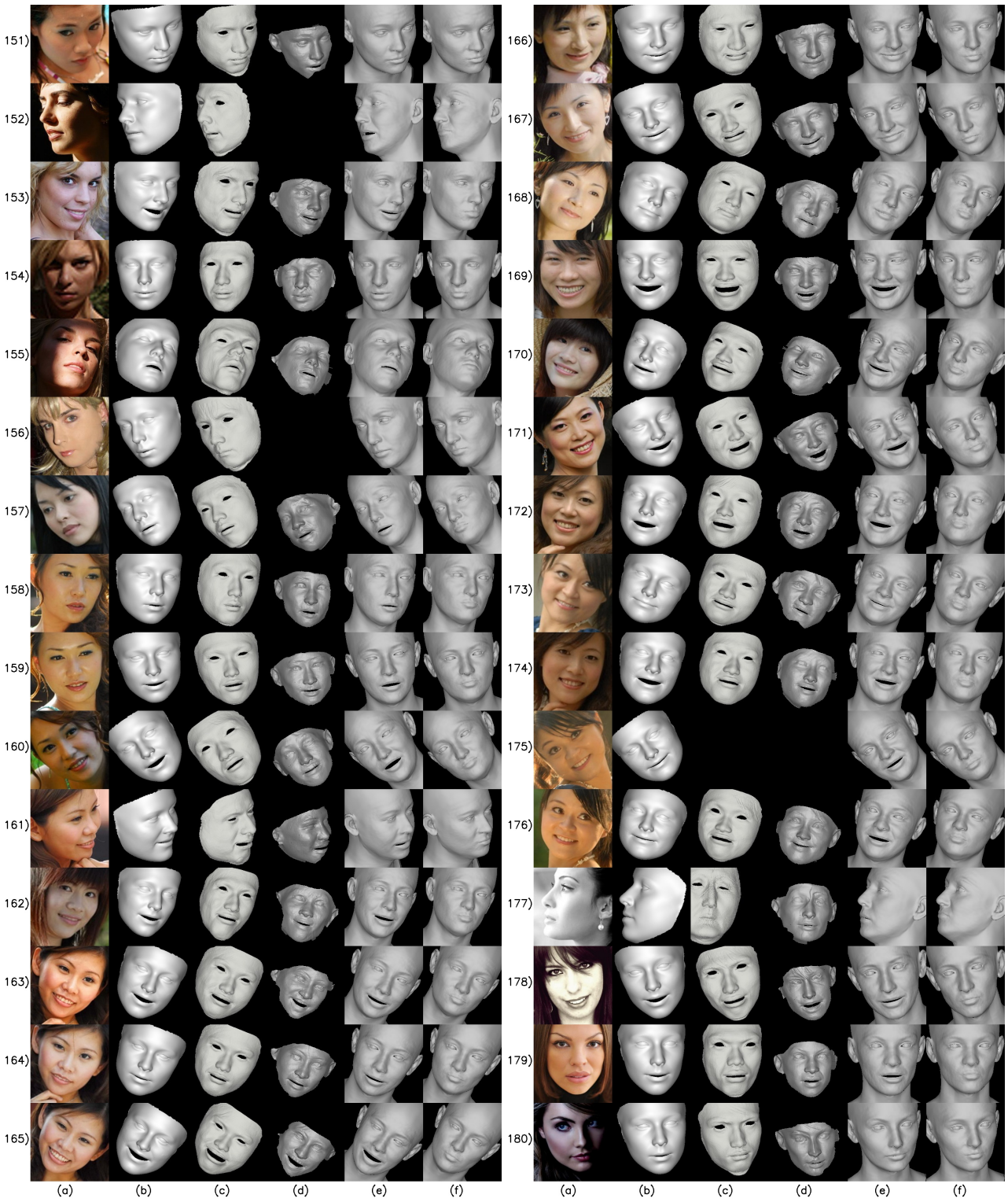


Figure 16: Qualitative comparisons on random ALFW2000 [84] samples. a) Input image, b) 3DDFA-V2 [29], c) FaceScape [82], d) Extreme3D [73], e) DECA detail reconstruction, and f) reposing (animation) of DECA's detail reconstruction to a common expression. The expression in (i) is from the source expression E in Figure 2 of the main paper. Blank entries indicate that the particular method did not return any reconstructed mesh.

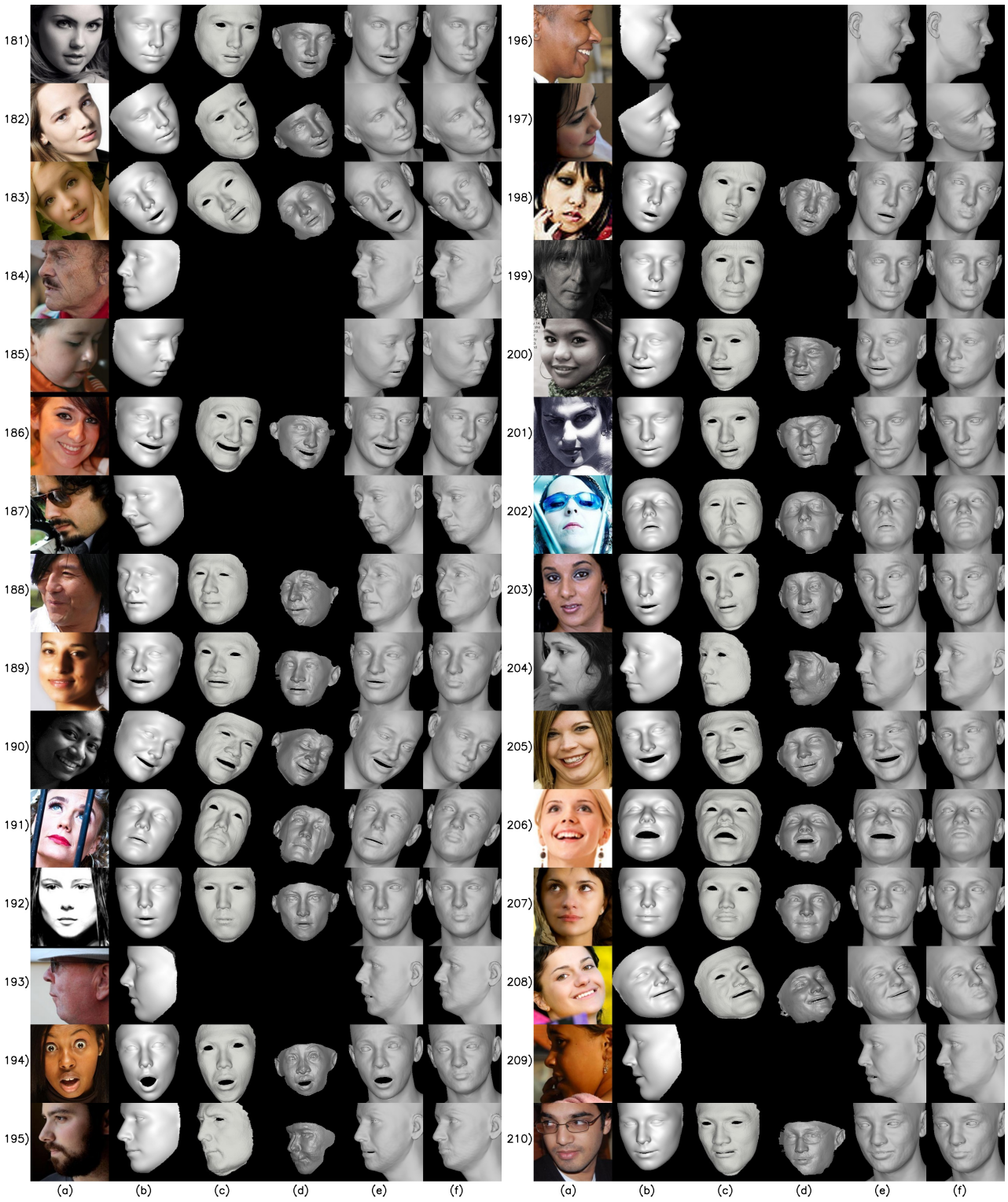


Figure 17: Qualitative comparisons on random ALFW2000 [84] samples. a) Input image, b) 3DDFA-V2 [29], c) FaceScape [82], d) Extreme3D [73], e) DECA detail reconstruction, and f) reposing (animation) of DECA's detail reconstruction to a common expression. The expression in (i) is from the source expression E in Figure 2 of the main paper. Blank entries indicate that the particular method did not return any reconstructed mesh.