This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3166666, IEEE Transactions on Visualization and Computer Graphics

Semantically Disentangled Variational Autoencoder for Modeling 3D Facial Details

Jingwang Ling, Zhibo Wang, Ming Lu, Quan Wang, Chen Qian and Feng Xu†

Abstract—Parametric face models, such as morphable and blendshape models, have shown great potential in face representation, reconstruction, and animation. However, all these models focus on large-scale facial geometry. Facial details such as wrinkles are not parameterized in these models, impeding accuracy and realism. In this paper, we propose a method to learn a Semantically Disentangled Variational Autoencoder (SDVAE) to parameterize facial details and support independent detail manipulation as an extension of an off-the-shelf large-scale face model. Our method utilizes the non-linear capability of Deep Neural Networks for detail modeling, achieving better accuracy and greater representation power compared with linear models. In order to disentangle the semantic factors of identity, expression and age, we propose to eliminate the correlation between different factors in an adversarial manner. Therefore, wrinkle-level details of various identities, expressions, and ages can be generated and independently controlled by changing latent vectors of our SDVAE. We further leverage our model to reconstruct 3D faces via fitting to facial scans and images. Benefiting from our parametric model, we achieve accurate and robust reconstruction, and the reconstructed details can be easily animated and manipulated. We evaluate our method on practical applications, including scan fitting, image fitting, video tracking, model manipulation, and expression and age animation. Extensive experiments demonstrate that the proposed method can robustly model facial details and achieve better results than alternative methods.

Index Terms—Detail reconstruction, facial animation, semantic disentanglement.

1 Introduction

Human faces always draw the most attention in our daily communication. Therefore, it usually takes the majority of the effort in generating CG characters for gaming and movie industries. Parametric face models are often used to efficiently represent facial shape variation and flexibly control its deformation. 3D morphable models are one kind of popular parametric face models, which can be represented by PCA [1], High-Order PCA [2], local PCA [3], Graph Convolution Network [4] or Skinned Multi-Person Linear model [5]. 3D faces can be reconstructed by fitting those morphable models to facial scans or images. However, these methods usually consider only the large-scale facial geometry. With the development of 3D acquisition techniques, more accurate facial details can be captured and lead to new face-related research trends. Therefore, it is crucial to devise an effective parametric model for 3D facial details.

This paper proposes the Semantically Disentangled Variational Autoencoder (SDVAE) to model 3D facial details. SDVAE aims to purely represent facial details and leave the large-scale geometry to traditional face models. We design a detail model decoupled from large-scale geometry based on the following considerations. First, large-scale parametric face models are widely used in academia and industry. A pure detail model can be easily integrated into many existing techniques based on parametric models. Second, the decoupling provides better flexibility and efficiency for face modeling. Existing works usually disentangle identity and expression when modeling large-scale geometry. As for fine-scale geometry, we argue that age is also essential as it could independently affect facial details. Therefore, we design a detail model when with independent control of identity, expression and age.

Some specific challenges need to be handled to train a model for facial details. First, facial details are highly nonlinear, thus methods like linear models [1], [2], [3] fail to generate reasonable results. Recent works [4], [5], [6] build a Graph Convolutional Neural Network (GCN) for 3D faces or bodies. However, we find GCNs cannot capture sufficient facial details in our preliminary experiments. Instead, we learn the model for facial details based on Variational Autoencoder (VAE). Specifically, we represent facial details as a displacement map, whose pixels represent offsets along the normal directions of the large-scale face. We can embed facial details to identity, expression, and age latent vectors via the bottleneck structure of VAE. Thus, we can also reconstruct the detailed 3D face by fitting the embedded latent vectors. Under the framework of a VAE, we can learn a controllable model with accurate facial details, which is essential for realistic 3D face reconstruction and animation.

To control the facial details with semantic attributes, we need to disentangle identity, expression and age. In order to explicitly improve the separability, we enforce the independence of the distributions in the latent space. To achieve this, we concatenate the latent vectors and forward it through a discriminator, which distinguishes the coupled and independent distributions. The encoder of VAE will be trained to generate latent distributions which are independent enough to fool the discriminator. Therefore, the correlation of identity, expression and age will be removed

- *J. Ling, Z. Wang and F. Xu are with School of software and BNRist, Tsinghua University.*
- *M. Lu is with Intel Labs China.*
- *Q. Wang and C. Qian are with Sensetime Research.*

†*Corresponding author. Email: xufeng2003@gmail.com.*

1077-2626 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: MICROSOFT. Downloaded on December 20,2022 at 08:57:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3166666, IEEE Transactions on Visualization and Computer Graphics

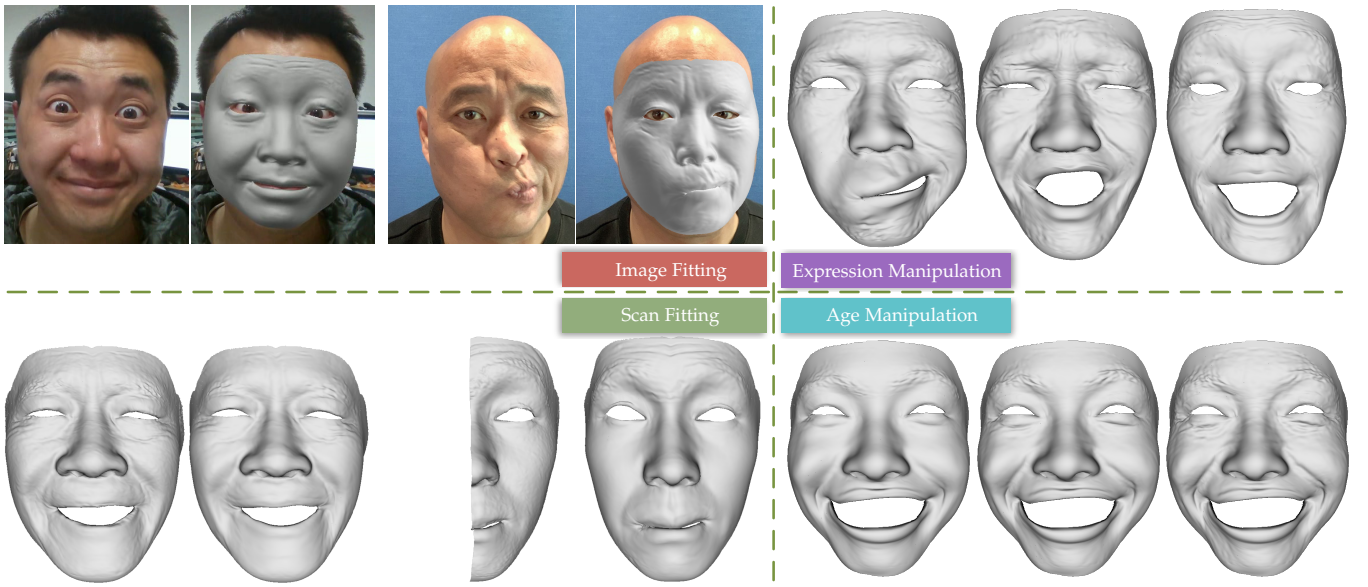

Fig. 1. Our parametric model can reconstruct facial details from an image or a facial scan, as shown by the image reconstruction results (top left) and scan fitting results (bottom left). We can accurately fit the given scan and recover the missing part when only a half is given. We show expression editing (top right) and the same identity at different ages (bottom right) to show the model's ability to edit the generated details by varying expression blendshape coefficients and ages.

in an adversarial manner.

Based on the proposed model, we apply it to several applications, including scan fitting, image fitting and model manipulation. We conduct extensive comparisons with previous methods both qualitatively and quantitatively, demonstrating the advantages of our method. Our contributions can be concluded as follows:

- We propose SDVAE to achieve the first semantically disentangled morphable detail model, which allows accurate reconstruction and semantic-specific manipulation of facial details.
- We present a novel approach to effectively disentangle the latent vectors of identity, expression and age under the framework of VAE in an adversarial manner.
- We apply our model to various applications of detail reconstruction and manipulation, demonstrating its advantages over previous methods.

2 Related Work

Linear Morphable Face Model. [1] adopts PCA to describe the variation in facial geometry and albedo. The learned statistical model can perform 3D reconstruction and controlled manipulation. [7] extends this approach to emotive facial scans by using an additional PCA model, resulting in a model for both identity and expression variation. [2], [8] propose to use High-Order PCA to construct a multilinear model, which models the combined effect of identity and expression variation. [9] presents a method to learn such multilinear models directly from raw scans by jointly performing model parameter optimization and scan registration. [10] devises a method to jointly learn the model and register the scans from 4D sequences, achieving riggable 3D face reconstruction with specific facial expressions. [3] presents a method to automatically segment the faces into areas based on the displacements of vertices. An individual linear model is learned the local information of each segmentation. [11] proposes to employ sparse PCA with a group sparsity term to construct a localized model from the training data. [9] combines local PCA with an anatomical constraint to learn the local deformation subspace. Blendshape [12] is a special parametric model for expression space. [13] presents a method to deform a specific subject given a set of predefined template blendshape models. [14] extends it by adapting the specific subject according to a small number of static face scans in different expressions. Compared with [13], [14] can construct more personalized blendshapes. [15] proposes a method to use a set of linear expression transfer operations to generate personalized blendshapes from neutral shapes. [15], [16] can simultaneously track the 3D face and adapt the blendshapes in real-time. All the above linear morphable models require establishing dense correspondences between the training set of facial meshes. However, facial details are difficult to align and struggle to be represented by linear models.

Nonlinear Morphable Face Model. Compared with widely used linear models, several methods are proposed to model the facial variations nonlinearly. [17] introduces a physical model based on passive flesh, active muscles, and rigid bone structures. Therefore, nonlinear animation effects are enabled with this physical model. With the development of Deep Neural Networks (DNNs), several methods are proposed to utilize the nonlinear capability of DNNs. [4] proposes the first autoencoder that performs graph convolutions on 3D meshes, obtaining more compact representations compared with linear models. [18] uses a VAE to model different levels of details by embedding shape variations in different layers of the network. [19] learns the statistical models of both shape and texture with graph convolution networks, thus can perform colored mesh reconstruction with very light-weight models. [20], [21] corre-

1077-2626 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: MICROSOFT. Downloaded on December 20,2022 at 08:57:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3166666, IEEE Transactions on Visualization and Computer Graphics

late identity and expression semantic attributes with facial geometry using a nonlinear autoencoder. Generative Adversarial Network (GAN) is also explored by the community for 3D face modeling. [22] proposes a method to map 3D face onto a 2D image domain and employ 2D convolutions to learn a GAN of facial texture. However, their facial geometry is still represented by a PCA model. [23] trains a GAN to model the facial geometry and decouple different factors like identity and expression. [24] presents a GAN to model both facial geometry and texture with a focus on detailed texture information. [25] proposes the first intrinsic GAN architecture operating directly on a 3D mesh instead of 2D image domain.

3D Facial Detail Reconstruction. Details are essential for realistic 3D face reconstruction. Although the industrial methods in [26], [27] can obtain high-quality facial details, their multi-view camera systems are too complicated for consumer-level usage. Plenty of methods [28], [29], [30] are proposed to use Shape from Shading (SfS) techniques to reconstruct the facial details without complicated industrial systems. However, the numerical optimization of the SfS technique is computationally complicated. [16] proposes a system to employ a high-end GPU to achieve real-time performance for facial detail tracking. [31], [32], [33], [34] use Deep Neural Networks (DNNs) to recover the facial details from input images. [35], [36] propose to train local regressors from high-resolution capture data and predict the detailed local geometry from large-scale shape or appearance. Although impressive detail reconstruction results have been achieved by deep neural networks, none of the method above can animate the details. [20], [37] are able to synthesize dynamic details, but the details only lie in the texture. [38] learns a synthesis network from a high-quality 3D scan dataset, which can infer facial geometric details under various key expressions given a single image. However, as they assume all the details in the input image do not change with expression, they can synthesize plausible new details but cannot deactivate existing details during animation. [39] proposes a variational autoencoder to synthesize plausible facial details from identity and expression coefficients of a bilinear 3DMM. They focus on synthesizing plausible details, but cannot reconstruct details of a specific person from an input image or video. [40] is the first method that can reconstruct riggable details from a single image. They can reconstruct facial details by forwarding the image through an encode-decoder network. As we will see in Section 5.4.2, their reconstruction cannot fully reflect the shading constraints in the image. On the contrary, our model more accurately reconstruct the details by fitting to those shading constraints. Moreover, our model can integrate a temporal smoothness term during fitting to reconstruct temporal coherent detail animation from an input video, which is nontrivial for an encode-decoder network. Regardless of various methods proposed for image detail reconstruction, our model uniquely serves as an extension of 3DMM, achieves accurate and controllable detail reconstruction, and can be integrated into comprehensive 3DMM-related tasks including image, video and scan fitting and expression and age animation, as shown in Figure 1.

3 Semantically Disentangled Facial Detail Modeling

Our goal is to build a model that can represent wrinkle-level details of various identities, expressions and ages. It is also compatible with an ordinary linear 3DMM [38], which models the large-scale geometry. To achieve this goal, we first extract facial details as a displacement map in UV space and train a conditional VAE. We effectively disentangle the latent factors, including identity, expression and age, in an adversarial manner. During test time, besides fitting the large-scale 3DMM, we also fit our VAE model to match the input constraints for detail reconstruction.

In this part, we first describe the dataset used for network training. Then, we introduce the network architecture and training strategy of SDVAE. Finally, we leverage our model to several practical applications, including scan fitting, image fitting and model manipulation.

3.1 Preprocessing and Detail Extraction

Our method uses a large 3D facial detail dataset with expression and age annotations. Each sample of facial details is represented as a 1-channel displacement map in a shared UV space, where each pixel encodes the displacement along the normal direction. In practice, these displacement maps can be obtained from a 3D scan dataset. A template mesh with predefined UV parameterization is first registered to each 3D facial scan, then the difference between the 3D scan and registered template is baked to deliver the displacement maps. For each sample, the expression annotation is represented as blendshape coefficients, predefined before scanning the user's expression. The age annotation is a single integer between 16 and 68, encoding a performer's age. The recently published FaceScape [38] dataset is a large-scale, high-quality dataset that meets our requirements, which are used as our training dataset.

To enable effective modeling with our semantically disentangled VAE, detail-irrelevant components in the training data should be removed. Besides, the displacement maps should be appropriately aligned to achieve better correspondences. The displacement map data released by FaceScape [38] is obtained by first smoothing a 3D scan via Laplacian filtering [41] and then subtracting smoothed scan from the original scan to obtain details. Since Laplacian filtering often causes mesh shrinking [42], low-frequency components are encoded in the displacement maps as shown in Figure 3, which will diverge the network's modeling ability towards low frequencies. Therefore, we perform Gaussian filtering on the displacement maps, removing low frequencies by subtracting the filtered displacement maps from the original ones. Figure 3 shows that removing those detail-irrelevant components does not affect mesh detail rendering, indicating that we do not lose any details in this preprocessing step. Since FaceScape does not delicately build detail correspondences during capture, the provided displacement maps have slight misalignment across different expressions of the same identity, which will cause artifacts such that identity-specific wrinkles shift with expression changes in section 4.3. Since FaceScape provides each displacement map with a corresponding color texture map, we compute the optical flow field between neutral texture and expression texture

1077-2626 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: MICROSOFT. Downloaded on December 20,2022 at 08:57:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3166666, IEEE Transactions on Visualization and Computer Graphics

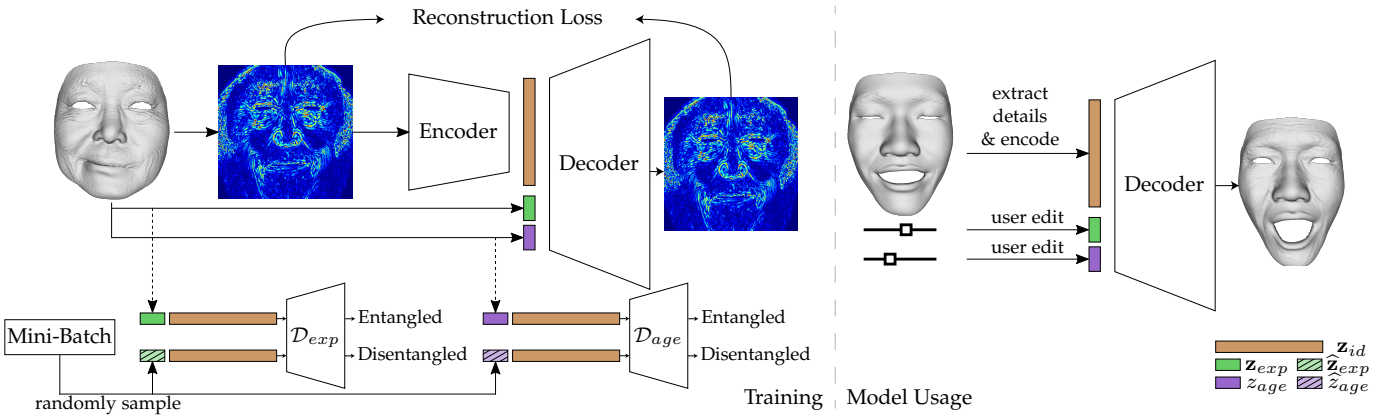

Fig. 2. Overview of our method. We represent facial details as a displacement map and disentangle their representation into identity, expression, and age latent codes. $\widehat{\mathbf{z}}_{exp}$ and $\widehat{z}_{age}$ are expression and age codes from random samples within the mini-batch. They are concatenated with the identity code to generate disentangled distributed data for discriminators. During model usage, we can obtain the identity code by extracting a detail displacement map from the original face and encoding the details, and let a user edit expression and age with a fixed identity.

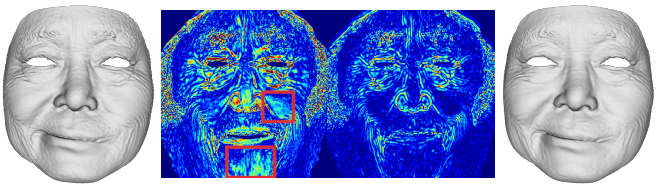

Fig. 3. Low-frequency components are encoded into displacement maps in the FaceScape dataset, shown in column 2 annotated by red boxes. We remove low-frequency components, and the resulting displacement map (column 3) and detailed face (column 4) has the same visual quality of the face using the original displacement map (column 1).

using [43] and warp each expression displacement map to align with its corresponding neutral displacement map using the computed flow field. We observe that the newly aligned displacement maps eliminate artifacts in expression animation and further improve training.

Our model can represent a detailed face together with a large-scale bilinear 3DMM. The large-scale 3DMM shares the same UV parameterization with detail displacement maps. Its expression latent code shares the same semantic information with training displacement maps' expression annotations. Then, we denote the large-scale 3DMM by

$$\mathbf{x}_c = \mathcal{G}_c(\alpha, \mathbf{z}_{exp}) \tag{1}$$

where $\mathbf{x}_c$ is the vertex positions of the generated face, $\mathbf{z}_{exp}$ and α are expression and large-scale identity latent codes, and $\mathcal{G}_c$ is a bilinear mapping that generates faces from latent vectors. We further propose a detail generation module

$$\mathbf{d}_f = \mathcal{G}_f(\mathbf{z}_{id}, \mathbf{z}_{exp}, z_{age}) \tag{2}$$

where $\mathbf{d}_f$ is the per-pixel displacements in UV space, $\mathbf{z}_{id}$ and z_{age} are detail identity and age latent codes, $\mathbf{z}_{exp}$ is the shared expression code consistent with the large-scale expression representation, and $\mathcal{G}_f$ is our novel proposed semantically disentangled VAE that generates facial details given identity, expression and age semantic codes. We share the expression representation between large-scale and details because the same muscles cause dynamic details and large-scale deformation. However, we do not share identity representation, as facial wrinkles contain extra identity information that large-scale shapes cannot represent. Finally, a detailed 3D face is composed via

$$\mathbf{x}_d = sdiv(\mathbf{x}_c) + N(sdiv(\mathbf{x}_c)) \cdot sample(\mathbf{d}_f, sdiv(\mathbf{x}_c)) \tag{3}$$

where $sdiv(\mathbf{x}_c)$ is the subdivided vertices from the large-scale face and $N(sdiv(\mathbf{x}_c))$ is the per-vertex normal directions of subdivided vertices. $sample(\mathbf{d}_f, sdiv(\mathbf{x}_c))$ is the displacements along normal directions sampled from displacement map $\mathbf{d}_f$.

3.2 Network and Training

Here we propose our semantically disentangled decoder $\mathcal{G}_f$ previously mentioned as the detail generation module. It is jointly trained with a variational encoder $\mathcal{E}$, and auxiliary discriminators $\mathcal{D}_{exp}$ and $\mathcal{D}_{age}$ to ensure disentanglement between identity, expression and age, as shown in Figure 2.

To both model facial details and obtain a smooth disentangled latent space capable of editing, we train $\mathcal{G}_f$ and $\mathcal{E}$ with the following objective:

$$L_{total} = L_{recon} + L_{KL} + L_{dis} \tag{4}$$

where L_{recon} is the reconstruction loss, L_{KL} is the KL Divergence, and L_{dis} is the disentanglement loss to ensure semantic separation between identity, expression and age. L_{recon} is given by

$$L_{recon} = E_{\mathbf{x}\sim p(\mathbf{x}), \mathbf{z}_{id}\sim q(\mathbf{z}_{id}|\mathbf{x})} ||\mathcal{G}_f(\mathbf{z}_{id}, \mathbf{z}_{exp}, z_{age}) - \mathbf{x}||_2^2 \tag{5}$$

where $\mathbf{x}$ is the input displacement map, $q(\mathbf{z}_{id}|\mathbf{x}) = \mathcal{N}(\mu, \sigma)$ and $\mu, \sigma = \mathcal{E}(\mathbf{x})$ are Gaussian distribution parameters regressed by the encoder $\mathcal{E}$ to approximate the posterior distribution of identity latent vectors $\mathbf{z}_{id}$. $\mathbf{z}_{exp}$ and z_{age} are not obtained from the encoder $\mathcal{E}$, but directly from the sample $\mathbf{x}$'s annotations available in the dataset. Also, L_{KL} is given by

$$L_{KL} = E_{\mathbf{x}\sim p(\mathbf{x})}[KL(q(\mathbf{z}_{id}|\mathbf{x})||\mathcal{N}(\mathbf{0}, \mathbf{I})] \tag{6}$$

1077-2626 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: MICROSOFT. Downloaded on December 20,2022 at 08:57:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3166666, IEEE Transactions on Visualization and Computer Graphics

to restrict the aggregated posterior to lie in a standard Gaussian distribution to ensure a smooth latent space for interpolation.

The latent representation should both contain sufficient information and divide it into independent factors of variation with known semantic meaning, i.e., identity, expression, and age. First, the latent should reconstruct the input displacement map, which is ensured by objective L_{recon}. Second, it should also make $\mathbf{z}_{id}$, $\mathbf{z}_{exp}$ and z_{age} disentangled and only contain information of their specified semantic labels, therefore becoming independent representations that allow interpretable control. Then we can vary $\mathbf{z}_{exp}$ or z_{age} and keep other semantic factors fixed to generate consistent expression or age editing results of the same identity. We add the L_{dis} that leads to this disentanglement and describe it as follows. Consider the aggregate posterior distribution of the latent $q(\mathbf{z}_{id}, \mathbf{z}_{exp})$ and its marginal distributions $q(\mathbf{z}_{id})$ and $q(\mathbf{z}_{exp})$. If the identity latent is encoded with redundant expression information, as the information of the two factors overlap, the joint distribution $q(\mathbf{z}_{id}, \mathbf{z}_{exp})$ will diverge from the product of marginals $\overline{q}(\mathbf{z}_{id}, \mathbf{z}_{exp}) = q(\mathbf{z}_{id})q(\mathbf{z}_{exp})$, which represents the ideal joint distribution when these two factors are independent. We train a discriminator $\mathcal{D}_{exp}$ to penalize the encoder by training it to tell coupled distributions from independent distributions. It can also be viewed as to minimize the divergence between $q(\mathbf{z}_{id}, \mathbf{z}_{exp})$ and $\overline{q}(\mathbf{z}_{id}, \mathbf{z}_{exp}) = q(\mathbf{z}_{id})q(\mathbf{z}_{exp})$ from a divergence minimization view of GANs [44]. Therefore, $\mathcal{E}$ is trained to predict disentangled identity latent vector, which minimizes the discriminator loss of $\mathcal{D}_{exp}$. Following [45], we randomly permute $\mathbf{z}_{exp}$ within the same batch to generate samples from the hypothetical independent distribution $\overline{q}(\mathbf{z}_{id}, \mathbf{z}_{exp}) = q(\mathbf{z}_{id})q(\mathbf{z}_{exp})$, therefore let the discriminator encourage the marginal distribution of latent to be factorial. Age information is disentangled from identity similarly. The disentanglement loss is formulated as

$$L_{dis} = L_G^{exp} + L_G^{age} \tag{7}$$

where

$$L_G^{exp} = -\log(\mathcal{D}_{exp}(\mathbf{z}_{id}, \mathbf{z}_{exp})) \tag{8}$$

$$L_G^{age} = -\log(\mathcal{D}_{age}(\mathbf{z}_{id}, z_{age})) \tag{9}$$

We train $\mathcal{D}_{exp}$ with the following objective to discriminate coupled and independent distributions

$$\begin{aligned} L_D^{exp} = &-\log(1 - \mathcal{D}_{exp}(\mathbf{z}_{id}, \mathbf{z}_{exp})) \\ &-\log(\mathcal{D}_{exp}(\mathbf{z}_{id}, \widehat{\mathbf{z}}_{exp})) \end{aligned} \tag{10}$$

and similarly train $\mathcal{D}_{age}$ with

$$\begin{aligned} L_D^{age} = &-\log(1 - \mathcal{D}_{age}(\mathbf{z}_{id}, z_{age})) \\ &-\log(\mathcal{D}_{age}(\mathbf{z}_{id}, \widehat{z}_{age})) \end{aligned} \tag{11}$$

where $\widehat{\mathbf{z}}_{exp}$ and $\widehat{z}_{age}$ are expression and age latent codes from randomly permuted samples within the mini-batch during training.

During training, $\mathcal{G}_f$, $\mathcal{E}$, $\mathcal{D}_{exp}$ and $\mathcal{D}_{age}$ are jointly trained via gradient reversal layers, which reverse gradients before $\mathcal{D}_{exp}$ and $\mathcal{D}_{age}$ by a reversal factor $-\lambda$. λ is annealed from 0 to λ_{max} during training to first train discriminators to tell coupled and independent distributions, and then instruct the encoder to generate independent latent distribution via the reversed gradients.

4 Model Fitting and Manipulation

4.1 Scan Fitting

Given a 3D facial scan, we want to obtain its large-scale and detail representation, which enables further manipulation of its expression and age attributes. We optimize both large-scale and detail model parameters α, $\mathbf{z}_{id}$, $\mathbf{z}_{exp}$, z_{age} and rigid pose to minimize the following energy function

$$E_{scan} = E_{lm} + E_{icp} + E_{DR} + E_{reg}^{\alpha} + E_{reg}^{exp} + E_{reg}^{id} \tag{12}$$

where E_{lm} is the 3D distance between vertices and landmarks, E_{icp} is the point-to-plane distance between vertices and scan points and E_{DR} as will be described below is the detail constraint between the extracted and generated displacements. E_{reg}^{α} and E_{reg}^{exp} are regularization terms of large-scale identity and expression codes. First, we initialize α and $\mathbf{z}_{exp}$ as the average identity and neutral expression respectively, and initialize $\mathbf{z}_{id}$ and z_{age} as their empirical means calculated on the training data. To make the energy function tractable, we use a two-step fitting algorithm by alternatively fitting the large scale and details in an iteration. In the large-scale fitting step, we optimize α, $\mathbf{z}_{exp}$ and rigid pose and keep the other parameters fixed by solving the 3DMM given a fixed displacement map. In the detail fitting step, we optimize $\mathbf{z}_{id}$, $\mathbf{z}_{exp}$ and z_{age} and initialize $\mathbf{z}_{exp}$ from the solution in the previous large-scale step. To avoid fitting to low-frequency mesh errors, for E_{DR} we extract high-frequency details as fitting constraints, by filtering and extracting high-frequency displacements from the input scan. Then the displacement values are projected to UV space using the correspondences between scan points and the current mesh obtained in the large-scale step. With the extracted displacements $\mathbf{d}_{extract}$ and the generated displacement map $\mathbf{d}_f$ using current detail parameters, we can formulate E_{DR} as

$$E_{DR} = \|\mathbf{d}_f - \mathbf{d}_{extract}\|_2^2 \tag{13}$$

After the detail step, the solved $\mathbf{z}_{exp}$ is used as initialization in the next large-scale step. We observe convergence after three iterations.

Note that the input displacement map may contain holes or scan noise. However, our detail module can generate complete reconstruction and remove noise via its data-driven prior E_{reg}^{id}. E_{reg}^{id} is formulated as the negative log-likelihood of latent codes in a Gaussian mixture distribution. Inspired by [46], we fit a Gaussian mixture model after training to approximate the prior distribution of latent codes.

4.2 Image and Video Fitting

To reconstruct a detailed 3D face from a single image, we fit a large-scale model to the detected landmarks and a detail model to shading constraints. The key difference from the traditional SfS method is that we replace the Laplacian smoothness term with our data-driven regularization term in latent space. We minimize the following energy objective

$$E_{image} = E_{lm} + E_{sfs} + E_{reg}^{\alpha} + E_{reg}^{exp} + E_{reg}^{id} \tag{14}$$

where E_{lm} is the distance between detected landmarks and projections of corresponding vertices, E_{reg}^{α}, E_{reg}^{exp}. and E_{reg}^{id} are regularization terms same as described in section 4.1,

1077-2626 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: MICROSOFT. Downloaded on December 20,2022 at 08:57:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3166666, IEEE Transactions on Visualization and Computer Graphics

and E_{sfs} minimizes the shading difference between input image and rendered image in gradient space

$$E_{sfs} = \sum_{\substack{u \in \mathbf{x}_d \\ v \in adj(u)}} \|(I_u - I_v) - (R(u) - R(v))\|_2^2 \quad (15)$$

where u and v are adjacent vertices on the generated detailed mesh $\mathbf{x}_d$, I_u and I_v are intensities of corresponding pixels on input image located at u's' and v's projections respectively, R is a rendering function that computes intensities of the rendered image given the vertex normal direction $N(u)$ and environment lighting ℓ_{sh}

$$R(u) = shading(N(u), \ell_{sh}) \quad (16)$$

The environment lighting is represented as the first two order environment spherical harmonics, and we solve ℓ_{sh} by first obtaining a large-scale face reconstruction by only fitting the 3DMM, and then solve ℓ_{sh} to minimize the rendering difference between the input image and rendered image.

Our model can also fit video data to track temporal-coherent detailed face animation. Our large-scale tracking is based on [16], and we feed RGB video frames to our detail fitting method. The fitting algorithm is similar to single image fitting, except that we use the previous frame's result as initialization and apply a temporal smoothness term on model parameters.

4.3 Model Manipulation

Since our reconstruction lies in a parametric model space, we can generate novel facial details by controlling semantic latent codes. First, we obtain the representation in terms of identity, expression, and age code from the reconstructed face. Then, by varying the expression code while fixing the identity and age codes, we can generate detailed 3D faces under novel specified expressions of the same person at the same age. We can also edit the subject's age by varying the age code while fixing identity and expression codes to achieve aging and de-aging effects.

5 Experiments

In this section, we first describe the implementation details. Then we compare the representation power of our method with popular parametric models. We evaluate the effects of our disentanglement loss on expression and age. Finally, we conduct a comparison with recent state-of-the-art method on image reconstruction, and show a broad range of applications of our models on scan, image and video fitting and animations of expression and age.

5.1 Implementation Details

We represent facial details as displacement maps at the resolution of 256x256, which can encode wrinkle-level details well. We use 51 blendshape coefficients to represent expression $\mathbf{z}_{exp} \in \mathbb{R}^{51}$. We normalize the age range from $[16, 68]$ to $[-1, 1]$ and feed it as $z_{age} \in \mathbb{R}$ into network.

Our model is trained using displacement maps from the FaceScape dataset, and we use 10920 samples from 546 identities for training and leave 1180 samples from 59 identities for testing. Both $\mathcal{E}$ and $\mathcal{G}_f$ are parameterized by convolutions followed by Instance Normalization [47] and ELU nonlinearity, with a fully connected layer in the middle to convert feature maps to and from latent vectors. Both $\mathcal{D}_{exp}$ and $\mathcal{D}_{age}$ are parameterized by a 3-layer MLP network. We train the network with Adam optimizer for 1000 epochs with a learning rate of 1e-4. We set KL divergence weight as 0.01 and set $\lambda_{max} = 9$ for the gradient reversal layer used for adversarial training. We set the batch size to 60, which is bounded by the memory of the graphics card. We repeat age and expression latent codes to be of similar length to the identity code before feeding them into discriminators to facilitate adversarial training.

For both scan fitting and image fitting, the energy function is optimized via LBFGS algorithm implemented in PyTorch. When doing scan fitting, we project scan details to displacement map by first extracting details via Laplacian filtering and then removing the low-frequency components by Gaussian filtering, similar to the method in section 3.1.

Our fitting method runs on a standard PC with an Intel Core i7-8700 CPU and an RTX 2080Ti graphics card. Our detail module achieves 0.62s per frame for video tracking. When doing model manipulation, our model can generate a displacement map in 1.43ms.

TABLE 1
Fitting errors of our method and baselines on FaceScape dataset at 128 dimensions.

	Ours	PCA	Bilinear
Training RMSE (mm)	0.0495	0.0517	**0.0480**
Test RMSE (mm)	**0.0498**	0.0539	0.0545
Number of Parameters (M)	**7.732**	8.0625	161.25

5.2 Comparison with Other Parametric Models

We propose to use VAE in facial details modeling rather than PCA and bilinear models, which are widely used for large-scale face modeling. We compare the representation power and robustness of our method with PCA and bilinear models here on the test set of the FaceScape [38] dataset. We first describe the methodologies of the baseline methods. We perform PCA on the training set of displacement maps to construct the PCA baseline. The bilinear baseline model is constructed similar to the method in [2] but on facial details rather than large-scale geometry. It is different from $\mathcal{G}_c$ as it is a separate detail model that can be used on top of $\mathcal{G}_c$. Its expression dimension is 19 since we have one neutral expression and 19 other expressions in the FaceScape dataset.

We test the representation power by autoencoding test displacement maps and evaluating the RMSE between inputs and decoded outputs. We test the reconstruction error at latent dimensions from 4 to 256, and our VAE-based method consistently achieves lower error than both PCA and bilinear model on all the dimensions, s shown in Figure 4 left and Table 1. Please also note that our method costs fewer model parameters. As shown in Figure 5, our method reconstructs more expressive details than the baseline methods, by reconstructing more accurate forehead wrinkles (row 2, row 3) and wrinkles around mouth (row 1, row 3, row 4).

1077-2626 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: MICROSOFT. Downloaded on December 20,2022 at 08:57:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3166666, IEEE Transactions on Visualization and Computer Graphics

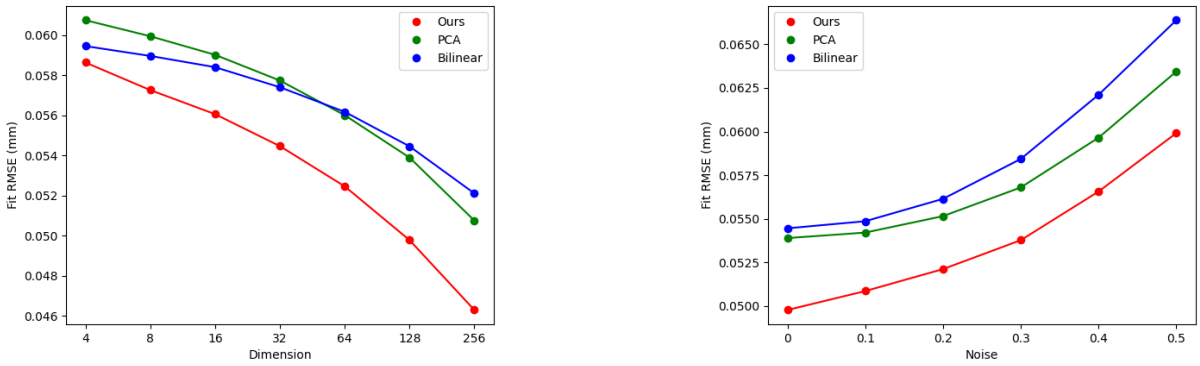

Fig. 4. Fitting errors of our method, PCA and bilinear model on test displacement maps in the FaceScape dataset. We show fitting errors at different dimensions on the left and fitting errors at dimension 128 when different levels of noise are added to input on the right.

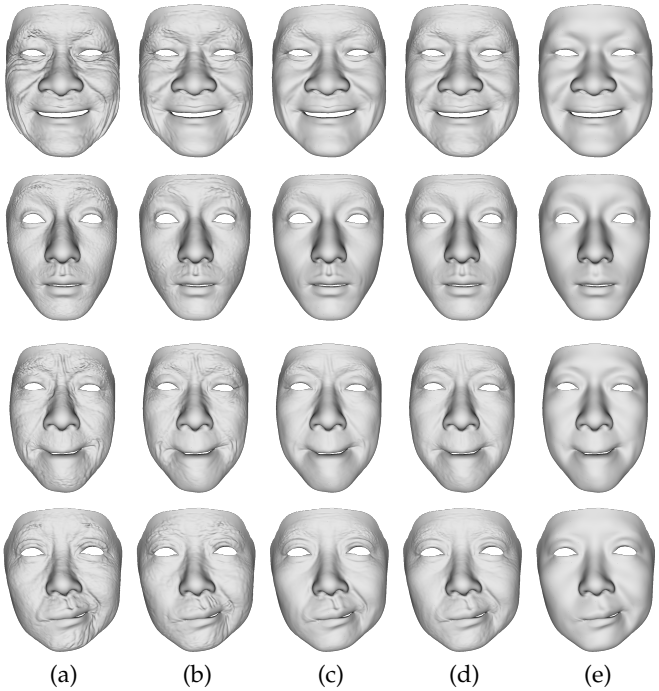

Fig. 5. Fitting results of different detail representation models. From left to right we show the input faces (a), our results (b), results of PCA (c), results of bilinear PCA (d), and large-scale fitting only (e).

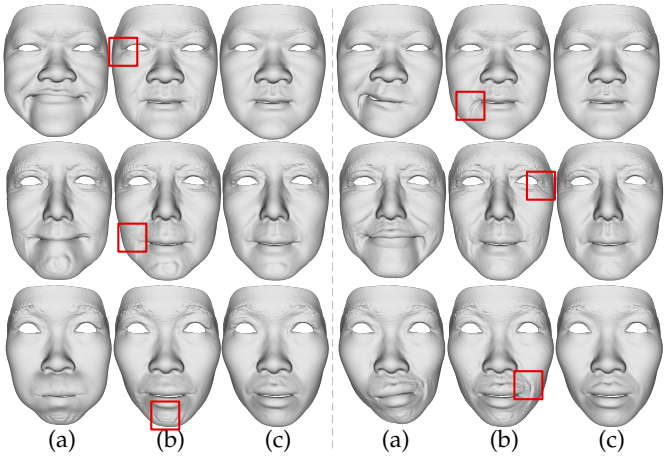

Fig. 6. Evaluation of our expression disentanglement loss. We use our method to convert the facial details in the inputs (a) to neutral expression while keeping the identity unchanged and get the results (b) and (c). The expression disentanglement loss is only used in (c) but not in (b). Artifacts caused by coupled latent space in the baseline are denoted as red boxes.

TABLE 2
Our expression disentanglement metric. We show the disentanglement metric of our model without and with expression disentanglement loss.

	w/o disentangle loss	w/ disentangle loss
Mean std (mm)	0.0221	**0.0159**
Median std (mm)	0.0191	**0.0130**

We also compare the robustness of prior spaces learned from different methods using noisy displacement maps. As each modeling method learns the prior information about facial details from the dataset, it should recover facial details and remove noise. The better the prior space is, the lower the error between reconstruction and noise-free ground truth should be. We add Gaussian noise to the inputs, feed the inputs to each method, and compute fitting errors between reconstructed results and the noise-free ground truth. We quantitatively compare our method with PCA and bilinear model under different noise levels, as shown in Figure 4 right. Our method consistently achieves lower error in the experiment, demonstrating that our model learns a better prior space than linear methods.

5.3 Disentanglement Evaluations

In this section, we conduct an ablation study on the adversarial disentangle loss, which is the key to disentangling different semantic information.

5.3.1 *Expression Evaluation*

To quantitatively evaluate the disentanglement of identity and expression in our model representation, we propose to use the following disentangle metrics following a large-scale mesh method [50]. For displacement maps of the same identity with different expressions in the test dataset, their identity representation should have the same semantic information. Therefore we first encode their identity latent

1077-2626 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: MICROSOFT. Downloaded on December 20,2022 at 08:57:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3166666, IEEE Transactions on Visualization and Computer Graphics

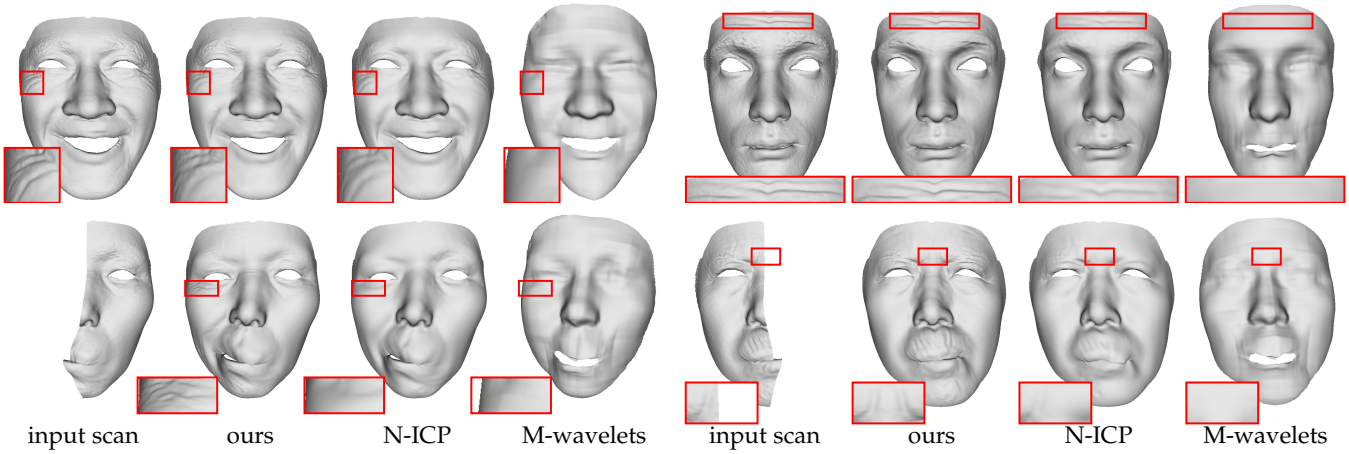

Fig. 7. Scan fitting results. From left to right: the original scans, our fitting results, fitting results from Nonrigid ICP [48] and Multilinear Wavelets [49].

vectors using encoder $\mathcal{E}$, then condition all those identity vectors with neutral blendshape coefficients and generate reconstructions by decoder $\mathcal{G}_f$. We compute the standard deviations of the generated displacement maps and choose the mean and median of the standard deviations of different identities as disentanglement metrics. The lower the metric values are, the more consistent identity codes can be encoded from different expressions, which means the model is more disentangled.

From Table 2, we can see that with the disentangle loss L_{dis}, our method achieves lower mean and median std values, indicating better disentanglement results. The benefit can also be noticed in the qualitative results. Figure 6 shows the visual results of expression disentanglement. We can see that without the disentangle loss, the reconstructed neutral expressions contain some details related to the input expressions (Figure 6(b)), which are not shown in the results with the disentangle loss (Figure 6(c)).

5.3.2 Age Evaluation

Since we do not have ground truth data of the same identity at different ages, we cannot evaluate the same disentanglement metric on identity and age. Therefore, we showcase its effectiveness via qualitative study. A model with better age disentanglement should correctly adjust age-related details to be compatible with the modified age, while keeping identity-related details fixed.

Figure 8 shows the visual results of age disentanglement. We can see that with the disentangle loss, the reconstructed young and old faces have more reasonable age-related details. For the results without the disentangle loss, the age-related details in the red rectangles fail to change with age. Note that the large-scale 3DMM we used cannot be controlled by age, therefore some age-related large-scale shape do not change according to age. We consider this as a limitation of the current large-scale 3DMM, and future work can extend our method by change the age of large-scale and details in the same time. We also show that our age changes are smooth and continuous, as we can produce smooth age interpolation in the accompanying video.

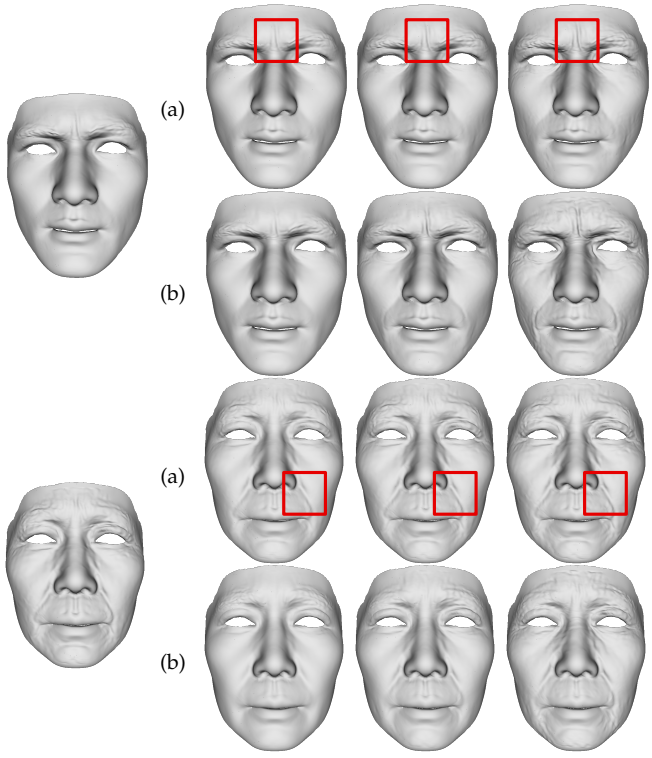

Fig. 8. Evaluation of our age disentanglement loss. We fix the input faces' identity and expression codes and change their age codes. Note that results generated by the model with disentangle loss are more consistent with target ages. Age-related details that are incorrectly fixed in the baseline are shown in red boxes.

5.4 Applications

5.4.1 Scan Fitting

We demonstrate one usage of our model to fit facial scans as shown in Figure 7. Here we do not show preprocessed scans in the FaceScape dataset, but use scans from [51] to showcase the ability to establish vertex correspondences using our fitting method. We also compare with two previous works Nonrigid ICP [48] and Multilinear Wavelets [49], both

1077-2626 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: MICROSOFT. Downloaded on December 20,2022 at 08:57:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3166666, IEEE Transactions on Visualization and Computer Graphics

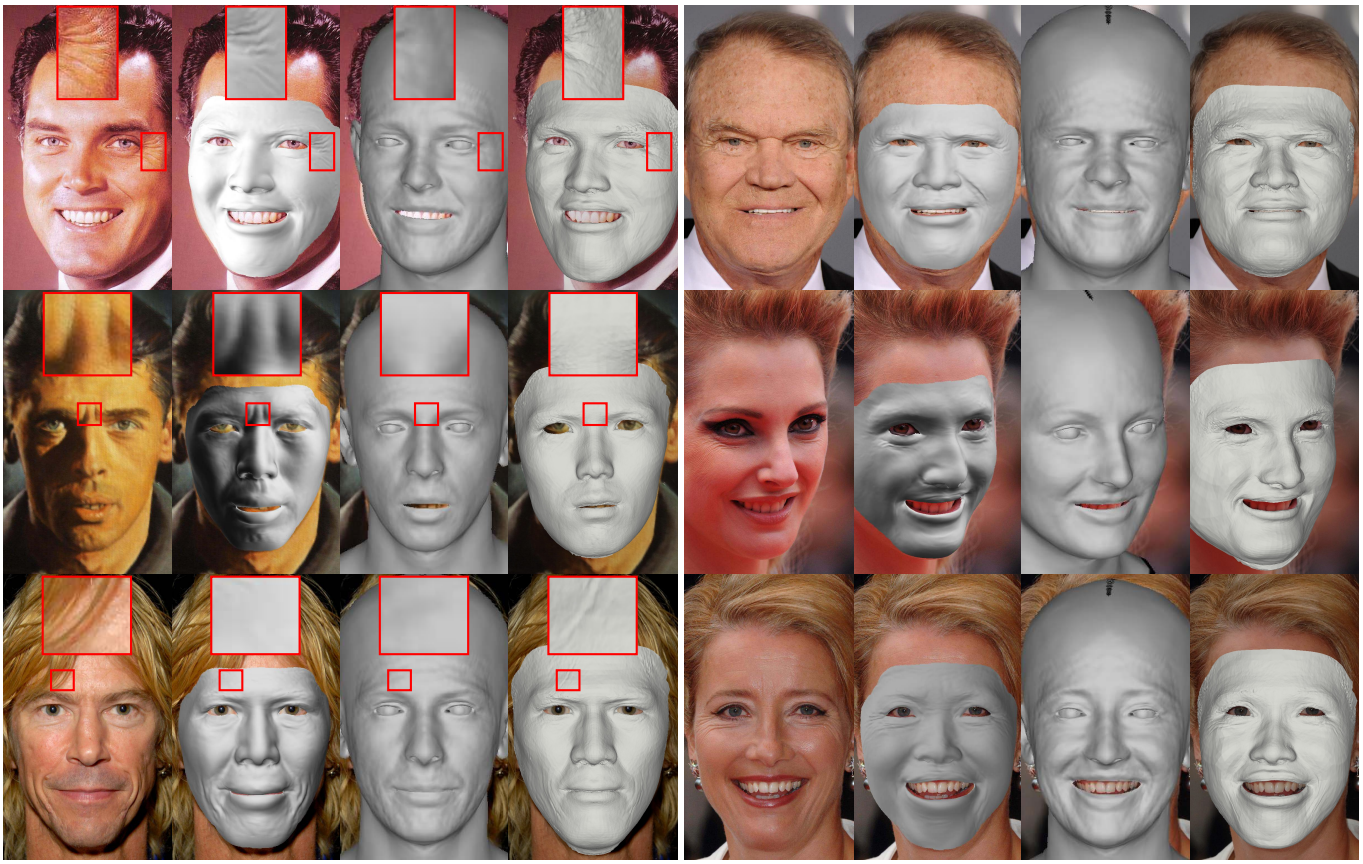

Fig. 9. Comparison on image reconstruction with DECA and FaceScape.

designed specifically for scan fitting. We use their open-source implementation for comparison, and initialize the Nonrigid ICP from the surface fitted by a 3DMM.

Nonrigid ICP is a popular method for facial surface registration. As it does not represent the face by a parametric model, it has high degrees of freedom and can fit the details on the visible surface as expected. However, it cannot edit the resulted face as it doesn't obtain the latent representation of the face. Our method can achieve results on par with Nonrigid ICP in the visible region. We can accurately reconstruct various facial wrinkles like crow's feet and forehead wrinkles, which can greatly enhance the realism compared to a large-scale face. Furthermore, given a challenging facial scan with only half visible constraint (bottom row), our model can also recover full details which accurately represent the observed parts and are plausible on the missing parts. Please notice the reconstructed details in the missing region match the age and expression of the inputs. Multilinear Wavelets is a parametric model-based method, thus can edit the expression of fitted results. However, wrinkle-level details are missing in its results, probably because it lacks the representation ability for facial details.

5.4.2 Image Fitting

Given an RGB face image, we can reconstruct facial details from the image via model fitting using the shading constraints in the image. We compare the reconstruction results with recent riggable detail reconstruction methods, FaceScape [38] and DECA [40]. We use their official implementation for comparison. Figure 9 shows the reconstruction results on in-the-wild images from [52].

We can see that our method has a wide coverage of facial details, and effectively matches the wrinkle shading in the input images. DECA generates plausible details, but they do not match the input image as well as ours, and generate less identity-specific details. We can generate more accurate details than FaceScape (row 1 left, row 2 left), and are more robust than FaceScape in handling occlusions like hair (row 3 left), because we model the possibility of wrinkle occurrence on each position in our model prior, while FaceScape mainly relies on local image patches. From the results, we can see our method both accurately reconstructs wrinkle details and are robust to various in-the-wild scenarios like hard shadows and occlusions, outperforming previous methods. Video tracking is similar to image fitting except we add a temporal smoothness term in the latent space. Figure 10 shows several frames of our tracking results. We also show tracking results of videos in our accompanying video.

5.4.3 Animation

Since our model also uses blendshape coefficients to manipulate facial details with expressions, we can perform face motion transfer using a sequence of blendshape coefficients and generate detailed face animation. The blendshape coefficients are extracted using our tracking system based on [16]. The results are shown in our accompanying video as well as in Figure 11. We can see that besides the static details

1077-2626 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: MICROSOFT. Downloaded on December 20,2022 at 08:57:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3166666, IEEE Transactions on Visualization and Computer Graphics

Fig. 10. Tracking results. For each sample we show the input video frame (a) and our tracking result (b).

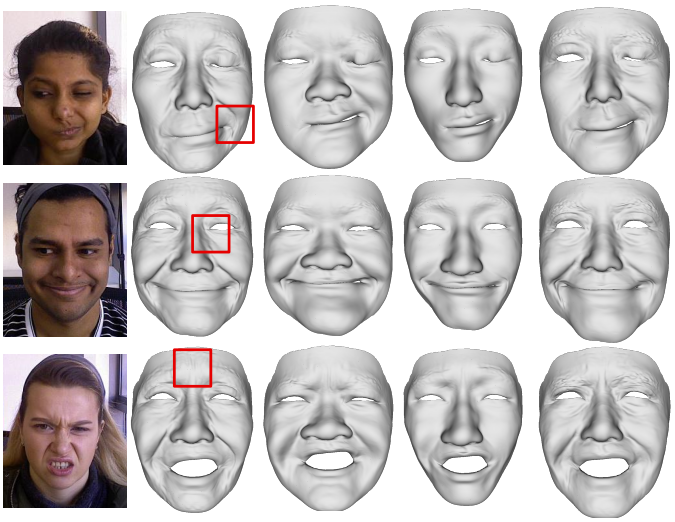

Fig. 11. Frames of blendshape animation of different identities. We show the reference images on the left, and for each column, we show detailed faces of the same identity performing the reference expressions. Regions where identity-specific dynamic wrinkles exist are marked by red rectangles.

in the neutral expressions, the user-specific dynamic details can also be generated by our method.

5.4.4 Age Editing

By varying the age coefficients from young to old, we can produce aging effects under different identities and expressions. Both the original face and the aged one can be driven to produce animations, and they show different dynamic details because of the changed age, which are also shown in the accompanying video and Figure 8.

5.4.5 Extrapolation

Our decoder $\mathcal{G}_f$ takes normalized ages in the range of $[-1, 1]$ and expression coefficients in the range of $[0, 1]$ as input. To explore the model's behavior beyond the predefined range, we visualize its extrapolation behavior in Figure 12. We vary the age value in the range of $[-3, 3]$ and scale the original blendshape coefficients in the range of $[-3, 3]$ respectively. Please note that to focus on details in expression extrapolation, we do not extrapolate large-scale blendshapes but set them to the nearest valid value (neutral or extreme expression), because out-of-range blendshapes will cause artifacts that interfere detail visualization.

We find that the detail model will exaggerate features related to aging or expression when the inputs exceed the predefined largest value. However, when we input values smaller than the lower bound, the results stay roughly the same with the youngest face or the neutral expression. The results finally degrade when we extrapolate far beyond the range.

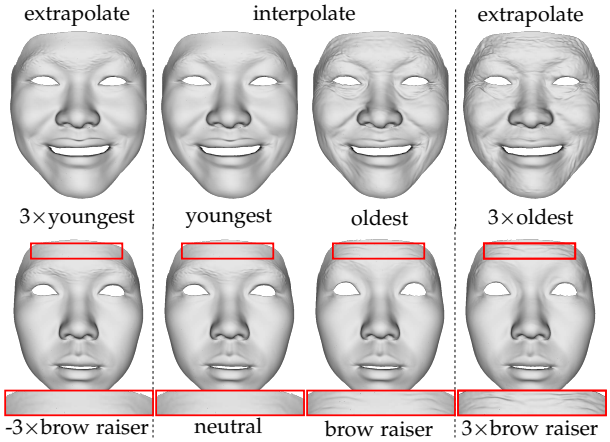

Fig. 12. Extrapolate results of facial details on age editing (the first row) and expression editing (the second row).

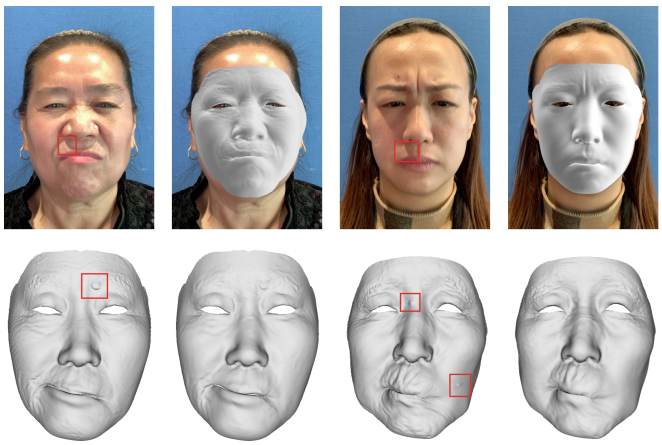

Fig. 13. Limitation of our method. The **top row** shows two image fitting results. For the old performer on the **left**, the shadow on the image leads to a wrong winkle in the result, while for the young performer on the **right**, the shadow does not lead to a winkle. This is due to different prior subspaces for different ages. The **bottom** row shows two scan fitting results. For the **left** result, the mole is correctly reconstructed, while for the **right** result, the two moles are not reconstructed. This is related to whether the training data contains moles on the specific positions.

6 Limitations

As a parametric model, we prefer to use fewer parameters to make the model compact, efficient, and easy to use in fitting,

1077-2626 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: MICROSOFT. Downloaded on December 20,2022 at 08:57:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3166666, IEEE Transactions on Visualization and Computer Graphics

manipulation, and other applications. However, pore-level details are inherent in high dimensions. So in this paper, since we aim to build a compact model of wrinkle-level details with 256 dimensions, we cannot model pore-level details well.

Our fitting robustness is affected by the ages of the performers. For the old performers, the prior subspace learned by our model will lead the results to generate more age-related details. As shown in the top row of Figure 13, the shadow boundary in the image is reconstructed as a wrinkle as older people tend to have wrinkles in this region. On the other hand, for a young performer, our model is robust to the noise in this region as young people tend not to have a wrinkle in this region. This also reflects that our method correctly learns the prior knowledge related to ages.

Our model is data-driven, thus it cannot model rare details. One example of this is mole reconstruction. For some moles which are in the similar positions with some training data, we can reconstruct them. Otherwise, we cannot (shown in the bottom row of Figure 13).

7 Conclusion

We propose a VAE-based model to represent facial details and control them with semantic attributes. Our technique by design disentangles fine-scale facial details from large-scale face shapes, which has the benefit of being compatible with the widely-used large-scale face models. By proposing our Semantically Disentangled Variational Autoencoder (SDVAE), we further disentangle identity, expression, and age for facial details. We propose an adversarial disentangle loss to achieve the aforementioned detail disentanglement. Quantitative and qualitative comparisons demonstrate the better representation power of our method compared to alternative parametric models. We achieve more expressive detail reconstruction compared to state-of-the-art riggable image reconstruction methods. Various applications like image and video fitting, full and partial scan fitting, age manipulation, and expression animation are performed to demonstrate the power of our technique.

Acknowledgments

This work was supported by Beijing Natural Science Foundation (JQ19015), the NSFC (No.62021002, 61727808), the National Key R&D Program of China 2018YFA0704000. This work was supported by THUIBCS, Tsinghua University and BLBCI, Beijing Municipal Education Commission. Feng Xu is the corresponding author.

References

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

[2] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2013.

[3] J. R. Tena, F. De la Torre, and I. Matthews, "Interactive region-based linear 3d face models," in *ACM SIGGRAPH 2011 papers*, 2011, pp. 1–10.

[4] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3d faces using convolutional mesh autoencoders," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 704–720.

[5] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans," *ACM Transactions on Graphics*, vol. 36, no. 6, 2017.

[6] B. Jiang, J. Zhang, J. Cai, and J. Zheng, "Disentangled human body embedding based on deep hierarchical neural network," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 8, pp. 2560–2575, 2020.

[7] B. Amberg, R. Knothe, and T. Vetter, "Expression invariant 3d face recognition with a morphable model," in *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 2008, pp. 1–6.

[8] D. Vlasic, M. Brand, H. Pfister, and J. Popovic, "Face transfer with multilinear models," in *ACM SIGGRAPH 2006 Courses*, 2006, pp. 24–es.

[9] C. Wu, D. Bradley, M. Gross, and T. Beeler, "An anatomically-constrained local deformation model for monocular face capture," *ACM transactions on graphics (TOG)*, vol. 35, no. 4, pp. 1–12, 2016.

[10] T. Bolkart and S. Wuhrer, "3d faces in motion: Fully automatic registration and statistical analysis," *Computer Vision and Image Understanding*, vol. 131, pp. 100–115, 2015.

[11] T. Neumann, K. Varanasi, S. Wenger, M. Wacker, M. Magnor, and C. Theobalt, "Sparse localized deformation components," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, pp. 1–10, 2013.

[12] J. P. Lewis, K. Anjyo, T. Rhee, M. Zhang, F. H. Pighin, and Z. Deng, "Practice and theory of blendshape facial models." *Eurographics (State of the Art Reports)*, vol. 1, no. 8, p. 2, 2014.

[13] R. W. Sumner and J. Popović, "Deformation transfer for triangle meshes," *ACM Transactions on graphics (TOG)*, vol. 23, no. 3, pp. 399–405, 2004.

[14] H. Li, T. Weise, and M. Pauly, "Example-based facial rigging," *Acm transactions on graphics (tog)*, vol. 29, no. 4, pp. 1–6, 2010.

[15] S. Bouaziz, Y. Wang, and M. Pauly, "Online modeling for realtime facial animation," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 4, pp. 1–10, 2013.

[16] Z. Wang, J. Ling, C. Feng, M. Lu, and F. Xu, "Emotion-preserving blendshape update with real-time face tracking," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2020.

[17] A.-E. Ichim, P. Kadleček, L. Kavan, and M. Pauly, "Phace: physics-based face modeling and animation," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–14, 2017.

[18] T. Bagautdinov, C. Wu, J. Saragih, P. Fua, and Y. Sheikh, "Modeling facial geometry using compositional vaes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3877–3886.

[19] Y. Zhou, J. Deng, I. Kotsia, and S. Zafeiriou, "Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1097–1106.

[20] P. Chandran, D. Bradley, M. H. Gross, and T. Beeler, "Semantic deep face models," *2020 International Conference on 3D Vision (3DV)*, pp. 345–354, 2020.

[21] J. Zhang, K. Chen, and J. Zheng, "Facial expression retargeting from human to avatar made easy," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 2, pp. 1274–1287, 2020.

[22] R. Slossberg, G. Shamai, and R. Kimmel, "High quality facial surface and texture synthesis via generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[23] V. F. Abrevaya, A. Boukhayma, S. Wuhrer, and E. Boyer, "A decoupled 3d facial shape model by adversarial training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9419–9428.

[24] G. Shamai, R. Slossberg, and R. Kimmel, "Synthesizing facial photometries and corresponding geometries using generative adversarial networks," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 3s, pp. 1–24, 2019.

[25] S. Cheng, M. Bronstein, Y. Zhou, I. Kotsia, M. Pantic, and S. Zafeiriou, "Meshgan: Non-linear 3d morphable models of faces," *arXiv preprint arXiv:1903.10384*, 2019.

[26] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross, "High-quality single-shot capture of facial geometry," in *ACM SIGGRAPH 2010 papers*, 2010, pp. 1–9.

[27] T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. W. Sumner, and M. Gross, "High-quality passive facial performance capture using anchor frames," in *ACM SIGGRAPH 2011 papers*, 2011, pp. 1–10.

1077-2626 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: MICROSOFT. Downloaded on December 20,2022 at 08:57:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3166666, IEEE Transactions on Visualization and Computer Graphics

[28] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz, "Total moving face reconstruction," in *European conference on computer vision*. Springer, 2014, pp. 796–812.
[29] L. Jiang, J. Zhang, B. Deng, H. Li, and L. Liu, "3d face reconstruction with geometry details from a single image," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 4756–4770, 2018.
[30] A. Tewari, M. Zollhöfer, F. Bernard, P. Garrido, H. Kim, P. Pérez, and C. Theobalt, "High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 357–370, 2020.
[31] Y. Guo, J. Cai, B. Jiang, J. 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*, vol. 41, no. 6, pp. 1294–1307, 2018.
[32] L. Tran, F. Liu, and X. Liu, "Towards high-fidelity nonlinear 3d face morphable model," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1126–1135, 2019.
[33] A. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. G. Medioni, "Extreme 3d face reconstruction: Seeing through occlusions," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3935–3944, 2018.
[34] A. Chen, Z. Chen, G. Zhang, Z. Zhang, K. Mitchell, and J. Yu, "Photo-realistic facial details synthesis from single image," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9428–9438, 2019.
[35] C. Cao, D. Bradley, K. Zhou, and T. Beeler, "Real-time high-fidelity facial performance capture," *ACM Transactions on Graphics (ToG)*, vol. 34, no. 4, pp. 1–9, 2015.
[36] W. Feng, J. Zhang, Y. Zhou, and S. Xin, "Gdr-net: A geometric detail recovering network for 3d scanned objects," *IEEE Transactions on Visualization and Computer Graphics*, 2021.
[37] C. Zheng and F. Xu, "Dtexfusion: Dynamic texture fusion using a consumer rgbd sensor," *IEEE Transactions on Visualization and Computer Graphics*, 2021.
[38] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 601–610.
[39] Q. Deng, L. Ma, A. Jin, H. Bi, B. H. Le, and Z. Deng, "Plausible 3d face wrinkle generation using variational autoencoders," *IEEE Transactions on Visualization & Computer Graphics*, no. 01, pp. 1–1, 2021.
[40] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3d face model from in-the-wild images," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–13, 2021.
[41] O. Sorkine, "Laplacian Mesh Processing," in *Eurographics 2005 - State of the Art Reports*, Y. Chrysanthou and M. Magnor, Eds. The Eurographics Association, 2005.
[42] G. Taubin, "Curve and surface smoothing without shrinkage," in *Proceedings of IEEE International Conference on Computer Vision*, 1995, pp. 852–857.
[43] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. dissertation, USA, 2009, aAI0822221.
[44] S. Nowozin, B. Cseke, and R. Tomioka, "f-gan: Training generative neural samplers using variational divergence minimization," *Advances in neural information processing systems*, vol. 29, 2016.
[45] H. Kim and A. Mnih, "Disentangling by factorising," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2649–2658.
[46] P. Ghosh, M. S. M. Sajjadi, A. Vergari, M. Black, and B. Schölkopf, "From variational to deterministic autoencoders," 2020.
[47] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2017.
[48] B. Amberg, S. Romdhani, and T. Vetter, "Optimal step nonrigid icp algorithms for surface registration," in *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 2007, pp. 1–8.
[49] A. Brunton, T. Bolkart, and S. Wuhrer, "Multilinear wavelets: A statistical shape space for human faces," in *European Conference on Computer Vision*. Springer, 2014, pp. 297–312.
[50] Z.-H. Jiang, Q. Wu, K. Chen, and J. Zhang, "Disentangled representation learning for 3d face shape," 2019.
[51] Z. Wang, X. Yu, M. Lu, Q. Wang, C. Qian, and F. Xu, "Single image portrait relighting via explicit multiple reflectance channel modeling," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–13, 2020.
[52] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Jingwang Ling is currently working toward the Ph.D. degree in Software Engineering at School of software and BNRist, Tsinghua University. He received his B.S degree in Software Engineering from Tsinghua University in 2020. His research areas and interests include face reconstruction and animation.

Zhibo Wang is currently working toward the Ph.D. degree in Software Engineering at School of software and BNRist, Tsinghua University. He received his B.S degree in Microelectronics Science and Technology from Nanjing University, Nanjing, China in 2017. His research interests include facial animation and face reconstruction.

Ming Lu received the Ph.D. degree in Information and Communication Engineering from Tsinghua University, Beijing, China, in 2019. He is currently a researcher at Intel Labs China. His research interests include computer vision and computer graphics. He is particularly interested in classification and detection, 3d face and body, image restoration and synthesis.

Quan Wang received the B.S. degree in electronic engineering from Tsinghua University, China. He is currently a research scientist at Sensetime. His research interests include computer vision and 3D reconstruction.

Chen Qian, currently the Executive Research Director of SenseTime, is responsible for leading the team in AI content generation, digital human and end-edge computing research. He received MPhil degree in information engineering from The Chinese University of Hong Kong. His research interest include AI content generation, face body animation and deep learning for end-edge computing.

Feng Xu is currently an associate professor in the School of Software at Tsinghua University. He received a B.S. degree in physics from Tsinghua University, Beijing, China in 2007, and a Ph.D. degree in automation from Tsinghua University, Beijing, China in 2012. His research interests include facial animation, performance capture, and 3D reconstruction.

1077-2626 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: MICROSOFT. Downloaded on December 20,2022 at 08:57:28 UTC from IEEE Xplore. Restrictions apply.