Learning to Aggregate and Personalize 3D Face from In-the-Wild Photo Collection

Zhenyu Zhang[1,2], Yanhao Ge[1], Renwang Chen[1], Ying Tai[1], Yan Yan[2], Jian Yang[2],
Chengjie Wang[1], Jilin Li[1], Feiyue Huang[1]
Tencent Youtu Lab, Shanghai, China
Nanjing University of Science and Technology, Nanjing, China
zhangjesse@foxmail.com yanyan, csjyang@njust.edu.cn
halege, renwangchen, yingtai, jasoncjwang, jerolinli, garyhuang@tencent.com

Abstract

Non-parametric face modeling aims to reconstruct 3D face only from images without shape assumptions. While plausible facial details are predicted, the models tend to over-depend on local color appearance and suffer from ambiguous noise. To address such problem, this paper presents a novel Learning to Aggregate and Personalize (LAP) framework for unsupervised robust 3D face modeling. Instead of using controlled environment, the proposed method implicitly disentangles ID-consistent and scene-specific face from unconstrained photo set. Specifically, to learn ID-consistent face, LAP adaptively aggregates intrinsic face factors of an identity based on a novel curriculum learning approach with relaxed consistency loss. To adapt the face for a personalized scene, we propose a novel attribute-refining network to modify ID-consistent face with target attribute and details. Based on the proposed method, we make unsupervised 3D face modeling benefit from meaningful image facial structure and possibly higher resolutions. Extensive experiments on benchmarks show LAP recovers superior or competitive face shape and texture, compared with state-of-the-art (SOTA) methods with or without prior and supervision.

1. Introduction

Monocular 3D reconstruction of human face is a long-standing problem with potential applications including animation, biometrics and human digitalization. It is an essentially ill-posed problem requiring strong assumption, e.g., shape-from-shading approaches [67]. With 3D Morphable Model (3DMM) [4] proposed, the reconstruction can be achieved through optimization on low-dimensional parameters [38, 37, 73]. Recently, deep neural networks are introduced to regress 3DMM parameters from 2D images with supervision [71, 35, 9, 12, 28] or improve 3DMM with non-linearity [49, 36, 51, 14, 48, 70]. Meanwhile, as these single-view guided methods may suffer from 2D ambiguity, other 3DMM-based works are proposed to leverage multi-view consistency [53, 47, 62, 57, 2]. While 3DMM provides reliable priors for 3D face modeling, it also brings potential drawbacks: as built from a small amount of subjects (e.g., BFM [33] with 200 subjects) and rigidly controlled conditions, models may be fragile to large variations of identity [72], and have limitations on building teeth, skin details or anatomic grounded muscles [10].

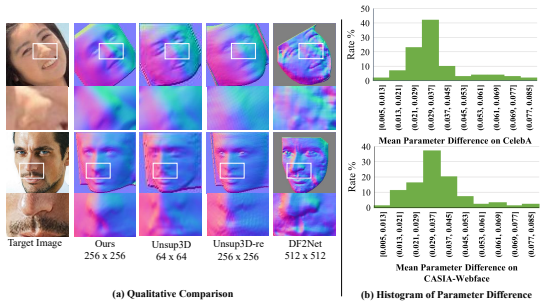


Figure 1. (a) Qualitative comparison between our method and Unsup3D [59] and DF2Net [65]. Our results show better shape of organs with finer details and less noise. (b) Distribution of parameter difference in expression basis [5] between all single-ID image pairs on face dataset [30, 61]. With the manually computed threshold of 0.04, it reveals that quite amounts (about 70%) of image pairs have similar expressions and non-rigid difference. Such conclusion inspires us to approximate expressions by mean conditions with reasonable relaxations, and learn a multi-image consistent face without 3DMM prior. Better showed by zooming in.

Due to the aforementioned limitations, an alternative approach is learning to model 3D face without 3DMM assumption, e.g., regressing face normal or depth directly from an input image [18, 52, 42, 65, 1] with ground truth scans and pseudo labels. Despite the efficiency of these approaches, they cannot model facial texture or a canon-

ical view without occlusion. Recent work Unsup3D [59] uses a weakly symmetric constraint to disentangle a face into intrinsic factors and accomplishes the canonical reconstruction in unsupervised manner. In summary, these non-parametric methods predict plausible facial structure via reconstruction or rendering loss [22]. However, without reliable 3DMM prior, they tend to suffer from ambiguity of image appearance. As illustrated in Fig. 1(a), results of Unsup3D [59] and DF2Net [65] have coarse or inconsistent shape of organs. Further, Unsup3D suffers from noise and discontinuity when reproduced for higher-scale reconstruction (Unsup3D-re), which makes the resolution less valuable. Such phenomenon is due to improper disentangling of albedo, illumination and geometry due to ambiguity of image details and noises as discussed in [50, 24, 10]. On top of these, we argue that *predicting meaningful and consistent facial structure* is a key point of unsupervised non-parametric 3D face modeling.

To achieve such goal, a better disentangling approach could be: first modeling *basic facial geometry and texture* of an identity, then adding *specific attributes and details* for a target scene. Actually, basic facial structure is mainly based on bones and anatomic grounded muscles of an identity, which can be enhanced by using 3DMM across unconstrained image set against ambiguous noise [40, 8, 13]. However, multi-image clues are difficult to introduce for non-parametric methods due to lack of shape topology. To tackle this problem, we make two assumptions for image collections: i) Besides the shape, the appearance of an identity due to basic facial structure, like winkles and occlusion of illumination, are similar enough; ii) Non-rigid shape deformation (mainly about expression) among faces are with limited extent. The first assumption has been demonstrated by works of [40, 8, 13]. For the second one, we compare the expression difference of all image pairs of photo sets in datasets. By using released SOTA 3DMM based model [8], we analyse the distribution of mean parameter difference of image pairs on expression basis [5] in Fig. 1(b). With the computed similarity threshold 0.04 from manually selected 1k separate image pairs with similar/dissimilar expression, we observe that about 70% pairs are below the threshold with mild non-rigid difference. Such conclusion makes it possible to approximate expressions by mean conditions with reasonable relaxations, and learn a multi-image consistent face without 3DMM prior.

In this paper, we propose a novel Learning to Aggregate and Personalize (LAP) framework for unsupervised non-parametric 3D face modeling. LAP first aggregates consistent face factors of an identity from in-the-wild photo collection, and then personalizes such factors to reconstruct a scene-specific face for a target image of the same ID. Concretely, LAP decodes a pair of ID-consistent albedo and depth by adaptively aggregating a global ID code from an image set, and reconstructs a 3D face aligned to each input image based on an estimated specific light and pose. Such aggregation model is optimized by a curriculum learning method with relaxed consistency loss, which helps to overcome large facial variations and lack of pre-defined topology. Moreover, to personalize a specific face, LAP modifies ID-consistent face through an attribute-refining network for modeling specific attributes and details. In this way, LAP achieves disentangling of ID-consistent facial structure and scene-specific local details in an unsupervised manner without 3DMM shape assumption. With LAP framework, we manage to model 3D face from arbitrary number of images, or even single image in superior quality and higher resolution than State-of-the-Art (SOTA) methods.

	MI-Consistency	Shape Assumption	Supervision
[71, 35, 9, 12, 72]	×	3DMM	3DMM paramter
[49, 36, 14, 51, 50, 48, 27]	×	3DMM	I
[47, 57, 2, 43]	Constrained	3DMM	I
[40, 8, 13]	In-the-Wild	3DMM	I
[18, 52, 1, 65]	×	No	3DMM parameter, 3D scan, I
[42, 39, 59]	×	No	I
Ours	In-the-Wild	No	I

Table 1. Comparison with selected existing method on different settings. Constrained/In-the-wild means the condition of image set, and **I** means image.

In summary, this paper has contributions in followings:

i) We propose a novel Learning to Aggregate and Personalize (LAP) framework to disentangle ID-consistent and scene-specific 3D face from multi or single image, without 3DMM assumptions in fully unsupervised manner.

ii) With a novel relaxed curriculum aggregation method, LAP is able to predict ID-consistent face factors against large facial variations of in-the-wild photo set.

iii) Based on the ID-consistent factors, LAP uses an attribute-refining network to model scene-specific 3D face with less noise and finer details of higher resolutions.

2. Related Works

In order to assess our contribution and illustrate contrast between LAP and existing methods, we make a comparison in Table 1. As illustrated, our method faces a more challenging setting, leveraging multi-image consistency from in-the-wild photo set without shape assumption or GT.

Parametric Method: With 3DMM [4] proposed, 3D face modeling can be formulated in a procedure of parametric optimization [38, 37, 73]. Recently, deep neural networks are introduced to regress 3DMM parameter from input image [71, 35, 9, 12, 72] by learning from generated ground truth. With neural rendering approach such as [22], methods are proposed to leverage image reconstruction loss to train the model in weakly or un-supervised manner [49, 36, 14], or improve 3DMM with more nonlinear feasibility [51, 50, 48, 27, 6]. Besides single-view method, multi-view based approaches [47, 62, 57, 2, 43] are proposed to model 3D face more robustly. While these meth-

ods are based on constrained conditions or video sequence, they may suffer from limitations for applications. To tackle this problem, methods [40, 8, 13] are proposed to use in-the-wild photo collection to improve the robustness of predicted facial shape. Although these approaches have similar motivation to LAP, they are developed based on 3DMM assumption. In contrast, LAP is proposed without such pre-defined topology, thus faces a more challenging problem.

Non-Parametric Method: As an alternative direction, 3D face can also be modeled without 3DMM, e.g., using Shape-from-Shading (SFS) method [67]. Recently, Sengupta *et al*. propose SFS-Net [42] to predict intrinsic factors from input images for modeling 3D faces. With the success of deep neural networks, data-driven methods [52, 1, 18, 65] are proposed to directly predict face geometry supervised by real and synthetic ground truth. Despite the efficiency of these approaches, they cannot model 3D geometry of full head or facial textures. A more recent work Unsup3d [59] uses weakly symmetric facial constrains to predict light, pose and albedo/depth of canonical view from facial image. Without 3DMM topology, these above non-parametric methods may suffer from ambiguity of image appearance, and predict facial geometry with incorrect details and noise. In contrast, by disentangling a face into ID-consistent and scene-specific factors, LAP models 3D face against such ambiguity and with finer details and structure.

Feature Disentangling in Face Reconstruction: With 3DMM assumptions, methods [20, 58] can disentangle faces into shapes, expressions and textures. For non-parametric methods, SFS-Net [42] and Unsup3d [59] decompose a face into albedo, light, pose and normal. Besides intrinsic decomposition, Deformation AutoEncoder (DAE) [46, 39] disentangles a face into appearance and deformation. Based on such framework, Xing *et al*. [60] propose a probabilistic method to improve the deformable geometry generation, and Li *et al*. [26] leverage videos to urge a better facial action unit. Different from these methods, LAP disentangles a face into global facial structure and scene-specific facial attribute without 3DMM prior from unconstrained photo collection.

3. Preliminary

To predict 3D faces without 3DMM assumption, we build our framework based on Unsup3D [59]. Given a face image $\mathbf{I}$, the framework disentangles it into four factors (d, a, ω, l) comprising a canonical depth map $d \in \mathbb{R}_+$, a canonical albedo image $a \in \mathbb{R}^3$, a global light direction $l \in \mathbb{S}^2$ and a viewpoint $\omega \in \mathbb{R}^6$. Each factor is predicted by a separate network which we denote as $\Phi^d, \Phi^a, \Phi^\omega, \Phi^l$. With these factors, the image $\mathbf{I}$ is reconstructed by lighting Λ and reprojection Π as follows:

$$\hat{\mathbf{I}} = \Pi(\Lambda(a, d, l), d, \omega). \tag{1}$$

Learning uses a reconstruction loss which encourages $\mathbf{I} \approx \hat{\mathbf{I}}$ with a differentiable renderer [22]. To constrain a canonical view of d and a to represent a full frontal face, the framework uses a weakly symmetric assumption by horizontally flipping:

$$\hat{\mathbf{I}}' = \Pi(\Lambda(a', d', l), d', \omega), \tag{2}$$

where a' and d' are the flipped version of a, d, and encourages $\mathbf{I} \approx \hat{\mathbf{I}}'$. To allow probably asymmetric facial region, the framework predicts confidence maps $\sigma, \sigma' \in \mathbb{R}_+$ by Φ^σ and calibrates the loss as follows:

$$\mathcal{L}(\hat{\mathbf{I}}, \mathbf{I}, \sigma) = -\frac{1}{|\Omega|} \sum \ln \frac{1}{\sqrt{2}\sigma} \exp -\frac{\sqrt{2}|\hat{\mathbf{I}} - \mathbf{I}|}{\sigma}, \tag{3}$$

where Ω is normalization factor. The flipped version $\mathcal{L}(\hat{\mathbf{I}}', \mathbf{I}, \sigma')$ is also calculated. In this way, 3D faces are modeled from images in unsupervised manner without 3DMM assumption. Note that, as Unsup3D extremely depends on single-image appearance, it cannot handle 2D ambiguity such as salient local color difference and noise. In contrast, LAP tackles this problem by further disentangling the face, which will be discussed in the following.

4. Methodology

In this section, we mainly describe the proposed Learning to Aggregate and Personalize (LAP) 3D face method. With a photo collection of a same identity, our aim is to accomplish a further disentangling: first modeling basic facial geometry/texture based on consistent facial structure, and then modifying it to personalized attributes and details. As illustrated in Fig. 2, such disentangling is achieved by two steps: Learning to Aggregate (in Sec. 4.1) and Learning to Personalize (in Sec. 4.2), without 3DMM priors.

4.1. Learning to Aggregate

As discussed in Sec. 1, appearance of an identity due to basic facial structure should be consistent across different images, and image collections contain limited non-rigid variation. Inspired by these facts, we propose depth/albedo aggregation network to adaptively aggregate facial factors from a photo collection and learn ID-consistent geometry/texture, and use such consistent factors to reconstruct each input image. We also propose a curriculum learning approach with relaxed consistency loss to suppress large facial variations for stable learning.

Aggregation Network: As illustrated in Fig. 2, the aggregation network has a shared encoder δ across multiple images and a global decoder ϕ for predicting consistent face. For modeling albedo and depth, we use two separate aggregation networks denoted as $\Phi^a = (\delta^a, \phi^a)$ and $\Phi^d = (\delta^d, \phi^d)$. Given a photo collection of N images $\{\mathbf{I}_i^k\}_{i=1}^N$ where k is the index of identity (omitted in the following for simplification), we feed each $\mathbf{I}_i$ into δ^a, δ^d

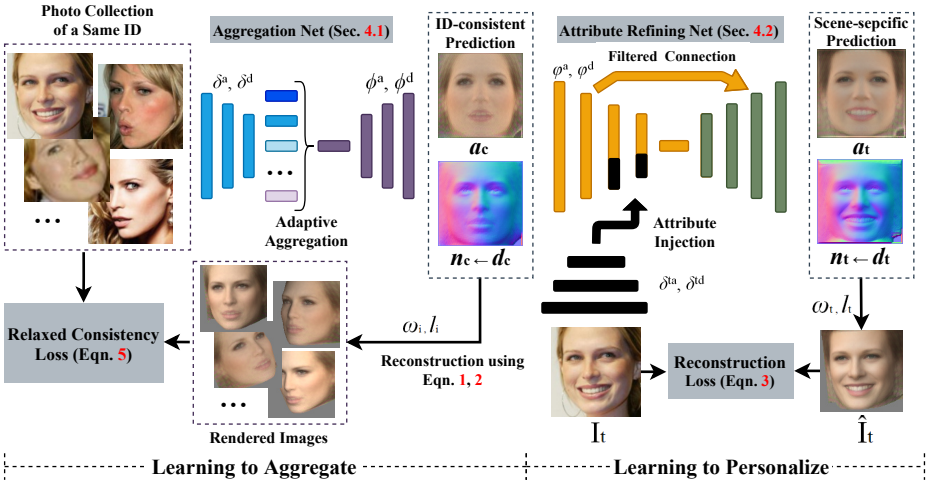


Figure 2. Overview of the proposed framework. Training is first conducted on Learning to Aggregate (Sec. 4.1) for modeling ID-consistent face, and then on Learning to Personalize (Sec. 4.2) for predicting scene-specific face aligned to the target image. δ^a, δ^d and ϕ^a, ϕ^d are the encoder and decoder of albedo/depth aggregation network. δ^{ta}, δ^{td} and φ^a, φ^d are the encoder of attribute injection module and attribute-refining network. The flipped operation and networks $\Phi^\omega, \Phi^l, \Phi^\sigma$ for predicting pose, lighting and confidence are omitted.

to get texture and geometry latent code $\mathbf{x}_i^a, \mathbf{x}_i^d$. Different from the encoder in [59], to get multi-level information, we downsample the feature of each scale through a convolutional layer with average pooling to a vector, and fuse the vectors through concatenation and a convolutional layer to get $\mathbf{x}_i^a, \mathbf{x}_i^d \in [1, 1, c]$. To learn a global representation of the identity based on $\{\mathbf{x}_i^a, \mathbf{x}_i^d\}_{i=1}^N$, inspired by [29, 45], we propose an adaptive aggregation method. Due to the quality of $\{\mathbf{I}_i\}_{i=1}^N$, the importance of each dimension in $\mathbf{x}_i^a, \mathbf{x}_i^d$ which reveals how correlated it is to the identity, is supposed to be different. Hence, we first learn channel-wise weights $\mathbf{w}_i^a, \mathbf{w}_i^d \in [1, 1, c]$ to represent the importance, and use softmax function to normalize them into $\{\bar{\mathbf{w}}_i^a\}_{i=1}^N, \{\bar{\mathbf{w}}_i^d\}_{i=1}^N$. Then the aggregation can be formulated as:

$$\mathbf{x}_c^a = \sum_{i=1}^{N} \bar{\mathbf{w}}_i^a \cdot \mathbf{x}_i^a, \quad \mathbf{x}_c^d = \sum_{i=1}^{N} \bar{\mathbf{w}}_i^d \cdot \mathbf{x}_i^d, \tag{4}$$

where $\mathbf{x}_c^a, \mathbf{x}_c^d$ are the combined global ID-code for texture and depth. Compared with naive average fusion, such adaptive aggregation method encourages the ID-correlated features to get larger weights in $\bar{\mathbf{w}}_i^a, \bar{\mathbf{w}}_i^d$, thus the fused code $\mathbf{x}_c^a, \mathbf{x}_c^d$ can better represent the consistent features of the identity (see Fig. 5). Next, we feed $\mathbf{x}_c^a, \mathbf{x}_c^d$ to the decoder ϕ^a, ϕ^d to get ID-consistent albedo a_c and depth d_c. With ω_i, l_i predicted by $\Phi^\omega(\mathbf{I}_i), \Phi^l(\mathbf{I}_i)$ from each input image, we can reconstruct rendered image $\hat{\mathbf{I}}_i, \hat{\mathbf{I}}_i'$ using Eqns. 1, 2. Then multi-image consistency is achieved by calculating Eqn. 3 between rendered and original images, which enhances ID-consistent facial structure in a_c, d_c and suppresses possibly ambiguous noise in each input image.

Curriculum Learning: As illustrated in Fig. 2, in-the-wild photo collection has different conditions on expression, make-ups, skins and noise, thus directly using such

Figure 3. Easier samples generated by Interface-GAN [44] pre-trained on FFHQ dataset [21].

photo collection for training urges corrupt a_c, d_c or even totally fails to find correspondence (see Fig. 5) without 3DMM pre-defined topology. This motivates us to perform a curriculum learning procedure [3, 17], i.e., training from easier samples to in-the-wild collections. A simple way is to use facial videos such as voxceleb [32] in constrained condition, but it may also have drawbacks: the quality of videos and various pose variations cannot be guaranteed. In contrast, we use a controllable GAN [44] to generate photo collections. As illustrated in Fig. 3, the generated samples have different poses and consistent facial structure with high quality and mild variation. To keep the ID consistency of generated images, we use Arcface [7] to filter out samples with cosine-similarity lower than 0.6 compared with the frontal face image. Note that, we do not expect the samples to have exactly same identity, but only similar facial structure which is sufficient enough for model to learn 3D correspondence. We generate images of 15 different poses from 30k different identities, with resolution of 1024×1024 to benefit a better learning.

Relaxed Consistency Loss: To further urge a stable training procedure, we propose a Relaxed Consistency Loss (RCL) to relax the most uncertain facial region in different conditions. We first train a BiSeNet [63] on CelebAMask-HQ dataset [25] for face parsing, and then define the visible facial region and background with constant 1 and 0 to get

an attention mask $\mathbf{M}$. As mouth, eyes and brow are more probably inconsistent compared to other regions across different conditions, we set these parts in $\mathbf{M}$ with lower value (e.g., 0.3) to get a new attention mask $\mathbf{M}_{re}$. Then the RCL can be formulated as:

$$\mathcal{L}_{RCL}(\hat{\mathbf{I}}_i, \mathbf{I}_i, \sigma_i) = -\frac{1}{|\mathbf{M}_{re}|} \sum \ln \frac{1}{\sqrt{2}\sigma} \exp -\frac{\sqrt{2}|\mathbf{M}_{re} \cdot (\hat{\mathbf{I}}_i - \mathbf{I}_i)|}{\sigma_i}. \tag{5}$$

Note that, although the confidence σ, σ' model the uncertainty to some extent, our RCL provides a certain and stronger constraint. Furthermore, for the extremely hard sample such as image with low quality, large occlusion and extreme lighting, the parsing model tends to predict a corrupt facial region which is much smaller than background, which helps us to naturally filter out such samples for stable learning. In this way, the aggregation network can learn valuable consistent feature against inconsistency.

4.2. Learning to Personalize

As illustrated in Fig. 2, while the learned ID-consistent albedo and depth have basic facial structure, they lack details and attribute (e.g., teeth and expression) aligned to a target image $\mathbf{I}_t \in \{\mathbf{I}_i\}_{i=1}^N$. Hence, we propose attribute-refining network to modify (a_c, d_c) to scene-specific (a_t, d_t), which is achieved via attribute injection and filtered connection.

Attribute Injection: Attribute injection module has encoder δ^{ta}, δ^{td} to extract albedo/depth attribute information from target image $\mathbf{I}_t$, and uses such information to guide the modifying. Denote the encoder of attribute refining network as φ^a, φ^d for encoding a_c, d_c respectively, a direct embedding approach is to fuse features of δ^{ta}, φ^a and δ^{td}, φ^d. However, such method brings two problems: Embedding too many low-level features of δ^{ta}, δ^{td} which are spatially aligned to $\mathbf{I}_t$ and rendered output $\hat{\mathbf{I}}_t$, urges φ^a, φ^d to ignore meaningful canonical texture/geometry information and degrades to trivial pure texture auto-encoder; Embedding too less feature (e.g., only the highest-level feature vector) loses necessary details of predictions. On top of these, we propose a balancing approach with selecting mechanism. Firstly, we only inject features of last three levels (i.e., features with height/width of $8 \times 8, 4 \times 4$ and 1×1) of δ^{ta}, δ^{td}; then, we learn a channel-wise weight for each level of feature to select valuable information, and fuse the re-weighted feature with corresponding one in φ^a, φ^d. In this way, we inject moderate attribute information from $\mathbf{I}_t$ to guide a finer prediction (see Fig. 6).

Filtered Connection: To modify a_c, d_c to a_t, d_t, the network needs to manipulate facial regions (e.g., mouth, eyes and cheek) which are different to target image, meanwhile suitably keep the structure of unchanged parts. Inspired by works of face editing [34, 66], we propose a Filtered Connection module to achieve such goal. As illustrated in Fig. 4, we firstly combine the features of encoder (yellow block) and decoder (green block) to learn a spatial attention mask $\mathbf{A} \in [h, w, 1]$ through convolution and sigmoid function, and then filter the feature by multiplying $\mathbf{A}$ and concatenate it to decoder. In this way, the changed regions of a_c, d_c are suppressed by lower weight in $\mathbf{A}$, while information of unchanged parts is propagated to the decoder. After getting a_t, d_t, we use ω_t, l_t predicted from I_t to reconstruct the output $\hat{\mathbf{I}}_t$. The learning to personalize framework is then optimized by using Eqn. 3 and its flipped version.

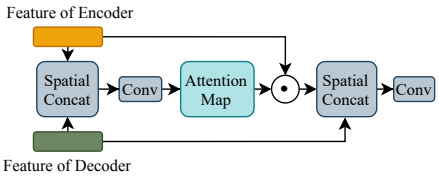


Figure 4. Illustration of the proposed Filtered Connection in attribute refining network.

4.3. Training and Inference

Training mainly contains three steps: firstly train $\Phi^\omega, \Phi^l, \Phi^\sigma$ and the aggregation network $(\delta^a, \phi^a), (\delta^d, \phi^d)$ using Eqn. 5; then freeze them and train attribute-refining network using Eqn. 3; finally jointly fine-tune all the networks for finer predictions. During each training stage, we select image set of a same identity with random size (from 1 to 6) as the input of aggregation network. Besides loss functions in Eqns. 3, 5, we also use the same perceptual loss as [59]. For back-propagation, we use the differentiable renderer [22]. For inference, using random size of image set or single image (i.e., only the target image) as input, the framework models 3D face for a target image.

5. Experiment

5.1. Setup

Dataset: We train our method mainly on the generated synthetic dataset (in Sec. 4.1), CelebA [30] and CASIA-WebFace [61]. To get photo collection of a same identity, we organize CelebA and CASIA-WebFace using ID-labels and keep identities with at least 6 photos. In this way, we get 16k different identities with 600k real face images, and select images of 12k/2k/2k identities as train/val/test set. For evaluation on facial geometry, following [59, 1], we perform testing on 3DFAW [15, 19, 68, 69], BFM [33] and Photoface [64] dataset. 3DFAW contains 23k images with 66 3D keypoint annotations, and we use the same protocol as [59] to perform testing. For BFM dataset, we use the same generated data released by [59] to evaluate predicted depth maps. Photoface dataset contains 12k images of 453 people with face/normal image pairs, and we follow the protocol of [42, 1] for testing. For evaluation on modeled texture, we perform fine-tuning and testing on CelebAMask-HQ [25] dataset which contains 30k real human facial im-

ages with high resolution (1024×1024). We organize it into 24k different identities using groundtruth ID-labels, and randomly select 20k/1k/3k identities as train/val/test set.

Implementation Details: We use the same architecture of $\Phi^{\omega}, \Phi^{l}, \Phi^{\sigma}$ as [59] to predict pose, light and confidence. Aggregation and attribute-refining network has the same encoder-decoder backbone as [59] to predict albedo and depth, respectively. As described in Sec. 4.3, we first train aggregation network and $\Phi^{\omega}, \Phi^{l}, \Phi^{\sigma}$ for 50 epochs on the proposed synthetic dataset, and then continue training on real photo sets for 50 epochs. Then we freeze them and train attribute-refining network for 100 epochs. Finally we fine-tune all the networks for 50 epochs. Training procedure has a batch size of 64 identities and learning rate of $1e-4$ with Adam solver [23] on one NVIDIA Tesla V100 GPU. We use image set of 128×128 as input of aggregation network and get the prediction a_c, d_c of the same size. For scene-specific face, we train different attribute-refining networks according to the size of target image ($64\times64, 128\times128, 256\times256$) and get the prediction a_t, d_t of that size. More details can be found in supp-material.

Evaluation Protocol: As our method can model 3D face for target image using photo set or single image, without special statement, we use single-image results to fairly compare with other methods. For predicted facial geometry, following [1, 59], we use Scale-Invariant Depth Error (SIDE) [11] and Mean Angle Deviation (MAD) for evaluating depth and normal. For evaluation on modeled texture, we calculate Structural Similarity Index (SSIM) [56] and cosine-similarity of encoded representation of Arcface [7] between original high-quality images and rendered ones.

5.2. Ablation Study

In this section, we perform experiments to analyse the effect of our method. To fairly compare with [59], we use the exact same network architecture as our method to build the model and train it using the same dataset but without multi-image consistency. We denote such reproduced model as **Unsup3D-re**. As original model of [59] has outputs of low resolution (64×64), Unsup3D-re can make valuable comparison on modeling ability of higher resolution.

Comparison with Baselines on Geometry: In Table 2 we make comparisons on different baselines and setups. Each model is trained on CelebA and CASIA-WebFace and then fine-tuned on BFM dataset. 'No-ft' means the model without fine-tuning. 'W/o adaptive aggregation' means using average fusion to fuse latent codes. 'W/o attribute injection' means only using feature vector of the highest level to inject target attribute without selecting. Methods with flag '-128' or '-256' mean the output size is 128×128 and 256×256, while other methods have an output of 64×64. As illustrated, Rows (1)-(6) reveal that our method has better ability on specialization and generalization, and outperforms Unsup3D [59] under the same settings. Rows (7)-(11) reveal the effect of each component of our methods, and demonstrate each of them contributes to the final prediction. Rows (1, 3) and (12)-(15) show the ability on modeling faces of higher resolution, which is more challenging due to more complex information on 3D space. The results demonstrate our method has more robust predictions when the resolution increases, while Unsup3D gets obvious performance decline. The above comparisons well support the effectiveness of our method.

No.	method	SIDE ($\times10^{-2}$) ↓	MAD (deg.) ↓
(1)	Ours-full	**0.721**$\pm$0.128	**15.53**$\pm$1.42
(2)	Unsup3D [59]	0.793$\pm$0.140	16.51$\pm$1.56
(3)	Unsup3D-re	0.785$\pm$0.152	16.44$\pm$1.63
(4)	Ours no-ft	**1.102**$\pm$0.205	**20.75**$\pm$2.06
(5)	Unsup3D [59] no-ft	1.295$\pm$0.233	21.84$\pm$2.56
(6)	Unsup3D-re no-ft	1.232$\pm$0.218	21.40$\pm$2.31
(7)	w/o curriculum learning	2.011$\pm$0.570	23.07$\pm$2.88
(8)	w/o RCL	0.738$\pm$0.135	15.66$\pm$1.50
(9)	w/o adaptive aggregation	0.764$\pm$0.142	16.21$\pm$1.76
(10)	w/o filtered connection	0.725$\pm$0.139	15.33$\pm$1.58
(11)	w/o attribute injection	0.750$\pm$0.157	16.01$\pm$1.20
(12)	Ours-full-128	0.708$\pm$0.121	15.42$\pm$1.38
(13)	Ours-full-256	**0.703**$\pm$0.137	**15.30**$\pm$1.26
(14)	Unsup3D-re-128	0.828$\pm$0.166	18.37$\pm$1.82
(15)	Unsup3D-re-256	0.930$\pm$0.182	19.79$\pm$1.95

Table 2. Comparison with Different Baselines and Settings.

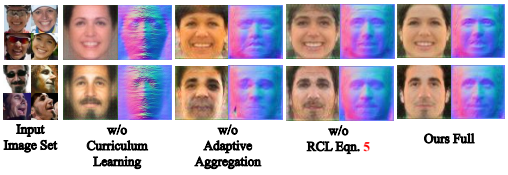


Figure 5. Predicted ID-consistent Face under Different Settings. We show canonical albedo and depth to make comparisons.

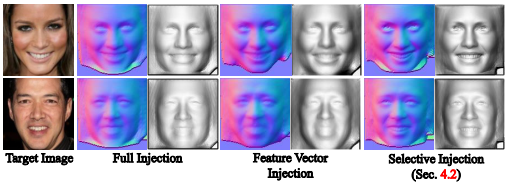
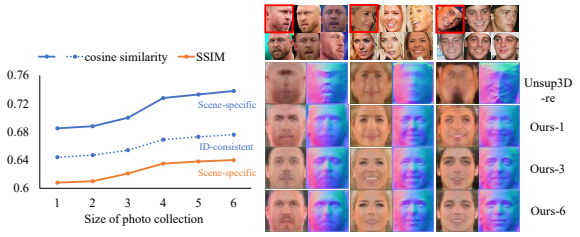


Figure 6. Comparison on predicted target facial geometry between baselines and our selective method in attribute injection module.

Qualitative Comparison: We first compare the predicted ID-consistent face under different settings in Fig. 5. As illustrated, the aggregation network cannot learn reasonable basic facial geometry without our curriculum learning method. Meanwhile, without adaptive aggregation or RCL, the learned ID-consistent depth and albedo are noisy and suffer from ambiguity. In contrast, our full method can model consistent geometry/texture with feature of the identity from the input image set with obvious higher quality. Secondly, we compare different injection methods described in Sec. 4.2 in Fig. 6. As illustrated, injecting full features of target image produces flat and over-smooth geometry, and injecting only the feature vector leads to coarse details. In contrast, our selective injection method models



(a) Performance with Different Size of Photo Collection

(b) ID-Consistent Canonical Albedo and Normal

Figure 7. Quality of prediction with different size of photo collection. (a) The quality of modeled texture on CelebAMask-HQ dataset. (b) Predicted ID-consistent face, the image with red frame is the one used for single-input model (Unsup3D-re and Ours-1).

Method	Depth Corr. ↑	Time (ms)
Ground Truth	66	-
AIGN [55] (supervised)	50.81	-
DepthNetGAN [31] (supervised)	**58.68**	-
MOFA [49] (3DMM based)	15.97	-
DepthNet [31]	35.77	-
Unsup3D [59] (CelebA pre-trained)	54.64	0.6
Unsup3D-re	55.83	2.0
Ours	**57.92**	2.0

Table 3. 3DFAW keypoint depth evaluation of different methods.

target facial geometry with superior details.

Multi vs. Single: We analyse the effect of photo collection size, i.e., the N in $\{\mathbf{I}_i\}_{i=1}^{N}$. We fine-tune the model on CelebAMask-HQ and evaluate the cosine-similarity and SSIM between $\mathbf{I}_t$ and rendered $\hat{\mathbf{I}}_t$ in Fig. 7(a). The model predicts target 3D face of 256×256, and we calculate SSIM on the same size in facial region. To compute cosine-similarity, we halve the rendered and target image and feed them into the pre-trained Arcface model after alignment for the requirement. As rendered image of ID-consistent face may have different expressions, we only compute cosine-similarity to analyse the predicted ID-feature. As illustrated, the quality of ID-consistent face increases with the size of image set, and this also contributes to better scene specific prediction. Such phenomenon demonstrates that the aggregated ID-consistent feature is crucial for high-quality modeling. Further, the quality of modeled texture is similar with 1 or 2 input images, but gets obvious increment with more images. This is due to the sufficiency of complementary information. The increment becomes lighter from 5 to 6, which reveals an potential upper bound. Qualitative results are shown in Fig. 7(b). Unsup3D-re fails to model reasonable face due to large pose and dis-alignment, while our single-input model (Ours-1) predicts robust results by learning constraint of multi-image consistency. With more input photos, the predictions get better performance.

5.3. Comparison with the State-of-the-Art

Analysis on geometry. We first evaluate the geometry of our predicted 3D face on 3DFAW dataset. To make fair comparison, following [59], we use the 2D keypoint locations to sample our predicted depth and calculate the depth correlation score [31] on frontal faces. As illustrated in Table 3, our method obviously outperforms AIGN, DepthNet and MOFA. For Unsup3D, our method also shows superiority. Though Unsup3D-re uses our architecture and dataset (CelebA and CASIA-Webface) for training and slightly improves the performance, our method gets further superior result which are closer to fully supervised approach. Inference time of our model is slightly slower than Unsup3D [59], but our method outperforms Unsup3D-re with the same time burden which demonstrate our implementation is efficient enough.

	MAD ↓	$< 20°$ ↑	$< 25°$ ↑	$< 30°$ ↑
Pix2V [41]	$33.9_{\pm 5.6}$	24.8%	36.1%	47.6%
Extreme [54]	$27.0_{\pm 6.4}$	37.8%	51.9%	47.6%
FNI [52]	$26.3_{\pm 10.2}$	4.3%	56.1%	**89.4%**
3DDFA [71]	$26.0_{\pm 7.2}$	40.6%	54.6%	66.4%
SfSNet [42]	$25.5_{\pm 9.3}$	43.6%	57.5%	68.7%
PRN [12]	$24.8_{\pm 6.8}$	43.1%	62.9%	74.1%
DF2Net [65] (GT)	$24.3_{\pm 5.7}$	42.2%	62.7%	74.5%
D3DFR [8]	$23.5_{\pm 6.1}$	46.1%	61.8%	73.3%
Cross-Modal [1] (GT)	$\mathbf{22.8}_{\pm 6.5}$	**49.0%**	62.9%	74.1%
Ours	$23.0_{\pm \mathbf{5.1}}$	48.2%	**63.1%**	74.9%
SfSNet-ft [42]	$12.8_{\pm 5.4}$	83.7%	90.8%	94.5%
Cross-Modal-ft [1] (GT)	$\mathbf{12.0}_{\pm 5.3}$	**85.2%**	92.0%	95.6%
Ours-ft	$12.3_{\pm \mathbf{4.5}}$	84.9%	**92.4%**	**96.3%**

Table 4. Facial Normal Evaluation on Photoface Dataset.

Method	Cosine-similarity ↑	SSIM ↑
Unsup3D [59] (64×64)	0.622	0.514
Unsup3D-re (256×256)	0.651	0.542
D3DFR [8]	0.398	0.335
Ours ID-consistent (128×128)	0.643	-
Ours (64×64)	0.695	0.618
Ours (128×128)	**0.697**	0.620
Ours (256×256)	0.692	**0.623**

Table 5. Quality of Rendered Image on CelebAMask-HQ.

We then evaluate predicted facial geometry on Photoface dataset. Following [1], we transform our predicted facial depth to normal map and resize it to 256×256 in order to compute MAD with ground truth. Results are illustrated in Table 4, where '-ft' means fine-tuning on Photoface. Our method outperforms most of the approaches and shows good generalization results. For Cross-Modal approach [1], our method gets competitive results with or without fine-tuning. Note that, the training procedure of [1] utilizes ground truth normal of 3D-scan which is crucial for understanding face geometry, while our model is fully unsupervised thus confronts more challenging conditions. These results demonstrate the effectiveness of our method.

Qualitative results are illustrated in Fig. 8. Note that Unsup3D has a limited size of output (64×64) and suffers from heavy noise when modeling larger output of 256×256 by our reproduced model (Unsup3D-re), while our method gets obvious superior results on the same resolution. Compared with non-parametric approaches [65, 59], our method obtains obviously better shape of organs in rows (1), (2) and (3). In rows (2) and (4), our method shows robustness on large pose and artifacts, and suffers from less ambiguity. Compared with 3DMM based methods [8, 16], our results have finer details and recovers better geometric correctness.

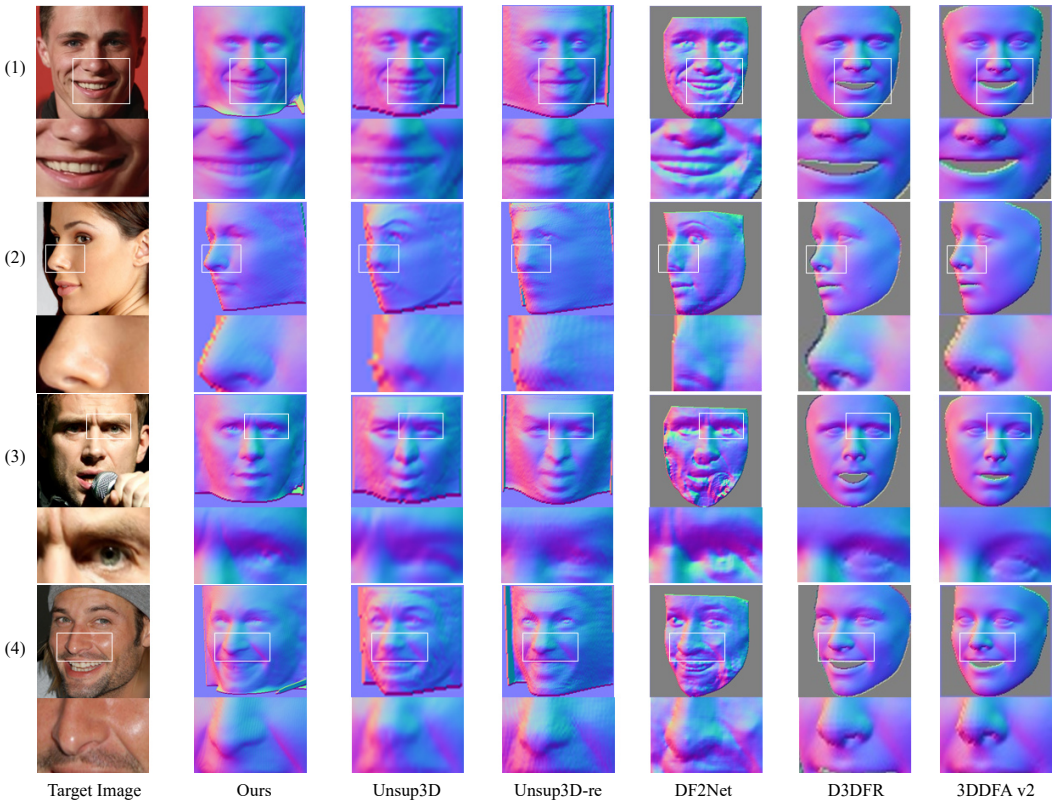


Figure 8. Qualitative results on predicted geometry. We compare our method with Unsup3D [59], DF2Net [65], D3DFR [8] and 3DDFA v2 [16], and the predictions are reproduced with their released code and pre-trained model. Unsup3D-re means our reproduced model with higher resolution. We use single input for our method to make fair comparison.

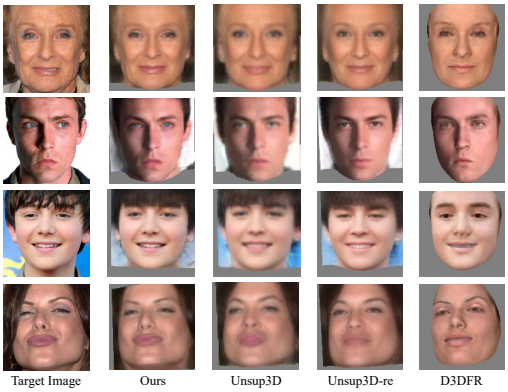


Figure 9. Qualitative Comparison on Rendered Image.

Analysis on Texture. We then analyse our modeled texture on high-quality CelebAMask-HQ dataset. We cover the target image with the modeled texture as the rendered image and use pre-trained Arcface [7] to compute consine-similarity. SSIM are only computed in facial region. We only compare our single-input results for fairness. As illustrated in Table 5, the rendered images of our method obtains the best performance. Note that, our ID-consistent predictions also get good consine-similarity, which reveals they aggregate reasonable features of target identities. Qualitative results can be viewed in Fig. 9, and our results have better perceptual quality.

6. Conclusion and Future Work

In this paper we propose a novel Learning to Aggregate and Personalize (LAP) framework for 3D face modeling without 3DMM prior or supervision. Based on statistical conclusion that non-rigid shape deformation is limited in face datasets, LAP adaptively aggregates consistent facial depth and albedo from in-the-wild photo collection, and learns multi-image consistency through a novel curriculum learning method with relaxation. For a face in specific scene, LAP personalizes the consistent face factors by attribute-refining network, improving finer details and attribute. Extensive experiments on benchmarks demonstrate LAP well leverages multi-image consistency and predict superior facial shape and texture. In the future, we may target on modeling 3D face on even higher resolution with reality, and leveraging unconstrained multi-image consistency by explicit algorithm beyond statistical assumption. [1]

[1]Acknowledgement: We thank the support from members of Tencent Youtu Lab for discussing and improving the ideas.

References

[1] Victoria Fernández Abrevaya, Adnane Boukhayma, Philip HS Torr, and Edmond Boyer. Cross-modal deep face normals with deactivable skip connections. In *CVPR*, pages 4979–4989, 2020. 1, 2, 3, 5, 6, 7

[2] Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, and Ping Tan. Deep facial non-rigid multi-view stereo. In *CVPR*, pages 5850–5860, 2020. 1, 2

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, pages 41–48, 2009. 4

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 1, 2

[5] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 1, 2

[6] Bindita Chaudhuri, Noranart Vesdapunt, Linda Shapiro, and Baoyuan Wang. Personalized face modeling for improved face reconstruction and motion retargeting. In *ECCV*, 2020. 2

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 4, 6, 8

[8] 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 *CVPRW*, 2019. 2, 3, 7, 8

[9] Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5908–5917, 2017. 1, 2

[10] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020. 1, 2

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, pages 2366–2374, 2014. 6

[12] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, pages 534–551, 2018. 1, 2, 7

[13] Zhongpai Gao, Juyong Zhang, Yudong Guo, Chao Ma, Guangtao Zhai, and Xiaokang Yang. Semi-supervised 3d face representation learning from unconstrained photo collections. In *CVPRW*, pages 348–349, 2020. 2, 3

[14] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *CVPR*, pages 8377–8386, 2018. 1, 2

[15] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 5

[16] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *ECCV*, 2020. 7, 8

[17] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *CVPR*, pages 5901–5910, 2020. 4

[18] 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 *ICCV*, pages 1031–1039, 2017. 1, 2, 3

[19] László A Jeni, Jeffrey F Cohn, and Takeo Kanade. Dense 3d face alignment from 2d videos in real-time. In *IEEE international conference and workshops on automatic face and gesture recognition*, volume 1, pages 1–8, 2015. 5

[20] Zi-Hang Jiang, Qianyi Wu, Keyu Chen, and Juyong Zhang. Disentangled representation learning for 3d face shape. In *CVPR*, pages 11957–11966, 2019. 3

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 4

[22] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, pages 3907–3916, 2018. 2, 3, 5

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[24] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction. In *CVPR*, pages 760–769, 2020. 2

[25] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 4, 5

[26] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *CVPR*, pages 10924–10933, 2019. 3

[27] Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *CVPR*, pages 5891–5900, 2020. 2

[28] Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Joint face alignment and 3d face reconstruction. In *ECCV*, pages 545–560. Springer, 2016. 1

[29] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *CVPR*, pages 5790–5799, 2017. 4

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 1, 5

[31] Joel Ruben Antony Moniz, Christopher Beckham, Simon Rajotte, Sina Honari, and Chris Pal. Unsupervised depth estimation, 3d face rotation and replacement. In *NeurIPS*, pages 9736–9746, 2018. 7

[32] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 4

[33] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009. 1, 5

[34] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *ECCV*, pages 818–833, 2018. 5

[35] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *International conference on 3D vision*, pages 460–469, 2016. 1, 2

[36] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, pages 1259–1268, 2017. 1, 2

[37] Sami Romdhani and Thomas Vetter. Efficient, robust and accurate fitting of a 3d morphable model. In *Computer Vision*, page 59. IEEE, 2003. 1, 2

[38] Sami Romdhani and Thomas Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR*, volume 2, pages 986–993, 2005. 1, 2

[39] Mihir Sahasrabudhe, Zhixin Shu, Edward Bartrum, Riza Alp Guler, Dimitris Samaras, and Iasonas Kokkinos. Lifting autoencoders: Unsupervised learning of a fully-disentangled 3d morphable model using deep non-rigid structure from motion. In *ICCVW*, 2019. 2, 3

[40] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *CVPR*, pages 7763–7772, 2019. 2, 3

[41] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *ICCV*, pages 1576–1585, 2017. 7

[42] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces in the wild. In *CVPR*, pages 6296–6305, 2018. 1, 2, 3, 5, 7

[43] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. *arXiv preprint arXiv:2007.12494*, 2020. 2

[44] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, pages 9243–9252, 2020. 4

[45] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *ICCV*, pages 6902–6911, 2019. 4

[46] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *ECCV*, pages 650–665, 2018. 3

[47] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Fml: Face model learning from videos. In *CVPR*, pages 10812–10822, 2019. 1, 2

[48] 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 *CVPR*, pages 2549–2559, 2018. 1, 2

[49] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCVW*, pages 1274–1283, 2017. 1, 2, 7

[50] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, pages 1126–1135, 2019. 2

[51] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *CVPR*, pages 7346–7355, 2018. 1, 2

[52] George Trigeorgis, Patrick Snape, Iasonas Kokkinos, and Stefanos Zafeiriou. Face normals in-the-wild using fully convolutional networks. In *CVPR*, pages 38–47, 2017. 1, 2, 3, 7

[53] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5163–5172, 2017. 1

[54] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *CVPR*, pages 3935–3944, 2018. 7

[55] Hsiao-Yu Fish Tung, Adam W Harley, William Seto, and Katerina Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *ICCV*, pages 4364–4372, 2017. 7

[56] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004. 6

[57] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *CVPR*, pages 959–968, 2019. 1, 2

[58] Rongliang Wu and Shijian Lu. Leed: Label-free expression editing via disentanglement. In *ECCV*, pages 781–798, 2020. 3

[59] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *CVPR*, pages 1–10, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[60] Xianglei Xing, Tian Han, Ruiqi Gao, Song-Chun Zhu, and Ying Nian Wu. Unsupervised disentangling of appearance and geometry by deformable generator network. In *CVPR*, pages 10354–10363, 2019. 3

[61] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 1, 5

[62] Jae Shin Yoon, Takaaki Shiratori, Shoou-I Yu, and Hyun Soo Park. Self-supervised adaptation of high-fidelity face models for monocular performance tracking. In *CVPR*, pages 4601–4609, 2019. 1, 2

[63] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, pages 325–341, 2018. 4

[64] Stefanos Zafeiriou, Mark Hansen, Gary Atkinson, Vasileios Argyriou, Maria Petrou, Melvyn Smith, and Lyndon Smith. The photoface database. In *CVPRW*, pages 132–139, 2011. 5

[65] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *ICCV*, pages 2315–2324, 2019. 1, 2, 3, 7, 8

[66] Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Generative adversarial network with spatial attention for face attribute editing. In *ECCV*, pages 417–432, 2018. 5

[67] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(8):690–706, 1999. 1, 3

[68] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2013. 5

[69] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 5

[70] Yuxiang Zhou, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In *CVPR*, pages 1097–1106, 2019. 1

[71] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *CVPR*, pages 146–155, 2016. 1, 2, 7

[72] Xiangyu Zhu, Fan Yang, Chang Yu Di Huang, Hao Wang, Jianzhu Guo, Zhen Lei, and Stan Z Li. Beyond 3dmm space: Towards fine-grained 3d face reconstruction. In *ECCV*, 2020. 1, 2

[73] Xiangyu Zhu, Dong Yi, Zhen Lei, and Stan Z Li. Robust 3d morphable model fitting by sparse sift flow. In *ICCV*, pages 4044–4049. IEEE, 2014. 1, 2