Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set

Yu Deng*[1,2] Jiaolong Yang[2] Sicheng Xu[3,2] Dong Chen[2] Yunde Jia[3] Xin Tong[2]

[1]Tsinghua University [2]Microsoft Research Asia [3]Beijing Institute of Technology

{v-denyu,jiaoyan,doch,xtong}@microsoft.com, {xusicheng,jiayunde}@bit.edu.cn

Abstract

Recently, deep learning based 3D face reconstruction methods have shown promising results in both quality and efficiency. However, training deep neural networks typically requires a large volume of data, whereas face images with ground-truth 3D face shapes are scarce. In this paper, we propose a novel deep 3D face reconstruction approach that 1) leverages a robust, hybrid loss function for weakly-supervised learning which takes into account both low-level and perception-level information for supervision, and 2) performs multi-image face reconstruction by exploiting complementary information from different images for shape aggregation. Our method is fast, accurate, and robust to occlusion and large pose. We provide comprehensive experiments on MICC Florence and Facewarehouse datasets, systematically comparing our method with fifteen recent methods and demonstrating its state-of-the-art performance. Code available at https://github.com/Microsoft/Deep3DFaceReconstruction

1. Introduction

Faithfully recovering the 3D shapes of human faces from unconstrained 2D images is a challenging task and has numerous applications such as face recognition [6, 51, 59], face media manipulation [5, 50], and face animation [10, 23]. Recently, there is a surge of interest in 3D face reconstruction from a single image using deep Convolutional Neutral Networks (CNN) in lieu of the complex and costly optimization used by traditional methods [37, 13, 38, 51, 25, 45, 49, 48, 53, 46, 14, 18]. Since ground truth 3D face data is scarce, many previous approaches resort to synthetic data or using 3D shapes fitted by traditional methods as surrogate shape labels [37, 57, 45, 31, 14, 18]. However, their accuracy may be jeopardized by the domain gap issue or the imperfect training labels.

To circumvent these issues, methods have been proposed to train networks without shape labels in an unsupervised or weakly-supervised fashion and promising results have been obtained [49, 48, 53, 46, 16]. The crux of unsupervised learning is a differentiable image formation procedure which renders a face image with the network predictions, and the supervision signal stems from the discrepancy between the input image and the rendered counterpart. For example, Tewari *et al*. [49] and Sengupta *et al*. [46] use pixel-wise photometric difference as training loss. To improve robustness and expressiveness, Tewari *et al*. [48] proposed a two-step reconstruction scheme where the second step produces a shape and texture correction with a neural network. Genova *et al*. [16] proposed to measure face image discrepancy on the perception level by using the distances between deep features extracted from a face recognition network.

Our goal in this paper is to obtain accurate 3D face reconstruction with weakly-supervised learning. We identified that using low-level information of pixel-wise color alone may suffer from local minimum issue where low error can be obtained with unsatisfactory face shapes. On the other hand, using only perceptual loss also lead to sub-optimal results since it ignores the pixel-wise consistency with raw image signal. In light of this, we propose a hybrid-level loss function which integrates both of them, giving rise to accurate results. We also propose a novel skin color based photometric error attention strategy, granting our method further robustness to occlusion and other challenging appearance variations such as beard and heavy make-up. We train an off-the-shelf deep CNN to predict 3D Morphable Model (3DMM) [5] coefficients, and accurate reconstruction is achieved on multiple datasets [1, 11, 56].

With a strong CNN model for single-image 3D face reconstruction, we take a further step and consider the problem of CNN-based face reconstruction aggregation with a set of images. Given multiple face images of a subject (*e.g*., from a personal album) captured in the wild under disparate

*This work was done when Yu Deng was an intern at MSRA.

conditions, it is natural to leverage all the images to build a better 3D face shape. To apply the deep neural networks on arbitrary number of images, one solution would be aggregating the single-image reconstruction results, and perhaps the simplest strategy is naively averaging the recovered shapes. However, such a native strategy did not consider the quality of the input images (*e.g.*, if some samples contain severe occlusion). Nor does it take full advantage of pose differences to improve the shape prediction.

In this paper, we propose to learn 3D face aggregation from multiple images, also in an unsupervised fashion. We train a simple auxiliary network to produce "confidence scores" of the regressed identity-bearing 3D model coefficients, and obtain final identity coefficients via confidence-based aggregation. Despite no explicit confidence label is used, our method automatically learns to favor high-quality (especially high-visibility) photos. Moreover, it can exploit pose difference to better fuse the complementary information, learning to more accurate 3D shapes.

To summarize, this paper makes the following two main contributions:

- We propose a CNN-based single-image face reconstruction method which exploits hybrid-level image information for weakly-supervised learning. Our loss consists of a robustified image-level loss and a perception-level loss. We demonstrate the benefit of combing them, and show the state-of-the-art accuracy of our method on multiple datasets [1, 11, 56], significantly outperforming previous methods trained in a fully supervised fashion [45, 14, 51]. Moreover, we show that with a low-dimensional 3DMM subspace, we are still able to outperform prior art with "unrestricted" 3D representations [45, 53, 48, 14] by an appreciable margin.

- We propose a novel shape confidence learning scheme for multi-image face reconstruction aggregation. Our confidence prediction subnet is also trained in a weakly-supervised fashion without ground-truth label. We show that our method clearly outperforms naive aggregation (*e.g.*, shape averaging) and some heuristic strategies [34]. To our knowledge, this is the first attempt towards CNN-based 3D face reconstruction and aggregation from an unconstrained image set.

2. Related Work

3D face reconstruction has been a longstanding task in computer vision and computer graphics. In the literature, 3D Morphable Models (3DMM) [5, 33, 7] have played a paramount role for 3D face modelling. With a 3DMM, reconstruction can be performed by an analysis-by-synthesis scheme using image intensity [5] and other features such as edges [39]. More recently, model fitting using facial landmarks has gained much popularity with the growth of face alignment techniques [4, 58, 21, 3]. However, sparse landmarks cannot well capture the dense facial geometry. Beyond 3DMM, another popular 3D face model is the multilinear tensor model [54, 10, 9, 43]. A few model-free single-image reconstruction methods have been proposed [20, 27, 19]; most require some reference 3D face shapes. For example, [20, 19] estimate image depth by building correspondences between the input image and one or a set of reference 3D faces. In [27], a shape-from-shading approach is proposed with a reference 3D face as prior.

The aforementioned approaches usually involve costly optimization to recover a quality 3D face. Recently, numerous method are proposed which employ CNNs for efficient face reconstruction. Some of them apply CNNs to regress 3DMM coefficients [37, 13, 2, 49, 51, 16], some use multi-step schemes to add correction or details onto coarse 3DMM predictions [38, 48, 52, 18], while others advocate direct model-free reconstruction [45, 53, 46, 14].

For all these CNN-based methods, one great hurdle is the lack of training data. Many methods resort to synthetic data or using 3D shapes fitted by traditional methods as surrogate labels [37, 57, 45, 31, 14, 18]. Others have attempted unsupervised or weakly-supervised training [49, 48, 53, 46, 16]. Our method is also based on weakly-supervised learning, for which our findings in this paper are threefold: **1)** the loss function is important for weakly-supervised learning and both low-level and perception-level information should be leveraged; **2)** the results obtained with weak supervision can be significantly better than those trained with synthetic data or pseudo-ground truth shapes, and **3)** somewhat surprisingly, the results confined in the low-dimensional 3DMM subspace can still be much better than state-of-the-art results with "unrestricted" representations.

We also studied the problem of reconstruction aggregation from multiple images. One related work is [34] which investigated the reconstruction quality measurement closest to human ratings and used it to fuse the reconstructions obtained with 3DMM fitting. We however show that their method is deficient in our case. Our method is also related to traditional methods working on unconstrained photo collections [27, 47, 40, 41]. While excellent results have been obtained by these methods, they typically consist of multiple steps such as face frontalization, photometric stereo, and local normal refinement. The whole pipeline is complex and may break down under severe occlusion and extreme pose. Our goal in this paper is not to replace these traditional methods, but to study the shape aggregation problem (similar to [34]) with a CNN and provide an extremely fast and robust alternative learned end-to-end.

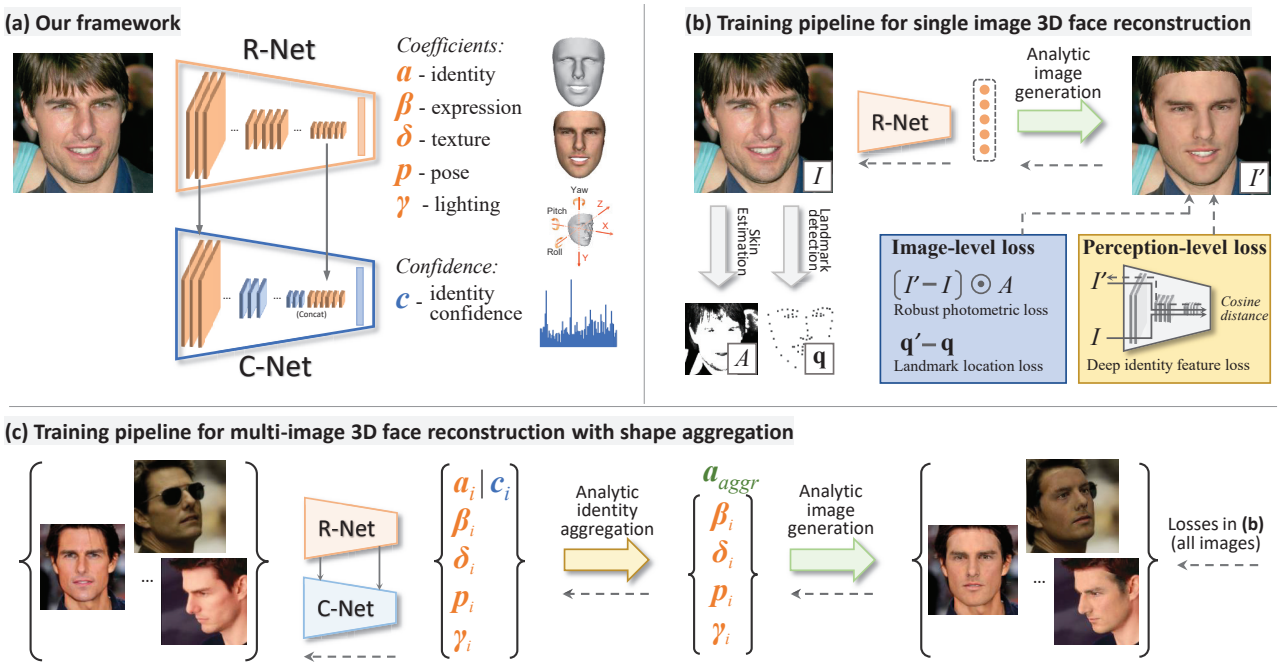

Figure 1. Overview of our approach. **(a)** The framework of our method, which consists of a reconstruction network for end-to-end single image 3D reconstruction and a confidence measurement subnet designed for multi-image based reconstruction. **(b)** The training pipeline for single images with our proposed hybrid-level loss functions. Our method does not require any ground-truth 3D shapes for training. It only leverages some weak supervision signals such as facial landmarks, skin mask and a pre-trained face recognition CNN. **(c)** The training pipeline for multi-image based reconstruction. Our confidence subnet learns to measure the reconstruction confidence for aggregation with out any explicit label. The dashed arrows denote error backpropagration for network training.

3. Preliminaries: Models and Outputs

As shown in Fig. 1 (a), we use a CNN to regress coefficients of a 3DMM face model. For unsupervised/weakly-supervised training [49, 48], we also regress the illumination and face pose to enable analytic image regeneration. We detail our models and CNN outputs as follows.

3D Face Model. With a 3DMM, the face shape $\mathbf{S}$ and the texture $\mathbf{T}$ can be represented by an affine model:

$$\begin{aligned} \mathbf{S} &= \mathbf{S}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \bar{\mathbf{S}} + \mathbf{B}_{id}\boldsymbol{\alpha} + \mathbf{B}_{exp}\boldsymbol{\beta} \\ \mathbf{T} &= \mathbf{T}(\boldsymbol{\delta}) = \bar{\mathbf{T}} + \mathbf{B}_{t}\boldsymbol{\delta} \end{aligned} \tag{1}$$

where $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ are the average face shape and texture; $\mathbf{B}_{id}$, $\mathbf{B}_{exp}$, and $\mathbf{B}_t$ are the PCA bases of identity, expression, and texture respectively, which are all scaled with standard deviations; $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\delta}$ are the corresponding coefficient vectors for generating a 3D face. We adopt the popular 2009 Basel Face Model [33] for $\bar{\mathbf{S}}$, $\mathbf{B}_{id}$, $\bar{\mathbf{T}}$, and $\mathbf{B}_t$, and use the expression bases $\mathbf{B}_{exp}$ of [18] which are built from FaceWarehouse [11]. A subset of the bases is selected, resulting in $\boldsymbol{\alpha} \in \mathbb{R}^{80}$, $\boldsymbol{\beta} \in \mathbb{R}^{64}$ and $\boldsymbol{\delta} \in \mathbb{R}^{80}$. We exclude the ear and neck region, and our final model contains 36K vertices.

Illumination Model. We assume a Lambertian surface for face and approximate the scene illumination with Spherical Harmonics (SH) [35, 36]. The radiosity of a vertex $\mathbf{s}_i$ with surface normal $\mathbf{n}_i$ and skin texture $\mathbf{t}_i$ can then be computed as $\mathbf{C}(\mathbf{n}_i, \mathbf{t}_i|\boldsymbol{\gamma}) = \mathbf{t}_i \cdot \sum_{b=1}^{B^2} \gamma_b \Phi_b(\mathbf{n}_i)$ where $\Phi_b : \mathbb{R}^3 \rightarrow \mathbb{R}$ are SH basis functions and γ_b are the corresponding SH coefficients. We choose $B = 3$ bands following [49, 48] and assume white lights such that $\boldsymbol{\gamma} \in \mathbb{R}^9$.

Camera Model. We use the perspective camera model with an empirically-selected focal length for the 3D-2D projection geometry. The 3D face pose $\mathbf{p}$ is represented by rotation $\mathbf{R} \in \mathrm{SO}(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$.

In summary, the unknowns to be predicted can be represented by a vector $\mathbf{x} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{p}) \in \mathbb{R}^{239}$. In this paper, we use a ResNet-50 network [22] to regress these coefficients by modifying the last fully-connected layer to 239 neurons. For brevity, we denote this modified ResNet-50 network for single image reconstruction as R-Net. We present how we train it in the following section.

4. Hybrid-level Weak-supervision for Single-Image Reconstruction

Given a training RGB image I, we use R-Net to regress a coefficient vector $\mathbf{x}$, with which a reconstructed image I' can be analytically generated with some simple, differentiable math derivations. Some examples of I' can be found

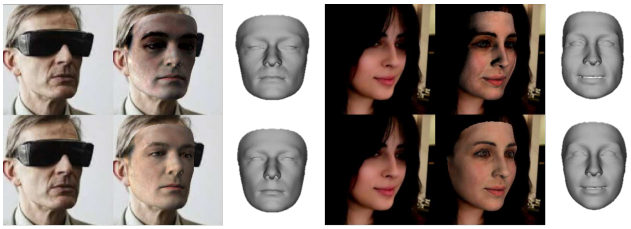

Figure 2. Comparison of the results without (*top row*) and with (*bottom row*) using our skin attention mask for training.

in Fig. 1. Our R-Net is trained without any ground truth labels, but via evaluating a hybrid-level loss on I' and back-propagate it.

4.1. Image-Level Losses

We first introduce our loss functions on low-level information including per-pixel color and sparse 2D landmarks.

4.1.1 Robust Photometric Loss

First, it is straightforward to measure the dense photometric discrepancy between the raw image and the reconstructed one [5, 50, 49, 48]. In this paper, we propose a robust, skin-aware photometric loss instead of a naive one, defined as:

$$L_{photo}(\mathbf{x}) = \frac{\sum_{i \in \mathcal{M}} A_i \cdot \|I_i - I'_i(\mathbf{x})\|_2}{\sum_{i \in \mathcal{M}} A_i} \tag{2}$$

where i denotes pixel index, $\mathcal{M}$ is reprojected face region which can be readily obtained, $\|\cdot\|$ denotes the l_2 norm, and A is a skin color based attention mask for the training image which is described as follows.

Skin Attention. To gain robustness to occlusions and other challenging appearance variations such as beard and heavy make-up, we compute a skin-color probability P_i for each pixel. We train a naive Bayes classifier with Gaussian Mixture Models on a skin image dataset from [26]. For each pixel i, we set $A_i = \begin{cases} 1, & \text{if } P_i > 0.5 \\ P_i, & \text{otherwise} \end{cases}$. We find that such a simple skin-aware loss function works remarkably well in practice without the need for a face segmentation method [43]. Figure 2 illustrates the benefit of using our skin attention mask.

It is also worth mentioning that our loss in Eq. 2 integrates over 2D image pixels as opposed to 3D shape vertices in [49, 48]. It enables us to easily identify self-occlusion via z-buffering thus our trained model can handle large poses.

4.1.2 Landmark Loss

We also use landmark locations on the 2D image domain as weak supervision to train the network. We run the state-of-the art 3D face alignment method of [8] to detect 68 landmarks $\{\mathbf{q}_n\}$ of the training images. During training, we project the 3D landmark vertices of our reconstructed shape onto the image obtaining $\{\mathbf{q}'_n\}$, and compute the loss as:

$$L_{lan}(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \omega_n \|\mathbf{q}_n - \mathbf{q}'_n(\mathbf{x})\|^2 \tag{3}$$

where ω_n is the landmark weight which we experimentally set to 20 for inner mouth and nose points and others to 1.

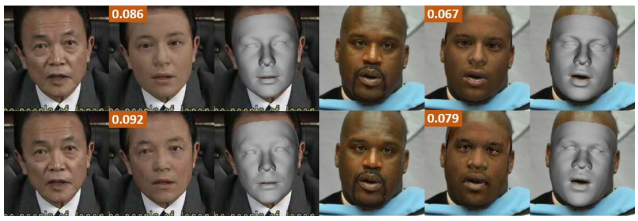

Figure 3. Comparison of the results with only image-level losses (*top row*) and with both image-level and perceptual losses (*bottom row*) for training. The numbers are the evaluated photometric errors. A lower photometric error does not guarantee a better shape.

4.2. Perception-Level Loss

While using the low-level information to measure image discrepancy can generally yields decent results [5, 50, 49, 48], we find using them alone can lead to local minimum issue for CNN-based 3D face reconstruction. Figure 3 shows that our R-net trained with only image-level losses generates smoother textures and lower photometric errors than the compared opponents, but the resultant 3D shapes are less accurate by visual inspection.

To tackle this issue, we introduce a perception-level loss to further guide the training. Inspired by [16], we seek for the weak-supervision signal from a pre-trained deep face recognition network. Specifically, we extract the deep features of the images and compute the cosine distance:

$$L_{per}(\mathbf{x}) = 1 - \frac{< f(I), f(I'(\mathbf{x})) >}{\|f(I)\| \cdot \|f(I'(\mathbf{x}))\|} \tag{4}$$

where $f(\cdot)$ denotes deep feature encoding and $< \cdot, \cdot >$ vector inner product. In this work, we train a FaceNet [44] structure using an in-house face recognition dataset with 3M face images of 50K identities crawled from the Internet, and use it as our deep feature extractor.

Figure 3 shows that with the perceptual loss, the textures are sharper and the shapes are more faithful. Quantitative results in the experiment section also show the benefit.

4.3. Regularization

To prevent face shape and texture degeneration, we add a commonly-used loss on the regressed 3DMM coefficients:

$$L_{coef}(\mathbf{x}) = \omega_\alpha \|\boldsymbol{\alpha}\|^2 + \omega_\beta \|\boldsymbol{\beta}\|^2 + \omega_\gamma \|\boldsymbol{\delta}\|^2 \tag{5}$$

which enforces a prior distribution towards the mean face. The balancing weights are empirically set to $\omega_\alpha = 1.0$, $\omega_\beta = 0.8$ and $\omega_\gamma = 1.7e{-}3$.

Although the face textures in the Basel 2009 3DMM [33] were obtained with special devices, they still contain some baked-in shading (*e.g.*, ambient occlusion). To favor a constant skin albedo similar to [48], we add a flattening constrain to penalize the texture map variance:

$$L_{tex}(\mathbf{x}) = \textstyle\sum_{c\in\{r,g,b\}} var(\mathbf{T}_{c,\mathcal{R}}(\mathbf{x})) \tag{6}$$

where $\mathcal{R}$ is a pre-defined skin region covering cheek, noise, and forehead.

In summary, our loss function $L(\mathbf{x})$ for R-Net is composed of two image-level losses, a perceptual loss and two regularization loss. Their weights are set to $w_{photo} = 1.9$, $w_{lan} = 1.6e{-}3$, $w_{per} = 0.2$, $w_{coef} = 3e{-}4$ and $w_{tex} = 5$ respectively in all our experiments.

5. Weakly-supervised Neural Aggregation for Multi-Image Reconstruction

Given multiple face images of a subject (*e.g.*, a photo album), it is natural to leverage all the images to build a better 3D face shape. Images captured under different conditions should contain information complementary to each other due to change of pose, lighting *etc*. Moreover, using an image set for reconstruction can gain further robustness to occlusion and bad lighting in some individual images.

Applying deep neural networks on an arbitrary number of orderless images is not straightforward. In this work, we use a network to learn a measurement of confidence or quality of the single-image reconstruction results, and use it to aggregate the individual shapes. Specifically, we seek to generate a vector $\boldsymbol{c} \in \mathbb{R}^{80}$ with positive elements measuring the confidence of the identity-bearing shape coefficients $\boldsymbol{\alpha} \in \mathbb{R}^{80}$. We do not consider other coefficients such as expression, pose, and lighting as they vary across images and fusion is unnecessary. We also bypass texture as we found the skin color of a subject can vary significantly across in-the-wild images. Let $\mathcal{I} := \{I^j | j = 1, \dots, M\}$ be an image collection of a person, $\mathbf{x}^j = (\boldsymbol{\alpha}^j, \boldsymbol{\beta}^j, \boldsymbol{\delta}^j, \mathbf{p}^j, \boldsymbol{\gamma}^j)$ the output coefficient vector from R-Net for each image j, and $\boldsymbol{c}^j$ the confidence vector for each $\boldsymbol{\alpha}^j$, we obtain the final shape via element-wise shape coefficient aggregation:

$$\boldsymbol{\alpha}_{aggr} = \left(\textstyle\sum_j \boldsymbol{c}^j \odot \boldsymbol{\alpha}^j\right) \oslash \left(\textstyle\sum_j \boldsymbol{c}^j\right) \tag{7}$$

where $\odot$ and $\oslash$ denote Hadamard product and division, respectively.

Next, we present how we train a network, denoted as C-Net, to predict $\boldsymbol{c}$ in a weakly-supervised fashion without labels. The structure of C-Net will be presented afterwards.

5.1. Label-Free Training

To train C-Net on image sets, we generate the reconstructed image set $\{I^{j\prime}\}$ of $\{I^j\}$ with $\{\hat{\mathbf{x}}^j\}$, where $\hat{\mathbf{x}}^j = (\boldsymbol{\alpha}_{aggr}, \boldsymbol{\beta}^j, \boldsymbol{\delta}^j, \mathbf{p}^j, \boldsymbol{\gamma}^j)$. We define the training loss as

$$\mathcal{L}(\{\hat{\mathbf{x}}^j\}) = \frac{1}{M}\sum_{j=1}^{M} L(\hat{\mathbf{x}}^j) \tag{8}$$

where $L(\cdot)$ is our hybrid-level loss function defined in Section 4 evaluated with $I^{j\prime}$ of I^j.

This way, the error can be backpropagated to $\boldsymbol{\alpha}_{aggr}$ thus further to $\boldsymbol{c}$ and C-Net weights, since Eq. 7 is differentiable. C-Net will be trained to produce confidences that lead to an aggregated 3D face shape consistent with the face image set as much as possible. The pipeline is illustrated in Fig. 1(c). In the multi-image training stage, the loss weights $\omega_{lm}, \omega_{photo}$ and ω_{id} are set to $1.6e{-}3$, 1.9, and 0.1 respectively.

Our aggregation design and training scheme are inspired by the set-based face recognition work of [55]. However, [55] used a scalar quality score for feature vector aggregation, whereas we produce element-wise scores for 3DMM coefficients. In Section 6.2.1, we show element-wise scores yield superior results and analyze how our network exploits face pose difference for better shape aggregation.

5.2. Confidence-Net Structure

Our C-Net is designed to be light-weight. Since R-Net is able to predict high-level information such as pose and lighting, it is natural to reuse its feature maps for C-Net. In practice, we take both shallow and deep features from R-Net, as illustrated in Fig. 1 (a). The shallow feature can be used to measure image corruptions such as occlusion.

Specifically, we take the features after the first residual block $F_{b1} \in \mathbb{R}^{28\times28\times256}$ and after global pooling $F_g \in \mathbb{R}^{2048}$ of R-Net as the input to C-Net. We apply three 3×3 convolution layers 256 channels and stride 2, followed by a global pooling layer on F_{b1} to get $F'_{b1} \in \mathbb{R}^{256}$. We then concatenate F'_{b1} and F_g, and apply two fully-connected layers with 512 and 80 neurons respectively. At last, we apply sigmoid function to make the confidence predictions $\boldsymbol{c} \in \mathbb{R}^{80}$ positive. Our C-Net has 3M parameters in total, which is about $1/8$ size of R-Net.

6. Experiments

Implementation Details. To train our R-Net, we collected in-the-wild images from multiple sources such as CelebA [32], 300W-LP [57], I-JBA [30], LFW [24] and LS3D [8]. We balanced the pose and race distributions and get ∼260K face images as our training set. We use the method of [12] to detect and align the images. The input image size is 224×224. We take the weights pre-trained in ImageNet [42] as initialization, and train R-Net using Adam

Table 1. Average reconstruction errors (mm) on MICC [1] and FaceWarehouse [11] datasets for R-Net trained with different losses. Our full hybrid-level loss function yields significantly higher accuracy than other baselines on both datasets.

Losses			MICC	Facewarehouse
L_{photo}	L_{lan}	L_{per}		
		✓	1.87±0.43	2.70±0.73
✓	✓		1.80±0.52	2.17±0.65
	✓	✓	1.71±0.43	2.11±0.48
✓	✓	✓	**1.67±0.50**	**1.81±0.50**

Table 2. Mean Root Mean Squared Error (RMSE) across 53 subjects on MICC dataset (in mm). We use ICP for alignment and compute point-to-plane distance between results and ground truth.

Method	Cooperative	Indoor	Outdoor
Tran *et al.* [51]	1.97±0.49	2.03±0.45	1.93±0.49
Genova *et al.* [16]	1.78±0.54	1.78±0.52	1.76±0.54
Ours	**1.66±0.52**	**1.66±0.46**	**1.69±0.53**

optimizer [29] with batch size of 5, initial learning rate of $1e{-}4$, and 500K total iterations.

To train C-Net, we construct an image corpus using 300W-LP [57], Multi-PIE [17] and part of our in-house face recognition dataset. For 300W-LP and Multi-PIE, we choose 5 images with rotation angles evenly distributed for each person. For the face recognition dataset, we randomly select 5 images for each person. The whole training set contains ∼50K images of ∼10K identities. We freeze the trained R-Net, and randomly initialize C-Net except for its last fully-connected layer which is initialized to zero (so that we start from average pooling). We train it using Adam [29] with batch size of 5, initial learning rate of $2e{-}5$ and 10K total iterations.

6.1. Results on Single Image Reconstruction

6.1.1 Ablation Study

To validate the efficacy of our proposed hybrid-level loss function, we conduct ablation study on two datasets: the MICC Florence 3D Face dataset [1] and the FaceWarehouse dataset [11]. MICC contains 53 subjects, each associated with a ground truth scan in neutral expression and three video sequences captured in cooperative, indoor, and outdoor scenarios. For FaceWarehouse, we use 9 subjects each with 20 expressions for evaluation.

Table 1 presents the reconstruction errors with various loss combinations. It shows that jointly considering image- and perception-level information gives rise to significantly higher accuracy than using them separately.

6.1.2 Comparison with Prior Art

Comparison on MICC Florence with [51, 16, 14, 25, 57, 31]. We first compare with the methods of Tran *et al.* [51] and Genova *et al.* [16]. For [51] and ours, we evaluate the error with the average shape from a sequence. [16] averaged their encoder embeddings from all frames before reconstruction and produce a single shape per sequence. Following [16], we crop the ground truth mesh to 95mm around the nose tip and run ICP with isotropic scale for alignment. The results of [51] only contains part of the forehead region, thus we further cut the ground truth meshes accordingly for fair comparison. Table 2 shows that our method significantly outperforms [51] and [16] on all three sequences. The qualitative comparison in Fig. 4 and Fig. 5 also demonstrates the superiority of our results. Note that [16] uses a perceptual loss similar to ours, but they ignores the low-level information such as photometric similarity.

We then compare with PRN [14], a recent CNN method with supervised learning that predicts unrestricted face shapes. Following [14], we render face images with 20 poses for each subject using pitch angles of -15, 20, and 25 degrees and yaw angles of -80, 40, 0, 40, 80 degrees. Figure 6 shows the point-to-plane RMSE averaged across subjects and pitch angles. Our method has a much lower error than PRN for all yaw angles. Also note that PRN has

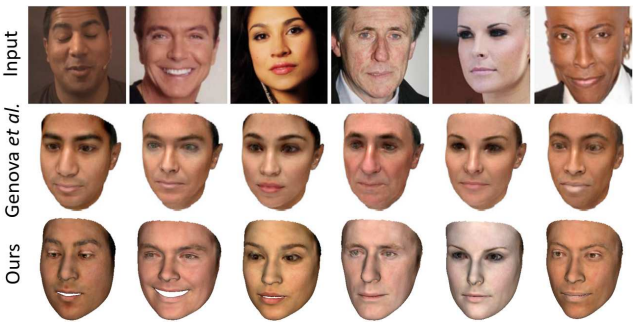

Figure 4. Comparison with Genova *et al.* [16]. Our texture and shape exhibit larger variance and are more consistent with the inputs. The images are from [16].

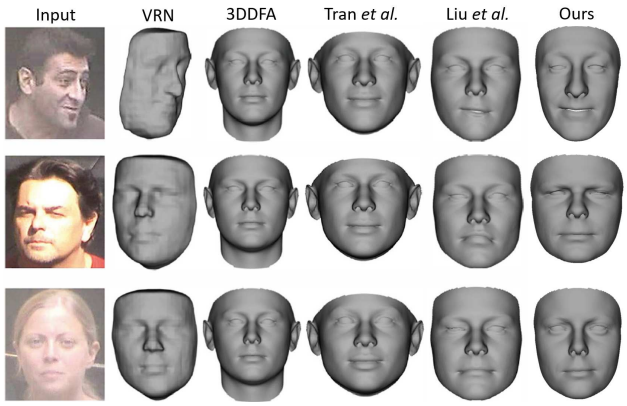

Figure 5. Comparison with VRN [25], 3DDFA [57], Tran *et al.* [51], Liu *et al.* [31] on three MICC subjects. Our results show largest variance and are visually most faithful among all methods. The input images and results of other methods are from [31].

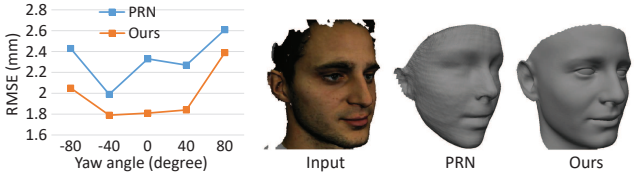

Figure 6. Comparison with PRN [14] on MICC. Leftmost: Mean RMSE of different yaw angles. Our method excels at all views. Right three images: qualitative result comparison.

a larger model size than ours (160MB vs. 92MB).

We further qualitatively compare with several learning-based methods including VRN [25], 3DDFA [57], and Liu *et al*. [31]. Figure 5 shows that our method can well-recover both identity and expression, whereas the results of other methods have very low shape variance.

Comparison on Facewarehouse with [48, 49, 28, 15]. We compare our results on the 9 Facewarehouse subjects selected by [48], with three learning-based approaches of Tewari *et al*. [49, 48], Kim *et al*. [28] and an optimization-based approach of Garrido *et al*. [15]. The evaluation protocol of [48] is used.

We evaluate two face regions: a smaller one same as [48]'s, and a larger one with more cheek areas included (see Fig. 7). The point-to-point errors are presented in Table 3. Our method achieved the lowest reconstruction error among all learning-based methods. Note that [49], [48]-C (coarse results), [28], and our method are all based on 3DMM representation, and we show significant improvement upon theirs. Our method is even better than [48]-F which uses a corrective space to refine the 3DMM shape. Our accuracy gets closer to the optimization-based approach of [15] while our method can be orders of magnitude faster.

Figure 7.

We further compare with [48] qualitatively in Fig 8. Our recovered shapes are of higher fidelity. Moreover, some artifacts from [48] can be observed under occlusion while our results are much more pleasing. Also note that our method can handle profile faces (see, *e.g*., Fig. 9), while the large-pose robustness of the above methods are unclear to us.

Comparison with other methods [37, 53]. Figure 10 compares our results with Richardson *et al*. [37], Tran and Liu [53] and Tewari *et al*. [49]. By visual inspection, our method produces better results.

6.2. Results on Multi Images Reconstruction

6.2.1 Ablation Study and Analysis

To test our multi-image shape aggregation method, we first conduct ablation study on render images of MICC. We render 20 poses for each of the 53 subjects as in Sec. 6.1.2. Table 4 presents the shape error of different aggregation strategies (S1 to S4). For S1, we train a C-Net similar to that described in Sec. 5.1 but modify the final FC layer to output a global confidence score $c^j \in \mathbb{R}^+$, and we aggregate the identity coefficients via $\boldsymbol{\alpha}_{aggr} = \sum_j c^j \cdot \boldsymbol{\alpha}^j / \sum_j c^j$. For

Table 3. Mean reconstruction error (mm) on 180 meshes of 9 subjects from FaceWarehouse. "-F" and "-R" denote the "fine" and "coarse" results of [48]. The face regions "S" (Smaller) and "L" (Larger) are shown in Fig. 7. Our error is lowest among the learning-based methods. *: due to the GPU parallel computing scheme, one forward pass of our R-Net takes 20ms with both batch-size 1 and batch-size 10 (evaluated with an NVIDIA TITAN Xp GPU). The times of other methods are quoted from [48].

	Learning					Optimization
	Ours	[48]-F	[48]-C	[49]	[28]	[15]
Region-S	**1.81**	1.84	2.03	2.19	2.11	1.59
Region-L	**1.91**	2.00	-	-	-	1.84
Time	20ms (2ms*)	4ms	4ms	4ms	4ms	120s

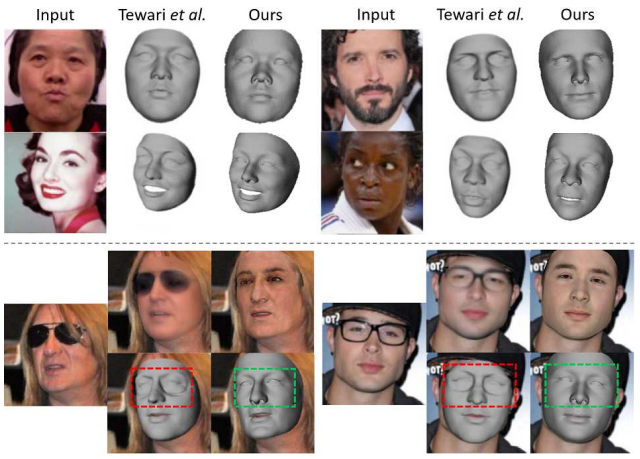

Figure 8. Comprison with Tewari *et al*. [48] (fine results). Top: results on different races. Bottom: results under occlusion. The images are from [48].

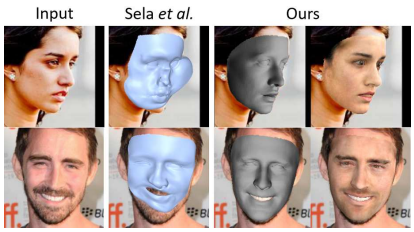

Figure 9. Comparison with Sela *et al*. [45] under large pose and challenging appearance.

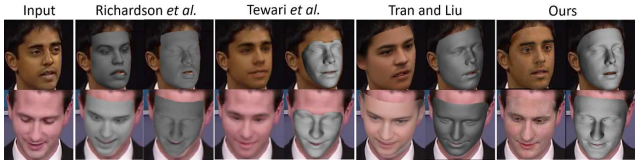

Figure 10. Comparison with Richardson *et al*. [37], Tewari *et al*. [49], and Tran and Liu [53]. Images are from [53].

Figure 11. Results on in-the-wild image sets. The leftmost bar chart shows the sorted value of confidence vector summation of each image in the set. Five images sampled from a set are shown in the middle with their confidence vector summations shown in the top left corner. The last two columns are our final results.

Table 4. Multi-image reconstruction errors on MICC rendered images with different aggregation strategies (see text for details).

Shape error mean	1.97±0.70
Shape averaging	1.78±0.59
Our S1: Global Aggr. with c^j	1.71±0.56
Our S2: Global Aggr. with $\sum_i \boldsymbol{c}_i^j$	1.70±0.55
Our S3: Max Conf. $j = \arg\max_j \sum \boldsymbol{c}_i^j$	1.71±0.50
Our S4: Elementwise Aggr. with $\boldsymbol{c}^j$	**1.67±0.54**

Table 5. Multi-image reconstruction error on the MICC dataset. We use same inputs and evaluation metric as in Table 2. "[34]-G" and "[34]-S" denote global and segment-based aggregation of our predicted shapes using the strategy of [34].

Method	Cooperative	Indoor	Outdoor	All
Shape averaging	1.66±0.52	1.66±0.46	1.69±0.53	1.62±0.51
[34]-G	1.68±0.57	1.67±0.47	1.73±0.53	1.65±0.55
[34]-S	1.68±0.58	1.67±0.48	1.72±0.52	1.65±0.55
Ours (S4)	**1.60±0.51**	**1.61±0.44**	**1.63±0.47**	**1.56±0.48**

S2, we sum all elements in the confidence vector $\boldsymbol{c}^j \in \mathbb{R}^{80}$ to get a global confidence for aggregation. For S3, we simply choose a single shape with largest confidence vector summation for error computation. For S4, we use our element-wise coefficient aggregation described in Sec. 5.1.

Table 4 shows that all shape aggregation methods including the naive shape averaging have a lower error than the mean of per-frame shape errors. Nevertheless, all our aggregation strategies yield better results than naive shape averaging, demonstrating the efficacy of our learning-based aggregation method. Among them, the element-wise coefficient aggregation (S4) performs best. We believe that with element-wise confidences, the network can exploit view difference for better reconstruction thus outstanding other strategies.

Figure 11 presents some examples of our confidence prediction on our test set (to ease presentation we show the confidence vector summation $\sum_i \boldsymbol{c}_i^j$ for each image). Our C-Net generally favors quality face images with frontal pose, high visibility, natural lighting etc. Occlusions like sunglasses, hat and hair decrease the confidence.

6.2.2 Comparison with Prior Art

To our knowledge, our method is the first one applying neutral networks for face reconstruction confidence prediction and aggregation. So here we compare with a heuristic strategy of Piotraschke and Blanz [34]. Table 5 shows that our method produced better results than shape averaging and [34] on the MICC dataset (we treat a sequence as an image set). The method of [34] underperformed. Its results are even slightly worse than shape averaging. We conjecture this is because [34] rely on the surface normal discrepancy with mean face to eliminate deficient reconstructions, yet our R-Net always produces a smooth, plausible face shape which renders their quality measurement ineffective.

7. Conclusions

We have proposed a CNN-based single-image face reconstruction method which exploits hybrid-level image information for weakly-supervised learning without ground-truth 3D shapes. Comprehensive experiments have shown that our method outperforms previous methods by a large margin in terms of both accuracy and robustness. We have also proposed a novel multi-image face reconstruction aggregation method using CNNs. Without any explicit label, our method can learn to measure image quality and exploit the complementary information in different images to reconstruct 3D faces more accurately.

References

[1] A. D. Bagdanov, A. Del Bimbo, and I. Masi. The florence 2d/3d hybrid face dataset. In *The Joint ACM Workshop on Human Gesture and Behavior Understanding*, pages 79–80, 2011. 1, 2, 6

[2] A. Bas, P. Huber, W. A. Smith, M. Awais, and J. Kittler. 3d morphable models as spatial transformer networks. In *International Conference on Computer Vision Workshop on Geometry Meets Deep Learning*, pages 904–912, 2017. 2

[3] A. Bas, W. A. Smith, T. Bolkart, and S. Wuhrer. Fitting a 3d morphable model to edges: A comparison between hard and soft correspondences. In *Asian Conference on Computer Vision (ACCV)*, pages 377–391, 2016. 2

[4] V. Blanz, A. Mehl, T. Vetter, and H.-P. Seidel. A statistical method for robust 3d surface reconstruction from sparse data. In *International Symposium on 3D Data Processing, Visualization and Transmission*, pages 293–300, 2004. 2

[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 187–194, 1999. 1, 2, 4

[6] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 25(9):1063–1074, 2003. 1

[7] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway. A 3d morphable model learnt from 10,000 faces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5543–5552, 2016. 2

[8] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision (ICCV)*, 2017. 4, 5

[9] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on graphics (TOG)*, 33(4):43, 2014. 2

[10] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):41, 2013. 1, 2

[11] 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 (TVCG)*, 20(3):413–425, 2014. 1, 2, 3, 6

[12] D. Chen, G. Hua, F. Wen, and J. Sun. Supervised transformer network for efficient face detection. In *European Conference on Computer Vision (ECCV)*, pages 122–138, 2016. 5

[13] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–26, 2017. 1, 2

[14] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 6, 7

[15] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt. Reconstruction of personalized 3d face rigs from monocular video. *ACM Transactions on Graphics (TOG)*, 35(3):28, 2016. 7

[16] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman. Unsupervised training for 3d morphable model regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 4, 6

[17] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing (IVC)*, 28(5):807–813, 2010. 6

[18] Y. Guo, J. Z. Zhang, J. Cai, B. Jiang, and J. Zheng. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 1, 2, 3

[19] T. Hassner. Viewing real-world faces in 3d. In *International Conference on Computer Vision (ICCV)*, pages 3607–3614, 2013. 2

[20] T. Hassner and R. Basri. Example based 3d reconstruction from single 2d images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2006. 2

[21] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4295–4304, 2015. 2

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3

[23] L. Hu, S. Saito, L. Wei, K. Nagano, J. Seo, J. Fursund, I. Sadeghi, C. Sun, Y.-C. Chen, and H. Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (TOG)*, 36(6):195, 2017. 1

[24] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 5

[25] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *International Conference on Computer Vision (ICCV)*, pages 1031–1039, 2017. 1, 6, 7

[26] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision (IJCV)*, 46(1):81–96, 2002. 4

[27] I. Kemelmacher-Shlizerman and S. M. Seitz. Face reconstruction in the wild. In *International Conference on Computer Vision (ICCV)*, pages 1746–1753, 2011. 2

[28] H. Kim, M. Zollhöfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt. Inversefacenet: Deep monocular inverse face rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4625–4634, 2018. 7

[29] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations (ICLR)*, 2015. 6

[30] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 1931–1939, 2015. 5

[31] F. Liu, R. Zhu, D. Zeng, Q. Zhao, and X. Liu. Disentangling features in 3d face shapes for joint face reconstruction and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5216–5225, 2018. 1, 2, 6, 7

[32] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 5

[33] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*, pages 296–301, 2009. 2, 3, 5

[34] M. Piotraschke and V. Blanz. Automated 3d face reconstruction from multiple images using quality measures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3418–3427, 2016. 2, 8

[35] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 497–500, 2001. 3

[36] R. Ramamoorthi and P. Hanrahan. A signal-processing framework for inverse rendering. In *Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 117–128, 2001. 3

[37] E. Richardson, M. Sela, and R. Kimmel. 3d face reconstruction by learning from synthetic data. In *International Conference on 3D Vision (3DV)*, pages 460–469, 2016. 1, 2, 7

[38] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5553–5562, 2017. 1, 2

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

[40] J. Roth, Y. Tong, and X. Liu. Unconstrained 3d face reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2606–2615, 2015. 2

[41] J. Roth, Y. Tong, and X. Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4197–4206, 2016. 2

[42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 5

[43] S. Saito, T. Li, and H. Li. Real-time facial segmentation and performance capture from rgb input. In *European Conference on Computer Vision (ECCV)*, pages 244–261, 2016. 2, 4

[44] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 4

[45] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1585–1594, 2017. 1, 2, 7

[46] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. SfSNet: learning shape, reflectance and illuminance of faces in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6296–6305, 2018. 1, 2

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

[48] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2549–2559, 2018. 1, 2, 3, 4, 5, 7

[49] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, and C. Theobalt. MoFa: model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 1274–1283, 2017. 1, 2, 3, 4, 7

[50] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, 2016. 1, 4

[51] A. T. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1493–1502, 2017. 1, 2, 6

[52] A. T. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[53] L. Tran and X. Liu. Nonlinear 3d face morphable model. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7346–7355, 2018. 1, 2, 7

[54] D. Vlasic, M. Brand, H. Pfister, and J. Popović. Face transfer with multilinear models. *ACM transactions on graphics (TOG)*, 24(3):426–433, 2005. 2

[55] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 4, page 7, 2017. 5

[56] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *International Conference on Automatic Face and Gesture Recognition*, pages 211–216, 2006. 1, 2

[57] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, 2016. 1, 2, 5, 6, 7

[58] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the

wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 787–796, 2015. 2

[59] S. Zulqarnain Gilani and A. Mian. Learning from millions of 3d scans for large-scale 3d face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1896–1905, 2018. 1