Coarse-to-Fine 3D Face Reconstruction

Leonardo Galteri*
University of Florence
leonardo.galteri@unifi.it

Claudio Ferrari*
University of Florence
claudio.ferrari@unifi.it

Giuseppe Lisanti
University of Bologna
giuseppe.lisanti@unibo.it

Stefano Berretti
University of Florence
stefano.berretti@unifi.it

Alberto Del Bimbo
University of Florence
alberto.delbimbo@unifi.it

Abstract

Reconstructing accurate 3D shapes of human faces from a single 2D image is a highly challenging Computer Vision problem that was studied for decades. Statistical modeling techniques, such as the 3D Morphable Model (3DMM), have been widely employed because of their capability of reconstructing a plausible model grounding on the prior knowledge of the facial shape. However, most of them derive a and smooth approximation of the real shape, without accounting for the surface details. In this work, we propose an approach based on a Conditional Generative Adversarial Network (CGAN) for refining the reconstruction provided by a 3DMM. The latter is represented as a three-channel image, where the pixel intensities represent, respectively, the depth and the azimuth and elevation angles of the surface normals. The network architecture is an encoder-decoder, which is trained progressively, starting from the lower-resolution layers; this technique allows a more stable training, which led to the generation of high quality outputs even when high-resolution images are fed during the training. Experimental results show that our method is able to produce detailed realistic reconstructions and obtain lower errors with respect to the 3DMM. Finally, a comparison with a state-of-the-art solution evidences competitive performance and a clear improvement in the quality of the generated models.

1. Introduction

The idea of deriving 3D information from 2D images using computer vision techniques is a research topic with a quite long tradition that dates back to ’80 [18]. However, estimating the 3D geometry from single or multiple images under general conditions, where no *a priori* knowledge is available about the imaged scene and the object of interest is a very challenging task. Hence, to make the problem solvable to some extent, priors are usually assumed. Face reconstruction is a particular case where such solution showed its viability. In this case, a 3D Morphable Model (3DMM) is used as a shape prior of the face; this statistical model limits the shape of the reconstructed face to the combination, according to a set of parameters, of an average face model and some deformation components. However, the results of such reconstructions appear generally over-smoothed, lacking of fine details.

To move a step further from the above solutions, a promising idea is that of first deriving a smooth approximation of the face shape, then add local details to it. A work that followed this idea, while keeping general in the assumptions, has been proposed in [28]. In that work, a *foundation shape* is generated by a deep learning based 3DMM [27], which is then refined by adding details generated by an encoder-decoder network. This idea brings quite naturally to the use of Generative Adversarial Networks (GANs) [9]; in the current literature of deep learning solutions, GANs have proved their capability of generating synthetic image data that are hardly distinguishable from real one [1]. Thanks to this specific prerogative, they have found successful application in tasks such as image super-resolution [17], image enhancement [21], image restoration [29], etc.

Grounding on the above considerations, in this work we propose a coarse-to-fine approach to reconstruct a detailed 3D face model from a single image. The approach develops on the idea of first deriving an approximated 3D shape by fitting a 3DMM to an image of the face. Then, such shape is refined using a Conditional Generative Adversarial Network (CGAN). To this end, the 3D shape is represented as a three-channel image, where the three channels are the depth, azimuth and elevation values of the vertices of the model. The CGAN is designed following the encoder-

*These authors contributed equally to this work.

decoder paradigm, which is trained progressively starting from the lower-resolution layers. Experimental results show that our method is able to produce reconstructions with realistic details and lower reconstruction errors with respect to the 3DMM. A comparison with a state-of-the-art solution reveals that the proposed approach is highly competitive, showing an evident superiority in generating detailed and realistic reconstructions. In summary, our contributions are as follows:

- We design an effective solution to reconstruct a realistic 3D face model from a single face image;
- We model the 3D face refinement step as the problem of training, with progressive growing, an encoder-decoder based Conditional GAN;
- We demonstrate that the 3D face obtained by using the proposed solution better approximates a realistic face with respect to state-of-the-art solutions.

The rest of the paper is organized as follows: in Section 2, we summarize the closely related work on 3D face reconstruction; in Section 3, we introduce the 3D Morphable Shape Model and illustrate how this serves to derive training image data with depth, azimuth and elevation channels; the GAN architecture we have designed and its training are detailed in Section 4; experimental results are presented in Section 5; finally, conclusions and future research directions are sketched in Section 6.

2. Related work

In the general case, reconstructing a 3D face model from 2D images is extremely challenging so that most of the existing solutions rely on some assumptions in the form of prior knowledge. Keeping aside methods that do not resort to any problem simplification, and that thus result in poor reconstructions, in the following, we briefly present the relevant literature on *model-based 3D face reconstruction*.

Methods belonging to this category keep the assumptions general and use priors in the form of a prototypical face model, thus reconstructing smooth shapes that usually lack of fine details. The most widely recognized examples in this category are the 3DMM based fitting methods, as originally proposed in [2], and subsequently refined in [23]. Also these methods emphasized more the appeal of rendered face images, rather than the quantitative evaluation of the accuracy of the reconstructed face shape. Among the 3DMM variants, the most successful was proposed in [19] that improved the 3DMM into the Basel Face Model with higher shape and texture accuracy and less correspondence artifacts. In [4, 3] an in-the-wild 3DMM was proposed by combining a statistical model of shape, which describes both identity and expression, with an in-the-wild texture model. Some other techniques fit the 3DMM surface to detected facial landmarks rather than to face intensities directly. These include solutions designed for videos, like in [24, 11], and the CNN based approaches of [14, 32].

An emerging trend in this category of methods is that of defining alternative solutions that are general but accurate. In most of the cases, this is obtained by applying a refinement step that adds details to an initially approximated shape; deep learning solutions are mostly used for this second step. Following this approach, in [22] a rather shallow network is trained on synthetic shapes with an iterative process, and facial details are also added by training an end-to-end system to additionally estimate shape-from-shading (SfS). Other methods in this category used deep networks by emphasizing more the aspect of estimating 3D shapes from unconstrained photos [27, 6, 13, 26]. These methods estimate shapes that are highly invariant to viewing conditions, but provide only coarse surface details.

We are not aware of methods that use GANs, either conditional or not, to generate detailed 3D models of the face starting from a raw estimation of the shape geometry. However, in designing our reconstruction solution, we leveraged on classical GAN-based methods applied to RGB images; therefore, in the following, we refer some relevant work that used GANs for image related tasks. GANs were first proposed in [9], and subsequently modified in a series of works, for improved training [25], or extended to unsupervised learning as with the Deep Convolutional GANs (DCGANs) [21]. Since their introduction, GANs have rapidly established as state-of-the-art solutions to improve the quality of generated 2D images in a variety of image synthesis tasks. In [12] conditional GANs were investigated as a general-purpose solution for image-to-image translation problems. These networks not only learn the mapping from input to output image, but also learn a loss function to train this mapping. This was extended in [31] for learning how to translate an image from a source domain to a target domain in the absence of paired examples. Though the methods above have been inspiring for our proposed solution, they are tailored for generating 2D RGB images, while we generate a three-channel image based on depth. azimuth and elevation. Despite our channels are disposed according to the same grid-like structure used for RGB images, the information carried out by each image channel is not the same, thus posing new and challenging problems about how to train GANs in a robust and effective way.

3. 3D reconstruction through 3DMM

Given a face image, we first aim at estimating an approximated 3D reconstruction exploiting the 3D Morphable Model (3DMM) technique; then, we represent the reconstructed geometry by a three channel 2D image, where the channels are the *depth*, *azimuth* and *elevation* angles of the surface normals of the reconstructed model. To obtain the models, we employed two 3DMM based solutions proposed in [7] and [27], called *Dictionary Learning*-3DMM (*DL-3DMM*) and *Deep3DMM*, respectively. The first adapts the 3DMM to a face image exploiting 2D-3D facial landmark pairs. This method can estimate the face shape fairly accurately even in the presence of strong facial expressions. The Binghamton University 3D Facial Expression dataset (BU-3DFE) [30] was used to build the average model and learn the deformation components. The second instead exploits a deep CNN to regress the 3DMM parameters directly from RGB images; this method does not model facial expressions but it is robust to the identity *i.e.* different images of the same individual generate the same parameters.

Actually, any other 3D face modeling technique could have fit our purposes; in fact, the proposed method aims to refine the reconstruction given as input. It thus results rather independent from the approximated model that is provided, and any method can be used in practice. Nonetheless, better input reconstructions lead to more accurate refined models.

3.1. Images in depth, azimuth and elevation format

The 2D representation of the reconstruction used in this work is inspired by the approach in [8]. Differently from the classic gray-scale depth image, this format transforms a 3D mesh into an RGB image. The first channel contains the *depth* value *i.e.* z coordinate of each 3D vertex; the other two contain, respectively, the *elevation* (or inclination, or polar angle) and *azimuth* angles of the normal vectors computed at each 3D vertex, represented in spherical coordinates. An example of the proposed representation based on *depth*, *azimuth* and *elevation* is shown in Figure 1.

The subsequent step in the image creation is the projection of the depth, azimuth and elevation values on the image plane, and rescaling of the values in the range $[0, 255]$. This procedure must be applied consistently both for the coarsely reconstructed 3DMM and the ground-truth so that the generated images are aligned. To this aim, we estimate an orthographic projection matrix $\mathbf{P} \in \mathbb{R}^{2\times 3}$ from 2D and 3D landmark correspondences. The 2D landmarks, which are detected on the RGB face images exploiting the method of [5], are both used to fit and project the 3DMM and, independently, estimate the projection matrix for the ground-truth model so as to account for the relative difference in the models' scale. The same procedure is applied for the Deep3DMM; in this latter case, the parameters to deform the 3DMM have been directly regressed from the RGB image. Thus, the landmarks are only used to estimate the projection matrix and map the 3D model onto the image plane. The projections are finally used to map the depth, azimuth and elevation values on the image plane and build the three-channel images of the 3DMM and ground-truth pair.

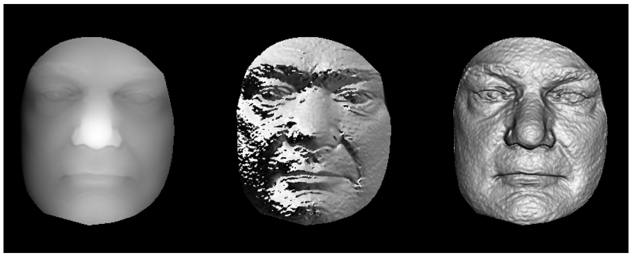

Figure 1. Representation of the 3D face model by a three-channel image. For visualization purposes, the depth, azimuth and elevation channels are shown as individual images, from left to right.

4. Deep generative refinement

The reconstruction described in Section 3 is usually obtained as a modification of an average, smooth, model; consequently, its surface usually lacks of fine grained details. In order to obtain such detailed reconstruction from a single RGB face image, we employ a Conditional Generative Adversarial Network (CGAN). Differently from classic CGAN, the architecture is trained progressively as described in [15]. Conditional GANs have been specifically designed for image-to-image translation, and this makes them particularly suited for our purpose. In our solution, indeed, the generator G aims at translating the approximated reconstruction, the *condition*, to the target domain, the *ground-truth*. The discriminator D, instead, has the objective of discriminating ground-truth images from the synthetically generated ones. Formally, the training procedure is supervised as the dataset contains paired images of the approximated model x and the correspondent detailed model y (*i.e.*, the ground-truth). The objective of conditional GANs is to learn a distribution of real detailed models given input conditions as:

$$\min_G \max_D \mathbb{E}_{(x,y)}\left[\log D(x,y)\right] + \mathbb{E}_x\left[\log\left(1 - D(x, G(x))\right)\right] \ . \tag{1}$$

In our particular case, x and y are the proposed image representations of Section 3.1 for, respectively, the 3DMM reconstruction and the ground truth model. The proposed solution is conditioned on x.

We aim to exploit the benefits of progressive growth of GANs [15] in a conditional context. For this reason, we design our generator as an encoder-decoder to transform a 3DMM into a high quality detailed face model. To ensure further stability to the training of our framework, we employ the improved version of Wasserstein GAN [10]. The set of weights for the discriminator are learned by minimizing the objective function:

$$\mathcal{L}_D = D(x,y) - D(x, G(x)) + \lambda(||\nabla_{\hat{x}} D(x, \hat{x})||_2 - 1)^2 \ , \tag{2}$$

where x and y are, as in Eq. (1), the proposed image representations of the 3DMM and the ground truth model, respectively, and $\hat{x}$ is sampled uniformly between pairs of points belonging to the real and the generator distribution.

Given the fact that our training is supervised, *i.e.*, each 3DMM is paired with the relative ground-truth image, we can define the loss for the generator as a combination of two contributions:

$$\mathcal{L}_G = L_p(y, G(x)) + \kappa L_{adv}(G(x)) \ , \qquad (3)$$

where

$$L_p(y, G(x)) = ||y - G(x)||_p \ ,$$

represents the pixel loss, and

$$L_{adv}(x, G(x)) = D(x, G(x)) \ ,$$

is the adversarial loss. The architectures for the components in our conditional GAN resemble the ones in [15]. The encoder part of the generator and the discriminator share the same architecture; the decoder differs from the encoder in as much as the down-sampling layers are substituted with up-sampling layers. We progressively train G and D starting from 4×4 down-scaled images up to 256×256, expanding G in both directions simultaneously, encoder and decoder, as shown in Figure 2.

5. Experiments

We performed a set of experiments in order to assess the validity of the proposed approach. In particular, we show how the proposed method effectively improves upon the approximated reconstruction provided by the 3DMM; further, we qualitatively compare our reconstruction results with the state-of-the-art solution proposed by [12].

All the experiments have been carried out on the Face Recognition Grand Challenge dataset (FRGC) [20]. In particular, the FRGC dataset has been split and used both for training and for testing. The FRGC dataset includes 4,007 scans of 466 individuals acquired with frontal view from the shoulder level, with very small pose variations. About 60% of the faces have neutral expression, while the others show spontaneous expressions of disgust, happiness, sadness, and surprise. Scans are given as matrices of 3D points of size 480×640, with a binary mask indicating the valid points of the face (about 40K on average). RGB images of the face are also available and aligned with the matrix of 3D points.

Data Augmentation - We augmented the training data by generating novel poses as follows: given a 3D face model from the training set (3DMM and ground-truth pair), we generated a random rotation matrix $\mathbf{R}_{rand} \in \mathbb{R}^{3\times3}$, with rotation angles (yaw, pitch, roll) in the range $[\pm45, \pm20, \pm20]$, and used it to build the orthographic projection matrix $\mathbf{P}$ using a fixed 2D translation vector $\mathbf{t} \in \mathbb{R}^2$ and scale parameters matrix $\mathbf{S} \in \mathbb{R}^{2\times3}$. We then used $\mathbf{P}$ to project the pose-augmented models onto the image plane.

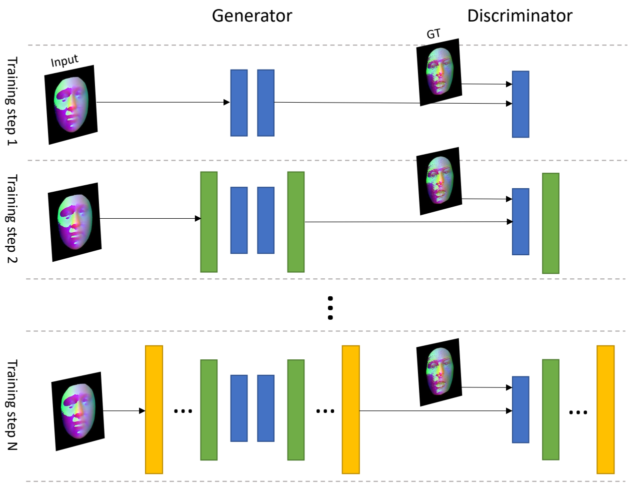

Figure 2. Schema of the proposed CGAN framework.

This process is repeated 5 times for each 3D model, which results in more than $14,000$ images. During training, pixel values of each channel have been normalized in the range $[-1, 1]$. To further strengthen the procedure, we randomly crop and pad the images online during training.

Training Details - The weights of the proposed architecture are initialized using a truncated normal distribution. Each resolution in our architecture has been separately trained for $10,000$ iterations with a batch size of 4 (*e.g.*, about 3 epochs with $14,000$ training samples). We train our networks using the Adam algorithm of [16], with a learning rate of 10^{-5}. We empirically found that a reasonable value for κ of Eq. (3) is $5 * 10^{-5}$.

Evaluation protocol and metric - We randomly split the FRGC individuals into three parts; the first $2/3$ are used for training, for a total of 310 individuals; the remaining $1/3$ of individuals and the relative models are used for test. In this way, we can ensure that an identity used for test has never been seen during the training. To quantitatively evaluate our approach, we employed the *Mean Absolute Error* (MAE) measure computed between the ground-truth depth image y and the estimated depth image $G(x)$.

5.1. Results

In this section we report qualitative and quantitative reconstruction results of our approach. Table 1 reports MAE computed with the two 3DMM models and the refined ones with respect to the ground-truth. Evidently, our refinement produces more accurate reconstructions, effectively improving upon both the 3DMM models. This holds for all the three channels. In Figure. 3 we also report some qualitative heatmaps comparing the 3DMM and refined models; the heatmaps consistently reveal a reduced general reconstruction error.

In order to present a more complete evaluation, we

Table 1. Mean absolute error computed on the test set of the FRGC v2.0 dataset. Results are shown for each channel separately. The average for the three channels is also reported.

	Depth		**Azimuth**		**Elevation**		**Avg**	
	Coarse	*Refined*	*Coarse*	*Refined*	*Coarse*	*Refined*	*Coarse*	*Refined*
DL-3DMM	0.110 ± 0.032	$\mathbf{0.084 \pm 0.026}$	0.183 ± 0.029	$\mathbf{0.151 \pm 0.022}$	0.159 ± 0.032	$\mathbf{0.132 \pm 0.019}$	0.150 ± 0.031	$\mathbf{0.122 \pm 0.023}$
Deep3DMM [27]	0.133 ± 0.039	$\mathbf{0.062 \pm 0.020}$	0.188 ± 0.027	$\mathbf{0.141 \pm 0.023}$	0.162 ± 0.027	$\mathbf{0.113 \pm 0.021}$	0.161 ± 0.031	$\mathbf{0.105 \pm 0.022}$

trained a recent CGAN architecture, namely *Pix2Pix* [12], to refine the approximated model considered in this work. From an architectural point of view, it adopts the U-Net as a generator, and it embodies a patch discriminator. We trained Pix2Pix on our 256×256 training images with the default settings for 20 epochs. Figure 4 reports some qualitative examples; we can immediately appreciate that, compared to our approach, the Pix2Pix method clearly introduces far more noise in the reconstructed models. Moreover, Figure 4 (a) highlights how the proposed method is able to effectively maintain the identity-specific traits of the subject portrayed, while Figure 4 (b) shows robustness to facial expressions, which instead wreck the Pix2Pix reconstructions.

6. Discussion and Future Work

In this work, we proposed an approach based on a Conditional Generative Adversarial Network (CGAN) for refining the reconstruction of face images provided by a 3DMM. The reconstruction is represented as an RGB image, where the pixel intensities represent the depth, azimuth and elevation values of the 3D model' vertices. We proposed an encoder-decoder architecture, which is trained progressively; this technique allowed a more stable training, which led to the generation of pleasant images even at higher resolutions. However, our approach is not exempt from limitations; first, if the shape of the 3DMM differs too much with respect to the ground-truth ones, the network might eventually overfit the data in the attempt of transforming the shapes and thus lose its generalization capabilities or, on the contrary, fail in generating pleasant outputs. Another limitation is that if we want to change the input 3D reconstruction model to be refined, a new instance of the network has to be trained from scratch. Even though the training procedure is rather fast and does not require as many images as other architectures, we still might want to investigate if a feasible solution to make it independent from the 3D input can be found. Overall, we demonstrated that a progressive CGAN can be effectively trained on distinctive image data and employed to generate highly detailed 3D surfaces from their smoother counterparts. The solutions that have been investigated and presented in this manuscript actually represent only a small portion of the possible alternatives, for which there is a lot of room for improvements. As an example, we will further investigate how to exploit the correlations that occur between the three channels encoding surface geometric properties to our advantage.

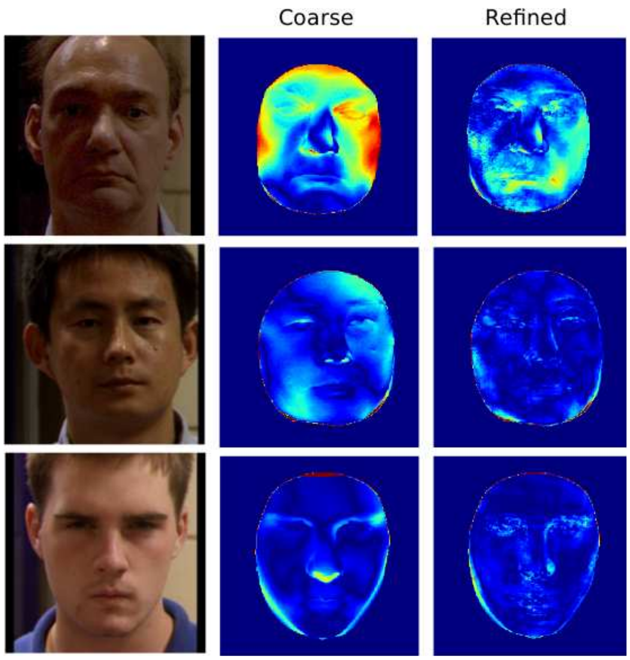

Figure 3. Absolute error heat maps with respect to the GT models.

Acknowledgements

The Titan Xp GPUs used for this research were donated by the NVIDIA Corporation.

References

[1] David Berthelot, Tom Schumm, and Luke Metz. BEGAN: Boundary equilibrium generative adversarial networks. *CoRR*, abs/1703.10717, 2017. 1

[2] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *ACM Conf. on Computer Graphics and Interactive Techniques*, 1999. 2

[3] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou. 3D face morphable models "in-the-wild". In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5464–5473, July 2017. 2

[4] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou. Large scale 3D morphable models. *Int. Journal of Computer Vision*, 126(2):233–254, April 2017. 2

[5] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *Int. Conf. on Computer Vision (ICCV)*, 2017. 3

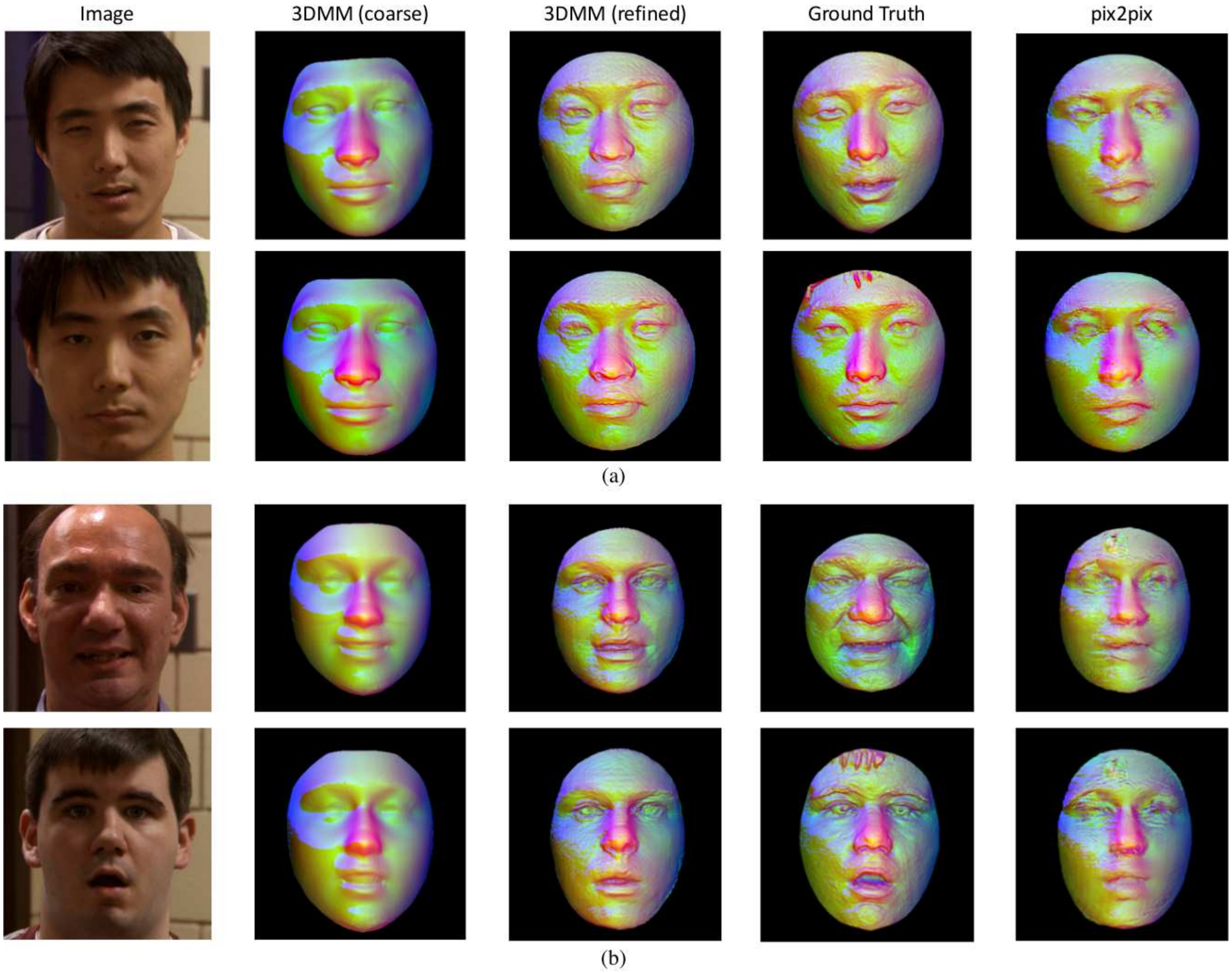

Figure 4. Qualitative results on two identities of the FRGC test set. In (a) the 3DMM model of [27], (b) DL-3DMM

[6] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3D face reconstruction with deep neural networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1503–1512, July 2017. 2

[7] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo. A dictionary learning-based 3D morphable shape model. *IEEE Trans. on Multimedia*, 19(12):2666–2679, Dec 2017. 3

[8] Syed Zulqarnain Gilani, Ajmal Mian, and Peter Eastwood. Deep, dense and accurate 3D face correspondence for generating population specific deformable models. *Pattern Recognition*, 69:238–250, 2017. 3

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 1, 2

[10] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein GANs. *arXiv 1704.00028*, 2017. 3

[11] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W.J. Christmas, M. Ratsch, and J. Kittler. A multiresolution 3D morphable face model and fitting framework. In *Int. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016. 2

[12] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, July 2017. 2, 4, 5

[13] 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 *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 1031–1039, Oct 2017. 2

[14] A. Jourabloo and X. Liu. Large-pose face alignment via CNN-based dense 3D model fitting. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4188–4196, June 2016. 2

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *arXiv 1710.10196*, 2017. 3, 4

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv 1412.6980*, 2014. 4

[17] C. Ledig, L. Theis, F. Huszr, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, July 2017. 1

[18] David G Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial intelligence*, 31(3):355–395, 1987. 1

[19] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. In *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, pages 296–301, 2009. 2

[20] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE Workshop Face Recognition Grand Challenge Experiments*, 2005. 4

[21] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. 1, 2

[22] E. Richardson, M. Sela, and R. Kimmel. 3D face reconstruction by learning from synthetic data. In *IEEE Int. Conf. on 3D Vision (3DV)*, pages 460–469, Oct 2016. 2

[23] S. Romdhani and T. Vetter. Efficient, robust and accurate fitting of a 3d morphable model. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 59–66, Oct 2003. 2

[24] Shunsuke Saito, Tianye Li, and Hao Li. Real-time facial segmentation and performance capture from RGB input. In *European Conf. Computer Vision (ECCV)*, pages 244–261. Springer International Publishing, 2016. 2

[25] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Int. Conf. on Neural Information Processing Systems (NIPS)*, pages 2234–2242, 2016. 2

[26] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. Jacobs. SfSNet: Learning shape, reflectance and illuminance of faces in the wild. *ArXiv 1712.01261*, 2017. 2

[27] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5163–5172, July 2017. 1, 2, 3, 5, 6

[28] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. Extreme 3D face reconstruction: Looking past occlusions. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[29] P. Wang, H. Zhang, and V. M. Patel. Generative adversarial network-based restoration of speckled sar images. In *IEEE Int. Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–5, Dec 2017. 1

[30] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato. A 3D facial expression database for facial behavior research. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2006. 3

[31] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 2242–2251, Oct 2017. 2

[32] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3D solution. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, June 2016. 2