

Disentangling Features in 3D Face Shapes for Joint Face Reconstruction and Recognition

Feng Liu^{*}, Ronghang Zhu^{*}, Dan Zeng^{*}, Qijun Zhao^{*}, and Xiaoming Liu^{*}
College of Computer Science, Sichuan University
Department of Computer Science and Engineering, Michigan State University

Abstract

This paper proposes an encoder-decoder network to disentangle shape features during 3D face reconstruction from single 2D images, such that the tasks of reconstructing accurate 3D face shapes and learning discriminative shape features for face recognition can be accomplished simultaneously. Unlike existing 3D face reconstruction methods, our proposed method directly regresses dense 3D face shapes from single 2D images, and tackles identity and residual (i.e., non-identity) components in 3D face shapes explicitly and separately based on a composite 3D face shape model with latent representations. We devise a training process for the proposed network with a joint loss measuring both face identification error and 3D face shape reconstruction error. To construct training data we develop a method for fitting 3D morphable model (DMM) to multiple 2D images of a subject. Comprehensive experiments have been done on MICC, BU DFE, LFW and YTF databases. The results show that our method expands the capacity of DMM for capturing discriminative shape features and facial detail, and thus outperforms existing methods both in 3D face reconstruction accuracy and in face recognition accuracy.

1. Introduction

3D face shapes reconstructed from 2D images have been proven to benefit many tasks, e.g., face alignment or facial landmark localization [18, 41], face animation [9, 13], and face recognition [5, 12]. Many prior work have been devoted to reconstructing 3D face shapes from a single 2D image, including shape from shading (SFS)-based methods [14, 20], 3D morphable model (DMM) fitting-



Figure 1. Comparison between the learning process of (a) existing methods and (b) our proposed method. GT denotes Ground Truth. (d) and (e) are 3D face shapes and disentangled identity shapes reconstructed by our method for the images in (c) from LFW [15].

based methods [4, 5], and recently proposed regression-based methods [22, 23]. These methods mostly aim to recover 3D face shapes that are loyal to the input 2D images or retain as much facial detail as possible (see Fig. 1). Few of them explicitly consider the identity-sensitive and identity-irrelevant features in the reconstructed 3D faces. Consequently, very few studies have been reported about recognizing faces using the reconstructed 3D face either by itself or by fusing with legacy 3D face recognition [5, 33].

Using real 3D face shapes acquired by 3D face scanners

This work is supported by the National Key Research and Development Program of China (2017YFB0802300) and the National Natural Science Foundation of China (61773270, 61703077).

Corresponding author. Email: qjzhao@scu.edu.cn.

for face recognition, on the other hand, has been extensively studied, and promising recognition accuracy has been achieved [6, 11]. Apple recently claims to use 3D face matching in its iPhone X for cellphone unlock [1]. All of these prove the discriminative power of 3D face shapes. Such a big performance gap between the reconstructed 3D face shapes and the real 3D face shapes, in our opinion, demonstrates that existing 3D face reconstruction methods seriously undervalue the identity features in 3D face shapes. Taking the widely used DMM fitting based methods as example, their reconstructed 3D faces are constrained in the limited shape space spanned by the pre-determined bases of DMM, and thus perform poorly in capturing the features unique to different individuals [39].

Inspired by the latest development in disentangling feature learning for 3D face recognition [26, 34], we propose to disentangle the identity and non-identity components of

3D face shapes, and more importantly, fulfill *reconstructing accurate 3D face shapes* loyal to input 2D images and *learning discriminative shape features* effective for face recognition in a *joint* manner. These two tasks, at the first glance, seem to contradict each other. On one hand, face recognition prefers identity-sensitive features, but not every detail on faces; on the other hand, 3D reconstruction attempts to recover as much facial detail as possible, regardless whether the detail benefits or distracts facial identity recognition. In this paper, however, we will show that by exploiting the ‘contradictory’ objectives of recognition and reconstruction, we are able to *disentangle identity-sensitive features from identity-irrelevant features in 3D face shapes*, and thus simultaneously robustly recognize faces with identity-sensitive features and accurately reconstruct 3D face shapes with both features (see Fig. 1).

Specifically, we represent 3D face shapes with a composite model, in which identity and residual (i.e., non-identity) shape components are represented with separate latent variables. Based on the composite model, we propose a joint learning pipeline that is implemented as an encoder-decoder network to disentangle shape features during reconstructing 3D face shapes. The encoder network converts the input 2D face image to identity and residual latent representations, from which the decoder network recovers its 3D face shape. The learning process is supervised by both reconstruction loss and identification loss, and based on a set of 3D face images with labelled identity information and corresponding 3D face shapes that are obtained by an adapted multi-image DMM fitting method. Comprehensive evaluation experiments prove the superiority of the proposed method over existing baseline methods in both 3D face reconstruction accuracy and face recognition accuracy. Our main contributions are summarized below.

(i) We propose a method which for the first time explicitly optimizes face recognition and 3D face reconstruction

simultaneously. The method achieves state-of-the-art 3D face reconstruction accuracy via joint discriminative feature learning and 3D face reconstruction.

(ii) We devise an effective training process for the proposed network that can disentangle identity and non-identity features in reconstructed 3D face shapes. The network, while being pre-trained by DMM-generated data, can surmount the limited 3D shape space determined by the DMM bases, in the sense that it better captures identity-sensitive and identity-irrelevant features in 3D face shapes.

(iii) We leverage the effectiveness of disentangled identity features in reconstructed 3D face shapes for improving face recognition accuracy, as being demonstrated by our experimental results. This further expands the application scope of 3D face reconstruction.

2. Related Work

In this section, we review existing work that is closely related to our work from two aspects: 3D face reconstruction for recognition and Convolutional Neural Network (CNN) based 3D face reconstruction.

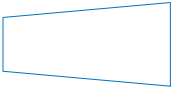
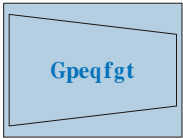
D Face Reconstruction for Recognition. 3D face reconstruction was first introduced for recognition by Blanz and Vetter [5]. They reconstructed 3D faces by fitting DMM to 2D face images, and used the obtained DMM parameters as features for face recognition. Their employed DMM fitting method is essentially an image-based analysis-by-synthesis approach, which does not consider the features unique to different individuals. This method was recently improved by Tran et al. [33] via pooling the DMM parameters of the images of the same subject and using a CNN to regress the pooled parameters. They experimentally proved the improved discriminative power of their obtained DMM parameters.

Instead of using DMM parameters for recognition, Liu et al. [23] proposed to recover pose and expression normalized 3D face shapes directly from 2D face landmarks via cascaded regressors and match the reconstructed 3D face shapes via the iterative closest point algorithm for face recognition. Other researchers [31, 36] utilized the reconstructed 3D face shapes for face alignment to assist extracting pose-robust features.

To summarize, *existing methods, when reconstructing 3D face shapes, do not explicitly consider recognition performance*. In [23] and [33], even though the identity of 3D face shapes in the training data is stressed, respectively, by pooling DMM parameters and by normalizing pose and expression, their methods of learning mapping from 2D images to 3D face shapes are *unsupervised* in the sense of utilizing identity labels of the training data (see Fig. 1).

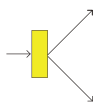
CNN-based 3D Face Reconstruction. Existing CNN-based 3D face reconstruction methods can be divided into two categories according to the way of representing 3D

4F"Kocig



Rtgflevgt"Kfgp^Wlv{

$\frac{k_{prwv}}{4F_{locig}}$



face images, and different residual D shape components that are unique to each of the subject's D images.

The DMM represents a D face shape as

(6)

where \mathbf{I} and \mathbf{E} are, respectively, the identity and expression shape bases, and \mathbf{a} and \mathbf{b} are the corresponding coefficients. In this paper, we use the shape bases given by the Basel Face Model [25] as \mathbf{I} , and the blendshape bases in FaceWarehouse [8] as \mathbf{E} .

To fit the DMM to \mathbf{I} images of a subject, we attempt to minimize the difference between \mathbf{I} , the landmarks detected on the images, and \mathbf{L} , the landmarks obtained by projecting the estimated D face shapes onto the images, under the constraint that all the images of the subject share the same \mathbf{L} . \mathbf{L} is computed from the estimated D face shape (let \mathbf{v}_i denote the vertices in \mathbf{I} corresponding to the landmarks) by $\mathbf{L} = \mathbf{P}(\mathbf{I} - \mathbf{a} - \mathbf{b})$, where \mathbf{P} is the scale factor, \mathbf{I} is the orthographic projection, and \mathbf{R} and \mathbf{T} are the rotation matrix and translation vector in D space. Mathematically, our multi image DMM fitting optimizes the following objective:

(7)

We solve the optimization problem in Eq. (7) in an alternating way. As an initialization, we set both \mathbf{a} and \mathbf{b} to zero. We first estimate the projection parameters \mathbf{P} , then expression parameters \mathbf{b} , and lastly identity parameters \mathbf{a} . When estimating one of the three sets of parameters, the rest two sets of parameters are fixed as they are. The optimization is repeated until the objective function value does not change. We have typically found this to converge within seven iterations.

3.3.3 Training Process

With the prepared training data, we train our encoder-decoder network in three phases. In Phase I, we train the encoder by setting the target latent representations as \mathbf{z}_i and \mathbf{z}_e and using Euclidean loss. In Phase II, we train the decoder for the identity and residual components separately. In Phase III, the end-to-end joint training is conducted based on the pre-trained encoder and decoder. Considering that the network already has good performance in reconstruction after pre-training, we first lay more emphasis on recognition in the joint loss function by setting λ to 1. When the loss function gets saturated (usually within 100 epochs), we continue the training by updating λ to 0.1. The joint training concludes in about another 100 epochs.

It is worth mentioning that the recovered DMM parameters are directly used as the latent representations during

pre-training. This provides a good initialization for the encoder-decoder network, but limits the network to the capacity of the pre-determined DMM bases. The joint training in Phase III alleviates such limitation by utilizing the identification loss as a complementary supervisory signal to the reconstruction loss. As a result, the learnt encoder-decoder network can better disentangle identity from non-identity information in D face shapes, and thus enhance face recognition accuracy without impairing the D face reconstruction accuracy.

4. Experiments

Two sets of experiments have been done to evaluate the effectiveness of the proposed method in D face reconstruction and face recognition. The MICC [2] and BU DFE [38] databases are used for experiments of D face reconstruction, and the LFW [15] and YTF [35] databases are used in face recognition experiments. Next, we report the experimental results¹.

4.1. D Shape Reconstruction Accuracy

The D face reconstruction accuracy is assessed by using D Root Mean Square Error (RMSE) [33], defined as $RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N ||\mathbf{I}_i - \mathbf{R}_i||^2}$ where N is the total number of testing samples, \mathbf{I}_i and \mathbf{R}_i are the ground truth and reconstructed D face shape of the i^{th} testing sample. To compute the RMSE, the reconstructed D faces are first aligned to ground truth via Procrustes global alignment based on D landmarks as suggested by [3], and then cropped at a radius of 0.1 around the nose tip.

We compare our method with four state-of-the-art D face reconstruction methods, DDFA [42], DMM-CNN [33], D shape regression based (DSR) method [23], and VRN [16]. Among them, the first two methods reconstruct D face shapes via estimating DMM parameters, while the other two directly regress D face shapes from either landmarks or D images. DMM-CNN method is the only existing method that takes into consideration the discriminative power of the estimated DMM parameters. DSR method generates pose and expression normalized D face shapes that are believed to be more beneficial to face recognition. For those methods that need facial landmarks on D images, we use the method in [7] to automatically detect the landmarks.

Results on MICC. The MICC database contains three challenging face videos and ground-truth D models acquired using a structured-light scanning system for each of 10 subjects. The videos span the range of controlled indoor to unconstrained outdoor settings. The outdoor videos are very challenging due to the uncontrolled lighting conditions. In this experiment, we randomly select

¹More experimental results are provided in the supplementary material.

Table 1. D face reconstruction accuracy (RMSE) under different yaw angles on the BU DFE database.

Method	Avg.				
VRN					
DDFA					
DMM-CNN	-	-	-	-	-
DSR					
Proposed					

Figure 5. Reconstruction results for three MICC subjects. The first column shows the input images, and the rest columns show the reconstructed D shapes that have the same expression as the input images, using the methods of VRN [16], DDFA [42], DMM-CNN [33], DSR [23] and the proposed method.

Figure 5. Reconstruction results for three MICC subjects. The first column shows the input images, and the rest columns show the reconstructed D shapes that have the same expression as the input images, using the methods of VRN [16], DDFA [42], DMM-CNN [33], DSR [23] and the proposed method.

Table 2. D face reconstruction accuracy on the MICC database.

Method	VRN	DDFA	DMM-CNN	DSR	Proposed
RMSE					

images from outdoor video frames of subjects. Table 2 shows the D face reconstruction error of different methods on the MICC database. As can be seen, our proposed method obtains the best accuracy due to its fine-grained processing of features in D face shapes. Note that VRN, the first method in the literature that regresses D face shapes directly from D images, has relatively high reconstruction error in terms of RMSE, mainly because it generates low-resolution D face shapes as volumetric representations. In contrast, we reconstruct high-resolution (dense) D face shapes as point clouds with help from low dimensional latent representations.

Results on BU DFE. The BU DFE database contains D faces of subjects displaying expression of neutral (NE), happiness (HA), disgust (DI), fear (FE), anger (AN), surprise (SU) and sadness (SA). All non-neutral expressions were acquired at four levels of intensity. We select neutral and the first intensity level of the rest six expressions as testing data, resulting in testing samples. Further, we render another set of testing images of neutral expression at different poses, i.e., to yaw angles with a interval. These two testing sets evaluate the reconstruction across

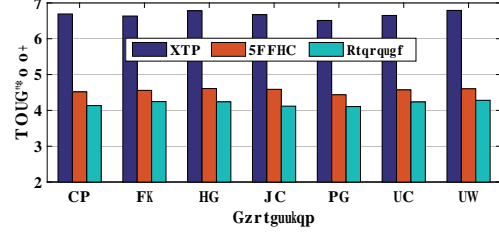


Figure 6. Reconstruction accuracy of D face shapes under different expressions on the BU DFE database. The mean RMSEs of three methods over all expressions are , , and respectively.

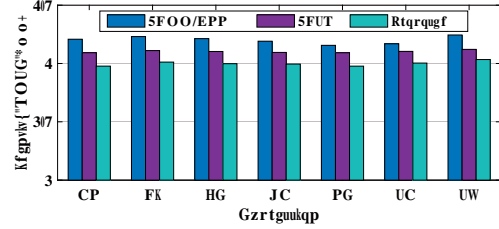


Figure 7. Reconstruction accuracy of the identity component of D face shapes under different expressions on the BU DFE database. The mean RMSEs of three methods over all expressions are , , and respectively.

expressions and poses, respectively.

Table 1 shows the *reconstruction error across poses* (i.e., yaw) of different methods. It can be seen that the RMSE of the proposed method is lower than that of baselines. Moreover, as the pose angle becomes large, the error of our method does not increase substantially. This proves the robustness of the proposed method to pose variations. Figure 6 shows the *reconstruction error across expressions* of VRN, DDFA, and the proposed method based on their reconstructed D face shapes that have the same expression as the input images. Figure 7 compares DMM-CNN, DSR, and the proposed method in terms of RMSE of their reconstructed identity or expression-normalized D face shapes. These results demonstrate the superiority of the proposed method over baselines in handling expressions.

Some example D face reconstruction results are shown in Fig. 5 and Fig. 8. From these results, we can clearly see that the proposed method not only performs well in reconstructing accurate D face shapes for in-the-wild D images, but also disentangles identity and non-identity (e.g.,

Table 3. Face recognition accuracy on the LFW and YTF databases.

Method	Shape	Texture	Accuracy	100%-EER	AUC	TAR-10%	TAR-1%
Labeled Faces in the Wild (LFW)							
DMM							
DDFA							
DMM-CNN							
Proposed							
YouTube Faces (YTF)							
DMM							
DDFA							
DMM-CNN							
Proposed							

Figure 8. Reconstruction results for an BU DFE subject under seven different expressions. The first column shows the input images. In the blue box, we show the reconstructed 3D shapes that have the same expression as the input images, using the methods of VRN [16], DDFA [42] and the proposed method. In the red box, we show the reconstructed *identity* 3D shapes obtained by DMM-CNN [33], DSR [23] and the proposed method. Our composite 3D shape model enables us to generate two types of 3D shapes.

expression) components in 3D face shapes. As we will show in the following face recognition experiments, the disentangled shape features contribute to face recognition.

4.2. Face Recognition Accuracy

To evaluate the effectiveness of our shape features (i.e., the identity representations) to face recognition, we compute the similarity of two faces using the cosine distances between their shape features extracted by the encoder of our method. To investigate the complementarity between our learnt shape features and existing texture features, we also fuse our method with existing methods via summation at the score level [21]. The counterpart methods we consider here include DMM [28], DDFA [42], DMM-CNN [33], and SphereFace [24]. We compare the methods in terms of verification accuracy, 100%-EER (Equal Error Rate), AUC (Area Under Curve) of ROC (Receiver Operating Characteristic) curves, and TAR (True Acceptance Rate) at FAR (False Acceptance Rate) of 10% and 1%.

Results on LFW. The Labeled Faces in the Wild (LFW) benchmark dataset contains 13,236 images collected from Internet. The verification set consists of 40 folders, each with 10 same-person pairs and 10 different-person pairs. The recognition accuracy of different methods on LFW is listed in Tab. 3. Among all the 3D face reconstruction methods, when using only shape features, our proposed method achieves the highest accuracy, improving TAR@FAR from 10.2% to 12.5% with respect to the latest DMM-based method [33].

Results on YTF. The YouTube Faces (YTF) database contains 2,181 videos of 1,271 individuals. Face images (video frames) in YTF have lower quality than those in LFW, due to larger variations in pose, illumination and expression, and low resolution as well. Table 3 summarizes the recognition accuracy of different methods on YTF. Despite the low-quality face images, our proposed method still outperforms the baseline methods in the sense of extracting discriminative shape features. By fusing with one of the state-of-the-art texture-based face recognition methods (i.e.,

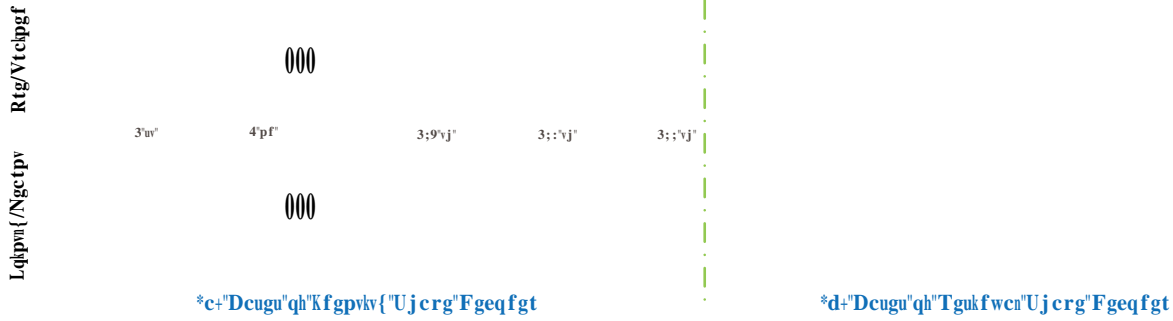


Figure 9. Comparing the pre-trained 3DMM-like and our jointly-learned bases defined by the weights of identity and residual shape decoders. (a) For the bases of identity shape decoder, the weights associated with each entry in are added to the mean shape, reshaped to a point cloud (), and shown as polygon meshes. (b) For the bases of residual shape decoder, the weights associated with each entry in are reshaped to a point cloud (), and shown as a heat map that measures the norm value of each vertex (i.e., the deviation from the identity shape). Red colors in the heat maps indicate larger deviations. It is important to note that the conventional DMM bases are trained from D face scans, while our bases are learnt from D images.

Table 4. Efficiency comparison of different methods.

Method	VRN	DDFA	DMM-CNN	DSR	Proposed
Time (ms)					

SphereFace [24]), our proposed method further improves the face recognition accuracy on YTF from to . This proves the complementarity of *properly reconstructed* shape features to texture features in face recognition. This is a notable result especially considering the D face recognition method of SphereFace [24] has already set a very high baseline (i.e.,).

4.3. Computational Efficiency

To assess the computational efficiency, we run the methods on a PC (with an Intel Core i - K @ GHz, GB RAM and an GeForce GTX) for images, and calculate the average runtime per image in Tab. 4. Note that DDFA and DMM-CNN estimate the DMM parameters in the first step, and we report their runtime of obtaining the final D faces. For VRN, DDFA and DMM-CNN, despite stand-alone landmark detection is required, the reported time does not include the landmark detection time. Our proposed method needs only milliseconds (ms) per image, which is an order of magnitude faster than baseline methods. This is owing to the light-weight network in our method. In contrast, baseline methods use either very deep networks [33], or cascade approaches [23, 27].

4.4. Analysis and Discussion

To offer insights into the learnt decoders, we visualize their weight parameters in Fig. 9. The weights associating one entry in the latent representations with all the neurons in the FC layer in the decoders are analogous to a DMM basis (see Fig. 4). Both pre-trained bases and jointly-learned

bases are shown for comparison in Fig. 9, from which the following observations can be made.

- (i) The pre-trained identity bases approximate the conventional DMM bases [4] that are ordered with latter bases capturing less shape variations. In contrast, our jointly-learned identity bases all describe rich shape variations.
- (ii) Some basis shapes in the jointly-learned bases do not look like regular face shapes. We believe this is due to the employed joint reconstruction and identification loss function. The bases trained from a set of D scans as in DMM, while optimal for reconstruction, might limit the discriminativeness of shape parameters. Our bases are trained with the classification in mind, which ensures the superior performance of our method in face recognition.
- (iii) The pre-trained residual bases, like the expression shape bases [8], appear symmetrical. The jointly-learned residual bases display more diverse shape deviation patterns. This indicates that the residual shape deformation captured by the jointly-learned bases is much beyond that caused by expression changes, and proves the effectiveness of our method in disentangling D face shape features.

5. Conclusions

We have proposed a novel encoder-decoder-based method for jointly learning discriminative shape features from a D face image and reconstructing its dense D face shape. To train the encoder-decoder network, we implement a multi-image DMM fitting method to construct training data, and develop an effective training scheme with a joint reconstruction and identification loss. We show with comprehensive experimental results that the proposed method can effectively disentangle identity and non-identity features in D face shapes and thus achieve state-of-the-art D face reconstruction accuracy as well as improved face recognition accuracy.

References

- [1] <https://support.apple.com/en-us/HT208109>. Accessed: 2017-11-15.
- [2] A. D. Bagdanov, A. Del Bimbo, and I. Masi. The florence 2D/3D hybrid face dataset. In *Workshop on Human gesture and behavior understanding*, pages 79–80. ACM, 2011.
- [3] A. Bas, W. A. Smith, T. Bolkart, and S. Wuhler. Fitting a 3D morphable model to edges: A comparison between hard and soft correspondences. In *ACCV*, pages 377–391, 2016.
- [4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, pages 187–194, 1999.
- [5] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *TPAMI*, 25(9):1063–1074, 2003.
- [6] K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3D and multi-modal 3D+ 2D face recognition. *CVIU*, 101(1):1–15, 2006.
- [7] 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 *ICCV*, 2017.
- [8] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3D facial expression database for visual computing. *TVCG*, 20(3):413–425, 2014.
- [9] C. Cao, H. Wu, Y. Weng, T. Shao, and K. Zhou. Real-time facial animation with image-based dynamic avatars. *TOG*, 35(4):126:1–126:12, 2016.
- [10] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3D face reconstruction with deep neural networks. In *CVPR*, 2017.
- [11] M. Emambakhsh and A. Evans. Nasal patches and curves for expression-robust 3D face recognition. *TPAMI*, 39(5):995–1007, 2016.
- [12] H. Han and A. K. Jain. 3D face texture modeling from uncalibrated frontal and profile images. In *BTAS*, pages 223–230, 2012.
- [13] X. Han, C. Gao, and Y. Yu. Deepsketch2face: A deep learning based sketching system for 3D face and caricature modeling. *TOG*, 36(4), 2017.
- [14] B. K. Horn and M. J. Brooks. *Shape from shading*. Cambridge, MA: MIT press, 1989.
- [15] 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, Technical Report 07-49, University of Massachusetts, Amherst, 2007.
- [16] 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 *ICCV*, 2017.
- [17] A. Jourabloo and X. Liu. Pose-invariant 3D face alignment. In *ICCV*, pages 3694–3702, 2015.
- [18] A. Jourabloo and X. Liu. Pose-invariant face alignment via CNN-based dense 3D model fitting. *IJCV*, in press, 2017.
- [19] A. Jourabloo, M. Ye, X. Liu, and L. Ren. Pose-invariant face alignment with a single cnn. In *ICCV*, 2017.
- [20] I. Kemelmacher-Shlizerman and R. Basri. 3D face reconstruction from a single image using a single reference face shape. *TPAMI*, 33(2):394–405, 2011.
- [21] J. Kittler, M. Hatef, R. P. Duin, and J. Matas. On combining classifiers. *TPAMI*, 20(3):226–239, 1998.
- [22] F. Liu, D. Zeng, J. Li, and Q. Zhao. Cascaded regressor based 3D face reconstruction from a single arbitrary view image. *arXiv:1509.06161*, 2015.
- [23] F. Liu, D. Zeng, Q. Zhao, and X. Liu. Joint face alignment and 3D face reconstruction. In *ECCV*, pages 545–560, 2016.
- [24] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017.
- [25] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. In *AVSS*, pages 296–301, 2009.
- [26] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *ICCV*, 2017.
- [27] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, 2017.
- [28] S. Romdhani and T. Vetter. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR*, pages 986–993, 2005.
- [29] J. Roth, Y. Tong, and X. Liu. Adaptive 3D face reconstruction from unconstrained photo collections. In *CVPR*, pages 4197–4206, 2016.
- [30] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *ICCV*, 2017.
- [31] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014.
- [32] 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 *CVPR*, 2017.
- [33] A. T. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *CVPR*, 2017.
- [34] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, pages 1283–1292, 2017.
- [35] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, pages 529–534, 2011.
- [36] D. Yi, Z. Lei, and S. Z. Li. Towards pose robust face recognition. In *CVPR*, pages 3539–3545, 2013.
- [37] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014.
- [38] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3D facial expression database for facial behavior research. In *FG*, pages 211–216, 2006.
- [39] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017.
- [40] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *SPL*, 23(10):1499–1503, 2016.

- [41] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Li. Face alignment across large poses: A 3D solution. In *CVPR*, pages 146–155, 2016.
- [42] 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 *CVPR*, pages 787–796, 2015.