# Face Reconstruction with Specific Weight Mask

Wentao Shi(✉), Tianji Ma, Nangyang Bai, and Lutao Wang

Chengdu University of Information Technology, No. 24, 1st Section, Xuefu Road, Southwest Airlines, Chengdu, Sichuan, China
waglt@cuit.edu.cn

**Abstract.** Create a 3D face model from a 2D face image, generally extract facial feature points, calculate a 3D deformation model, and perform deformation and stretching on the generated face database. However, this approach is not only time-consuming but also has no calculation errors. Ideally, neural networks' use to obtain deformation model parameters is also affected by factors such as pose, angle, and datasets. 3D face reconstruction methods rely excessively on the accuracy of the labeling and the face detector's accuracy. This article proposes a method that is not affected by pose. We adopt a feature point extractor that can obtain more features, design an hourglass network to get a model, and consider each feature area differently, effectively using the feature point information. Map from two-dimensional coordinates to three-dimensional space to achieve face reconstruction, and obtain a high-precision face model. We do experiments on the three-dimensional face datasets AFLW2000-3D and 300W-3D. The results show that this method can obtain good performance in face multi-angle reconstruction, and the accuracy is also improved.

**Keywords:** Computer vision · Face alignment · Dense alignment · Face reconstruction

## 1 Introduction

With the improvement of perception technology and the fermentation of deep learning, studys in face recognition improve rapidly in recent years, and the accuracy of 3D face reconstruction has also been continuously improved. Due to its wide range of applications, 3D face reconstruction has always attracted attention. Obtaining 3D information from 2D images is of great significance to various fields. It plays a crucial character on animation, games, smart shopping, information security, and other fields. However, 3D face reconstruction often relies on expensive capture equipment and professional technicians, and the cost is exceptionally high. For a long time, the loss of in-depth information and lack of prior knowledge has been a problem.

2D image recognition can easily obtain accurate information and has high robustness, while 3D face reconstruction is affected by various factors such as angle, illumination, skin texture, and lack of datasets with real marks. In order to get good results, the input picture has to perform a good angle, and the face of the image cannot be blocked. Also,


X. Wu et al. (Eds.): QShine 2020, LNICST 381, pp. 237–247, 2021.
https://doi.org/10.1007/978-3-030-77569-8_17

the method might be time-consuming, and the results might not work well. Therefore, the transition from two-dimensional to three-dimensional is a challenging topic in the computer vision area.

The face can be regarded as a three-dimensional object containing texture information and structure information. The three-dimensional face reconstruction method is mainly based on the optimization algorithm [1], and the corresponding face structure and texture information is obtained by obtaining the 3D Morphable Model (3DMM) coefficients [2]. 3DMM learns the prior knowledge of 3D face utilizing statistical analysis and obtains the required face model by controlling the average face database model's deformation. Reference [9] proposed an idea to separate the four areas of the face, find the best-fitting model in each area, and further deform and combine to find the best-fitting model corresponding to each area. However, these methods will be computationally time-consuming due to their high computational complexity and rely on prior knowledge, difficulty in initialization, and easy to fall into local optima. Even with these shortcomings in those methods, 3DMM is still proposing solutions to nonlinear regression functions. With the rise of machine learning and deep learning, most of the work is still based on 3DMM. Recently, methods of using CNNs to regress 3DMM coefficients have achieved good results [3], Zhu [4], End to End method [5]. However, many methods are restrained by poses, they need the input data to have a good angle, and the feature regions are not displaying well.

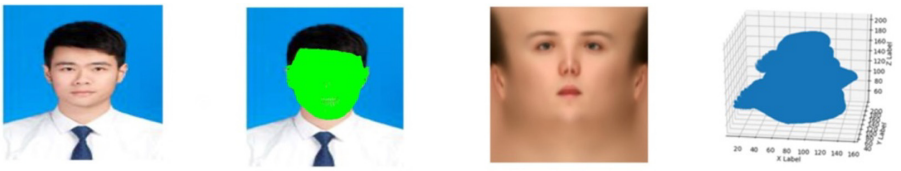

**Fig. 1.** Obtain 3D point cloud information from a single image

In order to solve the problem of insufficient robustness, limited by the rotation angle, restoration accuracy, we create a two-dimensional coordinate system that carries 3D semantic information and divide the face into different regions on the two-dimensional surface through feature points, and we give different weights to different feature areas. Figure 1 shows an example of obtaining 3D point cloud information from a single image. We achieved good results on different angles and performed robustness in different datasets.

## 2 Related Works

From 1999 to 2010, Blanz and Vetter [2] proposed a 3D Morphable Model (3DMM), whose method can construct a 3D face model based on 2D images. As pointed out above, the face is divided into the texture part and structure part. The texture coefficient and structure coefficient are shown in the Eq. (1) and Eq. (2), $\alpha$ is the structure coefficient and $\beta$ is the texture coefficient. These two coefficients control the transformation of

the face model. Generate an average face model that can be deformed according to the images, and change the deformable model's coefficients to stretch and deform to obtain the desired result.

$$S_{model} = \bar{S} + \sum_{i=1}^{m-1} \alpha_i S_i \tag{1}$$

$$\mathrm{T_{model}} = \bar{\mathrm{T}} + \sum_{\mathrm{i}=1}^{\mathrm{m}-1} \beta_\mathrm{i} \mathrm{T_i} \tag{2}$$

Later, in 2004, Blanz [10] proposed sparse facial feature points for model parameter estimation. Rara [11] and others proposed a model between 2D facial feature points and 3DMM parameters, using principal component regression analysis (PCR) to estimate 3DMM parameters. Due to the facial posture's influence, the possible accuracy of the detection of the detected 2D facial key feature points may be reduced. Dou [12] proposed a dictionary-based method to regress 3D face shape, using sparse coding to estimate model parameters from facial landmarks. Similarly, Zhou [13] also used a dictionary-based method and proposed a convex formula to estimate model parameters. Anbarjafari [9] proposed an end-to-end concept, dividing the face into four parts. All texture maps are distorted to fit the same UV map, and all the parts behind the face are discarded. For the four regions, corresponding facial models are obtained separately, and the four regions obtained by stitching are combined to obtain a complete face. This method is severely affected by noise and hair.

Recently, 3D face reconstruction methods based on deep learning came out. There are methods to add corrections or details to the rough 3DMM prediction [6], Tewari [15], Guo [16]. Many cutting-edge methods use CNNs to regress 3DMM parameters, please see the example, Richardson [6], Tuan[7], Jackson [8], Richardson [14], Feng Y [17], Tewari [20], Piotraschke [21], Huber [22], He K [23]. Figure 2 shows the process of face recognition, which performs excellently in many applications. However, 3D face reconstruction has many issues, such as lack of datasets, rely on the good pose of input data.

Reference [7] solves the problem of the insufficient training set and is more robust than previous methods. The author uses multiple pose photos of the same object to generate a high-accuracy 3D face model [21], and then uses the generated model as a training set and uses a threaded deep convolutional neural network to generate a robust face model. Reference [17] proposed to use 2D pictures with semantic information and non-equivalently consider the weights of different points for evaluation, but the texture information of this method is rough.

## 3 Network and Loss Function

Our proposed method uses more accurate facial feature extraction to construct a UV map regression 3D face model. The main steps are divided into the extraction of facial feature points, construction of UV maps, and a simple CNN network. In the feature point extraction, we use the face key point extraction method of more key points proposed by Niko. This method belongs to the branch of [22]. Compared with 68 key feature points, 13 key points are added (including forehead area). The difference between the two methods has been shown in Fig. 3 below.

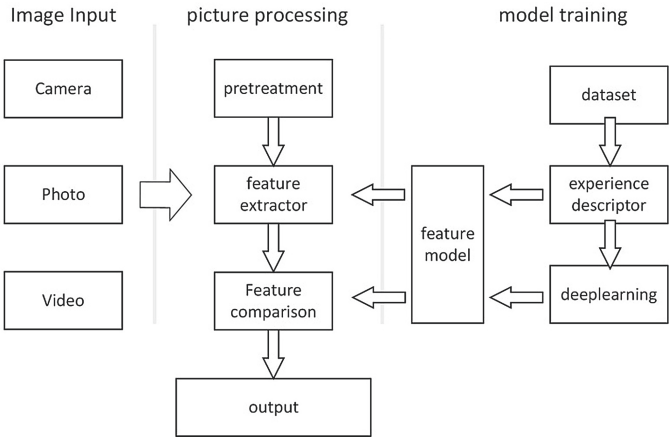


**Fig. 2.** Application of deep learning in the field of face recognition

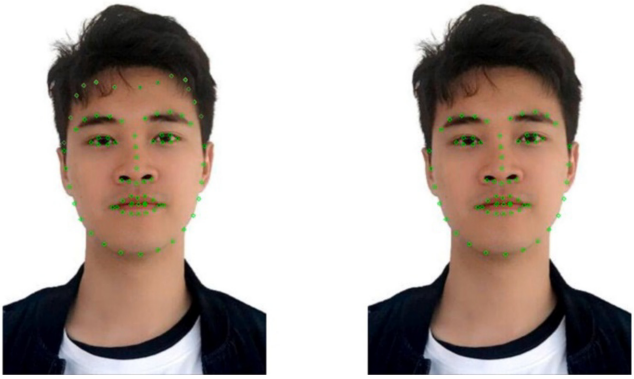

**Fig. 3.** 81 feature key point extractor (left), 68 feature key point extractor (right)

### 3.1 Network

We directly regress the parameters of a 3D face from a single 2D face through the CNN network. Meanwhile, we need to emphasize the robustness of the method. Therefore, we need a dataset containing 2D faces and corresponding 3D information and containing yaw angles as our training set. The 300W-3D data set contains large-angle face images, which satisfies our requirements very well. We choose 300W-3D as our training set. Simultaneously, to evaluate the superiority of the method, we manually annotated the 81 feature points of some of the acquired pictures and showed them in the experimental results section.

In order to ensure the effectiveness of the results, we should consider the facial feature regions differently. It is tough to consider features directly from 3D information. Converting a three-dimensional problem to a two-dimensional problem can solve this problem well.

The UV map is a two-dimensional image that records the information of all 3D point clouds. We refer to [19] to propose a method for constructing a UV location map. We design and use an hourglass network, mainly borrowed from the fully convolutional network and the residual network. According to the input color face photo, the UV position map is obtained from a single 2D face image.

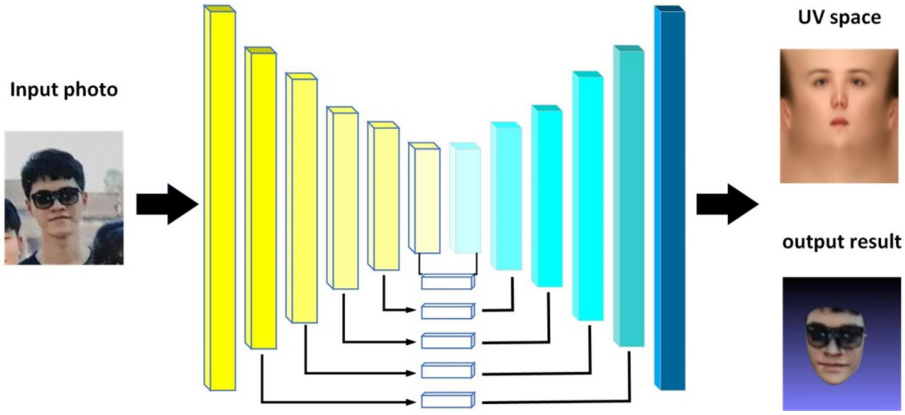


**Fig. 4.** CNN network, rectangle represents residual block.

Each rectangle is a residual block representing a feature. To ensure the dense geometric structure information of the face, the size of the UV map is 256, which can obtain a high-precision point cloud. The network structure is shown in Fig. 4. The network converts the input RGB image into the corresponding UV image, using encoder-decoder. The coding part is composed of cascaded residual blocks, and each layer is activated by the Relu function and input to the next convolutional layer. Finally, after activation of the Relu function, the output of each residual block structure is obtained, and the picture with the size of $256 \times 256 \times 3$ is reduced to the feature map of $8 \times 8 \times 512$, and the decoding part includes the transposed convolution layer, and the picture is restored to the UV map with the size $256 \times 256 \times 3$. The stride is 1, all kernel size is set to 4, and the Relu function is used to activate.

### 3.2 Loss Function

To display the information represented by the distinct regions, Anbarjafari [9] proposed an idea, considering the difference in the characteristic information of each region, modeling each region separately, and finally fusing each region.

Realize this vision by creating a UV map. The UV image is a 2D image that contains 3D point cloud information. Reference [9] Use UV maps to represent the texture information of the face. Different from others, Feng Y [17] uses UV space to store 3D point cloud information. The real 3D point cloud information can accurately match the projection of the 2D plane. The picture's RGB information becomes the x, y, z coordinate points in the texture map. Simultaneously, the UV coordinates also contain 3DMM

parameters, which carry enough 3D semantic information. We learn from the method they proposed. Since the UV map is obtained through the network, we can divide the face region. According to the difference of the region's feature information, different weights are assigned to each region when calculating the loss. Experiments show that doing so can get better experimental results.

$$\text{Loss} = \sum ||\text{P}(\text{x}, \text{y}) - \text{P}(\sim)(\text{x}, \text{y})|| \bullet \text{W}(\text{x}, \text{y}) \tag{3}$$

We divide the face into regions in the obtained UV space. Compared with the previous method, we distribute the weights more carefully according to the proportion. Equation (3) shows our loss function, P(x,y) is the predicted face coordinate point, P is the real face coordinate, and W(x,y) is the weight coefficient corresponding to the coordinate point. In our conception, the weight ratio is different according to the divided regions, and the calculation formula is shown in the above formula. Area 1 (81 facial feature points): Area 2 (eyes, nose): Area 3 (mouth): Area 4 (forehead, other areas): Area 5 (neck) = 16:5:4:3:0. Undoubtedly, the 81 key feature points of area 1 should have the highest weight, which is given to 16. Area 3, because the mouth is an important feature, it is assigned 4; the neck of area 5 is an outside area, assigned 0; area 2 and area 4 It belongs to a distinct area, where the eyes and nose of area 2 are iconic features, which are 1 higher than area 3; the forehead and other facial areas of area 4 are relatively less obvious, and 3 is assigned. In this way, we can consider each feature non-equivalently.

## 4 Experimental Results

Since the dataset requires both 2D pictures and their corresponding 3D semantic information, 300W-3D is selected as the training set because it contains 3DMM coefficients and different angle facial pictures, which enables it to store 3D point cloud information. Compare with 3DDFA [4], DeFA [4], and PRN [17] on the AFLW2000-3D and Florence datasets.

AFLW2000-3D is a dataset used to evaluate the performance of 3D face reconstruction for unconstrained images. This dataset contains the first two thousand face avatars in the AFLW dataset, which can be used for head deflection or head 2D or 3D detection, as well as large-angle faces. The dataset contains two data types. The first is JPG format data, which contains two-dimensional face pictures; MAT format data is a dictionary that contains feature points, 3DMM information, and various image parameters. Since the dataset carries 3DMM coefficients, the three-dimensional information is reconstructed by 3DMM and contains 68 feature points of three-dimensional information. It is the most commonly used dataset to evaluate the performance of facial reconstruction.

300W is a huge face dataset. The dataset has more than thousands of images, each image contains more than one face, but only one face is labeled, mainly used for face alignment. 300W-3D is a dataset that marked 300W data with 3DMM parameters, which can be used to train, test, and evaluate reconstruction performance. We show some examples in Fig. 5. Also, 300W-LP is suitable for our experiment. 300W-LP a subset of 300W samples containing large angles. It can be used to evaluate the robustness of the method to deflection angle rotation.

**Fig. 5.** Some large pose face in 300W-LP.

The Florence-3D dataset contains 53 labeled objects. Every object has different angles. In this section, we compare the performance of some methods, such as 3DDFA and DeFa.

We manually marked 81 key points of faces in some pictures and conducted a separate display experiment. The results have been shown at the rendering part, and we can see the performance intuitively.

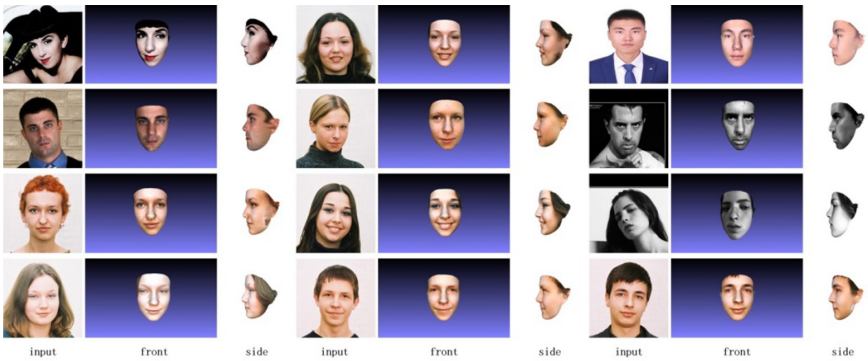


**Fig. 6.** Output of our method, including the front face and the side face. Some input faces are captured by ourselves

Use the carried 3DMM coefficients to generate the corresponding 3D position map and render it into the UV space. The size of the picture remains unchanged in the UV space, still 256 $\times$ 256. Use the hourglass network for training and Adam optimizer operation. The learning rate starts from 0.0001, decays by half every 20 batches, and the input face (rotation) is randomized. Use tensorflow to run. The graphics card used in the experimental hardware is a GTX1070Ti graphics card and I7-8700 central processing

unit. Figure 6 shows some examples of our outputs; some input images were captured and manually mark by ourselves.

Although the regularized mean square error (MSE) is generally used as the performance evaluation index, direct use of MSE will cause the loss of key information. To better evaluate the network performance, we evaluate the method at large pose datasets by calculating the regularized mean error (NME). The results show good robustness and well performance. Furthermore, we manually mark 81 key points in some photos and give the results at the rendering display part. In order to further study the performance of the method in this paper under different angles, some data are compared with the NME under small, different, and large deflection angles. The results are shown in Table 1.

**Table 1.** Performance compared with other methods at different angles (NME).

| Method | 0° to 30° | 30° to 60° | 60° to 90° | Mean |
|---|---|---|---|---|
| 3DDFA | 3.80 | 4.55 | 7.88 | 5.41 |
| 3DSTN | 3.13 | 4.43 | 5.78 | 4.45 |
| PRN | 2.75 | 3.50 | 4.60 | 3.62 |
| SDM | 3.67 | 4.90 | 9.67 | 6.08 |
| Ours | 2.69 | 3.48 | 4.58 | 3.58 |

We use CED (cumulative error distribution, which is used to obtain the sum of all variables below a specific value) to observe the experimental results intuitively. We evaluate the performance of our method on the AFLW2000-3D. The result has been shown in Fig. 7.

We divide the face into 5 regions according to the extracted feature points and assign different weights to each region. The 81 key points have the highest weights to ensure that the network can accurately learn these points' positions. Because the neck area is meaningless for the face model's regression, the weight is zero.

To verify the effect of dividing feature regions, we compare the method of not considering feature regions (all facial regions are equal to 1), the allocation method is shown in Table 2, and the results are shown in Fig. 8. Obviously, considering facial feature regions, giving different weights to different regions can make our network better.

The best way to evaluate the performance of the face model is to observe the rendering result directly. We manually mark 81 key points. We display the outputs below. See Fig. 9. Figure (a) is the reconstruction result of the front face, and figure (b) is the reconstruction result of the side face. Even if the picture uses a picture that includes a rotation angle, we can get satisfactory results.

As shown upon, all facial details, wrinkles, spots are basically restored, and the model will not be influenced by changing the angle. Even we zoom in on the picture, and we still can get a clear vision of the texture details.

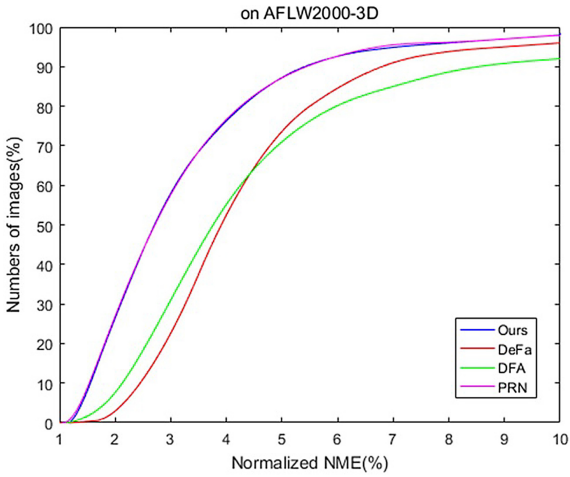


**Fig. 7.** 3D reconstruction performance comparison, the cumulative error distribution of 4 different methods on the AFLW2000-3D dataset, smooth the curve for better observing

**Table 2.** Different weight ratio

| | Area1 | Area2 | Area3 | Area4 | Area5 |
|---|---|---|---|---|---|
| Weight r1 | 16 | 5 | 4 | 3 | 0 |
| Weight r2 | 1 | 1 | 1 | 1 | 0 |

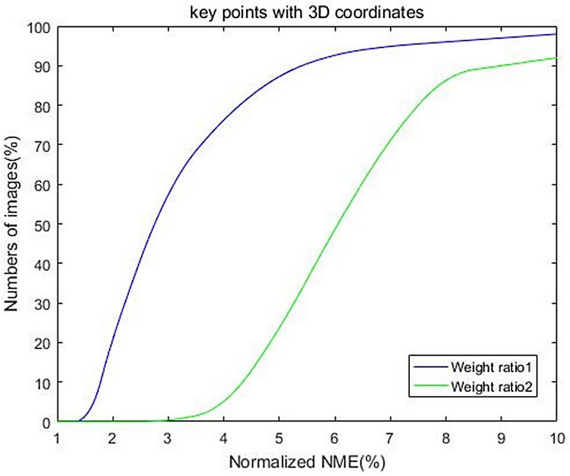


**Fig. 8.** Comparison of weight ratio 1 and weight ratio 2.

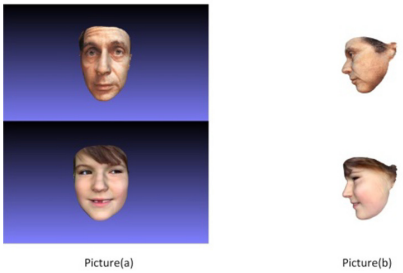


**Fig. 9.** Outputs of our methods, including both the front face and the side face

## 5 Conclusion

We provide a new idea for solving the 3D face reconstruction from a single 2D picture. This idea does not rely on expensive hardware equipment and complex networks and can fully return to face features. Experiments show that this method has an excellent performance in 3D face reconstruction.

In this paper, we show the CNN using the hourglass network structure to regress the UV parameters. During the training process, feature points are added as a guide to obtain a robust face image. Meanwhile, an excellent facial feature extractor is introduced to visually display the rendering results. Even if the input image is rotated, excellent results can be obtained without being affected by the rotation. There still has many ways to extend this work, such as applying more accurate feature area weights.

## References

1. Amberg, B., Romdhani, S., Vetter, T.: Optimal step nonrigid ICP algorithms for surface registration. In: Computer Vision and Pattern Recognition, pp. 1–8 (2007)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (1999)
3. Jourabloo, A., Liu, X.: Large-pose face alignment via CNN-based dense 3D model fitting. In: Computer Vision and Pattern Recognition (2016)
4. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: a 3D solution. In: Computer Vision and Pattern Recognition, pp. 146–155 (2016)
5. Dou, P., Shah, S.K., Kakadiaris, I.A.: End-to-end 3D face reconstruction with deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 21–26 (2017)
6. Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5553–5562 (2017)
7. Tran, A.T., Hassner, T., Masi, I., Medioni, G.: Regressing robust and discriminative 3D morphable models with a very deep neural network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1493–1502 (2017)
8. Jackson, A.S., Bulat, A., Argyriou, V., Tzimiropoulos, G.: Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In: International Conference on Computer Vision (ICCV), pp. 1031–1039 (2017)

9. Anbarjafari, G., Haamer, R.E., Lusi, I., et al.: 3D face reconstruction with region based best fit blending using mobile phone for virtual reality based social media. arXiv preprint arXiv: 1801.01089 (2017)
10. Blanz, V., Mehl, A., Vetter, T., Seidel, H.-P.: A statistical method for robust 3D surface reconstruction from sparse data. In: Proceedings of the International Symposium on 3D Data Processing Visualization and Transmission, Thessaloniki, Greece, pp. 293–300, 6–9 September 2004 (2004)
11. Rara, H., Farag, A., Davis, T.: Model-based 3D shape recovery from single images of unknown pose and illumination using a small number of feature points. In: Proceedings of the International Joint Conference on Biometrics, Washington, DC, pp. 1–7, 11–13 October 2011 (2011)
12. Dou, P., Wu, Y., Shah, S.K., Kakadiaris, I.A.: Robust 3D facial shape reconstruction from single images via twofold coupled structure learning. In: Proceedings of the British Machine Vision Conference, Nottingham, United Kingdom, pp. 1–13, 1–5 September 2014 (2014)
13. Zhou, X., Leonardos, S., Hu, X., Daniilidis, K.: 3D shape reconstruction from 2D landmarks: a convex formulation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, Massachusetts, pp. 4447–4455, 7–12 June 2015 (2015)
14. Richardson, E., Sela, M., Kimmel, R.: 3D face reconstruction by learning from synthetic data. In: International Conference on 3D Vision (3DV), pp. 460–469 (2016)
15. Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., Theobalt, C.: Self-supervised multi-level face model learning for monocular reconstruction at over 250 Hz. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2549–2559 (2018)
16. Guo, Y., Zhang, J.Z., Cai, J., Jiang, B., Zheng, J.: CNN-based real-time dense face reconstruction with inverse rendered photo-realistic face images. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **41**, 1294–1307 (2018)
17. Feng, Y., et al.: Joint 3D face reconstruction and dense alignment with position map regression network. In: Proceedings of the European Conference on Computer Vision (2018)
18. Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., Freeman, W.T.: Unsupervised training for 3D morphable model regression. In: IEEE Conference on Computer Vision and Pattern Recognition, June 2018
19. Bas, A., Huber, P., Smith, W.A., Awais, M., Kittler, J.: 3D morphable models as spatial transformer networks. In: International Conference on Computer Vision Workshop on Geometry Meets Deep Learning, pp. 904–912 (2017)
20. Tewari, A., Zollhöfer, M., Kim, H., Garrido, P., Bernard, F., Pérez, P., Theobalt, C.: MoFa model-based deep convolutionalface auto encoder for unsupervised monocular reconstruction. In: International Conference on Computer Vision, pp. 1274–1283 (2017)
21. Piotraschke, M., Blanz, V.: Automated 3D face reconstruction from multiple images using quality measures. In: Proceedings of the Conference on Computer Vision Pattern Recognition, June 2016
22. Huber, P., et al.: A multiresolution 3D morphable face model and fitting framework. In: Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (2016)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition (2016)