3D Face Reconstruction from Light Field Images: A Model-free Approach

Mingtao Feng[1], Syed Zulqarnain Gilani[2], Yaonan Wang[1], and Ajmal Mian[2]

[1] College of Electrical and Information Engineering, Hunan University,410006, China
{mintfeng,yaonan}@hnu.edu.cn
[2] Computer Science and Software Engineering, The University of Western Australia, 6009, Australia
{zulqarnain.gilani,ajmal.mian}@uwa.edu.au

Abstract. Reconstructing 3D facial geometry from a single RGB image has recently instigated wide research interest. However, it is still an ill-posed problem and most methods rely on prior models hence undermining the accuracy of the recovered 3D faces. In this paper, we exploit the Epipolar Plane Images (EPI) obtained from light field cameras and learn CNN models that recover horizontal and vertical 3D facial curves from the respective horizontal and vertical EPIs. Our 3D face reconstruction network (FaceLFnet) comprises a densely connected architecture to learn accurate 3D facial curves from low resolution EPIs. To train the proposed FaceLFnets from scratch, we synthesize photo-realistic light field images from 3D facial scans. The curve by curve 3D face estimation approach allows the networks to learn from only 14K images of 80 identities, which still comprises over 11 Million EPIs/curves. The estimated facial curves are merged into a single pointcloud to which a surface is fitted to get the final 3D face. Our method is model-free, requires only a few training samples to learn FaceLFnet and can reconstruct 3D faces with high accuracy from single light field images under varying poses, expressions and lighting conditions. Comparison on the BU-3DFE and BU-4DFE datasets show that our method reduces reconstruction errors by over 20% compared to recent state of the art.

1 Introduction

Three dimensional face analysis has the potential to address the challenges that confound its two dimensional counterpart such as variations in illumination, pose and scale [4]. This modality has achieved state-of-the-art performances on applications such as face recognition [14, 36, 39, 65], syndrome diagnosis [17, 16, 47, 55], gender classification [15] and face animation [9, 49]. Reconstructing 3D facial geometry from RGB images is, therefore, receiving a significant interest from the research community. However, using a single RGB image to recover the 3D face is an ill-posed problem [31] since the depth information is lost during the projection process. In fact, many different 3D shapes can result in similar 2D projections. The scale and bas-relief ambiguities [6] are common examples.

Most existing methods have resorted to the use of prior models such as the Basal Face Model (BFM) [43] and the Annotated Face Model(AFM) [12] to generate synthetic data with ground truth to train CNN [11, 40] models and to recover the model parameters at test time. However, model-based approaches are inherently biased and constrained to the space of the training data of the prior models.

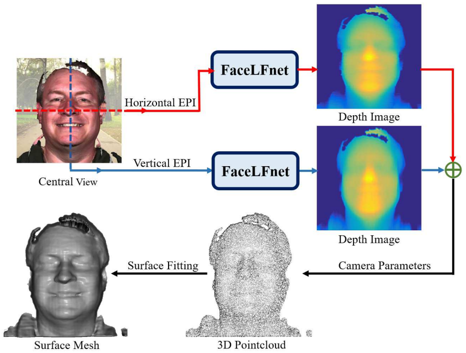


Fig. 1. Proposed pipeline for 3D face reconstruction from a single light field image. Using synthetic light field face images, we train two FaceLFnets for regressing 3D facial curves over their respective horizontal and vertical EPIs. The estimated depth maps are combined, using camera parameters, into a single pointcloud to which a surface is fitted to get the final 3D face.

A 4D light field image captures the RGB color intensities at each pixel as well as the direction of incoming light rays. High resolution plenoptic cameras [2, 3] are now commercially available. Plenoptic cameras use an array of micro-lenses to capture many sub-aperture images arranged in an equally spaced rectangular grid. Unlike most 3D scanners that use active light projection and are hence restricted to indoor use, plenoptic cameras are passive and can instantly acquire light field images outdoors as well, in a single photographic exposure. The sub-aperture light field images have been exploited to improve the performance of many applications such as saliency detection [32], hyperspectral light field imaging [57], material classification [53], image segmentation [62] and image restoration [50, 56] and in particular, depth estimation [26, 48, 34, 52, 46]. This paper focuses on reconstructing 3D faces from light field images under a wide range of pose, expression and illumination variations. Note that unlike stereo, the sub-aperture light field images are captured by the same camera with a single click.

Various methods have been proposed to solve the ill-posed problem of reconstructing 3D facial geometry from a single RGB image [31, 40, 11, 51, 44, 29]. These methods all use one or more common techniques. For instance, Shape from Shading (SfS) uses the shading variation to reconstruct 3D faces but the caveat is that the method is sensitive to lighting and RGB image texture and even under near ideal conditions, suffers from the bas-relief ambiguity [6]. 3D Morphable Models (3DMM) [11, 40] project the 3D faces in a low-dimensional subspace. However, the models are confined to the linear space of their training data and do not generalize well to all face shapes [13]. Landmark based methods use facial keypoints to guide the reconstruction process but rely heavily on accurate localization of the landmarks.

We propose a model-free approach (see Fig. 1) to reconstruct 3D faces directly from light field images using Convolutional Neural Networks (CNN). Our technique does not rely on model fitting or landmark detection. Training a CNN requires massive amount of photo-realistic labeled data. However, there is no publicly available 4D light field face dataset with corresponding ground truth 3D face models. We address this problem and propose a method of generating the training data. We use the BU-3DFE [58] and BU-4DFE datasets [60] to generate light field images from their ground truth 3D models. Figure 2 shows some examples. We randomly vary the light intensity and pose to make our dataset more realistic. Our dataset comprises approximately 19K photo-realistic light field images with ground truth depth maps [3]. Furthermore, we show that our method requires fewer training samples (facial identities) as it capitalizes on reconstructing 3D facial curves rather than the complete face at once. We believe that our synthesized dataset of 4D light field images with corresponding 3D facial scans can be applied to many other facial analysis problems such as pose estimation, recognition and alignment.

Equipped with a rich light field image dataset, we propose a densely connected CNN architecture (FaceLFnet) to learn 3D facial curves from Epipolar Plane Images (EPIs). We train two networks separately using horizontal and vertical EPIs to increase the accuracy of depth estimation. The densenet architecture is preferred as it can accurately learn the subtle slopes in low resolution EPIs[4]. FaceLFnets are trained using our synthetic light field face images for which the ground truth depth data is available. Once the face curve estimates are obtained independently from the horizontal and vertical FaceLFnets, we merge them into a single pointcloud based on the camera parameters and then use a surface fitting method to recover the final 3D face. The core idea of our work is a model-free approach, where the solution is not restricted to any statistical face space. This is possible by exploiting the shape information present in the Epipolar Plane Images.

Our contribution are: (1) A model-free approach for 3D face reconstruction from a single light field image. Our method does not require face alignment or landmark detection and is robust to facial expressions, pose and illumination variations. Being model-free, our method also estimates the peripheral regions of the face such as hair and neck. (2) A training technique that does not require massive number of facial identities. Exploiting the EPIs, we demonstrate that the proposed FaceLFnet can learn from only a few identities (80) and still outperform the state-of-the-art methods by a margin of 26%. (3) A data syntheses technique for generating a light field face image dataset which, to the best of our knowledge, is the first of its kind. This dataset will contribute to solving other face analysis problems as well.

[3] We use depth map to represent disparity map as they are related by light field camera parameters [22].

[4] Higher slope of lines in EPI corresponds to lower depth values.

2 Related Work

3D face reconstruction from a single image has attracted significant attention recently. Shape-from-shading (SfS) has been a popular approach for this task [61, 37, 18]. For example, WenYi et al. [61] proposed a symmetric SfS method to obtain illumination-normalized image and developed a face recognition system. Roy et al. [37] proposed an improved SfS method to enhance the depth map combining the RGB image and rough depth image to create more details. Yudeog et al. [18] estimated lighting variations with both global and local light models. SfS approach was then applied with the estimated lighting models for accurate shape reconstruction. Reconstruction using SfS requires priors of reflectance properties and lighting conditions and suffers from the bas-relief ambiguity [6].

A 3D Morphable Model (3DMM) was introduced by Blanz and Vetter [7] which represents a 3D face as a linear combination of orthogonal basis vectors obtained by PCA over 100 male and 100 female identities. James et al. [8] extended the concept and proposed a statistical model combined with a texture model for fitting the 3DMM on face images in *the wild*. 3DMM has also been used in [38, 5, 42, 30] for face reconstruction. The main limitation of such methods is that the 3DMM cannot model every possible face. Moreover, it is unable to extract facial details like wrinkles and folds because such details are not encoded in the linear subspace.

Recently, various attempts were made to integrate 3DMMs with CNN for facial geometry reconstruction from a single image. Elad et al. [40] employed an iterative CNN trained with synthetic data to estimate 3DMM vectors. The predicted geometry was then refined by the real-time shape-from-shading method. Matan et al. [41] extended the work [40] and introduced an end-to-end CNN framework that recovers the coarse facial shape using a *CoarseNet*, followed by a *FineNet* to refine the facial details. The two net parts are connected by a novel layer that renders the depth image from 3D mesh. Pengfei Dou et al. [11] proposed an end-to-end 3D face reconstruction method from a single RGB image. They trained a fusion-CNN with multi-task learning loss to simplify 3D face reconstruction into neutral and expressive 3D facial parameters estimation. Jourabloo et al. [29] proposed a 3DMM fitting method for face alignment, which uses a cascaded CNN to regress camera matrix and 3DMM parameters. Tuan Tran et al. [51] used multi image 3DMM estimates as ground truth and then trained a CNN to regress 3DMM shape and texture parameters from an input image.

Kemelmacher el at. [31] used the input image as a guide to build a single reference model to align with the face image and then refined the reference model using SfS method. Tal et al. [19] used a 3D neutral face as reference model to approximate the RGB image for face frontalization. Matan et al. [44] proposed a translation network that learns two maps (a depth image and a correspondence map), used for non-rigid registration with a template face, from a single RGB image. Fine-tuning is then performed for reconstructing facial details. In contrast to SfS and model fitting based face reconstruction methods, we learn 3D face curves from EPIs of the light field image. Our method does not require face alignment, dense correspondence or model fitting steps and is robust to facial pose, expressions and illumination.

To the best of our knowledge none of the existing methods is model-free and uses a prior face model at some stage of the reconstruction process. On the other hand our method is completely model-free. Similarly, we are unaware of any existing technique that uses light field images for 3D face reconstruction. However, literature points to some research in shape reconstruction from light field images using deep learning. Heber et al. [20] presented a method for reconstructing the shape from light field images that applies a CNN for pixel wise depth estimation from EPI patches. Although this method produces accurate scene depth, it uses a carefully designed dataset containing drastic slope changes in the EPIs. This method is unsuitable for non-rigid facial geometry reconstruction as faces are generally smooth and their EPIs contain only subtle slope variations. Heber et al. [21] proposed a U-shaped network architecture that automatically learns from EPIs to reconstruct their corresponding disparity images. However, training the network requires disparity maps of all the light field sub-views as labels, which is unrealistic for real datasets. Our approach differs in three ways. Firstly, we use one full EPI as input and its corresponding depth values as labels to overcome the problem of inaccurate depth estimation in the presence of subtle slope variations in the EPIs. Secondly, we train networks using horizontal EPIs and vertical EPIs separately to obtain a more accurate combined 3D pointcloud. Finally, our method does not require disparity maps of all the light field sub-views.

3 Facial Light Field Image Dataset Generation

The key to the success of CNN-based 3D face reconstruction from a single RGB image lies in the availability of large training datasets. However, there is no large scale dataset available that provides RGB face images and their corresponding high quality 3D models. Similarly, training a light field face reconstruction network requires a large-scale light field face dataset with corresponding ground truth 3D facial scans. Over the past few years, the computer vision community has made considerable efforts to collect light field images [22, 53, 33, 35, 1] for different applications. The only public light field face dataset [45] captured by Lytro Illum$^{\text{TM}}$ camera consists of 100 identities with 20 samples per person. However, depth maps of this dataset are generated using the Lytro Desktop Software$^{\text{TM}}$ and have low resolution as well as low depth accuracy. Therefore, this dataset is not suitable for training a network.

In the absence of large-scale 4D light field face datasets, we propose to generate a dataset of light field face images with ground truth 3D models. For this purpose, we use the public BU-3DFE [58] and BU-4DFE [60] databases to generate light field face images. The former is used for training and testing whereas the latter is used only for testing only. The BU-3DFE dataset consists of 2,500 3D scans from 100 identities (56$\%$ female, 44$\%$ male), with an age range from 18 to 70 years and multiple ethnicities. Each subject is scanned in one neutral and 6 non-natural expressions each with four intensity levels. The BU-4DFE dataset contains 3D video sequences of 101 identities (58 female and 43 male) in six different facial expressions. We select the most representative frame of each expression sequence. As a result, our dataset contains 606 3D scans. These models contain shape details such as wrinkles of not only the fiducial area, but also the hair, ears and neck area which pose challenges for conventional 3D face reconstruction methods. All 3D models have RGB texture.

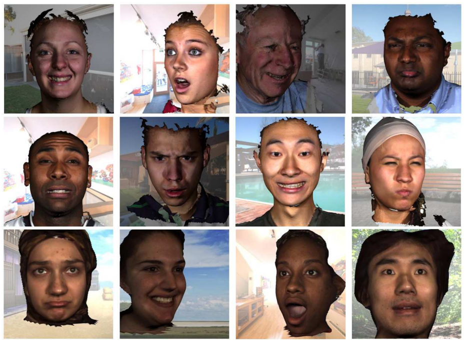


Fig. 2. Central view examples of our rendered light field images. The ground truth 3D scans are aligned with the central view. To make the dataset rich in variations, the generated light field images use random backgrounds and differ extensively in ethnicity, gender, age, pose and illumination.

To generate plausible synthetic light field face images, it is crucial to control the light field camera parameters, background and illumination properly during rendering. We use the open source Blender[5] software and the light field camera tool proposed by Katrin Honauer et al. [22] for this purpose. We place a virtual light field camera in Blender with 15×15 micro-lenses and set its field of view to capture the 3D facial scans. Both BU-3DFE and BU-4DFE databases provide 3D facial models in the near frontal pose. We load the 3D models along with their textures in Blender and apply two rigid rotations ($\pm 15°$) in pitch and four in yaw ($\pm 15°$ and $\pm 30°$). To synthesize photo realistic light field images, we apply randomly selected indoor and outdoor images as backgrounds. We place two lamps at different locations in the scene and randomly change their intensities to achieve lighting variations. The angular resolution of the synthetic light field image is 15×15 and the spatial resolution is 400×400. The ground truth depth maps are aligned with the central view of light field image. Examples of our synthetic light field images are shown in Figure 2.

We implement a Python script[6] in Blender on a 3.4 GHz machine with 8GB RAM to automatically generate the light field facial images. The process of synthesizing light field images can be parallelized since each sub-aperture image is rendered independently. In total, we use 80 identities from BU-3DFE dataset to synthesize 14,000 light field images with ground truth disparity maps. The remaining 20 subjects from BU-3DFE and all 101 subjects from the BU-4DFE dataset are used as test data to generate 1,451 light field facial images for evaluation.

4 Proposed Method

An overview of the proposed method for reconstructing facial geometry from light field image is shown in Figure 1 and the details follow.

[5] http://www.blender.org

[6] The script for light field facial image synthesis will be made public.

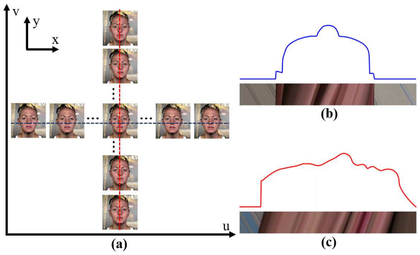


Fig. 3. EPIs corresponding to the 3D face curves. (a) Horizontal and vertical EPIs are obtained between the central view and sub-aperture images that are in the same row and column. (b) and (c) Visualization of the relationship between depth curves and slopes of lines in horizontal and vertical EPIs respectively.

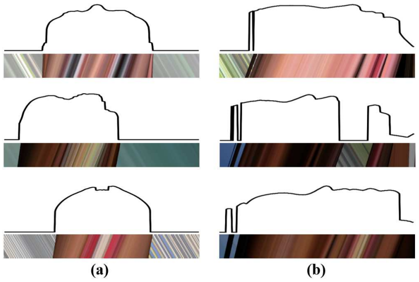


Fig. 4. Examples of EPIs and their corresponding 3D face curves.(a) Horizontal EPIs. (b) Vertical EPIs.

4.1 Training Data

A 4D light field image can be parameterized as $L(u, v, x, y)$, where (x, y) and (u, v) represent the spatial and angular coordinates respectively [54]. When we fix v and y, then $L(u, v^*, x, y^*)$ defines a 2D horizontal EPI. Similarly the 2D vertical EPI can be represented as $L(u^*, v, x^*, y)$ when we keep u and x constant. As shown in Figure 3, 2D EPIs demonstrate the linear characteristic of the light field image. The orientations of lines within the EPIs can infer the disparity of the corresponding 3D space points [54, 28, 59, 20, 21]. Equation (1) shows the relationship between the slope of the line and the disparity value where f is the light field camera parameter and k is the slope of the line.

$$Z = -f \times k, \tag{1}$$

As shown in Figure 3(b) and (c), EPIs correspond to the 3D facial curves from the ground truth. Different line slopes in the EPI indicate different curve shapes. We use 14,000 synthetic light field images corresponding to the 80 identities of BU-3DFE for training. All together we extract 11.2 Million horizontal and vertical EPIs as training samples. Figure 4 shows some example EPIs and their corresponding curves. Using EPI images as training data removes the need for a huge number of identities. Since each 3D face curve can be learned independently from its corresponding EPI, we are able

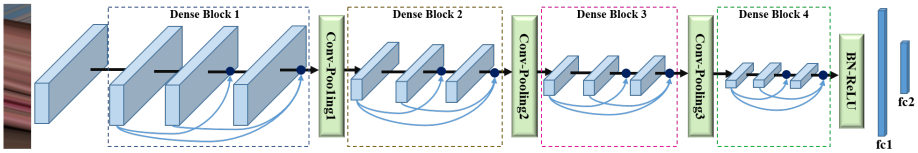


Fig. 5. Our proposed FaceLFnet for learning 3D face curves from EPIs. It contains 4 dense blocks, followed by two fully connected layers. The layers between two neighboring blocks are defined as transition layers and change feature map sizes via convolution and pooling [23].

to generate massive training data from a small number of 3D face scans. Note that we do not need any further data augmentation such as image inversion or multiple crops as our networks learn from the full EPIs.

4.2 FaceLFnet Architecture

Each EPI in our case corresponds to a 3D face curve as shown in Figure 3 and 4. The goal is to predict the full 3D curve from the EPIs using deep learning. CNNs can learn slope information of the pixels from individual EPIs, however, pixel wise prediction is very challenging. Heber et al. [20] divided each EPI into patches for 3D scene estimation. The authors estimated the depth value from each EPI patch independently as it contained the information pertaining to a single line at the center of the patch. In our case, pixel wise estimation is not practical as our network must learn the inter-relationship between the lines in one full EPI to estimate the complete 3D curve. Furthermore, in case of light field images for faces, some EPI patches especially in the quasi planar facial areas are devoid of lines and hence do not contain enough depth information leading to inaccurate depth estimation. Therefore, we propose using a complete EPI for depth prediction in order to exploit the correlations of adjacent pixels and mitigate the problem of inaccurate depth estimation due to pixel wise prediction.

The dimensions of each input EPI are $15 \times 400 \times 3$ (horizontal/vertical sub-aperture images $\times$ horizontal/vertical image pixels $\times$ RGB channels). Such a low resolution in the first dimension and size disparity in the first two dimensions pose challenges as the information of the input EPIs will reduce rapidly in one dimension than the other when passed through a deep network. To extenuate this problem and inspired by the success of Gao et al. [23], we propose a light field face network for estimating facial geometry from EPIs. The architecture of our network is illustrated in Figure 5. It is based on DenseNet that consists of multiple dense blocks and transition layers. We use four dense blocks and change the softmax classifier to a regressor. Before passing the EPIs through the first dense block, a 16 channels convolution layer with 3×3 kernel size is used. For each dense block, we use three convolutional layers and set the growth rate to 12. We also use convolution followed by average pooling as transition layers between two adjacent dense blocks. The sizes of feature-map in the four dense blocks are 15×400, 8×200, 4×100 and 2×50 respectively. The details of network configurations are given in Table 1.

Both horizontal and vertical FaceLFnets are trained from scratch using the Caffe deep learning framework [27]. The initial learning rate is set to 0.0003 which is divided by 10 at 30000 and 50000 iterations. Our networks require only one epoch for convergence. The caffe model for the trained networks will be made public.

4.3 3D Face Reconstruction

The output of our horizontal and vertical FaceLFnets are 3D facial curves that together make up a 3D face. We combine all the horizontal and vertical curves (in our case 400 each) of a face to form a horizontal and a vertical depth map separately. The next step is to reconstruct a 3D face from the two depth maps. A naive way to reconstruct the face is to take the average of both depth maps. However, such a methodology results in reconstruction error as each curve was learned independently. To mitigate this problem we propose a technique to project the depth maps on a 2D surface. First of all, we convert both the depth maps to 3D pointclouds, using the camera parameters. Next we give a slight jitter to the horizontal pointcloud by translating it $1mm$ to the left on x-axis only. We fit a single surface of the form $z(x, y)$ to both 3D pointclouds simultaneously using the *gridfit* algorithm [10]. Our method ensures that a smooth surface is fitted to the horizontal and vertical pointclouds taking into account the correlation between the curves resulting in a smooth reconstructed 3D face.

5 Experimental Results

To the best of our knowledge, there is no suitable real light field face dataset with accompanying 3D ground truth available in the literature. Hence, we present the evaluation of our method for 3D face reconstruction on light field images synthesized from the 3D scans of the remaining 20 subjects of BU-3DFE [58] and all 101 subjects from the BU-4DFE [60] dataset. We compare our subjective results with the recent state-of-the-art algorithm [44] for qualitative evaluation. We also present quantitative comparison with VRN-Guided [25] and other state-of-the-art methods [44, 41, 64, 31, 63, 24]

Layers	Output Size	FaceLFnet
Convolution	15×400	3×3 conv, stride 1
Dense Block 1	15×400	[3×3 conv, stride 1]$\times 3$
Transition Layer 1	15×400	3×3 conv, stride 1
	8×200	2×2 average pool, stride 2
Dense Block 2	8×200	[3×3 conv, stride 1]$\times 3$
Transition Layer 2	8×200	3×3 conv, stride 1
	4×100	2×2 average pool, stride 2
Dense Block 3	4×100	[3×3 conv, stride 1]$\times 3$
Transition Layer 3	4×100	3×3 conv, stride 1
	2×50	2×2 average pool, stride 2
Dense Block 4	2×50	[3×3 conv, stride 1]$\times 3$
Regression Layer	400	4096 fully-connected 400 fully-connected EuclideanLoss

Table 1. Our proposed FaceLFnet architecture. Note that each convolutional layer in the dense block corresponds to the sequence BN-ReLU. The growth rate of the four blocks is $k = 12$.

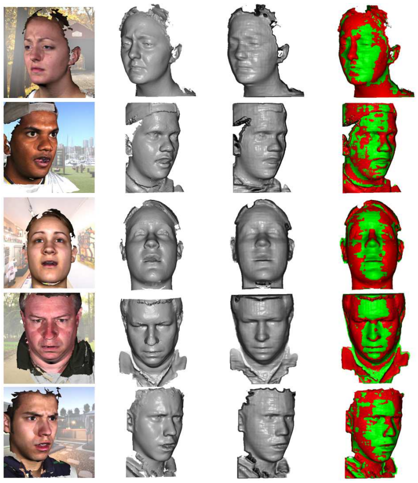


Fig. 6. Pose invariance. Columns one to four in each row respectively depict the input central view of the light field image, the ground truth 3D face, the reconstructed 3D face by our proposed method and the last two overlaid on each other.

on both datasets. Note that the VRN-Guided method incorporates facial landmarks in their proposed VRN architecture whereas we follow a marker-less strategy.

5.1 Qualitative Evaluation

For qualitative evaluation, we show our reconstruction results on light field images synthesized from BU-3DFE [58] and BU-4DFE [60] databases. We also show the ground truth and predicted 3D face shapes overlaid on each other using the Scanalyze software. Figure 6 shows the reconstructed 3D faces under different poses to demonstrate that our method is robust to pose variations. Unlike model based algorithms for 3D face reconstruction [11, 44] from a single RGB image, our method can recover the 3D model of the full head including the peripheral regions such as hair and neck and sometimes even part of the clothing. Figure 6 shows our results under pose invariance while Figures 7 shows our results under exaggerated expressions and illumination changes respectively. Note that our method is robust to variations in pose, expressions and illumination.

We use the code provided by Sela et al. [44] for qualitative comparison of the reconstructed faces. Figure 8 shows 3D faces reconstructed from light field images using our method and 3D faces reconstructed from single central view RGB images using the recent state-of-the-art method proposed by Sela et al. [44]. Since [44] estimate only the facial region, we also crop our reconstructed faces for better visual comparison. As demonstrated, our method produces more visually accurate reconstructions in the global geometry compared to [44]. As compared to methods based on fine-tuning, our method can not capture fine details since we use the output of our network directly without complex post-processing steps. Our proposed method performs better than [44] because, firstly, [44] relies on a face detector and crops the input RGB image based on the detected coordinates while our method does not need any face detection or cropping. Secondly, [44] synthesized their training data from 3DMM parameters and thus

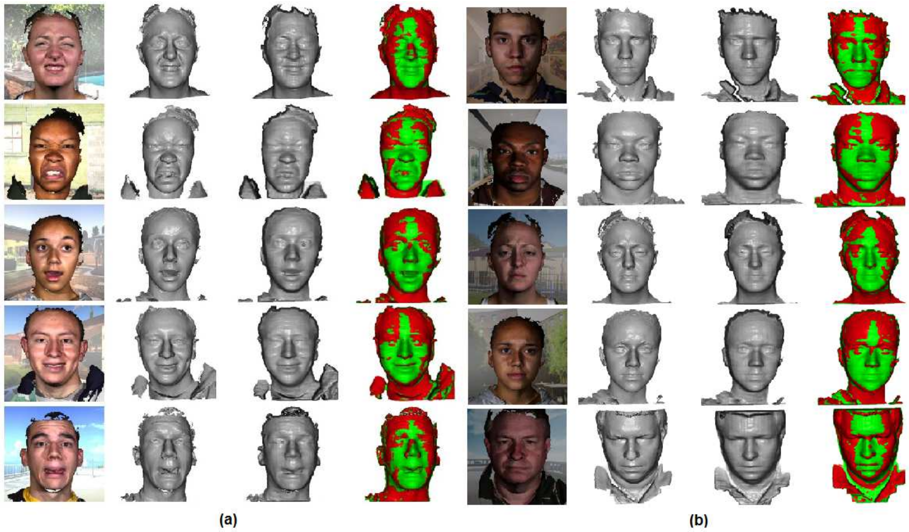


Fig. 7. (a) Expression invariance. As shown, our method can handle exaggerated expressions. (b) Invariance to illumination and skin color. Our method is robust to illumination variations and also works well in the case of dark skin (second row). Columns one to four in each row (in (a) and (b)) respectively depict the input central view of the light field image, the ground truth 3D face, the reconstructed 3D face by our proposed method and the last two overlaid on each other.

their training images do not have the neck and hair regions etc. When the input images are far from the model space, the global face shape will be unsatisfactory at some key facial regions like mouth, nose and eyes as can be seen in Figure 8. Finally, Sela et al. [44] use non-rigid registration to fit the 3DMM to the coarse output of the proposed network. The model fitting process deforms the facial shape when the model and the coarse shape estimated by the network are quite different.

5.2 Quantitative Evaluation

For quantitative comparison, we evaluate the 3D reconstruction on 3,500 light field images of 20 subjects from BU-3DFE [58] and 1,400 light field images of 101 subjects from the BU-4DFE dataset. To measure the affect of pose on the reconstruction accuracy, we use the 3,500 light field images from BU-3DFE dataset. There are 500 light field images for each pose. We use the Root Mean Square Error (RMSE) between the 3D point clouds of the estimated and ground truth reconstructions as a quantitative measure. Results of RMSE for different poses are depicted in Figure 9. Our method is robust to pose variations as the RMSE error increases by only $0.31mm$ when the pose is varied by 30 degrees.

To measure the affect of facial expressions on reconstruction accuracy, we synthesize frontal images in different expressions (Angry, Disgust, Fear, Happy, Sad and Surprise) from the BU-4DFE dataset and measure the reconstruction errors. Figure 10

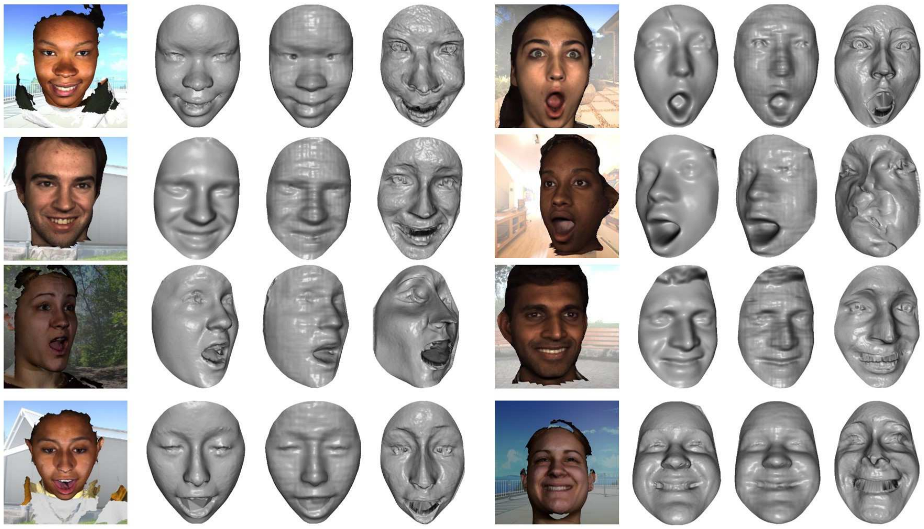


Fig. 8. Qualitative results. The columns contain (in order) central view image, the ground truth 3D face, 3D face reconstructed by our method and 3D face reconstructed by Sela et al. [44].

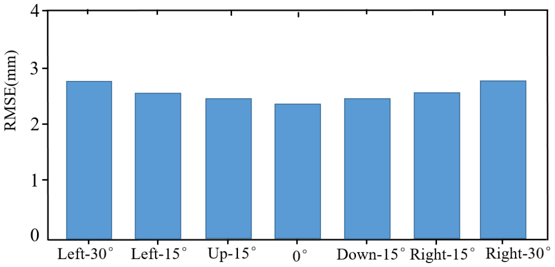


Fig. 9. Reconstruction errors for different facial poses on the BU-3DFE dataset [58]. Note that the RMSE increases from 2.62 to 2.93 (by 0.31 mm only) under extreme pose variations.

shows that the RMSE of 3D face reconstruction from our method is small even in the presence of exaggerated expressions.

We compare the absolute depth error of our proposed method with the state-of-the-art in Table 2, which shows that our proposed 3D reconstruction outperforms all existing methods. We report depth errors evaluated by mean, standard deviation, median and the average ninety percent largest error. Note that for a fair comparison with Sela et al. [44] we report the results obtained on the same dataset directly from their paper instead of calculating the reconstruction errors from our implementation of their work.

We also compare the results of our method with VRN-Guided [25], 3DDFE [63] and EOS [24] methods using the BU-4DFE dataset [60]. We use the Normalized Mean Error (NME) metric proposed by Aarson [25] to report the results for comparison with existing methods. NME is defined as the average per vertex Euclidean distance between the estimated and the ground truth reconstruction normalized by the outer 3D interocu-

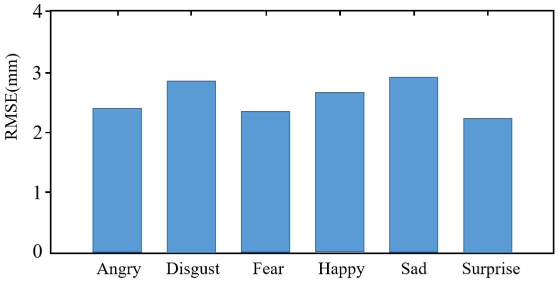


Fig. 10. Reconstruction errors for different facial expressions on the BU-4DFE dataset [60]. The RMSE increases from 2.49 to 2.98 (by only 0.49 mm) under extreme expression variations. Sad has the highest error whereas surprise has the lowest because of more edges around the lips which favors EPI based reconstruction.

	Error in mm			
	Mean	SD	Median	90% largest
Kemelmacher et al.[31]	3.89	4.14	2.94	7.34
Zhu et al.[64]	3.85	3.23	2.93	7.91
Richardson et al.[41]	3.61	2.99	2.72	6.82
Matan et al. [44]	3.51	2.69	2.65	6.59
Ours	**2.78**	**2.04**	**1.73**	**5.30**

Table 2. Comparative results on the BU-3DFE dataset [58]. The absolute RMSE between ground truth and predicted shapes evaluated by mean, standard deviation, median and the average ninety percent largest error of the different methods are presented.

lar distance:

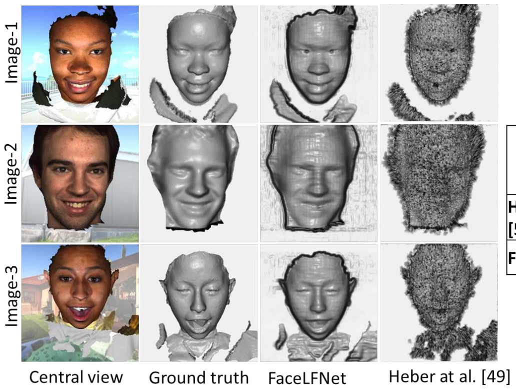
$$\mathrm{NME} = \frac{1}{n}\sum_{k=1}^{n}\frac{\|x_k - y_k\|_2}{d}, \tag{2}$$

where n is the total number of vertices per facial mesh and d is the interocular distance. x_k and y_k represent the coordinates of vertices from the estimated and ground truth meshes respectively. The NME is calculated on the face region only. As shown in Table 3, our method outperforms the state-of-the-art.

	3DDFA[63]	EOS[24]	VRN-Guided[25]	Ours
NME	5.14	5.33	4.71	**3.72**

Table 3. Reconstruction errors on the BU-4DFE dataset [60] in terms of NME defined in Eq. (2). ICP has been used to align the reconstructed face to the ground truth similar to [25].

We also compare our results with [20, 21] using our own implementation of their model as they did not make their codes/ trained models public. We trained the model [20] on synthetic data and then tested it on 10 light field face images of the test data. Figure 5.2 shows three best facial reconstructions by the model of [20]. These reconstructions are extremely noisy with high RMSE. The average reconstruction error of [20] for these 10 images is 27.23 ± 24.7 mm while ours is 2.79 ± 2.6 mm. The main reason for the poor performance of [20] (and [21]) is that the models were designed for 3D reconstruction of scenes where the textures and EPI slopes are drastic. Hence, these methods [49, 50] do not perfrom well at reconstructing 3D faces.



	RMSE ± SD (mm)		
	Image-1	Image-2	Image-3
Heber at al. [50]	17.39 ± 15.0	20.4 ± 17.4	16.53 ± 13.7
FaceLFNet	2.84 ± 2.4	3.15 ± 2.9	2.69 ± 2.1

Comparison of average and SD of RMSE for 3D reconstruction of each face (shown on left).

Fig. 11. Qualitative and quantitative comparison of 3D face reconstruction with Heber at al. [20]

6 Conclusion

We presented a model-free approach for recovering the 3D facial geometry from a single light field image. We proposed FaceLFnet, a densely connected network architecture that regresses the 3D facial curves over the Epipolar Plane Images. Using a curve by curve reconstruction approach, our method needs only a few training samples and yet generalizes well to unseen faces. We proposed a photo-realistic light field image synthesis method to generate a large-scale EPI dataset from a relatively small number of real facial identities. Our results show that 3D face reconstruction from light field images is more accurate and allows the use of a model-free approach which is robust to changes in pose, facial expressions, ethnicities and illumination. We conclude that light field cameras are a more appropriate choice as a passive sensor for 3D face reconstruction since they enjoy similar advantages to conventional RGB cameras in that they are point and shoot, portable and have low cost. These cameras are especially a better choice for medical applications where higher accuracy and model-free approaches are desirable. We will make our trained networks and dataset public which will become the first photo-realistic light field face dataset with ground truth 3D facial scans.

Acknowledgments

This research was partly supported by National Natural Science Foundation of China (NO.61401046, 61733004) and Australian Research Council (ARC) Discovery grant DP160101458. We are grateful to NVIDIA Corporation for donating the Titan Xp GPU used for this research.

References

1. (http://lightfieldstanfordedu/)
2. (https://wwwlytrocom/)
3. (https://wwwraytrixcom/)
4. Abate, A.F., Nappi, M., Riccio, D., Sabatino, G.: 2D and 3D face recognition: A survey. Pattern Recognition Letters **28**(14), 1885–1906 (2007)
5. Aldrian, O., Smith, W.A.: Inverse rendering of faces with a 3d morphable model. IEEE transactions on pattern analysis and machine intelligence **35**(5), 1080–1093 (2013)
6. Belhumeur, P.N., Kriegman, D.J., Yuille, A.L.: The bas-relief ambiguity. International Journal of Computer Vision **35**(1), 33–44 (Nov 1999)
7. 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. pp. 187–194. ACM Press/Addison-Wesley Publishing Co. (1999)
8. Booth, J., Antonakos, E., Ploumpis, S., Trigeorgis, G., Panagakis, Y., Zafeiriou, S.: 3d face morphable models ”in-the-wild”. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
9. Cao, C., Weng, Y., Lin, S., Zhou, K.: 3d shape regression for real-time facial animation. ACM Transactions on Graphics (TOG) **32**(4), 41 (2013)
10. D́Erico, J.: Surface fitting using gridfit. In: MATLAB Central File Exchange (2008)
11. Dou, P., Shah, S.K., Kakadiaris, I.A.: End-to-end 3d face reconstruction with deep neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
12. Fang, T., Zhao, X., Ocegueda, O., Shah, S.K., Kakadiaris, I.A.: 3d/4d facial expression analysis: An advanced annotated face model approach. Image and vision Computing **30**(10), 738–749 (2012)
13. Gilani, S.Z., Mian, A., Eastwood, P.: Deep, dense and accurate 3D face correspondence for generating population specific deformable models. Pattern Recognition **69**, 238–250 (2017)
14. Gilani, S.Z., Mian, A., Shafait, F., Reid, I.: Dense 3D face correspondence. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **40**(7), 1584–1598 (2018)
15. Gilani, S.Z., Rooney, K., Shafait, F., Walters, M., Mian, A.: Geometric facial gender scoring: Objectivity of perception. PloS one **9**(6) (2014)
16. Hammond, P., Forster-Gibson, C., Chudley, A., et al.: Face–brain asymmetry in autism spectrum disorders. Molecular Psychiatry **13**(6), 614–623 (2008)
17. Hammond, P.: The use of 3d face shape modelling in dysmorphology. In: Archives of disease in childhood. p. 92(12) (2007)
18. Han, Y., Lee, J.Y., So Kweon, I.: High quality shape from a single rgb-d image under uncalibrated natural illumination. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1617–1624 (2013)
19. Hassner, T., Harel, S., Paz, E., Enbar, R.: Effective face frontalization in unconstrained images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4295–4304 (2015)
20. Heber, S., Pock, T.: Convolutional networks for shape from light field. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3746–3754 (2016)
21. Heber, S., Yu, W., Pock, T.: U-shaped networks for shape from light field. In: BMVC (2016)
22. Honauer, K., Johannsen, O., Kondermann, D., Goldluecke, B.: A dataset and evaluation methodology for depth estimation on 4d light fields. In: Asian Conference on Computer Vision. pp. 19–34. Springer (2016)
23. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)

24. Huber, P., Hu, G., Tena, R., Mortazavian, P., Koppen, P., Christmas, W.J., Ratsch, M., Kittler, J.: 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)
25. Jackson, A.S., Bulat, A., Argyriou, V., Tzimiropoulos, G.: Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
26. Jeon, H.G., Park, J., Choe, G., Park, J., Bok, Y., Tai, Y.W., So Kweon, I.: Accurate depth map estimation from a lenslet light field camera. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1547–1555 (2015)
27. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 675–678. ACM (2014)
28. Johannsen, O., Sulc, A., Goldluecke, B.: What sparse light field coding reveals about scene structure. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3262–3270 (2016)
29. Jourabloo, A., Liu, X.: Pose-invariant face alignment via cnn-based dense 3d model fitting. International Journal of Computer Vision pp. 1–17 (2017)
30. Kazemi, V., Keskin, C., Taylor, J., Kohli, P., Izadi, S.: Real-time face reconstruction from a single depth image. In: 3D Vision (3DV), 2014 2nd international conference on. vol. 1, pp. 369–376. IEEE (2014)
31. Kemelmacher-Shlizerman, I., Basri, R.: 3d face reconstruction from a single image using a single reference face shape. IEEE transactions on pattern analysis and machine intelligence **33**(2), 394–405 (2011)
32. Li, N., Sun, B., Yu, J.: A weighted sparse coding framework for saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5216–5223 (2015)
33. Li, N., Ye, J., Ji, Y., Ling, H., Yu, J.: Saliency detection on light field. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014)
34. Lin, H., Chen, C., Bing Kang, S., Yu, J.: Depth recovery from light field using focal stack symmetry. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3451–3459 (2015)
35. Marwah, K., Wetzstein, G., Bando, Y., Raskar, R.: Compressive light field photography using overcomplete dictionaries and optimized projections. ACM Transactions on Graphics (TOG) **32**(4), 46 (2013)
36. Mian, A., Bennamoun, M., Owens, R.: An efficient multimodal 2d-3d hybrid approach to automatic face recognition. IEEE transactions on pattern analysis and machine intelligence **29**(11) (2007)
37. Or-El, R., Rosman, G., Wetzler, A., Kimmel, R., Bruckstein, A.M.: Rgbd-fusion: Real-time high precision depth recovery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5407–5416 (2015)
38. Patel, A., Smith, W.A.: 3d morphable face models revisited. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 1327–1334. IEEE (2009)
39. Queirolo, C., Silva, L., Bellon, O., Segundo, M.: 3D face recognition using simulated annealing and the surface interpenetration measure. IEEE TPAMI **32**(2), 206–219 (2010)
40. Richardson, E., Sela, M., Kimmel, R.: 3d face reconstruction by learning from synthetic data. In: 3D Vision (3DV), 2016 Fourth International Conference on. pp. 460–469. IEEE (2016)
41. Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)

42. Roth, J., Tong, Y., Liu, X.: Adaptive 3d face reconstruction from unconstrained photo collections. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4197–4206 (2016)
43. Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3d face analysis. Biometrics and identity management pp. 47–56 (2008)
44. Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
45. Sepas-Moghaddam, A., Chiesa, V., Correia, P.L., Pereira, F., Dugelay, J.L.: The ist-eurecom light field face database. In: Biometrics and Forensics (IWBF), 2017 5th International Workshop on. pp. 1–6. IEEE (2017)
46. Sheng, H., Zhao, P., Zhang, S., Zhang, J., Yang, D.: Occlusion-aware depth estimation for light field using multi-orientation epis. Pattern Recognition (2017)
47. Tan, D.W., Gilani, S.Z., Maybery, M.T., Mian, A., Hunt, A., Walters, M., Whitehouse, A.J.: Hypermasculinised facial morphology in boys and girls with autism spectrum disorder and its association with symptomatology. Scientific Reports **7**(1), 9348 (2017)
48. Tao, M.W., Srinivasan, P.P., Malik, J., Rusinkiewicz, S., Ramamoorthi, R.: Depth from shading, defocus, and correspondence using light-field angular coherence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1940–1948 (2015)
49. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2387–2395 (2016)
50. Tian, J., Murez, Z., Cui, T., Zhang, Z., Kriegman, D., Ramamoorthi, R.: Depth and image restoration from light field in a scattering medium. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
51. Tuan Tran, A., Hassner, T., Masi, I., Medioni, G.: Regressing robust and discriminative 3d morphable models with a very deep neural network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
52. Wang, T.C., Efros, A.A., Ramamoorthi, R.: Occlusion-aware depth estimation using light-field cameras. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3487–3495 (2015)
53. Wang, T.C., Zhu, J.Y., Hiroaki, E., Chandraker, M., Efros, A.A., Ramamoorthi, R.: A 4d light-field dataset and cnn architectures for material recognition. In: European Conference on Computer Vision. pp. 121–138. Springer (2016)
54. Wanner, S., Goldluecke, B.: Variational light field analysis for disparity estimation and super-resolution. IEEE transactions on pattern analysis and machine intelligence **36**(3), 606–619 (2014)
55. Whitehouse, A.J., Gilani, S.Z., Shafait, F., Mian, A., Tan, D.W., Maybery, M.T., Keelan, J.A., Hart, R., Handelsman, D.J., Goonawardene, M., et al.: Prenatal testosterone exposure is related to sexually dimorphic facial morphology in adulthood. In: Proc. R. Soc. B. vol. 282. The Royal Society (2015)
56. Wu, G., Zhao, M., Wang, L., Dai, Q., Chai, T., Liu, Y.: Light field reconstruction using deep convolutional network on epi. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
57. Xiong, Z., Wang, L., Li, H., Liu, D., Wu, F.: Snapshot hyperspectral light field imaging. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
58. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on. pp. 211–216. IEEE (2006)

59. Zhang, S., Sheng, H., Li, C., Zhang, J., Xiong, Z.: Robust depth estimation for light field via spinning parallelogram operator. Computer Vision and Image Understanding **145**, 148–159 (2016)
60. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P.: A high-resolution spontaneous 3d dynamic facial expression database. In: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. pp. 1–6. IEEE (2013)
61. Zhao, W.Y., Chellappa, R.: Illumination-insensitive face recognition using symmetric shape-from-shading. In: Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on. vol. 1, pp. 286–293. IEEE (2000)
62. Zhu, H., Zhang, Q., Wang, Q.: 4d light field superpixel and segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
63. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 146–155 (2016)
64. Zhu, X., Lei, Z., Yan, J., Yi, D., Li, S.Z.: High-fidelity pose and expression normalization for face recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 787–796 (2015)
65. Zulqarnain Gilani, S., Mian, A.: Learning from millions of 3D scans for large-scale 3D face recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)