

Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network

Yao Feng¹[0000-0002-9481-9783], Fan Wu²[0000-0003-1970-3470], Xiaohu Shao^{3,4}[0000-0003-1141-6020], Yanfeng Wang¹[0000-0002-3196-2347], and Xi Zhou^{1,2}[0000-0003-2917-0436]

¹ Cooperative Medianet Innovation Center, Shanghai Jiao Tong University

² CloudWalk Technology

³ CIGIT, Chinese Academy of Sciences

⁴ University of Chinese Academy of Sciences

Abstract. We propose a straightforward method that simultaneously reconstructs the 3D facial structure and provides dense alignment. To achieve this, we design a 2D representation called UV position map which records the 3D shape of a complete face in UV space, then train a simple Convolutional Neural Network to regress it from a single 2D image. We also integrate a weight mask into the loss function during training to improve the performance of the network. Our method does not rely on any prior face model, and can reconstruct full facial geometry along with semantic meaning. Meanwhile, our network is very light-weighted and spends only 9.8ms to process an image, which is extremely faster than previous works. Experiments on multiple challenging datasets show that our method surpasses other state-of-the-art methods on both reconstruction and alignment tasks by a large margin. Code is available at <https://github.com/YadiraF/PRNet>.

Keywords: 3D Face Reconstruction · Dense Face Alignment

1 Introduction

3D face reconstruction and face alignment are two fundamental and highly related topics in computer vision. In the last decades, researches in these two fields benefit each other. In the beginning, face alignment that aims at detecting a special 2D fiducial points [66, 64, 38, 46] is commonly used as a prerequisite for other facial tasks such as face recognition [59] and assists 3D face reconstruction [68, 27] to a great extent. However, researchers find that 2D alignment has difficulties [65, 30] in dealing with problems of large poses or occlusions. With the development of deep learning, many computer vision problems have been well solved by utilizing Convolution Neural Networks (CNNs). Thus, some works start to use CNNs to estimate the 3D Morphable Model (3DMM) coefficients [32, 67, 47, 39, 48, 40] or 3D model warping functions [4, 53] to restore the corresponding 3D information from a single 2D facial image, which provides both dense face



Fig. 1: The qualitative results of our method. Odd row: alignment results (only 68 key points are plotted for display). Even row: 3D reconstruction results (reconstructed shapes are rendered with head light for better view).

alignment and 3D face reconstruction results. However, the performance of these methods is restricted due to the limitation of the 3D space defined by face model basis or templates. The required operations including perspective projection or 3D Thin Plate Spline (TPS) transformation also add complexity to the overall process.

Recently, two end-to-end works [28] [9], which bypass the limitation of model space, achieve the state-of-the-art performances on their respective tasks. [9] trains a complex network to regress 68 facial landmarks with 2D coordinates from a single image, but needs an extra network to estimate the depth value. Besides, dense alignment is not provided by this method. [28] develops a volumetric representation of 3D face and uses a network to regress it from a 2D image. However, this representation discards the semantic meaning of points, thus the network needs to regress the whole volume in order to restore the facial shape, which is only part of the volume. So this representation limits the resolution of the recovered shape, and need a complex network to regress it. To sum up, model-based methods keep semantic meaning of points well but are restricted in model space, recent model-free methods are unrestricted and achieve state-of-the-art performance but discard the semantic meaning, which motivate us to find a new approach to reconstruct 3D face with alignment information in a model-free manner.

In this paper, we propose an end-to-end method called Position map Regression Network (PRN) to jointly predict dense alignment and reconstruct 3D face shape. Our method surpasses all other previous works on both 3D face alignment and reconstruction on multiple datasets. Meanwhile, our method is straightforward with a very light-weighted model which provides the result in one pass with

9.8ms. All of these are achieved by the elaborate design of the 2D representation of 3D facial structure and the corresponding loss function. Specifically, we design a UV position map, which is a 2D image recording the 3D coordinates of a complete facial point cloud, and at the same time keeping the semantic meaning at each UV place. We then train a simple encoder-decoder network with a weighted loss that focuses more on discriminative region to regress the UV position map from a single 2D facial image. Figure 1 shows our method is robust to poses, illuminations and occlusions.

In summary, our main contributions are:

- For the first time, we solve the problems of face alignment and 3D face reconstruction together in an end-to-end fashion without the restriction of low-dimensional solution space.
- To directly regress the 3D facial structure and dense alignment, we develop a novel representation called UV position map, which records the position information of 3D face and provides dense correspondence to the semantic meaning of each point on UV space.
- For training, we proposed a weight mask which assigns different weight to each point on position map and compute a weighted loss. We show that this design helps improving the performance of our network.
- We finally provide a light-weighted framework that runs at over 100FPS to directly obtain 3D face reconstruction and alignment result from a single 2D facial image.
- Comparison on the AFLW2000-3D and Florence datasets shows that our method achieves more than 25% relative improvements over other state-of-the-art methods on both tasks of 3D face reconstruction and dense face alignment.

2 Related Works

2.1 3D Face Reconstruction

Since Blanz and Vetter proposed 3D Morphable Model(3DMM) in 1999[6], methods based on 3DMM are popular in completing the task of monocular 3D face reconstruction. Most of earlier methods are to establish the correspondences of the special points between input images and the 3D template including landmarks[37, 68, 56, 27, 10, 29, 19] and local features[26, 49, 19], then solve the non-linear optimization function to regress the 3DMM coefficients. However, these methods heavily rely on the accuracy of landmarks or other feature points detector. Thus, some methods[22, 63] firstly use CNNs to learn the dense correspondence between input image and 3D template, then calculate the 3DMM parameters with predicted dense constrains. Recent works also explore the usage of CNN to predict 3DMM parameters directly. [32, 67, 47, 39, 48] use cascaded CNN structure to regress the accurate 3DMM coefficients, which take a lot of time due to iterations. [15, 57, 31, 36] propose end-to-end CNN architectures to directly estimate the 3DMM shape parameters. Unsupervised methods have been

also researched recently, [55, 3] can regress the 3DMM coefficients without the help of training data, which performs badly in faces with large poses and strong occlusions. However, the main defect of those methods is model-based, resulting in a limited geometry which is constrained in model space. Some other methods can reconstruct 3D faces without 3D shape basis, [24, 33, 20, 53, 51] can produce a 3D structure by warping the shape of a reference 3D model. [4] also reconstruct the 3D shape of faces by learning a 3D Thin Plate Spline(TPS) warping function via a deep network which warps a generic 3D model to a subject specific 3D shape. Obviously, the reconstructed face geometry from these methods are also restricted by the reference model, which means the structure differs when the template changes. Recently, [28] propose to straightforwardly map the image pixels to full 3D facial structure via volumetric CNN regression. This method is not restricted in the model space any more, while needs a complex network structure and a lot of time to predict the voxel data. Different from above methods, Our framework is model-free and light-weighted, can run at real time and directly obtain the full 3D facial geometry along with its correspondence information.

2.2 Face Alignment

In the field of computer vision, face alignment is a long-standing problem which attracts lots of attention. In the beginning, there are a number of 2D facial alignment approaches which aim at locating a set of fiducial 2D facial landmarks, such as classic Active Appearance Model(AMM)[43, 52, 58] and Constrained Local Models(CLM)[34, 1]. Then cascaded regression[14, 60] and CNN-based methods[38, 46, 9] are largely used to achieve state-of-the-art performance in 2D landmarks location. However, 2D landmarks location only regresses visible points on faces, which is limited to describe face shape when the pose is large. Recent works then research the 3D facial alignment, which begins with fitting a 3DMM[44, 67, 18] or registering a 3D facial template[51, 5] with a 2D facial image. Obviously, 3D reconstruction methods based on model can easily complete the task of 3D face alignment. Actually, [67, 63, 31] are specially designated methods to achieve 3D face alignment by means of 3DMM fitting. Recently [8, 9] use a deep network to directly predict the heat map to obtain the 3D facial landmarks and achieves state-of-the-art performance. Thus, as sparse face alignment tasks are highly completed by aforementioned methods, the task of dense face alignment begins to develop. Notice that, the dense face alignment means the methods should offer the correspondence between two face images as well as between a 2D facial image and a 3D facial reference geometry. [40] use multi-constraints to train a CNN which estimates the 3DMM parameters and then provides a very dense 3D alignment. [22, 63] directly learn the correspondence between 2D input image and 3D template via a deep network, while those correspondence is not complete, only visible face region is considered. Compared to prior works, our method can directly establish the dense correspondence of all regions once the position map is regressed. No intermediate parameters such as 3DMM coefficients and TPS warping parameters are needed in our method, which means our network can run very fast.

3 Proposed Method

This section describes the framework and the details of our proposed method. Firstly, we introduce the characteristics of the position map for our representation. Then we elaborate the CNN architecture and the loss function designed specially for learning the mapping from unconstrained RGB image to its 3D structure. The implementation details of our method are shown in the last subsection.

3.1 3D Face Representation

Our goal is to regress the 3D facial geometry and its dense correspondence information from a single 2D image. Thus we need a proper representation which can be directly predicted via a deep network. One simple and commonly used idea is to concatenate the coordinates of all points in 3D face as a vector and use a network to predict it. However, this projection from 3D space into 1D vector which discards the spatial adjacency information among points increases the difficulties in training deep neural networks. Spatially adjacent points could share weights in predicting their positions, which can be easily achieved by using convolutional layers, while the coordinates as a 1D vector needs a fully connected layer to predict each point with much more parameters that increases the network size and is hard to train. [16] proposed a point set generation network to directly predict the point cloud of 3D object as a vector from a single image. However, the max number of points is only 1024, far from enough to represent an accurate 3D face. So model-based methods[67, 15, 40] regress a few model parameters rather than the coordinates of points, which usually needs special care in training such as using Mahalanobis distance and inevitably limits the estimated face geometry to the their model space. [28] proposed 3D binary volume as the representation of 3D structure and uses Volumetric Regression Network(VRN) to output a $192 \times 192 \times 200$ volume as the discretized version of point cloud. By using this representation, VRN can be built with full convolutional layers. However, discretization limits the resolution of point cloud, and most part of the network output correspond to non-surface points which are of less usage.

To address the problems in previous works, we propose UV position map as the presentation of full 3D facial structure with alignment information. UV position map or position map for short, is a 2D image recording 3D positions of all points in UV space. In the past years, UV space or UV coordinates, which is a 2D image plane parameterized from the 3D surface, has been utilized as a way to express information including the texture of faces(texture map) [3, 13, 45, 61], 2.5D geometry(height map)[41, 42], 3D geometry(geometry image)[21, 54] and the correspondences between 3D facial meshes[7]. Different from previous works, we use UV space to store the 3D position of points from 3D face model aligned with corresponding 2D facial image. As shown in Figure 2, we assume the projection from 3D model to 2D image is weak perspective projection and define the 3D facial position in Left-handed Cartesian Coordinate system. The origin of the 3D space overlaps with the upper-left of the input image, with

the positive x-axis pointing to the right of the image and minimum z at origin. The ground truth 3D facial shape exactly matches the face in the 2D image when projected to the x-y plane. Thus the position map can be expressed as $Pos(u_i, v_i) = (x_i, y_i, z_i)$, where (u_i, v_i) represents the UV coordinate of i th point in face surface and (x_i, y_i, z_i) represents the corresponding 3D position of facial structure with (x_i, y_i) representing corresponding 2D position of face in the input RGB images and z_i representing the depth of this point. Note that, (u_i, v_i) and (x_i, y_i) represent the same position of face so alignment information can be reserved. Our position map can be easily comprehended as replacing the r, g, b value in texture map by x, y, z coordinates.

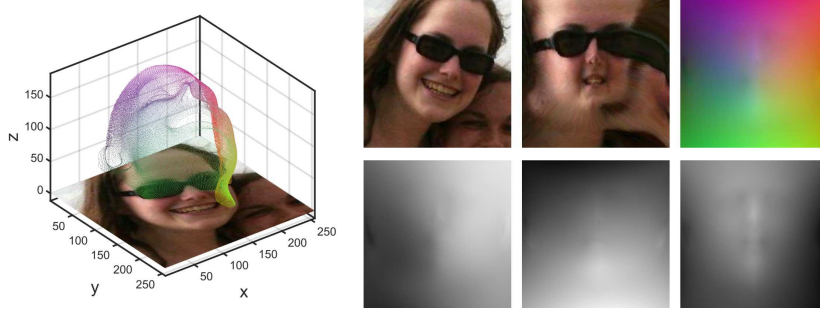


Fig. 2: The illustration of UV position map. Left: 3D plot of input image and its corresponding aligned 3D point cloud(as ground truth). Right: The first row is the input 2D image, extracted UV texture map and corresponding UV position map. The second row is the x, y, z channel of the UV position map.

Thus our position map records a dense set of points from 3D face with its semantic meaning, we are able to simultaneously obtain the 3D facial structure and dense alignment result by using a CNN to regress the position map directly from unconstrained 2D images. The network architecture in our method could be greatly simplified due to this convenience. Notice that the position map contains the information of the whole face, which makes it different from other 2D representations such as Projected Normalized Coordinate Code(PNCC)[67, 48], an ordinary depth image[53] or quantized UV coordinates[22], which only reserve the information of visible face region in the input image. Our proposed position map also infers the invisible parts of face, thus our method can predict a complete 3D face.

Since we want to regress the 3D full structure from 2D image directly, the unconstrained 2D facial images and their corresponding 3D shapes are needed for end-to-end training. 300W-LP[67] is a large dataset that contains more than 60K unconstrained images with fitted 3DMM parameters, which is suitable to form our training pairs. Besides, the 3DMM parameters of this dataset are based on the Basel Face Model(BFM)[6]. Thus, in order to make full use of this dataset,

we conduct the UV coordinates corresponding to BFM. To be specific, we use the parameterized UV coordinates from [3] which computes a Tutte embedding[17] with conformal Laplacian weight and then maps the mesh boundary to a square. Since the number of vertices in BFM is more than 50K, we choose 256 as the position map size, which get a high precision point cloud with negligible re-sample error.

3.2 Network Architecture and Loss Function

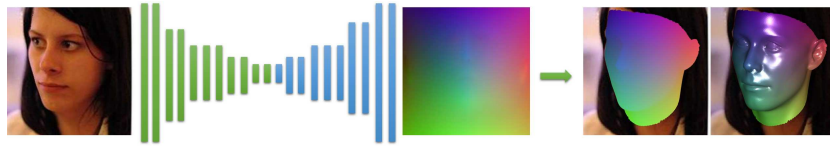


Fig. 3: The architecture of PRN. The Green rectangles represent the residual blocks, and the blue ones represent the transposed convolutional layers.

Since our network transfers the input RGB image into position map image, we employ an encoder-decoder structure to learn the transfer function. The encoder part of our network begins with one convolution layer followed by 10 residual blocks[25] which reduce the $256 \times 256 \times 3$ input image into $8 \times 8 \times 512$ feature maps, the decoder part contains 17 transposed convolution layers to generate the predicted $256 \times 256 \times 3$ position map. We use kernel size of 4 for all convolution or transposed convolution layers, and use ReLU layer for activation. Given that the position map contains both the full 3D information and dense alignment result, we don't need extra network module for multi-task during training or inferring. The architecture of our network is shown in Figure 3.

In order to learn the parameters of the network, we build a loss function to measure the difference between ground truth position map and the network output. Mean square error (MSE) is a commonly used loss for such learning task, such as in [63, 12]. However, MSE treats all points equally, so it is not entirely appropriate for learning the position map. Since central region of face has more discriminative features than other regions, we employ a weight mask to form our loss function. As shown in Figure 4, the weight mask is a gray image recording the weight of each point on position map. It has the same size and pixel-to-pixel correspondence to position map. According to our objective, we separate points into four categories, each has its own weights in the loss function. The position of 68 facial keypoints has the highest weight, so that to ensure the network to learn accurate locations of these points. The neck region usually attracts less attention, and is often occluded by hairs or clothes in unconstrained images. Since learning the 3D shape of neck or clothes is beyond our interests, we assign 0 weight to points in neck region to reduce disturbance in the training process.

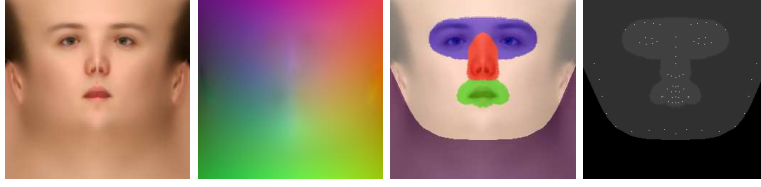


Fig. 4: The illustration of weight mask. From left to right: UV texture map, UV position map, colored texture map with segmentation information (blue for eye region, red for nose region, green for mouth region and purple for neck region), the final weight mask.

Thus, we denote the predicted position map as $Pos(u, v)$ for u, v representing each pixel coordinate. Given the ground truth position map $\tilde{Pos}(u, v)$ and weight mask $W(u, v)$, our loss function is defined as:

$$Loss = \sum \|Pos(u, v) - \tilde{Pos}(u, v)\| \cdot W(u, v) \quad (1)$$

Specifically, We use following weight ratio in our experiments, subregion1 (68 facial landmarks): subregion2 (eye, nose, mouth): subregion3 (other face area): subregion4 (neck) = 16:4:3:0. The final weight mask is shown in Figure 4.

3.3 Training Details

As described above, we choose **300W-LP**[67] to form our training sets, since it contains face images across different angles with the annotation of estimated 3DMM coefficients, from which the 3D point cloud could be easily generated. Specifically, we crop the images according the ground truth bounding box and rescale them to size 256×256 . Then utilize their annotated 3DMM parameters to generate the corresponding 3D position, and render them into UV space to obtain the ground truth position map, the map size in our training is also 256×256 , which means a precision of more than 45K point cloud to regress. Notice that, although our training data is generated from 3DMM, our network’s output, the position map is not restricted to any face template or linear space of 3DMM.

We perturb the training set by randomly rotating and translating the target face in 2D image plane. To be specific, the rotation is from -45 to 45 degree angles, translation changes is random from 10 percent of input size, and scale is from 0.9 to 1.2. Like [28], we also augment our training data by scaling color channels with scale range from 0.6 to 1.4. In order to handle images with occlusions, we synthesize occlusions by adding noise texture into raw images, which is similar to the work of [50, 63]. With all above augmentation operations, our training data covers all the difficult cases. We use the network described in section 3 to train our model. For optimization, we use Adam optimizer with a learning rate begins at 0.0001 and decays half after each 5 epochs. The batch size is set as 16.

4 Experimental Results

In this part, we evaluate the performance of our proposed method on the tasks of 3D face alignment and 3D face reconstruction. We first introduce the test datasets used in our experiments in section 4.1. Then in section 4.2 and 4.3 we compare our results with other methods in both quantitative and qualitative way. We then compare our method’s runtime with other methods in section 4.4. In the end, the ablation study is conducted in section 4.5 to evaluate the effect of weight mask in our method.

4.1 Test Dataset

To evaluate our performance on the task of dense alignment and 3D face reconstruction, multiple test datasets listed below are used in our experiments:

AFLW2000-3D is constructed by [67] to evaluate 3D face alignment on challenging unconstrained images. This database contains the first 2000 images from AFLW[35] and expands its annotations with fitted 3DMM parameters and 68 3D landmarks. We use this database to evaluate the performance of our method on both face reconstruction and face alignment tasks.

AFLW-LFPA is another extension of AFLW dataset constructed by [32]. By picking images from AFLW according to the poses, the authors construct this dataset which contains 1299 test images with a balanced distribution of yaw angle. Besides, each image is annotated with 13 additional landmarks as a expansion to only 21 visible landmarks in AFLW. This database is evaluated on the task of 3D face alignment. We use 34 visible landmarks as the ground truth to measure the accuracy of our results.

Florence is a 3D face dataset that contains 53 subjects with its ground truth 3D mesh acquired from a structured-light scanning system[2]. On experiments, each subject generates renderings with different poses as the same with [28]: a pitch of -15, 20 and 25 degrees and spaced rotations between -80 and 80. We compare the performance of our method on face reconstruction against other very recent state-of-the-art methods VRN-Guided[28] and 3DDFA[67] on this dataset.

4.2 3D Face Alignment

To evaluate the face alignment performance. We employ the Normalized Mean Error(NME) to be the evaluation metric, bounding box size is used as the normalization factor. Firstly, we evaluate our method on a sparse set of 68 facial landmarks, and compare our result with 3DDFA[67], DeFA[40] and 3D-FAN[9] on dataset AFLW2000-3D. As shown in figure 5, our result slightly outperforms the state-of-the-art method 3D-FAN when calculating per distance with 2D coordinates. When considering the depth value, the performance discrepancy between our method and 3D-FAN increases. Notice that, the 3D-FAN needs another network to predict the z coordinate of landmarks, while the depth value can be obtained directly in our method.

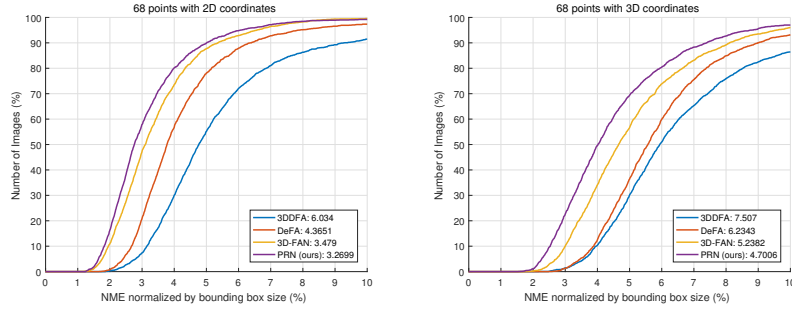


Fig. 5: Cumulative Errors Distribution (CED) curves on AFLW2000-3D. Evaluation is performed on 68 landmarks with both the 2D(left) and 3D(right) coordinates. Overall 2000 images from AFLW2000-3D dataset are used here. The mean NME% of each method is also showed in the legend.

To further investigate the performance of our method across poses and datasets, we also report the NME with small, medium and large yaw angles on AFLW2000-3D dataset and the mean NME on both AFLW2000-3D and AFLW-LPFA datasets. Table 1 shows the results, note that the numerical values are recorded from their published papers. Follow the work [67], we also randomly select 696 faces from AFLW2000 to balance the distribution. The result shows that our method is robust to changes of pose and datasets. Although all the state-of-the-art methods of 3D face alignment conduct evaluation on AFLW2000-3D dataset, the ground truth is still controversial[63, 9] due to its annotation pipeline which is based on Landmarks Marching method[68]. Thus, we visualize some results in Figure 6 that have NME larger than 6.5% and we find our results are more accurate than the ground truth in some cases. We also compare our dense alignment re-

Table 1: Performance comparison on AFLW2000-3D(68 landmarks) and AFLW-LPFA(34 visible landmarks). The NME (%) for faces with different yaw angles are reported. The first best result in each category is highlighted in bold, the lower is the better.

Method	AFLW2000-3D				AFLW-LPFA
	0 to 30	30 to 60	60 to 90	Mean	Mean
SDM[60]	3.67	4.94	9.67	6.12	-
3DDFA [67]	3.78	4.54	7.93	5.42	-
3DDFA + SDM [67]	3.43	4.24	7.17	4.94	-
PAWF[32]	-	-	-	-	4.72
Yu et al. [63]	3.62	6.06	9.56	-	-
3DSTN[4]	3.15	4.33	5.98	4.49	-
DeFA[40]	-	-	-	4.50	3.86
PRN (ours)	2.75	3.51	4.61	3.62	2.93

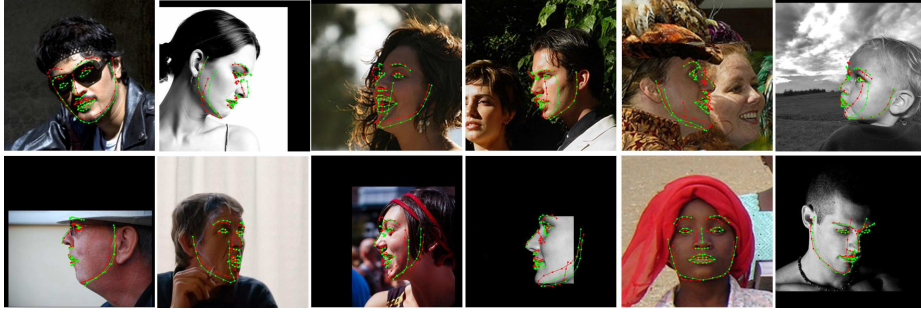


Fig. 6: Examples from AFLW2000-3D dataset show that our predictions are more accurate than ground truth in some cases. Green: predicted landmarks by our method. Red: ground truth from [67].

sults against other methods including 3DDFA[67] and DeFA[40] on the only test dataset AFLW2000-3D. In order to compare different methods with the same set of points, we select the points from the largest common face region provided by all methods, and finally around 45K points were used for the evaluation. As shown in figure 7, our method outperforms the best methods with a large margin of more than **27%** on both 2D and 3D coordinates.

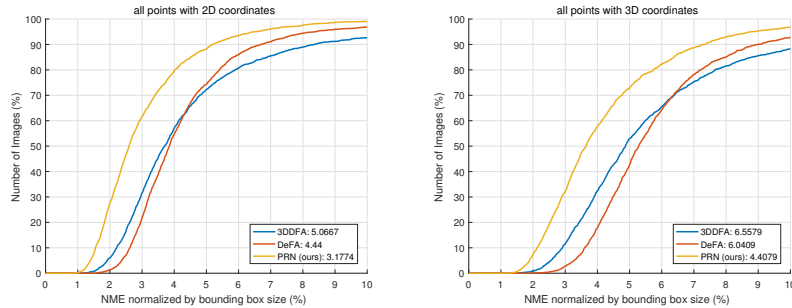


Fig. 7: CED curves on AFLW2000-3D. Evaluation is performed on all points with both the 2D (left) and 3D (right) coordinates. Overall 2000 images from AFLW2000-3D dataset are used here. The mean NME% is showed in the legend.

4.3 3D Face Reconstruction

In this part, we evaluate our method on 3D face reconstruction task and compare with 3DDFA[67], DeFA[40] and VRN-Guided[28] on AFLW2000-3D and Florence datasets. We use the same set of points as in evaluating dense alignment

and changes the metric so as to keep consistency with other 3D face reconstruction evaluation methods. We first use Iterative Closest Points(ICP) algorithm to find the corresponding nearest points between the network output and ground truth point cloud, then calculate Mean Squared Error(MSE) normalized by outer interocular distance of 3D coordinates.

The result is shown in figure 8. our method greatly exceeds the performance of other two state-of-the-art methods. Since AFLW2000-3D dataset is labeled

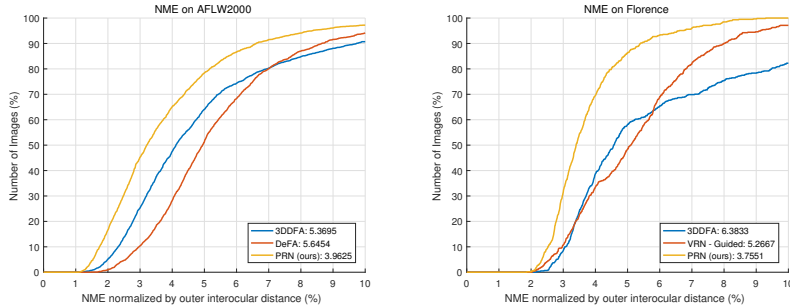


Fig. 8: 3D reconstruction performance(CED curves) on in-the-wild AFLW2000-3D dataset and Florence dataset. The mean NME% of each method is showed in the legend. On AFLW2000-3D, more than 45K points are used for evaluation. On Florence, about 19K points are used.

with results from 3DMM fitting, we further evaluate the performance of our method on Florence dataset, where ground truth 3D point cloud is obtained from structured-light 3D scanning system. Here we compare our method with 3DDFA and VRN-Guided[28], using experimental settings in [28]. The evaluation images are the renderings with different poses from Florence database, we calculate the bounding box from the ground truth point cloud and using the cropped image as network input. Although our method output more complete face point clouds than VRN, we only choose the common face region to compare the performance, 19K points are used for the evaluation. Figure 8 shows that our method achieves **28.7%** relative higher performance compared to VRN-Guided on Florence dataset, which is a significant improvement.

To better evaluate the reconstruction performance of our method across different poses, we calculated the NME for different yaw angle range. As shown in figure 9, all the methods perform well in near frontal view, however, 3DDFA and VRN-Guided fail to keep low error as pose becomes large, while our method keeps relatively stable performance in all pose ranges. We also illustrate the qualitative comparison in figure 9, our restored point cloud covers a larger region than in VRN-Guided, which ignores the lateral facial parts. Besides, due to the limitation on resolution of VRN, our method provides finer details of face, especially on the nose and mouth region.

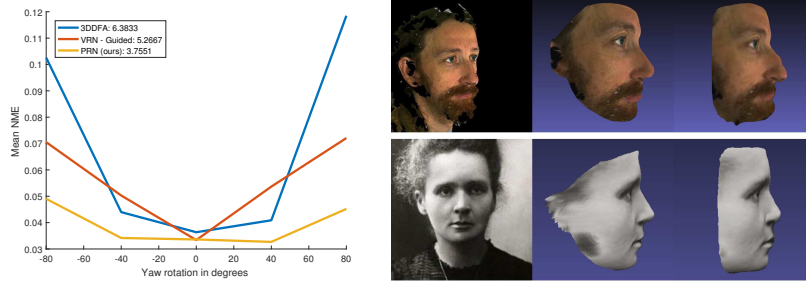


Fig. 9: Left: CED curves on Florence dataset with different yaw angles. Right: the qualitative comparison with VRN-Guided. The first column is the input images from Florence dataset and the Internet, the second column is the reconstructed face from our method, the third column is the results from VRN.

We also provide additional quantitative results on BU-3DFE[62] and qualitative results on 300VW[11] and Multi-PIE[23] datasets, please refer to supplementary material for full details.

4.4 Runtime

Surpassing the performance of all other state-of-the-art methods on 3D face alignment and reconstruction, our method is surprisingly more light-weighted and faster. Since our network uses basic encoder-decoder structure, our model size is only 160MB compared to 1.5GB in VRN[28]. We also compare the runtime, Table 2 shows the result. The results of 3DDFA and 3DSTN are directly recorded from their published papers and others are recorded by running their publicly available source codes. Notice that, We measure the run time of the process which is defined from inputting the cropped face image until recovering the 3D geometry(point cloud, mesh or voxel data) for 3D reconstruction methods or obtaining the 3D landmarks for alignment methods. The hardware used for evaluation is an NVIDIA GeForce GTX 1080 GPU and an Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz. Specifically, DeFA needs 11.8ms(GPU) to predict

Table 2: Run time in Milliseconds per Image

3DDFA[67]	DeFA[40]	3D-FAN[9]	3DSTN[4]	VRN-Guided[28]	PRN (ours)
75.7	35.4	54.7	19.0	69.0	9.8

3DMM parameters and another 23.6ms(CPU) to generate mesh data from predicted parameters, 3DFAN needs 29.1ms(GPU) to estimate 2D coordinates first and 25.6ms(GPU) to obtain depth value, VRN-Guided detects 68 2D landmarks

with 28.4ms(GPU), then regress the voxel data with 40.6ms(GPU), our method provides both 3D reconstruction and dense alignment result from cropped image in one pass in 9.8ms(GPU).

4.5 Ablation Study

In this section, we conduct several experiments to evaluate the influence of our weight mask on training and provide both sparse and dense alignment CED on AFLW2000 to evaluate different settings. Specifically, we experimented with three different weight ratios: (1) weight ratio 1 = 1:1:1:1, (2) weight ratio 2 = 1:1:1:0, (3) weight ratio 3 = 16:4:3:0. We could see that weight ratio 1 corresponds to the situation when no weight mask is used, weight ratio 2 and 3 are slightly different on the emphasis in loss function.

The results are shown in Figure 10. Network trained without using weight mask has worst performance compared with other two settings. By adding weights to specific regions such as 68 facial landmarks or central face region, weight ratio 3 shows considerable improvement on 68 points datasets over weight ratio 2.

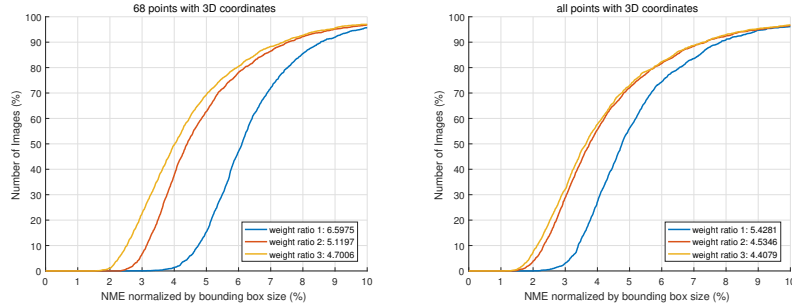


Fig. 10: The effect of weight mask evaluated on AFLW2000-3D dataset with 68 landmarks(left) and all points(right).

5 Conclusion

In this paper, we propose an end-to-end method, which well solves the problems of 3D face alignment and 3D face reconstruction simultaneously. By learning the position map, we directly regress the complete 3D structure along with semantic meaning from a single image. Quantitative and qualitative results demonstrate our method is robust to poses, illuminations and occlusions. Experiments on three test datasets show that our method achieves significant improvements over others. We further show that our method runs faster than other methods and is suitable for real time usage.

References

1. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: *Computer Vision and Pattern Recognition (CVPR)*, 2013 IEEE Conference on. pp. 3444–3451. IEEE (2013)
2. Bagdanov, A.D., Del Bimbo, A., Masi, I.: The florence 2d/3d hybrid face dataset. In: *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. pp. 79–80. ACM (2011)
3. Bas, A., Huber, P., Smith, W.A.P., Awais, M., Kittler, J.: 3d morphable models as spatial transformer networks. In: *ICCV 2017 Workshop on Geometry Meets Deep Learning* (2017)
4. Bhagavatula, C., Zhu, C., Luu, K., Savvides, M.: Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In: *The IEEE International Conference on Computer Vision (ICCV)*. vol. 2, p. 7 (2017)
5. de Bittencourt Zavan, F.H., Nascimento, A.C., e Silva, L.P., Bellon, O.R., Silva, L.: 3d face alignment in the wild: A landmark-free, nose-based approach. In: *European Conference on Computer Vision*. pp. 581–589. Springer (2016)
6. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. *international conference on computer graphics and interactive techniques* pp. 187–194 (1999)
7. Booth, J., Zafeiriou, S.: Optimal uv spaces for facial morphable model construction. In: *Image Processing (ICIP)*, 2014 IEEE International Conference on. pp. 4672–4676. IEEE (2014)
8. Bulat, A., Tzimiropoulos, G.: Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In: *European Conference on Computer Vision*. pp. 616–624. Springer (2016)
9. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d and 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks) (2017)
10. Cao, C., Hou, Q., Zhou, K.: Displaced dynamic expression regression for real-time facial tracking and animation. *ACM* (2014)
11. Chrysos, G.G., Antonakos, E., Zafeiriou, S., Snape, P.: Offline deformable face tracking in arbitrary videos. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. pp. 1–9 (2015)
12. Crispell, D., Bazik, M.: Pix2face: Direct 3d face model estimation (2017)
13. Deng, J., Cheng, S., Xue, N., Zhou, Y., Zafeiriou, S.: Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. *arXiv preprint arXiv:1712.04695* (2017)
14. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: *Computer Vision and Pattern Recognition (CVPR)*, 2010 IEEE Conference on. pp. 1078–1085. IEEE (2010)
15. Dou, P., Shah, S.K., Kakadiaris, I.A.: End-to-end 3d face reconstruction with deep neural networks (2017)
16. Fan, H., Su, H., Guibas, L.: A point set generation network for 3d object reconstruction from a single image pp. 2463–2471 (2016)
17. Floater, M.S.: Parametrization and smooth approximation of surface triangulations. *Computer Aided Geometric Design* **14**(3), 231–250 (1997)
18. Gou, C., Wu, Y., Wang, F.Y., Ji, Q.: Shape augmented regression for 3d face alignment. In: *European Conference on Computer Vision*. pp. 604–615. Springer (2016)
19. Grewe, C.M., Zachow, S.: Fully automated and highly accurate dense correspondence for facial surfaces. In: *European Conference on Computer Vision*. pp. 552–568. Springer (2016)

20. Gu, L., Kanade, T.: 3d alignment of face in a single image. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. vol. 1, pp. 1305–1312. IEEE (2006)
21. Gu, X., Gortler, S.J., Hoppe, H.: Geometry images. *ACM Transactions on Graphics (TOG)* **21**(3), 355–361 (2002)
22. Güler, R.A., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., Kokkinos, I.: Densereg: Fully convolutional dense shape regression in-the-wild. In: Proc. CVPR. vol. 2 (2017)
23. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. *Kybernetes* **30**(9/10), 1865 – 1872 (2003)
24. Hassner, T.: Viewing real-world faces in 3d. In: IEEE International Conference on Computer Vision. pp. 3607–3614 (2013)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition. pp. 770–778 (2016)
26. Huber, P., Feng, Z.H., Christmas, W., Kittler, J., Ratsch, M.: Fitting 3d morphable face models using local features. In: IEEE International Conference on Image Processing. pp. 1195–1199 (2015)
27. Huber, P., Hu, G., Tena, R., Mortazavian, P., Koppen, W.P., Christmas, W.J., Ratsch, M., Kittler, J.: A multiresolution 3d morphable face model and fitting framework pp. 79–86 (2016)
28. Jackson, A.S., Bulat, A., Argyriou, V., Tzimiropoulos, G.: Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 1031–1039. IEEE (2017)
29. Jeni, L.A., Cohn, J.F., Kanade, T.: Dense 3d face alignment from 2d videos in real-time. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. vol. 1, pp. 1–8. IEEE (2015)
30. Jeni, L.A., Tulyakov, S., Yin, L., Sebe, N., Cohn, J.F.: The first 3d face alignment in the wild (3dfaw) challenge. In: European Conference on Computer Vision. pp. 511–520. Springer (2016)
31. Jourabloo, A., Liu, X.: Pose-invariant 3d face alignment. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3694–3702 (2015)
32. Jourabloo, A., Liu, X.: Large-pose face alignment via cnn-based dense 3d model fitting. In: Computer Vision and Pattern Recognition (2016)
33. 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 (2011)
34. Kim, J., Liu, C., Sha, F., Grauman, K.: Deformable spatial pyramid matching for fast dense correspondences. In: Computer Vision and Pattern Recognition. pp. 2307–2314 (2013)
35. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. pp. 2144–2151. IEEE (2011)
36. Laine, S., Karras, T., Aila, T., Herva, A., Lehtinen, J.: Facial performance capture with deep neural networks. *arXiv preprint arXiv:1609.06536* (2016)
37. Lee, Y.J., Lee, S.J., Kang, R.P., Jo, J., Kim, J.: Single view-based 3d face reconstruction robust to self-occlusion. *Eurasip Journal on Advances in Signal Processing* **2012**(1), 1–20 (2012)

38. Liang, Z., Ding, S., Lin, L.: Unconstrained facial landmark localization with backbone-branches fully-convolutional networks. arXiv preprint arXiv:1507.03409 (2015)
39. Liu, F., Zeng, D., Zhao, Q., Liu, X.: Joint face alignment and 3d face reconstruction. In: European Conference on Computer Vision. pp. 545–560. Springer (2016)
40. Liu, Y., Jourabloo, A., Ren, W., Liu, X.: Dense face alignment. arXiv preprint arXiv:1709.01442 (2017)
41. Maninchedda, F., Häne, C., Oswald, M.R., Pollefeys, M.: Face reconstruction on mobile devices using a height map shape model and fast regularization. In: 3D Vision (3DV), 2016 Fourth International Conference on. pp. 489–498. IEEE (2016)
42. Maninchedda, F., Oswald, M.R., Pollefeys, M.: Fast 3d reconstruction of faces with glasses. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4608–4617. IEEE (2017)
43. Matthews, I., Baker, S.: Active appearance models revisited. International journal of computer vision **60**(2), 135–164 (2004)
44. McDonagh, J., Tzimiropoulos, G.: Joint face detection and alignment with a deformable hough transform model. In: European Conference on Computer Vision. pp. 569–580. Springer (2016)
45. Moschoglou, S., Ververas, E., Panagakis, Y., Nicolaou, M., Zafeiriou, S.: Multi-attribute robust component analysis for facial uv maps. arXiv preprint arXiv:1712.05799 (2017)
46. Peng, X., Feris, R.S., Wang, X., Metaxas, D.N.: A recurrent encoder-decoder network for sequential face alignment. In: European Conference on Computer Vision. pp. 38–56. Springer (2016)
47. Richardson, E., Sela, M., Kimmel, R.: 3d face reconstruction by learning from synthetic data. In: Fourth International Conference on 3d Vision. pp. 460–469 (2016)
48. Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image (2016)
49. Romdhani, S., Vetter, T.: Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 986–993 (2005)
50. Saito, S., Li, T., Li, H.: Real-time facial segmentation and performance capture from rgb input. In: European Conference on Computer Vision. pp. 244–261. Springer (2016)
51. Santa, Z., Kato, Z.: 3d face alignment without correspondences. In: European Conference on Computer Vision. pp. 521–535. Springer (2016)
52. Saragih, J., Goecke, R.: A nonlinear discriminative approach to aam fitting. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. pp. 1–8. IEEE (2007)
53. Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation (2017)
54. Sinha, A., Unmesh, A., Huang, Q., Ramani, K.: Surfnet: Generating 3d shape surfaces using deep residual networks. In: IEEE CVPR. vol. 1 (2017)
55. Tewari, A., Zollhfer, M., Kim, H., Garrido, P., Bernard, F., Prez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction (2017)
56. Thies, J., Zollhfer, M., Stamminger, M., Theobalt, C., Niener, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Computer Vision and Pattern Recognition. p. 5 (2016)

57. Tran, A.T., Hassner, T., Masi, I., Medioni, G.: Regressing robust and discriminative 3d morphable models with a very deep neural network (2016)
58. Tzimiropoulos, G., Pantic, M.: Optimization problems for fast aam fitting in-the-wild. In: Computer Vision (ICCV), 2013 IEEE International Conference on. pp. 593–600. IEEE (2013)
59. Wagner, A., Wright, J., Ganesh, A., Zhou, Z., Mobahi, H., Ma, Y.: Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(2), 372–386 (2012)
60. Xiong, X., Torre, F.D.L.: Global supervised descent method. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2664–2673 (2015)
61. Xue, N., Deng, J., Cheng, S., Panagakis, Y., Zafeiriou, S.: Side information for face completion: a robust pca approach. *arXiv preprint arXiv:1801.07580* (2018)
62. 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)
63. Yu, R., Saito, S., Li, H., Ceylan, D., Li, H.: Learning dense facial correspondences in unconstrained images (2017)
64. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: European Conference on Computer Vision. pp. 94–108. Springer (2014)
65. Zhao, R., Wang, Y., Benitez-Quiroz, C.F., Liu, Y., Martinez, A.M.: Fast and precise face alignment and 3d shape reconstruction from a single 2d image. In: European Conference on Computer Vision. pp. 590–603. Springer (2016)
66. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In: Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on. pp. 386–391. IEEE (2013)
67. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. *computer vision and pattern recognition* pp. 146–155 (2016)
68. Zhu, X., Lei, Z., Yan, J., Yi, D., Li, S.Z.: High-fidelity pose and expression normalization for face recognition in the wild pp. 787–796 (2015)