Deep Evolutionary 3D Diffusion Heat Maps for Large-pose Face Alignment

Bin Sun[1]
sun.bi@husky.neu.edu
Ming Shao[2]
mshao@umassd.edu
Siyu Xia[1,3]
xia081@gmail.com
Yun Fu[1]
yunfu@ece.neu.edu

[1] ECE, College of Engineering
Northeastern University
Boston, MA, USA

[2] CIS, College of Engineering
University of Massachusetts Dartmouth
North Dartmouth, MA, USA

[3] School of Automation
Southeast University, China

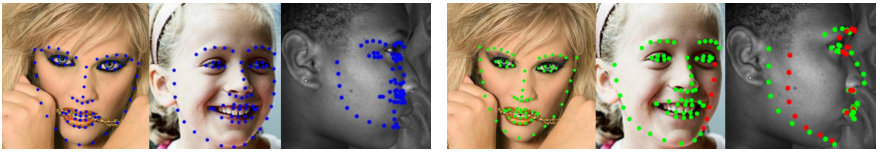
Abstract

As one of the fundamental vision tasks, face alignment has attracted a tremendous amount of efforts and achieved significant improvement over the decades. While the state-of-the-art works fairly well on the lab datasets and certain face images in the wild, it may easily fail in front of large pose variation, e.g., profile. In the worst case, the invisible landmarks may crash the initial models and thus limit many powerful models that only work well within a certain range using reliable features. To that end, we propose a new deep evolutionary model to integrate 3D Diffusion Heap Maps (DHM) to compensate for the invisible landmarks issue in large pose variation. Our contributions are summarized as: first, we introduce a sparse 3D DHM to assist the initial modeling under extreme pose conditions; second, a simple yet effective CNN feature is extracted and fed to recurrent neural networks for evolutionary learning. Additionally, we propose a Recurrent HourGlass (RHG) network that boost our evolutionary learning through HourGlass and LSTM module. Extensive experiments on three popular face alignment databases demonstrate the advantage of the proposed models over the state-of-the-art, especially under large-pose conditions. We also discuss and analyze the limitations of our models and future research work.

1 Introduction

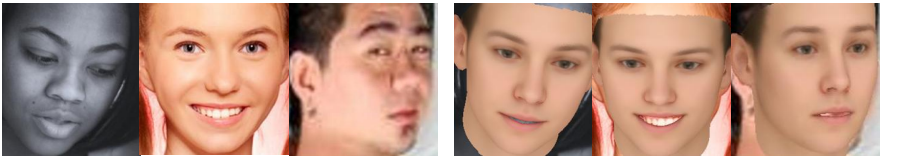
Face recognition and related application becomes increasingly popular, especially with the advances of deep learning. To name a few, face identification/verification [27], gaze detection [12], virtual face make-up [11], age synthesis [9], etc. Nonetheless, almost all of them heavily rely on face alignment that automatically locates predefined key points on a face. It has been treated as one of fundamental problems in real-world face recognition systems.

Recent research indicates that for moderate poses, illuminations, and expressions of face images in the wild, precisely detecting facial key points is feasible. Notably faces in photos are not always in medium poses where the yaw angle is less than 45° and all the landmarks



(a) (b)

Figure 1: Landmarks of (a) traditional method SDM and (b) our 3D DHM on different poses. The red dots indicate invisible landmarks.



(a) (b)

Figure 2: Comparison between (a) original faces and (b) misguidance faces generated by BFM.

are visible [29]. Faces in the wild rendering large poses will, however, fail even the most advanced face alignment algorithms. We analyze and detail the reasons as the followings:

Feature: Face alignment methods heavily rely on the features extracted from the image. 2D face images in large poses would hide some landmarks due to self-occlusion. When faces deviate from the frontal view, we can only trust those visible landmarks and use them to estimate the location of the invisible ones. Therefore, the alignment accuracy degrades significantly given more invisible landmarks.

Model: Face alignment can be treated as a non-linear optimization problem regarding to deformability of face. One solution is to map landmarks from 2D location space to feature space. This makes senses in medium poses but would fail in case of large poses where half of the key points are lost. Figure 1 shows the comparison between a representative traditional method [23] and ours in large-pose alignment tasks. To make up the information loss caused by self occlusion in large pose cases, 3D models are considered before regression. Considering the popular 3D approach Basel Face Model (BFM) [2] trained on only 200 people, the generated 3D dense model may misguide the regression process. From Figure 2, we can see the warped images by BFM fit the poses with minor difference over different races. Thus, we prefer a small and sparse 3D model for efficacy.

Data: While we can get access to many faces with landmarks from different face datasets nowadays, most of them are labeled by the human. Among them, most of the medium-pose data is labeled fairly well for training alignment models. Unfortunately, when the ground truth landmarks are self-occluded and become completely invisible, people have to guess the true location. As a result, those invisible manually labeled landmarks turn to be very unreliable, and confuse or even fail the model.

Contributions: We mainly focus on the first two challenges: *feature* and *model* in this work, thanks to the recently released large facial landmark datasets [4, 29]. To summarize, the contributions of this paper are:

- We propose a simple yet robust alignment feature learning paradigm using 3D Diffusion Heap Maps (DHM) and CNN to create high-level reliable features containing

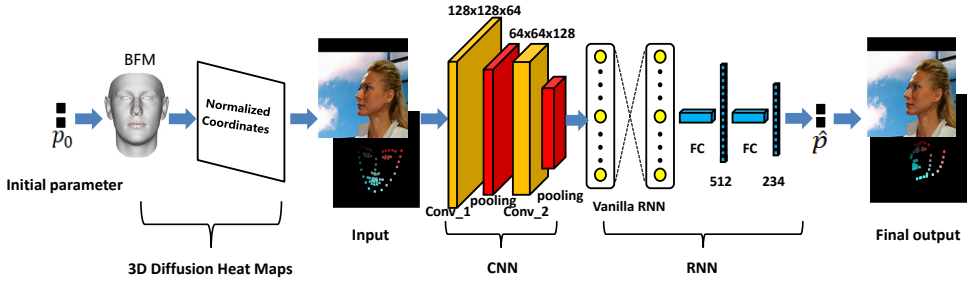


Figure 3: Framework overview. Note we also show the *3D diffusion heat maps* (under the input face image) calculated by Basel Face Model, and the final heat maps including the invisible landmarks which, however, are skipped by many 2D face alignment models.

both 2D and 3D information. We further investigate HourGlass [25] and LSTM [10] to upgrade our evolutionary learning paradigm, and achieve better performance. Note that our DHM is calculated from 3D model and only have 3 channels while 3DFAN [4] has 68 channels. This reduces the computation cost significantly.

- We cast face alignment to a deep evolutionary model with both 2D texture and 3D structure. Specifically, we use RNN to model the dynamics of the least square system. The system overview can be found in Figure 3.

- We conduct extensive experiments and improve the performance on a few benchmarks. We outperform the state-of-the-art by a large margin and show the robustness on both the original dataset and re-annotated AFLW2000-3D dataset.

2 Related work

2D Face Alignment: The first milestone work of 2D face alignment is ASM [6], followed by many successful non-deep algorithms including AAM [7] and Constrained Local Method (CLM) [1] that considered the local patches around the facial landmarks as the features and used constrained shape for initializing. Recently, critical works include tree-based models [15, 18] which improved the speed of face alignment to more than 1000 frames per second. Xiong et al. demonstrated the Supervised Descent Method [23] with the cascade of weak regressors for face alignment, and achieved the state-of-the-art performance [26]. Zhu et al. extended the work [26] and presented a new strategy [28] for large poses alignment by searching the best initial shape. Along with the spread of deep learning in AI is its successful applications on face alignment, specifically, Convolutional Neural Network (CNN). Sun et al. [20] firstly employed CNN model for face alignment tasks with a raw face as the input and conduct regression with high-level features. Differently, Trigeorgis et al. presented a RNN based approach with the philosophy of Xiong's work [23]. Another extension of SDM called Global Supervised Descent Methods (GSDM) [24] tried to solve the large poses problem by dividing the training space into different descent spaces. All these face alignment methods only use 2D information and most of them use cascade method [15, 23, 26] and local patch features [1, 7, 22, 23]. Differently, we suggest an integration of global 2D and 3D deep evolutionary network to overcome the information loss caused by 2D patch features.

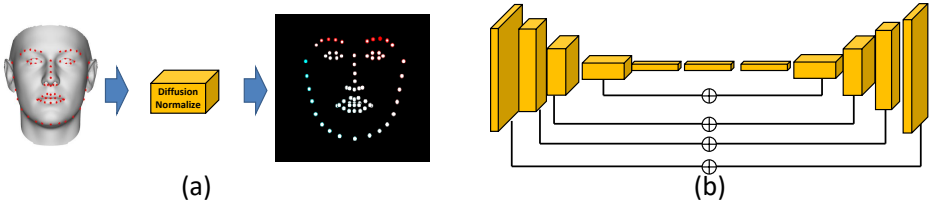


Figure 4: Illustration of (a) generating 3D diffusion heat maps; (b) the HourGlass module.

3D Face Alignment: As 3D face model can maintain the depth information well against pose issues, a bunch of 3D face alignment methods and 3D face datasets become increasingly popular. Dollár et al. [8] estimated the landmarks on large poses through a 3D Morphable model (3DMM) with cascade regression in 2010. Zhu and Jourabloo et al. [13, 29] presented CNN based model fitting a 3DMM to a 2D face through a cascade method, along with the facial key points. Besides, Zhu's work [29] also used three additional channels provided by the initial 3D face model in each iteration. A very dense 3D alignment model has been demonstrated by Liu et al. [14, 17] and achieved good performance. On the other hand, the release of a few benchmarks significantly promotes the research in this line. For example, 3D face training dataset called 300W-LP, and a testing dataset called AFLW-20003D were offered in [29] recently. Another 3D alignment dataset called LS3D-W was published by Bulat et al. [4] with about 230,000 images, and the deep learning models based on this dataset, e.g., HourGlass (HG) [3, 4, 25] have achieved impressive performance very recently. In this paper, we use sparse 3D heat maps together with the original image as input, whereas most of the previous works use dense 3D models. Jointly working with a plain RNN and a two-layer CNN, we achieve the state-of-the-art performance.

3 Algorithms

In this section we will detail our new framework (Figure 3) including two exclusive components: (1) 3D diffusion heat maps (DHM) generation; (2) deep evolutionary 3D heat maps.

3.1 3D Diffusion Heat Maps

To get feasible 3D landmarks in large poses, one of the reasonable ways is to build a 3D model of the face and simulate the details of a real face such as scale, expressions, and rotations, which can be formulated by the state-of-the-art 3DMM [2]. Typically, it represents factors of a 3D face by:

$$S = \overline{S} + E_{\text{id}} p_{\text{id}} + E_{\text{exp}} p_{\text{exp}} \tag{1}$$

where S is a predicted 3D face, $\overline{S}$ is the mean shape of the 3D face, E_{id} is principle axes based on neutral 3D face, p_{id} is shape parameters, E_{exp} is principle axes based on the increment between expressional 3D face and neutral 3D face, and p_{exp} is expression parameters. In our framework, the E_{id} and E_{exp} are calculated from a popular 3D face model named BFM [2]. Then we project the 3D face model by Weak Perspective Projection:

$$F(p) = f \times M \times R \times S + t_{2\text{d}} \tag{2}$$

where the $F(p)$ is the projected 3D face model, f is the scale, M is orthographic projected matrix $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, R is the rotation matrix written in $\begin{bmatrix} r_{\text{pitch}} & r_{\text{yaw}} & r_{\text{roll}} \end{bmatrix}$, S is the 3D shape model calculated from Eq. (1), and t_{2d} is the transition vector with the location coordinate x and y. Since F is the function of the parameter p, p can be written as $p = \begin{bmatrix} f & R & t_{2d} & p_{\text{id}} & p_{\text{exp}} \end{bmatrix}$. We can generate the aligned 3D shape through Eq. (1) and Eq. (2). Afterwards, with the key point index provided by BFM, we will have precise locations of key points on the 3D model. To generate sparse 3D features, we normalize the coordinates in the 3D model around the key points. For a specific color channel i, the process could be described as:

$$\text{map}_i(k) = \frac{S_j(k)}{\max(S_j) - \min(S_j)}, \quad j \in \{x, y, z\} \tag{3}$$

where "$\text{map}_i()$" is the 3D heat map with three channels R, G, and B, and the 3D coordinates triplet $\{x, y, z\}$ are mapped to the three channels. $S_j(k)$ means the value of the 3D shape at the location of the k_{th} landmarks. To incorporate the locality and increase the robustness of each landmark, we suggest to extend the 3D heat maps by a Gaussian diffusion map centered at each landmark's location, and thus, we obtain 3D DHM for robust representation. Specifically, we generate a set of heat maps centered at the landmarks and ranged by a 2D Gaussian with standard deviation equals to 1 pixel (See Figure 4(a)). This strategy assures our framework can extract the features of the whole image with different weights instead of discarding the features located at a non-landmarks position. Note we conduct the normalization and diffusion on each single channel/axis, independently.

Discussions: Recall the basic inputs of existing models usually include two separate parts: (1) an image; (2) initial mean landmarks. These methods usually depend on the initial landmarks for facial features for better performance, i.e., the features are usually extracted by cropping the image centered at initial landmarks. As discussed earlier, this strategy may degrade the performance in large pose situations, as the features centered at initial landmarks have significantly deviated from the ground truth. See the "Input" in Figure 3 (with face image and heap map). Thus, these features will misguide the learning model or regressor and probably converge to local trap. In contrast, we design a novel paradigm to address this issue. We keep the whole image as the input for robust facial feature learning and propose to employ sparse 3D shape information as the complement. This avoids the issues of misguidance by the incorrect initial landmarks that propagate to the local features.

3.2 Deep Evolutionary Diffusion Heat Maps

We offer an end-to-end trainable deep structure for face alignment in this section given the 3D heat maps and stacked image I. Since we have already generated DHM using 3D landmarks, we concentrate on: (1) discriminative and robust representation of image plus 3D DHM; (2) 3D DHM evolution; (3) recurrent HourGlass framework.

First, to formulate both discriminative and robust alignment features, we propose to use a plain CNN that absorbs both 2D and 3D information in a stacked structure. While handcraft features or a direct use of 3D information may work, our practice reflects that they are less competitive than the well developed CNN model. The CNN model here can extract high-level features critical to alignment, and we are especially interested in global CNN features. In our experiments, we also find that an off-the-shelf CNN model such as VGG-net [19] works fairly well in our case. Note we only use two convolutional layers to fuse the 3D information and RGB information which proves effective in our experiments.

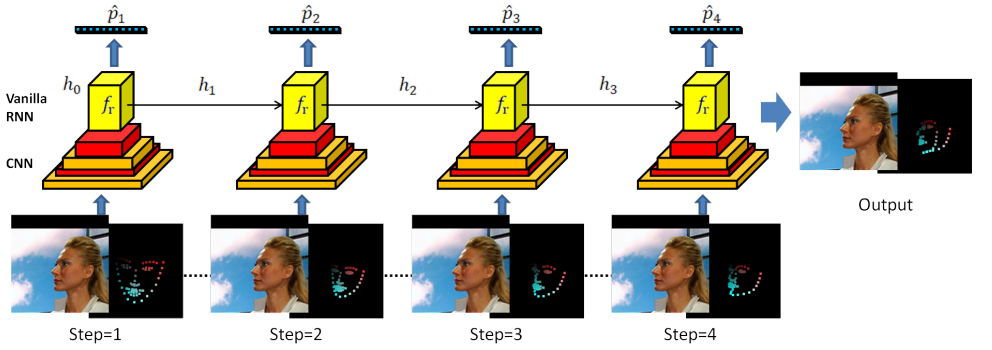


Figure 5: Intermediate results in RNN with state number $t = 4$ where changes of the heat maps can be seen. The heat map is scaled heavily in step 1 and deformed in step 2 to 4 with color being changed relatively. Finally, the heat map is well aligned on the face.

Algorithm 1 Deep 3D Evolutionary Diffusion Heat Maps

Inputs: Image I, initial values: $p_0 = [p_{\text{exp}}, p_{\text{id}}, f_0, R_0, t_{2d0}]$
for $i = 0$ to IterNum **do**
  generate S_i in 3D
  $map = \text{zeros}()$
  for $j = 0$ to 3 **do**
    $S_i[j] -= \min(S_i[j])$
    $S_i[j] /= (\max(S_i[j]) - \min(S_i[j]))$
    $map[S_i[0], S_i[1], j] = S_i[j]$
  end for
  extract features $\phi^k(cat(I, map), C_{\text{conv2}})$
  $\hat{p}_k = RNN(\phi^k(cat(I, map)))$
end for

Second, we resort to an evolutionary modeling for the alignment features. RNN has been widely applied to temporal data, as it is able to account for temporal dependencies. In training, RNN maintains the topology of feed forward networks while the feedback connections enable the representation of the current state of the system which encapsulates the information from the previous inputs, which can help update the parameter p in the loops within the networks. Mathematically, the update rules can be written as:

$$h_{t+1} = \tanh(W_{ih} C_{conv2}([I, \text{map}]) + W_{hh} h_t), \tag{4}$$

where h_t is the hidden state of the step t, $C_{conv2}([I, \text{map}])$ is the convolutional output features extracted from the input image I and the heat maps "map" generated by Eq. (2). W_{ih} is the weight from input to the hidden layer and the W_{hh} is the weight from hidden layer to hidden layer. With the hidden features h, we can model the update rule for $\hat{p}$ using $\hat{p}_{t+1} = \hat{p}_t + W_{ho} h_t$, where $\hat{p}_t$ is the parameters in the step t, W_{ho} is the weight from hidden layer to the output. Thus, from Eq. (4), we can prove that all the parameters in the step $t+1$ are based on the state of the step t. Since the whole image has been engaged as the input in each step, we may rectify the errors caused by the previous steps. Besides, we have the generated heat maps to emphasize the change in the previous step so that the whole network can converge. An

illustration of evolutions of 3D diffusion heat maps in four states can be found in Figure 5.

During the training, we define our loss function as

$$\min_{p} \|S_0 + \sum_{t=1}^{T} W_{ho} h_t(p) - S^*\|_F^2, \tag{5}$$

where S^* is the ground truth shape, F indicates the matrix Frobenius norm, T is the total number of the steps. The complete algorithm is shown in Algorithm 1.

The optimized p in the output layer will generate 3D landmarks for the test face which can be identified in the output of Figure 3. Here we suggest using Vanilla-RNN for simplicity. The evolutionary 3D DHM and intermediate results can be found in Figure 5. Note we update the parameters of the 3D model p instead of the landmarks themselves as we would encourage the model to restrict the spatial relation of each landmark. The input is a $224 \times 224 \times 3$ image stacked by heat maps. The output of RNN module is a 234-dimensional parameter, which will be casted to a 3D face model using Eq. (2). We can use specific vertices to find the landmark position.

Last, to explore the generality of our framework, we upgrade the CNN by a popular stacked HourGlass module (See Figure 4 (b)), and RNN by LSTM. In later experiments, we may compare with 3DFAN [4] that only uses HourGlass module resulting in inferior performance. This also demonstrates our evolutionary learning strategy is more general.

4 Experiments

In this section, we will first detail the evaluation datasets for large pose face alignment. Then, we conduct an analysis of the proposed method and evaluate different modules. Third, we will compare with the state-of-the-art face alignment methods. Finally, we will discuss the implementation details and some failure cases.

4.1 Dataset and Setting

We use 68-point landmarks to conduct fair comparisons with the state-of-the-art methods, though our method can adapt to any numbers of landmarks. Note in the training process, we may need 3D landmarks or parameters which are not always available. Thus, we estimate the 3D information through [4] in this situation. Evaluation datasets are detailed below:

- 300W-LP: The dataset has four parts, a total of 61,225 samples across large poses (1,786 from IBUG, 5,207 from AFW, 16,556 from LFPW and 37,676 from HELEN) [29]. Note we used 58,164 images for training and 3,061 as the validation.

- AFLW2000-3D: The dataset is essentially a reconstruction by Zhu et al. [29] given 2D landmarks. Note we use it for testing with 2000 images in total.

- Re-annotated AFLW2000-3D: The dataset is re-annotated by Bulat et al. [4] from AFLW2000-3D given 2D landmarks. We use it for testing with 2000 images in total.

- LS3D-W: The dataset is also a re-annotated by Bulat et al. [4]. We use it for training and testing to do a fair comparison. We use 218,595 images for training, and use its sub-dataset Menpo-3D (8,955 images) for testing. Note this dataset only has 2D landmarks projected from 3D space.

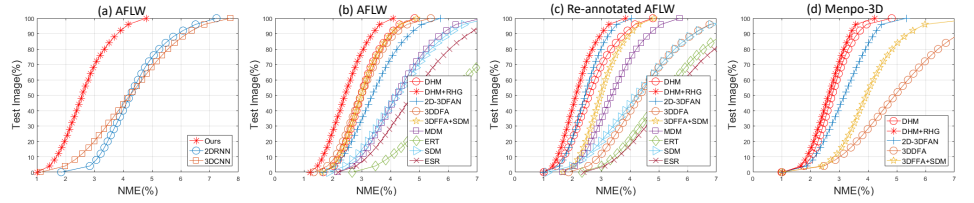


Figure 6: (a) shows analysis of replacing 3D diffusion heat map and RNN model. (b) shows the comparisons of AFLW3D dataset among ours and others. (c) shows the comparisons of Reannotated-AFLW3D dataset among ours and others. (d) is the comparison results among ours and some outstanding methods on Menpo3D dataset.

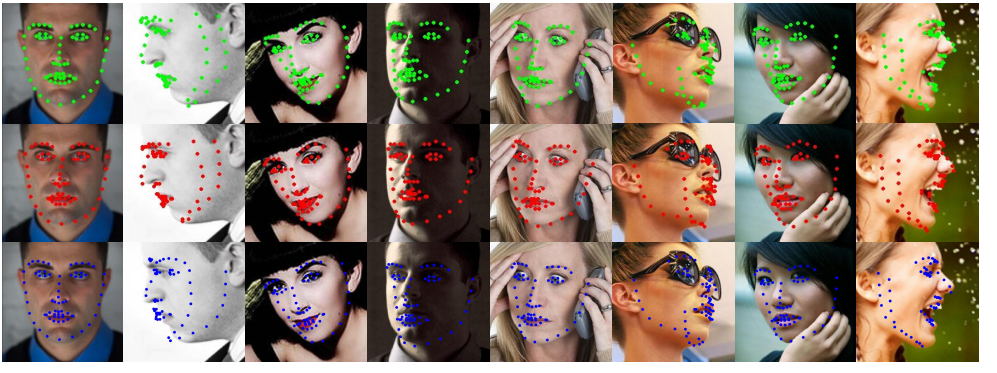


Figure 7: Results visualization of our method (top), 3DFAN (middle), and 3DDFA (bottom). Our method has more accurate results on eyes and contours than other two methods.

As our focus is face alignment, we should reduce the negative effects of face detection. Thus, the detected bounding box of each face is computed by ground-truth landmarks. To compare with other methods, we use the same metric "Normalized Mean Error (NME)" defined as $\text{NME} = \frac{1}{N}\sum_{i=1}^{N}\frac{\|\hat{X}_i - X_i^*\|_2}{d}$ where the $\hat{X}$ and X^* is predicted and ground truth landmarks, respectively, N is the number of the landmarks, d is normalized distance computed by the width and height of the bounding box using $d = \sqrt{w_{\text{bbox}} \times h_{\text{bbox}}}$.

To learn the weights of the network, we use Adam stochastic optimization [16] with default hyperparameters. The initial learning rate is 0.0001 for 300W-LP with exponential decay of 0.95 every 2000 iterations, an initial learning rate of 0.0001 is employed in our training process. The batch size is set to 50. The Recurrent HourGlass (HG) network starts with a 7×7 convolutional layer with stride 2. A residual module and a round of max pooling are added after it to bring the resolution down from 256 to 64. We use 3 stacked HG modules to extract features and a LSTM [10] as our recurrent module. The initial learning rate is 0.001 and we set weight decay at epoch 5, 15, 30. The total number of epochs is 40. We use RMSprop [21] as our optimizer. The training batch is 32 and validation batch is 16.

4.2 Performance Analysis of Our Model

In this section, we will demonstrate the advantage and necessity of two modules: (1) 3D diffusion heat maps; (2) RNN for deep evolution.

Table 1: Comparisons with state-of-the-art methods on ALFW2000-3D dataset and re-annotated AFLW2000-3D. We highlight the performance of our model in each setting.

Normalized Mean Error on ALFW2000-3D					NME on re-annotated ALFW2000-3D			
Method Name	[0°30°]	[30°60°]	[60°90°]	Mean	[0°30°]	[30°60°]	[60°90°]	Mean
DHM	**2.75**	**4.21**	**6.91**	**4.62**	**2.28**	**3.10**	**6.95**	**4.11**
DHM+RHG	**2.52**	**3.21**	**5.76**	**3.85**	**2.25**	**3.05**	**4.21**	**3.17**
3DFAN [4]	2.58	3.76	11.72	6.02	2.75	3.76	5.72	4.07
3DDFA [29]	3.78	4.54	7.93	5.42	4.82	5.71	10.93	7.15
3DDFA+SDM [29]	3.43	4.24	7.17	4.94	3.23	4.04	8.17	5.15
MDM [22]	3.67	5.94	10.76	6.45	3.27	5.91	9.77	6.31
ERT [15]	5.40	7.12	16.01	10.55	5.33	7.42	16.46	9.73
SDM [23]	3.67	4.94	9.76	6.12	3.47	4.91	9.81	6.06
ESR [6]	4.60	6.70	12.67	7.99	4.75	7.10	14.10	8.65

Figure 8: Failure cases of our model.

First, we evaluate the importance of evolutionary 3D maps. Instead of using RNN, we use a plain 3D-CNN structure. That is being said, we use the same 3D heat maps but replace the RNN by a conventional CNN structure. The rest parts remain the same. Note we use 300W-LP dataset for training and AFLW2000-3D dataset for evaluation. Results of this setting (3D-CNN) can be found in Figure 6(a).

Second, we keep the RNN structure and test the importance of 3D heat maps in our framework. We replace the 3D module by initial 2D landmarks. Accordingly, we change the output of the framework from a 234×1 vector to a 204×1 vector, which is the increment of locations of predicted landmarks. The rest of framework remains the same. Note we also use the same training and testing datasets. Result of this setting (2D-RNN) can be found in Figure 6(a). It is easy to find out that the combination of both module performs best.

4.3 Comparisons with Existing Methods

In this section, we conduct comprehensive evaluations with the state-of-the-art methods. Specifically, all methods are trained on the 300W-LP dataset including both ours and others. All of the input faces are cropped by the bounding box calculated from landmarks. All competitive methods have released their codes and thus, we optimize their models on 300W-LP for a fair comparison. The competitive methods include: (1) ERT [15], SDM [23], 3DDFA [29], MDM [22], 3DFAN [4]. After training with the same 300W-LP dataset, we evaluate all methods on the ALFW2000-3D dataset. The comparison results can be found in Figure 6 (b) and the quantitative results can be found in Table 1.

From the Figure 6(b) we can see that our method is the best on the AFLW2000-3D dataset. However, we found that the performance of 3DFAN could be better according to the work [4]. Bulat et al. claimed that the ground truth of the 300W-LP dataset and AFLW2000-3D dataset were not so accurate. After their re-annotation, the model achieved very high accuracy. For fair comparisons, we use their re-annotated data and results are shown in Fig-

ure 6(c). From the figure, we can find that the result of our model is competitive with the performance of 3DFAN. From the two experiments, we can prove that our method is very robust and reliable regardless of the minor errors of labels. The visualization of three methods are shown in Figure 7. We also use the large dataset LS3D-W to train our framework. Since there is no depth information, we use 2D-3D FAN model [4] to generate the depth coordinates. Then we change the output of the final fully connected layer from a 234×1 vector to a 204×1 vector to represent 3D landmarks. 3DFAN is trained on the same dataset for comparisons. The result is shown in Figure 6(d) where ours exhibits better performance.

We also illustrate some failure cases of our model in Figure 8 which indicate that our model may be fragile given large-angle rolled faces or heavily occlusions. It will also fail if the illumination is bad. The primary reason is the lack of faces of similar cases in the training dataset, which is a common issue for all other methods. Possible solutions include adding corresponding training data to approach the special case, or employ multiple initializations to avoid local minimas. We can also use data augmentation to make the model robust.

5 Conclusions

In this paper, we focused on improving the face alignment algorithms with the sparse 3D landmarks to approach the challenge of large poses. We presented a deep evolutionary framework to progressively update the 3D heat maps to generated target face landmarks. First, we proposed to use 3D diffusion heat maps as well as global 2D information as the robust representation. Second, we demonstrated that an RNN based evolutionary learning paradigm was able to model the dynamics of least square problems and optimize the facial landmarks. In addition to prove the universality of our evolutionary learning strategy, we also proposed a Recurrent HourGlass framework which achieved the state-of-the-art performance on popular face alignment benchmarks. The results show the possibility of further improvement with complicated structures in our evolutionary DHM strategy.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant 61671151 and 61728103.

References

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *CVPR*, pages 2610–2617. IEEE, 2012.

[2] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *TPAMI*, 25(9):1063–1074, 2003.

[3] Adrian Bulat and Georgios Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *International Conference on Computer Vision*, 2017.

[4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). *arXiv preprint arXiv:1703.07332*, 2017.

[5] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

[6] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *CVIU*, 61(1):38–59, 1995.

[7] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *TPAMI*, 23(6):681–685, 2001.

[8] Piotr Dollár, Peter Welinder, and Pietro Perona. Cascaded pose regression. In *CVPR*, pages 1078–1085. IEEE, 2010.

[9] Yun Fu, Guodong Guo, and Thomas S Huang. Age synthesis and estimation via faces: A survey. *TPAMI*, 32(11):1955–1976, 2010.

[10] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[11] Dong Guo and Terence Sim. Digital face makeup by example. In *CVPR*, pages 73–79. IEEE, 2009.

[12] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *TPAMI*, 32(3):478–500, 2010.

[13] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proc CVPR*, pages 4188–4196, 2016.

[14] Amin Jourabloo and Xiaoming Liu. Pose-invariant face alignment via cnn-based dense 3d model fitting. *International Journal of Computer Vision*, 124(2):187–203, 2017.

[15] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proc CVPR*, pages 1867–1874, 2014.

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

[17] Yaojie Liu, Amin Jourabloo, William Ren, and Xiaoming Liu. Dense face alignment. *arXiv preprint arXiv:1709.01442*, 2017.

[18] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *Proc CVPR*, pages 1685–1692, 2014.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proc CVPR*, pages 3476–3483, 2013.

[21] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[22] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proc CVPR*, pages 4177–4187, 2016.

[23] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proc CVPR*, pages 532–539, 2013.

[24] Xuehan Xiong and Fernando De la Torre. Global supervised descent method. In *Proc CVPR*, pages 2664–2673, 2015.

[25] Jing Yang, Qingshan Liu, and Kaihua Zhang. Stacked hourglass network for robust facial landmark localisation. In *CVPRW*, pages 2025–2033. IEEE, 2017.

[26] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *ECCV*, pages 1–16. Springer, 2014.

[27] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003.

[28] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *Proc CVPR*, pages 4998–5006, 2015.

[29] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proc CVPR*, pages 146–155, 2016.