Recurrent 3D-2D Dual Learning for Large-pose Facial Landmark Detection

Shengtao Xiao[1], Jiashi Feng[1], Luoqi Liu[2], Xuecheng Nie[1], Wei Wang[3], Shuicheng Yan[2,1], Ashraf Kassim[1]
[1]National University of Singapore, [2]Qihoo-360, [3]University of Trento
xiao_shengtao@u.nus.edu, elefjia@nus.edu.sg, liuluoqi@360.cn, niexuecheng@u.nus.edu
wei.wang@unitn.it {eleyans,ashraf}@nus.edu.sg

Abstract

Despite remarkable progress of face analysis techniques, detecting landmarks on large-pose faces is still difficult due to self-occlusion, subtle landmark difference and incomplete information. To address these challenging issues, we introduce a novel recurrent 3D-2D dual learning model that alternatively performs 2D-based 3D face model refinement and 3D-to-2D projection based 2D landmark refinement to reliably reason about self-occluded landmarks, precisely capture the subtle landmark displacement and accurately detect landmarks even in presence of extremely large poses. The proposed model presents the first loop-closed learning framework that effectively exploits the informative feedback from the 3D-2D learning and its dual 2D-3D refinement tasks in a recurrent manner. Benefiting from these two mutual-boosting steps, our proposed model demonstrates appealing robustness to large poses (up to profile pose) and outstanding ability to capture fine-scale landmark displacement compared with existing 3D models. It achieves new state-of-the-art on the challenging AFLW benchmark. Moreover, our proposed model introduces a new architectural design that economically utilizes intermediate features and achieves $4\times$ faster speed than its deep learning based counterparts.

1. Introduction

Facial landmark detection aims to locate key fiducial points such as eye corners, mouth, nose tips and face contour points for a given 2D face image. It is fundamental in many face-related applications, *e.g.*, face recognition [21], 3D face reconstruction [30] and face synthesis [22].

Despite remarkable progress, detecting landmarks on large-pose faces is still challenging for existing approaches. For instance, cascaded regression based approaches [25, 18, 27, 24] offer top performance among modern 2D facial landmark detection approaches [31, 25, 6, 18, 27, 24].

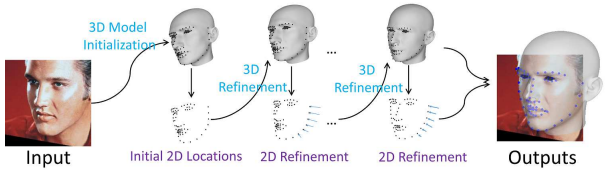

Figure 1. Flowchart of the proposed model. It first predicts an initial model and obtain initial 2D landmark location via direct 3D-to-2D projection for an input 2D face image, and then alternatively refines the 3D face model and 2D landmark locations in a mutual-boosting manner. The dual-refinement architecture ensures our model's robustness and accuracy for 2D landmark detection under challenging conditions.

However they still fail at detecting landmarks of challenging large poses, due to self-occlusion and unreliable features around invisible landmarks.

To tackle such limitations of 2D-based methods, several recent works [29, 11, 16] resort to 3D face models to improve detection robustness to large-pose facial landmarks. The 3D-based models generally align a 3D morphable face model [1] to the test 2D face image and infer landmark locations from the 3D face via single direction 3D-to-2D projection. Although the 3D model can effectively observe the entire face, it may fail to capture small variations over appearance and shape of facial components, *e.g.* mouth and eyes. Moreover, 3D faces fitted via morphable face model may have over-smoothed shapes and limited expressions. As a result, those models usually fail to precisely detect landmarks at fine scales, though being able to localize invisible landmarks.

In this work, to address the challenging large-pose facial landmark detection problem, we propose a novel Recurrent Dual Refinement (**RDR**) model that provides a closed-loop learning process for 2D landmark detection and its dual task of 3D face model refinement. Benefiting from informative feedback and mutual-boosting between these two learning steps, our model presents outstanding robustness to large face poses as well as strong ability at detecting landmarks at fine scale.

The proposed RDR model first introduces an effective

direct 3D parameter prediction module that reliably provides good initial 3D face models and facilitates the following refinement of both 3D face model and 2D landmark detection. A novel dual refinement module, consisting of 3D face model refinement and 2D landmark location refinement components, is then developed to alternatively update the 3D face fitting and correspondingly 2D landmark detection in a mutual-boosting manner. As illustrated in Fig. 1, at each refinement iteration, the 3D face component updates the 3D face model to strengthen RDR's robustness to large face poses and lift the 2D landmark refinement to improve the overall landmark detection performance. Then, the 2D refinement component improves the landmark locations inferred from 3D-to-2D projection by capturing the fine scale landmark displacement. The refined 2D landmark locations are then used for extracting deep shape-indexed features for dual refinement at next iteration. This 3D-2D dual refinement process is repeated recurrently until convergence, providing new-state-of-art landmark detection performance.

Besides superior performance, our RDR model also provides appealing efficiency — $4\times$ faster than the most recent deep neural network based 3D face models [29, 11] on single GPU. The efficiency comes from economically utilizing the learned features in the intermediate layers without passing the images through the entire architecture as before [29, 11].

RDR is designed to perform 3D faces fitting based on the dynamic expression model [23, 15, 4]. Thus it can also be used directly for virtual avatar manipulation [15, 4] by simply being fed the related 3D parameters, *i.e.*, coefficients of expression blendshapes, head poses and camera projection, which we will further explore in future.

The main contributions of this paper can be summarized as follows:

- We develop the first 3D-2D dual learning model, i.e., RDR, for addressing large-pose facial landmark detection within an end-to-end trainable framework.
- The proposed RDR jointly refines 3D face model and 2D landmark locations. It effectively utilizes the 3D guidance information and localizes facial landmarks accurately and robustly, even under challenging conditions.
- RDR offers an efficient solution to large-pose facial landmark detection, which shows great potential for practical usage. It runs at least $4\times$ more efficient than existing deep neural network models [29, 11] on a GPU in addition to its superior accuracy.

2. Related Work

2.1. Cascaded Regression for Landmark Detection

Methods of facial landmark detection via cascaded regression [25, 18, 13] refine landmark locations with features extracted around previously detected landmarks at each refinement stage. Various features can be used for regressing the landmark updates, *e.g.* SIFT, HOG [25] and binary features [18, 2, 28]. Deep features [20, 13] are also used to perform facial landmark detection but usually bring higher computational cost. Those 2D-based regression approaches have demonstrated appealing performance under moderate face conditions and a strong ability to capture fine-scale landmark displacement. However, their robustness deteriorates significantly when handling with large-pose face images that have unreliable local features due to self-occlusion.

2.2. 3D Face Alignment

Many existing 3D face fitting and alignment methods [4, 3, 30] require accurate locations of landmarks, either from manual annotation or off-the-shelf facial landmark detectors. For example, the algorithm in [4] needs to know accurate landmark locations of multiple facial images to off-line prepare expression blendshapes for 3D face reconstruction. Similarly, the algorithm proposed in [3] requires landmark locations for initialization, although off-line generation of blendshapes is not necessary.

Only very recently, different frameworks [29, 11] are proposed which directly predict the 3D face based on the 3D morphable model [1]. Both works employ a cascade of convolutional neural networks (CNNs) to extract features to predict 3D face parameters iteratively. Specifically, each regression step can be formulated as

$$Q_t := Q_{t-1} + \Delta Q_t = Q_{t-1} + \mathrm{CNN}_t(h(I, Q_{t-1})).$$

Here, Q_t denotes the related 3D fitting parameters. The CNN model CNN_t in the cascade takes the $h(I, Q_{t-1})$ as input for predicting the updates on Q_{t-1}. $h(I, Q_{t-1})$ is the process of generating input for the current iteration with 3D face information based on previous prediction Q_{t-1}.

The 3D Morphable Model (3DMM) performs PCA reconstruction to generate a 3D face. It is efficient but may limit the shape diversity and over-smooth the 3D face model if only a few principal components are used. Small appearance and expression variations of facial components are also challenging to be captured with 3DMM. Consequently, 2D landmark locations directly inferred via 3D-to-2D projection would be inaccurate.

3. 3D Face Fitting with Dynamic Expression Model

Before introducing our proposed model, we first explain the 3D face fitting approach used in the 3D-2D dual

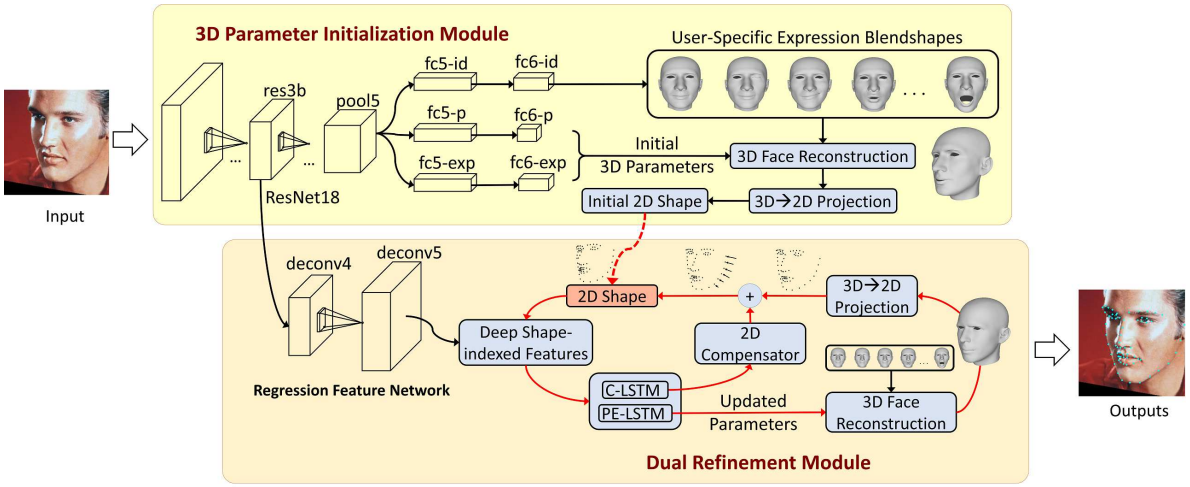

Figure 2. Overview of the proposed RDR model for large-pose facial landmark detection. Given an input 2D face image, RDR first directly predicts initial 3D face fitting parameters with the 3D parameter initialization module, generates an initial 3D face mesh and infers initial 2D landmark locations via 3D-to-2D projection. It then recurrently refines both the 3D face and 2D landmark locations with a dual refinement module consisting of a 3D face refinement component (PE-LSTM) and a 2D landmark refinement component (C-LSTM) with deep features directly extracted from the regression feature network.

learning. In the 3D face fitting process within our model, the Dynamic Expression Model (DEM) [23, 4] is employed due to its capability of virtual avatar manipulation. Details of DEM are introduced below.

In 3D face modeling, each 3D face is described by a set of N vertices $\mathbf{F} = [v_1, \ldots, v_N] \in \mathbb{R}^{3\times N}$. Each vertex v has a 3D-coordinate $[x, y, z]^\top \in \mathbb{R}^3$. The dynamic expression model [23, 4] represents each 3D face as a linear combination of d expression blendshapes $\mathbf{B} = [B_0, B_1, \ldots, B_d] \in \mathbb{R}^{3\times N\times d}$ with coefficients $\boldsymbol{\alpha}_{\text{exp}} \in \mathbb{R}^{1\times d}$, after proper affine transformation. An expression blendshape, $B_i \in \mathbb{R}^{3\times N}$, is formed by transforming a neutral 3D face to a 3D face that has a typical expression with the preserved identity. In this work, we use a neutral shape B_0 and another 46 expression blendshapes with different expressions to represent a 3D face, similar to [4]. Given a rotation matrix, $\mathbf{R}_{\phi,\gamma,\theta} \in \mathbb{R}^{3\times 3}$ and a translational vector $\mathbf{t}_{3\text{d}} \in \mathbb{R}^3$, the *3D face reconstruction* process of $\mathbf{F}$ is formally described as

$$\mathbf{F} = \mathbf{R}_{\phi,\gamma,\theta}(\mathbf{B} \times_3 \boldsymbol{\alpha}_{\text{exp}}) + \mathbf{t}_{3\text{d}}, \tag{1}$$

where the rotation parameters ϕ, γ, θ correspond to pitch, yaw and roll rotation angles respectively. $\times_i$ denotes the multiplication over the i-th mode of the tensor. Recall $\boldsymbol{\alpha}_{\text{exp}}$ is the vector of expression manipulation coefficients. As described in [5, 4], the expression blendshapes $\mathbf{B}$ of a specific person can be described as

$$\mathbf{B} = \mathcal{C}_B \times_2 \boldsymbol{\alpha}_{\text{id}}, \tag{2}$$

where $\mathcal{C}_B$ represents the *core tensor* obtained via decomposing a large collection of 3D faces along the identity mode (*i.e.*, mode 2). In this work, we use the 3D faces from the Face Warehouse [5] to obtain the core tensor. In addition, $\boldsymbol{\alpha}_{\text{id}}$ is the user-specific identity vector.

Using the 3D face model to detect 2D facial landmarks needs to project 3D faces onto the 2D image plane. Under weak perspective projection [3, 29, 11], the 2D locations of all vertices can be formulated as

$$\mathbf{M}(\mathbf{Q}) = \boldsymbol{\Pi}_f \mathbf{F} = \boldsymbol{\Pi}_f \left(\mathbf{R}_{\phi,\gamma,\theta}(\mathbf{B} \times_3 \boldsymbol{\alpha}_{\text{exp}}) + \mathbf{t}_{3\text{d}}\right), \tag{3}$$

where $\mathbf{Q} = \{\mathbf{P}, \boldsymbol{\alpha}_{\text{id}}, \boldsymbol{\alpha}_{\text{exp}}\}$ represents the parameters for 3D face fitting and 3D-to-2D projection. Within $\mathbf{Q}$, we use $\mathbf{P} = \{\phi, \gamma, \theta, \mathbf{t}_{3d}, f\}$ to collectively denote the rotations, translation and 3D-to-2D projection parameters. Assuming perfect pinhole model, the orthographic projection matrix $\boldsymbol{\Pi}_f$ is defined as

$$\boldsymbol{\Pi}_f = \begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \end{pmatrix},$$

where f is the scale factor. The 2D locations of landmarks $S_p = [s_1; \ldots; s_L]$ can be directly obtained from $\mathbf{M}(\mathbf{Q})$ as

$$S_p = \mathbf{M}(\mathbf{Q})^{\{\mathbf{v}\}} \in \mathbb{R}^{2\times L}, \tag{4}$$

where $\mathbf{v} = [v_{s_1}, \ldots, v_{s_L}]^\top$ denotes landmark indices.

4. Recurrent Dual Refinement Networks

With the above dynamic expression model and 2D locations of landmark vertices S_p, we now proceed to detail our RDR model that alternatively refines 3D face and 2D landmark locations in this section.

4.1. Overview

Fig. 2 shows the entire framework of the RDR model for large-pose 2D landmark detection. It first receives a 2D face image and passes it through a 3D parameter initialization module to directly predict parameters for accurate 3D face initialization and prepares deep facial

representation for 3D-2D dual refinement through the regression feature network. A dual refinement module has been developed to take deep features extracted around the predicted 2D facial landmarks from the regression feature network, and refine both 3D face and 2D landmark locations in a mutual-boosting manner, with mutual-boosting performance achieved. This dual refinement process is repeated recurrently until convergence.

4.2. Initial 3D Face Prediction

In our proposed RDR model and also in other cascaded regression ones (see Sec. 2), the quality of initial parameters is critical for the final landmark detection performance. Using good initial parameters can improve the overall performance and accelerate the refinement process. Our model initializes the 3D fitting parameters, *i.e.*, $\mathbf{Q} = \{\mathbf{P}, \boldsymbol{\alpha}_{\text{id}}, \boldsymbol{\alpha}_{\text{exp}}\}$, by direct regression via the 3D parameter initialization module, Fig. 2. We find this strategy indeed provides higher-quality parameters to start with and leads to better landmark detection performance, compared to other common initializations, *e.g.*, mean value initialization [18, 29, 11]. Moreover, such initialization does not bring significant computational overhead and can be seamlessly integrated with the following recurrent refinement process.

As shown in the upper panel of Fig. 2, the 3D parameter initialization module predicts the important initial parameters $\mathbf{Q}$ directly for a given face image. This network predicts the parameters upon the features from pool5 by using three independent fc5-fc6's, each of which predicts a specific 3D parameter of $\boldsymbol{\alpha}_{\text{id}}$, $\mathbf{P}$ and $\boldsymbol{\alpha}_{\text{exp}}$ respectively. Given the predicted identity parameter $\boldsymbol{\alpha}_{\text{id}}$, the user-specific expression blendshapes $\mathbf{B}$ are obtained via Eqn. (2) and used in the following 3D face refinement process. The initial 3D face model is generated with Eqn. (1) and the initial 2D landmark locations are inferred directly via 3D-to-2D landmark projection with Eqn. (4).

This direct 3D parameter initialization module is trained by minimizing the discrepancy between the predicted parameters $\widehat{\mathbf{Q}}^0$ and ground truth $\mathbf{Q}^*$:

$$\mathcal{L}_{\text{d}} = \|\mathbf{Q}^* - \widehat{\mathbf{Q}}^0\|_2^2. \tag{5}$$

We defer the details on preparing the ground truth to Section 5.1 in experiments. Although the direct prediction network provides good initial parameters, the output initial 3D face is still inaccurate in rotation and expression. To ensure the 3D face to provide a solid and accurate prior for 2D landmark detection, further refinement on the 3D face model is necessary.

4.3. Recurrent Dual 3D-2D Refinement

Before introducing details of recurrent dual 3D-2D refinement, we explain the inherent error in 2D facial landmark locations directly inferred from 3D faces via 3D-to-2D projection. Even using the ground truth 3D parameter $\mathbf{Q}^*$ with perfect projection, the 3D fitting error[1] is still at a significant level of around 5%. It comes from the inherent difficulty of fitting 3D face model to scattered 2D landmark in a single projection step. This also explains the limitations of recent 3D alignment methods [29, 11].

To resolve this critical issue and improve the accuracy of 2D landmark locations (which is also critical for further enhancing the 3D face model), we propose the dual refinement module. Concretely, we introduce a compensation term ΔS that accommodates future "fine-tuning" on the 2D landmarks locations obtained via 3D-to-2D projection:

$$S = S_p + \Delta S. \tag{6}$$

We propose to detect 2D landmarks through two steps: first the projection from the 3D key vertices gives S_p and then "fine-tuning" on S_p gives final 2D locations S. Eqn. (6) also indicates that the accuracy of S is jointly determined by S_p and ΔS. Dual refinement of the 3D face and the projected 2D landmark locations is therefore essential.

The dual refinement module, illustrated in Fig. 2, takes deep shape-indexed features [13] extracted from deconv5 around the predicted 2D landmark locations as input. Two key components within the dual refinement module, the 3D face refinement component and the 2D landmark refinement component, utilize the extracted deep features and perform refinement on 3D face model and the projected 2D landmark locations accordingly. Both refinement components are modelled by Long Short-Term Memory [8] units (LSTM), *i.e.* PoseExp-LSTM (PE-LSTM) and Compensator-LSTM (C-LSTM). The PE-LSTM refines the 3D face model by predicting the necessary updates $\Delta\mathbf{P}^k$ and $\Delta{\boldsymbol{\alpha}_{\text{exp}}}^k$. C-LSTM forecasts the suitable compensation ΔS^k to the projected 2D landmark locations S_p^k. The final facial landmark location, $S^T = S_p^T + \Delta S^T$, is obtained after T iterations of recurrent dual refinement.

Within the dual refinement module, we choose the Long-Short Term Memory as the recurrent unit based on following considerations. First, LSTM has an excellent capability of modeling historical information through adaptively memorizing and forgetting. Secondly, LSTM is easier for gradient descent optimization and allows long-term recurrent modeling. It is verified that utilizing historical information with LSTMs can effectively enhance landmark detection accuracy in [17, 13, 24].

Fast Recurrent Refinement In existing 3D face fitting methods [29, 11], updated input images have to go through the *entire* network to provide updates on the 3D fitting parameters at *every iteration*. This brings huge computational overhead and is the bottleneck for

[1]Measured by normalized distance between locations of landmark vertices on the 2D plane and the ground-truth 2D locations

Algorithm 1: Recurrent Dual Refinement.

Inputs: User-specific expression blendshapes $\mathbf{B}$; outputs of deconv5: D_{deconv5}; maximum refinement step K;
Initialization: $\boldsymbol{\alpha}_{\text{exp}}^0,\ \widehat{\mathbf{P}}^0,\ S^0 = S_p^0,\ k = 1$.
while $k \leq K$ **do**
  Extract deep shape-indexed feature: $\Phi(D_{\text{deconv5}}, S^{k-1})$;
  /* 3D Face Refinement */
  $\Phi_{PE}^k = \Phi(D_{\text{deconv5}}, S^{k-1}) * w_{PE}^k + b_{PE}^k$
  $\{\Delta\boldsymbol{\alpha}_{\text{exp}}^k, \Delta\mathbf{P}^k\} = \text{PE-LSTM}(\Phi_{PE}^k)$
  /* update $\widehat{\boldsymbol{\alpha}}_{\text{exp}}, \widehat{\mathbf{P}}$ */
  $\{\widehat{\boldsymbol{\alpha}}_{\text{exp}}^k, \widehat{\mathbf{P}}^k\} = \{\widehat{\boldsymbol{\alpha}}_{\text{exp}}^{k-1}, \widehat{\mathbf{P}}^{k-1}\} + \{\Delta\boldsymbol{\alpha}_{\text{exp}}^k, \Delta\mathbf{P}^k\}$
  Obtain 3D face based on parameters $\sigma(\widehat{\boldsymbol{\alpha}}_{\text{exp}}^k)$, $\mathbf{B}$ and $\widehat{\mathbf{P}}^k$ via Eqn. (1)
  Obtain projected 2D shape S_p^k via Eqn. (6)
  /* 2D Location Refinement */
  $\Phi_R^k = \Phi(D_{\text{deconv5}}, S^{k-1}) * w_R^k + b_R^k$
  Updated 2D shape: $S^k = S_p^k + \text{C-LSTM}(\Phi_R^k)$
  $k = k + 1$
end
Outputs: $S^K, \widehat{\mathbf{P}}^K, \widehat{\boldsymbol{\alpha}}_{\text{exp}}^K$

enhancing the efficiency. Inspired by [13], deep features extracted from hidden layers of the network are sufficiently informative for estimating head pose, focal length and expression related parameters (see Sec. 4.2).

By employing the deep shape-indexed features, our model does not need to feed-forward the image throughout the entire refinement process. At each iteration, it simply extracts regression features from the outputs of the deconv5 layer and passes them to the dual refinement module. This saves much computational cost and makes our model much faster than other deep neural network architectures.

Expression Control To ensure an artifact-free 3D face, the expression coefficients $\boldsymbol{\alpha}_{\text{exp}} = [\alpha_0, \ldots, \alpha_{46}]$ needs to satisfy the following natural requirements [14, 4]:

$$\alpha_0 = 1 - \sum_{i=1}^{46} \alpha_i,\ \text{s.t.}\ |\alpha_i| \leq 1, \forall i = 1, \ldots, 46. \tag{7}$$

To mitigate the difficulty caused by the hard constraint on estimation of $\boldsymbol{\alpha}_{\text{exp}}$, we introduce an extra sigmoid function $\sigma(x) = 1/(1+e^{-x})$ to normalize α_i such that they satisfy the above constraints. Instead of directly predicting the expression coefficients, we minimize $\sum_{i=1}^{46} \|\sigma(\widehat{\alpha}_i) - \alpha_i^*\|_2^2$ for all losses related to the expression parameters, where $\widehat{\alpha}_i$ is the estimated coefficient for the expression blendshape B_i. In the 3D face model refinement step, the updated expression coefficients at the k-th stage $\boldsymbol{\alpha}_{\text{exp}}^k$ are obtained via $\alpha_i^k := \sigma(\widehat{\alpha}_i^{k-1} + \Delta\alpha_i^k), \forall i = 1, \ldots, 46$. Here $\Delta\boldsymbol{\alpha}_{\text{exp}}^k$ is the update provided by the PoseExp-LSTM.

The above recurrent dual 3D-2D refinement process is summarized in Algorithm 1. The refinement process does not update the identity vector $\boldsymbol{\alpha}_{\text{id}}$ as the static face will not change during the recurrent refinement process. Similarly, user-specific expression blendshapes $\mathbf{B}$ generated at the initial stage can also be used throughout the entire process.

4.4. Joint Training of 3D Face Fitting and Landmark Detection

A 3D face generated during the refinement which well fits a given face image is essential for accurate 2D landmark detection. The fitness can be simply measured by the following cost function:

$$\mathcal{L}_p = \|S_p - S^*\|_2^2. \tag{8}$$

Here S^* is the manually annotated locations of 2D face landmarks. $S_p = \mathbf{M}(\widehat{\mathbf{Q}})^{\{\mathbf{v}\}}$ is the locations of facial landmarks on the 2D plane from 3D faces generated with $\widehat{\mathbf{Q}}$ (see Eqn. (3)). However, only optimizing this 2D landmark loss is not sufficient as it may lead to 3D faces with obvious artifacts when there is no constraint on the 3D parameters. The difference between predicted 3D fitting parameters $\widehat{\mathbf{P}}, \boldsymbol{\alpha}_{\text{exp}}$ and their ground truth is also considered in training the 3D refinement component. The overall loss becomes

$$\mathcal{L}_{\text{3D}} = \mathcal{L}_p + \lambda\|\widehat{\mathbf{P}} - \mathbf{P}^*\|_2^2 + \lambda\|\boldsymbol{\alpha}_{\text{exp}} - \boldsymbol{\alpha}_{\text{exp}}^*\|_2^2, \tag{9}$$

where λ is the trade-off parameter and $\boldsymbol{\alpha}_{\text{exp}} = \sigma(\widehat{\boldsymbol{\alpha}}_{exp})$.

Within the dual refinement module, the projected 2D shape S_p^k is further refined by the Compensator-LSTM. This can be achieved by optimizing the L2-distance between the updated shape $S = S_p + \Delta S$ and the ground-truth shape S^*. ΔS is the output of the Compensator-LSTM. The overall loss for dual 3D-2D refinement accumulated through K recurrent iterations is then formally described as

$$\begin{aligned}\mathcal{L}_{\text{RDR}} = \sum_{k=1}^{K} \Big(&\|S_p^k - S^*\|_2^2 + \beta\|S^k - S^*\|_2^2 \\ &+ \lambda\|\widehat{\mathbf{P}}^k - \mathbf{P}^*\|_2^2 + \lambda\|\sigma(\widehat{\boldsymbol{\alpha}}_{\text{exp}}^k) - \boldsymbol{\alpha}_{\text{exp}}^*\|_2^2 \Big)\end{aligned} \tag{10}$$

where the projected 2D shape S_p^k at the k-th iteration is given by $S_p^k = \mathbf{M}(\widehat{\mathbf{Q}}^k)^{\{\mathbf{v}\}}$ with $\widehat{\mathbf{Q}}^k = \{\widehat{\mathbf{P}}^k, \boldsymbol{\alpha}_{\text{id}}^0, \sigma(\widehat{\boldsymbol{\alpha}}_{\text{exp}}^k)\}$. Here β and λ are trade-off parameters to ensure that both 3D vertices and 2D landmarks are close to the ground-truth landmark locations.

For $\boldsymbol{\alpha}_{\text{exp}}$, only coefficients for expression blendshapes are predicted. The coefficient α_0 for a neutral face B_0 can be directly obtained through Eqn. (7).

4.5. 3D-to-2D Landmark Projection

We build the correspondence between 3D face and 2D landmarks based on the parallel line method [30], but our way is more robust to extreme conditions. Fig. 3 demonstrates that [30] may not work reliably when the face is with both large yaw and pitch angles. Unlike [30] searches contour vertices solely based on

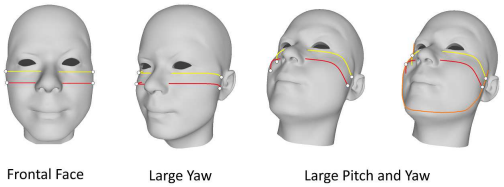

Figure 3. Illustration of applying parallel line method. Parallel lines (red and yellow) of 4 contour points are drawn. The key vertices are moving on corresponding parallel lines. (Best viewed in color)

vertices' horizontal locations on the parallel lines, we find the face contour points by looking for the intersection between true contour line (orange) and parallel lines of contour landmarks (red and yellow). This overcomes the limitation of the original method and gives more robust estimation of landmark correspondence at extreme conditions. More details are provided in Supplementary Material.

5. Experiment

5.1. Implementation Details

Architecture Details Our model is developed based on a variant of ResNet-18 [7]. Due to space limitation, the details of this network are provided in Supplementary Material.

Training Data 300W [19] and 300W-LP [29] datasets are used to train our RDR model. We perform flipping, shifting, scaling and rotation on the cropped face images for data augmentation while fixing the size as 256×256. During training data preparation, we find high 3D fitting error in these face images with extremely large poses (absolute yaw angle greater than $60°$) because of the unified identity constraint [4]. Therefore we only use a subset of 300W-LP for training and the data with 3D fitting error greater than 20% are not included in our training set. Then we obtain in total 510K training images with ground-truth 3D face fitting parameters $\mathbf{Q}^*$ and landmarks locations S^*. The average 3D fitting error with the ground truth parameters is about 5.1%. The 3D ground-truth parameters are obtained by the procedures described in [4].

Training Details We use the pre-trained ResNet-18 provided in [7] to initialize our network. In the experiments, we set the number of recurrent iterations as $K = 5$ and the extracted deep shape-indexed features are reduced to dimension 256 before passing to LSTM units at each iteration. The RDR model is trained via a standard stochastic gradient descent method with momentum of 0.9, mini-batch size of 8 and learning rate of 0.001. The weights of LSTM are randomly initialized with a uniform distribution of $[-0.1,\ 0.1]$. The whole network is trained end-to-end on Caffe [9].

Table 1. Landmark detection results of AFLW on different subsets by yaw angle. 3D approaches perform better than 2D ones on large-pose faces, *i.e.,* $30°-60°$ and $60°-90°$ categories.

Method	$0°-30°$	$30°-60°$	$60°-90°$	Mean
CDM [26]	8.15	13.02	16.17	12.44
ESR [6]	5.66	7.12	11.94	8.24
RCPR [2]	5.43	6.58	11.53	7.85
SDM [25]	4.75	5.55	9.34	6.55
3DDFA [29]	5.00	5.06	6.74	5.60
3DDFA+SDM [29]	4.75	4.83	6.38	5.32
RDR (Ours)	**3.63**	**4.29**	**5.31**	**4.41**

5.2. Datasets

We conduct experiments on two popular landmark detection datasets, AFLW [12] and 300W [19] with the model trained on data prepared in Sec. 5.1. The AFLW dataset contains 21,072 unconstrained faces images with various poses and annotated with 21 landmarks. We follow [29] and divide the face images into three groups based on their absolute ground-truth yaw angles: $[0°, 30°]$, $[30°, 60°]$ and $[60°, 90°]$, corresponding to 11,636, 5,478 and 3,957 face images respectively. These images are randomly selected such that each category has 3,957 samples. We also evaluate RDR on the *common set* and *challenge set* of 300-W testing images. The common set contains test images of unconstrained faces with moderate poses and expressions. The challenge subset contains face images with large poses, extreme expressions and occlusion. During testing, all initial face bounding boxes used are from those provided along with the datasets.

5.3. Performance Analysis

Results on AFLW The 21 landmarks used for evaluation are directly selected from the predicted 68 landmarks in our model. The accuracy of landmark detection on AFLW faces is computed via averaging visible point-to-point distance normalized by width of the provided face bounding boxes [11, 29]. The results are shown in Table 1. We compare RDR with most recent state-of-the-art methods. CDM [26], SDM [25], ESR [6] are 2D-based methods; 3DDFA [29] and D3PF [11] are the most recent 3D face alignment methods based on cascaded CNN regression. 3DDFA+SDM [29] performs SDM refinement on the results returned with 3DDFA.

From Table 1, one can observe that RDR significantly outperforms other methods on all categories. RDR brings around 17.1% relative performance improvement on average, compared with the second best method. RDR also significantly outperforms the state-of-the-art 3DDFA+SDM method [29] on the extreme pose condition, *i.e.* $[60°, 90°]$, by a large margin of 16.7%. D3PF [11] is evaluated on a subset of AFLW, *i.e*, AFLW-PIFA [10]. Our trained model is thus also evaluated on AFLW-PIFA testing images and archives NME of 4.11 which is significantly lower than [11]

Table 2. Landmark detection results on different subsets of 300-W dataset. Results of [6, 2, 25, 18, 27] are from [29].

Method	Common	Challenging	Full set
ESR [6]	5.28	17.00	–
RCPR [2]	6.18	17.26	8.35
SDM [25]	5.57	15.40	7.50
LBF [18]	4.95	11.98	6.32
CFSS [27]	**4.73**	9.98	**5.76**
3DDFA [29]	6.15	10.59	7.01
3DDFA+SDM [29]	5.53	9.56	6.31
RDR (Ours)	5.03	**8.95**	5.80

which has NME of 4.72. More results on AFLW-PIFA are provided in the Supplementary Material.

Results on 300W We further evaluate the performance of our model on the widely used 300-W benchmark. Some sample results are shown in 4-th row of Fig. 5. Following [18, 27, 29], we use the average point-to-point distance error normalized with ground-truth inter-ocular distance for performance evaluation. Table 2 shows that the proposed RDR achieves much better performance than the state-of-the-art deep 3D-face based models, 3DDFA and 3DFFA+SDM. Our model also performs comparably with the state-of-the art 2D-based method [27]. RDR provides the best performance on the challenging subset of 300W, demonstrating superiority on handling challenging faces under unconstrained conditions with large poses and extreme expressions. RDR performs slightly worse than CFSS [27] on the frontal images. This is because after the 3D model in RDR converges, the fitting error term ΔS in Eqn. (6) becomes fixed. Relying on following 2D refinement hardly eliminates such inherent error ΔS. The overall performance for frontal face images is therefore limited by this error from the 3D model.

Recently, a deep learning based methods with heavy structure [24, 13] also gives top performance on 300W but at the cost of high computational resources. Our model provides comparable performance on the challenge subset but requires 87% less computational resources as [24].

5.4. Ablation Study

We here investigate the effectiveness of the individual modules and components of our model, *i.e.* the parameter initialization module, 3D face refinement component and 2D landmark refinement component from the dual-refinement module. We perform the ablation study on the 300W testing set with the same evaluation metric used for 300W in Sec. 5.3. Performance of different network structures against the number of refinement iterations is plotted in Fig. 4.

Direct Initialization Compared with "Mean-3D" which uses mean parameters for initialization, there is a significant performance enhancement brought by "Dual-3D" and "PoseExp-3D" which both adopt direct parameter initialization. This verifies the effectiveness of the 3D parameter initialization module in boosting landmark detection performance of our RDR model.

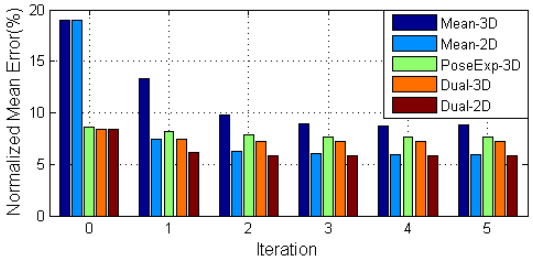

Figure 4. RDR regression performance for ablation study. "Mean-" uses mean parameters as initialization and "PoseExp" only performs refinement on the 3D parameters. "Dual-" represents our full network structure. Postfix "-3D" and "-2D" represent the predicted S_p (Eqn. (4)) and S (Eqn. (6)) of a model.

Dual refinement for Mutual-boosting Performance Fig. 4 shows that the 2D refinement can significantly improve the accuracy of landmark locations inferred from fitted 3D faces by comparing performance of "-3D"s and the corresponding "-2D"s. An obvious performance improvement from "PoseExp-3D" to "Dual-3D" in Fig. 4 shows that 2D refinement can also benefit 3D face fitting. The mutual-boosting performance validates our motivation for dual refinement.

Performance of 3D Face Reconstruction 200 face images of 10 randomly selected subjects from the FacewareHouse Dataset [5] are used to study the impact of dual refinement on 3D face reconstruction. The mean vertex-to-vertex 2D distance normalized by face bounding box is used for evaluation. With dual refinement, the 3D reconstruction error drops from 3.33% at the initial stage to 3.04% at the last iteration. It shows dual refinement effectively improves the quality of the reconstructed 3D face, which is also supported by the converging 3D fitting error shown in Fig. 4.

5.5. Robustness and Accuracy for Large-pose Facial Landmark Detection

The developed RDR model has demonstrated both strong robustness and outstanding accuracy in large-pose facial landmark detection, on $[60°, 90°]$ category of AFLW and Challenging Set of 300W. Obvious performance improvement over recent state-of-the-art 2D-based regression approaches, *e.g.* ESR, RCPR, SDM, LBF and CFSS, is observed for methods developed with 3D information, *i.e.* 3DDFA and RDR, from Table 1 and Table 2. This further verifies the effectiveness of introducing 3D face model for enhanced robustness to large poses in facial landmark detection. Compared with recent 3D alignment approaches [29, 11] that only have a 3D refinement component, our novel dual refinement module shows consistently superior performance on all evaluation benchmarks used. This is consistent with our earlier finding in Sec. 5.4 that 2D refinement can effectively improve

Figure 5. Results of RDR on AFLW (1st to 3th row) and 300W (last row) datasets. From the 1st to 2nd row: 3D-2D overlapped faces and 2D landmark locations. In the last two rows, more face images with large poses and their detected landmarks are shown. Our model can accurately detect the facial landmarks for faces with large poses, extreme expressions and challenging illuminations.

the final landmark detection performance. RDR offers a solution that combines both the robustness of 3D face alignment and the high accuracy for fine-scale landmark displacement of 2D refinement approaches.

Samples of facial landmark detection are shown in Fig. 5. Even for profile faces, *e.g.* 3rd row in Fig. 5, and huge expressions, *e.g.* last row in Fig. 5, our model can still accurately localize the landmarks.

5.6. Time Complexity

One appealing advantage of RDR is that it is much faster than existing deep learning based 3D face alignment methods. Table 3 presents the efficiency comparison of predicting the last step 2D landmark locations. We test our model on GeForce GTX Titan GPU and Intel i7-4930K CPU. It is shown that RDR is much faster than 3DDFA and D3PF. As our current implementation uses CPU to perform deep shape-indexed feature extraction and 3D face fitting, the time cost can be further reduced if using GPU for the computation in these two steps. Overall, RDR offers at least $4.0\times$ speedup over the 3DDFA method on GPU. Since our primary objective is to detect facial landmarks, 488 among 11k vertices are used in the 3D face reconstruction process. This also ensures our efficiency in the refinement process.

As indicated by Fig. 4, RDR rapidly converges within the first two steps. Therefore, we also test the speed of our model running for two iterations. Under this setting, RDR runs at around 45.5FPS on the GPU, which demonstrates great potential for industrial applications.

Table 3. Comparison of time cost (in seconds) for prediction of 2D landmark locations between our proposed RDR and existing deep learning based 3D approaches. Time complexities of 3DDFA [29] and D3PF [11] are cited from their corresponding papers.

	RDR (Ours)	3DDFA [29]	D3PF [11]
GPU (s)	0.031	0.126	-
CPU (s)	0.142	0.213	0.260

6. Conclusion

In this work we developed a novel recurrent dual refinement model that provides new state-of-the-art performance on the challenging AFLW benchmark. It alternatively refines 3D face model and 2D landmark locations and effectively utilizes the informative feedback for achieving mutual boosting. Benefitting from this novel approach, it shows both robustness to large-pose face images and accuracy to small landmark displacement. Moreover, our model economically utilizes deep features for refinement and shows to be much more efficient than current deep learning based 3D face alignment methods.

7. Acknowledgement

The work of Jiashi Feng was partially supported by National University of Singapore startup grant R-263-000-C08-133 and Ministry of Education of Singapore AcRF Tier One grant R-263-000-C21-112.

References

[1] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9):1063–1074, 2003. 1, 2

[2] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1513–1520, 2013. 2, 6, 7

[3] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics (TOG)*, 33(4):43, 2014. 2, 3

[4] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):41, 2013. 2, 3, 5, 6

[5] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. 3, 7

[6] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014. 1, 6, 7

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 6

[8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 6

[10] A. Jourabloo and X. Liu. Pose-invariant 3d face alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3694–3702, 2015. 6

[11] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4188–4196, 2016. 1, 2, 3, 4, 6, 7, 8

[12] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof. 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*, pages 2144–2151. IEEE, 2011. 6

[13] H. Lai, S. Xiao, Y. Pan, Z. Cui, J. Feng, C. Xu, J. Yin, and S. Yan. Deep recurrent regression for facial landmark detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016. 2, 4, 5, 7

[14] H. Li, T. Weise, and M. Pauly. Example-based facial rigging. In *ACM Transactions on Graphics (TOG)*, volume 29, page 32. ACM, 2010. 5

[15] H. Li, J. Yu, Y. Ye, and C. Bregler. Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph.*, 32(4):42–1, 2013. 2

[16] F. Liu, D. Zeng, Q. Zhao, and X. Liu. Joint face alignment and 3d face reconstruction. In *European Conference on Computer Vision*, pages 545–560. Springer, 2016. 1

[17] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas. A recurrent encoder-decoder network for sequential face alignment. In *European Conference on Computer Vision*, pages 38–56. Springer, 2016. 4

[18] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1685–1692, 2014. 1, 2, 4, 7

[19] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18, 2016. 6

[20] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3476–3483. IEEE, 2013. 2

[21] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014. 1

[22] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1, 2016. 1

[23] T. Weise, S. Bouaziz, H. Li, and M. Pauly. Realtime performance-based facial animation. In *ACM Transactions on Graphics (TOG)*, volume 30, page 77. ACM, 2011. 2, 3

[24] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *Proceedings of the European Conference on Computer Vision*, pages 57–72. Springer, 2016. 1, 4, 7

[25] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 532–539, 2013. 1, 2, 6, 7

[26] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1944–1951, 2013. 6

[27] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4998–5006, 2015. 1, 7

[28] S. Zhu, C. Li, C.-C. Loy, and X. Tang. Unconstrained face alignment via cascaded compositional learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3409–3417, June 2016. 2

[29] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, June 2016. 1, 2, 3, 4, 6, 7, 8

[30] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the

wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 787–796, 2015. 1, 2, 5

[31] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012. 1