2019 IEEE International Conference on Multimedia and Expo (ICME)

DYNAMIC CASCADED REGRESSION NETWORK WITH REINFORCEMENT LEARNING FOR ROBUST FACE ALIGNMENT

Zhihao Zhang, Liansheng Zhuang, Wengang Zhou, Houqiang Li

CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System
EEIS Department, University of Science and Technology of China
zzh3@mail.ustc.edu.cn, {lszhuang, zhwg, lihq}@ustc.edu.cn

ABSTRACT

Regression-based methods for facial landmark detection usually learn a series of regressors to update the landmark positions from an initial shape with a fixed number of iterations. Their accuracy is sensitive to the initial shape, and the fixed number of iterations always leads to massive unnecessary computation. In this paper, we propose a Dynamic Cascaded Regression Network (DCRN) with a two-stage architecture to address these issues. In the first stage, we introduce a Global Estimation Network (GEN) to provide a coarse landmark estimation. In the second stage, we propose a Local Regression Network (LRN) to iteratively refine the coarse estimation in a reinforcement learning (RL) paradigm. Our DCRN takes the face image as input, and adaptively learns facial landmarks. Extensive experiments on 300W, COFW, and AFLW datasets show the effectiveness of our proposed method and demonstrate that DCRN consistently achieves the state-of-the-art performance.

Index Terms— face landmark, dynamic cascaded regression, reinforcement learning

1. INTRODUCTION

Face alignment, which refers to facial landmark detection, has drawn significant attention in computer vision due to its wide applications, such as face recognition [1, 2, 3] and face verification [4, 5]. Though significant progress [6, 7, 8] has been made, face alignment remains a very challenging problem. For the face images with large view variations, different expressions, and partial occlusions, most existing methods fail to locate the face landmarks correctly.

Cascaded regression [7, 8, 9, 10, 11, 12, 13] has been a popular method for face alignment with significant progress in the past years. These methods directly learn a series of mapping functions (*i.e.*, regressors) to progressively update the estimation results towards the true locations in an iterative way. Nevertheless, the cascaded regression model suffers from their intrinsic shortcomings. First, current cascaded regression-based methods are sensitive to the initial landmark positions. They rely on the local feature descriptors. As a result, when the initialized shape is far from the true shape, they are prone to get trapped in local optima. Second, existing cascaded regression-based methods fix the number of iterations for the regression process, which typically causes extra computation cost when the initial shape is close to the true shape. Therefore, an adaptive number of iterations may benefit the cascaded regression-based methods.

To address the above issues, we propose a novel Dynamic Cascaded Regression Network (DCRN) with a two-stage architecture for face alignment. In the first stage, a Global Estimation Network (GEN) is proposed to estimate the landmark positions by extracting the global feature from a face image directly. Thanks to the GEN component, our DCRN is free of the landmark position initialization, which is different from most existing cascaded regression-based methods. In the second stage, a Local Regression Network (LRN) is adopted to iteratively refine the coarse estimated landmark positions from the first stage. The regression process stops when a termination criterion is met, so as to dynamically change the number of iterations. By virtue of the dynamic iterative LRN component, our DCRN reduces the computation cost of landmark refinement process. Essentially, the cascaded shape regression process with a termination criterion is formulated as a reinforcement learning (RL) problem, in which LRN corresponds to a policy network which outputs a series of shape increments as *actions*. To solve this continuous decision-making process, we adopt the deep deterministic policy gradient (DDPG) [14] algorithm to train our model.

In summary, our major contributions are listed as follows:

- We propose a novel Dynamic Cascaded Regression Network (DCRN) with a two-stage architecture for coarse-to-fine facial landmark detection. Compared with existing cascade regression-based methods, DCRN is free of landmark position initialization.

This work was supported in part to Dr. Houqiang Li by the 973 Program under Contract No. 2015CB351803 and NSFC under contract No. 61836011, in part to Dr. Liansheng Zhuang by NSFC under contract No. 61472379, and in part to Dr. Wengang Zhou by NSFC under contract No. 61632019 and Young Elite Scientists Sponsorship Program By CAST (2016QNRC001).

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00109

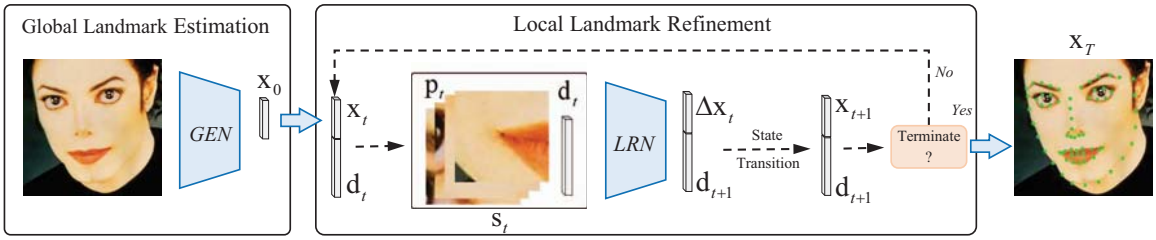

Fig. 1. Architecture of the proposed DCRN. GEN and LRN represent Global Estimation Network and Local Regression Network, respectively. GEN provides a coarse estimation of landmark positions as initialization $\mathbf{x}_0$ for the local regression process. We repeat the local landmark refinement process to update the landmark estimation $\mathbf{x}_t$ and historical actions feature $\mathbf{d}_t$ until the termination criterion is met. The input of LRN is state $\mathbf{s}_t$, which is composed of image patches $\mathbf{p}_t$ and historical actions feature $\mathbf{d}_t$.

- We formulate the cascaded regression process as a decision-making process and train it with reinforcement learning (RL). As a result, our method merely takes about 2∼3 iterations for landmark regression, largely reducing the computation cost.
- Extensive experiments demonstrate that our method is effective, and achieves the state-of-the-art performance on several standard datasets.

2. RELATED WORK

In this section, we briefly review the cascaded regression-based methods and the direct shape regression-based methods, which are closely related with our work.

Classic cascaded regression-based methods such as SDM [9] and CFSS [15] mainly use handcrafted features (*e.g.*, HoGs, SIFT, *etc.*) to drive the cascade process, which may be sub-optimal for face alignment task. To overcome this limitation, MDM [7] introduces the first end-to-end recurrent convolutional network for face alignment, and shows the powerful capability of neural networks. Furthermore, contrary to the methods that rely on local patches, DAN [16] extracts features from an entire face image and visual information about the estimated landmark locations. However, these methods are sensitive to the starting point of the regression process and adopt a fixed number of iterations.

In contrast, direct shape regression-based methods, which are free of landmark initialization, (*e.g.*, [17]) have drawn much attention recently. SHN [18] uses a two-part network, *i.e.*, a supervised transformation to normalize faces and a stacked hourglass network to get prediction heatmaps. In order to take advantage of the shape constraint and the geometric structure, PCD-CNN [13] imposes a shape constraint during the regression process by a dendritic structure of facial landmarks, and disentangles the head pose using a Bayesian framework. Furthermore, LAB [8] uses the stacked hourglass to estimate the facial boundary heatmap, and models the structure between facial boundaries to increase its robustness. Besides, DSRN [19] provides the first end-to-end learning architecture for direct face alignment. It constructs a strong representation to disentangle highly nonlinear relationships between images and shapes, and encodes the correlations of landmarks to improve the performance.

Our method leverages the advantages of the two kinds of methods mentioned above. We utilize a quite simple direct shape regression-based method to provide a coarse landmark estimation, and adopt a cascaded regression process, which is equipped with a termination criterion, to refine the coarse landmark estimation.

3. DYNAMIC CASCADED REGRESSION NETWORK

Face alignment aims to find a mapping from an input image I to a facial shape S represented by the coordinates of pre-defined landmarks in the form of a vector, $[x_1, y_1, \ldots, x_K, y_K]^{\mathrm{T}} \in \mathbb{R}^{2K}$, where K is the number of landmarks.

In this section, we introduce our Dynamic Cascaded Regression Network (DCRN). As shown in Fig. 1, DCRN predicts shapes from images in a two-stage architecture. In other words, GEN provides landmark initialization for LRN, which iteratively refines the coarse estimation. We start with GEN in Section 3.1 and describe LRN in Section 3.2 in detail.

3.1. Global Estimation Network

Cascaded regression-based methods usually update landmark positions from an initial shape, which makes them strongly dependent on initialization. Besides, most of them usually obtain initial landmark positions from a rectangular detection region with a referenced shape. Obviously, it's inappropriate to adopt this initialization method. To solve this problem, we introduce an estimation network to provide a coarse landmark estimation by extracting the landmark information from face image directly. However, a major challenge of this method lies in the highly nonlinear relationship between face images and associated facial shapes. Meanwhile, It requires a quite high generalization ability confronting large pose variation and various facial expression. Following the previous work, *i.e.*, DAN [16], we apply a concise network as our Global

Table 1. Structure of the global facial landmark estimation network. C, P and F represent 2 stacked Convolutional layers, max-Pooling layer and Fully-connected layer, respectively.

Name	C	P	C	P	C	P	C	P	F	F
Kernel	3	2	3	2	3	2	3	2	-	-
Channel	64	-	128	-	256	-	512	-	512	$2K$

Estimation Network (GEN), whose structure is shown in Table 1. Except for the max-pooling layers and the output layer, each convolutional or fully-connected layer is followed by a Rectified Linear Unit (ReLU) layer for activations.

Following the previous works [7, 8, 19] on face alignment, we use Euclidean distance normalized by "inter-ocular" distance as the global loss of landmark location for faster convergence:

$$L_{GEN} = NME(\mathbf{x}, \hat{\mathbf{x}}), \tag{1}$$

where $\mathbf{x}, \hat{\mathbf{x}}$ represent predicted coordinates and the ground truth, respectively, and *NME* represents the normalized mean error, which is defined as follows:

$$NME(\mathbf{x}, \hat{\mathbf{x}}) = \frac{\frac{1}{K}\sum_{i=1}^{K}\sqrt{(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2}}{d}, \tag{2}$$

where $(\hat{x}_i, \hat{y}_i)$ and (x_i, y_i) represent the i-th landmark's ground truth and the predicted coordinates, respectively, and d is the distance for normalization.

3.2. Local Regression Network

After the global estimation of landmark positions, we propose a LRN to refine the coarse prediction. We formulate this cascaded shape regression process as a reinforcement learning (RL) problem. In this subsection, we introduce our formulation for RL-based cascaded shape regression with a novel termination criterion and reward function firstly. Then, we introduce two correlative key networks used in RL algorithm, *i.e.*, actor network and critic network. Finally, we introduce the DDPG algorithm which adopts actor-critic architecture.

In decision-making process, there is an *agent* that interacts with the *environment*, and executes a series of *actions*, so as to optimize a *goal*. In face alignment, the goal is, given a face image I, to locate face landmarks progressively through a series of increments $\{\Delta\mathbf{x}_0, \Delta\mathbf{x}_1, \ldots, \Delta\mathbf{x}_{T-1}\}$, where $\Delta\mathbf{x}_t$ is the shape increments in t-th iteration and T is the number of iterations. Mathematically, a RL problem is defined by states $\mathbf{s} \in \mathcal{S}$, actions $\mathbf{a} \in \mathcal{A}$, state transition function $\mathbf{s}' = F(\mathbf{s}, \mathbf{a})$, reward $r(\mathbf{s})$ and termination criterion $G(\mathbf{s})$.

State & Action. The state $\mathbf{s}_t$ is defined as a tuple $(\mathbf{p}_t, \mathbf{d}_t)$, where $\mathbf{p}_t \in \mathbb{R}^{K\times30\times30\times3}$ denotes the image patches around the estimation landmarks $\mathbf{x}_t \in \mathbb{R}^{2K}$ and $\mathbf{d}_t \in \mathbb{R}^{512}$ represents the historical actions feature denoted by a vector. The action $\mathbf{a}_t \in \mathbb{R}^{2K}$ is defined as shape increments $\Delta\mathbf{x}_t \in \mathbb{R}^{2K}$. The initial conditions is set to $\|\mathbf{d}_0\|_1 = 0$.

State Transition. After decision of action $\mathbf{a}_t$ in state $\mathbf{s}_t$, the next state $\mathbf{s}_{t+1}$ is obtained by the state transition functions $(\mathbf{p}_{t+1}, \mathbf{d}_{t+1}) = F((\mathbf{p}_t, \mathbf{d}_t), \mathbf{a}_t)$: the patches $\mathbf{p}_{t+1}$ are cropped around the landmark estimation $\mathbf{x}_{t+1}$, which is updated by $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{a}_t$; The historical actions vector $\mathbf{d}_{t+1}$ is updated by $\mathbf{d}_{t+1} = LRN(\mathbf{s}_t)$, which is described below.

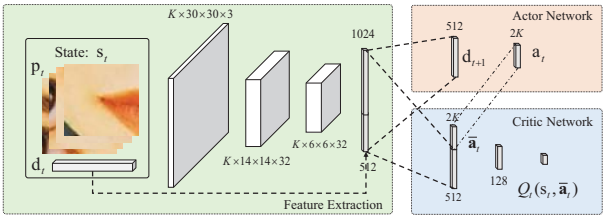

Fig. 2. An illustration of our policy netwok μ_π and state-action value network $Q_\theta(\mathbf{s}, \mathbf{a})$.

Termination Criterion. The cascaded regression process is supposed to be stopped when the distance between the predicted landmarks and the ground truth is within a certain threshold. To achieve a balance between the performance and computational complexity, we adopt the termination function:

$$G(\mathbf{s}_t) = \begin{cases} 1, & NME(\mathbf{x}_t, \hat{\mathbf{x}}) < \alpha_1 \ \&\ \|\mathbf{a}_{t-1}\|_1 < \alpha_2 \cdot K, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

During testing, since $\hat{\mathbf{x}}$ is unavailable, we hold the assumptions that $NME(\mathbf{x}_t, \hat{\mathbf{x}}) \equiv 0$. Note that $G(\mathbf{s}_t) = 1$ represents $\mathbf{s}_t$ is the terminate state.

Reward. The reward function is defined as a function of state $\mathbf{s}$, *i.e.*, $r(\mathbf{s})$, since the agent obtains the reward by the state $\mathbf{s}$ regardless of the action $\mathbf{a}$. In order to increase the stability of RL algorithm, we make a few changes in $r(\mathbf{s})$. When $\mathbf{s}_t$ is not the terminate state, $r(\mathbf{s})$ is assigned by

$$r(\mathbf{s}_t) = \begin{cases} 1, & (1-\beta_1)\cdot NME(\mathbf{x}_{t-1}, \hat{\mathbf{x}}) \geq NME(\mathbf{x}_t, \hat{\mathbf{x}}), \\ -1, & (1-\beta_2)\cdot NME(\mathbf{x}_{t-1}, \hat{\mathbf{x}}) \leq NME(\mathbf{x}_t, \hat{\mathbf{x}}), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

We enforce the refinement process to stop iterating if it does not arrive the terminate state within 10 iterations and set $r(\mathbf{s}_{10}) = -10$. At the terminate state $\mathbf{s}_T$, $r(\mathbf{s}_T) = 10$.

From the RL perspective, we regard LRN as the policy network μ_π (*i.e.*, actor network). The policy network μ_π provides action $\mathbf{a}_t$ at each state $\mathbf{s}_t$ according to regression policy π. As shown in Fig. 2, LRN outputs the shape increments $\Delta\mathbf{x}_t$ (*i.e.*, $\mathbf{a}_t$) and historical actions feature $\mathbf{d}_{t+1}$. To facilitate the RL training, we further introduce a state-action value network (*i.e.*, critic network), which assists to train the policy network and is removed during online estimation. The state-action value function $Q^\pi(\mathbf{s}_t, \mathbf{a}_t)$ is defined as the prediction of the total reward r with the discount rate γ from the observed state $\mathbf{s}_t$ after taking action $\mathbf{a}_t$, following the regression policy π, *i.e.*,

$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t) = \mathbb{E}[r(\mathbf{s}_{t+1}) + \gamma r(\mathbf{s}_{t+2}) + \ldots, |(\mathbf{s}_t, \mathbf{a}_t), \\ \mathbf{a}_{t+1,\ldots,T} \sim \pi]. \tag{5}$$

We approximate the value function using a state-action value network, $Q_\theta(\mathbf{s}, \mathbf{a}) \approx Q^\pi(\mathbf{s}, \mathbf{a})$. It serves as an evaluation

of state $\mathbf{s}_t$ after taking the action $\mathbf{a}_t$. As shown in Fig. 2, our state-action value network concatenates the feature of observed state $\mathbf{s}_t$ and given action $\overline{\mathbf{a}}_t$ to output the evaluated value $Q_\theta(\mathbf{s}_t, \overline{\mathbf{a}}_t)$.

We train LRN using the DDPG [14] approach, the core idea of which is to iteratively update the actor network and the critic network with training sample pairs collected based on the RL rule. It is not feasible to directly apply the original DDPG framework to train our model, since the state and action space are enormous in face alignment problem. To solve this problem, we pre-train LRN through a cascaded regression process with a fixed number of iterations. Obviously, this kind of pre-training inevitably holds some limitations, *e.g.*, it fixes the number of iterations, which leads to poor efficiency and bad interpretability of the fixed number of iterations. Therefore, we utilize the DDPG algorithm to fine-tune the policy network in turn, as elaborated in Algorithm 1.

Algorithm 1 DDPG algorithm

Initialize critic network $Q(\mathbf{s}, \mathbf{a}|\theta^Q)$ and actor network $\mu(\mathbf{s}|\theta^\mu)$ with weight θ^Q and θ^μ according to the pre-training process.
Initialize target network Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^Q, \theta^{\mu'} \leftarrow \theta^\mu$.
Initialize replay buffer R.
for episode=1,$\cdots$,M **do**
  Initialize a random process $\mathcal{N}$ for action exploration.
  Receive initial observation state $\mathbf{s}_0$ using GEN.
  for t = 1,$\cdots$,T **do**
    Select action $\mathbf{a}_t = \mu(\mathbf{s}_t|\theta^\mu, \mathcal{N}_t)$ according to the current policy and exploration noise.
    Store transition $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ in R.
    Sample a random minibatch of N transitions $(\mathbf{s}_i, \mathbf{a}_i, r_i, \mathbf{s}_{i+1})$ from R.
    Set $y_i = r_i + \gamma Q'(\mathbf{s}_{i+1}, \mu'(\mathbf{s}_{i+1}|\theta^{\mu'})|\theta^{Q'})$.
    Update critic by minimizing the loss:

$$L = \frac{1}{N}\sum_i (y_i - Q(\mathbf{s}_i, \mathbf{a}_i|\theta^Q))^2.$$

    Update the actor policy using the sampled policy gradient:

$$\nabla_{\theta^\mu} \approx \frac{\sum_i \nabla Q(\mathbf{s}, \mathbf{a}|\theta^Q)|_{\mathbf{s}=\mathbf{s}_i, \mathbf{a}=\mu(\mathbf{s}_i)} \nabla_{\theta^\mu} \mu(\mathbf{s}|\theta^\mu)|_{\mathbf{s}_i}}{N}.$$

    Update the target network:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'},$$
$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}.$$

  end for
end for

4. EXPERIMENTS

We conduct extensive experiments on three datasets to provide a comprehensive comparison with state-of-the-art methods and verify the significance of each component in our proposed DCRN through the ablation studies.

4.1. Dataset

We select three datasets with different characteristics to train and evaluate our proposed DCRN.

300W [20] dataset is the most widely used benchmark dataset with 68 landmarks. The training part includes 3,148 images and the testing dataset is split into four parts: the common subset (554 images), the challenging subset (135 images), the full set (689 images) and the private test set (600 images).

COFW [21] dataset is designed to depict faces in real-world conditions with partial occlusions. Each COFW face originally has 29 manually annotated landmarks. The training set includes 1,345 face images and 507 face images are used for testing.

AFLW [22] dataset contains a total of 24,386 face images collected in the wild, having large-scale pose variations and a large variety in face appearance. Each image is annotated with up to 21 landmarks. We follow [23] to adopt two settings on our experiments: (1) *AFLW-Full:* 20,000 and 4,386 images are used for training and testing, respectively. (2) *AFLW-Frontal:* 1,314 images are selected from 4,386 testing images for evaluation on frontal faces.

4.2. Experiment Setting

Evaluation Metric. We evaluate our algorithm using standard normalized landmarks mean error. Because of various profile faces on AFLW dataset, we follow [23] to use face size as the normalizing factor d. For other datasets, we follow MDM [7] to use outer-eye-corner distance as the "inter-ocular" normalizing factor d. Specially, to compare with the results with "inter-pupil" (eye-centre-distance) distance normalization, we report our results with both two normalizing factors on Table 3. In addition, the failure rate for a maximum error of 0.1 is also listed.

Implementation Details. All face images are cropped, resized to 256×256 and normalized according to the provided bounding boxes. We conduct data augmentation (*e.g.*, image rotation and image flip) for training. However, for a fair comparison with other methods, face images are re-initialized without any spatial transformation for testing. We employ the stochastic optimization algorithm Adam [24] for the neural network training. The minibatch size is set to 128. We list some key parameters in Table 2.

Table 2. The setting of key parameters in our experiments. We adopt different α_2 on 300W, COFW, and AFLW, respectively.

Parameter	α_1	α_2	β_1	β_2	γ	τ
Value	0.07	3.0 / 2.0 / 1.5	0.03	-0.02	0.9	0.8

4.3. Comparison with Existing Approaches

In this subsection, we provide extensive experimental results of our proposed DCRN and the state-of-the-art methods.

On 300W, as shown in Table 3, compared with recently proposed state-of-the-art algorithms, our DCRN achieves competitive performance on different criteria for different testsets. In particular, our DCRN achieves the state-of-the-art performance using the "inter-ocular" distance to normalize mean error.

On COFW, as shown in Table 4, our DCRN achieves the best performance on the criterion of Mean Error, which verifies its effectiveness of our method confronting the challenge with heavy occlusion. However, our method only achieves competitive performance on the criterion of Failure Rate. The reason may be that LRN cannot receive correct information from the image patches under partial occlusion, which is common on COFW dataset.

On AFLW, as shown in Table 5, our DCRN achieves the state-of-the-art performance, though AFLW has significant view changes and challenging shape variations. The results demonstrate the robustness of the proposed methods for the extreme diversity of samples on AFLW, *e.g.*, large pose, and exaggerated expressions.

Fig. 3 shows some of the difficult images and the predicted visible keypoints on the three datasets.

Table 3. Mean Error (%) on 300W Common Subset, Challenging Subset and Fullset (68 landmarks). The best performances are highlighted in bold.

Methods	Common Subset	Challenging Subset	Fullset
Inter-pupil Normalisation			
SDM [9]	5.57	15.40	7.50
CFAN [25]	5.50	16.78	7.69
CFSS [15]	4.73	9.98	5.76
MDM [7]	4.83	10.14	5.88
RAR [26]	**4.12**	8.35	**4.94**
TSR [12]	4.36	**7.56**	4.99
DCRN	4.64	9.21	5.54
Inter-ocular Normalisation			
DSRN [19]	4.12	9.68	5.21
PCD-CNN [13]	3.67	7.62	4.44
SAN [27]	3.34	6.60	3.98
DCRN	**3.32**	**6.35**	**3.92**

4.4. Ablation Studies

The above studies on datasets with different challenges, *e.g.*, limited training data, partial occlusions, large-scale pose variations, and various appearance, have proved the effectiveness and generality of DCRN. Furthermore, we conduct some additional experiments to verify the effectiveness and potential of different components in our two-stage architecture.

We first introduce two fundamental baselines, *i.e.*, GEN, LRN-m (stack LRN m times), which are proposed to verify the effectiveness of each component. Then, we introduce the baseline, *i.e.*, GEN-LRN8 (connect GEN with LRN stacked 8 times), which is designed to exhibit the strength of our two-stage architecture and the termination criterion. We train them with supervised learning using fully labeled face images. The results are shown in Table 6. Compared with GEN and LRN-m, GEN-LRN8 and DCRN show a significant performance boost in accuracy. This indicates that the two-stage architecture is effective, GEN makes them robust to the landmark initialization. Besides, compared with GEN-LRN8 which performs 8 iterations, DCRN achieves a competitive performance within approximate 2$\sim$3 iterations, which indicates that we can get results faster with the help of the termination criterion. We provide more experimental results in the supplementary materials.

Table 4. Mean Error (%) and Failure Rate (%) on COFW dataset (29 landmarks).

Methods	HPM [28]	DRDA [21]	RAR [15]	DAC-CSR [29]	PCD-CNN [12]	**DCRN Ours**
Mean Error	7.50	6.46	6.03	6.03	5.77	**5.42**
Failure Rate	13.00	6.00	4.14	4.73	**3.73**	4.34

Table 5. Mean Error (%) on AFLW dataset (19 landmarks).

Methods	RCPR [21]	CFSS [15]	DAC-CSR [29]	TSR [12]	SAN [27]	**DCRN Ours**
AFLW-Full	3.73	3.92	2.27	2.17	**1.91**	1.92
AFLW-Frontal	2.87	2.68	1.81	-	1.85	**1.78**

Table 6. Mean Error (%) on different dataset, *i.e.*, the Fullset on 300W, all the testset on COFW and the AFLW-Full on AFLW.

Methods	GEN	LRN			GEN	DCRN	
		-2	-4	-8	-LRN8	-	Steps
300W	6.57	5.16	4.53	4.44	4.02	**3.92**	**2.61**
COFW	6.37	27.32	17.17	16.73	**5.07**	5.42	**3.25**
AFLW	2.15	3.91	3.64	3.85	1.97	**1.92**	**2.09**

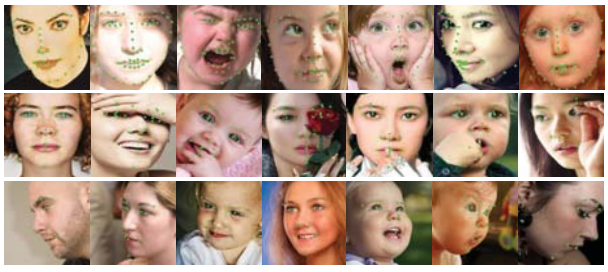

Fig. 3. Qualitative results generated from the proposed method. The green dots represent the predicted points. Each row shows some of the difficult samples from 300W, COFW, and AFLW, respectively, with all the visible predicted points.

5. CONCLUSION

In this paper, we propose the Dynamic Cascaded Regression Network (DCRN) for face alignment. DCRN is a two-stage architecture composed of GEN and LRN. The GEN directly outputs coarse landmark positions based on the global feature and the LRN iteratively refines the landmark estimation based on the local feature. Compared with existing cascaded regression-based methods of facial landmark detection, our DCRN is free of landmark position initialization and dynamically adapts the number of iterations. Extensive experiments show that our DCRN consistently yields high accuracy and it's worth mentioning that DCRN achieves competitive performance within approximate 2$\sim$3 iterations compared with the baseline (*i.e.*, GEN-LRN8) which performs 8 iterations.

6. REFERENCES

[1] Chao-Kuei Hsieh and Yung-Chang Chen, "Kernel-based pose invariant face recognition," in *ICME*, 2007.

[2] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation from predicting 10,000 classes," in *CVPR*, 2014.

[3] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning identity-preserving face space," in *ICCV*, 2013.

[4] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Hybrid deep learning for face verification," in *ICCV*, 2013.

[5] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014.

[6] Shen-Chi Chen, Chia-Hsiang Wu, Shih-Yao Lin, and Yi-Ping Hung, "2d face alignment and pose estimation based on 3d facial models," in *ICME*, 2012.

[7] 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 *CVPR*, 2016.

[8] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *CVPR*, 2018.

[9] Xuehan Xiong and Fernando De la Torre, "Supervised descent method and its applications to face alignment," in *CVPR*, 2013.

[10] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun, "Face alignment by explicit shape regression," *IJCV*, vol. 107, no. 2, pp. 177–190, 2014.

[11] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun, "Face alignment via regressing local binary features," *TIP*, vol. 25, no. 3, pp. 1233–1245, 2016.

[12] Jiang-Jing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, Xi Zhou, et al., "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection," in *CVPR*, 2017.

[13] Amit Kumar and Rama Chellappa, "Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment," in *CVPR*, 2018.

[14] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[15] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang, "Face alignment by coarse-to-fine shape searching," in *CVPR*, 2015.

[16] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," in *CVPR Workshop*, 2017.

[17] Zhiwen Shao, Hengliang Zhu, Yangyang Hao, Min Wang, and Lizhuang Ma, "Learning a multi-center convolutional network for unconstrained face alignment," in *ICME*, 2017.

[18] Jing Yang, Qingshan Liu, and Kaihua Zhang, "Stacked hourglass network for robust facial landmark localisation," in *CVPR*, 2017.

[19] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang, "Direct shape regression networks for end-to-end face alignment," in *CVPR*, 2018.

[20] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *ICCV Workshop*, 2013.

[21] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár, "Robust face landmark estimation under occlusion," in *ICCV*, 2013.

[22] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *ICCV Workshop*, 2011.

[23] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaoou Tang, "Unconstrained face alignment via cascaded compositional learning," in *CVPR*, 2016.

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

[25] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen, "Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment," in *ECCV*, 2014.

[26] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks," in *ECCV*, 2016.

[27] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang, "Style aggregated network for facial landmark detection," in *CVPR*, 2018.

[28] Xiang Yu, Junzhou Huang, Shaoting Zhang, Wang Yan, and Dimitris N Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *ICCV*, 2013.

[29] Zhen-Hua Feng, Josef Kittler, William Christmas, Patrik Huber, and Xiao-Jun Wu, "Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting," in *CVPR*, 2017.