Facial Landmark Point Localization using Coarse-to-Fine Deep Recurrent Neural Network

Shahar Mahpod, Rig Das, Emanuele Maiorana, Yosi Keller, and Patrizio Campisi,

Abstract—Facial landmark point localization is a typical problem in computer vision and is extensively used for increasing accuracy of face recognition, facial expression analysis, face animation etc. In recent years, substantial effort have been deployed by many researcher to design a robust facial landmark detection system. However, it still remains as one of the most challenging tasks due to the existence of extreme poses, exaggerated facial expression, unconstrained illumination, etc. In this paper, we propose a novel coarse-to-fine deep recurrent-neural-network (RNN) based framework, which uses heat-map images for facial landmark point localization. The use of heat-map images allows us using the entire face image instead of the face initialization bounding boxes or patch images around the landmark points. Performance of our proposed framework shows significant improvement in case of handling difficult face images with higher degree of occlusion, variation of pose, large yaw angles and illumination. In comparison with the best current state-of-the-art technique a reduction of 45% in failure rate and an improvement of 11.5% in area under the curve for 300-W private test set are some of the main contributions of our proposed framework.

Index Terms—Face Alignment, Facial Landmark Point Localization, Directed Acyclic Graph Neural Network, Deep Recurrent Neural Network.

1 Introduction

Facial landmark points (e.g., eyebrows, eyes, nose, mouth, jawline etc.) localization or face alignment has emerged as one of the most fundamental components for face recognition [1], face verification [2], and facial attribute inference [3], in recent years. Although a significant amount of progress has been recently observed in the area of facial landmark point localization since Sun et al. in [4] first proposed a deep convolutional-neural-network (CNN)-based approach for this problem, it still remains as a formidable challenge to develop a robust landmark point localization system which extracts landmark points from face images containing extreme poses, illumination, resolution variations, partial occlusion, and large head pose variation [5], [6].

Traditional approaches for the landmark point localization include template fitting methods such as [7], [8], [9], [10] and regression based methods such as [11], [12], [13], [14], [15]. Image features such as SIFT [16], [17] and learned features [18], [19] are extracted from the image patches around each and every landmark points and then iteratively used to refine the initial estimates of the landmark locations [20]. These proposed approaches can be successfully applied to facial landmark point localization of many facial images, but their performance on most challenging datasets such as i.bug [21], [22], 300-W [23], and Menpo [19], [24] leaves a significant scope for further improvement. This is due to the fact that the existing algorithms heavily depend upon the fact of initialization of face mean shape, such as improper positioning of face bounding boxes or failing to detect faces degrades the performance of subsequent landmark point localization a lot. Other than these, existing approaches do not perform well when face alignment is done with arbitrary poses, e.g., faces with yaw angle larger than $45°$, due to the failures in face detection and insufficient training samples [25]. Moreover, the features extracted at disjoint patches of non-frontal images do not provide sufficient information which can lead to improper face alignment [20].

In this paper, we address the aforementioned shortcomings by proposing a novel framework for facial landmark point localization or face alignment using coarse-to-fine deep recurrent neural network (CFDRNN), which uses directed acyclic graph neural network (DagNN) wrapper. DagNN is an object oriented neural network wrapper which allows construction of networks with a directed acyclic graph (Dag) topology. The proposed framework is based on multi-stage neural network strategy where each individual stage refines the landmark positions which are estimated at the previous stage and iteratively improves the estimated landmark locations. At each stage of our algorithm, except the initial one, the input is a face image normalized to a canonical pose, together with an image learned from its previous stage's dense layer. In order to use the entire face image during the face alignment process, we additionally input a landmark heat-map image at each stage [20]. This is one of the key features of our proposed framework. A landmark heat-map image generates high intensity values around the landmark locations where intensity decreases with the distance from the nearest landmark locations. The proposed CFDRNN network utilizes these heat-maps to infer the current estimates of landmark locations in an image and refine them. With the use of landmark heat-map images,

- *S. Mahpod & Y. Keller are with the Faculty of Engineering, Bar-Ilan University, Ramat Gan 52900, Israel (e-mail: {mahpod.shahar, yosi.keller}@gmail.com)*
- *R. Das, E. Maiorana & P. Campisi are with the Section of Applied Electronics, Department of Engineering, Roma Tre University, Via Vito Volterra 62, 00146, Rome, Italy (e-mail: {rig.das, emanuele.maiorana, patrizio.campisi}@uniroma3.it), Phone: +39.06.57337064, Fax: +39.06.5733.7026.*

our proposed CFDRNN framework is able to increase the area under the curve (AUC) on the 300-W private test set up to around 11.5% with respect to the current state-of-the-art techniques.

This paper is organized as follows: Section 2 provides detailed descriptions of the face-in-the-wild datasets that are considered in most of current literature for training and testing purposes, exploited also here to investigate the performance of our proposed framework. Section 3 provides a brief overview of the state-of-the-art techniques for facial landmark point localization or face alignment. Section 4 talks about our proposed CFDRNN and describes the adopted network's topology, while Section 5 details about the experimental setup and proposed CFDRNN training. Section 6 discusses the experimental results and comparison of state-of-the-art methods with the proposed CFDRNN framework, while the conclusions are eventually drawn in Section 7.

2 Face-in-the-Wild Datasets

Annotated facial landmark databases are extremely important in the area of computer vision. There are a number of databases which contains faces with different facial expressions, poses, illumination and occlusion variations, but, most of them do not include images under unconstrained conditions. Hence, recently a few number of databases containing faces in the wild have been collected and in the following a brief overview of few of these datasets are provided.

2.1 Labeled face parts in-the-wild (LFPW):

This database [26] contains 1035 images downloaded from different websites. These images contain large variations including pose, expression, illumination, and occlusion. The database contains 35 and 68 ground truth landmark points for every images.

2.2 Helen:

Helen dataset [27] contains 2330 annotated images collected mainly from Flickr. The images are of high resolutions and contains few images whose sizes are greater than 500×500 pixels. The provided annotations are very detailed in this database and it contains 68 and 194 landmark points for every image.

2.3 Annotated face in-the-wild (AFW):

This database [28] contains 337 images for each of which 68 landmark points are provided.

2.4 XM2VTS:

This database [29] is acquired within M2VTS (Multi Modal Verification for Tele-services and Security applications) project. This dataset contains one frontal view for each of the 295 subjects, with each subject acquired during four sessions, for a total of $2,360$ color images. The images are at resolution 720×576 pixels and 68 landmark points.

2.5 Intelligent behavior understanding group (i-bug):

This dataset [23] contains 135 complicated facial images with different facial expression, poses, illumination, and multiple faces in a single image.

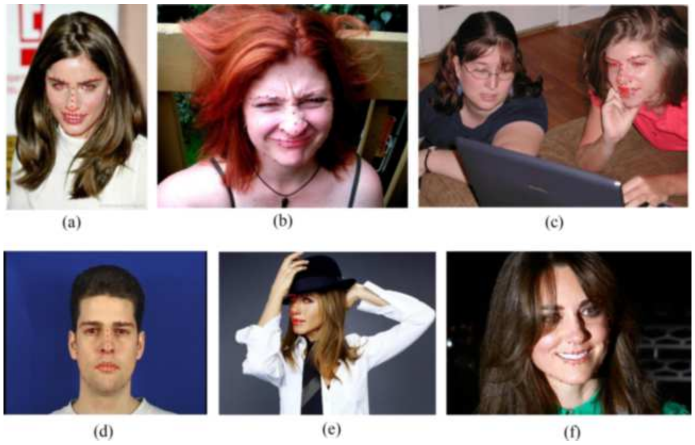

Fig. 1: Annotated face images from (a)LFPW, (b)Helen, (c)AFW, (d)XM2VTS, (e)i-bug, (f)300-W

2.6 300-W:

This dataset [22] is a well-known standard testing dataset with a collection of $3,837$ faces, also containing those from i-bug, with each face densely annotated with 68 landmark points. The test dataset of 300-W consists of 600 images, which can be divided into two parts containing 300 indoor and 300 outdoor images.

2.7 MENPO:

The MENPO challenge dataset [24] consists of a training set with 5658 semi-frontal and 1906 profile facial images, and a test set with 5335 frontal and 1946 profile facial images. The MENPO dataset has been introduced in order to overcome some of the major limitations of the 300-W. For example, the 300-W test set is quite small, only a few hundreds of pictures in total, whereas in MENPO it has been increased up to several thousands. Moreover, 300-W training and test datasets mostly consist of frontal face images, and there are only few side-face images. It is also noteworthy to mention over here that side-face images are also marked with 68 landmark points, which is more suitable for frontal faces.

Figure 1 shows annotated facial landmark images from all the databases. 300-W and i-bug datasets are usually considered for testing and others for training purposes

3 State-of-the-Art: Facial Landmark Point Localization

Face alignment or landmark point localization have already been investigated by many researchers in the past. It started with the active appearance models (AAM) [7], [30] and gradually moved on to constrained local models (CLM) [31], [32]. Whereas, in recent years the focus has been shifted towards the cascaded shape regression (CSR) [12], [16], [18], [33], [34], [35], [36] and deep-learning-based methods [19], [37], [38], [39], [40]. AAM is also known as template fitting model as it builds a face template to fit into the input images [41], [42]. Zhu et al. in [8] have shown that facial landmark point detection can be jointly addressed along with face detection and pose estimation. Moreover in [9], [32] part-based model have been used for face fitting.

CSR-based methods estimate the landmark points explicitly by regression using the image features. The face alignment starts with an initial guess of the landmark locations and then gets refined iteratively. The initial guessed shape is typically an average face shape placed inside the bounding box, which is returned by a face detector such as Viola-Jones [43]. In [44] Valstar et al. proposed a landmark localization prediction method which uses local image patches along with support vector regression. In [11] and [12] authors have employed a cascaded fern regression with pixel difference features for landmark points localization. Some of the articles such as [13], [14], [18], [45], [46], [47] uses random regression forest for casting votes for landmark locations based on local image patch with Haar-like features. The difference between the various CSR-based methods are in the choice of feature extraction and regression method, e.g. supervised descent method (SDM) [16] uses scale-invariant feature transform (SIFT) [48] features and simple linear regressor. Local binary features (LBF) [18] use sparse features which are generated using binary trees and individual pixel's intensity differences. LBF also uses sparse features along with support vector regression [49] and produces a very efficient method which runs at 3000 fps. In [17], Zhu et al. proposed a coarse-to-fine shape searching (CFSS) method, where SIFT features are extracted from landmark points and instead of the regression step of CSR method, a search over the space of possible face shapes is performed, running from coarse to fine scale over several number of iterations. This method considerably reduces the probability of plunging into a local minimum and hence improves the convergence. Tuzel et al. in [36] also use SIFT as feature extraction method, and with the help of mixture of invariant experts, regression is performed. Each of these experts are specialized in a certain part of the space of face shapes. Moreover, the proposed method wraps the input image before each iteration so that the current estimate of the face matches with a predefined canonical face.

Deep-learning-based method such as mnemonic descent method (MDM) [19] combines the feature extraction and regression steps of CSR together into a single recurrent neural network (RNN) and trains it. MDM also introduces memory into the process, allowing information to be passed between CSR iterations. Authors of [4], [15], [50] formulated the face alignment as a regression problem and used multiple deep learning models for locating the landmark points in a coarse-to-fine manner such as cascaded CNN. Cascaded CNN requires a pre-partition of faces into different parts and each of them are processed separately by deep CNN networks. The outputs are subsequently averaged and channeled into separate cascaded layers in order to process each landmark separately. Similarly, Zhang et al. in [15] used successive auto-encoder networks for performing coarse-to-fine face alignment. Authors of [38], [40] estimated the initial landmark points using the entire face image, and used local patches for further refinement. Heat maps for landmark prediction have also been proposed by few authors in recent past. The use of heat-maps for face alignment was initially proposed by Bulat et al. in [37], where a NN predicts the landmark locations.

In [25] authors have proposed a two-stage deep architecture for landmark point prediction. Initially five landmark points, corresponding to the two centers of pupils, nose tip and two mouth corners, are considered as basic landmarks. To detect these basic landmark points for each face they used a module in which heat-maps produced by all landmark points and their associated fields between every two associated landmarks, are detected via a sub-network of heat-map and affinity field prediction [51]. Finally, the whole shape regression sub-network is used for landmark point prediction. In [20] authors have proposed a method which uses heat-maps solely for the purpose of transferring information between different stages. Authors have also proposed a deep alignment network (DAN) which extracts features from entire face image, this being achieved by introducing landmark heat-map images indicating the current estimates of the landmark positions in the face image, and then transmitting the information between different stages.

Some recent state-of-the-art techniques for landmark point localization have been revisited to achieve the best performance in their class for i-bug challenging dataset, being it either common (LFPW and Helen) and full test dataset (i-bug, LFPW and Helen). Specifically, Ren et al. in [18] proposed a face alignment technique which achieves 3000 fps for locating the landmarks by using a set of local binary features and a locality principle for learning those features. The locality principle helps to learn a set of highly discriminative local binary features for each facial landmark independently and then the obtained local binary features are used to jointly learn a linear regression for the final output. They have trained their proposed system with the facial images collected from LFPW [26], Helen [27], AFW [28] and 300-W [22] datasets, and tested it both against the i-bug challenging dataset, a common testing dataset composed of test data collected from Helen and LFPW, and also against the full dataset which is the combination of previous two datasets. The performance of the proposed system are measured in percentage of normalized mean error (NME), and they have been able to achieve 11.98%, 4.95%, 6.32% NME for challenging, common and full dataset respectively.

Ranjan et al. in [5] have trained their proposed framework using only AFLW [52] training dataset and tested over the i-bug challenging dataset and AFLW dataset. They have proposed two different frameworks, namely HyperFace and HF-ResNet, where HyperFace is based on AlexNet [59] architecture and HF-ResNet is based on ResNet-101 [60] model. In case of 68 landmark points 10.88% and 8.18% NME are achieved for HyperFace and HF-ResNet model respectively. In [17], Zhu et al. proposed a face alignment framework based on coarse-to-fine shape searching (CFSS) where their framework begins with a coarse search over a shape space that contains diverse shapes, and employs the coarse solution to constrain subsequent finer search of shapes. Their training is performed over the dataset collected from LFPW, Helen, AFW, and 300-W training datasets. Testing is done over standard challenging, common and full dataset along with LFPW and Helen Test set separately. 9.98%, 4.73%, 5.76% for challenging, common and full dataset respectively whereas 4.87% and 4.63% NME is achieved for LFPW and Helen test dataset respectively. Zhang et al. in [6] have proposed a novel approach named as Tasks-Constrained Deep Convolutional Network (TCDCN), where they have shown that the landmark detection task is not an independent problem. Instead, its robustness can be greatly improved with auxiliary informa-

TABLE 1: Overview of state-of-the-art for facial landmark points localization, where performance is measured in % of normalized mean error (NME)

Paper	Training Databases	Test Databases	Landmarks	Method	Performance NME (%)
Ren et al. [18]	LFPW [26], Helen [27], AFW [28],300-W [22]	i-bug [23]	68	Local binary features and global linear regression	11.98
		Helen+LFPW [26], [27]			4.95
		i-bug+Helen+LFPW			6.32
Ranjan et al. [5]	AFLW [52]	i-bug	68	HyperFace	10.88
				HF-ResNet	8.18
		AFLW [52]	21	HyperFace	4.26
				HF-ResNet	2.93
Zhu et al. [17]	LFPW, Helen, AFW, 300-W	i-bug	68	CFSS	9.98
		Helen+LFPW			4.73
		i-bug+Helen+LFPW			5.76
		LFPW [26]			4.87
		Helen [27]			4.63
Zhang et al. [6]	MAFL [53], AFLW, COFW [11], Helen,300-W	i-bug	68	TCDCN	8.60
		Helen+LFPW			4.80
		i-bug+Helen+LFPW			5.54
		Helen			4.60
Xiao et al. [39]	LFPW, Helen, AFW, 300-W	i-bug	68	RAR	8.35
		Helen+LFPW			4.12
		i-bug+Helen+LFPW			4.94
		LFPW			3.99
		Helen			4.30
Lai et al. [54]	LFPW, Helen, AFW, 300-W	i-bug	68	Deep recurrent regression network	8.29
		Helen+LFPW			4.07
		i-bug+Helen+LFPW			4.90
		LFPW			4.49
		Helen			4.02
Shao et al. [25]	CelebA [55], 300-W, MENPO [24]	i-bug	68	Landmark heatmap & whole landmark regression	8.03
		Helen+LFPW			4.45
		i-bug+Helen+LFPW			5.15
Chen et al. [56]	Helen, 300-W, MENPO	i-bug	68	4-stage coarse-to-fine framework	7.12
		Helen+LFPW			3.73
		i-bug+Helen+LFPW			4.47
Kowalski et al. [20]	LFPW, Helen, AFW, 300-W	i-bug	68	DAN	7.57
		Helen+LFPW			4.42
		i-bug+Helen+LFPW			5.03
		300-W private test set [23]			4.30
	LFPW, Helen, AFW, 300-W, MENPO	i-bug		DAN-Menpo	7.05
		Helen+LFPW			4.29
		i-bug+Helen+LFPW			4.83
		300-W private test set [23]			3.97
He et al. [57]	LFPW,Helen,AFW,300-W,MENPO	i-bug	68	Robust FEC-CNN	6.56
Chen et al. [58]	LFPW,Helen,AFW,i-bug	300-W private test set	68	GAN	3.96

tion, such as gender, expression, and appearance attributes. They have trained their system with multi-attribute facial landmark (MAFL) [53], Caltech occluded faces in the wild (COFW) [11], AFLW along with the standard Helen, and 300-W dataset for 68 landmark points. Performance of 8.60%, 4.80%, 5.54% NME are achieved for i-bug challenging dataset, common dataset and full test dataset respectively, whereas 4.60% NME is achieved Helen test set .

Xiao et al. in [39] have proposed a novel recurrent attentive- refinement (RAR) network for facial landmark detection. RAR works similarly as cascaded regression, refining landmark locations in a sequential manner at each recurrent stage via multi-stage predictions. Their proposed framework is trained using the standard LFPW, Helen, AFW, and 300-W datasets and for 68 landmark points, being able to achieve 8.35%, 4.12%, 4.94% NME for challenging, common and full dataset respectively. For LFPW and Helen test sets, their method achieves 3.99% and 4.60% NME respectively. Using the same training and test dataset Lai et al in [54] have been able to achieve 8.29%, 4.07%, 4.90% NME for challenging, common, and full test set respectively, using a deep recurrent regression (DRR) network. Initially authors have encoded an input face image to resolution-preserved deconvolutional feature maps via a deep network with stacked convolutional and deconvolutional layers. Then, they estimated the initial coordinates of the facial key points by an additional convolutional layer on top of these deconvolutional feature maps and finally, by using the deconvolutional feature maps and the initial facial key points as input, they refine the coordinates of the facial key points by a recurrent network that consists of multiple long-short term memory (LSTM) components. Authors have also shown that their proposed method is able to achieve 4.49% and 4.02% NME for LFPW and Helen test dataset.

In [25] Shao et al. have proposed a deep architecture which directly detects facial landmarks without using face detection as an initialization step. Given an input image, basic landmarks such as the two centers of pupils, nose tip and two mouth corners of all faces are first detected by a sub-network of landmark heatmap and affinity field prediction. Then, the coarse canonical face and the pose are generated by a pose-splitting layer based on the visible basic landmarks. Finally, according to its pose, each canonical state is distributed to the corresponding branch of the shape regression sub-networks for the complete landmark detection. Training of the proposed network is performed using CelebA [55], 300-W and MENPO [24] training sets and 8.03%, 4.45% and 5.15% NME are achieved for three standard test sets respectively. Chen et al. in [56] used Helen, 300-W, and MENPO sets for training, and have been able to achieve NME of 7.12%, 3.73% and 4.47% respectively for the three corresponding standard testing sets, using their proposed 4-stage coarse-to-fine framework. Initially a pre-trained face detector is used to locate the face, and then a CNN is employed to predict the rotation angle of cropped image. The cropped images are then rotated to a horizontal-canonical position and fed into another CNN to predict the coarse landmark. Later, they separate the landmarks into several components, and predict each component's associated landmarks. Finally, they refine each point with multi-scale local patches cropped according to its 3, 5, and 7 nearest neighbors. The results are then fused using an attention gate network. A linear transformation is finally learned with the least square approximation to finally predict the landmark points.

Kowalski et al. in [20] have proposed a deep alignment network (DAN) for face alignment. It is based on multi-stage neural network where each stage refines the landmark positions estimated at the previous stage. The input to each stage are a face image normalized to a canonical pose and an image learned from the dense layer of the previous stage. To make use of the entire face image during the process of face alignment, they have additionally inputed a landmark heatmap at each stage. A landmark heatmap image comes up with a high intensity values around landmark locations where intensity decreases with the distance from the nearest landmark. The CNN can use the heatmaps to infer the current estimates of landmark locations in the image and thus refine them. Authors have proposed two networks named as DAN and DAN-Menpo, where for the first one the network is trained with standard 4 training datasets, and for the later one an additional MENPO [24] dataset is used for training. DAN-Menpo performs better than the DAN framework, being able to achieve 7.05%, 4.29%, and 4.83% NME for three standard test sets respectively. The authors have also tested their proposed framework's performance against the 300-W private test set [23], which consists of 300 indoor and 300 outdoor images. Using DAN framework they have been able to achieve 4.30% of NME whereas using DAN-Menpo 3.97% NME is achieved. He et al. in [57] have proposed a robust fully end-to-end cascaded CNN (Robust FEC-CNN) based on FEC-CNN [61] architecture. Using the 4 standard training datasets along with MENPO training data they have been able to achieve NME=6.56% for i-bug challenging dataset.

In [58], Chen et al. have proposed a generative-adversarial-network (GAN)-based landmark localization framework. They have incorporated priors about the structure of pose components, and proposed a novel structure-aware fully convolutional network to implicitly take priors of the structure of pose for the training of the deep network. The LFPW, Helen, AFW and i-bug datasets have been used for training, and the 300-W private test set for testing, achieving 3.96% NME for the 300-W private test set, which is merely 0.25% lower than the DAN-Menpo [20] method.

Table-1 provides a summary of the recent state-of-the-art techniques for facial landmark point localization, where the performance of every method is measured using normalized mean error (NME). The proposed framework in this paper deals with the problem in a unique way, by using two recurrent stages of different types of information. The first stage/layer produces the solution in a rough way using the heat-maps space and the second stage/layer refines the result in the physical location space. With this approach, it is possible to improve the resolution even for large images significantly more than most of the state-of-the-art papers which we discussed so far. In addition to that, this approach can also be used as a recurring problem, in which there is one stage providing primary results, and a secondary stage improving such results, while both stages are implemented in the same network. Also, unlike some of the other method, this approach performs the training of all coarse and subtle refinements within the same network and does not require the initial guessing of the landmarks.

4 Proposed Model Description

Face alignment is a retrieval problem that is defined to localize $N = 68$ points of interest or landmarks, i.e., $L^i = [x_i, y_i]$, with, $i = 1, ..., N$ in a facial image $I \in \mathbb{R}^{w \times h}$. Let us also denote a combination of N landmark points as $L_j = \{L_j^i\}$, where, $j = \{1, 2, 3, 4\}$ for our proposed CFDRNN method. Similar to some of the existing state-of-the-art methods, we also deal with the face alignment challenge through neural networks (NN) by proposing a novel CFDRNN-based framework. In this section we will describe our proposed structure which is designed for landmark point localization, in detail.

The architecture of CFDRNN is composed of two main levels. The first level is named as recurrent heat-map level (HML), which finds low resolution landmarks using heat-mapping technique, while the second level, i.e., recurrent landmark level (LML), improves the accuracy of the estimated landmark values that are obtained from HML, by resorting into regression technique.

Figure 2 describes the overall structure of our proposed framework. In general, the entry to the network is an image of $I_0 = 256 \times 256$ size and like in [62], in the first stage the image moves through the resizer blocks containing layers,

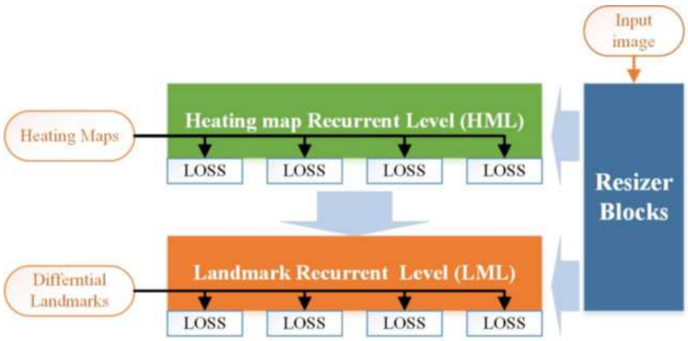

Fig. 2: General architecture of the CFDRNN framework. The network consists of a pre-block that is responsible for the resizing of image, the recurrent Heat-map Level, and recurrent Landmark Level.

which reduce it by a factor of $s = 4$ and produce two reduced image matrices (D_0 and D_r) of size $[64 \times 64 \times 68]$, where $[64 \times 64]$ corresponds to the image matrices size for $N = 68$ landmark points.

Image matrices are then passed through the HML stage, whose output are the heat-map images, i.e., $\{H^i\}_{i=1,...,N}$, of the landmark points. In order to extract the roughly estimated values of the landmarks $\hat{L}^i_j = [\hat{x}_i, \hat{y}_i]$, where $j = \{1, 2, 3, 4\}$ and the maximum point in each one of the heat-map images should be found and its indexes are multiplied by the s factor.

$$\hat{L}^i_j = [s \times max_x(H^i_j), s \times max_y(H^i_j)] \tag{1}$$

where $\hat{L}^i_j$ are the roughly estimated landmark points of L_j with the estimated values $[\hat{x}_i, \hat{y}_i]$. The functions max_x and max_y finds the indexes (x,y respectively) of the maximum value of a landmark point at heat-map H^i_j, where $i = \{1, ..., N\}$ landmark points and $j = \{1, 2, 3, 4\}$ stages. Which means H^i_j is the heat-map for ith landmark point for the jth stage of the network.

At the final step, the HML products, which are composed with reduced images, enter into the LML stage, whose output are the values $\Delta L^i_j = [\Delta x_i, \Delta y_i]$ which are the difference between the rough landmark values, $\hat{L}^i_j = [\hat{x}_i, \hat{y}_i]$, and the refined final landmark values, $L^i_j = [x_i, y_i]$, which the process attempts to produce. Therefore,

$$\Delta L^i_j = L^i_j - \hat{L}^i_j \tag{2}$$

and,

$$[\Delta x_i = x_i - \hat{x}_i, \Delta y_i = y_i - \hat{y}_i] \tag{3}$$

A full formulation of the procedure to restore the refined landmarks appears at Eq. 4,

$$L^i_j = [s \times max_x(H^i_j) + \Delta x_i, s \times max_y(H^i_j) + \Delta y_i] \tag{4}$$

where L^i_j with $j = \{1, 2, 3, 4\}$ are the estimated landmark points with the estimated values $[x_i, y_i]$, the functions max_x and max_y finds the indexes (x,y respectively) of the maximum value of a landmark point at heat-map H^i_4, and $[\Delta x_i, \Delta y_i]$ are as define at Eq. 3.

TABLE 2: Resizer Block - A sequence of layers that receives an I_0 image at the entrance and exports a small image D_0 in the middle of the chain and at the end produces the H_1 heat-maps series.

	IN	OUT	Type	Size	Stride
1	I_0		conv	[3x3]x64	1
2			relu	-	
3			conv	[3x3]x64	1
4			relu	-	
5			m-pool	[2x2]	2
6			conv	[3x3]x64	1
7			relu	-	-
8			conv	[3x3]x128	1
9			relu	-	-
10			m-pool	[2x2]	2
11			conv	[3x3]x128	1
12			relu	-	-
13			conv	[3x3]x128	1
14		T_0	relu	-	-
15	T_0	D_0	conv	[1x1]x68	1
16	T_0		conv	[9x9]x128	1
17			relu	-	-
18			conv	[9x9]x128	1
19			relu	-	-
20			conv	[1X1]x256	1
21			relu	-	-
22			conv	[1x1]x256	1
23			relu	-	-
24			dropout	-	-
25			conv	[1x1]x68	1
26		H_1	relu	-	-

4.1 Resizer Blocks

Inspired by Belagiannis et al. work in [62], the resizer block is responsible for two main functions. The first function, which also constitutes the first part of the block, which is shown as row no. 1-15 of Table-2, reduces the size of the image. It relies on layers with narrow filters of size 3 × 3, while reducing its products by using stride functions. The second function, which is expressed as the second part of the block and are shown as row no. 16-26 of Table-2 uses filters with a wide margin of 9 × 9 size to create connections between distant elements/pixels. The final output is the estimated heat-map collection $\{H^i\}_{i=1,...,N}$. In this part of the block, the layers preserve the product sizes, i.e., the stride is always 1, and the output width and height are the same as those of the first part. A point to be noted here is that a temporary intermediate block named, T_0, is produced in row no. 14 which also goes as an input to row no. 15 and 16.

A complete and detailed description of the resizer block layer's structure is given in Table 2, where the entry to this block is the image I_0, the resized image is D_0, and the final heat-map collection output of the block is $H_1 = \{H^i_1\}_{i=1,...,N}$, where H_1 is the first estimated heat-map image for all $N = 68$ points. The output D_0 and H_1 will be used as concatenated input data for the HML layer. It is noteworthy to mention over here that the "IN" and "OUT" columns of Table 2 represent the entry and exit of each layer, respectively. For convenience of reading, in cells, where there is no content, the entry to that layer is the output of its previous layer, unless it is otherwise specified.

In our network architecture, we use a double instance of the resizer block, where the second instance inserts its D_r and H_r outputs concatenated with D_0 into the LML layer

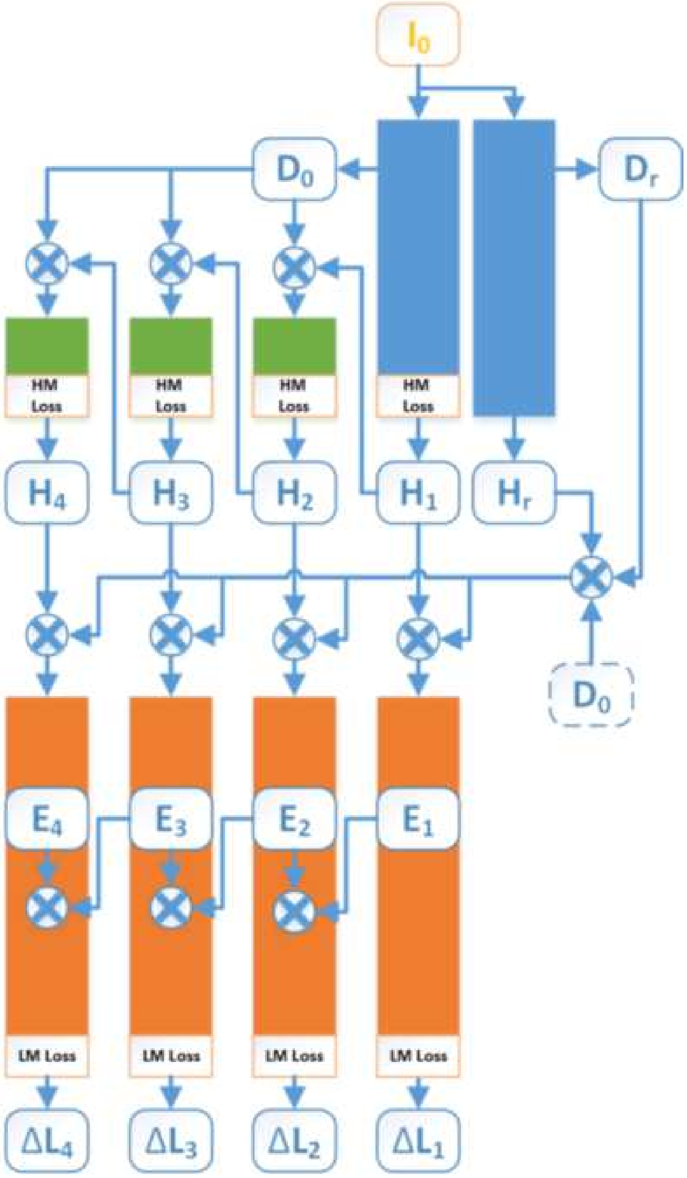

Fig. 3: A scheme of connections between the different blocks of the proposed network, where the two resizer blocks' instances painted in blue, the layers that make up the HML are painted in green, and the LML layers are painted in orange. The symbol $\oplus$ is used for concatenation between inputs in the third dimension.

only without using the loss function at the end of the block. Conversely, the first instances of the blocks, which end in a loss function and whose final product is H_1, enter the HML layer only. Figure 3 shows the link between two resizer blocks' instances for HML and LML stages and internal links between each of these levels.

4.2 Recurrent Heat-Map Level

Heat-map images are designed to describe the $[x_i, y_i]$ position of a particular point of interest, L_j^i, in an image by lighting a single pixel at $[x_i, y_i]$ where all other values in the image remain zero. Usually we will not use only a single point, but a narrow Gaussian, whose center is in this location. Formulation for heat-map H^i can be described as Eq.5.

TABLE 3: Recurrent Heat-map level (HML)- A sequence of layers that represents one branch of the HML, whose input is the reduced image D_0 concatenated with the heat-maps product H_j of the previous layer, and its outcome is an updated heat-maps series H_{j+1}.

	IN	OUT	Type	Size	Stride
1	$D_0 \oplus H_j$		conv	[7x7]x64	1
2			relu	-	-
3			conv	[13x13]x64	1
4			relu	-	-
5			conv	[1x1]x128	1
6			relu	-	-
7			conv	[1x1]x68	1
8		H_{j+1}	relu	-	-

TABLE 4: Recurrent Landmark Level (LML)- A sequence of layers representing one branch of LML, whose input is a collection of threaded entrances, from resizer blocks and HML layers. Each of the branches also has another entrance E_{j-1} in its middle that joins one of its layers, and is received from the middle of the previous branch. The result of each ΔL_j branch is the difference between the landmarks positions as can be recovered from the heat-maps H_j and the high-resolution truth values of the landmarks.

	IN	OUT	Type	Size	Stride
1	$D_0 \oplus H_j \oplus$ $D_r \oplus H_r$		conv	[7x7]x64	2
2			m-pool	[2x2]	1
3			conv	[5x5]x128	2
4			m-pool	[2x2]	1
5			conv	[3x3]x256	2
6		E_j	m-pool	[2x2]	1
7	$E_j \oplus E_{j-1}$		conv	[3x3]x512	2
8			m-pool	[2x2]	1
9			conv	[3x3]x1024	2
10			m-pool	[2x2]	1
11		ΔL_j	conv	[1x1]x136	2

In this network, the $[x_i, y_i]$ values are not for the original positions of input image but only a coarse solution for them, and at most will allow $1/4$ of the original resolution of the landmarks.

$$H^i = e^{-\frac{(x-x_i)^2+(y-y_i)^2}{2\sigma^2}} \tag{5}$$

The recurrent heat-map level (HML) is composed from several "branches", where each one of them is designed to get the estimated heat-maps H_j, with $j = \{1, 2, 3\}$, as input and improve it to H_{j+1}, which will be much closer to the ground truth values of heat-maps collection $\{H_j^i\}_{i=1,\ldots,N}$. Table 3 shows in detail a branch structure, and from the stride it can be seen that there is no reduction between layers. The input for each branch is the previous solution H_j, concatenated (denoted with $\oplus$ symbol in the table) with the resized image D_0, which means H_{j+1} can be produced by using HML branch on $D_0 \oplus H_j$.

We also use very wide filters ($[7 \times 7]$ and $[13 \times 13]$) for connection between far elements at the input data, especially because we want the filter to take into account the relationship between far landmarks. For example, the convolution filter might find a template that is very similar to an eye but because it does not find an ear (that mostly exists nearby an eye in images) in the wide spread of the filter it will not give high response for it.

The loss of those branches, and of the resizer block, is the $L2$ loss, when the heat-maps contain only the landmark positions without pairs of components. This is in contrast to the technique employed by Belagiannis and Zisserman in [62]. Figure 4 shows heat-map images for landmark points obtained for different faces starting from singular heat map point.

4.3 Recurrent Landmark Level

The LML is composed by several "branches", where each of them is designed to improve the results extracted by its previous branch. The goal of this layer is to improve the accuracy of the landmark locations, as they are converted from values marked on heat-maps to scalar values. Each of these branches has an entrance consisting of a composition of several links from the previous parts of the network.

As shown in Table 4, all the entrances of the branches have four components, three of which are identical between them, and the fourth is unique to each branch. The first component is D_0, which also serves as a reduced image at the entrance to the HML branches. Two more components, i.e., D_r and H_r, are obtained from the secondary resizer block (Fig. 2), and their purpose is to allow the LML to receive inputs similar to those of the HML, with no effect of the HML stages on them. The rationale behind using a secondary block of resizer is that the HML stage images contains multiple convergence variables relative to those produced in the LML stage i.e., ΔL_j values, where $j = \{1, 2, 3, 4\}$, so that the latter have a lower impact on the network training. When generating a separate additional input, the LML loss functions can independently affect these parts of the network. The fourth component is an entry that changes between the branches, with each branch having its own uniqueness that comes from the corresponding branch in the HML layer.

In addition to these entrances, we connect the branches by inserting the product of one of the layers in the middle of the branch E_j, and connecting it to the successive layer's product E_{j+1} in the next branch. In this way we allow a process in which each branch also has data on a preliminary evaluation of the result in the same layer. The end result of the network is the absolute landmark positions, calculated according to Eq. 4, where the heat-maps are obtained from the product H_4, and the relative LM values are extracted from ΔL_4.

5 Experimental Setup & Training

In this section we will provide detailed description of our network training and the experimental setup that have been used for our training and testing.

5.1 Training Parameters

As proposed by Belagiannis and Zisserman in [62], we also use augmentation to expand our data set. The augmentation includes color change of the image, rotation at small angles, enlargement, reduction, and spatial displacements. The processed inputs inserted into the network are RGB images with size of 256×256, and their pixel values are normalized to the range of values between $[-0.5, 0.5]$.

In our proposed work, we use the standard set by the 300-W competition, defining $N = 68$ locations for the landmarks. Therefore, the number of heat-maps generated from a branch in the HML phase is 68, and the number of pair of values generated from the LML phase is also 68. All heat-maps are 64×64 images with a background value of 0, each containing a representation of one landmark using a symmetric Gaussian, whose width is given by $\sigma = 1.3$. The LML stage products represents the fine difference between the final position of the landmark and the coarse position produced from the estimated heat-maps by finding the maximum within each and multiplying it by 4 (the ratio between the size of the original image and the heat-map image size). The values generated at the output of the network are normalized in the size of the image, in order to insert them into the range of values of $[0, 1]$.

In all of the network's linear layers (conv layers), we use the batch normalization method, and in the last layer of the resizer block a dropout layer with a 0.5 injection ratio is used. Due to constraints of memory limitations on the GPU card we use for the network training, the size of each batch is limited to only 2. The learning rate changes gradually depending on the number of epochs that is running at that particular time. At the start of the iteration/epoch, the learning rate is fixed at $1e^{-5}$ for the initial 30 epochs, whereas, it becomes $5e^{-6}$ for the next 5 epochs. After the first 35 epochs the learning rate is fixed at $1e^{-6}$ and remains the same till the end of our training, i.e., till the end of number of epochs. For the training we choose 1000 epochs/iteration as our limit.

5.2 Network Training

Our proposed CFDRNN network is trained using the training parameters as explained in the previous subsection, and using images from LFPW and HELEN training set, AFW, XM2VTS, and MENPO training set. We follow the most established approach for training of our CFDRNN network, and for that purposes divide the 300-W competition data into training and testing parts. The Menpo challenge training dataset consists of 6679 semi-frontal and profile face image datasets. These images are actually a collection of images selected from FDDB and AFLW datasets. The image were annotated with the same set of 68 landmarks as the 300-W competition data but without any face detector bounding boxes. Details of all these databases are provided in Section 2. This constitutes in total $11,007$ original images and their augmented images, generated as described in Section 5.1. For our proposed DagNN-wrapper based CFDRNN network, designing and training using MatConvNet-1.0-beta23 tool [63] is performed, and all the experiments are performed in MATLAB® (R2017a) with a system configuration of $64GB$ RAM; Titan X™(Pascal) graphics card; i7, $3.40GHz$ processor and Windows® 10 operating system. For validation we use a random subset of 2500 images from the training set.

6 Results & Discussion

In this section we perform an extensive evaluation of our proposed CFDRNN framework on several publicly available test datasets. The following subsections will provide

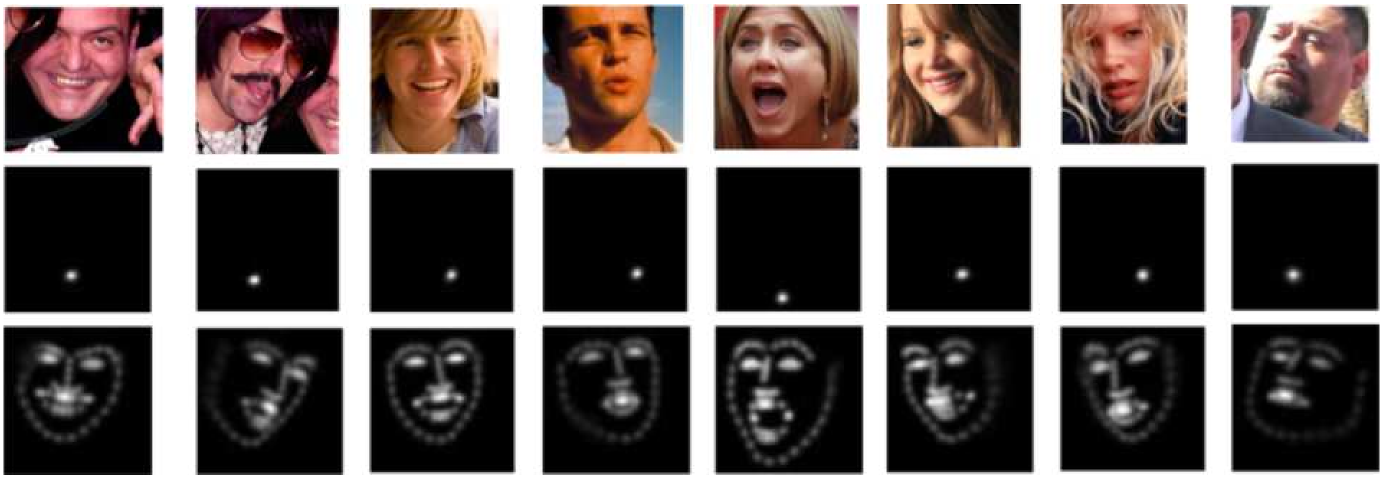

Fig. 4: Selected images from the training dataset and their heat-maps. The first row shows the images, second row corresponding one point heat-map images and the third row shows the corresponding 68 point heat-map images

extensive details about the test datasets, error measures and Section 6.3 compares our method with the state-of-the-art techniques.

6.1 Test Dataset

In order to evaluate the performance of our proposed CFDRNN framework we perform some exhaustive experiments on the data released for the 300-W competition [23]. The 300-W competition data is a combination of images from five datasets, such as LFPW, Helen, AFW, i-bug, and 300W private test set. The last dataset was originally used for evaluating competition entries and at that time it was private to the organizers of the competition, hence the name. Each image in these datasets is annotated with 68 landmarks and accompanied by a bounding box generated by a face detector. The test data consists of the remaining datasets, i.e., i-bug, 300W private test set and test sets of LFPW, Helen. In order to facilitate a fair comparison with state-of-the-art methods we split these test data into four subsets:

- the common subset which consists of the test subsets of LFPW and Helen (554 images),
- the challenging subset which consists of the i-bug dataset (135 images),
- the 300W public test set which consists of the test subsets of LFPW and Helen as well as the i-bug dataset (689 images),
- the 300W private test set, consists of indoor and outdoor images (600 images).

The annotation for the images in the 300-W public test set were originally published for the 300-W competition and we use them for testing as it became a common practice to do so in the recent years.

6.2 Error Measures

Several error measurement techniques for face alignment in an individual face image have been introduced in different state-of-the-art methods. The most common of them are the following two which we also use for our proposed CFDRNN method's error measurements. These two methods are as follows:

- the mean distance between the localized landmarks and the ground truth landmarks divided by the inter-ocular distance (the distance between the outer eye corners) [17], [18], [24],
- the mean distance between the localized landmarks and the ground truth landmarks divided by the inter-pupil distance (the distance between the eye centers) [19], [21],

For evaluating our method on the test datasets we use three metrics: the mean error, the area under the cumulative error distribution curve (AUC$_\alpha$) and the failure rate. Similar to [19], [36], we calculate AUC$_\alpha$ as the area under the cumulative distribution curve calculated up to a threshold α, then divided it by that threshold. As a result, the range of the AUC$_\alpha$ values is always in between 0 and 1. Following [19], we consider each image with an inter-ocular normalized error of 0.08 or greater as failure and use the same as threshold for AUC$_{0.08}$. In all the experiments we test on the full set of 68 landmarks.

6.3 Comparison with state-of-the-art

We compare the CFDRNN model with the state-of-the-art methods on all of the test sets of the 300-W competition data. Tables- 5 and 6 show the normalized mean error (NME). Table 7 shows the AUC$_{0.08}$ and failure rate of the proposed method in comparison with other methods on 300-W public and private test set. All of the experiments are performed on three of the most difficult test subsets i.e., the challenging subset or 300-W public test set and the 300W private test set. Therefore, the proposed method achieves, in comparison with DAN-Menpo [20] method:

- a failure rate reduction of 45% on the 300-W private test set,
- an improvement in AUC$_{0.08}$ of 11.5% on the 300-W private test set,
- a 13% improvement of the mean error on the challenging subset.

This shows that our proposed CFDRNN framework is particularly suited for handling difficult face images with a high

TABLE 5: Normalized mean error percent (NME in %) of face alignment methods on the 300-W public test set and its subsets. Best results are highlighted.

Method	Common Subset	Challenging Subset	Full Set
Inter-pupil normalization			
LBF [18]	4.95	11.98	6.32
CFSS [17]	4.73	9.98	5.76
TCDCN [6]	4.80	8.60	5.54
RAR [39]	4.12	8.35	4.94
DRR [54]	4.07	8.29	4.90
Shao et al. [25]	4.45	8.03	5.15
Chen et al. [56]	**3.73**	7.12	**4.47**
DAN [20]	4.42	7.57	5.03
DAN-Menpo [20]	4.29	7.05	4.83
Robust FEC-CNN [57]	-	6.56	-
CFDRNN	4.55	**6.05**	4.85
Inter-ocular normalization			
MDM [19]	-	-	4.05
k-cluster [34]	3.34	6.56	3.97
DAN [20]	3.19	5.24	3.59
DAN-Menpo [20]	**3.09**	4.88	**3.44**
CFDRNN	3.23	**4.24**	**3.44**

TABLE 6: Normalized mean error percent (NME in %) of face alignment methods on the LFPW, Helen and 300-W Private test sets separately. Best results are highlighted.

Method	LFPW	Helen	300-W Private Set
Inter-pupil normalization			
CFSS [17]	4.87	4.63	-
TCDCN [6]	-	4.60	-
DRR [54]	**4.49**	**4.02**	-
CFDRNN	4.63	4.51	**4.99**
Inter-ocular normalization			
RAR [39]	3.99	4.30	-
MDM [19]	-	-	5.05
DAN [20]	-	-	4.30
DAN-Menpo [20]	-	-	3.97
GAN [58]	-	-	3.96
CFDRNN	**3.30**	**3.20**	**3.49**

degree of occlusion and variation in pose, yaw angle and illumination.

Figures 5-7 plots the proposed CFDRNN framework's $\text{AUC}_{0.08}$ performance against the 300-W challenge in 2015. The plots shows the cumulative error distribution (CED) curves with respect to the landmark points for indoor test images in Fig.5, for outdoor test images in Fig.6 and in Fig.7 for combined indoor and outdoor images of 300-W private test set. All three figures show a significant improvement in the results over the top 5 results of 300-W challenge of 2015. Similarly, we have also tested the performance of our proposed CFDRNN network against the popular LFPW and Helen test set. Figures 8 and 9 shows the CED curve for LFPW and Helen testing sets respectively.

TABLE 7: AUC and failure rate of face alignment methods on the 300-W Public and Private test set.

Test Set	Method	$\text{AUC}_{0.08}$	Failure (%)
Inter-ocular normalization			
300-W Public	ESR [12]	43.12	10.45
	SDM [16]	42.94	10.89
	CFSS [17]	49.87	5.08
	MDM [19]	52.12	4.21
	DAN [20]	55.33	1.16
	DAN-Menpo [20]	57.07	**0.58**
	CFDRNN	**57.11**	0.73
300-W Private	ESR [12]	32.35	17.00
	CFSS [17]	39.81	12.30
	MDM [19]	45.32	6.80
	DAN [20]	47.00	2.67
	DAN-Menpo [20]	50.84	1.83
	GAN [58]	53.64	2.50
	CFDRNN	**56.68**	**1.00**

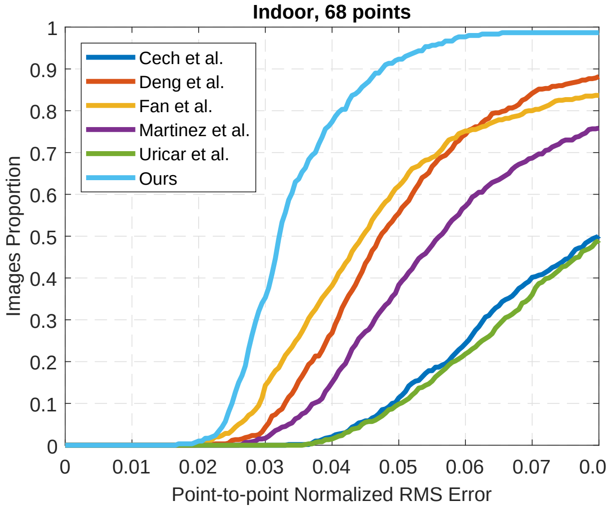

Fig. 5: Fitting results comparison with proposed method and state-of-the-art methods of 300-W challenge in 2015. The plots show the Cumulative Error Distribution (CED) curves with respect to the 68 landmarks points for the 'indoor' condition.

6.4 Testing Results

Figure 10 shows some of the estimated landmark localized images from 300-W indoor and outdoor test sets. Selected images are mainly with higher yaw angles or difficult facial expressions. These images show the effectiveness of our proposed CFDRNN framework over the so-called difficult face images, and rightfully justifies the results depicted in Fig.7.

7 Conclusion

In this paper, we introduced a robust coarse to fine deep recurrent neural network (CFDRNN) framework for facial landmark point localization. Contrary to the recently proposed state-of-the-art face alignment techniques, CFDRNN performs face alignment based on entire face images, without using any patch images or initial face mean shape, which makes it highly robust to large variations in head pose and its initialization. Using the entire face images instead of local patch images extracted around the landmarks, is possible, due to the use of novel heat-map images, which transmits the information about landmark locations

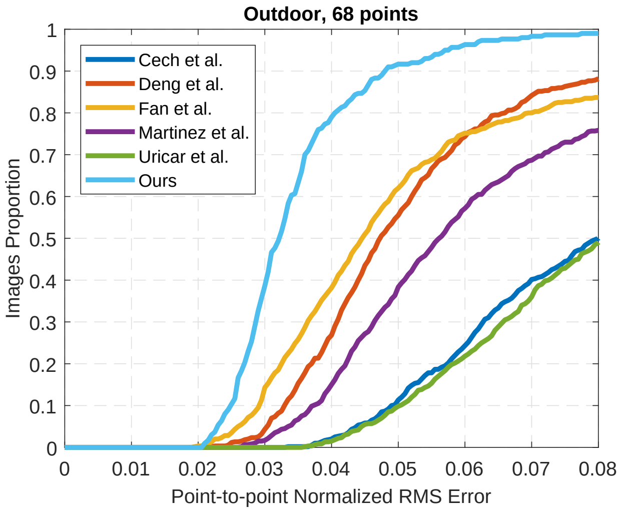

Fig. 6: Fitting results comparison with proposed method and state-of-the-art methods of 300-W challenge in 2015. The plots show the Cumulative Error Distribution (CED) curves with respect to the 68 landmarks points for the 'outdoor' condition.

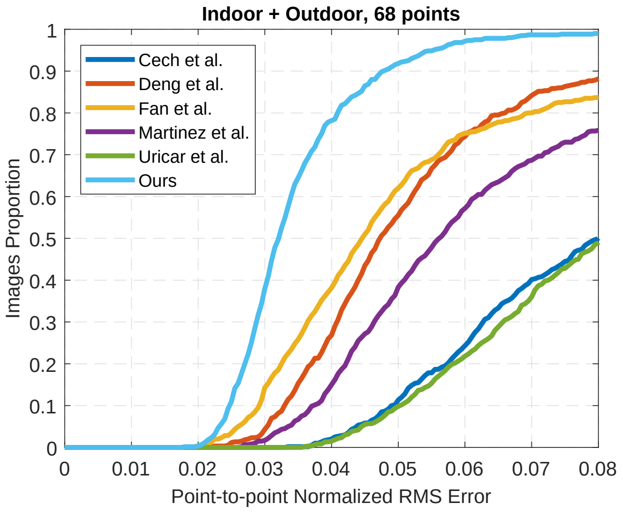

Fig. 7: Fitting results comparison with proposed method and state-of-the-art methods of 300-W challenge in 2015. The plots show the Cumulative Error Distribution (CED) curves with respect to the 68 landmarks points for the conditions (indoor+outdoor).

between different stages of CFDRNN. Exhaustive performance evaluation are also performed on several publicly available test datasets, especially on the challenging and 300-W private test set. Results show that our proposed CFDRNN framework reduces the failure rate by 45% and improves the $\mathrm{AUC}_{0.08}$ by 11.5% in comparison to the state-of-the-art techniques over the 300-W private test set. For challenging subset as well as improvement of 13% is achieved for the challenging subset. Hence, overall it can be said that our proposed CFDRNN framework can be considered as an useful contribution to the facial landmark localization research arena.

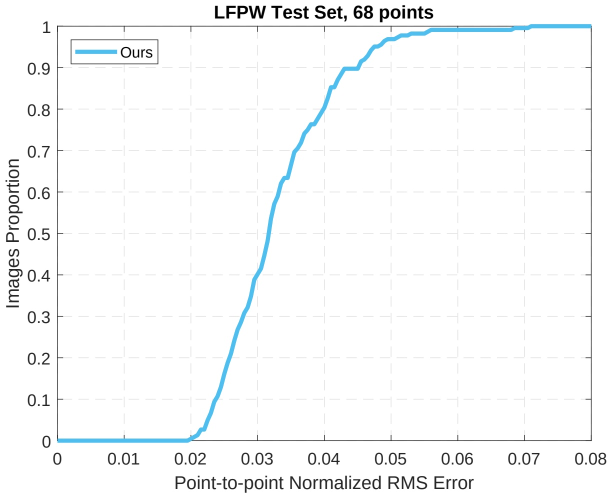

Fig. 8: Cumulative Error Distribution (CED) curves with respect to the landmarks for LFPW (68 points) test set.

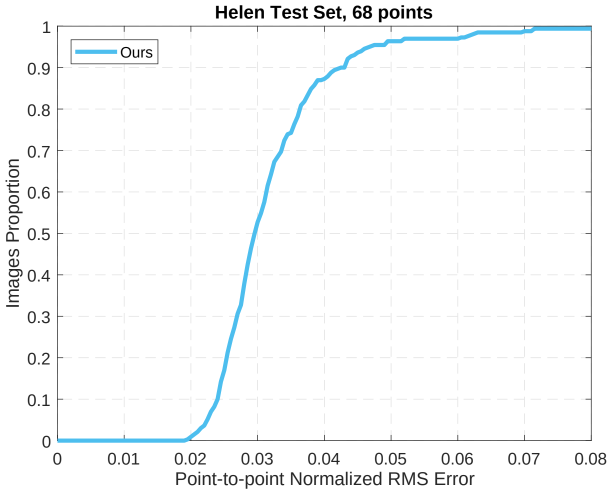

Fig. 9: Cumulative Error Distribution (CED) curves with respect to the landmarks for Helen (68 points) test set.

Acknowledgment

This work has been partially supported by COST Action 1206 "De-identification for privacy protection in multimedia content" and we gratefully acknowledge the support of NVIDIA® Corporation for providing the Titan X™ Pascal GPU for this research work.

References

[1] Z. Huang, X. Zhao, S. Shan, R. Wang, and X. Chen, "Coupling alignments with recognition for still-to-video face recognition," in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 3296–3303.

[2] C. Lu and X. Tang, "Surpassing human-level face verification performance on lfw with gaussian face," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI'15. AAAI Press, 2015, pp. 3811–3819. [Online]. Available: http://dl.acm.org/citation.cfm?id=2888116.2888245

[3] N. Kumar, P. Belhumeur, and S. Nayar, "Facetracer: A search engine for large collections of images with faces," in *Computer Vision – ECCV 2008*, D. Forsyth, P. Torr, and A. Zisserman, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 340–353.

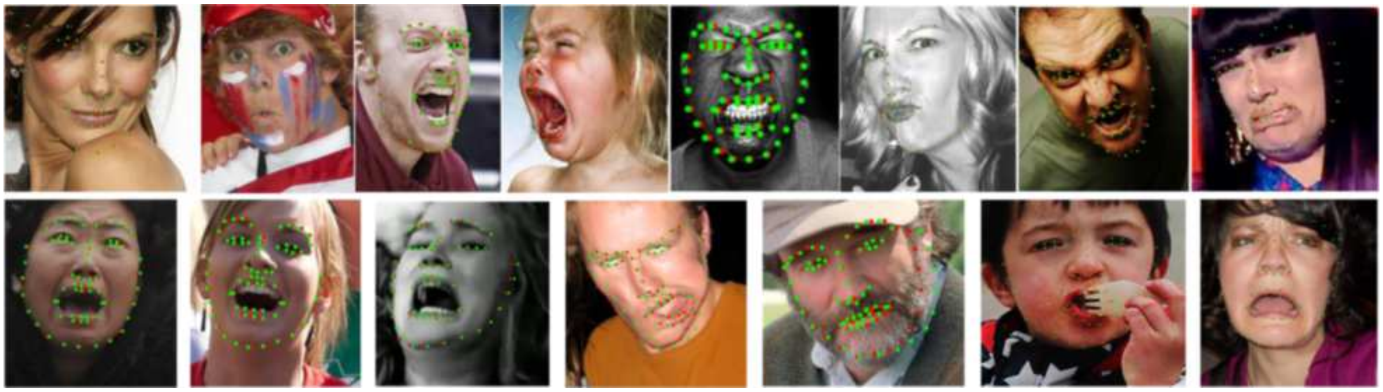

Fig. 10: Selected landmarked images from 300-W test set, where the red dots corresponds to ground truth landmark points and green dots signifies estimated landmarks by our proposed CFDRNN.

[4] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 3476–3483.

[5] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2017.

[6] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 5, pp. 918–930, May 2016.

[7] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, Jun 2001.

[8] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 2879–2886.

[9] 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 *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 1944–1951.

[10] G. Tzimiropoulos and M. Pantic, "Gauss-newton deformable part models for face alignment in-the-wild," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1851–1858.

[11] X. P. Burgos-Artizzu, P. Perona, and P. Dollr, "Robust face landmark estimation under occlusion," in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 1513–1520.

[12] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 2887–2894.

[13] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer, "Robust and accurate shape model fitting using random forest regression voting," in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 278–291.

[14] H. Yang and I. Patras, "Sieving regression forest votes for facial feature detection in the wild," in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 1936–1943.

[15] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine autoencoder networks (cfan) for real-time face alignment," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 1–16.

[16] X. Xiong and F. D. la Torre, "Supervised descent method and its applications to face alignment," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 532–539.

[17] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4998–5006.

[18] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1685–1692.

[19] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou, "Mnemonic descent method: A recurrent process applied for end-to-end face alignment," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 4177–4187.

[20] M. Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 2034–2043.

[21] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *2013 IEEE International Conference on Computer Vision Workshops*, Dec 2013, pp. 397–403.

[22] ——, "A semi-automatic methodology for facial landmark annotation," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2013, pp. 896–903.

[23] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: database and results," *Image and Vision Computing*, vol. 47, no. Supplement C, pp. 3 – 18, 2016, 300-W, the First Automatic Facial Landmark Detection in-the-Wild Challenge. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885616000147

[24] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen, "The menpo facial landmark localisation challenge: A step towards the solution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 2116–2125.

[25] X. Shao, J. Xing, J. Lv, C. Xiao, P. Liu, Y. Feng, and C. Cheng, "Unconstrained face alignment without face detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 2069–2077.

[26] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2930–2940, Dec 2013.

[27] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," *Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III*, pp. 679–692, 2012. [Online]. Available: https://doi.org/10.1007/978-3-642-33712-3_49

[28] D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, ser. CVPR '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 2879–2886. [Online]. Available: http://dl.acm.org/citation.cfm?id=2354409.2355119

[29] K. Messer, J. Kittler, M. Sadeghi, S. Marcel, C. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, J. Czyz, L. Vandendorpe, S. Srisuk, M. Petrou, W. Kurutach, A. Kadyrov, R. Paredes, B. Kepenekci, F. B. Tek, G. B. Akar, F. Deravi, and N. Mavity, "Face verification competition on the xm2vts database," *Audio- and Video-Based Biometric Person Authentication: 4th International Conference, AVBPA 2003 Guildford, UK, June 9–11, 2003 Proceedings*, pp. 964–974, 2003. [Online]. Available: https://doi.org/10.1007/3-540-44887-X_112

[30] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, Nov 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029666.37597.d3

[31] D. Cristinacce and T. Cootes, "Feature detection and tracking with

constrained local models," in *British Machine Vision Conference*, 2006, pp. 929–938.

[32] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 3444–3451.

[33] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1867–1874.

[34] M. Kowalski and J. Naruniec, "Face alignment using k-cluster regression forests with weighted splitting," *IEEE Signal Processing Letters*, vol. 23, no. 11, pp. 1567–1571, Nov 2016.

[35] D. Lee, H. Park, and C. D. Yoo, "Face alignment using cascade gaussian process regression trees," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4204–4212.

[36] O. Tuzel, T. K. Marks, and S. Tambe, "Robust face alignment using a mixture of invariant experts," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 825–841.

[37] A. Bulat and G. Tzimiropoulos, "Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge," in *Computer Vision – ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds. Cham: Springer International Publishing, 2016, pp. 616–624.

[38] H. Fan and E. Zhou, "Approaching human level facial landmark localization by deep learning," *Image Vision Comput.*, vol. 47, no. C, pp. 27–35, mar 2016. [Online]. Available: http://dx.doi.org/10.1016/j.imavis.2015.11.004

[39] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 57–72.

[40] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *2013 IEEE International Conference on Computer Vision Workshops*, Dec 2013, pp. 386–391.

[41] M. Pedersoli, R. Timofte, T. Tuytelaars, and L. V. Gool, "Using a deformation field model for localizing faces and facial points under weak supervision," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 3694–3701.

[42] X. Liu, "Generic face alignment using boosted appearance model," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[43] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–511–I–518 vol.1.

[44] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 2729–2736.

[45] P. Dollr, P. Welinder, and P. Perona, "Cascaded pose regression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 1078–1085.

[46] M. Dantone, J. Gall, G. Fanelli, and L. V. Gool, "Real-time facial feature detection using conditional regression forests," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 2578–2585.

[47] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 109–122.

[48] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[49] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Proceedings of the 9th International Conference on Neural Information Processing Systems*, ser. NIPS'96. Cambridge, MA, USA: MIT Press, 1996, pp. 155–161. [Online]. Available: http://dl.acm.org/citation.cfm?id=2998981.2999003

[50] B. Shi, X. Bai, W. Liu, and J. Wang, "Face alignment with deep regression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 1, pp. 183–194, Jan 2018.

[51] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 1302–1310.

[52] P. M. R. Martin Koestinger, Paul Wohlhart and H. Bischof, "Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization," in *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[53] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1891–1898.

[54] 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*, vol. PP, no. 99, pp. 1–1, 2017.

[55] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 3730–3738. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.425

[56] X. Chen, E. Zhou, Y. Mo, J. Liu, and Z. Cao, "Delving deep into coarse-to-fine framework for facial landmark localization," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 2088–2095.

[57] Z. He, J. Zhang, M. Kan, S. Shan, and X. Chen, "Robust fec-cnn: A high accuracy facial landmark detection system," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 2044–2050.

[58] Y. Chen, C. Shen, X. Wei, L. Liu, and J. Yang, "Adversarial learning of structure-aware fully convolutional networks for landmark localization," *CoRR*, vol. abs/1711.00253, 2017. [Online]. Available: http://arxiv.org/abs/1711.00253

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999134.2999257

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2016, pp. 770–778.

[61] Z. He, M. Kan, J. Zhang, X. Chen, and S. Shan, "A fully end-to-end cascaded cnn for facial landmark detection," in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, May 2017, pp. 200–207.

[62] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, May 2017, pp. 468–475.

[63] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," in *Proceeding of the ACM Int. Conf. on Multimedia*, 2015.