Disentangling 3D Pose in A Dendritic CNN for Unconstrained 2D Face Alignment

Amit Kumar Rama Chellappa
Department of Electrical and Computer Engineering, CFAR and UMIACS
University of Maryland-College Park,USA
akumar14@umiacs.umd.edu, rama@umiacs.umd.edu

Abstract

Heatmap regression has been used for landmark localization for quite a while now. Most of the methods use a very deep stack of bottleneck modules for heatmap classification stage, followed by heatmap regression to extract the keypoints. In this paper, we present a single dendritic CNN, termed as Pose Conditioned Dendritic Convolution Neural Network (PCD-CNN), where a classification network is followed by a second and modular classification network, trained in an end to end fashion to obtain accurate landmark points. Following a Bayesian formulation, we disentangle the 3D pose of a face image explicitly by conditioning the landmark estimation on pose, making it different from multi-tasking approaches. Extensive experimentation shows that conditioning on pose reduces the localization error by making it agnostic to face pose. The proposed model can be extended to yield variable number of landmark points and hence broadening its applicability to other datasets. Instead of increasing depth or width of the network, we train the CNN efficiently with Mask-Softmax Loss and hard sample mining to achieve upto 15% *reduction in error compared to state-of-the-art methods for extreme and medium pose face images from challenging datasets including AFLW, AFW, COFW and IBUG.*

1. Introduction

Face alignment or facial landmark estimation is the task of estimating keypoints such as eye-corners, mouth corners etc. on a face image. As shown in [5], accurate face alignment improves the performance of a face verification system, as well as other applications such as 3D face modelling, face animation etc. Currently, face alignment is dominated by regression-based approaches which yield a fixed number of points. Explicit Shape Regression (ESR) [13] and Supervised Descent Method (SDM) [48] have addressed the problem of face alignment for faces in medium pose. To achieve sub-pixel accuracy on such face images, coarse to fine approaches have also been proposed in the literature [31, 52, 54]. It is evident that such methods perform poorly on face images with extreme pose, expression and lighting mainly because they are dependent on bounding box and mean face shape intializations. On the other hand, Convolutional Neural Networks (CNNs) have achieved breakthroughs in many vision tasks including the task of keypoints estimation [35]. Lately, researchers have used heatmap regression extensively for the task of face alignment and pose estimation using an Encoder-Decoder architecture in the form of Convolution-Deconvolution Networks [14]. Most of the approaches in the literature perform heatmap classification followed by regression [6, 9–11]. In this paper, we propose the Pose Conditioned Dendritic Convolution Neural Network (PCD-CNN); which models the dendritic structure of facial landmarks using a single CNN (see Figure 1).

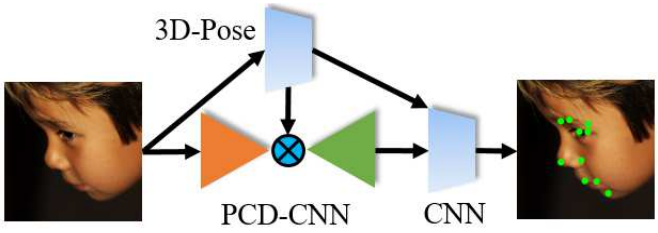


Figure 1: A bird's eye view of the proposed method. Dendritic CNN is explicitly conditioned on 3D pose. A generic CNN is used for auxiliary tasks such as fine-grained localization or occlusion detection.

Shape constraint: Methods such as ESR [13] and SDM [48] impose the shape constraint by jointly regressing over all the points. Such a shape constraint cannot be applied to a profile face as a consequence of extreme pose leading to a variable number of points. Tree structured part models (TSPM) [58] by Zhu et al. had two major limitations associated with it; namely pre-determined models and slower run-

time. With an intent to solve these, we propose a tree structure model in a single Dendritic CNN (PCD-CNN), which is able to capture the shape constraint in a deep learning framework.

Pose: Works such as Hyperface [36] and TCDCN [53] have used 3D pose in a multitask framework and demonstrated that learning pose and keypoints jointly using a deep network improves the performance of both tasks. However, in contrast to multi-tasking approaches, we condition the landmark estimates on the head pose, following a Bayesian formulation and demonstrate the effectiveness of the proposed approach through extensive experiments. We wish to point out that our primary goal is not to predict the head pose, instead, use 3D head pose to condition the landmark points. This makes our work different from multitask approaches.

Speed-vs-Accuracy: We observe that systems which process images at real time, such as [7, 25] have higher error rate as opposed to cascade methods which are accurate but slow. Researchers have proposed many different network architectures like Hourglass [35], Binarized CNN (based on hourglass) [10] in order to achieve accuracy in keypoints estimation. Although, such methods are fully convolutional , they suffer from slower run time as a result of cascaded deep bottleneck modules which perform a large number of FLOPs during test time. The proposed PCD-CNN works at the same scale as the input image and thus reduces the extrapolation errors. PCD-CNN is fully convolutional with fewer parameters and is capable of processing images almost at real time speed (20FPS). Limited generalizability as a consequence of smaller number of parameters is tackled by efficiently training the network using Mask-Softmax loss and difficult sample mining.

Generalizability: Methods for domain-limited face images have been developed, mostly following the cascade regression approach. [12, 46, 51] have been shown to work well for faces under extreme external object occlusion. On the other hand, [32, 38, 43–45, 54] achieved satisfactory results on the 300W [39] dataset which contains images in medium pose with almost no occlusion. [24, 30, 56] have demonstrated their effectiveness for extreme pose datasets with a limited number of fiducial points. However, they do not generalize very well to other datasets. We show that by a small increase in the number of parameters, PCD-CNN can be extended to most of the publicly available datasets including 300W, COFW, AFLW and AFW yielding variable number of points depending on the protocol.

Following the above discussion, the main contributions of this paper can be listed as:

- We propose the Pose Disentangled Dendritic CNN for unconstrained 2D face alignment, where the shape constraint is imposed by the dendritic structure of facial landmarks. The proposed method uses classification followed by classification approach as opposed to classification followed by regression. The second auxiliary network is modular and can be designed for fine grained localization or any other auxiliary tasks. Figure 2 shows the overall structure of PCD-CNN.
- The proposed method disentangles the head pose using a Bayesian framework and experimentally demonstrates that conditioning on 3D head pose improves the localization performance. The proposed method processes images at real-time speed producing accurate results.
- With a recursive extension, the proposed method can be extended to datasets with arbitrarily different number of points and different auxiliary tasks.
- As a by-product, the network outputs pose estimates of the face image where we achieve close to state-of-the-art result on pose estimation on the AFW dataset. In another experiment, the auxiliary classification network is trained for occlusion detection where we obtain state-of-the-art result for occlusion detection on COFW dataset.

2. Prior Work

We briefly review prior work in the area of keypoint localization under the following two categories: Deep Learning-based and Hand crafted features-based methods.

Parametric part-based models such as Active Appearance Models (AAMs) [16] and Constrained Local Models [17] are statistical methods which perform keypoint detection by maximizing the confidence of part locations in a given input image using ***handcrafted features*** such as SIFT and HOG. The tree structure part model (TSPM) proposed in [58] used deformable part-based model for simultaneous detection, pose estimation and landmark localization of face images modeling the face shape in a mixture of trees model. Later, [3] proposed learning a dictionary of probability response maps followed by linear regression in a Constrained Local Model (CLM) framework. Early cascade regression-based methods such as [4, 13, 40, 43, 45, 48, 54] also used hand crafted features such as SIFT to capture appearance of the face image. The major drawback of regression-based methods is their inability to learn models for unconstrained faces in extreme pose.

Deep learning-based methods have achieved breakthroughs in a variety of vision tasks including landmark localization. One of the earliest works was done in [31, 41] where a cascade of deep models was learnt for fiducial detection. 3DDFA [57] modeled the depth of the face image in a Z-buffer, after which a dense 3D face model was fitted to the image via CNNs. Pose Invariant Face Alignment (PIFA) [24] by Jourabloo et al. predicted the coefficients of 3D to 2D projection matrix via deep cascade regressors. [7] used 3D spatial transformer networks to capture 3D to 2D

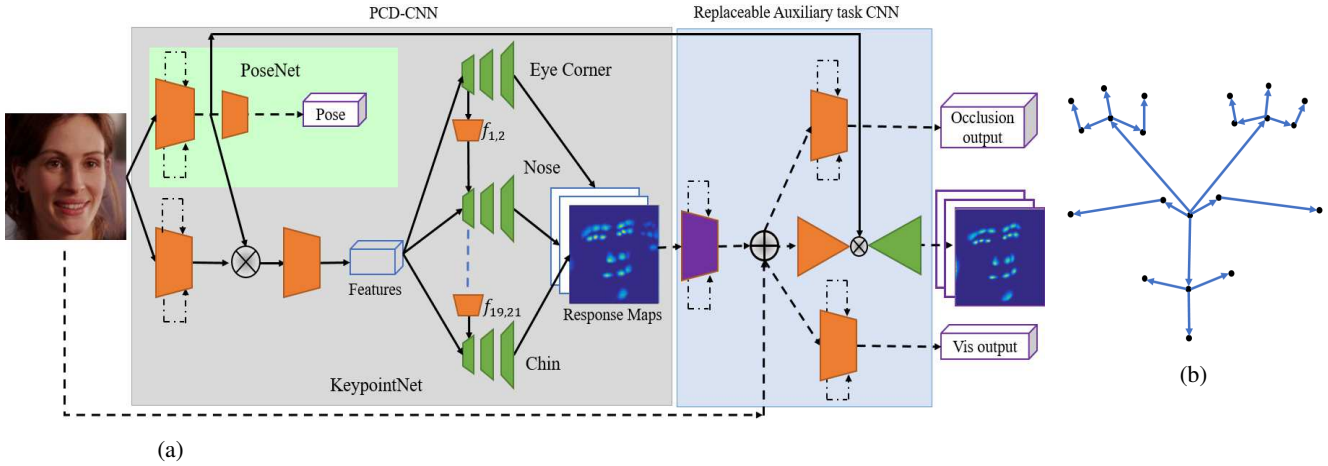


Figure 2: (a) Details of the proposed method. The dotted lines on top of convolution layers denote residual connections. Dendritic KeypointNet is conditioned on PoseNet. The network inside the grey box represents the proposed PCD-CNN, whereas the second network inside the blue box is modular and can be replaced for an auxiliary task. A conv-deconv network for finer localization is used alongside these auxiliary networks. (b) Proposed dendritic structure of facial landmark points for effective information sharing among landmark points. The nodes of the dendritic structure are the outputs of deconvolutions while the edges between nodes i and j are modeled by convolution functions f_{ij}. For the architecture of deconvolution network refer to Figure 3.

projection. [22, 27, 33] extended [24] by using CNNs to directly learn the dense 3D coordinates. The proposed method has a dendritic structure which looks at the global appearance of the image while the local interactions are captured by pose conditioned convolutions. PCD-CNN does not assume that all the keypoints are visible and the interactions between keypoints are learned. PCD-CNN is entirely based on 2D images, which captures the 3D information by conditioning on 3D head pose.

Formulating keypoint estimation as the per-pixel labeling task, Hourglass networks [35] and Structured feature learning [15] were proposed. Hourglass networks use a stack of 8 very deep hourglass modules and hence, even though based entirely on convolution can process only 8-10 frames per second. [15] implemented message passing between keypoints, however was able to process images at lower resolution due to large number of parameters. PCD-CNN models the dendritic structure in branched deconvolution networks where each network is implemented in Squeezenet [21] fashion and hence has fewer parameters, contributing to real-time operation at full image scale.

In the next few sections, we describe Pose Conditioned Dendritic-CNN in detail where we discuss the different concepts introduced, and then present ablative studies to arrive at the desired architecture.

3. Pose Conditioned Dendritic CNN

The task of keypoint detection is to estimate the 2D coordinates of, say N landmark points, given a face image. Observing the effectiveness of deep networks for a variety of vision tasks, we present a single end-to-end trainable deep neural network for landmark localization.

Conditioning on 3D pose: Keypoints are susceptible to variations in external factors such as emotion, occlusion and intrinsic face shape. On the other hand, 3D pose is fairly stable to them and can be estimated directly from 2D image [30]. Reasonably accurate 2D keypoint coordinates can be also inferred given 3D pose and a generic 3D model of a human face. However, the converse problem of estimating 3D pose from 2D keypoints is ill posed. Therefore, we make use of the probabilistic formulation over the variables including the image $\boldsymbol{I} \in \mathbb{R}^{w \times h \times 3}$ of height h and width w, $3D$ head pose denoted by $\boldsymbol{P} \in \mathbb{R}^3$, $2D$ keypoints $\boldsymbol{C} \in \mathbb{R}^{N \times 2}$, where N is the number of keypoints. Following the natural hierarchy between the two tasks, the joint and the conditional probabilities can be written as:

$$p(\boldsymbol{C}, \boldsymbol{P}, \boldsymbol{I}) = p(\boldsymbol{C}|\boldsymbol{P}, \boldsymbol{I}) p(\boldsymbol{P}|\boldsymbol{I}) p(\boldsymbol{I}) \tag{1}$$

$$\begin{aligned} p(\boldsymbol{C}, \boldsymbol{P}|\boldsymbol{I}) &= \frac{p(\boldsymbol{C}, \boldsymbol{P}, \boldsymbol{I})}{p(\boldsymbol{I})} \\ &= \underbrace{p(\boldsymbol{P}|\boldsymbol{I})}_{\text{CNN}} . \underbrace{p(\boldsymbol{C}|\boldsymbol{P}, \boldsymbol{I})}_{\text{PCD-CNN}} \end{aligned} \tag{2}$$

We implement the first factor with an image-based CNN learned to predict the 3D pose of the face image. The second factor is implemented through a ConvNet and multiple DeconvNets arranged in a dendritic structure. The convolution network maps the image to lower dimension, after which the outputs of several deconvolution networks are stacked to form the keypoint-heatmap. The models are tied

together by element-wise product (as (1) and (2)) to condition the measurement of 2D coordinates on 3D pose. We choose element-wise product as the operation to condition on the head pose as keypoint heatmaps can be interpreted as probability distribution over the keypoints. The visibility of each keypoint is learnt implicitly as the invisible points are labeled as background.

Multi-tasking-vs-Conditioning: In a multi-tasking method such as [30], several tasks are learnt synergetically and backpropagation impacts all the tasks. On the other hand, in the proposed PCD-CNN, the error gradients back-propagated from keypoint network affect both, keypoint network and pose network; however, the pose network affects the keypoint network only during the forward pass. In other words, multi-tasking approaches try to model the joint distribution $p(\boldsymbol{C}, \boldsymbol{P}|\boldsymbol{I})$, whereas the proposed approach explicitly models the decomposed form $p(\boldsymbol{P}|\boldsymbol{I})p(\boldsymbol{C}|\boldsymbol{P}, \boldsymbol{I})$ by learning the individual factors.

Proposed Pose Conditioned Dendritic CNN : To capture the structural relationship between different keypoints, we propose the dendritic structure of facial landmarks as shown in figure 2b where the nose tip is assumed to be the root node. Such a structure is feasible even in faces with extreme pose. Following this, the keypoint network is modeled with a single CNN in a tree structure composed of convolution and deconvolution layers. The pairwise relationships between different keypoints are modeled via specialized functions, $f_{i,j}$, which are implemented through convolutions and are analogous to the spring weights in the spring-weight model of Deformable Part Models [18]. A low confidence of a particular keypoint is reinforced when the response of $f_{i,j}$ corresponding to the adjacent node is added. With experimental justifications we show that such a deformable tree model outperforms the recently published works [7, 25, 27, 33] which use 3D models and 3D spatial transformer networks to supplement keypoint detection models. Figure 2 shows the overall architecture of the proposed PCD-CNN and the proposed dendritic structure of the facial landmarks.

Instead of going deeper or wider [10, 35] with deep networks, we base our work on the Squeezenet-11 [21] architecture, attributing to its capability to maintain performance with fewer parameters. We use two Squeezenet-11 networks; one for pose and other for keypoints, named as -PoseNet and KeypointNet respectively, as shown in Fig 2a. Convolutions are performed on the $pool_8$ activation maps of the PoseNet, the response of which is then multiplied element-wise to the response maps of $pool_8$ layers of the KeypointNet. Each convolution layer is followed by ReLU non-linearity and batch normalization. In table 1a, we show that keypoint localization error reduces when conditioned on 3D head pose.

The design of deconvolution network is non-trivial. To maintain the same property as of SqueezeNet, we first upsample the feature maps using parametrized strided convolutions and then squeeze the output features maps using 1x1 convolutions. We call this network as Squeezenet-DeconvNet. Figure 3 shows the detailed architecture of the Squeezenet-DeconvNet. Since, each keypoint in the proposed network is modeled by a separate Squeezenet-DeconvNet, it alleviates the need for large number of deconvolution parameters (256 and 512 3×3 in Hourglass networks). In fact, in the practical version of PCD-CNN, there are only 32 and 16 deconvolution filters which results in the design of networks, which are small enough to fit in a single GPU. The design of networks with fewer filters is motivated by real-time processing consideration. With experiments we show that disentangling the pose by conditioning on it, reinforces the learning of the proposed PCD-CNN with fewer parameters (Table 1a).

In order to obtain fine grained localization results, we concatenate to the input data, a learned function of the predicted probabilities (represented as purple box in Figure 2a) and pass them through the second Squeezenet based conv-deconv network. This function is modeled by a residual unit with 1×1 and 3×3 filters, which are learned end-to-end with the second classification network (while keeping the weights PCD-CNN frozen). For experimental purposes, we replace the second conv-deconv by another regression network designed along the lines of GoogleNet [42]. Table 1b shows a comparison between two stage classification approach versus classifcation followed by regression

Method	**Normalised Error**
Without pose conditioning	3.45
With pose conditioning	2.85

(a)

Method	**Normalised Error**
Classification+Regression	3.93
Classification+Classification	3.09

(b)

Method	**Normalised Error**
Softmax	4.56
Using Mask-Softmax	2.85

(c)

Table 1: Root mean square error normalized by bounding box size, calculated on the AFLW validation set following the PIFA protocol. (a) With and without conditioning on pose. (b) Comparison showing that PCD-CNN when followed by another classification stage results in lower localization error compared to classification followed by regression. Note that conditioning on pose is not used in both the cases above for fair comparison. (c) Comparison indicating the effect of using Mask-softmax over Softmax

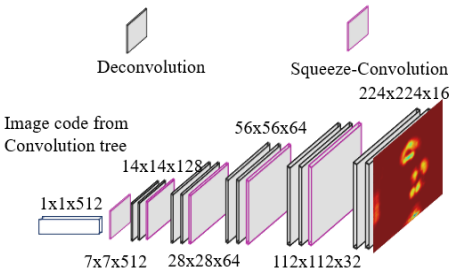


Figure 3: Detailed description of a single Squeezenet-DeconvNet network. Note the fewer number of deconvolution filters. Each deconvolution network is identical to the one shown above.

approaches [1].

One of the goals of this work is to generalize the facial landmark detection to other datasets in order to broaden its applicability. A trivial extension would be to increase the number of deconvolution branches, which however is infeasible due to limited GPU memory. However, PCD-CNN can be extended to yield more landmark points arranged in different configurations. In figure 4 we show the proposed tree structures for COFW and 300W datasets with 29 and 68 landmark points respectively. Keeping the basic **Dendritic Structure of Parts** intact, first the number of output response maps in the last deconvolution layer are increased and then network slicing is performed to produce the desired number of keypoints. For instance, the output of the deconvolution network for eye-center is sliced to produce four outputs as required by the 300W dataset. Depending on the dataset, the second network can be replaced to perform auxiliary tasks resulting in a modular architecture; for instance in the case of COFW dataset we replace the second conv-deconv network with another Squeezenet network to detect occlusion. We direct the readers to the supplementary material for more details on network surgery and a magnified view of figures 2b and 4.

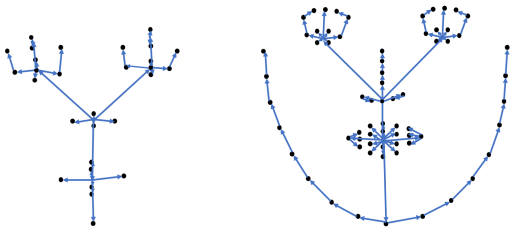


Figure 4: The proposed extension of the dendritic structure from Figure 2 generalizing to other datasets (COFW and 300W) each with different number of points.

Each branch of PCD-CNN is designed according to the proposed Squeezenet-Deconv networks shown in Figure 3. Due to fewer parameters in the Squeezent-Deconv, we hypothesize limited generalization capacity of the deconvolution network. By means of experiments, we show that effective training methods such as Mask-Softmax and Hard sample mining improves the performance of PCD-CNN by a large margin as a result of better generalization capacity.

Mask-Softmax Loss: To train the network, the localization of fiducial keypoints is formulated as a classification problem. The label for an input image of size $h \times w \times 3$ is a label tensor of same size as the image with $N+1$ channels, where N is the number of keypoints. The first N channels represent the location of each keypoint whereas the last channel represents the background. Each pixel is assigned a class label with invisible points being labeled as background. The objective is to minimize the following loss function:

$$L_0(\boldsymbol{p}, \boldsymbol{g}) = \sum_{i=1}^{h}\sum_{j=1}^{w} m(i,j) \sum_{k=1}^{N+1} g_k(i,j) log\left(\frac{e^{p_k(i,j)}}{\sum_l e^{p_l(i,j)}}\right) \tag{3}$$

where $k \in \{1, 2 \dots N\}$ is the class index and $g_k(i,j)$ represents the ground truth at location (i,j). $p_l(i,j)$ is the score obtained for location (i,j) after forward pass through the network. Since the number of negative examples is orders of magnitudes larger than the positives, we design a strategic mask $m(i,j)$ which selects all the positive pixel samples, and keeps only 50% of the 4-neighborhood pixels and 0.025% of the negative background samples by random selection. During backward pass, the gradients are weighed accordingly. We experimentally show the effect of using Mask-Softmax Loss by training two separate PCD-CNN; with and without the Mask-Softmax Loss; trained under identical training policies(Table 1c) .

Hard Sample Mining: [28] by Kabkab et al. showed that effective sampling of data improves the classification performance of the network. Following [28], we use an offline hard sample mining procedure to train the proposed PCD-CNN. The histogram of error on the training data is plotted after the network is trained for 10 epochs by random sampling (refer supplementary material). We denote the mode of the distribution as C, and categorize all the training samples producing errors larger than C as hard samples. Next we retrain the proposed PCD-CNN with hard and easy samples, sampled at the respective proportion. This effectively results in retraining the network by reusing the hard samples. Table 2a shows that such hard sample mining improves the performance of PCD-CNN (with fewer parameters) by a large margin.

In the next set of experiments, we train PCD-CNN by increasing the number of deconvolution filters to 128 and 64 in each deconvolution network. We follow the same strategy of Mask-Softmax and hard sample mining to train this network. Unsurprisingly, we see an improvement in performance for the task of keypoint localization (Table 2b), although, increasing the number of deconvolution filters leads

Method	Normalised Error
Without Hard Mining	2.85
With Hard Mining	2.49

(a)

Method	Normalised Error
Less Filters+Hard Mining	2.49
More Filters+Hard Mining	2.40

(b)

Table 2: Root mean square error normalized by bounding box calculated on the AFLW validation set following PIFA protocol. (a) depicts the effect of offline hard sample mining. (b) shows the effect of offline hard-mining and quadrupling the number of deconvolution filters.

to slower run time of 11FPS as opposed to 20FPS.

4. Experiments

We select four different datasets with different characteristics to train and evaluate the proposed two stage PCD-CNN.

AFLW [29]and **AFW** [58] are two *difficult* datatsets which comprises of images in extreme pose, expression and occlusion. AFLW consists of $24,386$ in-the-wild faces (obtained from *Flickr*) with head pose ranging from $0°$ to $120°$ for yaw and upto $90°$ for pitch and roll. AFLW provides at most 21 points for each face. It excludes coordinates for invisible landmarks and in our method such invisible points are labelled as background. For AFLW we follow the PIFA protocol; i.e. the test set is divided into three groups corresponding to three pose groups with equal number of images in each group.

AFW which is a popular benchmark for the evaluation of face alignment algorithms, consisting of 468 in-the-wild faces (also obtained from Flickr) with yaw up to $90°$. The images are diverse in terms of pose, expression and illumination and was considered the most difficult publicly available dataset, until AFLW. The number of visible points varies depending on the pose and occlusion with a maximum of 6 points per face image. We use AFW only for evaluation purposes.

A *medium* pose dataset from the popular **300W** face alignment competition [39]. The dataset consists of re-annotated five existing datasets with 68 landmarks: iBug, LFPW, AFW, HELEN and XM2VTS. We follow the work [54] to use $3,148$ images for training and 689 images for testing. The testing dataset is split into three parts: common subset (554 images), challenging subset (135 images) and the full set (689 images).

Another dataset showing extreme cases of external and internal object *occlusion*; **COFW** [47]. COFW is the most challenging dataset that is designed to depict faces in real-world conditions with partial occlusions [12]. The face images show large variations in shape and occlusions due to differences in pose, expression, hairstyle, use of accessories or interactions with other objects. All 1,007 images were annotated using the same 29 landmarks as in the LFPW dataset, with their individual visibilities. The training set includes 845 LFPW faces + 500 COFW faces, that is 1,345 images in total. The remaining 507 COFW faces are used for testing.

Evaluation Metric: Following most previous works, we obtain the error for each test sample via averaging normalized errors for all annotated landmarks. We illustrate our results with mean error over all samples, or via Cumulative Error Distribution (CED) curve. For AFLW and AFW, the obtained error is normalized by the ground truth bounding box size over all visible points whereas for 300W and COFW, error is normalized by the inter-occular distance. Wherever applicable NME stands for Normalized Mean Error.

Training: The PCD-CNN was first trained using the AFLW training set which was augmented by random cropping, flipping and rotation. The network was trained for 10 epochs where the learning rate starting from 0.01 was dropped every 3 epochs. Keeping the weights of PCD-CNN fixed, the auxiliary network for fine grained classifcation was trained for another 10 epochs using the hard mining strategy explained in section 3. PoseNet was kept frozen while training the network for COFW and 300W datasets. All the experiments including training and testing were performed using the Caffe [23] framework and Nvidia TITAN-X GPUs and p6000 GPUs. Being a non-iterative and single shot keypoint prediction method, our method is fast and can process **20** frames per second on 1 GPU only in batch mode. (Refer to supplementary material for more training details)

4.1. Results

Table 3a compares the performance of proposed method over other existing methods on AFLW-PIFA and AFW dataset. Table 3b compares the performance on AFLW-PIFA with respect to each pose group. Tables 4a and 4b compares the mean normalized error on the 300W and COFW datasets respectively. It is clear from the tables that while the proposed PCD-CNN performs comparable to previous state-of-the-art method [10], the two stage PCD-CNN outperforms the state-of-the-art methods on all three datasets: AFLW, AFW and COFW by large margins. It is not surprising that increasing the number of deconvolution filters improves the performance on all the datasets. Figures 5a, 5b and 5c show the cumulative error distribution for landmark localization in AFLW, AFW and COFW test sets. From the plots, we observe that the proposed PCD-CNN leads to a significant increase in the percentage of images

	AFLW	AFW
Method	**NME**	**NME**
TSPM [58]	-	11.09
CDM [2]	12.44	9.13
RCPR [12]	7.85	-
ESR [13]	8.24	-
PIFA [24]	6.8	9.42
3DDFA [57]	5.32	-
LPFA-3D [26]	4.72	7.43
EMRT [55]	4.01	3.55
Hyperface [36]	4.26	-
Rec Enc-Dec [1]	>6	-
PIFAS [27]	4.45	6.27
FRTFA [7]	4.23	-
CALE [11]	2.63	-
KEPLER [30]	2.98	3.01
Binary-CNN [10]	2.85	-
PCD-CNN(Fast) Pre Test Aug	**2.85**	**2.80**
PCD-CNN(Fast) Post Test Aug	**2.81**	**2.66**
PCD-CNN(C+C) Pre Test Aug	**2.49**	**2.52**
PCD-CNN(C+C+more filters)	**2.40**	**2.47**
PCD-CNN(C+C) Post Test Aug (Best)	**2.40**	**2.36**

(a)

Method	[0,30]	[30,60]	[60,90]	Mean
HyperFace [36]	3.93	4.14	4.71	4.26
AIO [37]	2.84	2.94	3.09	2.96
Binary-CNN [10]	2.77	2.86	2.90	2.85
PCD-CNN(C+C)	**2.33**	**2.60**	**2.64**	**2.49**

(b)

Table 3: Comparison with previous methods on (a) AFLW-PIFA test set and AFW test set. (b) AFLW-PIFA categorized by absolute yaw angles. In (a) C+C stands for classification+classification. For AFLW, numbers for other methods are taken from respective papers following the PIFA protocol. For AFW, the numbers are taken from respective published works following the protocol of [58]. The numbers represent the normalized mean error.

Method	Common	Challenge	Full
RCPR [12]	6.18	17.26	8.35
SDM [48]	5.57	15.40	7.52
ESR [13]	5.28	17.00	7.58
CFAN [52]	5.50	16.78	7.69
LBF [38]	4.95	11.98	6.32
CFSS [54]	4.73	9.98	5.76
TCDCN [53]	4.80	8.60	5.54
DDN [50]	-	-	5.59
MDM [43]	4.83	10.14	5.88
TSR [34]	4.36	**7.56**	4.99
PCD-CNN	**3.67**	7.62	**4.44**

(a)

Method	NME	Failure Rate
RCPR [12]	8.5	20%
OFA [51]	6.46	-
HPM [19]	8.48	6.99%
ERCLM [8]	6.49	6.3%
RPP [49]	7.52	16.2%
Human [12]	5.6	0%
PCD-CNN Pre Test Aug	**6.02**	**4.53%**
PCD-CNN Post Test Aug	**5.77**	**3.73%**

(b)

Table 4: Comparison of the proposed method with other state-of-the-art methods on (a) 300W dataset (b) COFW testset. The NMEs for comparison on 300W dataset are taken from the Table 3 of [34].

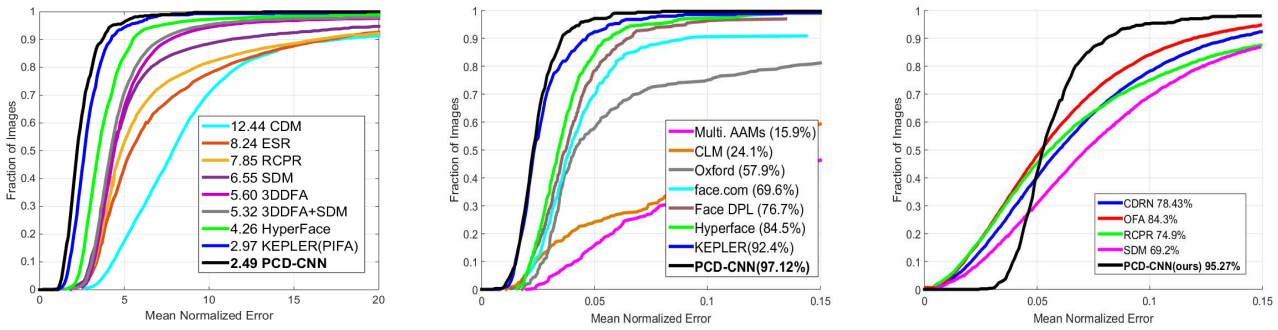
with mean normalized error less than 5%. On AFW, fraction of images having an error of less than 15° for pose estimation is 87.22% compared to 82% in the recent work [20]. On COFW dataset, the NME reduces to 6.02 (close human performance of 5.6) bringing down the failure rate to 4.53%. PCD-CNN achieves a higher recall of 44.7% at the precision of 80% as opposed to RCPR's [12] 38.2%. (refer to the supplementary material for more results.)

Improvement in localization by augmentation during testing : For a fair evaluation, we compare with the previous state-of-the-art methods with and without augmentation during testing. In the next set of experiments along with the test image, we also pass the flipped version of it and the final output is taken as the mean of the two outputs. With experimentation we observe that data augmentation while testing also improves the localization performance. While on AFLW-PIFA the error rate of 2.40 is achieved, the effect of test set augmentation is more prominent in AFW dataset, where the error rate of 2.36 is achieved. Similarly, on 300W (challenging) error rate drops to 7.17 from 7.62 as a result of test set augmentation. On COFW, error rate and failure rate of 5.77 and 3.73% respectively are achieved as the best results.

Figure 6 shows some of the difficult images and the predicted visible keypoints on the four datasets. We also achieve state of the art results on the performance of auxiliary tasks, such as pose estimation on AFW and occlusion prediction on COFW dataset.

5. Conclusion and Future Work

In this paper, we present a dendritic CNN which processes images at full scale looking at the images globally and capturing local interactions through convolutions. The proposed PCD-CNN is able to precisely localize landmark



(a) (b) (c)

Figure 5: Cumulative error distribution curves for landmark localization on AFLW, AFW and COFW dataset respectively. (a) Numbers in the legend represents mean error normalized by the face size. (b) Numbers in the legend are the fraction of testing faces that have average normalized error below 5%. (c) The numbers in the legend are the fraction of testing faces that have average normalized error below 10%.

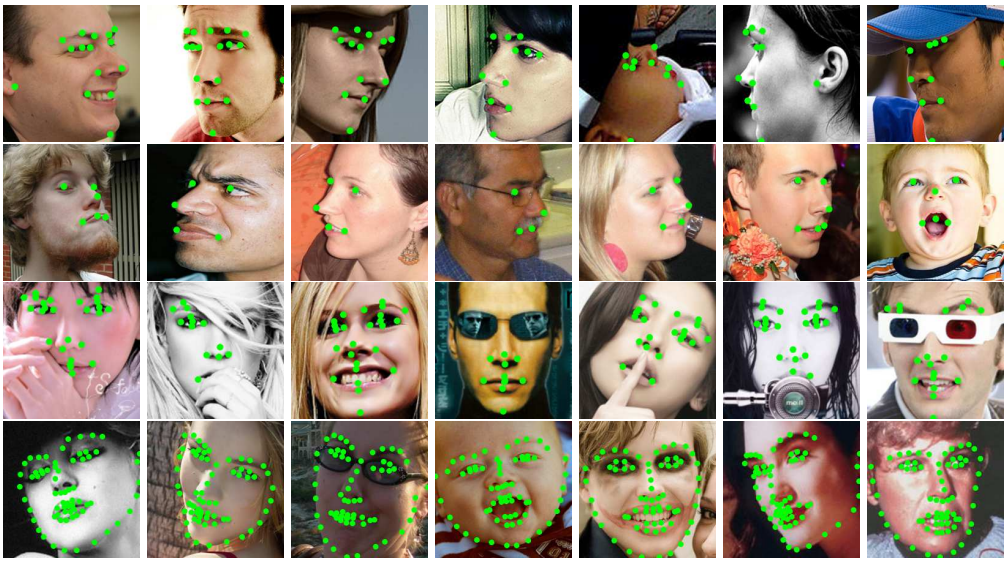


Figure 6: Qualitative results generated from the proposed method. The green dots represent the predicted points. Each row shows some of the difficult samples from AFLW, AFW, COFW, and 300W respectively with all the visible predicted points.

points on unconstrained faces without using any 3D morphable models. We also demonstrate that disentangling pose by conditioning on it can influence the localization of landmark points by reducing the mean pixel error by a large margin. Due to effective design choices made, the proposed model is not limited to yield a fixed number of points and can be extended to other datasets with different protocols. With the help of ablative studies, impact of effective training of the convolutional network by using sampling strategies such as Mask-Softmax and hard instance sampling is shown. Using smaller and fewer convolution filters, the proposed network is able to process images close to real-time and can be deployed in a real life scenario. The proposed method can be easily extended to 3D dense face alignment and other tasks, which we plan to pursue in the future.

6. Acknowledgment

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. We also thank our colleagues for all the discussion sessions.

References

[1] A recurrent autoencoder-decoder for sequential face alignment. http://arxiv.org/abs/1608.05477. Accessed: 2016-08-16. 5, 7

[2] *Pose-free Facial Landmark Fitting via Optimized Part Mixtures and Cascaded Deformable Shape Model*, 2013. 7

[3] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, CVPR '13, pages 3444–3451, Washington, DC, USA, 2013. IEEE Computer Society. 2

[4] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *CVPR 2014*, 2014. 2

[5] A. Bansal, C. Castillo, R. Ranjan, and R. Chellappa. The do's and don'ts for cnn-based face verification. *arXiv preprint arXiv:1705.07426*, 2017. 1

[6] V. Belagiannis and A. Zisserman. Recurrent human pose estimation. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 468–475, May 2017. 1

[7] C. Bhagavatula, C. Zhu, K. Luu, and M. Savvides. Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 4, 7

[8] V. N. Boddeti, M. Roh, J. Shin, T. Oguri, and T. Kanade. Face alignment robust to pose, expressions and occlusions. *CoRR*, abs/1707.05938, 2017. 7

[9] A. Bulat and G. Tzimiropoulos. *Human Pose Estimation via Convolutional Part Heatmap Regression*, pages 717–732. Springer International Publishing, Cham, 2016. 1

[10] A. Bulat and G. Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2, 4, 6, 7

[11] A. Bulat and Y. Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. In E. R. H. Richard C. Wilson and W. A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 86.1–86.12. BMVA Press, September 2016. 1, 7

[12] X. P. Burgos-Artizzu, P. Perona, and P. Dollar. Robust face landmark estimation under occlusion. *ICCV*, 0:1513–1520, 2013. 2, 6, 7

[13] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *IJCV*, 107(2):177–190, 2014. 1, 2, 7

[14] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR*, abs/1412.7062, 2014. 1

[15] X. Chu, W. Ouyang, H. Li, and X. Wang. Structured feature learning for pose estimation. In *CVPR*, 2016. 3

[16] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE T-PAMI*, 23(6):681–685, Jun 2001. 2

[17] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *Pattern Recogn.*, 41(10):3054–3067, Oct. 2008. 2

[18] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, Sept. 2010. 4

[19] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1899–1906, June 2014. 7

[20] G.-S. Hsu, K.-H. Chang, and S.-C. Huang. Regressive tree structured model for facial landmark localization. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 7

[21] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. *arXiv:1602.07360*, 2016. 3, 4

[22] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. *International Conference on Computer Vision*, 2017. 3

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

[24] A. Jourabloo and X. Liu. Pose-invariant 3d face alignment. In *ICCV*, Santiago, Chile, December 2015. 2, 3, 7

[25] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proc. IEEE Computer Vision and Pattern Recogntion*, Las Vegas, NV, June 2016. 2, 4

[26] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *CVPR*, Las Vegas, USA, June 2016. 7

[27] A. Jourabloo, X. Liu, M. Ye, and L. Ren. Pose-invariant face alignment with a single cnn. In *In Proceeding of International Conference on Computer Vision*, Venice, Italy, October 2017. 3, 4, 7

[28] M. Kabkab, A. Alavi, and R. Chellappa. Dcnns on a diet: Sampling strategies for reducing the training set size. *CoRR*, abs/1606.04232, 2016. 5

[29] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011. 6

[30] A. Kumar, A. Alavi, and R. Chellappa. Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 258–265, May 2017. 2, 3, 4, 7

[31] A. Kumar, R. Ranjan, V. M. Patel, and R. Chellappa. Face alignment by local deep descriptor regression. *CoRR*, abs/1601.07950, 2016. 1, 2

[32] D. Lee, H. Park, and C. D. Yoo. Face alignment using cascade gaussian process regression trees. In *CVPR*, pages 4204–4212, June 2015. 2

[33] Y. Liu, A. Jourabloo, W. Ren, and X. Liu. Dense face alignment. In *In Proceeding of International Conference on Computer Vision Workshops*, Venice, Italy, October 2017. 3, 4

[34] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 7

[35] A. Newell, K. Yang, and J. Deng. *Stacked Hourglass Networks for Human Pose Estimation*, pages 483–499. Springer International Publishing, Cham, 2016. 1, 2, 3, 4

[36] 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. *CoRR*, abs/1603.01249, 2016. 2, 7

[37] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 17–24, May 2017. 7

[38] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 FPS via regressing local binary features. In *CVPR*, pages 1685–1692, 2014. 2, 7

[39] 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*, pages 397–403, Dec 2013. 2, 6

[40] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, pages 3476–3483, June 2013. 2

[41] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, CVPR '13, pages 3476–3483, Washington, DC, USA, 2013. IEEE Computer Society. 2

[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. 4

[43] 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 *CVPR*, Las Vegas, USA, June 2016. 2, 7

[44] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[45] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *CVPR*, pages 1851–1858, June 2014. 2

[46] Y. Wu, C. Gou, and Q. Ji. Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[47] Y. Wu and Q. Ji. Robust facial landmark detection under significant head poses and occlusion. In *ICCV*, pages 3658–3666, Dec 2015. 6

[48] Xuehan-Xiong and F. De la Torre. Supervised descent method and its application to face alignment. In *CVPR*, 2013. 1, 2, 7

[49] H. Yang, X. He, X. Jia, and I. Patras. Robust face alignment under occlusion via regional predictive power estimation. *IEEE Transactions on Image Processing*, 24(8):2393–2403, Aug 2015. 7

[50] X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. *CoRR*, abs/1605.01014, 2016. 7

[51] J. Zhang, M. Kan, S. Shan, and X. Chen. Occlusion-free face alignment: Deep regression networks coupled with de-corrupt autoencoders. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 7

[52] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *ECCV*, volume 8690 of *Lecture Notes in Computer Science*, pages 1–16. Springer International Publishing, 2014. 1, 7

[53] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, pages 94–108, 2014. 2, 7

[54] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. June 2015. 1, 2, 6, 7

[55] S. Zhu, C. Li, C. C. Loy, and X. Tang. Towards arbitrary-view face alignment by recommendation trees. *CoRR*, abs/1511.06627, 2015. 7

[56] S. Zhu, C. Li, C.-C. Loy, and X. Tang. Unconstrained face alignment via cascaded compositional learning. In *CVPR*, June 2016. 2

[57] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. *CoRR*, abs/1511.07212, 2015. 2, 7

[58] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886, June 2012. 1, 2, 6, 7