HyperFace: A Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition

Rajeev Ranjan, *Member, IEEE,* Vishal M. Patel, *Senior Member, IEEE,* and Rama Chellappa, *Fellow, IEEE*

Abstract—We present an algorithm for simultaneous face detection, landmarks localization, pose estimation and gender recognition using deep convolutional neural networks (CNN). The proposed method called, HyperFace, fuses the intermediate layers of a deep CNN using a separate CNN followed by a multi-task learning algorithm that operates on the fused features. It exploits the synergy among the tasks which boosts up their individual performances. Additionally, we propose two variants of HyperFace: (1) HyperFace-ResNet that builds on the ResNet-101 model and achieves state-of-the-art performance, and (2) Fast-HyperFace that uses a high recall fast face detector for generating region proposals to improve the speed of the algorithm. Extensive experiments show that the proposed models are able to capture both global and local information in faces and performs significantly better than many competitive algorithms for each of these four tasks.

Index Terms—Face Detection, Landmarks Localization, Head Pose Estimation, Gender Recognition, Deep Convolutional Neural Networks, Multi-task Learning.

1 Introduction

Detection and analysis of faces is a challenging problem in computer vision, and has been actively researched for applications such as face verification, face tracking, person identification, etc. Although recent methods based on deep Convolutional Neural Networks (CNN) have achieved remarkable results for the face detection task [12], [42], [60], it is still difficult to obtain facial landmark locations, head pose estimates and gender information from face images containing extreme poses, illumination and resolution variations. The tasks of face detection, landmark localization, pose estimation and gender classification have generally been solved as separate problems. Recently, it has been shown that learning correlated tasks simultaneously can boost the performance of individual tasks [71] , [70], [6].

In this paper, we present a novel framework based on CNNs for simultaneous face detection, facial landmarks localization, head pose estimation and gender recognition from a given image (see Figure 1). We design a CNN architecture to learn common features for these tasks and exploit the synergy among them. We exploit the fact that information contained in features is hierarchically distributed throughout the network as demonstrated in [63]. Lower layers respond to edges and corners, and hence contain better localization properties. They are more suitable for learning landmarks localization and pose estimation tasks.

- *R. Ranjan and R. Chellappa are with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, 20742.*
  E-mail: {rranjan1,rama}@umiacs.umd.edu
- *V. M. Patel is with Rutgers University.*

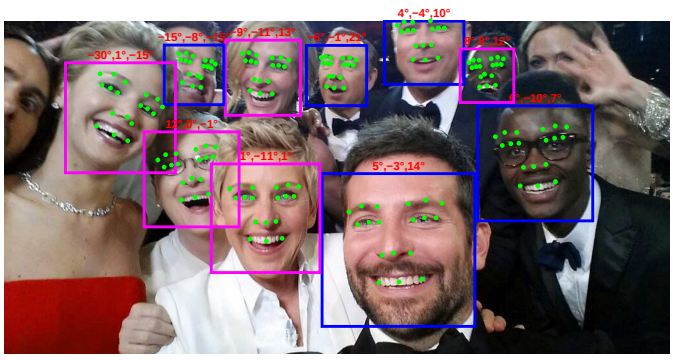


Fig. 1. Our method can simultaneously detect the face, localize landmarks, estimate the pose and recognize the gender. The blue boxes denote detected male faces, while pink boxes denote female faces. The green dots provide the landmark locations. Pose estimates for each face are shown on top of the boxes in the order of roll, pitch and yaw.

On the other hand, deeper layers are class-specific and suitable for learning complex tasks such as face detection and gender recognition. It is evident that we need to make use of all the intermediate layers of a deep CNN in order to train different tasks under consideration. We refer the set of intermediate layer features as *hyperfeatures*. We borrow this term from [1] which uses it to denote a stack of local histograms for multilevel image coding.

Since a CNN architecture contains multiple layers with hundreds of feature maps in each layer, the overall dimension of hyperfeatures is too large to be efficient for learning multiple tasks. Moreover, the hyperfeatures must be associated in a way that they efficiently encode the

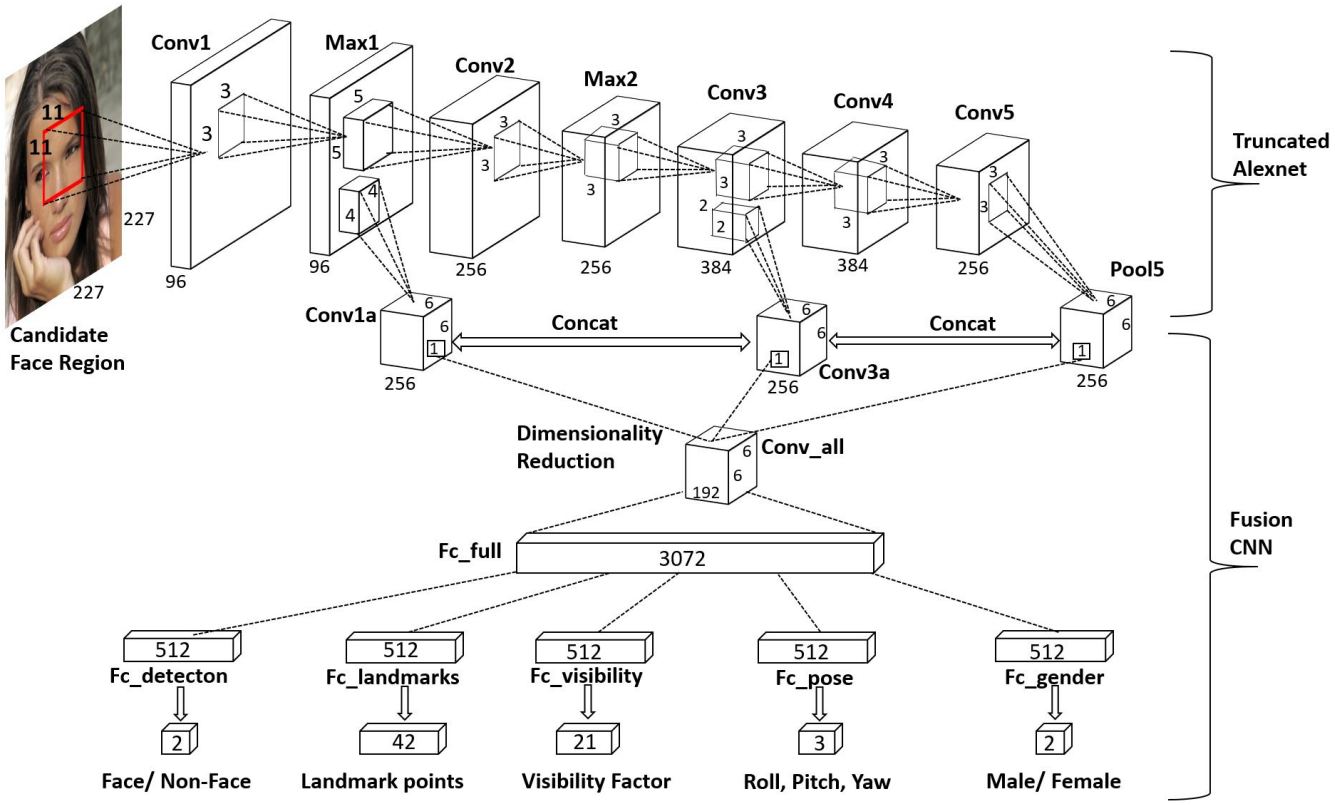


Fig. 2. The architecture of the proposed HyperFace. The network is able to classify a given image region as face or non-face, estimate the head pose, locate face landmarks and recognize gender.

features common to the multiple tasks. This can be handled using feature fusion techniques. Features fusion aims to transform the features to a common subspace where they can be combined linearly or non-linearly. Recent advances in deep learning have shown that CNNs are capable of estimating an arbitrary complex function. Hence, we construct a separate fusion-CNN to fuse the hyperfeatures. In order to learn the tasks, we train them simultaneously using multiple loss functions. In this way, the features get better at understanding faces, which leads to improvements in the performances of individual tasks. The deep CNN combined with the fusion-CNN can be learned together in an end-to-end fashion.

We also study the performance of face detection, landmarks localization, pose estimation and gender recognition tasks using off-the-shelf Region-based CNN (R-CNN [15]) approach. Although R-CNN for face detection has been explored in DP2MFD [42], we provide a comprehensive study of all these tasks based on R-CNN. Furthermore, we study the multitask approach without fusing the intermediate layers of CNN. Detailed experiments show that multitask learning performs better than methods based on individual learning. Fusing the intermediate layer features provides additional performance boost. This paper makes the following contributions.

1) We propose two novel CNN architectures that perform face detection, landmarks localization, pose estimation and gender recognition by fusing the intermediate layers of the network. The first one called HyperFace is based on AlexNet [29] model, while the second one called HyperFace-ResNet (HF-ResNet) is based on ResNet-101 [18] model.
2) We propose two post-processing methods: Iterative Region Proposals (IRP) and Landmarks-based Non-Maximum Suppression (L-NMS), which leverage the multitask information obtained from the CNN to improve the overall performance.
3) We study the performance of R-CNN based approaches for individual tasks and the multitask approach without intermediate layer fusion.
4) We achieve new state-of-the-art performances on challenging unconstrained datasets for all of these four tasks.

This paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed *HyperFace* framework in detail. Section 4 describes the implementation of R-CNN based approach, Multitask_Face and HF-ResNet. Section 5 provides the results of HyperFace and HF-ResNet along with R-CNN baselines on challenging datasets. Finally, Section 6 concludes the paper with a brief summary and discussion.

2 Related Work

Multi-Task Learning: Multi-task learning (MTL) was first analyzed in detail by Caruana [5]. Since then, several approaches have adopted MTL for solving different problems in computer vision. One of the earlier approaches for jointly

addressing the tasks of face detection, pose estimation, and landmark localization was proposed in [70] and later extended in [71]. This method is based on a mixture of trees with a shared pool of parts in the sense that every facial landmark is modeled as a part and uses global mixtures to capture the topological changes due to viewpoint variations. A joint cascade-based method was recently proposed in [6] for simultaneously detecting faces and landmark points on a given image. This method yields improved detection performance by incorporating a face alignment step in the cascade structure.

Multi-task learning using CNNs has also been studied recently. Eigen and Fergus [9] proposed a multi-scale CNN for simultaneously predicting depth, surface normals and semantic labels from an image. They apply CNNs at three different scales where the output of the smaller scale network is fed as input to the larger one. UberNet [27] adopts a similar concept of simultaneously training low-, mid- and high-level vision tasks. It fuses all the intermediate layers of a CNN at three different scales of the image pyramid for multi-task training on diverse sets. Gkioxari et al. [16] train a CNN for person pose estimation and action detection, using features only from the last layer. The use of MTL for face analysis is somewhat limited. Zhang et al. [65] used MTL-based CNN for facial landmark detection along with the tasks of discrete head yaw estimation, gender recognition, smile and glass detection. In their method, the predictions for all theses tasks were pooled from the same feature space. Instead, we strategically design the network architecture such that the tasks leverage from both low level as well as high level features of the network. We also jointly predict the task of face detection and landmark localization. These two tasks always go hand-in-hand and are used in most end-to-end face analysis systems.

Feature Fusion: Fusing intermediate layers from CNN to bring both geometry and semantically rich features together has been used by quite a few methods. Hariharan et al. [17] proposed Hypercolumns to fuse pool2, conv4 and fc7 layers of AlexNet [29] for image segmentation. Yang and Ramanan [61] proposed DAG-CNNs, which extracts features from multiple layers to reason about high, mid and low-level features for image classification. Sermanet et al. [46] merge the 1st stage output of CNN to the classifier input after sub-sampling, for the application of pedestrian detection.

Face detection: Viola-Jones detector [54] is a classic method which uses cascaded classifiers on Haar-like features to detect faces. This method provides realtime face detection, but works best for full, frontal, and well lit faces. Deformable Parts Model (DPM) [13]-based face detection methods have also been proposed in the literature where a face is essentially defined as a collection of parts [70], [39]. It has been shown that in unconstrained face detection, features like HOG or Haar wavelets do not capture the discriminative facial information at different illumination variations or poses. To overcome these limitations, various deep CNN-based face detection methods have been proposed in the literature [42], [33], [60], [12], [59]. These methods have produced state-of-the-art results on many challenging publicly available face detection datasets. Some of the other recent face detection methods include NPDFaces [36], PEP-Adapt [32], and [6].

Landmarks localization: Fiducial points extraction or landmarks localization is one of the most important steps in face recognition. Several approaches have been proposed in the literature. These include both regression-based [4], [57], [56], [55], [26], [52] and model-based [7] , [40], [35] methods. While the former learns the shape increment given a mean initial shape, the latter trains an appearance model to predict the keypoint locations. CNN-based landmark localization methods have also been proposed in recent years [50], [65], [30] and have achieved remarkable performance.

Although much work has been done for localizing landmarks for frontal faces, limited attention has been given to profile faces which occur more often in real world scenarios. Jourabloo and Liu recently proposed PIFA [25] that estimates 3D landmarks for large pose face alignment by integrating a 3D point distribution model with a cascaded coupled-regressor. Similarly, 3DDFA [69] fits a dense 3D model by estimating its parameters using a CNN. Zhu et al. [68] proposed a cascaded compositional learning approach that combines shape prediction from multiple domain specific regressors.

Pose estimation: The task of head pose estimation is to infer the orientation of person's head relative to the camera view. It is useful in face verification for matching face similarity across different orientations. Non-linear manifold-based methods have been proposed in [2], [19], [48] to classify face images based on pose. A survey of various head pose estimation methods is provided in [41].

Gender recognition: Previous works on gender recognition have focused on finding good discriminative features for classification. Most previous methods use one or more combination of features such as LBP, SURF, HOG or SIFT. In recent years, attribute-based methods for face recognition have gained a lot of traction. Binary classifiers were used in [31] for each attribute such as male, long hair, white etc. Separate features were computed for different attributes and they were used to train individual SVMs for each attribute. CNN-based methods have also been proposed for learning attribute-based representations in [38], [64].

3 HyperFace

We propose a single CNN model for simultaneous face detection, landmark localization, pose estimation and gender classification. The network architecture is deep in both vertical and horizontal directions, i.e., it has both top-down and lateral connections, as shown in Figure 2. In this section, we provide a brief overview of the system and then discuss the different components in detail.

The proposed algorithm called *HyperFace* consists of three modules. The first one generates class independent region-proposals from the given image and scales them to 227×227 pixels. The second module is a CNN which takes in the resized candidate regions and classifies them as face or non-face. If a region gets classified as a face, the network additionally provides facial landmarks locations, estimated head pose and gender information. The

third module is a post-processing step which involves Iterative Region Proposals (IRP) and Landmarks-based Non-Maximum Suppression (L-NMS) to boost the face detection score and improve the performance of individual tasks.

3.1 HyperFace Architecture

We start with Alexnet [29] for image classification. The network consists of five convolutional layers along with three fully connected layers. We initialize the network with the weights of R-CNN_Face network trained for face detection task as described in Section 4. All the fully connected layers are removed as they encode image-classification specific information, which is not needed for pose estimation and landmarks extraction. We exploit the following two observations to create our network. 1) The features in CNN are distributed hierarchically in the network. While the lower layer features are informative for landmarks localization and pose estimation, the higher layer features are suitable for more complex tasks such as detection or classification [63]. 2) Learning multiple correlated tasks simultaneously builds a synergy and improves the performance of individual tasks as shown in [6], [65]. Hence, in order to simultaneously learn face detection, landmarks, pose and gender, we need to fuse the features from the intermediate layers of the network (hyperfeatures), and learn multiple tasks on top of it. Since the adjacent layers are highly correlated, we do not consider all the intermediate layers for fusion.

We fuse the max_1, $conv_3$ and $pool_5$ layers of Alexnet, using a separate network. A naive way for fusion is directly concatenating the features. Since the feature maps for these layers have different dimensions $27 \times 27 \times 96$, $13 \times 13 \times 384$, $6 \times 6 \times 256$, respectively, they cannot be easily concatenated. We therefore add $conv_{1a}$ and $conv_{3a}$ convolutional layers to $pool_1$, $conv_3$ layers to obtain consistent feature maps of dimensions $6 \times 6 \times 256$ at the output. We then concatenate the output of these layers along with $pool_5$ to form a $6 \times 6 \times 768$ dimensional feature maps. The dimension is still quite high to train a multi-task framework. Hence, a 1×1 kernel convolution layer ($conv_{all}$) is added to reduce the dimensions [51] to $6 \times 6 \times 192$. We add a fully connected layer (fc_{all}) to $conv_{all}$, which outputs a 3072 dimensional feature vector. At this point, we split the network into five separate branches corresponding to the different tasks. We add $fc_{detection}$, $fc_{landmarks}$, $fc_{visibility}$, fc_{pose} and fc_{gender} fully connected layers, each of dimension 512, to fc_{all}. Finally, a fully connected layer is added to each of the branch to predict the individual task labels. After every convolution or a fully connected layer, we deploy the Rectified Linear Unit (ReLU). We did not include any pooling operation in the fusion network as it provides local invariance which is not desired for the face landmark localization task. Task-specific loss functions are then used to learn the weights of the network.

3.2 Training

We use the AFLW [28] dataset for training the HyperFace network. It contains $25,993$ faces in $21,997$ real-world images with full pose, expression, ethnicity, age and gender variations. It provides annotations for 21 landmark points per face, along with the face bounding-box, face pose (yaw, pitch and roll) and gender information. We randomly selected 1000 images for testing, and used the rest for training the network. Different loss functions are used for training the tasks of face detection, landmark localization, pose estimation and gender classification.

Face Detection: We use the Selective Search [53] algorithm in R-CNN [15] to generate region proposals for faces in an image. A region having an Intersection over Union (IOU) overlap of more than 0.5 with the ground truth bounding box is considered a positive sample ($l = 1$). The candidate regions with IOU overlap less than 0.35 are treated as negative instances ($l = 0$). All the other regions are ignored. We use the softmax loss function given by (1) for training the face detection task.

$$loss_D = -(1 - l) \cdot \log(1 - p) - l \cdot \log(p), \tag{1}$$

where p is the probability that the candidate region is a face. The probability values p and $1-p$ are obtained from the last fully connected layer for the detection task.

Landmarks Localization: We use 21 point markups for face landmarks locations as provided in the AFLW [28] dataset. Since the faces have full pose variations, some of the landmark points are invisible. The dataset provides the annotations for the visible landmarks. We consider bounding-box regions with IOU overlap greater than 0.35 with the ground truth for learning this task, while ignoring the rest. A region can be characterized by $\{x, y, w, h\}$ where (x, y) are the co-ordinates of the center of the region and w,h are the width and height of the region respectively. Each visible landmark point is shifted with respect to the region center (x, y), and normalized by (w, h) as given by (2)

$$(a_i, b_i) = \left(\frac{x_i - x}{w}, \frac{y_i - y}{h} \right). \tag{2}$$

where (x_i, y_i)'s are the given ground truth fiducial co-ordinates. The (a_i, b_i)'s are treated as labels for training the landmark localization task using the Euclidean loss weighted by the visibility factor. The loss in predicting the landmark location is computed from (3)

$$loss_L = \frac{1}{2N} \sum_{i=1}^{N} v_i((\hat{x}_i - a_i)^2 + ((\hat{y}_i - b_i)^2), \tag{3}$$

where $(\hat{x}_i, \hat{y}_i)$ is the i^{th} landmark location predicted by the network, relative to a given region, N is the total number of landmark points (21 for AFLW [28]). The visibility factor v_i is 1 if the i^{th} landmark is visible in the candidate region, else it is 0. This implies that there is no loss corresponding to invisible points and hence they do not take part during back-propagation.

Learning Visibility: We also learn the visibility factor in order to test the presence of the predicted landmark. For a given region with overlap higher than 0.35, we use a simple Euclidean loss to train the visibility as shown in (4)

$$loss_V = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{v}_i - v_i \right)^2, \tag{4}$$

where $\hat{v}_i$ is the predicted visibility of i^{th} landmark. The true visibility v_i is 1 if the i^{th} landmark is visible in the candidate region, else it is 0.

Pose Estimation: We use the Euclidean loss to train the head pose estimates of roll (p_1), pitch (p_2) and yaw (p_3). We compute the loss for a candidate region having an overlap more than 0.5 with the ground truth, from (5)

$$loss_P = \frac{(\hat{p_1} - p_1)^2 + (\hat{p_2} - p_2)^2 + (\hat{p_3} - p_3)^2}{3}, \quad (5)$$

where $(\hat{p_1}, \hat{p_2}, \hat{p_3})$ are the estimated pose labels.

Gender Recognition: Predicting gender is a two class problem similar to face detection. For a candidate region with overlap of 0.5 with the ground truth, we compute the softmax loss given in (6)

$$loss_G = -(1-g) \cdot \log(1-p_g) - g \cdot \log(p_g), \quad (6)$$

where $g = 0$ if the gender is male, or else $g = 1$. Here, (p_0, p_1) is the two dimensional probability vector computed from the network.

The total loss is computed as the weighted sum of the five individual losses as shown in (7)

$$loss_{full} = \sum_{i=1}^{i=5} \lambda_{t_i} loss_{t_i}, \quad (7)$$

where t_i is the i^{th} element from the set of tasks $T = \{D, L, V, P, G\}$. The weight parameter λ_{t_i} is decided based on the importance of the task in the overall loss. We choose $(\lambda_D = 1, \lambda_L = 5, \lambda_V = 0.5, \lambda_P = 5, \lambda_G = 2)$ for our experiments. Higher weights are assigned to landmark localization and pose estimation tasks as they need spatial accuracy.

Fig. 3. Candidate face region (red box on left) obtained using Selective Search gives a low score for face detection, while landmarks are correctly localized. We generate a new face region (red box on right) using the landmarks information and feed it through the network to increase the detection score.

3.3 Testing

From a given test image, we first extract the candidate region proposals using [53]. For each region, we predict the task labels by a forward-pass through the HyperFace network. Only those regions, whose detection scores are above a certain threshold, are classified as face and processed for subsequent tasks. The predicted landmark points are scaled and shifted to the image co-ordinates using (8)

$$(x_i, y_i) = (\hat{x_i} w + x, \hat{y_i} h + y), \quad (8)$$

where $(\hat{x_i}, \hat{y_i})$ are the predicted locations of i^{th} landmark from the network, and $\{x, y, w, h\}$ are the region parameters defined in (2). Points obtained with predicted visibility less than a certain threshold are marked invisible. The pose labels obtained from the network are the estimated roll, pitch and yaw for the face region. The gender is assigned according to the label with maximum predicted probability.

There are two major issues while using proposal-based face detection. First, the proposals might not be able to capture small and difficult faces, hence reducing the overall recall of the system. Second, the proposal boxes might not be well localized with the actual face region. It is a common practice to use bounding-box regression [15] as a post processing step to improve the localization of the detected face box. This adds an additional burden of training regressors to learn the transformation from the candidate detected box to the annotated face box. Moreover, the localization is still weak since the regressors are usually linear. Recently, Gidaris and Komodakis proposed LocNet [14] which tries to solve these limitations by refining the detection bounding box. Given a set of initial bounding box proposals, it generates new sets of bounding boxes that maximize the likelihood of each row and column within the box. It allows an accurate inference of bounding box under a simple probabilistic framework.

Instead of using the probabilistic framework [14], we solve the above mentioned issues in an iterative way using the predicted landmarks. The fact that we obtain landmark locations along with the detections, enables us to improve the post-processing step so that all the tasks benefit from it. We propose two novel methods: Iterative Region Proposals (IRP) and Landmarks-based Non-Maximum Suppression (L-NMS) to improve the performance. IRP improves the recall by generating more candidate proposals by using the predicted landmarks information from the initial set of region proposals. On the other hand, L-NMS improves the localization by re-adjusting the detected bounding boxes according to the predicted landmarks and performing NMS on top of them. No additional training is required for these methods.

Iterative Region Proposals (IRP): We use a fast version of Selective Search [53] which extracts around 2000 regions from an image. We call this version $Fast\_SS$. It is quite possible that some faces with poor illumination or small size fail to get captured by any candidate region with a high overlap. The network would fail to detect that face due to low score. In these situations, it is desirable to have a candidate box which precisely captures the face. Hence, we generate a new candidate bounding box from the predicted landmark points using the FaceRectCalculator provided by [28], and pass it again through the network. The new region, being more localized yields a higher detection score and improves the corresponding tasks output, thus increasing the recall. This procedure can be repeated (say T time), so that boxes at a given step will be more localized to faces as compared to the previous step. From our experiments, we found that the localization component saturates in just one step (T =

1), which shows the strength of the predicted landmarks. The pseudo-code of IRP is presented in Algorithm 1. The usefulness of IRP can be seen in Figure 3, which shows a low-resolution face region cropped from the top-right image in Figure 15.

Algorithm 1 Iterative Region Proposals

boxes $\leftarrow selective\_search($**image**$)$
scores $\leftarrow get\_hyperface\_scores($**boxes**$)$
detected_boxes $\leftarrow$ **boxes**$(scores \geq threshold)$
new_boxes $\leftarrow$ **detected_boxes**
for `stage = 1 to T` **do**
  fids $\leftarrow get\_hyperface\_fiducials($**new_boxes**$)$
  new_boxes $\leftarrow FaceRectCalculator($**fids**$)$
  deteced_boxes $\leftarrow$ [**deteced_boxes**|**new_boxes**]
end
final_scores $\leftarrow get\_hyperface\_scores($**detected_boxes**$)$

Landmarks-based Non-Maximum Suppression (L-NMS): The traditional approach of non-maximum suppression involves selecting the top scoring region and discarding all the other regions with overlap more than a certain threshold. This method can fail in the following two scenarios: 1) If a region corresponding to the same detected face has less overlap with the highest scoring region, it can be detected as a separate face. 2) The highest scoring region might not always be localized well for the face, which can create some discrepancy if two faces are close together. To overcome these issues, we perform NMS on a new region whose bounding box is defined by the boundary co-ordinates as $[\min_i x_i, \min_i y_i, \max_i x_i, \max_i y_i]$ of the landmarks for the given region. In this way, the candidate regions would get close to each other, thus decreasing the ambiguity of the overlap and improving the localization.

Algorithm 2 Landmarks-based NMS

Get **detected_boxes** from Algorithm 1
fids $\leftarrow get\_hyperface\_fiducials($**detected_boxes**$)$
precise_boxes $\leftarrow [min_x, min_y, max_x, max_y]($**fids**$)$
faces $\leftarrow nms($**precise_boxes**$, overlap)$
for each `face in` **faces do**
  top-k_boxes $\leftarrow$ Get top-k scoring boxes
  final_fids $\leftarrow median(fids($**top-k_boxes**$))$
  final_pose $\leftarrow median(pose($**top-k_boxes**$))$
  final_gender $\leftarrow median(gender($**top-k_boxes**$))$
  final_visibility $\leftarrow median(visibility($**top-k_boxes**$))$
  final_bounding_box $\leftarrow FaceRectCalculator($**final_fids**$)$
end

We apply landmarks-based NMS to keep the top-k boxes, based on the detection scores. The detected face corresponds to the region with maximum score. The landmark points, pose estimates and gender classification scores are decided by the median of the top k boxes obtained. Hence, the predictions do not rely only on one face region, but considers the votes from top-k regions for generating the final output. From our experiments, we found that the best results are obtained with the value of k being 5. The pseudo-code for L-NMS is given in Algorithm 2.

4 Network Architectures

To emphasize the importance of multitask approach and fusion of the intermediate layers of CNN, we study the performance of simpler CNNs devoid of such features. We evaluate four R-CNN-based models, one for each task of face detection, landmark localization, pose estimation and gender recognition. We also build a separate Multitask_Face model which performs multitask learning just like HyperFace, but does not fuse the information from the intermediate layers. These models are described as follows:

R-CNN_Face: This model is used for face detection task. The network architecture is shown in Figure 4(a). For training R-CNN_Face, we use the region proposals from AFLW [28] training set, each associated with a face label based on the overlap with the ground truth. The loss is computed as per (1). The model parameters are initialized using the Alexnet [29] weights trained on the Imagenet dataset [8]. Once trained, the learned parameters from this network are used to initialize other models including Multitask_Face and HyperFace as the standard Imagenet initialization doesn't converge well. We also perform a linear bounding box regression to localize the face co-ordinates.

R-CNN_Fiducial: This model is used for locating the facial landmarks. The network architecture is shown in Figure 4(b). It simultaneously learns the visibility of the points to account for the invisible points at test time, and thus can be used as a standalone fiducial extractor. The loss functions for landmarks localization and visibility of points are computed using (3) and (4), respectively. Only region proposals which have an overlap> 0.5 with the ground truth bounding box are used for training. The model parameters are initialized from R-CNN_Face.

R-CNN_Pose: This model is used for head pose estimation task. The outputs of the network are roll, pitch and yaw of the face. Figure 4(c) presents the network architecture. Similar to R-CNN_Fiducial, only region proposals with overlap> 0.5 with the ground truth bounding box are used for training. The training loss is computed using (5).

R-CNN_Gender: This model is used for face gender recognition task. The network architecture is shown in Figure 4(d). It has the same training set as R-CNN_Fiducial and R-CNN_Pose. The training loss is computed using (6).

Multitask_Face: Similar to HyperFace, this model is used to simultaneously detect face, localize landmarks, estimate pose and predict its gender. The only difference between Multitask_Face and HyperFace is that HyperFace fuses the intermediate layers of the network whereas Multitask_Face combines the tasks using the common fully connected layer at the end of the network as shown in Figure 5. Since it provides the landmarks and face score, it leverages iterative region proposals and landmark-based NMS post-processing algorithms during evaluation.

The performance of all the above models for their respective tasks are evaluated and discussed in details in Section 5.

4.1 HyperFace-ResNet

The CNN architectures have improved a lot over the years, mainly due to an increase in number of layers [18],

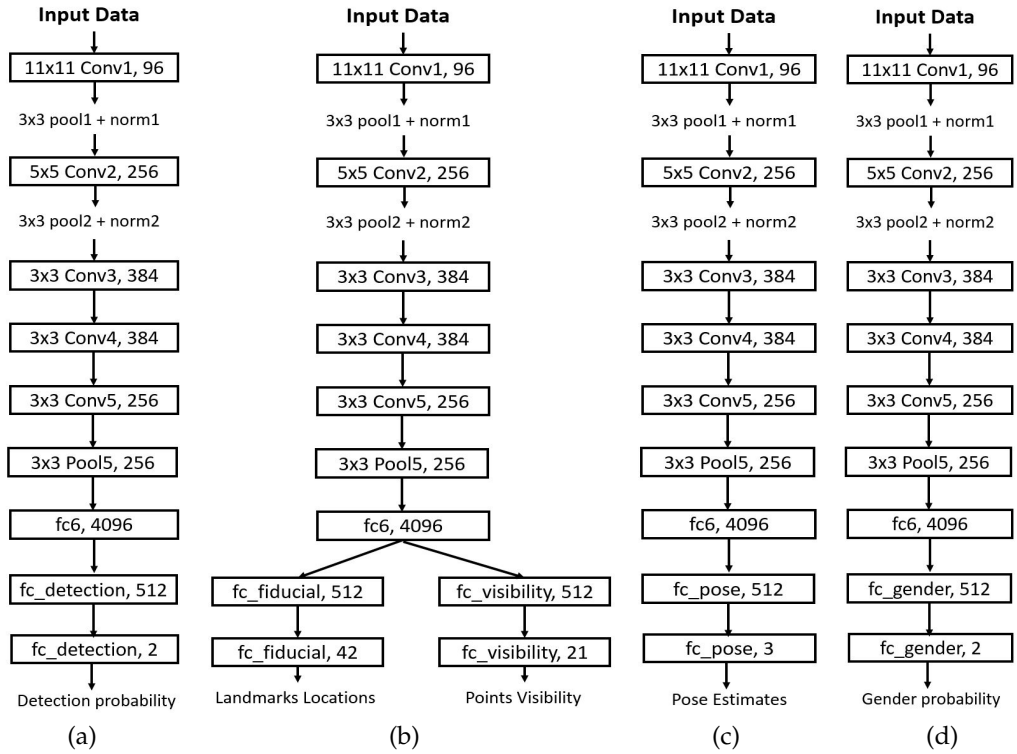


Fig. 4. R-CNN-based network architectures for (a) Face Detection (R-CNN_Face), (b) Landmark Localization (R-CNN_Fiducial), (c) Pose Estimation (R-CNN_Pose), and (d) Gender Recognition (R-CNN_Gender). The numbers on the left denote the kernel size and the numbers on the right denote the cardinality of feature maps for a given layer.

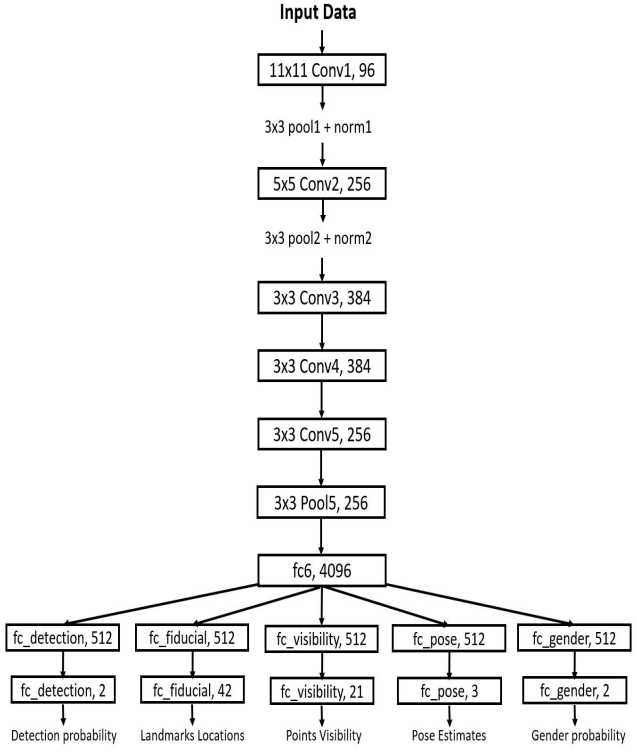


Fig. 5. Network Architecture of Multitask_Face. The numbers on the left denote the kernel size and the numbers on the right denote the cardinality of feature maps for a given layer.

effective convolution kernel size [47], batch normalization [22] and skip connections. Recently, He et al. [18] proposed a deep residual network architecture with more than 100 layers, that achieves state-of-the-art results on the ImageNet challenge [8]. Hence, we propose a variant of HyperFace that is built using the ResNet-101 [18] model instead of AlexNet [29]. The proposed network called HyperFace-ResNet (HF-ResNet) significantly improves upon its AlexNet baseline for all the tasks of face detection, landmarks localization, pose estimation and gender recognition. Figure 6 shows the network architecture for HF-ResNet.

Similar to HyperFace, we fuse the geometrically rich features from the lower layers and semantically strong features from the deeper layers of ResNet, such that multi-task learning can leverage from their synergy. Taking inspiration from [20], we fuse the features using hierarchical element-wise addition. Starting with 'res2c' features, we first reduce its resolution using a 3×3 convolution kernel with stride of 2. It is then passed through the a 1×1 convolution layer that increases the number of channels to match the next level features ('res3b3' in this case). Element-wise addition is applied between the two to generate a new set of fused features. The same operation is applied in a cascaded manner to fuse 'res4b22' and 'res5c' features of the ResNet-101 model. Finally, average pooling is carried out to generate 2048-dimensional feature vector that is shared among all the tasks. Task-specific sub-networks are branched out separately in a similar way as HyperFace. Each convolution layer is followed by a Batch-Norm+Scale [22] layer and ReLU activation unit. We do not use dropout in HF-ResNet.

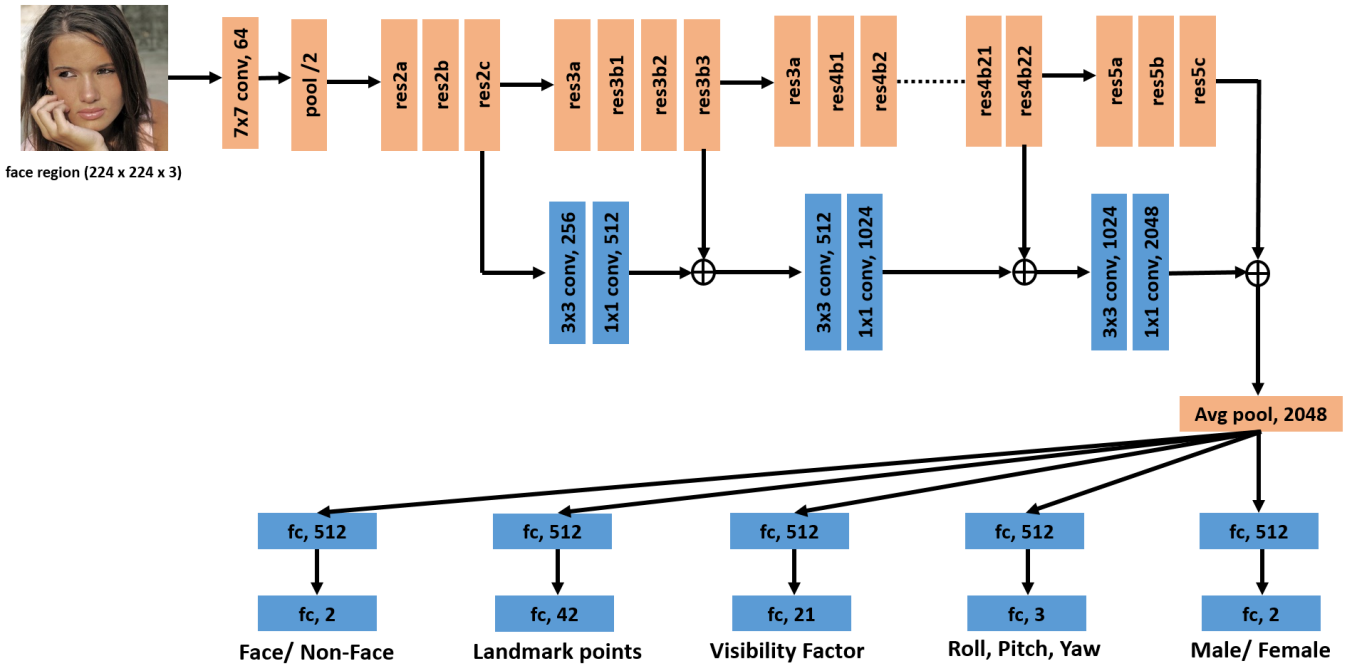


Fig. 6. The architecture of the proposed HyperFace-Resnet (HF-ResNet). ResNet-101 model is used as the backbone network, represented in color orange. The new layers added are represented in color blue. The network is able to classify a given image region as face or non-face, estimate the head pose, locate face landmarks and recognize gender.

The training loss functions are the same as described in Section 3.2.

HF-ResNet is slower than HyperFace since it performs more convolutions. This makes it difficult to be used with Selective Search [53] algorithm which generates more than 2000 region proposals to be processed. Hence, we use a faster version of region proposals using high recall SSD [37] face detector. It produces 200 proposals, needing just 0.05s. This considerably reduces the total runtime for HF-ResNet to less than 1s. The fast version of HyperFace is discussed in Section 5.6.

5 Experimental Results

We evaluated the proposed HyperFace method, along with HF-ResNet, Multask_Face, R-CNN_Face, R-CNN_Fiducial, R-CNN_Pose and R-CNN_Gender on six challenging datasets:

- Annotated Face in-the-Wild (AFW) [70] for evaluating face detection, landmarks localization, and pose estimation tasks
- 300-W Faces in-the-wild (IBUG) [44] for evaluating 68-point landmarks localization.
- Annotated Facial Landmarks in the Wild (AFLW) [28] for evaluating landmarks localization and pose estimation tasks
- Face Detection Dataset and Benchmark (FDDB) [23] and PASCAL faces [58] for evaluating the face detection results
- Large-scale CelebFaces Attributes (CelebA) [38] and LFWA [21] for evaluating gender recognition results.

Our method was trained on randomly selected $20,997$ images from the AFLW dataset using Caffe [24]. The remaining 1000 images were used for testing.

5.1 Face Detection

We present face detection results for AFW, PASCAL and FDDB datasets. The AFW dataset [70] was collected from Flickr and the images in this dataset contain large variations in appearance and viewpoint. In total there are 205 images with 468 faces in this dataset. The FDDB dataset [23] consists of 2,845 images containing 5,171 faces collected from news articles on the Yahoo website. This dataset is the most widely used benchmark for unconstrained face detection. The PASCAL faces dataset [58] was collected from the test set of PASCAL person layout dataset, which is a subset from PASCAL VOC [10]. This dataset contains 1335 faces from 851 images with large appearance variations. For improved face detection performance, we learn a SVM classifier on top of $fc_{detection}$ features using the training splits from the FDDB dataset.

Some of the recent published methods compared in our evaluations include DP2MFD [42], Faceness [60], HeadHunter [39], JointCascade [6], CCF [59], SquaresChnFtrs-5 [39], CascadeCNN [33], Structured Models [58], DDFD [12], NPDFace [36], PEP-Adapt [32], TSM [70], as well as three commercial systems Face++, Picasa and Face.com.

The precision-recall curves of different detectors corresponding to AFW and PASCAL faces datasets are shown in Figures 7 (a) and (b), respectively. Figure 8 compares the performance of different detectors using the Receiver Operating Characteristic (ROC) curves on the FDDB dataset. As can be seen from these figures, both HyperFace and HF-ResNet outperform all the reported academic and commercial detectors on the AFW and PASCAL datasets. HyperFace achieves a high mean average precision (mAP) of 97.9% and 92.46%, for AFW and PASCAL datasets respectively. HF-ResNet further improves the mAP to 99.4% and 96.2%

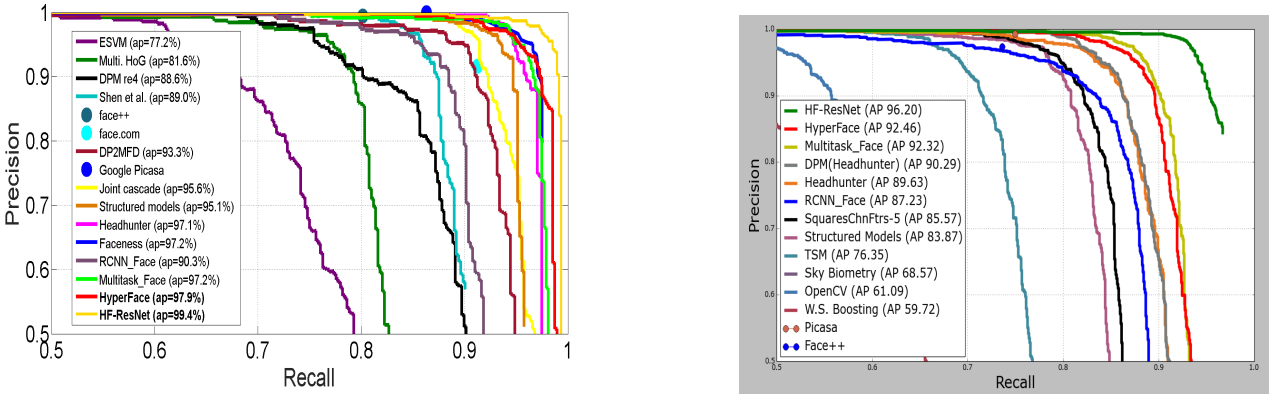


Fig. 7. **Face Detection** performance evaluation on (a) the AFW dataset, (b) the PASCAL faces dataset. The numbers in the legend are the mean average precision (mAP) for the corresponding datasets.

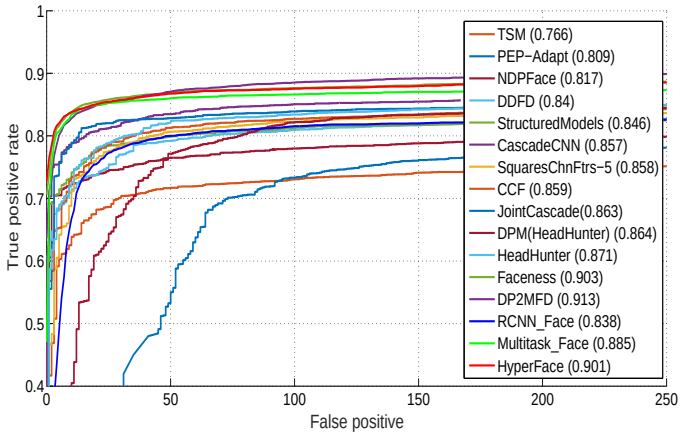


Fig. 8. **Face Detection** performance evaluation on the FDDB dataset. The numbers in the legend are the mean average precision.

respectively.

The FDDB dataset is very challenging for HyperFace and any other R-CNN-based face detection methods, as the dataset contains many small and blurred faces. First, some of these faces do not get included in the region proposals from selective search. Second, re-sizing small faces to the input size of 227×227 adds distortion to the face resulting in low detection score. In spite of these issues, HyperFace performance is comparable to recently published deep learning-based face detection methods such as DP2MFD [42] and Faceness [60] on the FDDB dataset [1] with mAP of 90.1%.

It is interesting to note the performance differences between R-CNN_Face, Multitask_Face and HyperFace for the face detection tasks. Figures 7, and 8 clearly show that multi-task CNNs (Multitask_Face and HyperFace) outperform R-CNN_Face by a wide margin. The boost in the performance gain is mainly due to the following two reasons. First, multi-task learning approach helps the network to learn improved features for face detection which is evident from their mAP values on the AFW dataset. Using just the linear bounding box regression and traditional NMS, the HyperFace obtains a mAP of 94% (Figure 13) while R-CNN_Face achieves a mAP of 90.3%. Second, having landmark information associated with detection boxes makes it easier to localize the bounding box to a face, by using IRP and L-NMS algorithms. On the other hand, HyperFace and Multi-task_Face perform comparable to each other for all the face detection datasets which suggests that the network does not gain much by fusing intermediate layers for the face detection task.

1. http://vis-www.cs.umass.edu/fddb/results.html

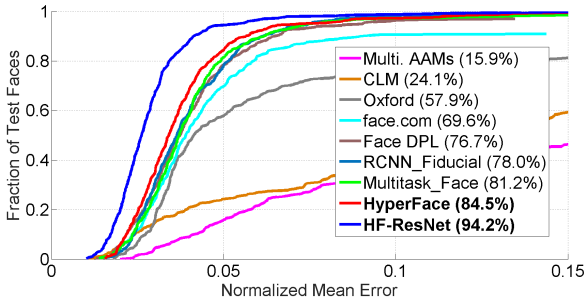


Fig. 9. **Landmarks Localization** cumulative error distribution curves on the AFW dataset. The numbers in the legend are the fraction of testing faces that have average error below (5%) of the face size.

5.2 Landmarks Localization

We evaluate the performance of different landmarks localization algorithms on AFW [70] and AFLW [28] datasets. Both of these datasets contain faces with full pose variations. Some of the methods compared include Multiview Active Appearance Model-based method (Multi. AAM) [70], Constrained Local Model (CLM) [45], Oxford facial landmark detector [11], Zhu [70], FaceDPL [71], JointCascade [6], CDM [62], RCPR [3], ESR [4], SDM [56] and 3DDFA [69]. Although both of these datasets provide ground truth bounding boxes, we do not use them for evaluating on HyperFace, HF-ResNet, Multitask_Face and R-CNN_Fiducial. Instead we use the respective algorithms to detect both the face and its fiducial points. Since, the R-CNN_Fiducial cannot detect faces, we provide it with the detections from the HyperFace.

Figure 9 compares the performance of different landmark localization methods on the AFW dataset using the protocol defined in [71]. In this figure, (*) indicates that models that are evaluated on near frontal faces or use hand-initialization [70]. The dataset provides six keypoints for each face which are: left_eye_center, right_eye_center, nose_tip, mouth_left, mouth_center and mouth_right. We compute the error as the mean distance between the predicted and ground truth keypoints, normalized by the face size. The plots for comparison were obtained from [71].

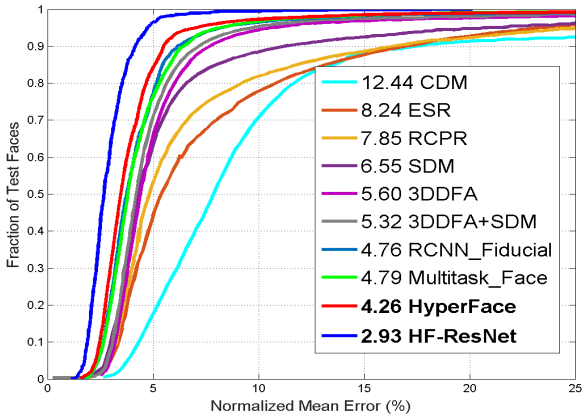


Fig. 10. **Landmarks Localization** cumulative error distribution curves on the AFLW dataset. The numbers in the legend are the average NME for the test images. The test samples are chosen such that samples with absolute yaw angles between [0°,30°], [30°,60°] and [60°,90°] are $1/3$ each.

For the AFLW dataset, we calculate the error using all the visible keypoints. For AFW, we adopt the same protocol as defined in [69]. The only difference is that our AFLW testset consists of only 1000 images with 1132 face samples, since we use the rest of the images for training. To be consistent with the protocol, we randomly create a subset of 450 samples from our testset whose absolute yaw angles within $[0°, 30°]$, $[30°, 60°]$ and $[60°, 90°]$ are $1/3$ each. Figure 10 compares the performance of different landmark localization methods. We obtain the comparison plots from [69] where the evaluations for RCPR, ESR and SDM are carried out after adapting the algorithms to face profiling. Table 1 provides the Normalized Mean Error (NME) for AFLW dataset, for each of the pose group.

TABLE 1
The NME(%) of face alignment results on AFLW test set with the best results highlighted.

	AFLW Dataset (21 pts)				
Method	[0, 30]	[30, 60]	[60, 90]	mean	std
CDM [62]	8.15	13.02	16.17	12.44	4.04
RCPR [3]	5.43	6.58	11.53	7.85	3.24
ESR [4]	5.66	7.12	11.94	8.24	3.29
SDM [56]	4.75	5.55	9.34	6.55	2.45
3DDFA [69]	5.00	5.06	6.74	5.60	0.99
3DDFA [69]+SDM	4.75	4.83	6.38	5.32	0.92
R-CNN_Fiducial	4.49	4.70	5.09	4.76	0.30
Multitask_Face	4.20	4.93	5.23	4.79	0.53
HyperFace	**3.93**	**4.14**	**4.71**	**4.26**	**0.41**
HF-ResNet	**2.71**	**2.88**	**3.19**	**2.93**	**0.25**

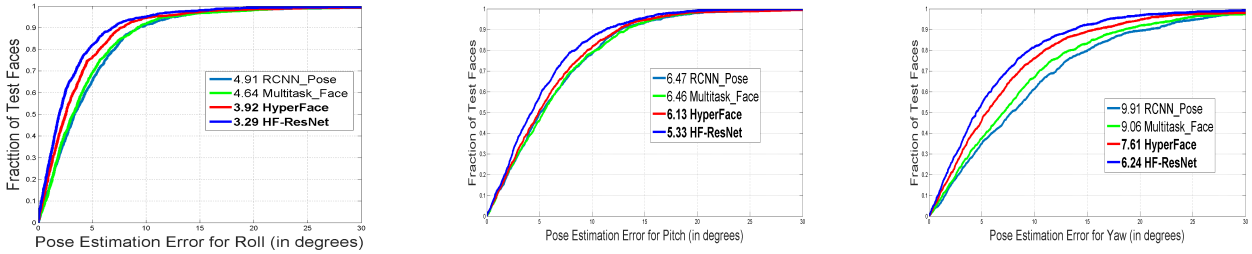
As can be seen from the figures, R-CNN_Fiducial, Multitask_Face, HyperFace and HF-ResNet outperform many recent state-of-the-art landmark localization methods including FaceDPL [71], 3DDFA [69] and SDM [56]. Table 1 shows that HyperFace performs consistently accurate over all pose angles. This clearly suggests that while most of the methods work well on frontal faces, HyperFace is able to predict landmarks for faces with full pose variations. Moreover, we find that R-CNN_Fiducial and Multitask_Face attain similar performance. The HyperFace has an advantage over them as it uses the intermediate layers for fusion. The local information is contained well in the lower layers of CNN and becomes invariant as depth increases. Fusing the layers brings out that hidden information which boosts the performance for the landmark localization task. Additionally, we observe that HF-ResNet significantly improves the performance over HyperFace for both AFW and AFLW datasets. The large margin in performance can be attributed to the larger depth for the HF-ResNet model.

We also evaluate our models on the challenging subset of the 300-W [44] landmarks localization dataset (IBUG). The dataset contains 135 test images with wide variations in expression and illumination. The head-pose angle varies from $-60°$ to $60°$ in yaw. Since the dataset contains 68 landmarks points instead of 21 used in AFLW [28] training, the model cannot be directly applied for evaluating IBUG. We retrain the network for predicting 68 facial key-points as a separate task in conjunction with the proposed tasks in hand. We implement it by adding two fully-connected layers in a cascade manner to the shared feature space (fc-full), having dimensions 512 and 136, respectively.

Following the protocol described in [43], we use $3,148$ faces with 68-point annotations for training. The network is trained end-to-end for the localization of 68-points landmarks along with the other tasks mentioned in Section 3.2. We use standard Euclidean loss function for training. For evaluation, we compute the average error of all 68 landmarks normalized by the inter-pupil distance. Table 2 compares the Normalized Mean Error (NME) obtained by HyperFace and HF-ResNet with other recently published methods. We observe that HyperFace achieves a comparable NME of 10.88, while HF-ResNet achieves the state-of-the-art result on IBUG [44] with NME of 8.18. This shows the effectiveness of the proposed models for 68-point landmarks localization.

TABLE 2
Normalized Mean Error (in %) of 68-point landmarks localization on IBUG [44] dataset.

Method	Normalized Mean Error
CDM [62]	19.54
RCPR [3]	17.26
ESR [4]	17.00
SDM [56]	15.40
LBF [43]	11.98
LDDR [30]	11.49
CFSS [67]	9.98
3DDFA [69]	10.59
TCDCN [66]	8.60
HyperFace	10.88
HF-ResNet	**8.18**



(a) (b) (c)

Fig. 11. **Pose Estimation** performance evaluation on AFLW dataset for (a) roll (b) pitch and (c) yaw angles. The numbers in the legend are the mean error in degrees for the respective pose angles.

5.3 Pose Estimation

We evaluate R-CNN_Pose, Multitask_Face and HyperFace on the AFW [70] and AFLW [28] datasets for the pose estimation task. The detection boxes used for evaluating the landmark localization task are used here as well for initialization. For the AFW dataset, we compare our approach with Multi. AAM [70], Multiview HoG [70], FaceDPL[2] [71] and face.com. Note that multiview AAMs are initialized using the ground truth bounding boxes (denoted by *). Figure 12 shows the cumulative error distribution curves on AFW dataset. The curve provides the fraction of faces for which the estimated pose is within some error tolerance. As can be seen from the figure, both HyperFace and HF-ResNet outperform existing methods by a large margin. For the AFLW dataset, we do not have pose estimation evaluation for any previous method. Hence, we show the performance of our method for different pose angles: roll, pitch and yaw in Figure 11 (a), (b) and (c) respectively. It can be seen that the network is able to learn roll, and pitch information better than yaw.

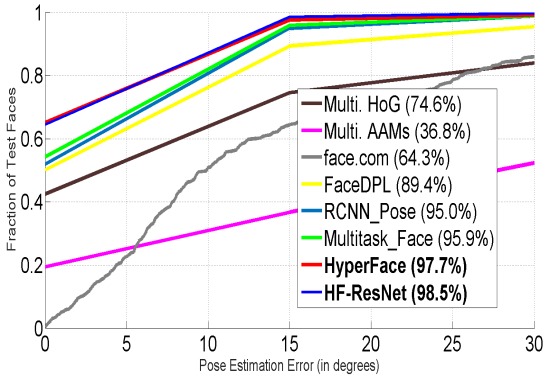


Fig. 12. **Pose Estimation** cumulative error distribution curves on AFW dataset. The numbers in the legend are the percentage of faces that are labeled within $\pm 15^{\circ}$ error tolerance.

The performance traits of R-CNN_Pose, Multitask_Face, HyperFace and HF-ResNet for pose estimation task are similar to that of the landmarks localization task. R-CNN_Pose and Multitask_Face perform comparable to each other whereas HyperFace achieves a boosted performance due to the intermediate layers fusion. It shows that tasks which rely on the structure and orientation of the face work well with features from lower layers of the CNN. HF-ResNet further improves the performance for roll, pitch as well as yaw.

2. Available at: http://www.ics.uci.edu/~dramanan/software/face/face_journal.pdf

5.4 Gender Recognition

We present the gender recognition performance on CelebA [38] and LFWA [21] datasets since these datasets come with gender information. The CelebA and LFWA datasets contain labeled images selected from the CelebFaces [49] and LFW [21] datasets, respectively [38]. The CelebA dataset contains 10,000 identities and there are 200,000 images in total. The LFWA dataset has 13,233 images of 5,749 identities. We compare our approach with FaceTracer [31], PANDA-w [64], PANDA-1 [64], [34] with ANet and [38]. The gender recognition performance of different methods is reported in Table 3. On the LFWA dataset, our method outperforms PANDA [64] and FaceTracer [31], and is equal to [38]. On the CelebA dataset, our method performs comparably to [38]. Unlike [38] which uses $180,000$ images for training and validation, we only use $20,000$ images from validation set of CelebA to fine-tune the network.

TABLE 3
Performance comparison (in %) of **gender recognition** on CelebA and LFWA datasets.

Method	CelebA	LFWA
FaceTracer [31]	91	84
PANDA-w [64]	93	86
PANDA-1 [64]	97	92
[34]+ANet	95	91
LNets+ANet [38]	**98**	**94**
R-CNN_Gender	95	91
Multitask_Face	97	93
HyperFace	97	**94**
HF-ResNet	**98**	**94**

Similar to the face detection task, we find that gender recognition performs better for HyperFace and Multitask_Face as compared to R-CNN_Gender proving that learning related tasks together improves the discriminating capability of the individual tasks. Again, we do not see

much difference in the performance of Multitask_Face and HyperFace suggesting intermediate layers do not contribute much for the gender recognition task. HF-ResNet achieves state-of-the-art results on both CelebA [38] and LFWA [21] datasets.

5.5 Effect of Post-Processing

Figure 13 provides an experimental analysis of the post-processing methods: IRP and L-NMS, for face detection task on the AFW dataset. *Fast SS* denotes the quick version of selective search which produces around 2000 region proposals and takes $2s$ per image to compute. On the other hand, *Quality SS* refers to its slow version which outputs more than $10,000$ region proposals consuming more than $10s$ for one image. The HyperFace with a linear bounding box regression and traditional NMS achieves a mAP of 94%. Just by replacing them with L-NMS provides a boost of 1.2%. In this case, bounding-box is constructed using the landmarks information rather linear regression. Additionaly, we can see from the figure that although *Quality SS* generates more region proposals, it performs worse than *Fast SS* with ierative region proposals. IRP adds 300 new regions for a typical image consuming less than $0.5s$ which makes it highly efficient as compared to *Quality SS*.

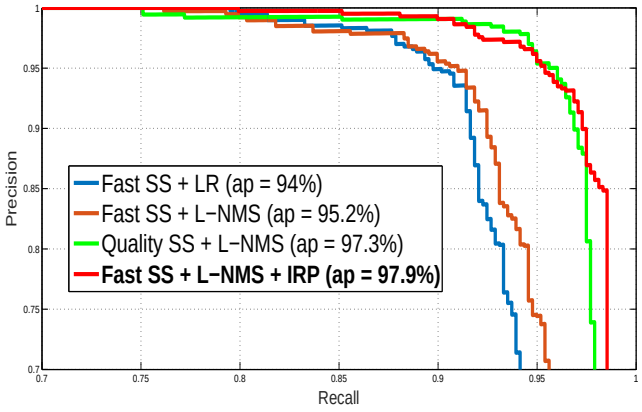


Fig. 13. Variations in performance of HyperFace with respect to the Iterative Region Proposals and Landmarks-based NMS. The numbers in the legend are the mean average precision.

5.6 Fast-HyperFace

The Hyperface method is tested on a machine with 8 cores and GTX TITAN-X GPU. The overall time taken to perform all the four tasks is $3s$ per image. The limitation is not because of CNN, but due to Selective Search [53] algorithm which takes approximately $2s$ to generate candidate region proposals. One forward pass through the HyperFace network for 200 proposals takes merely 0.1s.

We also propose a fast version of HyperFace which uses a high recall fast face detector instead of Selective Search [53] to generate candidate region proposals. We implement a face detector using Single Shot Detector (SSD) [37] framework. The SSD-based face detector takes a 512×512 dimensional input image and generates face boxes in less than 0.05s, with confidence probability scores ranging from 0 to 1. We use a probability threshold of 0.01 to select high recall detection boxes. Unlike traditional SSD, we do not use non-maximum suppression on the detector output, so that we have more number of region proposals. Typically, the SSD face detector generates 200 proposals per image. These proposals are directly passed through HyperFace to generate face detection scores, localize face landmarks, estimate pose and recognize gender for every face in the image. Fast-HyperFace consumes a total time of 0.15s (0.05s for SSD face detector, and 0.1s for HyperFace) on a GTX TITAN X GPU. The Fast-HyperFace achieves a mAP of 97.6% on AFW face detection task, which is comparable to the HyperFace mAP of 97.9%. Thus, Fast-HyperFace improves the speed by a factor of 12 with negligible degradation in performance.

6 Discussion

We discuss few crucial observations from our experiments. First, all the face related tasks benefit from using the multi-task learning framework. The gain is mainly due to the network's ability to learn more discriminative features, and post-processing methods which can be leveraged by having landmarks as well as detection scores for a region. Secondly, fusing intermediate layers improves the performance for structure dependent tasks of pose estimation and landmarks localization, as the features become invariant to geometry in deeper layers of CNN. The HyperFace exploits these observations to improve the performance for all the four tasks.

We also visualize the features learned by the HyperFace network. Figure 14 shows the network activation for a few selected feature maps out of 192 from the $conv_{all}$ layer. It can be seen that some feature maps are dedicated solely for a single task while others can be used to predict different tasks. For example, feature map 27 and 186 can be used for face detection and gender recognition, respectively. The former distinguishes the face and non-face regions whereas the latter outputs high activation for the female faces. Similarly, feature map 19 shows high activation near eyes and mouth regions, while feature map 96 gives a rough contour of the face orientation. Both of these features can be used for landmark localization and pose estimation tasks.

Several qualitative results of our method on the AFW, PASCAL and FDDB datasets are shown in Figure 15. As can be seen from this figure, our method is able to simultaneously perform all the four tasks on images containing extreme pose, illumination, and resolution variations with cluttered background.

7 Conclusion

In this paper, we presented a multi-task deep learning method called HyperFace for simultaneously detecting faces, localizing landmarks, estimating head pose and identifying gender. Extensive experiments using various publicly available unconstrained datasets demonstrate the effectiveness of our method on all four tasks. In future, we will evaluate the performance of our method on other applications such as simultaneous human detection and human pose estimation, object recognition and pedestrian detection.

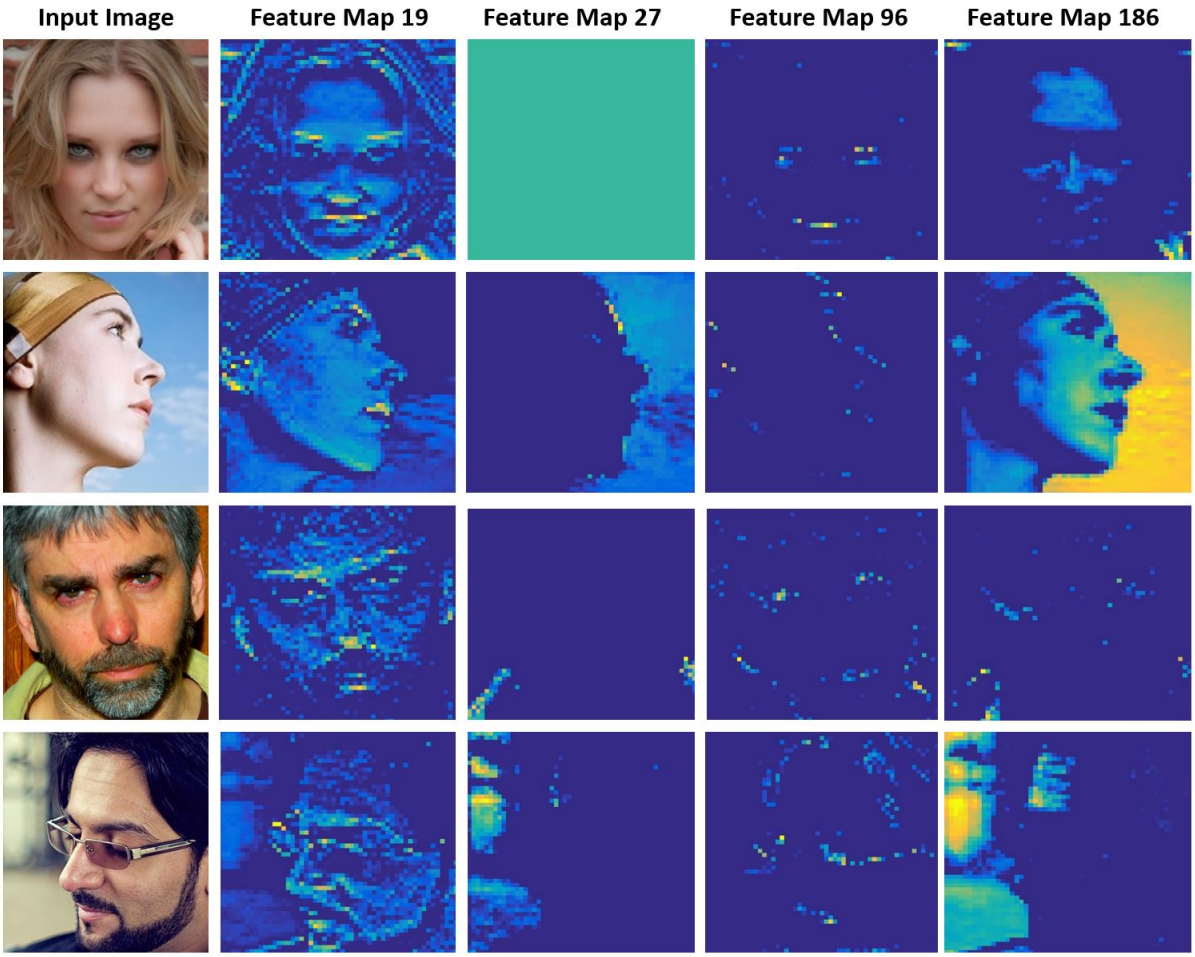


Fig. 14. Activations of selected feature maps from **conv_all** layer of the HyperFace architecture. Green and yellow colors denote high activation whereas blue denotes low activation units. These features depict the distinguishable face traits for the tasks of face detection, landmarks localization, pose estimation and gender recognition.

ACKNOWLEDGMENTS

We thank Dr. Jun-Cheng Chen for implementing the SSD512-based face detector used in the Fast-HyperFace pipeline. 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.

REFERENCES

[1] A. Agarwal and B. Triggs. Multilevel image coding with hyperfeatures. *International Journal of Computer Vision*, pages 15–27, 2008.
[2] V. Balasubramanian, J. Ye, and S. Panchanathan. Biased manifold embedding: A framework for person-independent head pose estimation. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–7, June 2007.
[3] X. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1513–1520, 2013.
[4] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.
[5] R. Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.
[6] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *European Conference on Computer Vision*, volume 8694, pages 109–122. 2014.
[7] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. *Comput. Vis. Image Underst.*, 61(1):38–59, Jan. 1995.
[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.
[9] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.
[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.
[11] M. R. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... buffy" – automatic naming of characters in tv video. In *Proceedings of the British Machine Vision Conference*, pages 92.1–92.10, 2006.
[12] S. S. Farfade, M. Saberian, and L.-J. Li. Multi-view face detection using deep convolutional neural networks. In *International Conference on Multimedia Retrieval*, 2015.
[13] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based mod-

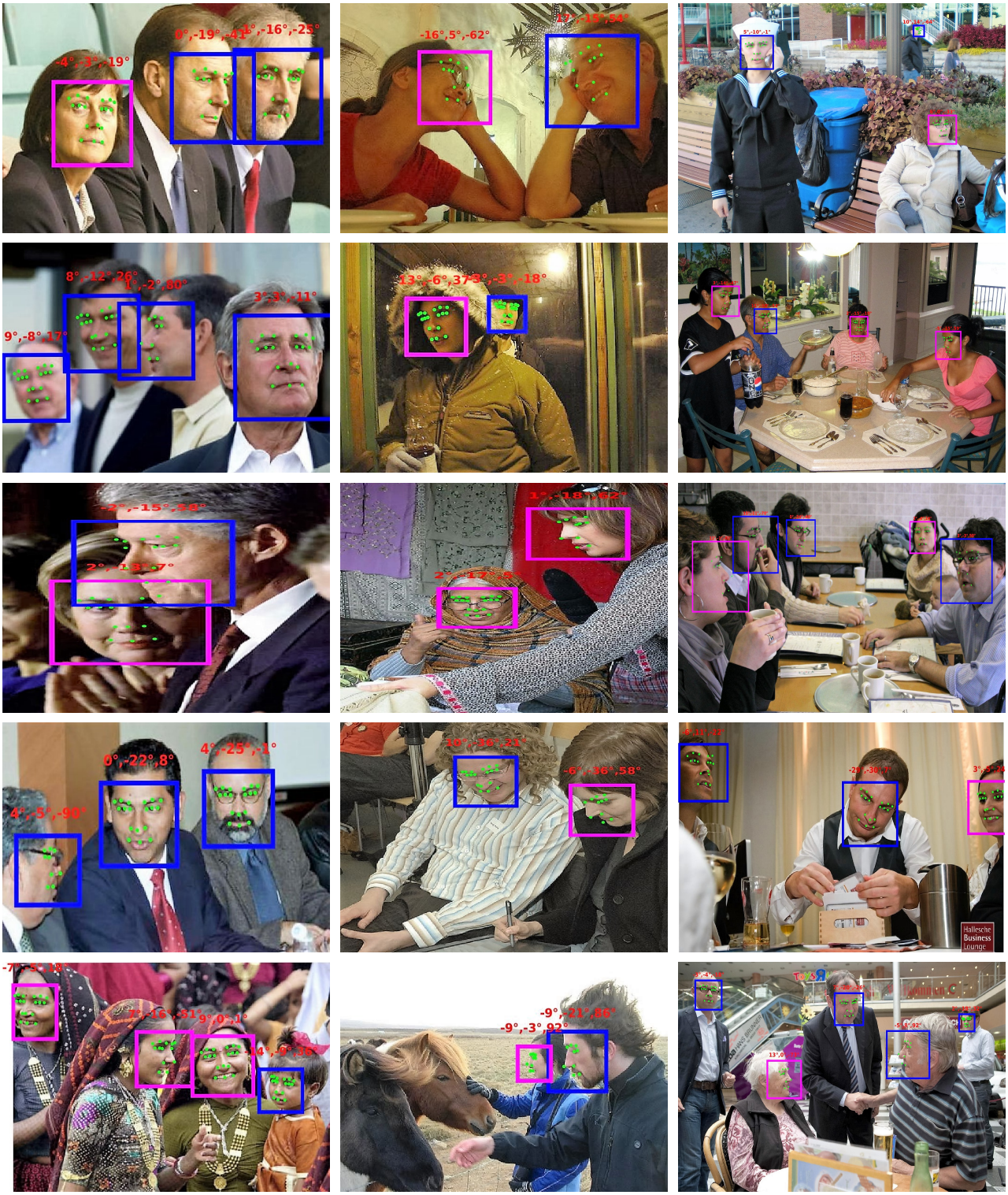


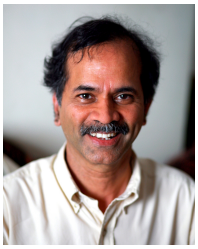
Fig. 15. Qualitative results of our method. The blue boxes denote detected male faces, while pink boxes denote female faces. The green dots provide the landmark locations. Pose estimates for each face are shown on top of the boxes in the order of roll, pitch and yaw.

els. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sept 2010.
[14] S. Gidaris and N. Komodakis. Locnet: Improving localization accuracy for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 789–798, 2016.
[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, 2014.
[16] G. Gkioxari, B. Hariharan, R. B. Girshick, and J. Malik. R-cnns for pose estimation and action detection. *CoRR*, abs/1406.5212, 2014.
[17] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.
[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
[19] N. Hu, W. Huang, and S. Ranganath. Head pose estimation by non-linear embedding and mapping. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 2, pages II–342–5, Sept 2005.
[20] P. Hu and D. Ramanan. Finding tiny faces. *arXiv preprint arXiv:1612.04402*, 2016.
[21] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, Oct. 2007.
[22] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.
[23] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.
[24] 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.
[25] A. Jourabloo and X. Liu. Pose-invariant 3d face alignment. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.
[26] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, June 2014.
[27] I. Kokkinos. Ubernet: Training auniversal'convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. *arXiv preprint arXiv:1609.02132*, 2016.
[28] M. Kostinger, P. Wohlhart, P. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *IEEE International Conference on Computer Vision Workshops*, pages 2144–2151, Nov 2011.
[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
[30] A. Kumar, R. Ranjan, V. M. Patel, and R. Chellappa. Face alignment by local deep descriptor regression. *CoRR*, abs/1601.07950, 2016.
[31] N. Kumar, P. N. Belhumeur, and S. K. Nayar. FaceTracer: A Search Engine for Large Collections of Images with Faces. In *European Conference on Computer Vision (ECCV)*, pages 340–353, Oct 2008.
[32] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic part model for unsupervised face detector adaptation. In *IEEE International Conference on Computer Vision*, pages 793–800, Dec 2013.
[33] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5325–5334, June 2015.
[34] J. Li and Y. Zhang. Learning surf cascade for fast and accurate object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3468–3475, 2013.
[35] L. Liang, R. Xiao, F. Wen, and J. S. 0001. Face alignment via component-based discriminative search. In D. A. Forsyth, P. H. S. Torr, and A. Zisserman, editors, *ECCV (2)*, volume 5303 of *Lecture Notes in Computer Science*, pages 72–85. Springer, 2008.
[36] S. Liao, A. Jain, and S. Li. A fast and accurate unconstrained face detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.
[37] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
[38] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, Dec. 2015.
[39] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, volume 8692, pages 720–735. 2014.
[40] I. Matthews and S. Baker. Active appearance models revisited. *Int. J. Comput. Vision*, 60(2):135–164, Nov. 2004.
[41] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):607–626, April 2009.
[42] R. Ranjan, V. M. Patel, and R. Chellappa. A deep pyramid deformable part model for face detection. In *International Conference on Biometrics Theory, Applications and Systems*, 2015.
[43] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.
[44] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.
[45] J. Saragih, S. Lucey, and J. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.
[46] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. Lecun. Pedestrian detection with unsupervised multi-stage feature learning. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 3626–3633, Washington, DC, USA, 2013. IEEE Computer Society.
[47] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
[48] S. Srinivasan and K. Boyer. Head pose estimation using view based eigenspaces. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 4, pages 302–305 vol.4, 2002.
[49] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pages 1988–1996. 2014.
[50] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 3476–3483, Washington, DC, USA, 2013. IEEE Computer Society.
[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.
[52] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, June 2014.
[53] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 1879–1886, Washington, DC, USA, 2011. IEEE Computer Society.
[54] P. A. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.
[55] X. Xiong and F. D. la Torre. Global supervised descent method. In *CVPR*, 2015.
[56] Xuehan-Xiong and F. De la Torre. Supervised descent method and its application to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
[57] J. Yan, Z. Lei, D. Yi, and S. Z. Li. Learn to combine multiple hypotheses for accurate face alignment. In *Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops*, ICCVW '13, pages 392–396, Washington, DC, USA, 2013. IEEE Computer Society.
[58] J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Face detection by structural models. *Image and Vision Computing*, 32(10):790 – 799, 2014.
[59] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Convolutional channel features. In *IEEE International Conference on Computer Vision*, 2015.
[60] S. Yang, P. Luo, C. C. Loy, and X. Tang. From facial parts responses to face detection: A deep learning approach. In *IEEE International Conference on Computer Vision*, 2015.
[61] S. Yang and D. Ramanan. Multi-scale recognition with dag-cnns. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.
[62] X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded

deformable shape model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1944–1951, 2013.
[63] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.
[64] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1644, 2014.
[65] Z. Zhang, P. Luo, C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108, 2014.
[66] 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*, 38(5):918–930, 2016.
[67] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015.
[68] S. Zhu, C. Li, C.-C. Loy, and X. Tang. Unconstrained face alignment via cascaded compositional learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3409–3417, 2016.
[69] 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.
[70] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, June 2012.
[71] X. Zhu and D. Ramanan. FaceDPL: Detection, pose estimation, and landmark localization in the wild. preprint 2015.

Rajeev Ranjan received the B.Tech. degree in Electronics and Electrical Communication Engineering from Indian Institute of Technology Kharagpur, India, in 2012. He is currently a Research Assistant at University of Maryland College Park. He is pursuing Ph.D. under the supervision of Dr. Rama Chellappa. His research interests include face detection, face recognition and machine learning. He is a recipient of UMD Outstanding Invention of the Year award, 2015.

Vishal M. Patel received the B.S. degrees in electrical engineering and applied mathematics (Hons.) and the M.S. degree in applied mathematics from North Carolina State University, Raleigh, NC, USA, in 2004 and 2005, respectively, and the Ph.D. degree in electrical engineering from the University of Maryland College Park, MD, USA, in 2010. He is currently an A. Walter Tyson Assistant Professor in the Department of Electrical and Computer Engineering (ECE) at Rutgers University. Prior to joining Rutgers University, he was a member of the research faculty at the University of Maryland Institute for Advanced Computer Studies (UMIACS). His current research interests include signal processing, computer vision, and pattern recognition with applications in biometrics and imaging. He has received a number of awards including the 2016 ONR Young Investigator Award, the 2016 Jimmy Lin Award for Invention, A. Walter Tyson Assistant Professorship Award, the Best Paper Award at IEEE BTAS 2015, and Best Poster Awards at BTAS 2015 and 2016. He is an Associate Editor of the IEEE Signal Processing Magazine, IEEE Biometrics Compendium, and serves on the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society. He is a member of Eta Kappa Nu, Pi Mu Epsilon, and Phi Beta Kappa.



Rama Chellappa is a Minta Martin Professor of Engineering and Chair of the ECE department at the University of Maryland. Prof. Chellappa received the K.S. Fu Prize from the International Association of Pattern Recognition (IAPR). He is a recipient of the Society, Technical Achievement and Meritorious Service Awards from the IEEE Signal Processing Society and four IBM faculty Development Awards. He also received the Technical Achievement and Meritorious Service Awards from the IEEE Computer Society. At UMD, he received college and university level recognitions for research, teaching, innovation and mentoring of undergraduate students. In 2010, he was recognized as an Outstanding ECE by Purdue University. Prof. Chellappa served as the Editor-in-Chief of PAMI. He is a Golden Core Member of the IEEE Computer Society, served as a Distinguished Lecturer of the IEEE Signal Processing Society and as the President of IEEE Biometrics Council. He is a Fellow of IEEE, IAPR, OSA, AAAS, ACM and AAAI and holds four patents.