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

Rajeev Ranjan
University of Maryland
College Park, MD 20742
rranjan1@umd.edu

Vishal M. Patel
Rutgers University
New Brunswick, NJ 08901
vishal.m.patel@rutgers.edu

Rama Chellappa
University of Maryland
College Park, MD 20742
rama@umiacs.umd.edu

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 and trains multi-task loss on the fused features. It exploits the synergy among the tasks which boosts up their individual performances. Extensive experiments show that the proposed method is able to capture both global and local information of faces and performs significantly better than many competitive algorithms for each of these four tasks.

1. Introduction

Detection and analysis of faces is a challenging problem in computer vision, and has been actively researched for its extensive applications on face verification, face tracking, person identification, etc. Although recent methods based on deep Convolutional Neural Networks (CNN) have achieved remarkable results for face detection tasks [9], [32], [47], it is still difficult to obtain facial landmark locations, head pose estimates and gender information from face images containing extreme pose, 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 the individual tasks [52], [51], [4].

In this paper, we present a novel framework based on CNNs for simultaneous face detection, facial landmark 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 build the synergy among them. We exploit the fact that information contained in features is hierarchically distributed throughout the network as demonstrated in [48]. Lower layers respond to edges and corners, and hence contain better localization features. They would be more suitable for learning landmark localization and pose estimation tasks. On the other hand, higher layers are class-specific and suitable for learning complex tasks which is desired for 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 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 any 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 understanding of faces, which leads to improvements in the performances of the individual tasks. The deep CNN combined with the fusion-CNN can be learned together end-to-end. Since the network performs hyperfeatures fusion, we name it Hyper-net. This paper makes the following key contributions.

1. We propose a novel CNN architecture that performs the multiple tasks of face detection, landmarks localization, pose estimation and gender recognition.
2. We propose two post-processing methods: recursive region proposal and landmark-based non-maximum suppression, which leverage the multitask information obtained from the CNN to improve the overall performance.
3. 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 provides the multitask results on challenging datasets. Finally, Section 5 concludes the paper with a brief summary and discussion.

2. Related Work

One of the earlier approaches for jointly addressing the tasks of face detection, pose estimation, and landmark localization was proposed in [51] and later extended in [52]. 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 topological changes due to viewpoint changes. A joint cascade-based method was recently proposed in [4] 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 lately. [50] learns gender and other attributes to improve landmark localization, while [12] trains a CNN for person pose estimation and action detection, using features only from the last layer. The intermediate layer features have been used for solving the tasks of image segmentation [13] and pedestrian detection [34].
Face detection: Viola-Jones detector [41] is a classic method which uses cascaded classifiers on Haar-like features to detect faces. This method provides real-time face detection, but works best for full, frontal, and well-lit faces. Deformable Parts Model (DPM) [10]-based face detection methods have also been proposed in the literature where a face is essentially defined as a collection of parts [31], [29]. 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 this limitations, various deep CNN-based face detection methods have been proposed in the literature [32], [24], [47], [9], [46]. 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 NDFaces [27], PEP-Adapt [23], and [4].

Landmark localization: Fiducial point extraction or landmark localization is one of the most important steps in face recognition. Several approaches have been proposed in the literature. These include both regression-based [3], [44], [43], [42] and model-based [5], [30], [26]. 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 [37], [50] and have achieved remarkable performance.

Pose estimation: The task of head pose estimation is to infer the orientation of person’s head relative to the camera view. It is extremely useful in face verification for matching face similarity in different orientations. However, not much research has gone into pose estimation from unconstrained images. Non-linear manifold-based methods have been proposed in [2], [14], [35] to classify face images based on pose. A survey of various head pose estimation methods is provided in [31].

Gender recognition: Previous works on gender recognition have focused on finding good discriminative features for classification. Most previous methods use one or a 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 [22] for each attribute such as male, long hair, white etc. Different features were computed for different features and they were used to train a different SVM for each attribute. CNN-based methods have also been proposed for learning attribute-based representations [28], [49].

3. HyperFace

We propose a single CNN model called Hyper-net for simultaneous face detection, landmark localization, pose estimation and gender classification. The network architecture is deep in both vertical and horizontal directions and is shown in Figure 2. In this section, we provide a brief overview of the system pipeline and then discuss the different components in this pipeline in detail.

The proposed framework 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 the facial landmarks locations, estimated head pose and the gender information for it. The third module is a post-processing step which involves recursive region-proposals and landmarks-based k-NMS to boost the face detection score and improve the performance of individual tasks.

3.1. Hyper-net Architecture

We start with the Alexnet[21] network for image classification. The network consists of five convolutional layers along with three fully connected layers. We initialize the network with the Imagenet Challenge (ILSVRC2012 [6]) pre-trained weights distributed with the Caffe[18] implementation. All the fully connected layers are removed as they encode image-classification task specific information, which is not desirable for face related tasks. We utilize 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 localization tasks such as landmarks and pose estimation, the higher layer features are suitable for more complex tasks such as detection or classification[48]. 2) Learning multiple correlated tasks simultaneously builds a synergy and improves the performance of individual tasks as shown in [4, 50]. 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 max1, conv3 and pool5 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 27×27×96, 13×13×384, 6×6×256, respectively, they cannot be concatenated directly. We therefore add conv1a and conv3a convolutional layers to pool1, conv3 layers to obtain consistent feature maps of dimensions 6×6×256 at the output. We then concatenate the output of these layers along with pool5 to form a 6×6×768 dimensional feature maps. The dimension is still quite high to train a multi-task framework. Hence, a 1×1 kernel convolution layer (convall) is added to reduce the dimensions [38] to 6×6×192. We add a fully connected layer (fcall) to convall, 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 \( f_{\text{detection}}, f_{\text{landmarks}}, f_{\text{visibility}}, f_{\text{pose}} \) and \( f_{\text{gender}} \) fully connected layers, each of dimension 512, to \( f_{\text{all}} \). Finally, a fully connected layer is added to each of the branch to the individual task labels. After every convolution or fully connected layer, we deploy Rectified Layer Unit (ReLU) non-linearity. We did not include any pooling operation in the fusion network as they provide 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 AFLW[20] dataset for training the Hyper-net. 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 select 1000 images for testing, and keep the rest for training the network. We formulate different loss functions for training the tasks of face detection, landmark localization, pose estimation and gender classification.

**Face Detection:** We use the Selective Search[40] algorithm in a similar manner as RCNN [11] to generate region proposals for faces in an image. A region having an overlap of more than 0.5 with the ground truth bounding box is considered a positive sample (\( l = 1 \)). The candidate regions with overlap less than 0.35 are treated as negative instance (\( l = 0 \)). All the other regions are ignored. We use the softmax loss function given by (1) for training the face detection task.

\[
\text{loss}_D = -(1 - l) \cdot \log(1 - p) - l \cdot \log(p),
\]

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.

**Landmark Localization:** We use the 21 point markup for face landmark location as provided in the AFLW[20] 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 the regions with overlap greater than 0.35 with the ground truth for learning this task, while ignore the rest. A region can be characterized by \( \{x, y, w, h\} \) where \( (x, y) \) is the co-ordinate of the center of the region and \( w, h \) is 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)

\[
(a_i, b_i) = \left(\frac{x_i - x}{w}, \frac{y_i - y}{h}\right),
\]

where \( (x_i, y_i) \)'s are the given ground truth fiducial coordinates. 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 labels for landmarks which are not visible are taken to be \((0, 0)\). The loss in predicting the landmark location can be computed from (3)

\[
\text{loss}_L = \frac{1}{2N} \sum_{i=1}^{N} v_i ((\hat{x}_i - a_i)^2 + (\hat{y}_i - b_i)^2),
\]

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[20]). 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)

\[
\text{loss}_V = \frac{1}{N} \sum_{i=1}^{N} (\hat{v}_i - v_i)^2,
\]

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)

\[
\text{loss}_P = \frac{(\hat{p}_1 - p_1)^2 + (\hat{p}_2 - p_2)^2 + (\hat{p}_3 - p_3)^2}{3},
\]

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)

\[
\text{loss}_G = -(1 - g) \cdot \log(1 - p_0) - g \cdot \log(p_1),
\]

where \(g = 0\) if the gender is male, or else \(g = 1\). \((p_0, p_1)\) 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)

\[
\text{loss}_{\text{all}} = \sum_{t=1}^{5} \lambda_t \text{loss}_t,
\]

where \(\text{loss}_t\) is the individual loss corresponding to the \(t^{th}\) task. The weight parameter \(\lambda_t\) 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.

3.3. Testing

From a given test image, we first extract the candidate region proposals using \([40]\). For each of the regions, we predict the task labels by a forward-pass through HyperNet. Only those regions with detection scores above a certain threshold are classified as face and considered for the subsequent tasks. The predicted landmark points are scaled and shifted to the image co-ordinates using (8)

\[
(x_i, y_i) = (\bar{x}_i + x, \bar{y}_i + y),
\]

where \((\bar{x}_i, \bar{y}_i)\) is the predicted location 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.

The fact that we obtain the 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: recursive region proposals and landmark-based \(k\)-NMS to improve the recall performance.

Recursive Region Proposals: We use a fast version of Selective Search[40] which extracts around 2000 regions from an image. 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 \([20]\). The new region, being more localized yields a higher detection score and the corresponding tasks output, thus increasing the recall. The usefulness of this step can be appreciated from Figure 3.

Landmarks-based \(k\)-NMS: Traditional way 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

\[
[x_i, y_i, max_i x_i, max_i y_i].
\]

In this way, the candidate regions would get close to each other, thus decreasing the ambiguity of the overlap and improving the face localization.

We apply landmarks-based \(k\)-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 best results are obtained with the value of \(k\) being 5.

4. Experimental Results

We evaluated the proposed HyperFace method on six challenging datasets: Annotated Face in-the-Wild (AFW) \([51]\) for evaluating face detection, landmark localization, and pose estimation; Annotated Facial Landmarks in the Wild (AFLW) \([20]\) for evaluating landmark localization and pose estimation; Face Detection Dataset and Benchmark (FDB) \([17]\) and PASCAL faces \([45]\) for evaluating face detection; Large-scale CelebFaces Attributes (CelebA) \([28]\) and LFWA \([16]\) for evaluating gender recognition. Our method was trained on the AFLW dataset using Caffe \([18]\).

4.1. Face Detection

We show face detection results for AFW, FDB and PASCAL datasets. The AFW dataset \([51]\) 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 [17] 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 [45] was collected from the test set of PASCAL person layout dataset, which is a subset from PASCAL VOC [7]. 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 features using the training splits from the FDDB dataset.

Some of the recent published methods compared in our evaluations include DP2MFD [32], Faceness [47], HeadHunter [29], JointCascade [4], CCF [46], SquaresChnFtrs-5 [29], CascadeCNN [24], Structured Models [1], DDFD [9], NDPFace [27], PEP-Adapt [23], TSM [51], as well as three commercial systems Face++, Picasa and Face.com.

The precision-recall curves of different detectors corresponding to the AFW and the PASCAL faces datasets are shown in Figures 4 (a) and (c), respectively. Figure 4 (b) compares the performance of different detectors using the Receiver Operating Characteristic (ROC) curves on the FDDB dataset. As can be seen from these figures, our method outperforms all the academic and commercial detectors on the AFW and PASCAL datasets and performs comparably to recently published deep learning-based face detection methods such as DP2MFD [32] and faceness [47] on the FDDB dataset 1.

4.2. Landmark Localization

We evaluate the performance of different landmark localization algorithms on the AFW [51] and AFLW [20] datasets. Some of the methods compared include Multiview Active Appearance Model-based method (Multi. AAM) [51], Constrained Local Model (CLM) [33], Oxford facial landmark detector [8], Zhu [51], FaceDPL [52], JointCascade [4], Convolutional Latent Variable Models (CLVM) [15], Ensemble of Regression Trees (ERT) [19], Gaussian-Newton Deformable Part Models (GN-DPM) [39], and Tasks-Constrained Deep Convolutional Network (TCDCN) [50].

Figure 5 compares the performance of different landmark localization methods. In this figure, (*) indicates that models that are evaluated on near frontal faces or use hand-initialization [51]. We compute the error as the mean distance between the predicted and ground truth keypoints, normalized by the face size. For AFLW, we calculate the error using all the visible keypoints. To evaluate on AFW, we adopt the same protocol as defined in [52]. The number in the legend indicates the fraction of test faces with less than 5% error. As can be seen from this figure, our method outperforms many recent state-of-the-art landmark localization methods including FaceDPL [52], TCDCN [50], ERT [19], GN-DPM [39] and CLVM [15] on both datasets. This clearly shows that while most of the methods work well on frontal faces, the proposed method is able to predict landmarks for faces with full pose variations.

---

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

We show the gender recognition performance of our method on the CelebA [28] and LFW [16] datasets since these datasets come with gender information. The CelebA and LFW datasets contain labeled images selected from the CelebFaces [36] and LFW [16] datasets, respectively [28]. 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 [22], PANDA-w [49], PANDA-1 [49], [25] with ANet and [28]. The gender recognition performance of different methods is reported in Table 1. On the LFW dataset, our method outperforms [28] as well as PANDA-1 and FaceTracer [22]. On the CelebA dataset, method performs comparably to [28]. Unlike [28] which uses 180,000 images for training, we only use 20,000 images from validation dataset to train a SVM classifier on $f_{gender}$ features.

| Method         | CelebA | LFWA  |
|----------------|--------|-------|
| FaceTracer [22]| 91     | 84    |
| PANDA-w [49]   | 93     | 86    |
| PANDA-1 [49]   | 97     | 92    |
| [25]+ANet      | 95     | 91    |
| LNet+ANet [28] | 98     | 94    |
| HyperFace      | 96.33  | 96.25 |

Table 1. Performance comparison (in %) of gender recognition on CelebA and LFW datasets.

4.4. Pose Estimation

We evaluate our method on the AFW [51] and AFLW [20] datasets. For AFW dataset, we compare our approach with Multi. AAM [51], Multiview HoG [51], FaceDPL 2 [52] and face.com. Note that multiview AAMs are initialized using the ground truth bounding boxes (denoted by *). Figures 6 (a) and (b) show the cumulative error distribution curves on AFW and AFLW, respectively. These curves show the fraction of faces for which the estimated pose is within some error tolerance. As can be seen from this figure, the HyperFace method achieves the best performance on both of these datasets and beats FaceDPL by a large margin. For AFLW dataset, we show the performance of our method for different pose angles: roll, pitch and yaw. It can be seen from the results that the network is able to learn roll, and pitch information better than yaw. This experiment clearly shows that the hyperfeatures do contain enough information to estimate pose reliably.

Several qualitative results of our method on the AFW, AFLW and FDDB dataset are shown in Figure 7. 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.

4.5. Runtime

The Hyperface method was tested on a machine with 8 cores and GTX TITAN-X gpu. The overall time taken to perform all the four tasks was 4s per image. The limitation was not because of CNN, but due to selective search which takes approximately 3s to generate candidate region proposals. One forward pass through the Hyper-net takes only 0.2s.

5. Conclusions

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.

Acknowledgments

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.

Figure 6. Cumulative error distribution curves for pose estimation on (a) AFLW dataset. The numbers in legend show the mean error in degrees while estimating roll, pitch and yaw. (b) AFW dataset. The numbers in the legend are the percentage of faces that are labeled within ±15° error tolerance.
Figure 7. 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.

References

[1] A. Agarwal and B. Triggs. Multilevel image coding with hyperfeatures. *International Journal of Computer Vision*, pages 15–27, 2008. 1

[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

[3] 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. 3

[4] 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. 1, 2, 3, 6

[5] 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. 3

[6] 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. 3

[7] 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. 6

[8] 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. 6

[9] 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. 1, 3, 6

[10] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sept 2010. 3

[11] 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*. 4

[12] G. Gkioxari, B. Hariharan, R. B. Girshick, and J. Malik. R-cnn for pose estimation and action detection. *CoRR*, abs/1406.5212, 2014. 2

[13] 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. 2

[14] 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. 3

[15] P. Hu and D. Ramanan. Bottom-up and top-down reasoning with convolutional latent-variable models. *CoRR*, abs/1507.05699, 2015. 6

[16] 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. 5, 6, 7

[17] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UMCS-2010-009, University of Massachusetts, Amherst, 2010. 5, 6

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolu-
J. Li and Y. Zhang. Learning surf cascade for fast and accurate object detection with an ensemble of regression trees. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1867–1874, June 2014. 6

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. 4, 5, 6, 7

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. 3

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. 3, 7

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. 3, 6

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. 3, 6

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. 7

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. 3

S. Liao, A. Jain, and S. Li. A fast and accurate unconstrained face detector. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015. 3, 6

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In International Conference on Computer Vision, Dec 2015. 3, 5, 6, 7

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. 3, 6

I. Matthews and S. Baker. Active appearance models revisited. Int. J. Comput. Vision, 60(2):135–164, Nov. 2004. 3

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. 3

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. 1, 3, 6

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. 6

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. 2

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. 3

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. 7

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. 3

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. 3

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. 6

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. 4, 5

P. A. Viola and M. J. Jones. Robust real-time face detection. International Journal of Computer Vision, 57(2):137–154, 2004. 3

X. Xiong and F. D. la Torre. Global supervised descent method. In CVPR, 2015. 3

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. 3

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. 3

J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Face detection by structural models. Image and Vision Computing, 32(10):790–799, 2014. 5, 6

B. Yang, J. Yan, Z. Lei, and S. Z. Li. Convolutional channel features. In IEEE International Conference on Computer Vision, 2015. 3, 6

S. Yang, P. Luo, C. C. Loy, and X. Tang. From facial parts responses to face detection: A deep learning approach. In...
[48] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. CoRR, abs/1311.2901, 2013. 1, 3

[49] 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. 3, 7

[50] 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. 2, 3, 6

[51] 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. 1, 2, 3, 5, 6, 7

[52] X. Zhu and D. Ramanan. FaceDPL: Detection, pose estimation, and landmark localization in the wild. preprint 2015. 1, 2, 6, 7