Deep Features for Recognizing Disguised Faces in the Wild

Ankan Bansal Rajeev Ranjan Carlos D. Castillo Rama Chellappa
University of Maryland
College Park, MD, USA
{ankan,rranjan1,carlos,rama}@umiacs.umd.edu

Abstract

Unconstrained face verification is a challenging problem owing to variations in pose, illumination, resolution of image, age, etc. This problem becomes even more complex when the subjects are actively trying to deceive face verification systems by wearing a disguise. The problem under consideration here is to identify a subject under disguises and reject impostors trying to look like the subject of interest.

In this paper we present a DCNN-based approach for recognizing people under disguises and picking out impostors. We train two different networks on a large dataset comprising of still images and video frames with L2-softmax loss. We fuse features obtained from the two networks and show that the resulting features are effective for discriminating between disguised faces and impostors in the wild. We present results on the recently introduced Disguised Faces in the Wild challenge dataset.

1. Introduction

Face recognition performance on constrained datasets, like Labeled Faces in the Wild (LFW) [14], has improved significantly in the past few years due to the proliferation of deep convolutional neural networks. However, the performance on completely unconstrained datasets like IJB-A [21], Youtube Face (YTF) [45], and UMDFaces [3] continues to remain low at low false alarm rates. These datasets contain significant variations in view-point, pose, illumination, occlusion, resolution, age etc. This shows that the problem of face recognition is far from 'solved'. The recently announced Disguised Faces in the Wild (DFW) dataset and challenge [7, 24] aims to study another covariate of the face verification pipeline - 'disguises'.

Disguises and impersonations are part of a sub-field of face recognition where the subjects are non-cooperative and are actively trying to deceive the system. A disguise is defined as a means of altering one's appearance or concealing one's identity. This means that the subject is actively trying to adopt a new identity in order to hide his or her own. Similarly, an impersonation is the act of pretending to be another person. A subject might be trying to disguise his or her identity by adopting another identity or another person might be trying to impersonate the subject of interest.

This is an extremely challenging face verification problem. The aim of a face verification system in such cases is to be able to identify disguises and reject impersonators. Building such a system will be extremely helpful in law enforcement applications. The DFW challenge [7, 24] was introduced keeping such a target in mind.

In this paper, we present a deep learning-based face verification pipeline which is a step towards solving the challenging problem of disguised face recognition in the wild. We build an ensemble of two deep CNNs and achieve very good preliminary results for this task. Our approach could be a building block for future solutions to this problem. We use a large amount of data for training our models and report results on the DFW challenge test set [7, 24]. We build upon recent progress in face verification systems to solve this relatively less studied problem.

There are several factors to consider while designing a face verification system. One of these is the loss function used to train the deep networks. Most current methods use softmax loss for training their deep network. Softmax presents several advantages for training CNNs. It can be easily implemented using existing functions in various deep learning libraries [1, 6, 16] and does not have any restrictions on batch-size. It converges quickly. However, it is biased to the sample distribution in the training set. Unlike triplet loss, it does not specifically attend to hard samples. It fits well on high quality data and ignores the difficult samples in the mini-batch. To overcome this limitation, L_2-constrained Softmax Loss was introduced in [31]. It pushes samples from the same class closer and samples from different classes apart. In this work, we use the L_2-constrained softmax loss for training our networks.

With the increasing popularity of deep networks, large datasets are required for training these networks. Keeping this in mind, recently, several large-scale face recognition

Figure 1. Various disguises worn by Gary Oldman throughout his career as an actor. Recognizing people under disguises is clearly a challenging problem, even for humans. Designing autonomous systems for such a problem will be an important step towards complete face understanding.

datasets have been released publicly. MS-Celeb-1M [9] is a very large face dataset containing 10 million images of celebrities. However, there is a large amount of label noise in the dataset. Similarly, the CASIA-WebFace dataset contains about 500,000 images of celebrities. Again, it contains some label noise. Another large-scale dataset targeted towards training deep CNNs is the UMDFaces dataset [3]. The authors of this dataset claim that there is very little noise. The authors of [2] released a dataset of over 22,000 videos as an extension of the UMDFaces dataset. We combine a cleaned version of the MSCeleb dataset and the UMDFaces image and video frame dataset to create our training set, in this work. This training set contains about 5.6 million images of about 58,000 subjects.

The rest of the paper is organized as follows. In section 2 we briefly describe some related work. We present our method and results in section 3. Finally, conclusions and avenues for further research are presented in section 4.

2. Related Work

A standard face verification pipeline is shown in figure 2 and consists of the following steps: (i) Face detection; (ii) facial landmark detection and alignment; (iii) feature representation of a face; (iv) metric learning. We briefly review some recent related work in all these areas next.

2.1. Face Detection

Face detection is the process of localizing all faces present in an image. Typical face detection methods output bounding box coordinates for each face in an image. With the popularity of deep CNNs, several CNN-face detectors have been proposed in recent years [5, 17, 25, 29, 32, 33, 47, 50]. Many of these approaches begin with a region-proposal step which gives several hundreds to thousands of generic object proposals per image [17, 25, 32, 33]. Such object proposal generators include Selective Search [42] and Edge-Boxes [52]. Proposal-based face detectors are reminiscent of the proposal-based object detection methods [8, 34]. Other face detection methods [47, 50] are based on the Single Shot Detection (SSD) framework [26]. Since small faces in an image are particularly difficult to detect, some recent works have given specific attention to finding them [13, 29].

Significant improvements in unconstrained face detection performance have also been supported by the availability of large training datasets *e.g*. FDDB [15], WIDER [46]. The FDDB dataset consists of 2,845 images containing 5,171 faces. The WIDER face dataset is much larger and comprises of a total of 32,203 images. Both of these datasets contain faces with large variations in pose, illumination, scale, resolution, occlusion etc.

2.2. Landmark Localization

Fiducial keypoint localization is the next step in the pipeline. Such landmarks can include eye centers, nose top, mouth corners, ear lobe tips, chin, etc. These landmarks are used for aligning the detected faces *i.e*. to transform a given face into a canonical view of the face without losing the identity information. The authors in [2] showed that selecting good keypoints and a good face alignment method are important for achieving good verification performance. Several recently proposed facial landmark localization and face alignment approaches also use DCNNs. The all-in-one CNN [33] and Hyperface model [32] are multi-modal CNNs, which include keypoint localization as one of their modalities. Other CNN-based landmark localization models include [22, 23, 39].

The methods proposed in [23] uses a single DCNN for generating a unique keypoint descriptor which is used to localize keypoints on the face bounding box. In [22], the authors adopt a multi-modal framework and present an iterative method for landmark localization and pose predication using heat-map based DCNN regressors. These heat maps give the probability of the presence of keypoints at different locations. A cascade of deep CNNs was proposed by Sun *et al*. in [39] for effective landmark localization. The keypoints generated using any of the above mentioned approaches can be used as anchor points to transform the faces into canonical views.

Fiducial landmark localization and face alignment approaches can either use only $2D$ information from images [4, 49] or can incorporate $3D$ information for improved alignment [18, 19, 51]. The coarse-to-fine auto-encoder networks (CFAN) proposed in [49] cascade a few successive

Stacked Auto-encoder Networks (SANs). Successive SANs progressively refine the landmarks predictions by taking inputs from previous steps at higher and higher resolutions. Bulat *et al*. [4] proposed an approach which first performs facial part detection and provides confidence scores for the location of facial keypoints. In the second stage, these score maps are aggregated along with early CNN features for refining the predicted locations.

Unlike the methods described above which output only the $2D$ locations of fiducial landmarks, [18, 19] estimate both $2D$ and $3D$ landmarks and their $2D$ visibilities. The authors integrate a $3D$ point distribution model to design a cascaded coupled-regressor approach to estimate both the camera projection matrix and the $3D$ landmarks. Zhu *et al*. fit a dense $3D$ face model to an image using CNNs. They also propose a method to generate large-scale training samples in profile views.

2.3. Feature Representation

As shown in figure 2, the next step is to extract features from the aligned face images. Again, deep CNNs are currently the most popular and best performing methods for extracting features. These feature representations are usually learned using large training sets [2, 3, 9, 45, 48]. The requirement for such features is that the features for different images of the same subject should be close (in some metric) and the features for images of different subjects should be far.

The DeepFace system proposed by Taigman *et al*. [41] used $3D$ model based alignment along with deep CNNs for efficient feature learning. It used a 9-layer CNN with 120 million parameters. The system was trained on a privately held dataset of 4 million facial images of 4,000 identities. Sun *et al*. introduced the DeepID3 framework which consists of an ensemble of deep CNNs. These CNNs were trained on about 200,000 images of 10,177 subjects and the framework was the first to achieve super-human performance on the LFW test set. Unlike other deep learning methods, the FaceNet model [36] directly optimizes the embedding itself. The authors used face triplets generated using an online triplet mining approach for training the model. They used a private dataset of about 200 million images of about 8 million identities. Parkhi *et al*. [30] released a large face dataset of 2.6 million images and 2,600 identities along with a deep CNN trained on this dataset. They used the popular VGGNet [37] architecture and triplet embedding for face verification.

To learn more discriminative features, several works have used loss formulations other than the commonly used softmax loss. Wen *et al*. [44] introduced a new supervision signal called the center loss. The center loss simultaneously learns a center for each class and penalizes the distance between deep features of a class and its corresponding class center. This ensures that the deep features for a class are close to each other and far from deep features from other classes. A similar effect was achieved in [27], which proposed the angular softmax loss. This enables the CNNs to learn angularly discriminative features. Ranjan *et al*. [31] added an L_2-constraint on the feature descriptors to restrict them to lie on a hypersphere of a fixed radius. Their model achieved state-of-the-art performance on LFW [14] and very good performance on YTF [45].

2.4. Metric Learning

Metric learning is the process of learning a classifier or similarity measure from data and is an important step for enhancing performance of face verification systems. FaceNet [36] and VGGFace [30] embed the the DCNN features into a discriminative subspace by using triplet loss. Triplet loss was also used in [35] for learning a more discriminative low-dimensional embedding of a high dimension feature. Hu *et al*. [12] presented a discriminative deep metric learning method for face verification. In [38], the authors took full advantage of the training batches by lifting the vector of pairwise distances within the batch to the matrix of pairwise distances.

In addition to the steps described above, the datasets used for training the models play a crucial role in the process. Several large datasets targeted at face recognition and verification have been made publicly available in the past few years [2, 3, 9, 14, 20, 28, 30, 45, 48]. Along with datasets for training, significant attention has been paid to releasing new and challenging evaluation protocols [3, 14, 20, 21, 28, 45]. This is driving significant progress in different domains of face recognition and verification. Also, decisions about network architectures and pre-processing steps can impact the face verification performance significantly [2]. The current DFW dataset and challenge [7, 24] will be another step towards better face recognition systems.

3. Learning Deep face Representations

In this section, we describe our approach to face verification and present results on the DFW challenge dataset [7, 24]. We build an ensemble of two deep CNNs for the task. The two networks are trained on the same dataset. We give a brief description of the dataset used for training in section 3.1. Next, in section 3.2 we give a brief overview of the loss function used for training our networks. Then, we describe the architectures and training details of the two deep CNNs used in our method in section 3.3 and finally, in section 3.4, we report the results for the DFW challenge. An overview of a typical face verification pipeline is shown in figure 2.

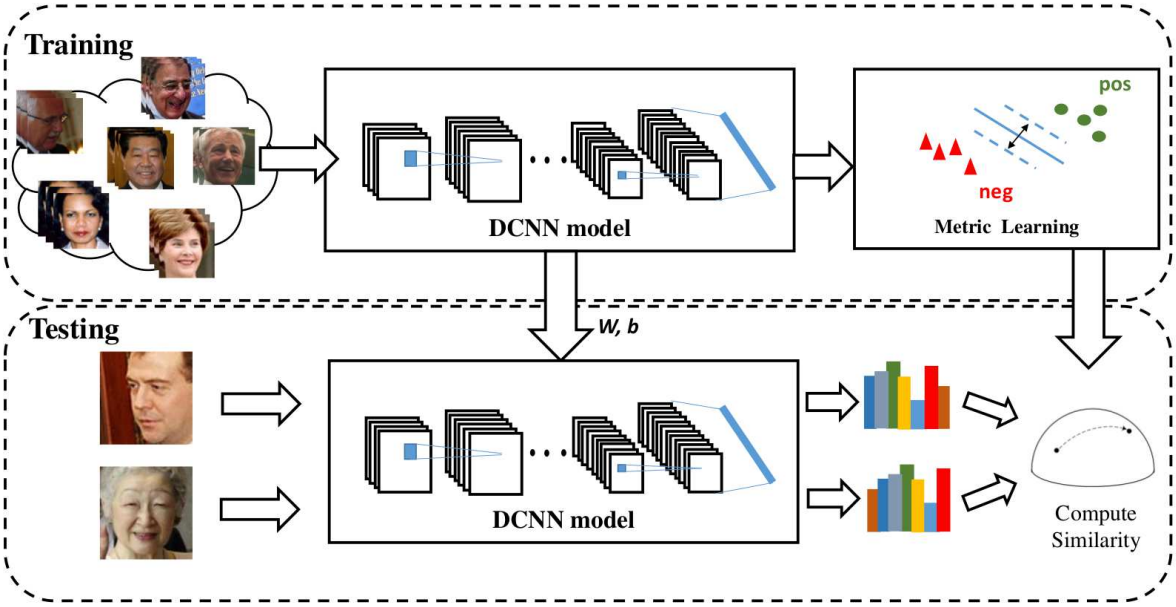

Figure 2. A typical face verification system pipeline. During training, a deep network is trained for classification using a large training dataset (*e.g*. UMDFaces [2, 3], MS-Celeb-1M [9]). After training the network, a metric learning framework (*e.g*. triplet embedding) is used to embed the features obtained from the deep CNN into a discriminative subspace. At test time, given two faces, the features from the deep CNN are computed and embedded into the embedding subspace. Finally, a similarity score (*e.g*. cosine similarity) is calculated between the two embedded features.

3.1. Dataset

We used the Universe Face dataset, which is a combination of curated MS-Celeb-1M [9], UMDFaces [3], and frames from UMDFaces-Videos [2]. We removed subject overlaps from the DFW dataset while creating the Universe Face dataset. This dataset contains about 3.5 million images from the curated MS-Celeb-1M dataset, about $300,000$ still images from the UMDFaces dataset, and about 1.8 million video frames from the UMDFaces-Videos dataset for a total of over 5.6 million face images and about $58,000$ identities. Combining data from various sources enables training of more robust networks. Also, using a combination of video frames and still photographs has been shown to improve generalization [2].

3.2. Loss Function

We now give a brief description of the L_2-constrained softmax (L2SM) loss function [31] used to train our networks. The aim behind L2SM is to improve the softmax loss to give high similarity scores for positive pairs and low similarity scores for negative pairs. L2SM adds an L_2-constraint on the feature descriptor, forcing them to lie on a hypersphere of a fixed radius. This has two advantages. First, it forces both good quality and bad quality images to have the same norm, unlike the softmax loss which gave good quality images a higher norm than poor quality images. This means that L2SM gives similar attention to both good and bad quality faces, unlike softmax. Second, L2SM forces the features from the same subject to be closer and features from different subjects to be far from each other. Therefore, it maximizes the margin netween positive and negative pairs.

L2SM [31] achieved state-of-the-art performance on the IJB-A challenge [21] and the LFW dataset [14]. It also achieved competitive performance on the Youtube Face (YTF) dataset. A more detailed description of the loss function can be found in [31].

3.3. Architectures

We describe the architectures of the two networks in our ensemble. We also give the training details and present the fusion algorithm for combining the outputs of the two networks.

Pre-processing: We use the all-in-one CNN [33] for face detection and alignment. We crop and resize each aligned face to each network's corresponding input size before sending them through the network. We applied a random horizontal flip as a data augmentation strategy.

ResNet

We use ResNet-101 [11] for training our face recognition system. The network contains 101 convolutional layers followed by a fully-connected layer of dimension 512. We use PReLU [10] activation function after every convolutional layer. We use the Universe Face dataset for training

the network. In total, the training data contains $57,779$ subjects and $5,554,906$ images. The network was trained using L_2-Softmax Loss [31] with α parameter set to 50. The initial learning rate was set to 0.1, which was reduced after every 50k iterations by a factor of 0.2. The training was carried out till $250,000$ iterations with a batch size of 128. We use the Triplet Probabilistic Embedding (TPE) [35] to learn a 128-dimensional embedding using images from UMDFaces [3] dataset.

Inception ResNet-v2

We adapt the Inception-ResNet-v2 model proposed in [40] for face recognition by removing the 1000 dimensional softmax layer and adding two fc layers on top - 512-D and $57,779$-D. This network has a total of 244 convolution layers. We use the L_2-constrained softmax loss [31] with α parameter 40. The model was trained for $120,000$ iterations with an initial learning rate of 0.1, which was reduced by a factor of 0.2 after every $50k$ iterations. We used a batch size of 120. For training, each of the aligned face is cropped and resized to the input size of the network ($299\times299\times3$). The total training time for the model was 4 days with 8 Nvidia Quadro P6000 GPUs. For final inference, we use TPE [35] to learn a 128-dimensional embedding using images from UMDFaces [3].

For fusion, we take the average of the scores obtained from the two networks as our final scores for each pair of images. More sophisticated fusion strategies will be explored in future.

3.4. Results

We first evaluate our approach on the relatively simple Disguised and Makeup Faces Database [43]. This dataset contains 2460 images for 410 identities. The images in this dataset are mostly celebrities with different disguises and makeups. Our method achieves significant performance improvements over the baseline results reported in [43]. The method achieves a true accept rate (TAR) of 92% at a false accept rate (FAR) of 0.0001, and a TAR of 96.4% at FAR 0.001. This shows that our method can recognize people with make-up and disguises with high confidence.

We then evaluate our approach on the recently announced Disguised Faces in the Wild (DFW) challenge [24]. The DFW challenge provides about $7,800$ test images for about 600 identities containing both disguises and impersonations. Each identity in the test set contains a normal image, some validation images, a few images with the subject in disguise, and a few images of impersonators *i.e.* other people who look like the subject under consideration. The aim of this challenge is to recognize disguised faces as belonging to the subject under consideration and reject the impersonators. The challenge follows a standard face verification evaluation strategy. Each pair in the test set is assigned a similarity score by the algorithm and has an associated ground-truth label ('positive', 'negative', or 'do not care'). The evaluation criterion is a standard *ROC* curve which plots the True Acceptance Rate (TAR) against False Acceptance Rate (FAR).

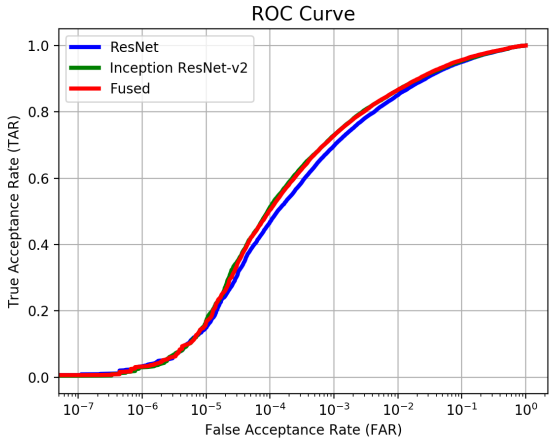

Figure 3. Results ROC

Feature	$FAR = 0.001$	$FAR = 0.01$
ResNet-101	70.0	85.1
Inception ResNet-v2	73.2	86.6
Fused	72.9	86.7

Table 1. TAR (%) at different FAR values for the three different features used in this work.

In this paper, we present results for our two networks separately and also for the combination strategy highlighted in section 3.3. All the results presented in this paper are self-generated using the evaluation codes provided by the organizers of the DFW challenge.

Figure 3 shows the ROC curves for both our networks and for the final fused scores. Table 1 gives the TAR values at $FAR = 0.01$ and $FAR = 0.001$ for the three kinds of features.

4. Conclusion

In this paper, we presented an approach for general face verification and evaluated it on the Disguised Faces in the Wild challenge. Recognizing disguised faces is an important practical problem for law enforcement and identity protection. We presented an ensemble of two deep CNNs trained on a large face dataset of about 5.6 million images. We showed that the proposed approach achieves promising preliminary results on this challenging problem and provides direction for future work for better face understanding.

Acknowledgement

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] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] A. Bansal, C. Castillo, R. Ranjan, and R. Chellappa. The dos and donts for cnn-based face verification. *arXiv preprint arXiv:1705.07426*, 5, 2017.

[3] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa. Umdfaces: An annotated face dataset for training deep networks. In *Biometrics (IJCB), 2017 IEEE International Joint Conference on*, pages 464–473. IEEE, 2017.

[4] A. Bulat and G. Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. 2016.

[5] D. Chen, G. Hua, F. Wen, and J. Sun. Supervised transformer network for efficient face detection. In *European Conference on Computer Vision*, pages 122–138. Springer, 2016.

[6] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS workshop*, number EPFL-CONF-192376, 2011.

[7] T. I. Dhamecha, R. Singh, M. Vatsa, and A. Kumar. Recognizing disguised faces: Human and machine evaluation. *PloS one*, 9(7):e99212, 2014.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[9] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[11] 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.

[12] J. Hu, J. Lu, and Y.-P. Tan. Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1875–1882, 2014.

[13] P. Hu and D. Ramanan. Finding tiny faces. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1522–1530. IEEE, 2017.

[14] 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, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[15] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. *University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009*, 2(7):8, 2010.

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[17] H. Jiang and E. Learned-Miller. Face detection with the faster r-cnn. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 650–657. IEEE, 2017.

[18] A. Jourabloo and X. Liu. Pose-invariant 3d face alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3694–3702, 2015.

[19] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4188–4196, 2016.

[20] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016.

[21] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1939, 2015.

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

[23] A. Kumar, R. Ranjan, V. Patel, and R. Chellappa. Face alignment by local deep descriptor regression. *arXiv preprint arXiv:1601.07950*, 2016.

[24] V. Kushwaha, M. Singh, R. Singh, M. Vatsa, N. Ratha, and R. Chellappa. Disguised faces in the wild. Technical report, IIIT Delhi, March 2018.

[25] Y. Li, B. Sun, T. Wu, and Y. Wang. Face detection with end-to-end integration of a convnet and a 3d model. In *European Conference on Computer Vision*, pages 420–436. Springer, 2016.

[26] 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.

[27] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song.

Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 2017.

[28] D. Miller, E. Brossard, S. Seitz, and I. Kemelmacher-Shlizerman. Megaface: A million faces for recognition at scale. *arXiv preprint arXiv:1505.02108*, 2015.

[29] M. Najibi, P. Samangouei, R. Chellappa, and L. Davis. Ssh: Single stage headless face detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4875–4884, 2017.

[30] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[31] R. Ranjan, C. D. Castillo, and R. Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017.

[32] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

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

[34] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[35] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa. Triplet probabilistic embedding for face verification and clustering. In *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*, pages 1–8. IEEE, 2016.

[36] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[38] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 4004–4012. IEEE, 2016.

[39] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3476–3483. IEEE, 2013.

[40] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.

[41] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[42] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[43] T. Y. Wang and A. Kumar. Recognizing human faces under disguise and makeup. In *Identity, Security and Behavior Analysis (ISBA), 2016 IEEE International Conference on*, pages 1–7. IEEE, 2016.

[44] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[45] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 529–534. IEEE, 2011.

[46] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5533, 2016.

[47] S. Yang, Y. Xiong, C. C. Loy, and X. Tang. Face detection through scale-friendly deep convolutional networks. *arXiv preprint arXiv:1706.02863*, 2017.

[48] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[49] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision*, pages 1–16. Springer, 2014.

[50] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S^3fd: Single shot scale-invariant face detector. *arXiv preprint arXiv:1708.05237*, 2017.

[51] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2016.

[52] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, pages 391–405. Springer, 2014.