2018 13th IEEE International Conference on Automatic Face & Gesture Recognition

Deep Unsupervised Domain Adaptation for Face Recognition

Zimeng Luo[1], Jiani Hu[1], Weihong Deng*[1], Haifeng Shen[2]
[1]School of Information and Communication Engineerin, Beijing University of Posts and Telecommunications
[2]AI Lab, Didi Chuxing
{luozimeng, jnhu, whdeng}@bupt.edu.cn shenhaifeng@didichuxing.com

Abstract— **Face recognition is challenge task which involves determining the identity of facial images. With availability of a massive amount of labeled facial images gathered from Internet, deep convolution neural networks(DCNNs) have achieved great success in face recognition tasks. Those images are gathered from unconstrain environment, which contain people with different ethnicity, age, gender and so on. However, in the actual application scenario, the target face database may be gathered under different conditions compered with source training dataset, e.g. different ethnicity, different age distribution, disparate shooting environment. These factors increase domain discrepancy between source training database and target application database and make the learnt model degenerate in target database. Meanwhile, for the target database where labeled data are lacking or unavailable, directly using target data to fine-tune pre-learnt model becomes intractable and impractical. In this paper, we adopt unsupervised transfer learning methods to address this issue. To alleviate the discrepancy between source and target face database and ensure the generalization ability of the model, we constrain the maximum mean discrepancy (MMD) between source database and target database and utilize the massive amount of labeled facial images of source database to training the deep neural network at the same time. We evaluate our method on two face recognition benchmarks and significantly enhance the performance without utilizing the target label.**

I. INTRODUCTION

With availability of a massive amount of labeled images gathered from Internet, deep neural networks have significantly improved the performance in many computer vision applications, such as object detection [3], [15], object classification [7], [17], face recognition [11], [19], [16], [8], [9] and so on. Most of typical techniques are to train a deep neural network with massive images and then apply it to target test dataset. These methods are valid when the training data and test data are independently and identically drawn from the same or similarity distribution. However, in actual application scenario, the distribution of target and training data is always dissonant, which degenerates model performance on target test data. To accommodate the distribution of target data and enhance model performance, one of the most direct approach is to fine-tune a pre-trained deep neural network's parameter on target database with the supervision of data label. This strategy turns out to be problematic for a target task where labeled data is lacking or even unavailable. Meanwhile, the deep neural networks easily suffer from a significant amount of over-fitting under supervision of small amount of labeled data, which usually degenerates the generalization ability of the model.

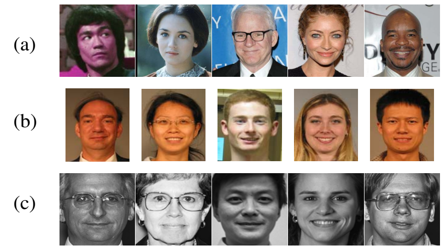

Fig. 1. We show several sample images of three face databases. (a) is CASIA-WebFace [22]. (b) is GBU face challenge dataset [13], (c) is FERET [14].

This problem also exists in the face recognition task. Most of deep models such as VGG-face [11], Facenet [16], SphereFace [8] et al. , are firstly trained on a large scale face database and then evaluated on other face databases like LFW [4], Youtube [20], MegaFace [6] et.al. These databases are all gathered from Internet for the convenience of data collection. However, in real world applications, the target test data may contain people with specific ethnicity, age group, gender, imaging quality, pose of faces etc. and the shooting environment of target test data and source training data may vary greatly. These factors increase domain discrepancy and degenerate face recognition performance on target application. As shown in Fig. 1, there are significant differences in the picture of different databases. The images in CASIA-WebFace [22] are collected from Internet under unconstrained environment and most of the figures are celebrities and public. The GBU [13] contains still frontal facial images acquired with digital camera. FERET [14] is collected under constrained environment and the pictures are all gray. Different data collection methods and application environments cause a significant discrepancy between different databases. Our experiment results also show a poor performance on target test database when directly adopting the pre-trained model.

The approaches which are proposed to address the challenge of discrepancy between training data and test data are often referred to domain adaptation. Generally, the databases with large scale of labeled data are called source domain while the databases with little or no labeled data are called target domain. Source domain and target domain usually share the same task but with different distribution. Depending

*Corresponding author

978-1-5386-2335-0/18/$31.00 ©2018 IEEE
DOI 10.1109/FG.2018.00073

on whether the label information is available for target data, domain adaptation can be categorized into supervised or unsupervised domain adaptation. In this work, we focus on unsupervised domain adaptation in face recognition problems. The main contribution of this paper is applying the domain adaptation approaches to face recognition problems when the target test dataset exists huge distribution discrepancy compared with training data. We explore the training methods of unsupervised domain adaptation without utilizing the label of target data in deep neural networks, which significantly alleviates the discrepancy between source and target face database and enhances model performance on target test data.

II. RELATED WORK

Recently, all the top performing methods for face recognition were all based on DCNN architectures. Facenet [16] presented a triplet embedding loss to learn a mapping from face images to a compact Euclidean space. The loss aimed to separate the positive pair from the negative by a distance margin. VGG-face model [11] is a typical application based on VGG architectures [17]. It was trained on a large scale datset of 2.6M images from 2622 subjects. [8] proposed a angular softmax loss to learn angularly discriminative features on a hypersphere manifold. These method were all focus on utilizing a massive amount of labeled facial images to train a DCNN of strong generalization ability and testing on common benchmarks. All these approaches obtained excellent performance on many benchmarks e.g. LFW [4], YTF [20], MegaFace [6], which were also collected from the Internet like the training data.

Deep learning architectures have been explored for domain adaptation problems and obtained significant performance gains. [23] comprehensively explored the transferability of deep neural networks. It focused on the scenario where a sufficient labeled data was available in target domain. However, in practical scenario, labeled target data is usually limited or unacquirable. To address this issue, many unsupervised domain adaptation approaches were proposed. [18] proposed a model that automatically learnt a representation jointly trained to optimize for classification and domain invariance. A single-kernel MMD was adapted as domain discrepancy metric and only one layer was selected to transfer. Our work is primarily motivated by [10], which further proposed a novel deep neural network architecture for unsupervised domain adaptation and transferred all the layers corresponding to task-specific features by utilizing multi-kernel MMD. [21] designed a weighted MMD model to alleviate the effect of class weight bias in domain adaptation. This work was restricted to the situation that source domain and target domain shared the same categories, which is unsuitable for open-set face recognition problems. [5] proposed a Bi-shifting Auto-Encoder network, which could shift source domain samples to target domain and also shift the target domain samples to source domain. This method solved many special cross domain recognition problems in face recognition scenarios such as cross view angle recognition, cross ethnicity recognition and cross imaging sensor recognition.

III. METHOD

A. Maximum Mean Discrepancy

In the field of domain adaptation, maximum mean discrepancy(MMD) has been widely adopted as a standard distribution distance metric to measure the discrepancy between source and target domains. Assuming that there are two datasets *S*(represents source data) and *T*(represents target data), the data are firstly mapped into a reproducing kernel Hilbert space (RKHS) using function $\phi(\cdot)$. Let Φ be a class of function $\phi(\cdot)$. The MMD between *S* and *T* is defined as (1),

$$MMD^2[\Phi, S, T] := \left[\sup_{\phi\in\Phi}\left(\mathbf{E}_S[\phi(x^S)] - \mathbf{E}_T[\phi(x^T)]\right)\right]^2 \quad (1)$$

E represents the expectation with regard to the distribution. We have $MMD[\Phi, S, T] = 0$ when *S* and *T* share the same distribution based on the statistic tests defined by MMD. Let $x^S = (x_1^S, ..., x_M^S)$ and $x^T = (x_1^T, ..., x_N^T)$ denote two set of samples drawn from *S* and *T*, respectively. MMD can be estimated empirically given by (2)

$$MMD^2[\Phi, S, T] := \left[\sup_{\phi\in\Phi}\left(\frac{1}{M}\sum_{i=1}^{M}\phi(x_i^S) - \frac{1}{N}\sum_{i=1}^{N}\phi(x_i^T)\right]^2 \quad (2)$$

$\phi(\cdot)$ is the mapping to the RKHS, and $k(\cdot,\cdot) = \langle\phi(\cdot), \phi(\cdot)\rangle$ is the universal kernel associated with this mapping Generally, $k(\cdot,\cdot)$ is defined as the convex combination of m kernels $\{k_i\}$ $i \in [1, m]$,

$$k = \sum_{i=1}^{m}\beta_i k_i, s.t. \sum_{i=1}^{m}\beta_i = 1, \beta_i \geq 0 \quad (3)$$

$\{\beta_i\}$ is the coefficients of ith kernel.

Obviously, as the definition above, the calculation of MMD does not require the category information of source data or target data. Hence, the MMD can be utilized in unsupervised domain adaptation when the target labeled is lacking or unavailable.

B. Deep Unsupervised Domain adaptation

Supervised CNNs based feature representations have been shown to be extremely effective for many computer vision tasks. However, as the unavailability of target labeled data, directly adapting CNNs to the target domain via fine-tuning is impractical. To address this issue, training a joint source and target CNN architecture is adapted in our approaches. The joint architecture we used is as Fig. 2 shows. A massive amount of source labeled facial images are utilized to guarantee the generalization ability of deep CNNs through training a classification network. Meanwhile, the domain loss minimizes the domain discrepancy by constraining the MMD between source data and target data.

According to the observation of [23], the transfer ability drops in higher layers with increasing domain discrepancy and transfer learning method would obtain better performance when transferring higher layers of the deep neural network. Selecting suitable transfer layers can significantly enhance the transfer efficiency. In order to better measure the domain discrepancy and enhance transfer efficiency, we simultaneously transfer several higher layers of deep neural network as [10] mentioned, which is called multi-layer adaptation. We extend the VGG-Face architecture proposed in [11] and adopt a multi-layer adaptation regularizer based

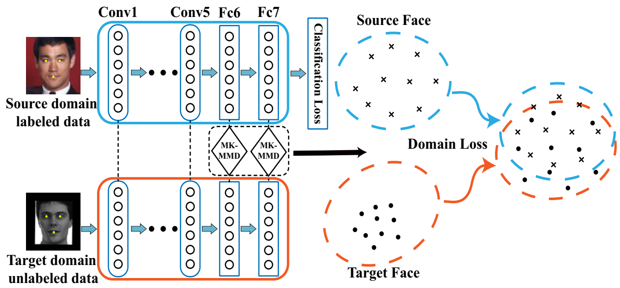

Fig. 2. The joint deep neural network architecture for unsupervised domain adaptation. The inputs of the upper network are source labeled images while the lower are target unlabeled data. All the source face and target face are aligned to the same reference point(the yellow dotes in the facial images). The blue dotted oval represents the source domain distribution and the crosses inside donate source face. The orange dotted oval represents the target domain distribution and the dotes inside donate target face. Domain loss aims at minimizing the distribution discrepancy of two domains.

on MMD on layer *FC6* and *FC7*. The loss function is as (4) shows,

$$L = L_c(x^s, y^s) + \lambda \sum_{l=l_1}^{l_2} MMD_k^2(S,T) \tag{4}$$

where L_c is the classification loss function on the source labeled data and we adopt softmax loss function in our experiments. x^s represents the source data and its ground truth label is y^s. MMD_k^2 denotes the distance between the source data, S, and the target data, T. In the architecture of VGG-FACE, l_1 and l_2 respectively represent layer *FC6* and *FC7*. λ is a penalty parameter to balance the contribution of classification loss and domain discrepancy loss. If λ is set too low, MMD regularizer may have no effect on eliminating domain discrepancy while if λ is set too high, the heavily regularization may get all points in feature space together. In order to better measure the domain discrepancy, we adopt a multi-kernel-MMD as the definition of (3).

IV. Experiment

In this section, we evaluate our unsupervised domain adaptation methods on two famous face recognition benchmarks. We will begin with introducing the detailed information and evaluation protocol of the datasets we utilized, followed by illustrating the training details of our experiments and reporting result of on two face recognition benchmarks.

A. Dataset and Evaluation Protocol

CASIA-WebFace: CASIA-WebFace dataset is a large scale face dataset gathered from Internet. It contains 10,575 subjects and 494,414 images. The large scale of labeled facial data does great help to train DCNNs. In our experiments, we adopt this dataset as the source domain data for training the classification network.

GBU: Its full name is *The Good, the Bad, and the Ugly Face Challenge*. This dataset consists of three partitions. Different partitions contain pairs of images with different difficulty levels based on the performance of three top performers in the FRVT 2006. The Good partition consists images which are easy to match. The Bad partition contains pairs of images of average difficulty to recognize. The Ugly partition contains pairs of images considered difficult to recognize. Each of partitions consists of a target set and a query set. Both target set and query set contain 1085 images for 437 distinct people. Following the evaluation protocol of [13], we use ROC analysis and compare the the verification rate for each partition at a FAR(false accept rate) of 0.001 of different algorithms. We utilize part of images from FRGC [25] as the target training data, which consists of 19270 still front faces. The subjects in target training set do not appear in the three partitions.

FERET: The FERET database consists of a total of 14051 gray-scale images representing 1,199 individuals. The facial images in FERET are divided into five set, respectively is *fa*(used as a gallery set), *fb*(images with different expression), *fc*(images with different illumination), *dup I*(the images were taken later in time) and *dup II*(the images were taken at least a year after the corresponding gallery images). The recognition rates at rank 1 are reported in our experiments.

B. Training Details

As Fig. 2 shows, all the images of different datasets are aligned to the same reference point using three facial landmarks(left eye, right eye and center of mouth). The images are firstly resized to 250×250 and then a random crop of 224×224 is fed to DCNNs. We also augment the data by flipping it horizontally with 50% probability. The VGG-face model was train on 2.6 millon facial images. The reported results on LFW and YTF shown the excellent performance of this model. However, we know nothing about the face aligned method in [11]. Inconsistent alignment methods between training data and test data may cause a poor performance while testing. To address this issue, we firstly use the CASIA-WebFace dataset to train classification network based on pre-train VGG-FACE model. The CASIA-WebFace dataset and other target datasets share the uniform alignment methods as we mentioned before. As the last classifier is trained from scratch, we set its learning rate to be 10 times that of the lower layers. The based learning rate is fixed at 10^{-4}. The batch size is set to 32 and the network is trained for 2×10^4 iterations.

After training the basic classification networks on CASIA-WebFace, we add the target data and joint classification loss and domain loss to train the domain adaptation network. As Fig. 2 shows, our network architecture is comprised of two basic DCNNs which are identical in structure and shared by parameters. One is for classification on source data and the other is for feature calculating on target data. We extend the VGG-Face [11] network as our basic architectures and adopt the model we trained on CASIA -WebFace before to initialize network parameters. The preprocessing of target data is the same as source data we mentioned before. At this stage, the learning rate of all layers is fixed to 10^{-4}.

The kernel we use in MMD is Gaussian kernel $k(x_i, x_y) = e^{-\|x_i - x_y\|^2/\gamma}$ where γ donates the bandwidth. In our experiment, we use both single-kernel and multi-kernel MMD as the domain discrepancy metric. When using multi-kernel M-MD, we consider five Gaussian kernels by setting bandwidth to $\gamma_m \cdot (1, 2^1, 2^2, 2^3, 2^4)$, respectively, where γ_m is set to the median pairwise distances [24] on training data. When adopting single-kernel MMD, we only utilize one Gaussian kernel which bandwidth is set to γ_m. To be fair, the hyper-parameter λ in (4) is fixed at 0.5 in all experiments. To evaluate the effectiveness of multi-layer and multi-kernel adaptation, we make several variants of the deep adaptation networks, respectively are $\mathbf{DAN_{s7}}$ (using single-kernel and transferring *FC7* layer only), $\mathbf{DAN_{s6,7}}$ (using single-kernel and transferring both *FC6* and *FC7* layer), $\mathbf{DAN_{m7}}$ (using multi-kernel and transferring *FC7* layer only) and $\mathbf{DAN_{m6,7}}$ (using multi-kernel and transferring both *FC6* and *FC7* layer).

C. Experiment Results

In the experiment of GBU challange dataset, we report the VR(verification rate) for three partitions, the Good, the Bad and the Ugly at a FAR(false accept rate) of 0.001. We evaluate the performance of the four variants of deep adaptation networks and compared them with other methods. *Fusion-baseline* in [12] donated the the FRVT 2006 fusion algorithm and the result *VGG-face* was reported in [12] by utilizing the pre-trained VGG-face model. The *VGG-CASIA* represents the model we trained in stage 1 on CASIA-WebFace dataset and we choose it as our baseline. The exact results are shown in Table I.

TABLE I
VR AT FAR = 0.001 FOR GBU PARTITIONS.

Method	**Ugly**	**Bad**	**Good**
Fusion-baseline[12]	15.00%	80.00%	98.00%
VGG[12]	26.00%	52.00%	85.00%
VGG-CASIA	48.80%	73.55%	95.57%
DAN_{s7}	60.78%	84.95%	98.00%
$DAN_{s6,7}$	63.42%	87.08%	98.54%
DAN_{m7}	68.42%	87.68%	98.67%
$DAN_{m6,7}$	**69.42%**	**88.87%**	**98.93%**

As the result reported in Table 1, the baseline model trained on CASIA-WebFace only obtains much better performance compared the result reported in [12]. The result suggests that the uniform face aligned algorithm of source domain and target domain images is the key to ensuring model performance in the face recognition problem. The $DAN_{m6,7}$ performs the best with an improvement up to 20.62% on the Ugly partition, 14.13% on the Bad partition and 3.1% on the Good partition compared with our baseline. The models trained with single-kernel MMD obtain a little bit worse results compared with multi-kernel MMD. We can observe that the unsupervised domain adaptation network can significantly improve the performance on target dataset and the multi-kernel MMD combined with transferring multi-layer is more beneficial to alleviate the domain discrepancy. We further present the ROC curve in Fig. 3. As the the trend of the ROC curve shows, the performance of deep adaptation network is superior on all three partitions.

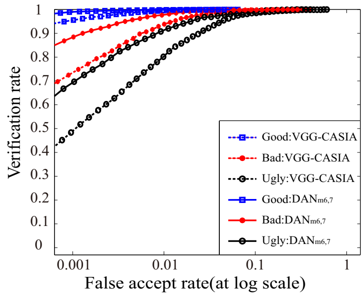

Fig. 3. ROCs on the GBU partitions. ROCs are shown for the VGG-CASIA baseline and the unsupervised domain adaptation algorithms.

In the experiments of FERET, we report the recognition rates of four sets and their average. The exact results are shown in Table II. We firstly reported two classical methods performance on FERET. One is LBP[1], which designed facial image representation based on local binary pattern (LBP) texture features. [2] proposed a simple deep networks based on principal component analysis (PCA) for image classification. We still choose the model trained on CASIA webface as the baseline in this experiment. After the unsupervised domain adaptation training, we obtain a bit of performance boost compared with our baseline and other approaches, especially in the subsets of dup I and dup II. Similar to the experiment on GBU, the network $DAN_{m6,7}$ obtains the best accuracy of 98.70% on average.

TABLE II
RECOGNITION RATES ON FERET.

Method	**fb**	**fc**	**dup I**	**dup II**	**Avg.**
LBP[1]	97.00%	79.00%	66.00%	64.00%	81.00%
PCANet[2]	99.58%	**100%**	95.43%	94.02%	97.26%
VGG-CASIA	99.67%	98.45%	95.15%	95.72%	97.25%
DAN_{s7}	99.74%	98.96%	97.09%	97.43%	98.31%
$DAN_{s6,7}$	**99.75%**	98.97%	96.40%	97.00%	98.03%
DAN_{m7}	99.58%	99.48%	96.54%	97.01%	98.15%
$DAN_{m6,7}$	99.67%	99.48%	**97.37%**	**98.29%**	**98.70%**

V. CONCLUSIONS

In this paper, we focus on the issue of domain discrepancy between source training data and target test data in face recognition scenario. We adopt a deep unsupervised domain adaptation neural network and jointly utilize the labeled large scale source data and unlabeled target data to alleviate the domain discrepancy. We show the transferability between source face and target face by the multi-kernels MMD constraining on multi-layers representation. Empirical results show that the method can significantly enhance model performance on target test data without utilizing the label information.

VI. ACKNOWLEDGMENTS

This work was partially supported by the National Natural Science Foundation of China under Grant Nos. 61471048, 61573068, and 61375031, and Beijing Nova Program under Grant No. Z161100004916088.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041, 2006.

[2] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma. Pcanet: A simple deep learning baseline for image classification? *IEEE Transactions on Image Processing*, 24(12):5017–5032, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904, 2015.

[4] 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.

[5] M. Kan, S. Shan, and X. Chen. Bi-shifting auto-encoder for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3846–3854, 2015.

[6] 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.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*, pages 1097–1105, 2012.

[8] 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)*, July 2017.

[9] B. Chen, W. Deng, and J. Du. Noisy softmax: Improving the generalization ability of dcnn via postponing the early softmax saturation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[10] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015.

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

[12] P. J. Phillips. A cross benchmark assessment of a deep convolutional neural network for face recognition. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 705–710. IEEE, 2017.

[13] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer. The good, the bad, and the ugly face challenge problem. *Image and Vision Computing*, 30(3):177–185, 2012.

[14] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, 16(5):295–306, 1998.

[15] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137, 2016.

[16] 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.

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.

[18] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[19] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. *A Discriminative Feature Learning Approach for Deep Face Recognition*. Springer International Publishing, 2016.

[20] 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.

[21] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

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

[23] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[24] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213, 2012.

[25] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 947–954, 2005.