OPEN SOURCE FACE RECOGNITION PERFORMANCE EVALUATION PACKAGE

Xiang Xu and Ioannis A. Kakadiaris

Computational Biomedicine Lab
University of Houston
4849 Calhoun Road, Houston, TX, 77204

ABSTRACT

Biometrics-related research has been accelerated significantly by deep learning technology. However, there are limited open-source resources to help researchers evaluate their deep learning-based biometrics algorithms efficiently, especially for the face recognition tasks. In this work, we design and implement a light-weight, maintainable, scalable, generalizable, and extendable face recognition evaluation toolbox named FaRE that supports both online and offline evaluation to provide feedback to algorithm development and accelerate biometrics-related research. FaRE consists of a set of evaluation metric functions and provides various APIs for commonly-used face recognition datasets including LFW, CFP, UHDB31, and IJB-series datasets, which can be easily extended to include other customized datasets. The package and the pre-trained baseline models will be released for public academic research use after obtaining university approval.

Index Terms— Face Recognition, Evaluation, Toolbox

1. INTRODUCTION

Face recognition (FR) is a general concept for the applications of face identification and verification. In the scenario of face identification, the system classifies human identity according to a single facial image or a set of images while in the scenario of face verification, a binary decision is made by the computer to decide whether two images belong to the same identity. Recently, the deployment of deep neural network has produced impressive results in both face identification and verification tasks. Methods such as ArcFace [1] have significantly pushed the frontier of face recognition performance. Figure 1 depicts a general face recognition system consists of three stages including enrollment, matching, and evaluation stages. During the enrollment, the algorithm should generate a template from a facial image or a set of images for each subject. Most of the current research work is focusing on learning to generate a discriminative template for each identity. In the matching stage, the distance or similarity score is computed and a decision algorithm determines the identity in an identification application or to accept/reject the person in a verification scenario. In the evaluation stage, the overall
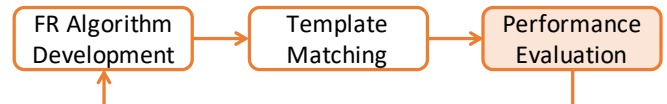
Fig. 1. Depiction of the general FR development pipeline consists of FR algorithm development (template generation), template matching, and performance evaluation. Performance evaluation can provide feedback to algorithm development. Our work fits in the performance evaluation block.

face recognition performance is assessed using quantitative measurements. However, in the biometrics community, there are limited existing resources that provide a consistent and general measurement for evaluating and developing a deep learning-based face recognition system. This need becomes more acute, especially for the on-boarding of new researchers.

In this work, we design and implement a light-weight, maintainable, scalable, generalizable, and extendable face recognition evaluation package named FaRE, which can be generalized to the data and algorithms developed from other modalities such as fingerprint, *etc.*. The whole package, which is easy to install, is written in Python since most of the current deep learning frameworks support Python language. We adopt the most commonly-used FR datasets, analyze their metrics, and then generalize an evaluation pipeline to evaluate the performance of FR algorithms. To support offline evaluation, a file management module is implemented to organize and match the generated template file with meta-data for each dataset. To support online evaluation, data loaders are implemented to feed the data to the neural network and generate a template from a facial image or an image set. The similarity matrix is obtained by computing the similarity of the templates from probe and gallery set based on the evaluating dataset protocol. Based on the similarity matrix and the ground-truth label provided in datasets, different quantitative measurement functions are used according to the protocols provided in the datasets. To visualize the quantitative results, comparison figures can be plotted using FaRE. With our evaluation package, the new datasets and protocols can be easily extended and evaluated. In addition, new fusion func-

tions can be easily added for set-based FR. FaRE is being used to validate the biometrics algorithms developed in the UH Computational Biomedicine Lab (CBL). To provide the pre-trained models for the biometrics community, two face template generators are trained using ResNet-101 [2] and DenseNet-121 [3] architectures on VGG-Face2 [4] dataset. The FR experiments including open-set face identification and face verification are performed and are evaluated by FaRE. In summary, our contributions are two-fold: (i) A light-weighted, maintainable, scalable, generalizable, and extendable FR package is designed and implemented; (ii) Two networks are trained and evaluated on IJB-C, which can serve pre-trained models in other face-related tasks.

2. RELATED WORK

FR Algorithms, Datasets, and Protocols: A significant part of the success of recent FR algorithms can be attributed to the large-scale image collections [4] that have become available in the past few years. Researchers improve the FR performance by deploying the generative model to generate the frontal images [5] or developing new loss functions [6, 1] to learn discriminative face representations. To evaluate the improvement of unconstrained FR algorithms, several benchmarks [7, 8, 9] with new protocols have been proposed. Unlike other face-related tasks such as detection [10], alignment [11, 12], reconstruction [13], and soft-biometrics [14], FR tasks can be evaluated with various metrics and protocols for the different scenarios. Some datasets [15, 16] designed for face verification task are using a verification protocol comparing Receiver Operating Characteristic (ROC) curve or Precision-recall (PR) curve. A closed-set dataset named UHDB31 [17] is collected and designed for evaluating FR performance in the presence of variations of pose and illuminations. Recently, the IJB-series datasets [7, 8, 9] has become available for open-set-based FR by adding more identities and variations. As opposed to closed-set protocol, open-set FR protocols contain the novel identities in the probe but not in the gallery set in the face identification scenario. ROC curves are reported in the 1:1 verification protocols while Cumulative Match Characteristic (CMC) curves and Decision/Identification Error Trade-Off (DET/IET) curves are reported in the 1:N mixed recognition protocols for both still images and images captured from video.

FR Systems and Evaluation: OpenBR [18] is a well known open-sourced computer vision and pattern recognition library in the biometrics community, which contains several FR algorithms. However, these algorithms only work well on frontal controlled faces and the evaluation provided in OpenBR is not user-friendly. Bob [19, 20] is another toolbox that aims to reproduce the research in signal processing and machine learning, but is difficult to install due to a large number of dependencies. A C++ based 3D-aided 2D FR was proposed by Xu *et al.* [21], which used an estimated 3D model to frontal-

ize the 2D images and matched the templates generated from the local patches according to the occlusion masks, significantly improving the performance on face images with pose variations. However, that work only focuses on the facial template enrollment and matching stages, which does not provide a systematic evaluation package.

In summary, all these works either only provide the pre-trained model or contain some evaluation functions, which cannot be generalized to other face recognition scenarios for other researchers use.

3. FACE RECOGNITION EVALUATION PACKAGE

To help researchers obtain quick feedback from evaluation and accelerate research process, we design and implement a *light-weight*, *extendable*, *generalizable*, *scalable*, and *maintainable* face recognition evaluation package, which can be easily generalized to evaluate other biometrics applications. In this section, we first present a quick overview of the design and implementation of FaRE package. Then, we describe and analyze each main component module in this package and the advantages in detail.

Overview: Figure 2 illustrates the main component modules, functionality, and the designed system architecture of FaRE. Considering the generalization for commonly-used face verification datasets such as LFW and CFP and face identification datasets such as IJB-A, IJB-B, and IJB-C, two main protocols are defined in these datasets: a comparison protocol for face verification and a search protocol for face identification. We abstract the protocols into three parts: comparison protocol, closed-set protocol, and open-set protocol. Each protocol calls their intrinsic metric functionality to measure FR performance. The datasets consist of the generated templates from online training or loaded templates in offline mode and the corresponding labels. Therefore, a custom dataset can be easily extended by inheriting the existing dataset, which mainly requires to feed the templates and labels into the system. In addition, to fit the set-based FR, a template is defined and a custom template fusion function can be easily added to generate one template from a set of feature vectors. As depicted in Fig. 2 (b), to organize the files and templates in datasets, some classes are defined for managing the data or meta-data. On top of the system, the users can easily call the dataset wrapper and perform the evaluation.

Metrics: As one of the basic functions defined in FaRE, metrics class defines and manages several commonly used metrics including ROC, PR, Accuracy (ACC), and Equal Error Rate (EER) for face verification comparison protocol, CMC and DET/IET for face identification search protocol. These functions perform the main job of computing the quantitative metrics for researchers, which summarized in Table 1.

Protocols: With pre-defined metric functions, in the comparison protocol, the system considers the ground-truth labels and the similarity vectors as input while the system requires the
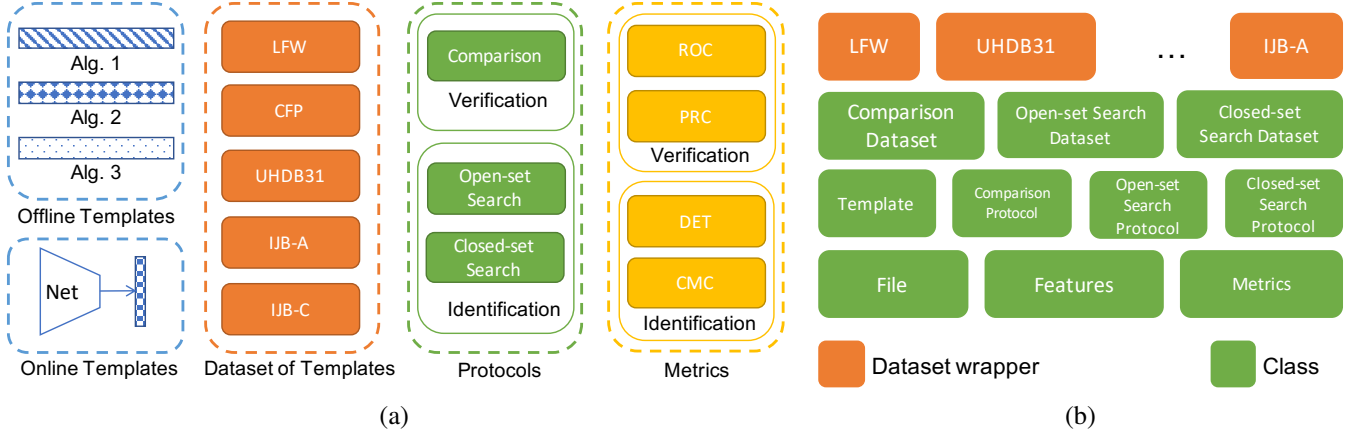
Fig. 2. (a) Depiction of different main component modules in FaRE: To support both offline and online evaluation, the templates and metadata should be mapped according to the datasets' protocol. The system can compute the similarity matrix and generate quantitative results according to the corresponding protocol used in the dataset. (b) Depiction of the system architecture of FaRE. The system is implemented in a bottom-to-up manner. The details can be found in Sec. 3.

Protocol	Implemented Metrics
Verification / Compare	ROC, PR, ACC, EER, AUC
Closed-set Identification / Search	CMC
Open-set Identification / Search	CMC, DET/IET

Table 1. Summary of the current supported metrics in FaRE

ground-truth labels and similarity matrix in the search protocol. The search protocol using includes both the closed-set protocol and an open-set protocol. In the closed-set protocol, the identities in the probe are assumed to be within the identities set in the gallery, forcing the system to assign a label from the gallery to the testing probe according to similarity ranking. In the open-set protocol, the identities in the probe might be out of the range of identities in the gallery, which allows the system the ability to reject some samples based on their similarity scores and defined threshold.

Datasets: Some dataset APIs are implemented and provided for users to quickly evaluate their algorithm on common-used datasets (LFW, CFP, IJB-A, IJB-B, and IJB-C) with different purposes. Each dataset supports both offline and online evaluation: In the offline mode, the dataset will load the features from the disk and compute the similarities. To evaluate the training process of the deep neural network, several data loaders are implemented to load the image data and forward them to the trained network to obtain the templates.

Light-weight: Unlike other libraries such as Bob [19], our package is implemented in Python and only requires a few basic dependencies such as numpy for array operation, matplotlib for visualization, scikit-learn [22] for metric computing, and MXNet [23] for deep learning. Therefore, FaRE is a light-weight package because it only requires few dependencies and it is implemented by a limited number of codes, which makes FaRE extremely easy to install.

Extensibility: FaRE features four extensibility aspects: adding new template fusion functions, new metrics, new protocols, and new datasets. In set-based FR, a common way is to compute the mean feature vectors or assign different weights to compute weighted average feature vectors as the template for a set, which are implemented in FaRE. In addition, it supports adding new template fusion functions to fuse the features from a set of images and new metrics functions to compute new quantitative measurement. Extending current protocols or datasets is suggested to inherit the corresponding super-class and adjust the protocol process based on customized requirements, which can be quickly extended.

Generalization ability: The generalization ability we define here is that the system can incorporate to different datasets, running modes, and template generators. The package is abstracted to fit the requirements of various datasets and template generators. FaRE supports both online and offline performance evaluation. The online evaluation mode can be used in validating the training process while the offline evaluation mode is designed for evaluating existing algorithms.

Scalability: The system can process one image or a set of images at the same time. Several data loaders are designed and implemented to process a batch of images at the same time for online evaluation. The researchers have options to use multiple CPUs and GPUs for evaluation in this package.

Maintenance: Due to the separation of different modules and implemented logger in FaRE, the system can easily track errors and help the developer to quickly update this package.

4. EXPERIMENTAL EVALUATION

In this section, we train two baseline models and evaluate them using our proposed package to perform the quantitative measurement and generate comparison plots.
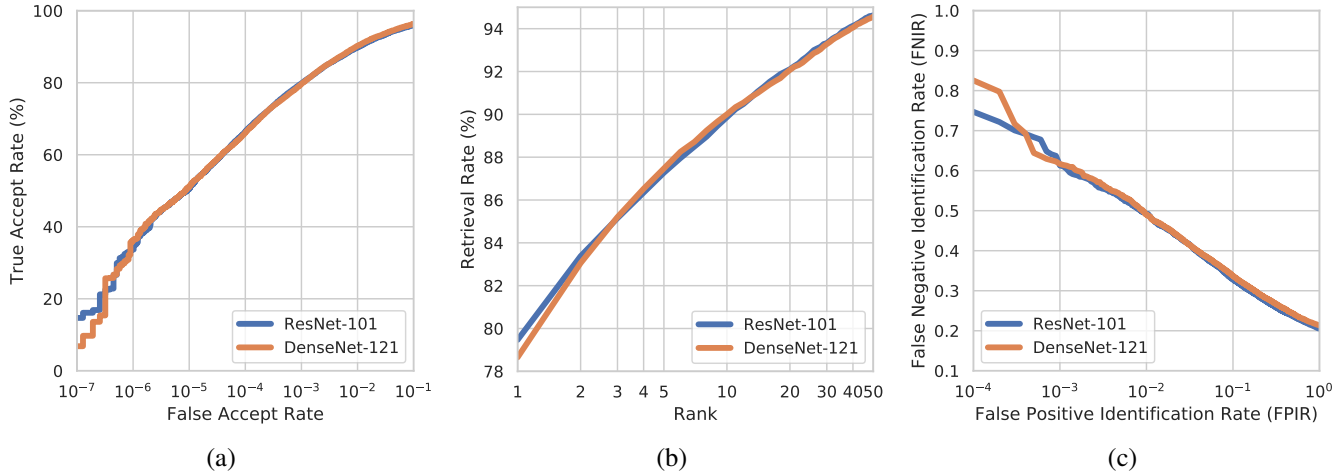
Fig. 3. Depiction of (a) the average ROC curve for the 1:1 face verification protocol, (b) the average CMC curve and IET curve for the 1:N open-set face identification protocol on the IJB-C dataset.

VGG-Face [24] is a well-known FR template generator and is provided the public access to the pre-trained model. ResNet-101 [2] and DenseNet-121 [3] are also two famous network architectures for the object classification task. However, there are no public pre-trained models of these two networks for the FR task. Therefore, two baselines using ResNet-101 and DenseNet-121 are trained on VGG-Face2 dataset to generate a facial template from a single image or a set of images. The average of the feature representations generated from a set of images of a subject is computed and treated as the template of that subject. We evaluate FaRE using two baselines on the IJB-C [9] dataset for both face verification and identification tasks to present the advantages in additional two aspects: generalization ability and scalability. The models are trained on a GPU cluster and all evaluation experiments are performed on a local machine with a CPU of Intel Core i7-6700K and a GPU of GeForce GTX 1080-Ti. In the online evaluation mode, it takes around one hour to generate the templates and compute the similarity scores for the mix identification task with a two-fold evaluation according to the IJB-C protocol.

To demonstrate the generalization ability, here, we set the ResNet-101 to use offline mode while DenseNet-101 is using online mode. The mean feature vectors are computed from the set of features as the final facial template. The average ROC performance across gallery sets for 1:1 mixed verification protocol and average CMC and IET performance across gallery sets for 1:N mixed identification protocol are computed by FaRE. In addition, the corresponding figures are generated by FaRE as depicted in Fig. 3.

To demonstrate the scalability, for simplicity, we directly performed 10-fold evaluation using FaRE on LFW dataset in the online evaluation mode. The relation of a number of images processed at the same time with the total processing

time of generating the templates and computing the similarity scores is depicted in Fig. 4. It takes approximately 35 seconds to finish generating templates using DenseNet-121 and comparing all pairs in LFW by processing 32 images at the same time, which can provide quick feedback for algorithm development stage.
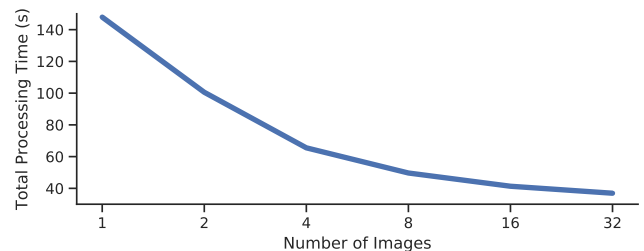
Fig. 4. Depiction the scalability of FaRE.

5. CONCLUSIONS

In this work, we designed and implemented a light-weight, maintainable, scalable, generalizable, and extendable face recognition evaluation toolbox in Python named FaRE that supports both online and offline evaluation to benefit the biometrics research community and to accelerate the biometrics-related research. FaRE is designed to evaluate general FR systems, which consists of current commonly-used evaluation metrics functions, closed-set, and open-set FR datasets, and can be extended to other customized datasets. Two baselines are evaluated on the IJB-C datasets to provide baselines and pre-trained model for other face-related tasks.

6. REFERENCES

[1] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *arXiv preprint arXiv:1801.07698*, 2018.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, Jun. 26 – Jul. 1 2016, pp. 770–778.

[3] Gao Huang, Zhuang Liu, Van der Maaten Laurens, Kilian Q Weinberger, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, July 22–25 2017, pp. 2261–2269.

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. IEEE Conference on Automatic Face and Gesture Recognition*, Xi'an, China, May 15–19 2018.

[5] Luan Tran, Xi Yin, and Xiaoming Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, Jul. 21–26 2017.

[6] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Jul. 21–26 2017.

[7] Brendan F. Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, Mark Burge, and Anil K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, Jun. 8–10 2015.

[8] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, Jordan Cheney, and Patrick Grother, "IARPA Janus Benchmark-B face dataset," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Honolulu, Hawaii, Jul. 21–26 2017, pp. 90–98.

[9] Brianna Maze, Jocelyn Adams, James A. Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother, "IARPA Janus Benchmark-C: Face dataset and protocol," in *Proc. IEEE International Conference on Biometrics*, Queensland, Australia, Feb. 20–23 2018, pp. 158–165.

[10] Lei Shi, Xiang Xu, and Ioannis A. Kakadiaris, "SSFD: a face detector via a single-scale feature map," in *Proc. IEEE International Conference on Biometrics: Theory Applications and Systems*, Los Angeles, CA, USA, Oct. 2018.

[11] Xiang Xu, Shishir Shah, and Ioannis A Kakadiaris, "Face alignment via an ensemble of random ferns," in *Proc. IEEE International Conference on Identity, Security and Behavior Analysis*, Sendai, Japan, 2016.

[12] Xiang Xu and Ioannis A. Kakadiaris, "Joint head pose estimation and face alignment framework using global and local CNN features," in *Proc. IEEE Conference on Automatic Face and Gesture Recognition*, Washington, DC, May 30–Jun. 3 2017, pp. 642–649.

[13] Xiang Xu, Ha Le, and Ioannis A. Kakadiaris, "On the importance of feature aggregation for face reconstruction," in *Proc. Winter Conference on Applications of Computer Vision*, Waikoloa Village, HI, Jan. 8–11 2019, pp. 1–10.

[14] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *Proc. European Conference on Computer Vision*, Munich, Germany, Sep. 8–14 2018.

[15] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep., University of Massachusetts, Amherst, MA, 2007.

[16] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs, "Frontal to profile face verification in the wild," in *Proc. IEEE Winter Conference on Applications of Computer Vision*, Lake Placid, NY, 2016.

[17] Ha A. Le and Ioannis A. Kakadiaris, "UHDB31: A dataset for better understanding face recognition across pose and illumination variation," in *Proc. IEEE International Conference on Computer Vision Workshops*, Venice, Italy, Oct. 22–29 2017, pp. 2555–2563.

[18] Joshua C. Klontz, Brendan F. Klare, Scott Klum, Anil K. Jain, and Mark J. Burge, "Open source biometric recognition," in *Proc. IEEE Conference on Biometrics: Theory, Applications and Systems*, Washington DC, Sep. 29–Oct. 2 2013.

[19] Andre Anjos, Laurent El Shafey, Roy Wallace, Manuel Gunther, Chris McCool, and Sebastien Marcel, "Bob: a free signal processing and machine learning toolbox for researchers," in *Proc. ACM Conference on Multimedia Systems*, Nara, Japan, Oct. 2012.

[20] Andre Anjos, Manuel Gunther, Tiago de Freitas Pereira, Pavel Korshunov, Amir Mohammadi, and Sebastien Marcel, "Continuously reproducing toolchains in pattern recognition and machine learning experiments," in *Proc. International Conference on Machine Learning*, Sydney, Australia, Aug. 6 – Aug. 11 2017.

[21] Xiang Xu, Ha Le, Pengfei Dou, Yuhang Wu, and Ioannis A. Kakadiaris, "Evaluation of 3D-aided pose invariant face recognition system," in *Proc. International Joint Conference on Biometrics*, Denver, CO, Oct. 1–4 2017, pp. 446–455.

[22] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang, "MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems," in *Proc. Neural Information Processing Systems, Workshop on Machine Learning Systems*, Barcelona, Spain, Dec. 7–12 2015.

[24] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," in *Proc. British Machine Vision Conference*, Swansea, United Kingdom, Sep. 7–10 2015, pp. 1–12.