Ear images classification based on data augmentation and ResNeXt50

Thinh Le Duc[1], Linh Nguyen Hoang Anh[1], Trung Nguyen Quoc[1,2], and Vinh Truong Hoang[3]

[1] Department of Information Technology, FPT University, Ho Chi Minh city, Vietnam
{thinhldse140160,linhnhase150300,trungnq46}@fpt.edu.vn
[2] VSB, Technical University of Ostrava
quoc.trung.nguyen.st@vsb.cz
[3] Faculty of Information Technology, Ho Chi Minh City Open University, Vietnam
vinh.th@ou.edu.vn

Abstract. Biometric researchers have recently paid a lot of attention to recognizing persons by their voices. This trend can be attributed to a number of factors, including the fact that ear recognition does not suffer from some of the drawbacks of contactless biometrics, such as facial recognition, that it is the most promising face matching candidate in the context of multi-position face recognition, and that ears can be used to identify people in surveillance videos where faces may be completely or partially obscured. In addition, the ear appears to age more slowly. Although ear detection and recognition technology has advanced to a certain point, it has only been successful in controlled indoor environments. In this study, machine vision experiments were utilized to recognize the ears of well-known musicians in Vietnam using the EarVN1.0 dataset. The data set is divided into training and test sets using the ratios 70:30 and 60:40, respectively, during the model-building process. According to experimental findings, the ResNeXt50 neural network with image enhancement approach produced good results, scoring 93% for a 70–30 ratio and 90% for a 60–40 ratio.

Keywords: EarVN1.0 · HOG · LPQ · SVM · RF · ResNeXt50.

1 Introduction

In the early stages of civilization, each person had unique physical characteristics, such as their look, gait, voice, etc., which were crucial to the organization of human society. In tiny towns, it is simple for people to recognize one another due to their distinguishing traits. The creation of identity management systems to be able to record, successfully preserve, and remove personal identities of individuals, however, has been prompted by the explosion in population growth, combined with commerce expansion and increased mobility in modern society.

High-accuracy user identification systems have been necessary with the advancement of contemporary technological techniques in recent years in order to

guarantee the security of the application or requirements. of society. The most effective choice is without a doubt the biometric identification system. A person's identification can be verified using semi- or fully automated methods based on behavioral traits like voice or signature, or physical attributes like face, iris, and fingerprint [16]. In ordinary applications, this method has several advantages over the password method because biological data cannot be misplaced, stolen, or copied. Studies on biometric data fall into two categories: intrinsic biometric features like veins and external biometric features like face, iris, and fingerprints [24, 19]. Eye surface vein or nail on the palm and finger. While internal qualities cannot be influenced by the exterior environment, external environmental elements can have an impact on external traits because they are visible [21, 3].

The ability to discriminate between biometric traits has been evolved into a variety of systems that are utilized in security and forensic investigations, among other uses. Face recognition has been unsuccessful during the present global pandemic because users are hiding their faces. The entire human race is currently transitioning into a disguised community on a global scale. Face recognition systems suffer substantially as a result, and they need to be improved in current systems. Due to COVID-19's contact-based extraction, fingerprint and palette-based recognition are not appropriate in this situation. But because it is visible, the human ear has turned out to be more practical. There have been numerous studies on the biometric properties of the auricle, and the personal identity system based on ear recognition is a current biometrics research topic [1]. Accurately assessing the ear's details involves many challenges. These include hiding the ear with clothing, hairstyles, ear jewelry, and accessories. Another conclusion might be that the photograph was taken from a different angle, which obscured important anatomical details of the ear. Due to these challenges, ear recognition has been relegated to a supporting position in identifying procedures and systems that are frequently utilized for identification and verification [7].

The rest of this work is divided into the following sections: the vast majority of the research-accessible ear information bases are presented in Section 2. Section 3 presents the dataset that was used for the investigation. A study of ear recognition method is presented in Section 4, the results of the experiments are presented in Section 5, and the conclusion is presented in Section 6.

2 Related Work

The majority of 2D ear image-based biometric recognition systems have features that allow for the extraction of the extracted vector and comparison with the trained models. This allows us to divide ear identification techniques into two groups: manual feature extraction techniques and deep neural networks(DNNs).

Investigating some established machine learning techniques will be our first step. The researchers used information from compression networks and ear geometry to create a classification model. For ear biometrics, Hurley et al. [15] employed force-field feature extraction. Then, various auditory recognition methods were used. Principal component analysis (PCA) was used to ear photographs,

and while the results were encouraging, it was clear that the face is a more trustworthy biometric than the ear.

Methods for ear biometrics that rely on local features and have been successful in other biometric applications are also suggested [4]. These methods identify a group of relevant points (keypoints) from the image, and then compute an independent descriptor for each keypoint in order to extract local features of the ear [13]. Numerous feature extraction strategies have been created to acquire distinct and intrinsic local structures from images because feature extraction is the foundation of any recognition system. These methods include texture descriptors, encodings, and features for the image's texture data. The success of LBP in these fields encouraged us to apply it to ear recognition. Few academics have examined the application of LBP in ear recognition as a standalone image representation technique or in conjunction with other techniques in this regard. The LBP descriptor is employed in [6] to extract ear characteristics. The recognition rate was tested using ear pictures from 125 people in the IIT Delhi ear dataset, and the recognition rate was 93%.

The following research papers examine the second track, which is based on deep learning techniques. However, ear recognition using deep learning has only lately been applied [10, 11, 7, 1, 12]. The lack of training data with labels is one challenge with ear recognition issues. Emersic et al [10] .'s solution makes use of data augmentation.

The CNN created by Tian et al. [23] was composed of three convolutional layers, a fully connected layer, and a softmax classifier, and it was applied to the task of ear recognition. The USTB ear database, which included 79 people at different posture angles, was utilised. The photographs used occlusions like headsets, earrings, and other similar objects. According to Raveane et al. [20], it is challenging to accurately identify and locate an ear within an image. This difficulty increases when working under variable conditions, and it may also be due to the peculiar shape of human ears, as well as the lighting conditions, as well as the shifting profile shape of an ear when photographed. The ear detection system determined the presence and location of an ear using a combination of numerous CNNs and a detection grouping technique. When compared against clear, purpose-taken photos, the suggested method performs similarly to existing methods, with an accuracy of up to 98%. A deep learning model for unconstrained ear recognition was put forth by the authors in [9]. For ear recognition, they provided the features to a shallow classifier. The best outcomes were obtained using an ensemble of ResNet18 models, which offered consistent performance across the tested datasets, and they recommended a deep learning-based averaging ensemble to limit the over fitting.

The authors of [2] presented a six later deep CNN, which was tested on the IIT Delhi II and AMI ear datasets with recognition rates that were accurate to 97.36% and 96.99% for 1000 epochs, respectively. The outcomes are duplicated using the AMI dataset, where the ear images are rotated at various angles, with various lighting conditions, as well as when random noise is included. The

combined variation conditions resulted in a reduction in recognition accuracy, which fell to 91.99%.

3 Data Preparation

3.1 *EARVN1.0*

One of the largest ear datasets obtained under open conditions is the recently released EarVN1.0 [14] dataset. There are 28,412 photos of great artists' ears in all, including the left and right ears, from both genders. The spatial resolution of the RGB image ranges from 15×15 to 200×200 pixels. Cropped ears from face photos show significant differences in scale, viewing angle, illumination, contrast, and other properties. They were captured using a variety of acquisition methods in uncontrolled environments, tiny artifact in the background This dataset for ear recognition is complex and challenging. When used in practical applications, it does, however, represent reality. Figure 2 shows representative ear pictures for two people to demonstrate the complexity of the EarVN1.0 dataset.

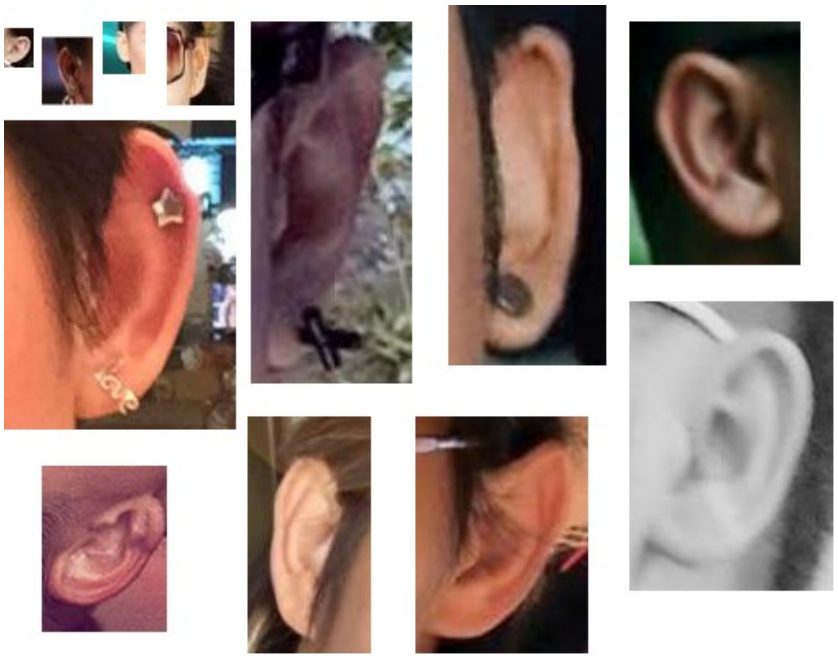

Fig. 1. The EarVN1.0 dataset contains sample ear pictures for a few different people.

3.2 *Data Augmentation*

Large corpora of annotated examples are needed for deep CNN training in order to accurately extract additional class-specific features. Deep CNNs have a tendency to overfit on tiny datasets when the large-scale training data is not available. A series of label-preserving transformations that enhance the training examples by perturbing them by gently altering their appearance before supplying them to the networks for training is an efficient way to solve the problem. Simple changes like horizontal flipping, color space augmentations, and random cropping are the early examples used to demonstrate the usefulness of data augmentations.

4 Methods

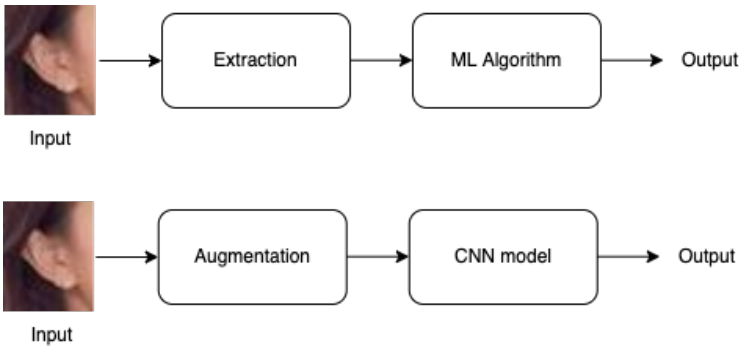

Fig. 2. The process of applying handcraft algorithm (above) and CNN model. (under)

4.1 *Handcraft*

We applied handcraft techniques to obtain an observation. The process of using the handcraft approach is described in figure 2.

4.1.1 ***Extraction*** We apply the extraction method to get the critical feature in the input image.

Histogram of oriented gradients: We investigate the issue of feature sets for reliable visual object recognition using a test case of linear SVM-based human detection. In this experimental demonstration, we demonstrate that grids of histograms of oriented gradient (HOG) descriptors greatly outperform existing feature sets for human detection after examining existing edge and gradient based descriptors, featuring a wide variety of poses and backgrounds [8].

GIST: [18] GIST were taken into account as local features. To enhance the quality of the input image, we pre-processed it. Support Vector Machines are used for classification (SVM). Separate characteristics are extracted for GIST. When features are employed separately, classification produces subpar results. Higher precision is achieved later when the two features are combined to generate a hybrid feature.

4.1.2 *ML Algorithm*

A machine learning model is a program that can find patterns or make decisions from a previously unseen dataset.

In this process, we implemented outstanding techniques of Machine learning algorithms, namely Linear Support Vector, Random Forest, and Support Vector Machine [5]. These algorithms play an important role in our project on the way to obtaining the outcome.

4.2 *Deep learning approach*

The primary research methodologies are effectively explained in this part with positive findings. In this study, we practice with VGGNet, ResNeXt50, and ResNeXt50 neural networks utilizing image enhancement approaches. Deep learning models have currently achieved many high outcomes. The process of applying the deep learning approach have show in figure 2.

VGGNets: The Visual Geometry Group networks (VGGNets) [22] are the best image recognition and object localization performers in the ILSVRC-2014. Using a network depth ranging from 11 to 19 layers, the authors examined the impact of the network's depth on the recognition accuracy. VGGNets have introduced two crucial traits that set them apart from earlier CNNs like AlexNet. First, the entire network utilizes tiny 3×3 receptive fields. Larger receptive fields, such 5×5 or 7×7, can be covered while the number of trainable parameters is greatly decreased by stacking numerous convolutional layers. Second, to create deeper networks, many layers with the same traits are piled, [22] provides additional information.

ResNeXt: In the ILSVRC-2016 competition for image classification and localization, ResNeXt [25] took second place with its highly modularized CNN architecture. ResNeXt builds deep networks with the same ease of design as VGGNet and ResNet. First, by stacking several layers or building pieces with the same number of channels and filter sizes that make up a similar architecture. Second, the number of channels doubles when the spatial dimension is shrunk by a factor of two. ResNeXt uses the split-transform-merge technique from the Inception module as well, but uses the same set of transformations throughout all routes, making it simple to increase and study the number of paths as a separate hyperparameter. Cardinality, the measure of the size of the set of transformations, is thought to be a crucial factor in network performance. Figure 3 shows the ResNeXt building block with a cardinality of 32.

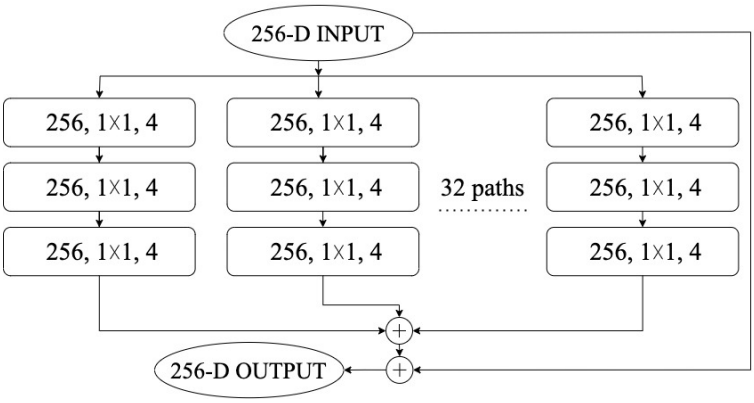

Fig. 3. A Block of ResNeXt with Cardinality equal to 32.

4.3 *Experiments*

After loading images from the disk, we applied types of augmentation randomly. First, we rotated images ranging from -20 degrees to 20 degrees. Next, we utilized Gaussian Blur and Gaussian Noise at the same percent of 20 %, and 20 % is also the ratio of changing brightness, contrast, saturation and hue which were applied. Moreover, we made use of Random Horizontal Flip with a ratio of 50 % to obtain upside down images for testing models. Lastly, we normalized the data.

All of the experiments were performed on tensorflow 2.2 and pytorch 1.10 in python and were run on GPU RTX 6000 which minimized the time and strengthened the training result by heightening size batch. Towards the handcraft techniques, we put on model by model to learn from features that are suitable for it and evaluate accuracy.

With Convolutional Neural Network (CNN), we prepared images with the dimension of 257 x 161 at first to train with both ResNeXt50 and VGG16 which optimal function is adam, cross-entropy loss, a weight decay of 0.001 and in 200 epochs. After that, we applied augmentation data process for the model that gave better performance which was what we estimated.

5 Result

Table 1 presents a comparison of the findings achieved using state-of-the-art techniques. We present the accuracy of 4 ratios of data set with 3 outstanding extraction, namely Histograms of oriented gradients, GIST and CNN. In each extractions, there is always one peak result, such as Linear Support Vector model in Histograms of oriented gradients and Random Forest in Local phase quantization and GIST are the highest results among the other models in the same extractions which are presented obviously in the Table 1.

However, our new approach, ResNeXt50 with Augmentation, proved the dominant result with significant numbers like 92.6 %, 90.5%. Those are the highest numbers out of all of our implemented techniques. Even the result of the Wavelet method proposed by Mewada et al [17] that we have compared in Table 2.

Extraction	Model	Accuracy	
		70_30	60_40
Histogram of oriented gradients	**Linear Support Vector**	**75.1**	**66.8**
	Random Forest	73.7	64.2
	Support vector machine	66.2	60.0
GIST	Linear Support Vector	77.2	67.9
	Random Forest	**76.3**	**67.4**
	Support vector machine	46.9	42.7
CNN	VGG	73.2	61.1
	ResNeXt50	89.9	80.6
	ResNeXt50 with Augmentation	**92.6**	**90.5**

Table 1. Comparison of recognition accuracy in our experiments.

Model	Accuracy
	60_40
Wavelet + CNN [17]	82.2
ResNeXt50 with Augmentation	**90.5**

Table 2. Comparison of our model recognition accuracy with Wavelet model that introduce in Mewada et al [17].

6 Conclusion

On the EarVN1.0 dataset, which is one of the biggest datasets gathered under unrestricted circumstances, we provided the findings of the overall experimental investigation of ear recognition in this publication. They demonstrated remarkable identification abilities in a variety of vision tasks. We create a baseline result by some handcraft method with the Histogram of oriented gradients and GIST as the extraction method and Linear Support Vector, Random Forest, and Support Vector Machine as the ML method to learn features from extraction output. On the other hand, we use VGGNets and ResNeXt as deep learning methods and have a comparison with the baseline result. Notice the complete domination of the deep learning methods with higher accuracy which show in Table 1, we continue to apply some augmentation methods to the deep learning methods with applied intensity based on experience and trial and error. The final result has improved over the method Wavelet + CNN which show in Table 2.

References

1. Abaza, A., Ross, A., Hebert, C., Harrison, M.A.F., Nixon, M.S.: A survey on ear biometrics. ACM Computing Surveys **45**(2), 1–35 (Feb 2013). https://doi.org/10.1145/2431211.2431221, https://dl.acm.org/doi/10.1145/2431211.2431221
2. Ahila Priyadharshini, R., Arivazhagan, S., Arun, M.: A deep learning approach for person identification using ear biometrics. Applied Intelligence **51**(4), 2161–2172 (Apr 2021). https://doi.org/10.1007/s10489-020-01995-8, http://link.springer.com/10.1007/s10489-020-01995-8
3. Alashik, K.M., Yildirim, R.: Human Identity Verification From Biometric Dorsal Hand Vein Images Using the DL-GAN Method. IEEE Access **9**, 74194–74208 (2021). https://doi.org/10.1109/ACCESS.2021.3076756, https://ieeexplore.ieee.org/document/9420051/
4. Annapurani, K., Sadiq, M., Malathy, C.: Fusion of shape of the ear and tragus – A unique feature extraction method for ear authentication system. Expert Systems with Applications **42**(1), 649–656 (Jan 2015). https://doi.org/10.1016/j.eswa.2014.08.009, https://linkinghub.elsevier.com/retrieve/pii/S0957417414004850
5. Biship, C.M.: Pattern recognition and machine learning (information science and statistics) (2007)
6. Boodoo-Jahangeer, N.B., Baichoo, S.: LBP-based ear recognition. In: 13th IEEE International Conference on BioInformatics and BioEngineering. pp. 1–4. IEEE, Chania, Greece (Nov 2013). https://doi.org/10.1109/BIBE.2013.6701687, http://ieeexplore.ieee.org/document/6701687/
7. Booysens, A., Viriri, S.: Ear Biometrics Using Deep Learning: A Survey. Applied Computational Intelligence and Soft Computing **2022**, 1–17 (Aug 2022). https://doi.org/10.1155/2022/9692690, https://www.hindawi.com/journals/acisc/2022/9692690/
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 886–893 vol. 1 (June 2005). https://doi.org/10.1109/CVPR.2005.177
9. Dodge, S., Mounsef, J., Karam, L.: Unconstrained ear recognition using deep neural networks. IET Biometrics **7**(3), 207–214 (May 2018). https://doi.org/10.1049/iet-bmt.2017.0208, https://onlinelibrary.wiley.com/doi/10.1049/iet-bmt.2017.0208
10. Emersic, Z., Stepec, D., Struc, V., Peer, P.: Training Convolutional Neural Networks with Limited Training Data for Ear Recognition in the Wild. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). pp. 987–994. IEEE, Washington, DC, DC, USA (May 2017). https://doi.org/10.1109/FG.2017.123, http://ieeexplore.ieee.org/document/7961853/
11. Galdámez, P.L., Raveane, W., González Arrieta, A.: A brief review of the ear recognition process using deep neural networks. Journal of Applied Logic **24**, 62–70 (Nov 2017). https://doi.org/10.1016/j.jal.2016.11.014, https://linkinghub.elsevier.com/retrieve/pii/S1570868316300684
12. Hasan, U., Hussain, W., Rasool, N.: AEPI: insights into the potential of deep representations for human identification through outer ear images. Multimedia Tools and Applications **81**(8), 10427–10443 (Mar 2022). https://doi.org/10.1007/s11042-022-12025-9, https://link.springer.com/10.1007/s11042-022-12025-9

13. Hassaballah, M., Abdelmgeid, A.A., Alshazly, H.A.: Image Features Detection, Description and Matching, pp. 11–45. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-28854-3_2, $https$: $//doi.org/10.1007/978-3-319-28854-3_2$
14. Hoang, V.T.: EarVN1.0: A new large-scale ear images dataset in the wild. Data in Brief **27**, 104630 (Dec 2019). https://doi.org/10.1016/j.dib.2019.104630, https://linkinghub.elsevier.com/retrieve/pii/S2352340919309850
15. Hurley, D.J., Nixon, M.S., Carter, J.N.: Force field feature extraction for ear biometrics. Computer Vision and Image Understanding **98**(3), 491–512 (Jun 2005). https://doi.org/10.1016/j.cviu.2004.11.001, https://linkinghub.elsevier.com/retrieve/pii/S1077314204002024
16. Jain, A.K., Ross, A.A., Nandakumar, K.: Introduction to Biometrics. Springer US, Boston, MA (2011). https://doi.org/10.1007/978-0-387-77326-1, http://link.springer.com/10.1007/978-0-387-77326-1
17. Mewada, H.K., Patel, A.V., Chaudhari, J., Mahant, K., Vala, A.: Wavelet features embedded convolutional neural network for multiscale ear recognition. Journal of Electronic Imaging **29**(4), 043029 (2020)
18. Mustafa, R., Dhar, P.: A method to recognize food using gist and surf features. In: 2018 Joint 7th International Conference on Informatics, Electronics Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision Pattern Recognition (icIVPR). pp. 127–130 (June 2018). https://doi.org/10.1109/ICIEV.2018.8641072
19. Rathgeb, C., Tolosana, R., Vera-Rodriguez, R., Busch, C. (eds.): Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks. Advances in Computer Vision and Pattern Recognition, Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-030-87664-7, https://link.springer.com/10.1007/978-3-030-87664-7
20. Raveane, W., Galdámez, P.L., González Arrieta, M.A.: Ear Detection and Localization with Convolutional Neural Networks in Natural Images and Videos. Processes **7**(7), 457 (Jul 2019). https://doi.org/10.3390/pr7070457, https://www.mdpi.com/2227-9717/7/7/457
21. Shaheed, K., Liu, H., Yang, G., Qureshi, I., Gou, J., Yin, Y.: A Systematic Review of Finger Vein Recognition Techniques. Information **9**(9), 213 (Aug 2018). https://doi.org/10.3390/info9090213, http://www.mdpi.com/2078-2489/9/9/213
22. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition (Apr 2015), http://arxiv.org/abs/1409.1556, arXiv:1409.1556 [cs]
23. Tian, L., Mu, Z.: Ear recognition based on deep convolutional network. In: 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). pp. 437–441. IEEE, Datong, China (Oct 2016). https://doi.org/10.1109/CISP-BMEI.2016.7852751, http://ieeexplore.ieee.org/document/7852751/
24. Vasanthi, M., Seetharaman, K.: Facial image recognition for biometric authentication systems using a combination of geometrical feature points and low-level visual features. Journal of King Saud University - Computer and Information Sciences **34**(7), 4109–4121 (Jul 2022). https://doi.org/10.1016/j.jksuci.2020.11.028, https://linkinghub.elsevier.com/retrieve/pii/S1319157820305577
25. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated Residual Transformations for Deep Neural Networks (Apr 2017), http://arxiv.org/abs/1611.05431, arXiv:1611.05431 [cs]