Deep Transfer Learning for Detecting Face Images with Moustache

Filip Babić, Petra Grd, Igor Tomičić, Ena Barčić
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, 42 000 Varaždin
fbabicucg@gmail.com, {petra.grd, itomicic, enbarcic}@foi.unizg.hr

Abstract. *Face detection is one of the most researched fields in deep learning, but some aspects have not yet been sufficiently researched. In this paper, we will focus on the application of deep transfer learning for detecting face images with moustache. For the purpose of this paper, we use a version of the MobileNetV2 convolutional neural network which was pretrained on the ImageNet dataset and the provided weights are used for transfer learning to moustache detection problem. In addition, we created a balanced dataset for moustache detection from FFHQ dataset. Using our approach we were able to achieve a testing F1-score of 89%.*

Keywords. face biometrics, neural networks, moustache detection, MobileNetV2, FFHQ dataset

1 Introduction

Deep transfer learning is a valuable technique in the field of computer vision that allows the transfer of knowledge learned from one task to another, and in recent years, it has been used to achieve state-of-the-art performance in various tasks, including object detection, image classification, and face recognition.

In this paper we will focus on the application of the deep transfer learning for detecting face images with moustache, where using deep transfer learning in such a context may have several advantages. Firstly, it can lead to improved accuracy as it leverages knowledge obtained from large datasets of face images, leading to improved performance in detecting face images with moustache. Secondly, it reduces training time since the model can be fine-tuned using a smaller dataset rather than being trained from scratch. Furthermore, deep transfer learning algorithms are designed to generalize well, resulting in improved generalization on new, untested data. They can also be engineered to be robust to variations in lighting, facial expressions, and camera viewpoint for example. Additionally, deep transfer learning algorithms require less computational resources compared to building a model from scratch. This makes them well-suited for deployment in resource-constrained environments, where efficiency is crucial. By leveraging pre-trained weights and knowledge from existing models, deep transfer learning minimizes the computational burden, enabling effective moustache detection on face images without excessive resource requirements.

Due to their cultural and historical relevance, moustaches in particular have been extensively studied as an important visual cue for gender classification and face recognition. Detecting face images with moustache can be challenging due to the variations in moustache shape, size, and color, and influenced by factors such as facial expressions, lighting, and camera viewpoint.

The further motivation for moustache detection can be found in various potential use-cases. For instance, it can enhance gender classification algorithms by detecting moustaches on face images, thus improving accuracy. Moustaches can also significantly impact face recognition performance, making moustache detection relevant in the field. In surveillance systems, identifying individuals based on their facial hair through moustache detection can be valuable in settings like security checkpoints, border control, and criminal investigations. Moustache detection also holds potential for personalized experiences, such as within augmented reality applications. Moreover, analyzing trends in facial hair and targeting specific demographics based on their facial hair preferences can be advantageous in marketing and advertising endeavors.

In this paper, we will present a state of the art research efforts within the field, and propose a solution by applying a deep transfer learning for detecting moustache on face images.

Section 2 will provide a related work overview; Section 3 will illustrate the network architecture of the proposed model; Section 4 will provide more detailed information on the used dataset, training, validation and testing protocol, and demonstrate and discuss achieved results. Section 5 will conclude and discuss the implications of the research and possible future work.

2 Related work

When talking about facial hair in the context of face recognition we are most commonly talking about two separate fields of research: facial hair detection and facial hair segmentation. Facial hair detection, as the name implies focuses on detecting facial hair in images of individuals or group images (Yang et al., 2018).

The authors in (Yang et al., 2018) present a DCNN for face detection leveraging facial attributes-based supervision called Faceness-Net. Some of the attributes they focus on are beard, no beard, goatee, 5 o'clock shadow, mustache, and sideburns. They use the Labeled Faces in the Wild (LFW) dataset (Huang et al., 2007) for training, and chose to combine the evaluation to one category 'hair+beard'. They use cropped images achieving a detection accuracy of 95.56% and uncropped images achieving a detection accuracy of 94.57%. They conclude that hair, eye, and nose parts perform much better than the mouth and beard.

In (Gudi, 2016) the authors focus on recognizing semantic features in faces using deep learning. They test the effects of various factors and hyper-parameters of deep neural networks for an optimal network configuration that can accurately recognize semantic facial features like emotions, age, gender, ethnicity, etc. In their research, they include 'add-on' features like glasses and facial hair (beard and moustache). The baseline architecture relies on a ReLu model and is trained and tested on the VV Dataset ("Vicarious Perception Technologies" 2023) and Facial Expression Recognition (FERC) Dataset (Goodfellow et al., 2015). The features of beards and moustaches are further classified into three categories consisting of non, light, or heavy. The authors achieve a final accuracy of 88.1% for the beards and 89.13% for the moustaches. They conclude that their network's precision for detecting glasses and heavy moustaches is very high, but the network does not learn very precise light beards and no beard classifiers.

In (Hand and Chellappa, 2016) the authors present a multi-task network for attribute classification. They present attribute relationships in three ways: using a multi-task deep convolutional neural network (MCNN) sharing the lowest layers amongst all attributes, sharing the higher layers for related attributes, and building an auxiliary network on top of the MCNN which utilizes the scores from all attributes to improve the final classification of each attribute. In addition, the authors have separated facial hair into five categories consisting of 5 o'clock shadow, moustache, no beard, sideburns, and goatee, using the CelebA (Liu et al., 2015) and LFWA (Huang et al., 2007) datasets for training and testing. They conclude that on the CelebA dataset using all three approaches the best results were achieved on the sideburns samples achieving 97.77% on the Independent CNN, 97.82% on the MCNN, and 97.85% on the MCNN-AUX, while the lowest results were achieved on the 5 o'clock Shadow categories achieving 93.94% on the Independent CNN, 94.41% on the MCNN, and 94.51% on the MCNN-AUX. On the LFWA dataset using all three approaches the best results were achieved on the moustache samples achieving 93.69% on the Independent CNN, 93.53% on the MCNN, and 93.43% on the MCNN-AUX, while the lowest results were achieved on the 5 o'clock Shadow categories achieving 77.39% on the Independent CNN, 77.70% on the MCNN, and 77.06% on the MCNN-AUX. On the LFWA dataset.

The authors in (Sghaier and Elfaki, 2021) present an efficient technique for human face occlusion detection and extraction. They extract three types of occlusions using artificial intelligence techniques consisting of a Viola-Jones algorithm based on Vision Cascade Detector and combined two performance methods: fuzzy C-means method and filters and morphological operations. This is done to boost the efficiency of face recognition systems. They divide occlusions into three categories: beard, moustache, and eyeglasses and use the Faces94 ("Face Recognition Data" n.d.) and CMU-PIE (Gross et al., 2008) datasets. The authors were able to achieve 100% accuracy with both the FCM method and Morphological operations on the Faces94 dataset, and a 99% accuracy with both on the CMU-PIE dataset.

Other approaches are more focused on facial hair segmentation such as the authors of (Le, Luu, Seshadri, et al., 2012) which present a novel system for beard and moustache detection and segmentation in challenging facial images. They use pre-processes images using the self-quotient algorithm and then use a sparse classifier to detect and segment regions containing facial hair from the MBGC, color FERET databases (Phillips et al., 1997). Facial hair is classified into two categories: beard and moustache. The authors are able to achieve 96.2% accuracy on beards, and 98.8% accuracy on moustaches on the MBGC dataset and 95.8% accuracy on beards, and 97.0% accuracy on moustaches on the color FERET dataset.

Some approaches combine both facial hair detection and facial hair segmentation like in (Ubaid et al., 2022) where the authors present an efficient state-of-the-art system for beard and hair detection and segmentation for changing color in challenging facial images based on a modified version of a Mask R-CNN model. They create their own dataset of 1 500 images which are equally divided into both hair and beard images. In the experiment phase, the authors were able to achieve a final accuracy of 91.2%.

The authors in (Le, Luu, Zhu, et al., 2017) present a robust, fully automatic, and semi-self-training system to detect and segment facial beard/moustache simultaneously in challenging facial images. The authors use the superpixel together with a combination of two classifiers, Random Ferns (rFerns) and Support Vector Machines (SVM) to obtain good classification performance as well as improve time efficiency.

They divide facial hair into three categories: facial hair, non-facial hair (male), and non-facial hair (female). In the end, they propose a facial hair detection algorithm and test it on the Multiple Biometric Grand Challenge (MBGC) (“Multiple Biometric Grand Challenge (MBGC) Presentations” 2010) and FERET (Phillips et al., 1997). They achieve the highest accuracy on the MBGC dataset with an accuracy of 88.5% for facial hair, 91% for non-facial hair (male), and 90.8% for non-facial hair (female), and the lowest accuracy on the FERET dataset with 85.1% for facial hair and 84.5% for non-facial hair (female), and the lowest accuracy for non-facial hair (male) was achieved on the PINELLAS dataset with 84.2%.

Among different deep learning models, MobileNetV2 architecture has emerged as a choice in different computer vision tasks such as face recognition, face mask detection and similar. Leveraging its streamlined structure, MobileNetV2 demonstrates exceptional prowess in accurately and rapidly discerning facial features, while optimizing the resources necessary for training and testing. The authors in (Almghraby and Elnady*, 2021) use mobilenetv2 and transfer learning for real-time mask detection. They used stochastic gradient descent (SGD) for optimization with 12 epochs in each experiment, a learning rate of 0.001, and a momentum of 0.85. They were able to achieve 99% training and 98% validation accuracy. Similarly the authors in (Kumar and Bansal, 2023) present their approach for mask detection on photos and video images using Caffe-MobileNetV2 transfer learning. They used the MobileNetV2 mask identification and added five different layers to the pre-trained MobileNetV2 architecture. They achieved an accuracy of 99.64% on photos.

The authors in (Huu et al., 2022) propose a model capable of distinguishing between masked and nonmasked faces using deep learning (DL)—MobileNetV2. The authors used Retina Face as their face detector model, next they used MaskTheFace to modify their dataset and finally they trained MobileNetV2. They were able to achieve an accuracy of up to 99.37%. In (Narduzzi et al., 2022) the authors focus on the adaptation of MobileNetV2 for the purpose of face detection on ultra-low power platforms. They used Tiny-Face as a starting point and computed a set of canonical bounding boxes derived by clustering them into 25 different sizes. Detection results were obtained after applying non-maximum suppression (NMS). MobileNetV2 was adapted to a fully convolutional architecture, and two breakpoints and three outputs were used. They concluded that MobileNetV2 can successfully be applied as an alternative to classical machine learning methods for embedded face detection.

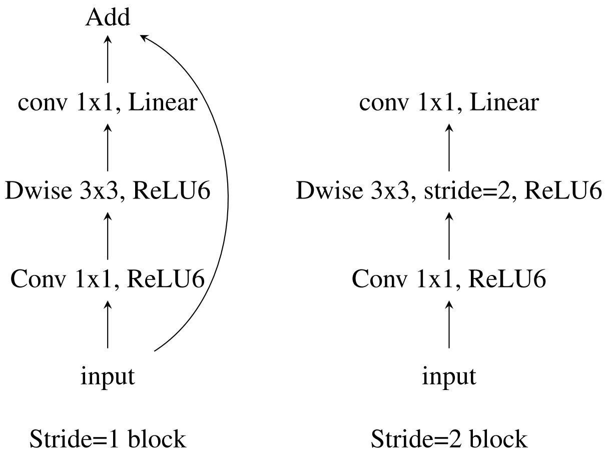

Figure 1: Convolutional blocks of MobileNetV2 (Sandler et al., 2018)

3 Network Architecture

The artificial neural network architecture used in the process of this research is a version of the MobileNetV2 convolutional neural network (Sandler et al., 2018). This particular architecture was chosen for its provided benefit in terms of low resource consumption during the training and testing processes. Lower resource consumption is the result of a trade-off in the manner of a reduced number of input parameters, leading to a higher efficiency score.

The basic building block of MobileNetV2 is "a bottleneck depth-separable convolution with residuals" (Sandler et al., 2018) (illustrated in Figure 1). The first layer of the model is a fully convolutional layer with 32 filters. 19 bottleneck residual layers follow it. The most prominently used activation function is ReLU6 (Howard et al., 2017), a modified rectified linear unit with the constrained maximum activation size of 6, primarily for the robustness it provides. A constant expansion rate is present in every layer, excluding the first layer. The utilised optimizer was AdamOptimizer (Table 1).

Table 1: CNN architecture hyperparameters

Hyperparameter	Value
Batch size	32
Activation function	ReLu6 and SoftMax
Loss function	binary cross entropy
Optimizer	AdamOptimizer
Learning rate	0.0001
Dropout	0.6

4 Experiments

4.1 Dataset

The dataset selected for this research is Flickr-Faces-HQ Dataset. This image dataset exhibits a substantial variation in terms of the faces it comprises. It encompasses individuals from different age groups, ethnicities, and backgrounds, thus offering a diverse representation. Moreover, the dataset also contains numerous images featuring faces adorned with accessories like glasses, sunglasses, hats, caps, and similar items. The presence of such accessories adds an additional level of complexity and diversity to the dataset, making it suitable for studying and developing algorithms that can effectively handle and recognize various facial attributes and accessories. The images are taken from the Flickr page and automatically aligned and cropped (Karras et al., 2018). The dataset contains more than 70 000 images with different resolutions. In this paper, a resolution of 128x128 pixels was chosen because it proved to be sufficient for the goal of the paper and manageable using available computer resources.

The selected dataset does not have labels for moustaches in images, so in order to train the neural network, the dataset has been manually divided into two parts. The first part consists of images without moustaches and the second part of 6 954 images with moustaches (Figure 2). For images without moustaches, a randomly selected subset of 7 060 images without moustaches was selected to create a balanced dataset.

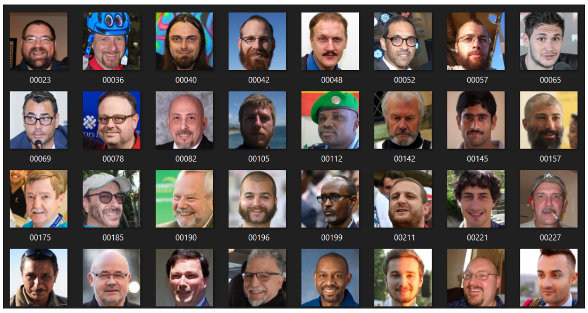

Figure 2: Examples of images with moustaches

The end results is the dataset with two subfolders corresponding to the classes necessary, one with moustache (49,62%) and one without moustache (50,38%).

4.2 Training and Validation

In order to train and later test the results of the network for moustache detection, the holdout method is used where 80% of images in each class is used for training and 20% for testing. The network was pretrained on ImageNet dataset and the provided weights are used for transfer learning to moustache detection. The lower part of the network is freezed and only the top of the network is trained with the training part of the dataset described in previous section.

To further increase the accuracy and reduce losses, artificial data augmentation is used to increases the diversity of the data available for model training without having to collect new data. Image cropping, rotating, shearing and horizontal flipping is used, which resulted with 17 500 images in the training set.

The network was trained on the machine whose main features are: processor - AMD Ryzen 7 3800X, graphics card - Nvidia GTX 1060 6 GB and RAM - 32 GB. CNNs were trained using a graphics card. Libraries on which neural networks are based and which are the basis of this paper are OpenCV (OpenCV, 2022), Keras (*Keras* 2022), TensorFlow (*TensorFlow* 2022) and scikit-learn (*scikit-learn* 2022).

In order to train the network, different hyperparameter values, shown in Table 1, have been utilized and the best results have been obtained by using learning rate of 0.0001, batch size of 32, dropout rate of 0.6 and training on 100 epochs. The training and validation results can be seen in Figure 3.

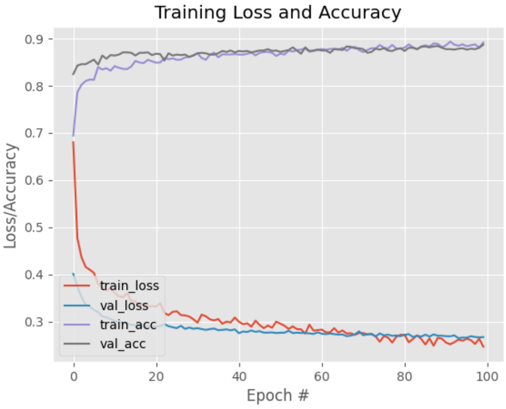

Figure 3: Training Loss and Accuracy

4.3 Testing and Results

As mentioned in the previous section, 20% of the dataset is used to test the model performance, while 80% of dataset has been used for training and validation. During the network validation, the accuracy was calculated, but accuracy is not always sufficient, especially if we are dealing with an unbalanced dataset. To this end, a comprehensive testing and creation of a confusion matrix has been done from which Precision, Recall, Accuracy and F1-score measures were calculated.

Information concerning actual and predicted classifications made by a classification system is contained in the confusion matrix. The data in the matrix is frequently used to evaluate the performance of such systems.

Precision is the proportion of positive cases that were correctly identified (Pedregosa et al., 2011):

$$Precision = \frac{TP}{TP + FP}. \tag{1}$$

Table 2: General Confusion matrix (Pedregosa et al., 2011)

	Predict Class 1	Predict Class 2
Actual Class 1	TP	FN
Actual Class 2	FP	TN

Recall is the proportion of actual positive cases which are correctly identified (Pedregosa et al., 2011):

$$Recall = \frac{TP}{TP + FN}. \tag{2}$$

Accuracy is the proportion of all cases that were correctly identified (Pedregosa et al., 2011):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3}$$

F1-score is a harmonic mean of Precision and Recall, calculated as (Pedregosa et al., 2011):

$$F1\text{-}score = 2 * \frac{Precision * Recall}{Precision + Recall}. \tag{4}$$

Table 3 shows the testing results for the proposed approach and the examples of moustache detection can be seen in Figure 4.

Table 3: Test results

	Precision	Recall	F1-score
Moustache	0.87	0.90	0.89
No moustache	0.91	0.87	0.89
Overall Accuracy			0.89

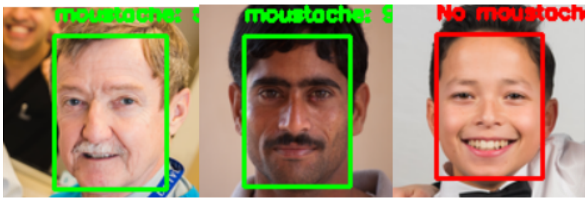

Figure 4: Examples of moustache detection

The results achieved within the context of this study cannot be directly compared with prevailing state-of-the-art achievements. This divergence emanates primarily from the datasets employed and the categories used for classification. Most datasets used in literature are private datasets or parts of the publicly available dataset are used without specification which parts or how many images. It is essential to underscore that a definitive, universally acknowledged benchmark dataset for facilitating direct comparisons of moustache detection outcomes, has not yet been established within the scientific community. Presently, a conspicuous gap exists in the domain of moustache detection evaluations, largely due to the scarcity of a standardized dataset that engenders a level playing field for comprehensive result assessment. An inherent predicament pertains to the prevalent datasets, often characterized by a notable class imbalance, wherein the number of non-moustache images significantly outweighs their moustache-bearing counterparts. This skewed distribution invariably renders conventional accuracy metrics insufficient for performance evaluation.

5 Conclusion

In this paper, we have provided an overview of the current state of the art in the field of moustache detection. Additionally, we have presented a solution that applies deep transfer learning techniques for detecting face images with moustaches. We have employed a version of the MobileNetV2 convolutional neural network and Flickr-Faces-HQ Dataset with a large variation of faces of different ages, ethnicities, backgrounds, and faces with accessories. We used the holdout method for training and testing our model. Specifically, we divided the dataset into two parts: 80% of the images from each class were utilized for training, while the remaining 20% were reserved for testing and evaluating the performance of the model. The network was pretrained on ImageNet dataset, and the weights provided were used in transfer learning for moustache detection.

For this research, a new balanced dataset specifically designed for moustache detection was created. This dataset was used to train and test the model in order to evaluate its performance. The evaluation of the model's performance was reported in Section 4, focusing on precision, recall, accuracy and F1-score as the performance metrics.

Deep transfer learning showed to be a promising technique for detecting moustaches on face images, where by leveraging the knowledge learned from large datasets of face images, it could improve the accuracy of moustache detection, even in challenging conditions. Future research in this area should focus on improving the accuracy and robustness of deep transfer learning algorithms for moustache detection, as well as exploring new applications of moustache detection in various fields.

Acknowledgments

This research was funded by the project "Development of CSTI platform for retrieval and analysis of structured and unstructured data". The project received funding from the European Regional Development Fund through OP Competitiveness and Cohesion 2014–2020 within the Call for Proposals "Development of the products and services arising from research and development activities—Phase II", under grant number KK.01.2.1.02.0310.

References

Almghraby, M., & Elnady*, A. O. (2021). Face Mask Detection in Real-Time using MobileNetv2. *International Journal of Engineering and Advanced Technology*, *10*(6), 104–108. https://doi.org/10.35940/ijeat.F3050.0810621

Face Recognition Data. (n.d.). https://cmp.felk.cvut.cz/spacelib/faces/faces94.html.

Goodfellow, I. J., Erhan, D., Luc Carrier, P., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., ... Bengio, Y. (2015). Challenges in representation learning: A report on three machine learning contests [Special Issue on "Deep Learning of Representations"]. *Neural Networks*, *64*, 59–63. https://doi.org/https://doi.org/10.1016/j.neunet.2014.09.005

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2008). Multi-pie. *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, 1–8. https://doi.org/10.1109/AFGR.2008.4813399

Gudi, A. (2016). Recognizing semantic features in faces using deep learning. https://doi.org/10.48550/arXiv.1512.00743

Hand, E. M., & Chellappa, R. (2016). Attributes for improved attributes: A multi-task network for attribute classification. https://doi.org/10.48550/arXiv.1604.07360

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. https://doi.org/10.48550/ARXIV.1704.04861

Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). *Labeled faces in the wild: A database for studying face recognition in unconstrained environments* (tech. rep. No. 07-49). University of Massachusetts, Amherst.

Huu, P. N., Quang, V. T., Le Bao, C. N., & Minh, Q. T. (2022). Proposed Detection Face Model by MobileNetV2 Using Asian Data Set. *Journal of Electrical and Computer Engineering*, *2022*, e9984275. https://doi.org/10.1155/2022/9984275

Karras, T., Laine, S., & Aila, T. (2018). A style-based generator architecture for generative adversarial networks. https://doi.org/10.48550/ARXIV.1812.04948

Keras. (2022). https://keras.io/

Kumar, B. A., & Bansal, M. (2023). Face Mask Detection on Photo and Real-Time Video Images Using Caffe-MobileNetV2 Transfer Learning. *Applied Sciences*, *13*(2), 935. https://doi.org/10.3390/app13020935

Le, T. H. N., Luu, K., Seshadri, K., & Savvides, M. (2012). Beard and mustache segmentation using sparse classifiers on self-quotient images. *2012 19th IEEE International Conference on Image Processing*, 165–168. https://doi.org/10.1109/ICIP.2012.6466821

Le, T. H. N., Luu, K., Zhu, C., & Savvides, M. (2017). Semi self-training beard/moustache detection and segmentation simultaneously. *Image and Vision Computing*, *58*, 214–223. https://doi.org/10.1016/j.imavis.2016.07.009

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. *Proceedings of International Conference on Computer Vision (ICCV)*.

Multiple Biometric Grand Challenge (MBGC) Presentations. (2010). *NIST*
Last Modified: 2021-06-02T18:40-04:00.

Narduzzi, S., Türetken, E., Thiran, J.-P., & Dunbar, L. A. (2022). Adaptation of MobileNetV2 for Face Detection on Ultra-Low Power Platform. *2022 9th Swiss Conference on Data Science (SDS)*, 1–6. https://doi.org/10.1109/SDS54800.2022.00008

OpenCV. (2022). *Opencv*. https://opencv.org/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, *12*(null), 2825–2830.

Phillips, P., Moon, H., Rauss, P., & Rizvi, S. (1997). The feret evaluation methodology for face-recognition algorithms. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 137–143. https://doi.org/10.1109/CVPR.1997.609311

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. https://doi.org/10.48550/ARXIV.1801.04381

Scikit-learn. (2022). https://scikit-learn.org/

Sghaier, S. M., & Elfaki, A. O. (2021). Efficient techniques for human face occlusions detection and extraction. *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)*, 1–5. https://doi.org/10.1109/WiDSTaif52235.2021.9430214

Tensorflow. (2022). https://www.tensorflow.org/

Ubaid, M. T., Khalil, M., Khan, M. U. G., Saba, T., & Rehman, A. (2022). Beard and hair detection, segmentation and changing color using mask r-cnn. In A. Ullah, S. Anwar, Á. Rocha, & S. Gill (Eds.), *Proceedings of international conference on information technology and applications* (pp. 63–73). Springer Nature. https://doi.org/10.1007/978-981-16-7618-5_6

Vicarious Perception Technologies. (2023). *VicarVision*. https://vicarvision.nl/.

Yang, S., Luo, P., Loy, C. C., & Tang, X. (2018). Faceness-net: Face detection through deep facial part responses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(8), 1845–1859. https://doi.org/10.1109/TPAMI.2017.2738644