Attention-based Gender Recognition on Masked Faces

Vincenzo Carletti, Antonio Greco, Alessia Saggese and Mario Vento
Dept. of Information Eng., Electrical Eng. and Applied Mathematics (DIEM), University of Salerno, Italy

Keywords: Gender Recognition, Attention Mechanism, Masked Faces.

Abstract: Gender recognition from face images can be profitably used in several vertical markets, such as targeted advertising and cognitive robotics. However, in the last years, due to the COVID-19 pandemic, the unreliability of such systems when dealing with faces covered by a mask has emerged. In this paper, we propose a novel architecture based on attention layers and trained with a domain specific data augmentation technique for reliable gender recognition of masked faces. The proposed method has been experimentally evaluated on a huge dataset, namely VGGFace2-M, a masked version of the well known VGGFace2 dataset, and the achieved results confirm an improvement of around 4% with respect to traditional gender recognition algorithms, while preserving the performance on unmasked faces.

1 INTRODUCTION

Face analysis, such as gender (Ng et al., 2015) and ethnicity recognition (Greco et al., 2020a) or age estimation (Carletti et al., 2020), attracted a growing interest of the scientific community in the last decade. Indeed, the above tasks can be considered as the base for a huge amount of different applications, such as targeted advertising (Greco et al., 2020b), forensic search for video surveillance or social robots (Foggia et al., 2019). Among the different face analysis tasks, in this paper we will focus on the gender recognition problem.

In the last years plenty of methods have been proposed for gender recognition from face images, which adopt color, shape and texture features (Zhang et al., 2016) (Azzopardi et al., 2016a) (Azzopardi et al., 2017) (Carletti et al., 2020), trainable features (Azzopardi et al., 2016b) (Simanjuntak and Azzopardi, 2019) (Azzopardi et al., 2018a), a combination of handcrafted and trainable features (Azzopardi et al., 2018b), single convolutional neural networks (CNNs) (Levi and Hassner, 2015) (Antipov et al., 2017) (Greco et al., 2020), multi-task CNNs (Ranjan et al., 2017) (Dehghan et al., 2017) (Gurnani et al., 2019) or ensemble of CNNs (Afifi and Abdelhamed, 2019) (Antipov et al., 2016). These approaches achieve over 95% of accuracy on facial images collected in controlled environments, but their performance decrease in presence of strong variations (pose, blur, noise, partial occlusions and so on). To further investigate the effect of image corruptions over gender recognition algorithms, a framework has been recently proposed and made publicly available (Carletti et al., 2020).

The spread of the COVID-19 pandemic has made the gender recognition problem even more difficult, since the masks occlude part of the faces. A deep analysis, involving seven different deep architecture devised for gender recognition, has been conducted in (Greco et al., 2021), where it was demonstrated that the performance drop for this specific problem is about 10%. Even if this is not dramatic, like in the case of emotion recognition, for which the accuracy drop is instead about 50%, some effort is still required, aiming at improving the performance of gender recognition algorithms in presence of masked faces, but mostly important without degrading the performance in case of unmasked faces.

The problem of occlusions, also including scarves or glasses, is among the main challenges of face analysis tasks, and in recent years several solutions have been proposed for solving this problem, especially in face recognition tasks. One of the first approaches for dealing with occlusions has been proposed in (Min et al., 2011): the face ROI is firstly split into two horizontal half and, for each part, the presence of scarf/sunglasses is detected by Gabor wavelets, letting the model processing only the non-occluded facial regions with a SVM classifier. In 2019 and in 2020, an increasing number of face detection and recognition and expression analysis algorithm have

Carletti, V., Greco, A., Saggese, A. and Vento, M.
Attention-based Gender Recognition on Masked Faces.
DOI: 10.5220/0010978700003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 672-678
ISBN: 978-989-758-555-5; ISSN: 2184-4321

been proposed for dealing with occlusions, due to the COVID-19 pandemic. A common trend has been to introduce in the architectures visual attention mechanisms (Li et al., 2019)(Xu et al., 2020)(Yuan, 2020). The attention tries to emulate a cognitive process adopted by the human brain when it focuses only on some areas of the image most promising for the extraction of the salient features concerning the problem to be solved. To emulate this behaviour, the attention mechanism associates different weights to the different features extracted by the network.

Another commonly adopted strategy has been the use of augmentation techniques for improving age and gender recognition algorithms in presence of occluded faces (Hsu et al., 2021). The authors introduce three occlusion techniques, namely blackout, random brightness and blur, designed to simulate different challenges that could be experienced in real-world applications. Anyway, even if interesting, these policies are not domain dependent and, thus, can not really solve the issue related to the presence of the mask covering the face. The idea could be instead to augment the training set with some domain specific policies, such as faces covered by a mask, obtained by adding synthetic occlusions to the image.

Starting from the above considerations, in this paper we propose a reliable gender recognition algorithm, based on the attention mechanism and trained with a domain specific data augmentation technique, able to achieve a remarkable accuracy both in presence and in absence of masks occluding the faces. According to our knowledge, this is the first time that domain specific techniques data augmentation and attention mechanisms are used together to improve performance over masked faces of gender recognition algorithms.

2 THE PROPOSED APPROACH

The method is based on the standard processing pipeline for these types of systems: face detection and gender recognition. The architecture of the proposed approach is depicted in Figure 1.

In the face detection step we localize the face in the whole image and crop the region that must be given as input to the classifier. The most popular face detectors recently demonstrated a great accuracy (Li et al., 2021), but they are not so effective on faces covered by a mask, since the common datasets adopted for face detection do not include masked faces. Therefore, for our method we trained our own face detector based on MobileNetv2-SSD (Sandler et al., 2018), by using all the images available in the training sets of WiderFace (Yang et al., 2016) and MAFA (Ge et al., 2017); the evaluation of the accuracy of this module is out of the scope of this paper, but it has been able to locate and crop all the masked and unmasked faces in the dataset we used to train, validate and test our method.

In the gender recognition step we apply our CNN on the face image, resized to 224×224. The proposed CNN is a custom version of ResNet50 (He et al., 2016), in which we have modified the Residual Unit (RU) to add an attention module. In particular, we adopt the visual attention mechanism (Liu and Milanova, 2018) to emulate the way in which the human brain works: humans are able to dynamically locate the region of interest and analyze the scene by selectively processing subsets of the visual input. In the visual attention mechanisms, the main idea is to associate different weights to the different features extracted by the network, with the aim to focus not on the whole image but instead only on some areas of the image, namely the most promising ones. For our problem at hands, the reason why we propose to introduce attention layer is to use the whole face when available, while only the top part of the face when it is covered by a mask.

To this aim, we followed the method proposed in (Hu et al., 2018), namely we weigh the output feature block of the RU with weights computed by an additional attention module (see Figure 1). There are different visual attention mechanisms, but we selected the ones that use or combine channel and spatial attention, that are suited for our purposes. In particular, we chose three of them, widely adopted in the scientific literature, namely the Convolutional Block Attention Module (CBAM) (Woo et al., 2018), the spatial and channel Squeeze and Excitation (scSE) (Roy et al., 2018) and the Efficient Channel Attention (ECA) (Wang et al., 2020).

CBAM infers, given an initial feature map, the attention weights along the channel and spatial dimensions separately. The computed attention maps are then multiplied with the input feature map to refine the features. scSE still uses both channel and spatial dimensions, but concurrently; indeed, scSE computes spatial and channel-wise attention maps by concurrently recalibrating the input spatially and channel-wise. This module is an extended version of the well known SE attention module, which is the base of SENet architecture (Hu et al., 2018). It is important to highlight that both CBAM and scSE compute spatial attention using a 2D convolution of kernel size $K \times K$. Differently, ECA investigates a 1D convolution with adaptive kernel size to replace the fully connected layers in the channel attention module.

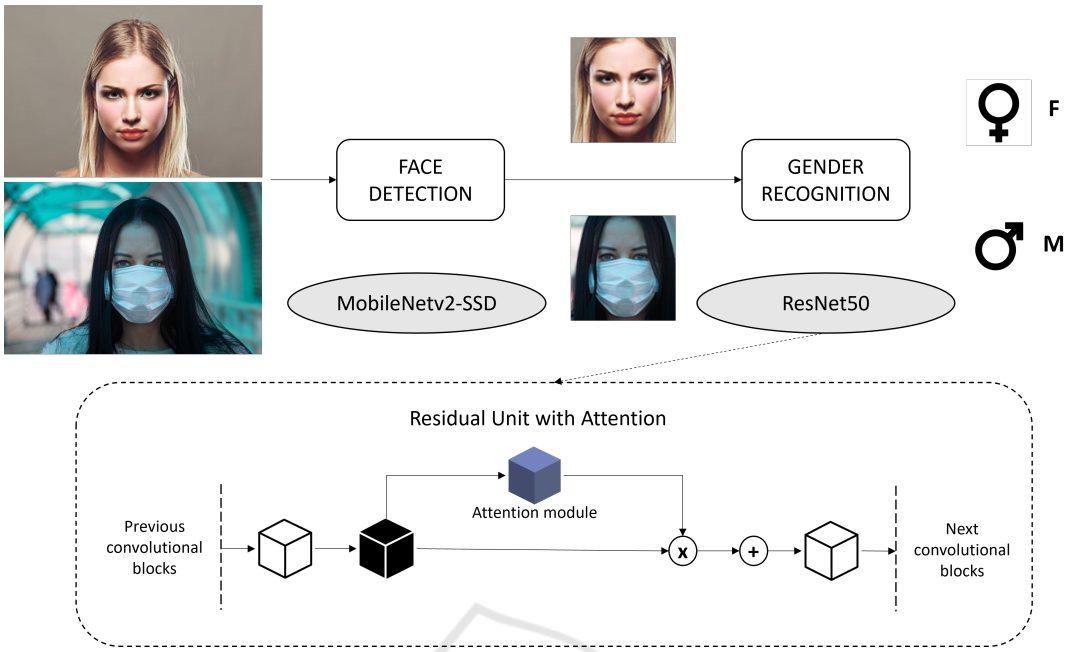


Figure 1: Overview of the proposed method. Face detection is performed with a custom detector based on MobileNet-SSD and trained with masked and unmasked faces. Gender recognition is carried out with a modified version of ResNet50, in which the Residual Unit is extended with an attention module; this change allows to dynamically focus on the visible portion of the face.

3 EXPERIMENTAL RESULTS

In this Section we will detail the dataset used in our experiments (see Subsection 3.1), the training procedure (see Subsection 3.2) and finally the achieved results (see Subsection 3.3) and a visual explanation of the experimental findings (see Subsection 3.4).

3.1 Dataset

We conducted our experimentation over the VGGFace2-M dataset, that we recently proposed in (Greco et al., 2021). As well as the original VGGFace2 (Cao et al., 2018) dataset, it contains over 3 millions of face images belonging to 9,131 different subjects, annotated with gender.

Basically, it includes the same face images of the original dataset, but covered with a synthetically added mask. It has been produced by using the approach proposed in (yuan Wang et al., 2020): the dlib face detector (Kazemi and Sullivan, 2014) has been applied and 68 facial landmarks have been identified on each face. Thus, the following four points have been used to determine the position and the shape of the mask on the face, namely the nose point, the chin bottom point, the chin right point and the chin left point. The left part of the mask is built by using

Figure 2: Images from the VGGFace2-M dataset.

the chin left point, the chin bottom point and the nose point. The right part is obtained symmetrically with the chin right point. The two parts are then merged, by obtaining a mask whose height is equal to the distance between the nose point and the bottom chin point and whose width is equal to the sum of the widths of the two parts. The mask is finally rotated of the angle between the nose point and the extremity of one of the eyes. Some example of masked faces belonging to the VGGFace2-M dataset are reported in Figure 2.

3.2 Training Procedure

The proposed architecture has been trained through a SGD optimizer (Liu et al., 2020). The weights have been pre-trained over the ImageNet dataset and a fine tuning of all the layers has been performed. The network has been trained for 70 epochs with batches of 128 samples and a learning rate starting from 0.005 with a decay of factor equal to 5 every 20 epochs. The adopted loss function is the categorical cross-entropy, with a regularization weight decay of 0.005.

Different types of general purpose data augmentation strategies commonly adopted for face analysis tasks are applied: random rotation ($\pm 10^{\circ}$), shear (up to 10%), cropping (up to 5% variation over the original face region), horizontal flipping (with 50% probability), change of brightness (up to $\pm 20\%$ of the available range) and contrast(up to $\pm 50\%$ of the maximum value). The chosen ranges have been experimentally chosen in order to reproduce conditions that are likely to occur in real scenarios. Also, a domain specific data augmentation technique has been performed: indeed, masked faces from the VGGFace2-M dataset have been included in the dataset for training.

3.3 Overall Results

The obtained results are summarized in Table 1, where we report the results in terms of accuracy on the test set of both the datasets VGGFace2 and VGGFace2-M. The drop over masked faces with respect to unmasked face is also in the table (Drop), together with the average accuracy value (Avg).

Also, as we can see from column *Training set*, two different policies have been considered for training:

- *VGGFace2-M*: only the masked version of VGGFace2 dataset, namely VGGFace2-M, without any unmasked face, has been used for training. In this case, we suppose that any subject is wearing a mask. Anyway, we also evaluated the performance over unmasked faces.
- *VGGFace2-UM*: the union of the two datasets, namely VGGFace2 and VGGFace2-M, have been used for training.

We also include in the table results achieved by the baseline solution, namely the model with the best performance over both VGGFace2 and VGGFace2-M, as reported in (Greco et al., 2021). It is a VGG-16 CNN trained for gender recognition over the VGGFace2 dataset. We can note that this result is among the best ones on VGGFace2 (97.60%), but also, as expected, the worst one on masked faces (92.99%), with a performance drop of 4.61%.

In average, the best result is obtained when ResNet50+CBAM model is trained on VGGFace2-UM, achieving the best average accuracy of 97.23% and the best performance over unmasked faces of 97.61%, with a drop of 0.76% (vs 4.61% of the baseline solution). It is important to highlight that the accuracy achieved over the unmasked VGGFace2 is also higher than the VGG baseline network trained over the unmasked training set. Even if this architecture does not achieve the best result over masked faces, it is very close to it, being 96.85% vs 97.02% the accuracy achieved by the same architecture (ResNet50+CBAM) trained with VGGFace2-M.

Also, it is important to note that all the architectures based on attention layers achieve better results than the baseline solution. This achievement further confirms that the attention mechanism allows to simultaneously and reliably manage both masked and unmasked faces, reaching state of the art accuracy when dealing with unmasked faces but reporting only a very slight drop in the performance (lower than 1%) also over masked faces. Therefore, it can be considered definitively a viable solution for the problem at hand.

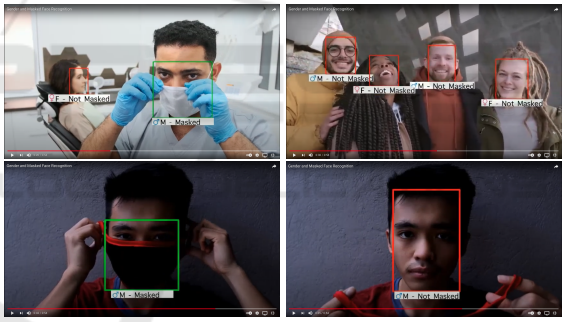


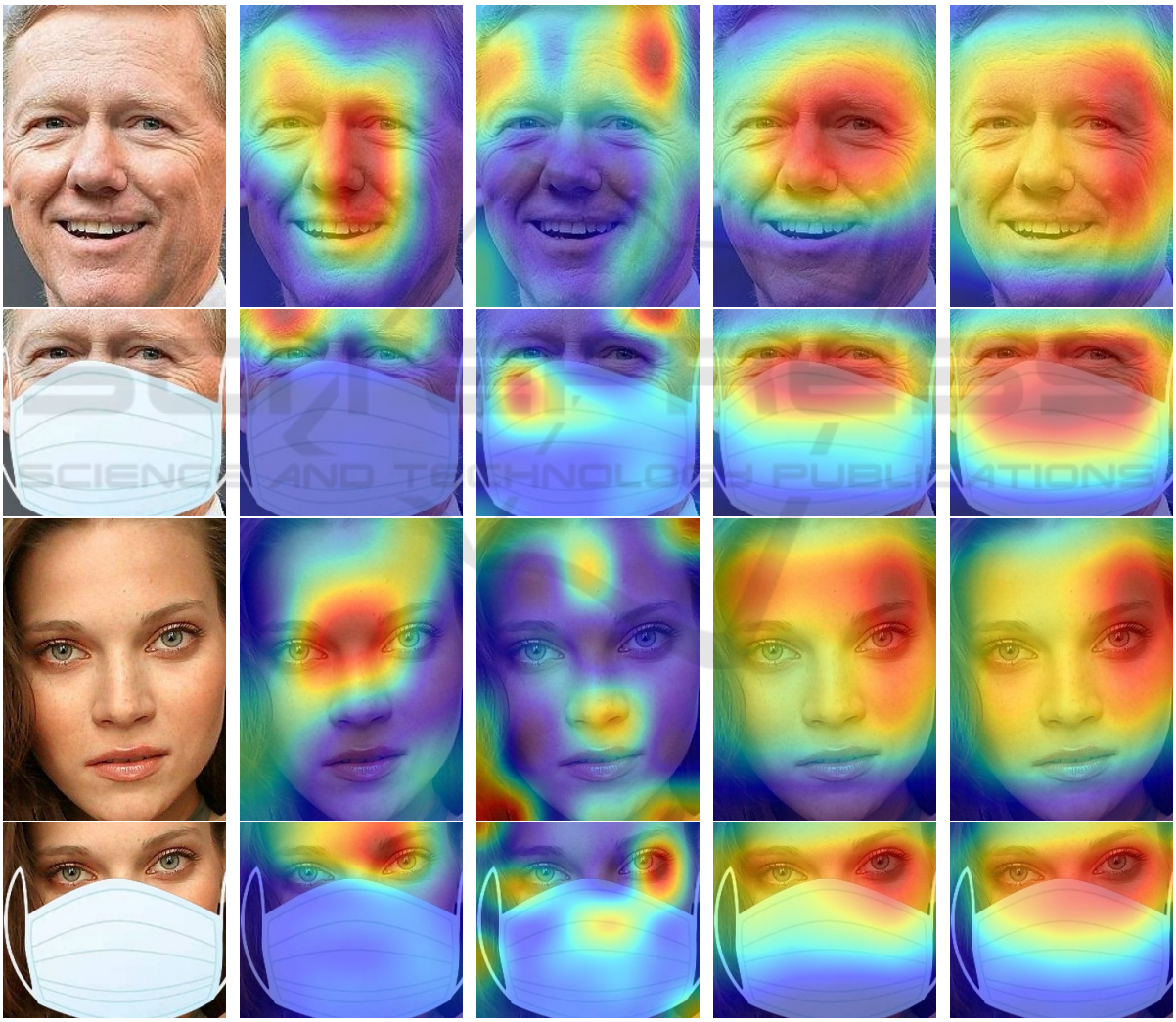
Figure 3: Examples of the proposed system in action. M and F in the image indicate Male and Female, respectively. For each face, the presence of the mask is represented with a green bounding box, while the absence is indicated with a red rectangle.

3.4 Explainability of the Results

In this section we aim to explain the results obtained by the proposed network through the visualization of the discriminative features learned by the CNN. To this purpose, we computed the class activation maps (Zhou et al., 2016) to identify the region of the face mostly used by the CNNs for recognizing the gender. The class activation map is a heat map computed for every pixel of the image, in which the ”hot” pixels, coloured with red gradations, represent regions of the image mostly used by the CNN to extract discriminative features for the classification. We used Grad-CAM (Selvaraju et al., 2017) to compute the

Table 1: Accuracy achieved over both VGGFace2 and VGGFace2-M by varying the specific model considered and the training set. In bold we report the best result, namely the best accuracy over VGGFace2 and VGGFace2-M, the minimum drop and the highest average accuracy.

Model	**Training set**	**Accuracy (%)**			
		VGGFace2	**VGGFace2-M**	**Drop**	**Avg.**
VGG	VGGFace2	97.60	92.99	4.61	95.30
VGG	VGGFace2-M	95.72	96.37	-0.66	96.05
ResNet50+ECA	VGGFace2-M	96.58	96.87	-0.29	96.73
ResNet50+CBAM	VGGFace2-M	96.81	**97.02**	-0.21	96.92
ResNet50+scSE	VGGFace2-M	96.76	96.96	**-0.20**	96.86
ResNet50+ECA	VGGFace2-UM	97.54	96.77	0.78	97.15
ResNet50+CBAM	VGGFace2-UM	**97.61**	96.85	0.76	**97.23**
ResNet50+scSE	VGGFace2-UM	97.54	96.78	0.76	97.16



(a) Reference (b) VGG-U (c) VGG-M (d) CBAM-M (e) CBAM-UM

Figure 4: Class activation maps computed on (a) male and female unmasked and masked reference faces for (b) VGG16 trained on VGGFace2, (c) VGG16 trained on VGGFace2-M, (d) ResNet50+CBAM trained on VGGFace2-M and (e) ResNet50+CBAM trained on VGGFace2-UM.

class activation maps of the two versions of VGG and ResNet50+CBAM and reported some examples in Figure 4.

It is evident that the two versions of VGG do not use the whole part of the face visible in presence of mask; the version trained with VGGFace2-M not only suffer of the above mentioned problem, but also uses very small parts of the face even when it is completely visible. On the other hand, the proposed ResNet50+CBAM is able to focus the attention on the whole region of the face that is visible; in presence of mask, it uses the region of the eyes and the upper part of the nose, while on unmasked faces the CNN takes advantage of the features of the whole face to perform the classification. It is particularly evident on the version trained with VGGFace2-UM, while the other focuses more on the upper part of the face, being trained only with masked faces.

4 CONCLUSIONS

State of the art algorithms for gender recognition from face images, even if very reliable, suffer a performance drop in presence of occlusions covering the face. This has a great impact especially during the COVID-19 pandemic, when wearing a mask became mandatory in several countries all around the world. Starting from this assumption, in this paper we have presented a novel algorithm exploiting the attention mechanism and a domain specific data augmentation policy able to achieve on average a 4% improvement with respect to traditional state of the art gender recognition algorithm, while preserving the performance over unmasked faces. It is due to the capability of the network, demonstrated through a visual investigation with class activation maps, to adaptively select the discriminative region of interests, namely the visible parts of the face.

Future works include the extension of this work to real masked faces and to other face analysis tasks that still suffer from the presence of the mask, such as age estimation or ethnicity recognition.

REFERENCES

Afifi, M. and Abdelhamed, A. (2019). Afif4: deep gender classification based on adaboost-based fusion of isolated facial features and foggy faces. *Journal of Visual Communication and Image Representation*, 62:77–86.

Antipov, G., Baccouche, M., Berrani, S.-A., and Dugelay, J.-L. (2017). Effective training of convolutional neural networks for face-based gender and age prediction. *Pattern Recognition*, 72:15–26.

Antipov, G., Berrani, S.-A., and Dugelay, J.-L. (2016). Minimalistic cnn-based ensemble model for gender prediction from face images. *Pattern Recognition Letters*, 70:59–65.

Azzopardi, G., Foggia, P., Greco, A., Saggese, A., and Vento, M. (2018a). Gender recognition from face images using trainable shape and color features. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1983–1988. IEEE.

Azzopardi, G., Greco, A., Saggese, A., and Vento, M. (2017). Fast gender recognition in videos using a novel descriptor based on the gradient magnitudes of facial landmarks. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE.

Azzopardi, G., Greco, A., Saggese, A., and Vento, M. (2018b). Fusion of domain-specific and trainable features for gender recognition from face images. *IEEE Access*, 6:24171–24183.

Azzopardi, G., Greco, A., and Vento, M. (2016a). Gender recognition from face images using a fusion of svm classifiers. In *International Conference on Image Analysis and Recognition*, pages 533–538. Springer.

Azzopardi, G., Greco, A., and Vento, M. (2016b). Gender recognition from face images with trainable cosfire filters. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 235–241. IEEE.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 67–74.

Carletti, V., Greco, A., Percannella, G., and Vento, M. (2020). Age from faces in the deep learning revolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9):2113–2132.

Carletti, V., Greco, A., Saggese, A., and Vento, M. (2020). An effective real time gender recognition system for smart cameras. *J. Ambient Intell. Humaniz. Comput.*, 11(6):2407–2419.

Dehghan, A., Ortiz, E. G., Shu, G., and Masood, S. Z. (2017). Dager: Deep age, gender and emotion recognition using convolutional neural network. *arXiv preprint arXiv:1702.04280*.

Foggia, P., Greco, A., Percannella, G., Vento, M., and Vigilante, V. (2019). A system for gender recognition on mobile robots. In *Proceedings of the 2nd International Conference on Applications of Intelligent Systems*, pages 1–6.

Ge, S., Li, J., Ye, Q., and Luo, Z. (2017). Detecting masked faces in the wild with lle-cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2682–2690.

Greco, A., Percannella, G., Vento, M., and Vigilante, V. (2020a). Benchmarking deep network architectures for ethnicity recognition using a new large face dataset. *Machine Vision and Applications*.

Greco, A., Saggese, A., and Vento, M. (2020b). Digital signage by real-time gender recognition from face images. In *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*, pages 309–313.

Greco, A., Saggese, A., Vento, M., and Vigilante, V. (2020). A convolutional neural network for gender recognition optimizing the accuracy/speed tradeoff. *IEEE Access*, 8:130771–130781.

Greco, A., Saggese, A., Vento, M., and Vigilante, V. (2021). Performance assessment of face analysis algorithms with occluded faces. In *International Conference on Pattern Recognition*, pages 472–486. Springer.

Gurnani, A., Shah, K., Gajjar, V., Mavani, V., and Khandhediya, Y. (2019). Saf-bage: Salient approach for facial soft-biometric classification-age, gender, and facial expression. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hsu, C.-Y., Lin, L.-E., and Lin, C. (2021). Age and gender recognition with random occluded data augmentation on facial images. *Multimedia Tools and Applications*.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141.

Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874.

Levi, G. and Hassner, T. (2015). Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42.

Li, X., Lai, S., and Qian, X. (2021). Dbcface: Towards pure convolutional neural network face detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Li, Y., Zeng, J., Shan, S., and Chen, X. (2019). Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450.

Liu, X. and Milanova, M. (2018). Visual attention in deep learning: a review. In *ICRA 2018*.

Liu, Y., Gao, Y., and Yin, W. (2020). An improved analysis of stochastic gradient descent with momentum. *arXiv: Optimization and Control*.

Min, R., Hadid, A., and Dugelay, J. (2011). Improving the recognition of faces occluded by facial accessories. In *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 442–447.

Ng, C.-B., Tay, Y.-H., and Goi, B.-M. (2015). A review of facial gender recognition. *Pattern Analysis and Applications*, 18(4):739–755.

Ranjan, R., Patel, V. M., and Chellappa, R. (2017). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135.

Roy, A. G., Navab, N., and Wachinger, C. (2018). Concurrent spatial and channel squeeze & excitation in fully convolutional networks. *ArXiv*, abs/1803.02579.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Simanjuntak, F. and Azzopardi, G. (2019). Fusion of cnn- and cosfire-based features with application to gender recognition from face images. In *Science and Information Conference*, pages 444–458. Springer.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q. (2020). Eca-net: Efficient channel attention for deep convolutional neural networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11531–11539.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I.-S. (2018). Cbam: Convolutional block attention module. In *ECCV*.

Xu, X., Sarafianos, N., and Kakadiaris, I. (2020). On improving the generalization of face recognition in the presence of occlusions. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3470–3480.

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

Yuan, Z. (2020). Face detection and recognition based on visual attention mechanism guidance model in unrestricted posture. *Scientific Programming*, 2020:1–10.

yuan Wang, Z., Wang, G., Huang, B., Xiong, Z., Hong, Q., Wu, H., Yi, P., Jiang, K., Wang, N., Pei, Y., Chen, H., Miao, Y., Huang, Z., and Liang, J. (2020). Masked face recognition dataset and application. *ArXiv*, abs/2003.09093.

Zhang, W., Smith, M. L., Smith, L. N., and Farooq, A. (2016). Gender and gaze gesture recognition for human-computer interaction. *Computer Vision and Image Understanding*, 149:32–50.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.