Gender recognition in the wild: a robustness evaluation over corrupted images

Antonio Greco¹ · Alessia Saggese¹ · Mario Vento¹ · Vincenzo Vigilante²

Received: 28 August 2020 / Accepted: 23 November 2020 / Published online: 23 December 2020
© The Author(s) 2020

Abstract

In the era of deep learning, the methods for gender recognition from face images achieve remarkable performance over most of the standard datasets. However, the common experimental analyses do not take into account that the face images given as input to the neural networks are often affected by strong corruptions not always represented in standard datasets. In this paper, we propose an experimental framework for gender recognition “in the wild”. We produce a corrupted version of the popular LFW+ and GENDER-FERET datasets, that we call LFW+C and GENDER-FERET-C, and evaluate the accuracy of nine different network architectures in presence of specific, suitably designed, corruptions; in addition, we perform an experiment on the MIVIA-Gender dataset, recorded in real environments, to analyze the effects of mixed image corruptions happening in the wild. The experimental analysis demonstrates that the robustness of the considered methods can be further improved, since all of them are affected by a performance drop on images collected in the wild or manually corrupted. Starting from the experimental results, we are able to provide useful insights for choosing the best currently available architecture in specific real conditions. The proposed experimental framework, whose code is publicly available, is general enough to be applicable also on different datasets; thus, it can act as a forerunner for future investigations.

Keywords Gender recognition · Image corruptions · Deep learning

1 Introduction

Face analysis has attracted in the last years several scientists working in artificial intelligence and computer vision, and a huge amount of contributions has been produced. Nowadays,

plenty of researchers are working on face recognition (Guo and Zhang 2019), gender recognition (Ng et al. 2015), age estimation (Carletti et al. 2020b), emotion analysis (Li and Deng 2020) and ethnicity recognition (Fu et al. 2014). Among them, gender recognition may be considered the easiest face analysis task, being a binary classification problem. Nevertheless, the recent literature shows that the accuracy achieved by the modern convolutional neural networks (CNNs) for gender recognition strongly depends on the environmental conditions, the quality of the image and other face variations (occlusions, different orientations) (Greco et al. 2020). Table 1 shows the top-3 results achieved over the most popular datasets, namely LFW+, GENDER-FERET, Adience and MIVIA-Gender, described in the following.

Labeled Faces in the Wild, popular with the name LFW (Huang et al. 2008), is the most widely used dataset for gender recognition. The term “in the wild” refers to the fact that people are not framed in controlled laboratory conditions. However, it is far from representing the challenging conditions happening in real environments, since the images depict famous people photographed with professional cameras. Therefore, excluding pose, age and ethnicity variations,

This research was partially supported by the Italian MIUR within PRIN 2017 grants, Projects Grant 20172BH297 002CUP D44I17000200005 and by A.I. Tech srl (www.aitech.vision).

✉ Alessia Saggese
asasgese@unisa.it

Antonio Greco
agreco@unisa.it

Mario Vento
mvento@unisa.it

Vincenzo Vigilante
vvigilante@unisa.it; vvigilante@aitech.vision

¹ Department of Computer Engineering, Electrical Engineering and Applied Mathematics (DIEM), University of Salerno, Salerno, Italy

² A.I. Tech srl, Fisciano, Italy

Table 1 Methods achieving the top-3 accuracy (Acc.) over LFW+, GENDER-FERET (GF), Adience and MIVIA-Gender (MG)

Dataset	Method	Acc.
LFW+	Antipov et al. (2017)	0.993
	Antipov et al. (2016)	0.971
	Jia and Cristianini (2015)	0.968
GF	Simanjuntak and Azzopardi (2019)	0.974
	Basulaim and Dabash (2019)	0.958
	Azzopardi et al. (2018b)	0.947
Adience	Gurnani et al. (2019)	0.918
	Dehghan et al. (2017)	0.910
	Afifi and Abdelhamed (2019)	0.906
MG	Azzopardi et al. (2018b)	0.915
	Carletti et al. (2020a)	0.902
	Azzopardi et al. (2017)	0.802

the dataset includes few corruptions that are common in real environments, such as blur, noise and challenging lighting conditions. Zhang et al. (2016) achieve 0.925 accuracy on the LFW dataset by using a SVM classifier trained with a fusion of texture and shape features. Ranjan et al. (2017) show that Hyperface, their multi-task network able to perform face detection, landmark localization, pose estimation and gender recognition, is able to obtain an accuracy of 0.940 on the aligned version of the dataset, i.e. LFWA (Huang et al. 2007). Afifi and Abdelhamed (2019) achieve an accuracy of 0.960 with an ensemble of CNNs trained with various facial components (eyes, mouth, nose and so on). Azzopardi et al. (2018b) hold the top-1 0.994 accuracy on the LFW dataset with their stacked classifier combining trainable COSFIRE filters and local SURF features.

It is important to mention that these methods have been trained and tested on the same dataset, without performing cross-dataset evaluation. The latter protocol has been adopted by Han et al. (2017), which tested their multi-task CNN, trained on the CASIA-WebFace images (Yi et al. 2014), over an extended version called LFW+ (Jia and Cristianini 2015), achieving an accuracy of 0.965. Following the same protocol, Jia and Cristianini (2015) show that their method based on multiscale LBP, trained with 4 millions of images, was able to achieve an accuracy of 0.968 on the LFW+ dataset. Antipov et al. (2016) demonstrate that with an ensemble of 3 lightweight convolutional neural networks (CNNs) trained with the CASIA-WebFace images is possible to obtain an accuracy of 0.971 on the same dataset. One year after, Antipov et al. (2017) achieve on the LFW+ the impressive 0.993 accuracy with a CNN inspired by the ResNet-50 architecture.

Another dataset used for recent gender recognition experiments is the GENDER-FERET (Azzopardi et al. 2016a); it is acquired in controlled laboratory conditions, but it

includes only grayscale images. On this dataset, Azzopardi et al. (2016a) achieve an accuracy of 0.926 with a fusion of SVM classifiers fed with raw, texture and shape features. SVM classifiers trained with COSFIRE filters (Azzopardi et al. 2016b) or with a fusion of these features and local SURF descriptors (Azzopardi et al. 2018b) obtain 0.936 and 0.947 of accuracy, respectively. Basulaim and Dabash (2019) further improve the performance to 0.958 by using COSFIRE filters and a cubic SVM. Simanjuntak and Azzopardi (2019) retain the current state of the art performance on this dataset (0.974) with a fusion of CNN and COSFIRE-based features. With a similar approach based on the combination of edge-based and color-based COSFIRE filters Azzopardi et al. (2018a) achieve an accuracy of 0.964 on the GENDER-COLOR-FERET, a RGB version of the GENDER-FERET. In all these cases, the training and the test sets are obtained from the whole GENDER-FERET, so without cross-dataset evaluation.

The performance substantially decreases when the experimental analysis is carried out on more challenging datasets, even by adopting a portion of them for training the methods. An example is the Adience dataset, proposed by Levi and Hassner (2015), which includes strong variations in terms of age, pose, brightness, contrast and image quality; on their set, the authors achieved an accuracy of 0.868 by using a lightweight CNN with a 5-fold cross validation. Van de Wolfshaar et al. (2015) increase the accuracy of 0.872 with a CNN based on AlexNet for extracting the features fed to a SVM classifier. The above mentioned approach proposed by Afifi and Abdelhamed (2019) achieve on this dataset an accuracy of 0.906, 5.4% less than the performance obtained on LFW. A similar accuracy of 0.910 has been achieved by Dehghan et al. (2017) with DAGER, a deep network for age, gender and emotion recognition trained with 4 millions of images. The best performance on the Adience dataset is hold by Gurnani et al. (2019), obtaining an accuracy of 0.918 with SAF-BAGE, a deep network for age, gender and expression recognition.

The MIVIA-Gender dataset, proposed by Azzopardi et al. (2017), was recorded with a standard surveillance camera; therefore, the available images present most of the variations happening in real environments, namely blur, noise and brightness and contrast change. Azzopardi et al. (2017) achieve an accuracy of 0.802 with a SVM fed with a face descriptor based on the gradient magnitudes of facial landmarks, while Carletti et al. (2020a) substantially increase the performance to 0.902 with an algorithm based on HOG features optimized for running on board of smart cameras. Even on the MIVIA-Gender dataset Azzopardi et al. (2018b) hold the top-1 0.915 accuracy, but still a few experiments have been done on these data.

From the analysis of the state of the art, we can note that peaks exceeding 0.95 are reached even with relatively

small architectures on LFW, LFWA, LFW+, GENDER-FERET and GENDER-COLOR-FERET (Del Coco et al. 2016); on the other hand, it emerges the evidence that on different test sets the performance can drop below 0.90, especially with cross-dataset evaluations, as happens for Adience and MIVIA-Gender. Only in a few cases the analysis is carried out on more than one benchmark (Azzopardi et al. 2018b) (Foggia et al. 2019) (Afifi and Abdelhamed 2019) to evaluate the generalization capabilities of the methods and, in most of the papers, the experimentation is performed on images not representing the real conditions. Think, as an example, to the method proposed by Haider et al. (2019), namely a CNN optimized for smartphones; despite the necessity of dealing with the challenging faces captured with the camera of a smartphone, they evaluate the performance on datasets recorded in controlled laboratory conditions (CAS-PEAL-R1 and FEI), achieving an average accuracy of 0.95 that is not easily reproducible in the operating phase.

The main considerations arising from the experimental analyses available in the literature are the following: (i) the conditions causing the performance degradation of the existing neural networks for gender recognition have not yet been deeply investigated; (ii) the methods are not evaluated in the real operating conditions. It is clear that the most challenging datasets contain variations not covered by the simplest ones, that should be further analyzed.

The robustness of the neural networks to image corruptions happening in real environments has been recently deeply studied in Hendrycks and Dietterich (2019). The authors of this research demonstrated that the images acquired with classic surveillance cameras are affected by various types of blur, noise and digital corruptions, which cause a dramatic performance drop of the neural networks.

This is particularly true for faces of unaware people acquired in standard video-surveillance scenarios. Suddenly moving people or individuals walking towards the camera produce face images with motion and zoom blur, in addition to the variations of pose and occlusions that can occur in similar conditions. Blur and noise are due to the movement of the people or to the intrinsic characteristics of the cameras (e.g. specific configurations of the automatic gain control) and the variability of the operating environment (e.g. illumination conditions). The lighting conditions also affect the contrast and the brightness of the face: for example, a face image very dark can be obtained when a person is framed with a strong back-light. All these variations, together with partial occlusions (e.g. dirt on the camera lens), image compression (e.g. Motion JPEG for server-side elaboration) and rescaling (e.g. very small face of a person far from the camera upsampled before feeding it to the neural network), can strongly penalize the gender recognition classifiers; nevertheless, these factors have never been deeply investigated

and should be instead reproduced in a realistic experimental analysis.

Even if such conditions are in part covered by some of the existing datasets, they are not globally considered into common gender recognition experimental frameworks. It is thus required to define a protocol for analyzing the performance of these methodologies in real conditions.

In this scenario, the contribution of the paper is three-fold:

- We make publicly available the code for generating corrupted versions of existing datasets. With this code we produced LFW+C and GENDER-FERET-C, obtained by corrupting the images available in the well known LFW+ and GENDER-FERET datasets, but the procedure can be applied also on different datasets. The experimental framework allows to analyze the robustness of gender recognition approaches to the corruptions affecting face images in real environments.
- We perform an extensive experimental analysis of nine state of the art CNNs on LFW+, LFW+C, GENDER-FERET, GENDER-FERET-C and MIVIA-Gender, in order to evaluate the performance degradation of the gender classifiers when dealing with facial image corruptions in the wild.
- We separately analyze the impact of the different corruption categories on the classification performance, in order to provide useful insights for choosing the best gender recognition methods according to the specific conditions of the operating environment.

2 Experimental framework

2.1 Methods

In our analysis, we selected nine different convolutional neural networks, reported in Table 2. Three of them are the most popular architectures in this field, namely VGG, SENet and DenseNet, since they demonstrated to be very effective in various face analysis and gender recognition tasks. The other six are lightweight architectures (three versions of MobileNet, ShuffleNet, SqueezeNet and Xception), recently proposed for gender recognition since the claimed simplicity of the problem pushed the researchers to propose smaller networks in order to gain processing time, memory and storage space without losing accuracy; we want to investigate the generalization capability of these lightweight networks with respect to corruptions.

In the following, we provide a brief description of all the convolutional neural networks considered in our experimental framework.

Table 2 Convolutional neural networks used in our experimental framework

Network	Input size	Parameters	FPS
VGG	224×224	138.3 M	0.5
SENet	224×224	21.5 M	–
DenseNet	224×224	7.0 M	–
MobileNet-A	224×224	2.3 M	1.4
MobileNet-B	96×96	1.4 M	4.8
MobileNet-C	64×64	0.1 M	14.1
ShuffleNet	224×224	4.0 M	1.8
SqueezeNet	224×224	0.7 M	5.3
Xception	71×71	20.9 M	1.6

For each network, we report the size of the input, the number of parameters and the processing frame rate (FPS) on embedded devices using an ARM v7 CPU 1.2 GHz and 1 GB RAM (not available for models that do not fit in the memory)

VGG. Proposed by Simonyan and Zisserman (2014), it can be considered the most used network for face analysis. One of its versions, namely VGG-Face (Parkhi et al. 2015), was the first architecture trained on a public large scale dataset for the problem of face recognition, and it has been widely applied for face analysis problems. Thanks to the small size of the convolutional kernels (3×3), it is able to achieve good generalization capabilities also for quite small data sets.

SENet. It has been designed by Hu et al. (2018) and recently adopted for face recognition by Cao et al. (2018) and for emotion recognition by Albanie et al. (2018); it is based on the well known ResNet-50 architecture (He et al. 2016), using residual blocks to allow very deep CNNs to be trained, avoiding the vanishing gradient problem. Differently from the original ResNet architecture, SENet adds *Squeeze and Excitation* modules, introducing new connections for computing the feature maps by adaptively weighting the input channels; this architectural change allows to substantially reduce the error in various classification tasks (Hu et al. 2018).

DenseNet. This network, proposed by Huang et al. (2017), is based on two main design choices: dense connectivity and transition layers. Dense connectivity allows to share feature maps between consecutive layers; this mechanism includes a growth rate, that is used to establish how much each layer contributes to the global state. To enable the possibility of sharing features from different layers, the transition layers are in charge of convolution and pooling between connected dense blocks, with the aim to normalize the size of the feature maps computed by the different layers. Among the various DenseNet architectures, we chose the smallest version (DenseNet-121), which demonstrated impressive performance on ImageNet.

MobileNet. Designed by Sandler et al. (2018), it is among the most efficient networks available in the literature, thought for running on board of mobile and embedded devices. The second version (v2) is based on residual blocks and depthwise separable convolutions, i.e. a standard convolution decomposed into a depthwise and a pointwise component. We evaluate three versions of MobileNet v2: the vanilla one (hereinafter MobileNet-A), which has an input size 224×224 and 17 residual blocks; Mobilenet-B, that has the same architecture -A but reduced input size (96×96) and 25% less feature maps for each layer; MobileNet-C, with an even smaller input size (64×64) and number of feature maps (50% less) but only 8 residual blocks instead of 17.

ShuffleNet. It is a very efficient architecture as well, still optimized for embedded and mobile devices (Ma et al. 2018). ShuffleNet is based on pointwise group convolutions and bottleneck-like structures, combined with a channel split. There are various versions of this network designed to achieve a different trade-off between processing speed and representation capability. In this work we use the improved version of ShuffleNet (V2) with an input size of 224×224 and a scale factor of 1.0.

SqueezeNet. Proposed by Iandola et al. (2016), it is a network widely used when there are strict requirements in terms of RAM and storage space. In fact, it is a compressed version of AlexNet designed for reducing the model size through pruning and quantization, in order to make it suitable for embedded and mobile devices.

Xception. It is a simplified version of the Inception network (Chollet 2017), optimized with design choices inspired by MobileNet and ResNet. In fact, Xception is composed by depthwise separable convolutional layers structured into modules, all of which have linear residual connections around them, except for the first and the last modules. In our implementation, we use Xception with an input size of 71×71 , since we want to evaluate the performance of its lightweight version.

2.2 Training strategy

As recently done in the literature by Antipov et al. (2017), we perform a cross-dataset evaluation, i.e. we train the CNNs by using the images of a dataset and test the learned model on different test sets without fine tuning. Such experimental protocol allows to evaluate the generalization capability of the CNNs on different datasets and it is more realistic for applications that have to run in real environments.

We use the largest gender recognition dataset in the world, namely VGG-Face2 (Cao et al. 2018), for training our CNNs. It includes 9131 subjects and 3.31 millions of images, providing variations in terms of pose, age, ethnicity and brightness. It is annotated with identity and gender, so it is suitable for our purposes.

For all the models, we used the SGD optimizer with the following parameters: momentum=0.9, batch-size=128, learning rate=0.005 with a decay of a factor 5 every 20 epochs. As loss function we adopted the cross-entropy with a weight decay of 0.005 for regularization. The networks are trained for 70 epochs; after each training epoch, we use the VGG-Face2 validation set to evaluate the performance and we choose the checkpoint with the highest validation accuracy.

As for the input of the network, we crop the faces by using a face detector based on SSD (Liu et al. 2016) and ResNet-10 as backbone network. Then, as suggested by the authors of the VGG-Face2 dataset, we subtract the average face image computed on the training set to obtain zero-centered data over all the channels. Finally, we rescale the images to adapt them to the native input size of each network.

To increase the generalization capability of the CNNs, we apply various types of data augmentation strategies that are commonly adopted in the state of the art. Those strategies include 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 ranges used in our procedure are chosen empirically to reproduce variations that are likely to occur in real scenarios.

All the models were fine-tuned from pretrained ImageNet weights, except for MobileNet v2-C, that was never trained on the ImageNet dataset, and VGG, for which we have adopted the VGG-Face weights.

2.3 Corrupted test sets

We adopt two of the most popular gender recognition datasets for producing corrupted images, namely LFW+ (Han et al. 2017) and GENDER-FERET (Azzopardi et al. 2016a).

LFW+ is composed by 15,699 face images belonging to 8000 different unique subjects. Since there is not a standard division between training and test set, we adopted the whole dataset for our experiments. GENDER-FERET consists of 946 grayscale face images, already divided in training (474 images) and test set (472 images), balanced to have the same number of male and female subjects.

Starting from LFW+ and the test set of GENDER-FERET, we produced two new datasets, that we called LFW+C and GENDER-FERET-C. They have been generated from the original images by applying 13 different types of corruptions, namely the ones described by Hendrycks and Dietterich (2019) that can occur on faces acquired in real environments. Each type of corruption $C \in \{C1, \dots, C13\}$ is applied to the original images with five levels of severity $s \in \{1, 2, 3, 4, 5\}$; increasing the severity value, the effect

of the corruption is stronger over the original image. Therefore, LFW+C consists of $15,699 \times 13 \times 5 = 1,020,435$ images, while GENDER-FERET-C is composed by $472 \times 13 \times 5 = 6,136$ samples.

We partitioned the corruptions into three main categories, namely *blur*, *noise* and *digital*. More details about their implementation and parameters are reported in Table 3, while an overview of the effect of the different corruptions is shown in Fig. 1.

Blur corruptions. Surveillance cameras installed in real environments, differently from the setup reproduced in controlled laboratory conditions, frame people moving spontaneously, namely walking, talking to an interlocutor or on the phone, looking around, and so on. The sudden movements of the face can cause image blur, which can be very strong if the shutter time of the camera is not set properly. In addition, blur can be also manually added to attenuate the effects of a noisy image acquisition. The blur corruptions can therefore be partitioned in four categories. *Gaussian blur* (C1) is a pre-preprocessing operation often adopted by smart cameras to reduce the noise introduced during the image acquisition. *Defocus blur* (C2) is the effect typically obtained when the acquisition is performed by cameras with limited depth of field (DoF) in scenarios characterized by a large DoF. *Zoom blur* (C3) is caused by a person moving towards the camera. Finally, *motion blur* (C4) happens when a person suddenly moves his face, quickly changing the pose and causing the face blur.

Noise corruptions. The digital cameras typically acquire images 24 h a day and could be installed in places not protected from heat. Moreover, they are often pointed outdoor, so as to frame the people entering inside the shop. Thus, the sensor is subjected to high temperature and exposure, which cause two categories of noise corruptions. *Gaussian noise* (C5) substantially degrades the image quality with random speckles appearing when the temperature of the sensor increases. *Shot noise* (C6) has a similar effect but it occurs in case of high exposure.

Digital corruptions. The image acquired by the camera is also subjected to corrections aimed at improving its rendering. For example, modern cameras have a dynamic contrast control which allows to improve the contrast when the difference between the brightest and the darkest pixels in the image is high; therefore, we consider *contrast increase* (C7) and *contrast decrease* (C8) as possible digital corruptions. Similarly, we include in this category *brightness increase* (C9) and *brightness decrease* (C10), since the automatic gain control (AGC) available on the cameras dynamically change the image brightness according to the environmental lighting conditions. To simulate partial occlusions of the face, which can be due to smartphone, tissue, scarf, sunglasses, parts of the body or other people, we include *spatter* (C11) among the digital corruptions. The occlusions are

Table 3 Description of the considered corruptions and parameter values adopted to implement the various levels of severity (s)

Corruption	Parameter	Values					Description
		$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$	
Gaussian blur	σ	1	1.8	2.6	3.4	4	$I_c(x, y, d) = I(x, y, d) * G_\sigma(x, y), \forall d \in \{R, G, B\} G_\sigma(x, y) = e^{-(x^2+y^2)/2\sigma^2}$
Defocus blur	r	1.5	2	2	2.5	3	$I_c(x, y, d) = I(x, y, d) * (D_r(x, y) * G_\sigma(x, y))$, being D_r a disc shaped kernel ^{la} with radius r .
	σ	0.1	0.2	0.3	0.4	0.4	
Zoom blur	z	1.11	1.18	1.26	1.32	1.4	$I_c(x, y, d) = \frac{1}{ T } \sum_t \text{zoom}_t(I(x, y, d)), \quad t \in T = \{t = 1 + ne, t \leq z, \forall n \in \mathbb{N}\}$ $\text{zoom}_t(I)$ enlarges the image I by a factor t using linear interpolation ^b .
	ϵ	0.01	0.01	0.02	0.02	0.03	
Motion blur	r	3.3	5	5	5	6.7	Implementation from the ImageMagick library ^c , with random angle ^a .
	σ	1	1.7	2.7	4	5	
Gaussian noise	σ	0.08	0.12	0.18	0.24	0.3	$I_c(x, y, d) = I(x, y, d) + \mathcal{N}(0, \sigma^2)$
Shot noise	q	60	29	15	8	5	$I_c(x, y, d) = \text{Poisson}(I(x, y, d) * q)/q$
Contrast inc.	q	1.5	1.9	2.6	3.3	5	$I_c(x, y, d) = (I(x, y, d) - \mu_d) * q + \mu_d$,
Contrast dec.	q	0.4	0.33	0.24	0.16	0.1	where μ_d is the average value of the original image I over the channel d .
Brightness inc.	q	0.1	0.2	0.3	0.4	0.5	$I_c(x, y, v) = I(x, y, v) + q, \quad \forall x, y,$
Brightness dec.	q	-0.1	-0.2	-0.3	-0.4	-0.5	where v is the v channel in hsv image representation.
Spatter	c_0	0.65	0.65	0.65	0.65	0.67	Implementation by Hendrycks and Dietterich (2019).
	c_1	0.3	0.3	0.3	0.3	0.4	
	c_2	4	3	2	1	1	
	c_3	0.69	0.68	0.68	0.65	0.65	
	c_4	0	0	0	1	1	
JPEG compr.	quality	25	18	15	10	7	Implementation from the Pillow library ^d .
Pixelation	q	0.6	0.5	0.41	0.3	0.25	$I_c(x, y, d) = I(\lfloor xf \rfloor / q, \lfloor yf \rfloor / q, d)$,

$I(x, y, d)$ refers to the original image and $I_c(x, y, d)$ to the corrupted image, while $d \in \{R, G, B\}$ represents the channel in the RGB color space

^aThe r and σ parameter value are expressed in pixels and referred to a 48×48 image. The value of the parameter is scaled proportionally with larger images.

^bImplementation from the SciPy library: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.ndimage.zoom.html>

^c<https://imagemagick.org/api/effect.php#MotionBlurImage>

^d<https://pillow.readthedocs.io/en/3.1.x/handbook/image-file-formats.html#jpeg>

Fig. 1 Examples of corrupted images available in LFW+C. Each image is obtained by applying the considered corruptions (one for each column) with increasing value of severity s (one for each row from 1 to 5)

rendered by adding bright random patterns on the image for low corruption severity and dark elements for higher corruption severity, as proposed by Hendrycks and Dietterich (2019). When the processing is performed on external servers, the acquired image is compressed to reduce the bandwidth consumption and sent through the network. Thus, *JPEG compression* (C12) can cause artifacts simulated by gradually decreasing the quality of the resulting image, which is a configurable parameter, inversely proportional to the severity. Finally, we consider image *pixelation* (C13), namely the effect arising when a low resolution image is upscaled, as the last digital corruption. It typically occurs when the size in pixels of the face is smaller than the input size of the convolutional neural network and the upscaling causes a pixelation effect.

2.4 Real corruptions

The experiments on LFW+C and GENDER-FERET-C allow to determine the negative impact of image corruptions on existing datasets; the considered corruption patterns are manually added one at a time so that we can evaluate the effect of each of the corruption categories separately. However, it is also important to analyze the performance of the considered convolutional neural networks even in real extreme conditions, in which the corruptions may also be mixed.

To the best of our knowledge, the only publicly available gender recognition dataset acquired in real environments is the MIVIA-Gender. It has been collected in two different scenarios, inside a university and at the entrance of a supermarket. The image acquisition has been performed in different illumination conditions, in order to produce strong noise and digital corruptions. The people, unaware of the presence of the surveillance camera, suddenly moved their faces causing severe blur effects.

The whole dataset is composed by 5832 face images of 900 subjects. Following on the publicly available partition adopted in (Carletti et al. 2020a), we use one half of the subjects as test set for evaluating the robustness of the considered convolutional neural to corruptions happening in real scenarios.

2.5 Performance metrics

For all the datasets, we evaluate the performance by computing the gender classification accuracy, namely the ratio between the number of correct classifications and the total number of samples. This protocol is different with respect to the one proposed by Hendrycks and Dietterich (2019), adopting indeed metrics for evaluating the absolute and the relative classification error with respect to a baseline. The reason behind this choice lies in the fact that we find more

intuitive for a binary classification problem the direct evaluation of the accuracy for each network.

3 Results

In this Section we analyze the results achieved by the considered convolutional neural networks, summarized in Table 4. First, we analyze the performance on LFW+ and GENDER-FERET, and on their corrupted versions LFW+C and GENDER-FERET-C, identifying the possible motivations of the drop and the CNNs more resilient to the considered corruptions. Then, we analyze the accuracy achieved on MIVIA-Gender, which includes images corrupted due to the real conditions of the environment, and show that the results are coherent with the findings over LFW+C and GENDER-FERET-C.

3.1 Performance on LFW+ and LFW+C

The accuracy of the considered methods over LFW+ is comparable with the best in the state of the art (0.993). As shown in Table 4, most of them, except MobileNet-C, Xception and SqueezeNet, achieve an accuracy higher than 0.98; this result points out the generalization capability obtained with the proposed training procedure and is not negligible, since we do not fine tune the CNNs over the images of this dataset. The best accuracy of 0.989 is achieved by VGG and SENet.

As expected, we can note that all the methods suffer a substantial influence from the application of corruptions. SENet seems to be the most resilient, having an average accuracy of 0.968 on the corrupted samples, which also correspond to the lowest drop (2.12%) compared with the other architectures. VGG achieves a similar accuracy (0.963), with a slightly higher drop with respect to LFW+ (2.63%). The

Table 4 Comparison of the accuracy (%) achieved by the experimented CNNs on the considered datasets

Method	LFW+	LFW+C	GF	GF-C	MG
VGG	0.989	0.963	0.936	0.887	0.954
SENet	0.989	0.968	0.949	0.865	0.918
MobileNet-A	0.984	0.958	0.943	0.858	0.904
DenseNet	0.985	0.955	0.951	0.867	0.898
ShuffleNet	0.983	0.946	0.943	0.834	0.893
MobileNet-B	0.983	0.941	0.936	0.831	0.891
MobileNet-C	0.966	0.929	0.818	0.760	0.848
Xception	0.957	0.923	0.811	0.739	0.820
SqueezeNet	0.948	0.906	0.773	0.697	0.746

For the sake of readability, GENDER-FERET is abbreviated with GF, as well as MIVIA-Gender with MG. The best results for each dataset are highlighted in bold. The methods are ordered for descending accuracy over the MIVIA-Gender dataset

accuracy on LFW+C is still higher than 0.95 for MobileNet-A (0.958, 2.64% less) and DenseNet (0.955, 3.05% less).

While the performance on the original LFW+ dataset has a slight variability, on the LFW+C we note a higher variance. In particular, we notice that smaller networks such as ShuffleNet (0.946), MobileNet-B (0.941), MobileNet-C (0.929) and Xception (0.923) have more substantial performance drops of 3.76%, 4.27%, 3.83% and 3.55%. SqueezeNet has the highest decrease (4.43%), with an accuracy on LFW+C of 0.914, the lowest one.

By looking at Fig. 2, we can note that most of the networks, namely MobileNet-A, ShuffleNet, MobileNet-B, MobileNet-C and Xception, suffer a performance drop mainly caused by noise corruptions. SENet and VGG achieve a balanced accuracy over all the corruptions, while DenseNet and SqueezeNet are less effective on samples affected by blur corruptions.

Table 5 reports more detailed results, which allow to refine the analysis of the impact of the various corruptions. In particular, we notice that SENet and VGG are substantially more effective than the other methods when dealing with noise corruptions (average accuracy 0.969); the type of noise, namely gaussian and shot, seems to be not crucial, but the negative impact is relevant on the accuracy achieved by the other networks. As for the blur corruptions, SENet obtains the best accuracy in average (0.964) and on zoom (0.980) and motion (0.964) categories; VGG retains the highest performance on samples affected by gaussian blur (0.987), while ShuffleNet is the most robust to defocus blur (0.933). The latter has a very negative impact on the performance of all the networks; this corruption, as evident from Fig. 1, causes a strong loss of details in the face image, making very difficult the gender classification. SENet achieves the best average accuracy even over digital corruptions (0.969); in this case VGG obtains the highest accuracy on most of the categories (contrast increase and decrease and brightness increase and decrease), but SENet overcomes VGG for the accuracy on spatter (0.897 vs 0.851). Indeed, this digital corruption causes a strong performance drop to all the methods: the highest accuracy is achieved by MobileNet-A (0.913), but the average is lower than 0.90. Apparently, the presence of occlusions can substantially complicate the recognition of the gender.

3.2 Performance on GENDER-FERET and GENDER-FERET-C

From Table 4 we can note that DenseNet (0.951) and SENet (0.949) achieve the best accuracy on GENDER-FERET, while MobileNet-A (0.943), ShuffleNet (0.943), VGG (0.936) and MobileNet-B (0.936) are very close to the top rank. The remaining methods, coherently with the previous analysis over LFW+, achieve a very low accuracy (under

Table 5 Accuracy of the considered CNNs on the corruption categories in LFW+C

Methodm	LFW+C		Blur		Noise			Digital			Pixelation	Average					
	Gaussian	Defocus	Zoom	Motion	Average	Gaussian	Shot	Average	Cont.Inc.	Cont.Dec.			Brig.Inc.	Brig.Dec.	Spatter	JPEG Comp.	
SENet	0.968	0.986	0.926	0.980	0.964	0.964	0.970	0.969	0.969	0.980	0.978	0.980	0.979	0.897	0.985	0.989	0.969
VGG	0.963	0.987	0.930	0.967	0.928	0.953	0.969	0.969	0.969	0.984	0.989	0.982	0.986	0.851	0.985	0.989	0.966
MobileNet-A	0.958	0.980	0.926	0.974	0.953	0.958	0.932	0.928	0.930	0.965	0.973	0.976	0.971	0.913	0.974	0.983	0.965
DenseNet	0.955	0.980	0.910	0.971	0.918	0.945	0.966	0.964	0.965	0.970	0.962	0.978	0.974	0.857	0.978	0.984	0.958
ShuffleNet	0.946	0.980	0.933	0.976	0.961	0.962	0.927	0.922	0.924	0.970	0.954	0.976	0.966	0.777	0.971	0.982	0.942
MobileNet-B	0.941	0.978	0.930	0.966	0.954	0.957	0.877	0.859	0.868	0.956	0.963	0.972	0.962	0.875	0.967	0.981	0.954
MobileNet-C	0.929	0.963	0.894	0.960	0.939	0.939	0.879	0.873	0.876	0.941	0.922	0.952	0.944	0.891	0.950	0.965	0.938
Xception	0.923	0.954	0.897	0.947	0.932	0.932	0.902	0.896	0.899	0.952	0.869	0.947	0.936	0.872	0.943	0.954	0.925
SqueezeNet	0.906	0.940	0.793	0.939	0.893	0.891	0.913	0.911	0.912	0.934	0.902	0.933	0.898	0.837	0.934	0.948	0.912

The best results for each corruption type are highlighted in bold

The accuracy on the whole dataset is reported in the second column, while the remaining columns are divided in three blocks, one for each corruption category (blur, noise, digital). For each block, we report the accuracy achieved by the CNNs on the samples affected by a specific corruption type, and the average accuracy in the last column. The methods are ordered for descending accuracy over LFW+C

Fig. 2 Accuracy of the considered methods on the original LFW+ dataset, over LFW+C and on the different categories of corrupted samples, namely blur, noise and digital. The methods are ordered, from the left to the right, for descending accuracy on LFW+C

0.82) and demonstrate a poor generalization capability also on this dataset.

We notice that on GENDER-FERET-C the performance drop in case of corruptions is even higher than the loss observed for LFW+C; this is probably caused by the additional challenge due to the presence of only grayscale images. VGG is the convolutional neural network more robust to corruptions on this dataset, achieving an accuracy of 0.887 and the lowest drop (5.23%). DenseNet (0.867, 8.83% less) and SENet (0.865, 8.85% less) suffer substantially more the effect of the corruptions, even obtaining the best performance on the original dataset. MobileNet-A (0.858, 9.01% less) still achieves accuracy comparable with CNNs substantially more complex; on the other hand, ShuffleNet (0.834, 11.56% less) and MobileNet-B (0.831, 11.22% less) seem to have a not sufficient representative power to effectively classify the images of GENDER-FERET-C. As previously observed, MobileNet-C (0.760, 7.09% less), Xception (0.739, 9.83% less) and SqueezeNet (0.697, 8.88% less) are very far from the other networks in terms of accuracy.

The chart in Fig. 3 shows that the effect of noise corruptions is disruptive on the GENDER-FERET-C for almost all the considered methods. Only VGG and DenseNet seem to be more resilient to this type of corruption, which is probably more challenging when present on grayscale images. Blur and digital corruptions are substantially less harmful than the noise for this specific dataset.

The detailed results reported in Table 6 further reinforce such experimental evidence. VGG and DenseNet achieve an average accuracy on noise corruptions substantially higher than the other methods (0.870 and 0.865); SENet obtains

Table 6 Accuracy of the considered CNNs on the corruption categories in GENDER-FERET-C (GF-C)

Method	GF-C		Noise			Digital			Spatter	JPEG Comp.	Pixelation	Average					
	Gaussian	Blur	Average	Motion	Shot	Cont. Inc.	Cont. Dec.	Brig. Inc.					Brig. Dec.				
VGG	0.887	0.918	0.865	0.889	0.882	0.888	0.866	0.874	0.870	0.903	0.912	0.933	0.820	0.834	0.902	0.928	0.890
DenseNet	0.867	0.903	0.807	0.879	0.860	0.862	0.859	0.872	0.865	0.922	0.836	0.953	0.836	0.734	0.889	0.924	0.870
SENet	0.865	0.939	0.817	0.860	0.908	0.881	0.778	0.826	0.802	0.899	0.894	0.953	0.819	0.707	0.908	0.931	0.873
MobileNet-A	0.858	0.903	0.859	0.900	0.905	0.892	0.686	0.714	0.700	0.898	0.922	0.958	0.832	0.770	0.884	0.919	0.883
ShuffleNet	0.834	0.882	0.814	0.853	0.859	0.852	0.686	0.682	0.684	0.889	0.860	0.950	0.828	0.757	0.867	0.910	0.866
MobileNet-B	0.831	0.877	0.794	0.927	0.901	0.875	0.700	0.672	0.686	0.864	0.853	0.948	0.764	0.732	0.853	0.913	0.847
MobileNet-C	0.760	0.859	0.842	0.842	0.821	0.841	0.630	0.639	0.635	0.756	0.715	0.822	0.701	0.670	0.761	0.823	0.750
Xception	0.739	0.783	0.680	0.747	0.755	0.741	0.694	0.701	0.697	0.777	0.784	0.799	0.741	0.620	0.752	0.778	0.750
SqueezeNet	0.697	0.814	0.725	0.713	0.781	0.758	0.592	0.590	0.591	0.789	0.578	0.758	0.636	0.589	0.696	0.795	0.692

The best results for each corruption type are highlighted in bold

The average accuracy on the whole dataset is reported in the second column, while the remaining columns are divided in three blocks, one for each corruption category (blur, noise, digital). For each block, we report the accuracy achieved by the CNNs on the samples affected by a specific corruption type, and the average accuracy in the last column. The methods are ordered for descending accuracy over GF-C

Fig. 3 Accuracy of the considered methods on the original GENDER-FERET (GF) dataset, over GENDER-FERET-C (GF-C) and on the different categories of corrupted samples, namely blur, noise and digital. The methods are ordered, from the left to the right, for descending accuracy on GF-C

0.802, while the remaining CNNs do not overcome 0.700. As observed for LFW+C, there are no significant differences between the impact of gaussian and shot noise. MobileNet-A achieves the best average accuracy over blur corruptions (0.892), followed by VGG (0.888), that is the most robust to defocus blur (0.865) and SENet (0.881), which obtains the highest accuracy on samples affected by gaussian and motion blur (0.939 and 0.908). As already noticed for LFW+C, the defocus blur seems to be the most challenging type for this corruption category, while spatter is the counterpart among the digital corruptions. VGG achieves the highest average accuracy over digital corruptions (0.890) and most of the success is due to the best performance on samples affected by spatter (0.834). In fact, the leadership for contrast increase and brightness decrease is got by DenseNet (0.922 and 0.836), for JPEG compression and pixelation by SENet (0.908 and 0.931) and for contrast decrease and brightness increase by MobileNet-A (0.922 and 0.958). In addition, on this dataset we can also note a low average accuracy on images affected by brightness decrease; this is reasonable, since in grayscale images the loss of brightness can substantially degrade the details and make very hard the recognition of facial traits useful to recognize males and females.

3.3 Performance on MIVIA-Gender

The accuracy achieved by the considered methods on the MIVIA-Gender dataset is reported in Table 4. We notice a strong superiority of VGG, which widely overcomes the best current state of the art performance on this dataset (0.915), obtaining an accuracy of 0.954. This result confirms the

robustness demonstrated by this method over LFW+C and GENDER-FERET-C; its higher resilience to noise with respect to other CNNs and the good generalization capability over different types of corruptions allows VGG to achieve remarkable performance on images acquired in real environments.

The second top rank is obtained by SENet (0.918), which probably pays its lower robustness to noise with respect to VGG, especially on grayscale images. MobileNet-A (0.904) and DenseNet (0.898) hold the third and the fourth positions; the better resilience of MobileNet-A to defocus blur and spatter slightly overcomes the higher robustness to noise demonstrated by DenseNet. ShuffleNet (0.893) and MobileNet-B (0.891) achieve performance very close to the ones obtained by more complex methods; they are interesting alternative candidates when the available computing resources are limited.

Finally, MobileNet-C (0.848), Xception (0.820) and SqueezeNet (0.746) confirm their inadequacy for effectively performing gender recognition in real environments.

4 Discussion

The analysis of the results achieved by the considered convolutional neural networks allows to deduce some insights about the general behaviour of the gender recognition methods when dealing with image corruptions.

It is evident that the noise has the highest negative effect on most of the networks, while blur and digital corruptions are less disrupting. It may be counter intuitive if we look at the dataset sample shown in Fig. 1; in fact, we notice that blur corruptions subtract a great amount of information from the image with respect to noise corruptions, being the facial traits sufficiently evident to the human eye. However, it is well known that the sensitivity to high frequency noise is still a substantial problem with CNN architectures (Hendrycks and Dietterich 2019).

VGG and DenseNet appear to be the exceptions to this rule: the former probably achieves robustness to noise thanks to its simple structure, which often demonstrated to generalize well in different scenarios; the latter uses average pooling in its transition layers, making it more resilient to high frequency noise than architectures which rely more heavily on max pooling. Both the architectures seem to have more problems with blur. However, VGG obtains an impressive accuracy over the MIVIA-Gender, while DenseNet does not generalize so well on this dataset acquired in real environments.

SENet and MobileNet-A obtained very positive results; the former, in terms of accuracy, is probably the best alternative to VGG; the latter, being a lightweight network optimized for mobile devices, is the best solution when no GPU

is available or a higher frame rate is required. Similar considerations may be done for ShuffleNet and MobileNet-B, but the expected accuracy is slightly lower.

Most of the tested architectures are reasonably resilient to blur, meaning that the training procedure relied on low frequency features more than on high frequency ones. A possible solution for reducing the average classification error and its variance between the different categories of corruptions may be the adoption of more sophisticated augmentation techniques, as the ones proposed by Lim et al. (2019), which demonstrated to provide the CNNs with a strong generalization capability. Since defocus (blur) and spatter (digital) demonstrated to be the most challenging corruption types, a data augmentation strategy which takes into account these image corruptions may be very useful for increasing the performance.

5 Conclusion

In this paper we proposed and applied a novel experimental framework for investigating the performance of CNNs for gender recognition from face images. It allows to evaluate the classification accuracy on images corrupted as happens in real environments. We produced the LFW+C and the GENDER-FERET-C datasets, the corrupted versions of the popular LFW+ and GENDER-FERET, in order to evaluate the robustness of nine different convolutional neural networks to specific image corruptions. Then we analyzed the accuracy achieved by the same networks using a dataset acquired in real environments, namely the MIVIA-Gender. The experimental analysis confirmed the expected performance drop. VGG and SENet result the most robust solutions in absolute, but a lightweight networks such as MobileNet-A is an effective alternative when limited processing hardware is available.

From the analysis of the results we deduced interesting insights, for example the disruptive impact of noise, defocus and spatter. We discussed about the capability of the various networks to generalize with respect to these corruptions. These observations may suggest possible countermeasures for dealing with corruptions happening in operating environments. A possible future direction may be the investigation of more sophisticated data augmentation strategies or specific changes to the network architectures, which may further improve the robustness of the gender recognition solutions to image corruptions.

The experimental framework, publicly available, can be easily extended for including more datasets and more convolutional neural networks, in order to extensively evaluate the performance of gender recognition methods on new datasets.

Funding Open access funding provided by Università degli Studi di Salerno within the CRUI-CARE Agreement.. This research was partially supported by the Italian MIUR within PRIN 2017 grants, Projects Grant 20172BH297 002CUP D44I17000200005 and by A.I. Tech srl (www.aitech.vision).

Availability of data and material The authors do not provide supplementary data and material.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

Code availability The code is publicly available at: <https://github.com/MiviaLab/GenderRecognitionFramework>.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Afifi M, Abdelhamed A (2019) Afif4: deep gender classification based on adaboost-based fusion of isolated facial features and foggy faces. *J Vis Commun Image Represent* 62:77–86
- Albanie S, Nagrani A, Vedaldi A, Zisserman A (2018) Emotion recognition in speech using cross-modal transfer in the wild. *arXiv preprint arXiv:1808.05561*
- Antipov G, Berrani S-A, Dugelay J-L (2016) Minimalistic cnn-based ensemble model for gender prediction from face images. *Pattern Recognit Lett* 70:59–65
- Antipov G, Baccouche M, Berrani S-A, Dugelay J-L (2017) Effective training of convolutional neural networks for face-based gender and age prediction. *Pattern Recognit* 72:15–26
- Azzopardi G, Greco A, Vento M (2016a) Gender recognition from face images using a fusion of svm classifiers. In: *International conference on image analysis and recognition*. Springer, pp 533–538
- Azzopardi G, Greco A, Vento M (2016b) Gender recognition from face images with trainable cosfire filters. In: *IEEE International conference on advanced video and signal based surveillance (AVSS)*, IEEE, pp 235–241
- Azzopardi G, Greco A, Saggese A, 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)*, IEEE, pp 1–6
- Azzopardi G, Foggia P, Greco A, Saggese A, Vento M (2018a) Gender recognition from face images using trainable shape and color features. In: *2018 24th international conference on pattern recognition (ICPR)*, IEEE, pp 1983–1988

- Azzopardi G, Greco A, Saggese A, Vento M (2018b) Fusion of domain-specific and trainable features for gender recognition from face images. *IEEE Access* 6:24171–24183
- Basulaim KO, Dabash AA (2019) Performance enhancement for gender recognition using trainable bank of gabor filters and nca. In: 2019 first international conference of intelligent computing and engineering (ICOICE), IEEE, pp 1–5
- Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A (2018) Vggface2: A dataset for recognising faces across pose and age. In: IEEE conference on face and gesture recognition, pp 67–74
- Carletti V, Greco A, Saggese A, Vento M (2020a) An effective real time gender recognition system for smart cameras. *J Ambient Intell Human Comput* 11(6):2407–2419
- Carletti V, Greco A, Percannella G, Vento M (2020b) Age from faces in the deep learning revolution. *IEEE Transact Pattern Anal Mach Intell* 42(9):2113–2132
- Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258
- Dehghan A, Ortiz EG, Shu G, Masood SZ (2017) Dager: deep age, gender and emotion recognition using convolutional neural network. *arXiv preprint arXiv:1702.04280*,
- Del Coco M, Carcagnì P, Leo M, Mazzeo PL, Spagnolo P, Distanti C (2016) Assessment of deep learning for gender classification on traditional datasets. In: 2016 13th IEEE international conference on advanced video and signal based surveillance (AVSS), IEEE, pp 271–277
- Foggia P, Greco A, Percannella G, Vento M, Vigilante V (2019) A system for gender recognition on mobile robots. In: Proceedings of the 2nd international conference on applications of intelligent systems, pp 1–6
- Fu S, He H, Hou Z-G (2014) Learning race from face: a survey. *IEEE Transact Pattern Anal Mach Intell* 36(12):2483–2509
- Greco A, Saggese A, Vento M, Vigilante V (2020) A convolutional neural network for gender recognition optimizing the accuracy/speed tradeoff. *IEEE Access* 8:130771–130781
- Guo G, Zhang N (2019) A survey on deep learning based face recognition. *Comput Vis Image Underst* 189:102805
- Gurnani A, Shah K, Gajjar V, Mavani V, Khandhediya Y (2019) Saffage: salient approach for facial soft-biometric classification-age, gender, and facial expression. In: 2019 IEEE winter conference on applications of computer vision (WACV), IEEE, pp 839–847
- Haider KZ, Malik KR, Khalid S, Nawaz T, Jabbar S (2019) Deep-gender: real-time gender classification using deep learning for smartphones. *J Real-Time Image Process* 16(1):15–29
- Han H, Jain AK, Wang F, Shan S, Chen X (2017) Heterogeneous face attribute estimation: a deep multi-task learning approach. *IEEE Transact Pattern Anal Mach Intell* 40(11):2597–2609
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
- Hendrycks D, Dietterich T (2019) Benchmarking neural network robustness to common corruptions and perturbations. In: Proceedings of the international conference on learning representations
- Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
- Huang GB, Jain V, Learned-Miller E (2007) Unsupervised joint alignment of complex images. In: 2007 IEEE 11th international conference on computer vision, IEEE, pp 1–8
- Huang GB, Mattar M, Berg T, Learned-Miller E (2008) Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In: Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition
- Huang G, Liu Z, Der Maaten LVan, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
- Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K (2016) Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size. *arXiv preprint arXiv:1602.07360*
- Jia S, Cristianini N (2015) Learning to classify gender from four million images. *Pattern Recognit Lett* 58:35–41
- Levi G, Hassner T (2015) Age and gender classification using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 34–42
- Li S, Deng W (2020) Deep facial expression recognition: a survey. In: *IEEE transactions on affective computing*
- Lim S, Kim I, Kim C, Kim S (2019) Fast autoaugment. *arXiv :abs/1905.00397*
- Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: single shot multibox detector. In: Proceedings of the European conference on computer vision (ECCV), Springer, pp 21–37
- Ma N, Zhang X, Zheng H-T, Sun J (2018) Shufflenet v2: practical guidelines for efficient cnn architecture design. In: Proceedings of the European conference on computer vision (ECCV), pp 116–131
- Ng C-B, Tay Y-H, Goi B-M (2015) A review of facial gender recognition. *Pattern Anal Appl* 18(4):739–755
- Parkhi OM, Vedaldi A, Zisserman A, et al. (2015) Deep face recognition. In *BMVC*, vol 1, p 6
- Ranjan R, Patel VM, Chellappa R (2017) Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transact Pattern Anal Mach Intell* 41(1):121–135
- Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) Mobilenetv2: inverted residuals and linear bottlenecks. *arXiv preprint arXiv:1801.04381*
- Simanjuntak F, Azzopardi G (2019) Fusion of cnn-and cosfire-based features with application to gender recognition from face images. In: Science and information conference, Springer, pp 444–458
- Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*
- Van de Wolfshaar J, Karaaba MF, Wiering MA (2015) Deep convolutional neural networks and support vector machines for gender recognition. In: 2015 IEEE symposium series on computational intelligence, IEEE, pp 188–195
- Yi D, Lei Z, Liao S, Li SZ (2014) Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*
- Zhang W, Smith ML, Smith LN, Farooq A (2016) Gender and gaze gesture recognition for human-computer interaction. *Comput Vis Image Underst* 149:32–50

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.