# Selfie Periocular Verification using an Efficient Super Resolution Approach

Juan Tapia[a,], Rodrigo Lara[b,], Andres Valenzuela[b,], Marta Gomez-Barrero[c,], Christoph Busch[a,]

[a] da/sec-Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany
[b] TOC Biometrics, Santiago, Chile.
[c] Hochschule Ansbach, Germany.

***The following paper is a pre-print- The publicaction is currently under review for Pattren Recognition Journal - SI: Biometrics COVID-19.***



**Abstract**

Selfie-based biometrics has great potential for a wide range of applications from marketing to higher security environments like online banking. This is now especially relevant since e.g. periocular verification is contactless, and thereby safe to use in pandemics such as COVID-19. However, selfie-based biometrics faces some challenges since there is limited control over the data acquisition conditions. Therefore, super-resolution have to be used to increase the quality of the captured images. Most of the state of the art super-resolution methods use deep networks with large filters, thereby needing to train and store a correspondingly large number of parameters, and making their use difficult for mobile devices commonly used for selfie-based.

In order to achieve an efficient super-resolution method, we propose an Efficient Single Image Super-Resolution (ESISR) algorithm, which takes into account a trade-off between the efficiency of the deep neural network and the size of its filters. To that end, the method implements a novel loss function based on the Sharpness metric. This metric turns out to be more suitable for increasing the quality of the eye images. Our method drastically reduces the number of parameters when compared with Deep CNNs with Skip Connection and Network (DCSCN): from 2,170,142 to 28,654 parameters when the image size is increased by a factor of x3. Furthermore, the proposed method keeps the sharp quality of the images, which is highly relevant for biometric recognition purposes. The results on remote verification systems with raw images reached an Equal Error Rate (EER) 8.7% for FaceNet and 10.05% for VGGFace. Where embedding vectors were used from periocular images the best results reached an EER of 8.9% (x3) for FaceNet and 9.90% (x4) for VGGFace.

*Keywords:* Super-Resolution, Periocular Verification, Selfie Biometric.

## 1. Introduction

Smartphones, and mobile devices in general, play nowadays a central role in our society. We use them a daily basis not only for communication purposes, but also to access social media and for sensitive tasks such as online banking. In order to increase the security level of those more sensitive applications, verifying the subject's identity represents a key. To tackle it, many companies are currently working towards creating applications to verify the subject's identity by comparing a face image stored in the embedded chip of an ID-Card/Passport and a selfie image using Near Field Communication (NFC) from smartphones [1]. This represents a user-friendly identity verification process, which can be easily embedded into numerous processes. However, this verification form also faces some challenges: that selfie image is captured in an uncontrolled scenario, where occlusions due to wearing a scarf in winter or a hygienic facial mask in a pandemic such as COVID-19 may hinder the performance of general face recognition algorithms. Therefore, there is a reinforced need to explore alternatives which can deal with such occluded images successfully, such as utilising the periocular region for recognition purposes.

The aforementioned reasons have increased the interest on periocular based biometrics in the last decade in different scenarios [2, 3]. In particular, it has been shown that periocular images captured with mobile devices for recognition purposes are mainly coming from selfie face images. And the number of digital photos will increase every year: it is anticipated that in 2020 1.4 trillion images will be taken, and 90% of them will come from smartphones[1]. In order to recognise individuals from a selfie, the periocular region needs to be cropped, and the resulting periocular sample usually has often a very low-resolution [4]. Moreover, the subjects capture selfie images in multiple places and backgrounds, using selfie sticks, alone, or with others. This translates into a high variability within the images, in terms of size, lighting conditions, and face pose.

With the aim of improving the quality of such low-resolution images, several Single Image Super-Resolution (SISR) methods have been recently proposed [5, 6], mainly based on convolutional neural networks. Even though some authors have enhanced such networks to do more efficient the reconstruction results of the super-resolution [7], most approaches still use deep

*Corresponding author. E-mail address: juan.tapia-farias@h-da.de
*Email addresses:* rodrigo.lara@toc.cl (Rodrigo Lara), andres.valenzuela@toc.cl (Andres Valenzuela), marta.gomez-barrero@hs-ansbach.de (Marta Gomez-Barrero), christoph.busch@h-da.de (Christoph Busch)

[1] https://focus.mylio.com/tech-today/how-many-photos-will-be-taken-in-2020

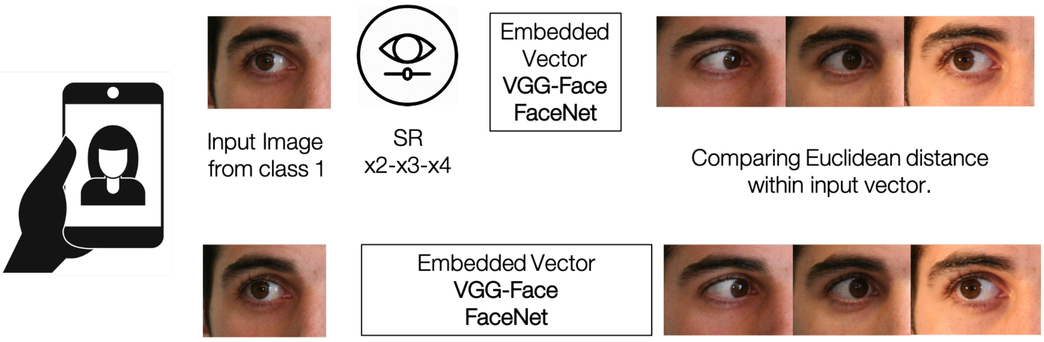



Figure 1: Block diagram of the verification framework. Top: Verification system proposed, including a super-resolution approach. Bottom: Traditional periocular verification systems.

models, which demand larger resources and are thus not suitable for mobile or IoT devices. Furthermore, the loss function used in most techniques is based on structural similarity (SSIM) and Peak Signal to Noise Ratio (PSNR) metrics. Even though those metrics are appropriate for increasing the resolution of general purpose images (e.g., landscapes, cities, or birds) they are not that suitable for increasing the quality of iris based biometrics applications. In contrast, the ISO/IEC 29794 standard on biometric sample quality — Part 6: Iris image data describes the sharpness as one relevant quality.

In this work, we propose a method to verify the identity from a smartphone selfie periocular image in the visible spectrum (VIS) and an efficient super-resolution approach (see Fig. 1). As already mentioned, this is a challenging task since there is limited control of the quality of the images taken: selfies can be captured from different distances, light conditions, and resolutions. Therefore, to tackle these issues, our work proposes a single image super-resolution algorithm method with a novel loss function based on the sharpness LoG metric and a lightweight CNN. This model takes into account the trade-off between the number of layers and filter sizes in order to achieve a light model suitable for mobile devices applications. Additionally, we explore pixel-shuffle and transposed convolutions in order to recover the fine details of the periocular eye images. To validate the proposed approach, we trained our best super-resolution method on a dataset and then tested it on a totally different database. Our method drastically reduces the number of parameters when compared with Deep CNNs with Skip Connection and Network (DCSCN): from 2,170,142 to 28,654 parameters when the image size is increased by a factor of x2.

This paper is an extension of our previous work [8]. In that work, we focused on achieving an accurate ESISR algorithm for periocular eye images taken from selfie images, reporting results in terms of image similarity for the recovered images on a smaller Samsung dataset. In this paper, we evaluate in more detail this new ESISR architecture and also benchmark it with two new state of the art methods: WDSR-A and SRGAN. A full explanation of the reasons that lead us to such architecture is discussed in this work since the reduction of layers in the architecture is not trivial. As an additional contribution, this paper includes the evaluation and performance of our proposed methods on periocular verification systems using two pre-trained CNNs: FaceNet and VGGFace. The larger MOBIO database was used to evaluated the SR methods and periocular recognition systems. Detection Error Trade-off (DET) curves are included to show the performance and the efficiency of our proposal. All these new experiments are benchmarked with those previously obtained in [9, 10, 11].

Therefore, the main contributions from this article can be summarised as follows:

- An efficient architecture using only 7 layers with a feature extractor and one block based on recursive learning of reconstruction is proposed to reduce the number of parameters in comparison with the state-of-the-art WDSR-A, SRGAN and DCSCN algorithms (Section 2).
- A recursive pixel-shuffle technique is introduced over a transposed convolution in order to extract and keep high details of periocular images.
- A novel database for selfie periocular eye images is prepared and will be available for researchers upon request.
- A novel loss function that includes a sharpness metric along with the SR loss function was proposed. This metric is a quality metric for iris.
- A periocular verification system based on embedded vector from two pre-trained models FaceNet and VGGFace with a SR-based pre-processing of the samples (x2, x3 and x4) was tested.

The rest of the article is organised as follows. Sect.2 summarises the related works on periocular recognition and super resolution. Recognition and super-resolution method is described in Sect. 3. The experimental framework is then presented in Sect. 4 and the results are discussed in Sect. 5. We conclude the article in Sect. 6.

## 2. Related work

### 2.1. Super-Resolution (SR)

Super-resolution (SR) is the process of recovering a high-resolution (HR) image from a low-resolution (LR) one [12, 5].

In contrast, supervised machine learning approaches learn mapping functions from LR images to HR images from a large number of examples. The mapping function learned by these models is the inverse of a downgrade function that transforms HR images into LR images. Such downgrade functions can be known or unknown.

Many state-of-the-art SR models learn most of the mapping function in LR space followed by one or more upsampling layers at the end of the network. This is called post-upsampling. Earlier approaches first upsampled the LR image with a predefined up-sampling operation and then learned the mapping in the HR space (pre-upsampling SR). A disadvantage of this approach is that more parameters per layer are required, which in turn leads to higher computational costs and limits the construction of deeper neural networks [5]. SR requires that most of the information contained in an LR image must be preserved in the SR image. SR models therefore mainly learn the residuals between LR and HR images. Residual network designs are therefore of high importance: identity information is conveyed via skip connections whereas reconstruction of high frequency content is done on the main path of the network [5].

Dong *et al.* [12] proposed several SISR algorithms which can be categorized into four types: prediction models, edge-based methods, image statistical methods, and patch-based (or example-based) methods. This method uses 2 to 4 convolutional layers to prove that the learned model performs well on SISR tasks. The authors concluded that *using a larger filter size is better than using deeper Convolutional Neural Networks (CNNs)*.

Kim *et al.* [13] proposed an image SR method using a Deeply-Recursive Convolutional Network (DRCN), which contains deep CNNs with up to 20 layers. Consequently, the model has a huge number of parameters. However, the CNNs share each other's weights to reduce the number of parameters to be trained, thereby being able to succeed in training the deep CNN network and achieving a significant performance. The authors conclude in their work *that deeper networks are better than large filters*.

Yamanaka *et al.* [9] proposed a Deep CNN with a Residual Net, Skip Connection and Network (DCSCN) model achieving a state of the art reconstruction performance while reducing by at least 10 times the computational cost. According to the existing literature, deep CNNs with residual blocks and skip connections are suitable to capture fine details in the reconstruction process. In the same context, [14] and [15] propose the pixel-shuffle and transposed convolution algorithm in order to extract the most relevant features from the images. The transposed convolutional layer can learn up-sampling kernels. However, the process is similar to the usual convolutional layer and the reconstruction ability is limited. To obtain a better reconstruction performance, the transposed convolutional layers need to be stacked, which means the whole process needs high computational resources [9]. Conversely, pixel-shuffle extracts features from the low-resolution images. The authors [9] argue that batch normalization loses scale information of images and reduces the range flexibility of activations. Removal of batch normalization layers not only increases SR performance but also reduces GPU memory 40%. This way, significantly larger models can be trained.

Ledig *et al.* [11] proposed a deep residual network which is able to recover photo-realistic textures from heavily down-sampled images on public benchmarks. An extensive Mean-Opinion-Score (MOS) test shows significant gains in perceptual quality using SR based on Generative Adversarial Network (SRGAN). In addition, the authors present a new perceptual loss based on content loss and adversarial loss.

Yu *et al.* [10] proposed the key idea of wide activation to explore efficient ways to expand features before ReLU, since simply adding more parameters is inefficient for real-time image SR scenarios. The authors present two new networks named Wide Activation for Efficient and Accurate Image Super-Resolution (WSDR). These networks (WDSR-A and WDSR-B) yielded better results on the large-scale DIV2K image super resolution benchmark in terms of PSNR with the same or lower computational complexity. Similar results but with a larger number of parameters are presented by Lim *et al.* [16] in a model called Enhanced Deep Residual Networks for Single Image Super Resolution (EDSR).

Specifically for biometric applications, some papers have explored the use of SR in iris recognition in the visible and near-infrared spectrum. Ribeiro *et al.* [17] proposed a SISR method using CNNs for iris recognition. In particular, the authors test different state of the art CNN architectures and use different training databases in both the near-infrared and visible spectra. Their results are validated on a database of 1,872 near-infrared iris images and on a smartphone image database. The experiments show that using deeper architectures trained with texture databases that provide a balance between edge preservation and the smoothness of the method can lead to good results in the iris recognition process. Furthermore, the authors used PSNR and SSIM to measure the quality of the reconstruction. More recently, Alonso-Fernandez *et al.* [18] presented a comprehensive survey of iris SR approaches. They also described an Eigen-patches reconstruction method based on the principal component analysis ans Eigen-transformation of local image patches. The inherent structure of the iris is reproduced by building a patch-position-dependent dictionary. The authors also used PSNR and SSIM to measure the quality of the reconstruction in the NIR spectrum and in the VSIRIS database in the visible spectrum [19].

#### *2.1.1. Metrics*

Deep learning-based methods for SISR significantly outperform conventional approaches in terms of Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity(SSIM). SRCNN was the first work utilising an end-to-end CNN as a mapping function from low-resolution images to their high-resolution counterparts. Since then, various CNN architectures were proposed in order to improve both the accuracy and the efficiency. In this section, we review these two metrics.

SSIM is a subjective metric used for measuring the structural similarity between images from the perspective of the human visual system. It is based on three relatively independent properties, namely: luminance, contrast, and structure. Abstractly,

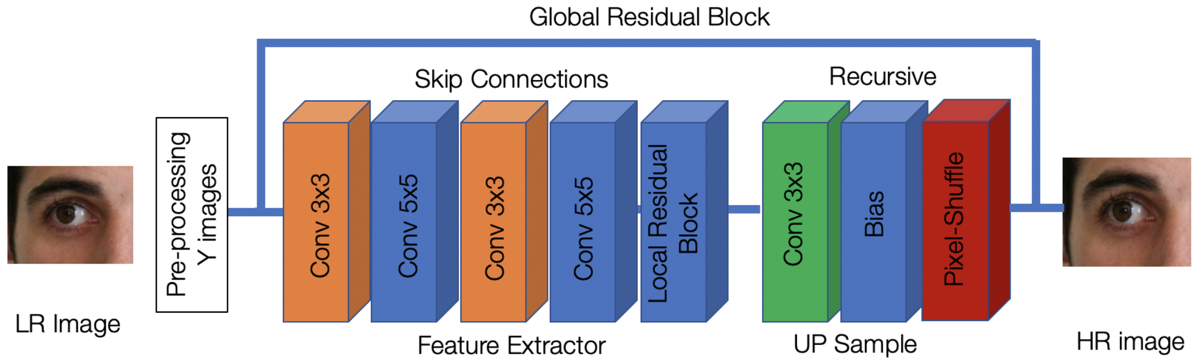



Figure 2: Proposed Super-Resolution method ESISR.

the SSIM formula can be seen as a weighted product of the comparison of luminance, contrast, and structure computed independently. Therefore, SSIM can be defined as:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (1)$$

where $\mu$ and $\sigma$ represent the average and variance of x and y, respectively; and $C_1$ and $C_2$ are two variables to stabilise the division with a weak denominator.

PSNR is a common objective metric to measure the reconstruction quality of a lossy transformation. It is inversely proportional to the logarithm of the Mean Squared Error (MSE) between the ground truth image and the generated image:

$$\mathrm{PSNR} = 10 \log_{10}\left(\frac{\max^2}{\mathrm{MSE}}\right) \quad (2)$$

where max denotes the maximum pixel value, and MSE the mean of the square of differences between the pixel values of the erroneous output (due to soft errors and frame drops) and the correctly reconstructed output (without errors). Therefore, this metric measures pixel differences and not the quality of the images.

### 2.2. Periocular recognition

Periocular recognition has been explored from traditional feature extraction such as intensity, shape, texture, fusion, and off-the-shelf CNN features with pre-trained models. However, only a few papers that previously (according to our state of the art) explore the use of the SR method to improve the quality of the RGB images coming from periocular selfies captures. These remote captures present uncontrolled real scenarios.

Kumari [6] *et al.*˜ provided a survey of periocular biometrics and deep insight of various aspects such as the periocular region utility as a stand-alone modality, its fusion with iris, application in the smartphone authentication, and the role of the in soft biometric classification. No SR methods are reported.

Ahuja [20] *et al.*˜ proposed a hybrid convolution-based model, for verifying pairs of periocular RGB images. They compose the hybrid model as a combination of an unsupervised and a supervised convolution neural network, and augment the combination SIFT model.

Chandrashekhar [21] *et al.*˜ propose a new initialization strategy for the definition of the periocular region-of-interest and the performance degradation factor for periocular biometric and the influence of HOG, LBP, SIFT, Fusion at the Score Level, Effect of Reference Points of the eyes, Covariates, Occlusion Performance and Pigmentation Level Performance.

Kiran [22] *et al.*˜ explore multi-modal biometrics as a means for secure authentication. The proposed system employs face, periocular, and iris images, all captured with embedded smartphone cameras. As the face image is captured from a close distance, one can always obtain periocular and iris information with significant details. It also explores various score level fusion schemes of complementary information from all three modalities.

Diaz [23] *et al.*p̃roposed a method to apply existing pre-trained architectures, proposed in the context of the ImageNet Large Scale Visual Recognition Challenge, to the task of periocular recognition. These networks have proven to be very successful for many other computer vision tasks apart from the detection and classification tasks for which they were designed. They demonstrate that these off-the-shelf CNN features can effectively recognize individuals based on periocular images, despite being trained to classify generic objects.

## 3. Proposed method

As mentioned in Sect. 1 and depicted in Fig. 1, we focus in this work in a two stage system in order to improving SR approaches for periocular images in order to enhance the recognition performance of periocular-based biometric systems. Therefore, we describe in Sect. 3.1 the proposed ESISR technique proposed and in Sect. 3.2 the feature extraction and comparison methods utilised for periocular recognition.

### 3.1. Stage-1: Super-Resolution

Since SR in general is an image-to-image translation task where the input image is highly correlated with the target image, researchers try to learn only the residuals between them (i.e. global residual learning). This process avoids learning a complicated transformation from a complete image to another. Instead, it only requires learning a residual map to restore the missing high-frequency details. Since most regions' residuals are close to zero, the model complexity and learning difficulty are thus greatly reduced.

This local residual learning is similar to ResNet to alleviate the degradation problem caused by ever-increasing network depths, reduce training difficulty, and improve the learning ability. For these reasons, we are using recursive learning to learn higher-level features without introducing an overwhelming number of parameters, which means applying the same modules multiple times.

In addition to choosing an appropriate network architecture, the definition of the perceptual loss function is critical for the performance of the proposed method based on the DCSCN network, as mentioned in Sects. 1 and 2. While SR is commonly based on the MSE, PSNR, and SSMI metrics, we have designed a loss function that incorporates as well the sharpness with respect to perceptually relevant features. The function thus balances between reconstructing images by minimising the difference of the sharpness values and weights the results of SSIM and PSNR.

In this section, we present an efficient image SR network that is able to recover periocular images from selfies (ESISR). Our network includes two building blocks, as it can be observed in Fig. 2: A feature extraction and a reconstruction stage based on DCSCN. The description of each stage of the algorithm are described in the remainder of this section.

### *3.1.1. Pre-processing*

The original RGB images captured with a smartphone represent an additive color-space where colors are obtained by a linear combination of Red, Green, and Blue values. The three channels are thus correlated by the amount of light on the surface. In order to avoid such correlations, all the images were converted from RGB to YCbCr. The YCrCb color space is derived from RGB, and separates the luminance and chrominance components into different channels. In particular, it has the following three components: i) Y, Luminance or Luma component obtained from RGB after gamma correction; ii) $Cr = R \neg Y$, how far is the red component from Luma; and iii) $Cb = B \neg Y$, how far is the blue component from Luma. We only use Y component in this work. The periocular image areas were automatically cropped from faces to the size of $250 \times 200$ pixels.

### *3.1.2. Feature extraction*

As mentioned above, the Y component of the converted image is used as input of our model. Several patches of $32 \times 32$ and $48 \times 48$ pixels were extracted from the image and used to grasp the features efficiently. We look for the features that achieve a better trade-off between the number and size of filters of each CNN layer. Seven blocks of $5 \times 5$ and $3 \times 3$ have been selected. The information is extracted using small convolutional blocks with residual connections and stride convolutions in order to preserve both the global and the fine the details in the periocular images. Only the final features from $3 \times 3$ and $5 \times 5$ pixels are concatenated, following the recursive pixel-shuffle approach (see Fig. 3). These local skip connections in residual blocks make the network easier to optimize, thereby supporting the construction of deeper networks.

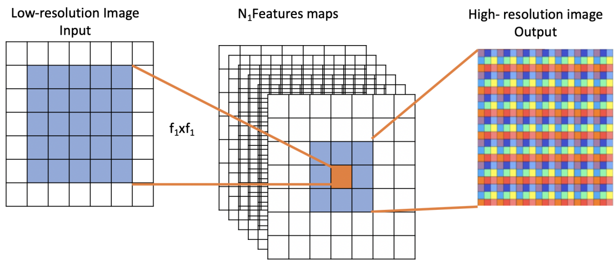



Figure 3: Pixel-shuffle convolution layer that aggregates the feature maps from LR space and builds the SR image in a single step. Based on [24].

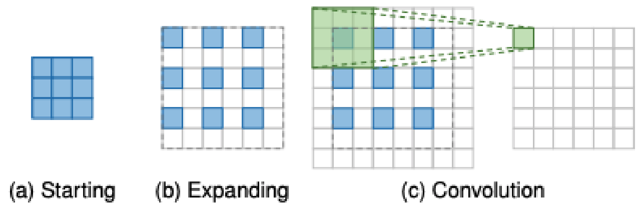



Figure 4: Transpose-convolution operation representation. (a) The starting matrix represents the input image. (b) Expanding operation adds zeros to the images in order to increase the size. c) The convolution operation is performed again in a new resolution. Based on [5].

A model with transpose convolution instead of pixel-shuffle was trained to explore the quality of the reconstruction images [5]. See Fig. 4. Transpose convolution operates conversely to normal convolution, predicting the input based on feature maps sized like convolution output. It increases image resolution by expanding the image by adding zeros and performing convolution operations.

### *3.1.3. Reconstruction*

Our reconstruction stage uses only one convolutional block with 2 layers (Conv + Relu + Conv) in a recursive path. This block includes $3 \times 3$ convolutions and pixel-shuffle algorithm (see Fig. 2) to create a high-resolution image from a low-resolution input. Batch normalization was removed. An optimized sub-pixel convolution layer that learns a matrix of upscaling filters to increase the final LR feature maps into the SR output was used.

### *3.1.4. Perceptual loss function*

The ISO/IEC 29794-6[2] standard methods used "to quantify the quality of iris images, normative requirements on software and hardware measuring the utility of iris images, terms, and definitions for quantifying iris image quality, and standardized

[2] `https://www.iso.org/standard/54066.html`

encoding of iris image quality:

$$LoG(x,y) = -\frac{1}{\pi\sigma^4}\left[1 - \frac{x^2+y^2}{2\sigma^2}\right]e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3}$$

The Laplacian of Gaussian operator(LoG) is thus the sharpness metric used in this work. Calculation of the sharpness of an image is determined by the power resulting from filtering the image with a Laplacian of Gaussian kernel(F). The standard deviation of the Gaussian is 1,4.

Now, it is important to highlight that the loss function aims to improve the quality of the reconstruction. To that end, we combine the SSIM and PSNR classical SR metrics with the sharpness metric for iris images recommended, as follows:

$$\begin{aligned} L(I_{LR}, I_{HR}) = 0.5 \cdot \mathrm{LoG}\,(I_{LR}, I_{HR}) \cdot [0.25 \cdot \mathrm{SSIM}\,(I_{LR}, I_{HR}) \\ + 0.25 \cdot \mathrm{PSNR}\,(I_{LR}, I_{HR})] \end{aligned} \tag{4}$$

where $I_{LR}$ represents a low-resolution image, $I_{HR}$ the corresponding high-resolution image recovered, and LoG the sharpness as defined in Eq. 3. The best values of the weights for each specific metric (i.e., 0.25, 0.25 and 0.50) were estimated in a grid search with a train dataset.

### 3.2. Stage-2: Periocular recognition

Most traditional methods in the state of the art are based on machine learning techniques with different feature extraction approaches such as HOG, LBP and BSIF or the fusion of some of them [6]. However, today we have powerful pre-trained deep learning methods based on faces images. Using transfer learning techniques, the information extracted from some layers using fine-tuning techniques or embedding approaches could be suitable to perform periocular verification. This is the approach followed in this article.

This task involves information from periocular images estimating an eye embedding vector for a new given eye from a selfie image. An eye embedding is a vector that represents the features extracted from the eyes periocular images and comparing it with the embedding vector. This comparison occurs using euclidean distance to verify if the distance is below a predefined threshold, often tuned for a specific dataset or application. For this paper, a VGGFace[25] and FaceNet [26] models have been used as a feature extractor for periocular recognition.

## 4. Experimental Setup

### 4.1. Experimental Protocol

In order to assess the soundness of the proposed method, we have first evaluated the SR approaches and then the complete pipeline including the periocular recognition stage.

*Experiment 1* we have trained traditional DCSCN, WDSR-A and SRGAN methods as a baseline for benchmarking purposes. The main properties and default parameters of those methods are summarised in the following.

**DCSCN**: Number of CNN layers = 12, Number of first CNN filters = 196, Number of last CNN filters = 48, Decay Gamma = 1.5, Self Ensemble = 8, Batch images for training epoch = 24,000, Dropout rate = 0.8, Optimizer function = Adam, Image size for each Batch = 48, Epochs = 100, Early stopping = 10.

**WDSR-A** Number of residual blocks = 8, Number of CNN layers in the main branch = 6, Number of expansion of residual blocks = 4, Number of filters main branch = 64, Number of filters residual blocks = 256, Activation function = Relu, Optimization Function = Adam, Learning Rate = 1e-4 and 1-e-5, Beta = 1e-7, Size of batch images = 96, Number of steps = 60,000.

**SRGAN** The SRGAN has two stages:

*Generator*: This stage is used for learning the inverse function for downsampling the image and to generate the LR images from their respective HR, This stage is based in a pre-trained VGG-54. The following parameters are used: Number of residual blocks = 16, Number of CNN layers with residual blocks = 2, activation function residual block = PRelu, Kernel size residual block = 3, CNN layers = 3, kernel size = 9, 3 and, 9. Filters numbers = 64, Optimization function = Adam, Learning rate = 1e-4 and 1e- 5, batch image size = 96, Steps = 100,000, mini size batches = 16.

*Discriminator*: In order to evaluate the similarity between the images generated by the SR generator (VGG-54) and the HR images, the architecture discriminator is trained with the following parameters: CNN layers = 8, Filter numbers:64, 64, 128, 128, 256, 256, 512 and 512. Kernel size = 3, activation function = Relu, Momentum batch normalization = 0.8, Optimization function = Adam, Learning Rate = 1e-5 and, 1e-6, Batch size = 16, Steps = 100,000.

In order to test our ESISR method we carried out two different experiments for the SR.

*Experiment 2* evaluates our ESISR method using the pixel-shuffle technique.

*Experiment 3* analyses our proposal further improving the efficiency of the Experiment 2, and also studies the transpose convolution instead of pixel-shuffle. All the experiments measure the quality of the produced SR images using the sharpness function defined in Eq. 3, and the efficiency in terms of the number of features and parameters. It should be noted that the True Sharpness represents the sharpness of the original image, and Output Sharpness represents the sharpness of the output image created by ESISR. Therefore, the goal is to achieve an Output Sharpness as close as possible to the True Sharpness. From those experiments, we selected the configuration achieving the best performance. All methods were trained using the Samsung database and tested with *SET-5E* dataset.

For periocular verification system, first we extract the embedded information from original periocular images with SR. Afterwards feature extraction was applied to best super-resolved images using x2, x3 and x4 increased sizing. All the SR methods for periocular verification were tested using the MOBIO dataset. This dataset is totally different that was used to train the SR stage.

A PC with Intel I7, 32 GB RAM, and GPU-1080TI was used for all the experiments.

*4.2. Databases*

In order to analyse the performance of the SR algorithm, three databases were used. A new dataset was acquired in a collaborative effort with subjects from different countries with Samsung smartphones using the app, specially designed for this purpose `visualselfie.org`[3]. This app was designed in order to capture different variations of selfie scenarios in three distances, as depicted in Fig. 5. In more detail, 800 images were selected to be used for training and 100 for testing[4].

From the training dataset, 228,700 patches of $48 \times 48$ px. were created for experiment 2 and $32 \times 32$ for experiment 3.

Figure 5: Example of Samsung databases. Left: closest position. Middle: half arm extended. Right: full arm extended.

A second dataset called *Set-5E* was created to validate the results. This database has 100 images from different subjects acquired with different smartphones extracted from the CSIP database in the visual spectrum [27]. It has 2004 images, depicting 50 subjects over 10 different mobile setups.

A third database MOBIO was used to super-resolved the size of the images with the best pre-trained super-resolution model (ESISR). And was used to measure the performance of the eyes verification system. The MOBIO dataset comprises the biometric data from 152 volunteers. Each subject provided samples of face, iris, and voice. There are in average 8 images for each subject from a NOKIA N93i mobile. Some examples are presented in Fig. 6

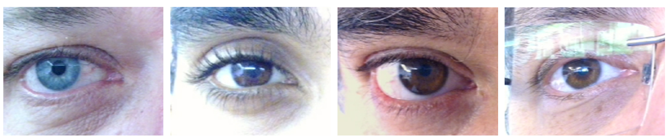


Figure 6: MOBIO database examples.

## 5. Results and Discussion

*5.1. Super-resolution models*

**Experiment 1**: This experiment was used as a baseline in order to evaluate the state of the art SR methods and our efficient proposal. The DCSCN, WSDR-A, and SR-GAN were tested with default parameters.

**Experiment 2**: Our proposed and efficient ESISR method was tested using *pixel-shuffle* and the new loss function including the Sharpness metric (see Eqs. 3 and 4). The best parameters for our proposal were: Number of CNN layers = 7, Number of first CNN filters = 32, Number of last CNN filters = 8, Decay Gamma = 1.2, Self Ensemble = 8, Batch images for training epoch = 24,000, Dropout rate = 0.5, Optimizer function = Adam, Image size for each Batch = 32, Epochs= 100, Early stopping = 10.

Table 1 summarizes the results: Rows 1-3 show the results for traditional SR methods (DSCN with 12 layers and $96 \times 96$ patches, WDSR-A with 8 residual blocks and $62 \times 62$ patches, SR-GAN with 16 residual blocks and $96 \times 96$ patches). Rows 4 up to 6 present the results of our proposed method: ESISR-1 using the pixel-shuffle algorithm with only 7 convolutions layers and $48 \times 48$ patches, ESISR-2 using the pixel-shuffle algorithm with only 7 convolutions layers and $32 \times 32$ patches, and ESISR-3 using the transposed convolution algorithm with only 7 convolutions layers.

First we should note that all the image enlargement x2, x3, and x4 extract the same number of features for each method (i.e., 1,301 for DCSNN and 1,000 for ESISR). The bigger difference lies on the number of parameters of each method. The DCSCN, WSDR-A and SR-GAN methods need a large number of parameters: for images increased for x2, the parameters numbers are: 1,754,942; 597,000 and 24,864,000; for images increased for x3, the parameter numbers are: 2,170,142; 603,000 and 25.131.000 respectively, and for images increased for x4m the parameter numbers are: 2,087,102; 610,000 and 26,939,000, respectively. These numbers are drastically reduced by the our ESISR proposed method, we needs only 27.209 parameters when the image is increased by x2, 28.654 parameters when increased by x3, and 64.201 parameters when increased by x4.

In addition to that gain in terms of efficiency, we may observe in Table 1 that the newly proposed loss function based on sharpness allows us to get a good reconstruction. The Output sharpness for each scale value is similar to the values obtained by DSCN (e.g. 16.85 *vs*. 16.70 for x2), and also close to the target True Sharpness of 17.04. Therefore, we may conclude that the proposed method keeps the sharpness quality of the images, thereby making it suitable for SR applications for mobile devices.

**Experiment 3**: in addition to the configuration of ESISR tested in Experiment 2, we also evaluated two additional approaches in our experiment. First, the most efficient implementation of ESISR with a big reduction of features (down to 131) and a number of parameters with pixel-shuffle and $32 \times 32$ was analysed (Table 1, row 5). Then, we also tested the method using *transposed convolution* with the same number of 131 features (Table 1, row 6). The Transpose convolutions layer is an inverse convolutions layer that will both up-sample input and learn how to fill in details during the model training process, at the cost of increasing the number of parameters (i.e., less efficient than pixel-shuffling). As we may observe in Table 1, the

[3]Only available from smartphones

[4]A similar number of images are used in SOTA for general-purpose method. For instances, DIV2K database.

pixel-shuffle with $32 \times 32$ px. uses the same number of parameters as with $48x \times 48$ px. In contrast, the transposed convolution requires 100,316 parameters when the image is increased by 2 (x2), 109,564 parameters when increased by 3 (x3), and 100,318 parameters when increased by 4 (x4). In spite of this increase, the ESISR is still 10 to 20 times more efficient than the traditional DCSCN.

Regarding the quality of the SR iris images, we can observe that both configurations tested in this last experiment (row 5-6) achieve a similar sharpness for the x3 and x4 scale values (14.43, 14.38, 15.46 and 16.32), but not for x2. In the latter case, the pixel-shuffle approach clearly outperforms the transpose-convolution method (15.43 *vs*. 14.38). The lower result of reconstruction was reached for the SRGAN method with a higher number of parameters and a relevant difference of the value of output sharpness. Reconstruction examples are presented in appendix section 8 at the end of the paper.

### 5.2. Periocular SR verification

Our periocular verification systems including a SR stage analyzed in the previous section. In order to measure the quality of the super-resolved images the MOBIO dataset was used to evaluated the performance of the reconstruction (x2, x3 and x4) using the best SR method proposed in the section 5.1 ESISR with pixel-shuffle.

Two experiments were defined in order to explore the quality of the verification systems as following.

*Experiment 1* A FaceNet pre-trained model was used to extract the embedding information.

*Experiment 2* A VGGFace pre-trained model was used to extract the embedding information.

For VGGFace the feature vector has a size of 2,622. For FaceNet the feature vector has a size of 1,722.

Fig. 7 shows the Probability Density Functions of the embedding data-vector for FaceNet and VGGFace. The VGGFace data-vector is more spread between 0.1 and 1.0 instead of the FaceNet data-vector which is concentrated between 0.1 and 0.4. Both distributions shown an overlap when one embedded vector from a selfie is compared with the same identity of the gallery and with other other users.

Fig. 8, shows the DET curves results of periocular verification system with a normal resolution in dark-blue and for SR images by x2, x3 and x4 using the ESISR proposed method in other tones of green and dashed lines.

Table 2, shows the results for different sizes of SR images increased by a factor of x2, x3, and x4 and its benchmark with the pre-trained FaceNet model and VGGFace as a feature extractor. VGGFace reached the best results with a lower Equal Error Rate (EER) of 0.145. Row 2, shows the results of x3. The results reported show the EER and False Not Match Rate (FNMR) based on False Match Rate (FMR) at 1%.

FaceNet reached the best results with an EER of 8.7% for images without SR. Row 1, shows the results of Low resolution size. The lower result with an EER 9.5% was for SR x4. Row 4, shows the results of x4.

The best results for VGGFace reached an EER of 9.90% for SR x4. Row 4, shows the results of x4. The lower result reached

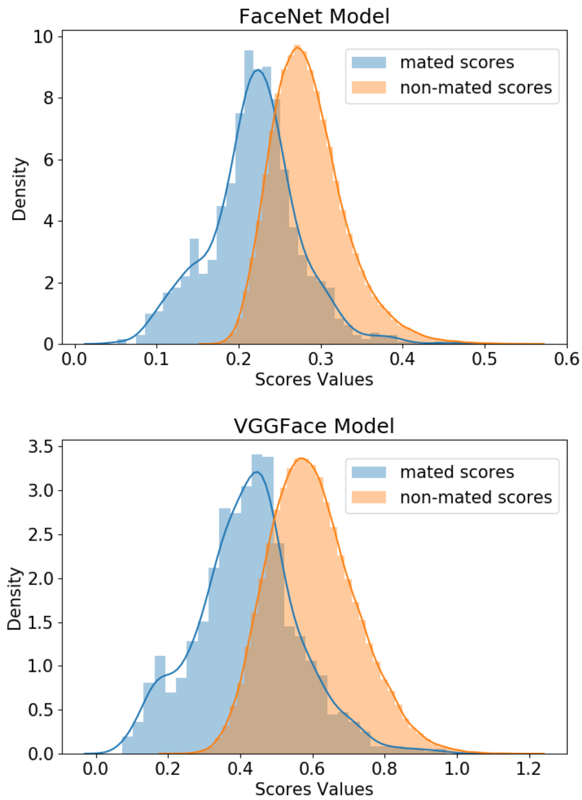



Figure 7: Mated and Non-mated score distributions for FaceNet (left) and VGGFace (right)

.

an EER 10.05% for images without SR. Row 1, shows the results of normal size. Overall, FaceNet reached the best results in all the models with x2, x3 and x4 in comparison with VGGFace. This is interesting for high security applications, since operating points are usually defined at small FNMR values.

The results are related with the size of the embedded vector extracted from pre-trained model showed that the feature extracted from FaceNet are more representative and general purpose than VGGFace.

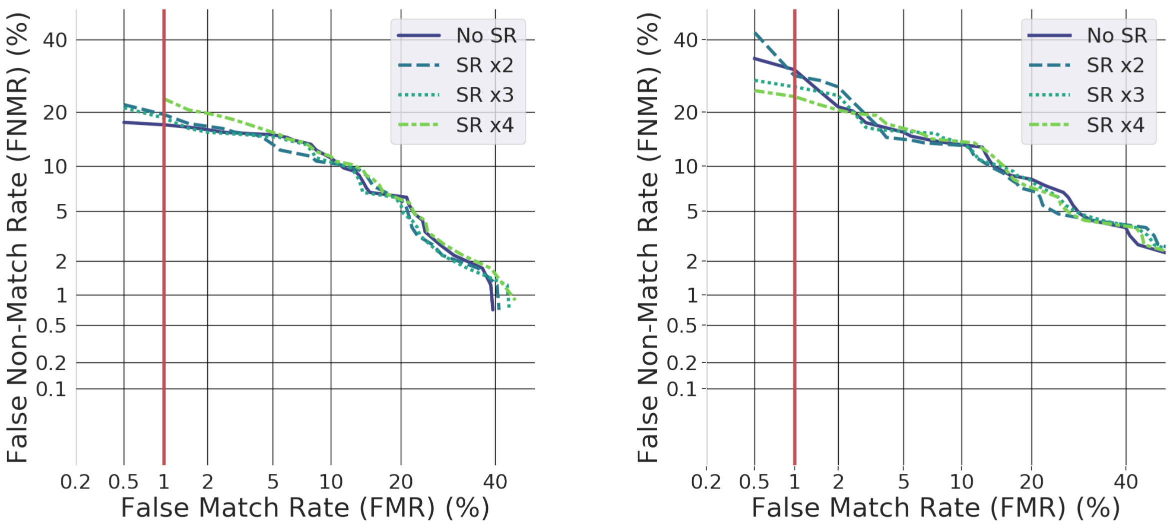
It is important to highlight that the three scales keep the quality of the periocular verification based on sharpness perceptual loss proposed. Thus, a weighted perceptual loss help to keep the quality of the images based on Sharpness metrics. Clearly, this is the most suitable metric for been applied in periocular iris images with SR than the traditional SNR and SSIM.

## 6. Conclusion

In this paper, we have proposed an efficient and accurate Image Super Resolution method focused on the generation of enhanced eyes images for periocular verification purposes using selfie images. To that end, we developed a two-stage approach based on a CNN with pixel-shuffle, a new loss function based on a sharpness metric (see Eq. 3), compliant with the ISO/IEC 29794-6 standard and a selfie periocular verification proposal.

Table 1: Summary of the results for 3 different scales (x2, x3, and x4) for our system (ESISR) with different configurations and the benchmark with DCSCN, WDSR-A, and SRGAN. True Sharpness denotes the sharpness for the original image (LR), and Output Sharpness the sharpness for reconstructed SR images.

| Method | Conv. | #Feature | Scale | #Param | PSNR | SSIM | True Sharp. | Output Sharp. |
|---|---|---|---|---|---|---|---|---|
| DCSCN [9] | | 1,301 | x2 | 1,754,942 | 37.11. | 0.95 | 17.04 | 16.85 |
| | | | x3 | 2,170,142 | 32.82 | 0.91 | 18.05 | 16.45 |
| | | | x4 | 2,087,102 | 30.52 | 0.86 | 16.90 | 12.47 |
| WDSR-A [10] | Pixel-shuffle | | x2 | 597,000 | 47.87 | 0.98 | 17.04 | 10.89 |
| | | | x3 | 603,000 | 46.59 | 0.97 | 18.05 | 10.82 |
| | | | x4 | 610,000 | 43.92 | 0.94 | 16.90 | 10.72 |
| SRGAN [11] | Pixel-shuffle | | x2 | 24.864.000 | 39.66 | 0.96 | 17.04 | 10.82 |
| | | | x3 | 25.131.000 | 38.72 | 0.94 | 18.05 | 10.95 |
| | | | x4 | 26.930.000 | 34.09 | 0.88 | 16.90 | 10.64 |
| ESISR-1 | Pixel-shuffle 48x48 | 1,000 | x2 | 27,209 | 36.49 | 0.95 | 17.04 | 16.70 |
| | | | x3 | 28,654 | 32.89 | 0.90 | 18.05 | 16.01 |
| | | | x4 | 64,201 | 29.08 | 0.86 | 16.90 | 12.00 |
| ESISR-2 | Pixel-shuffle 32x32 | 131 | x2 | 27,209 | 38.91 | 0.90 | 17.04 | 15.43 |
| | | | x3 | 28,654 | 36.78 | 0.85 | 18.05 | 15.46 |
| | | | x4 | 64,201 | 35.47 | 0.81 | 16.90 | 16.34 |
| ESISR-3 | Transpose Convolution | 131 | x2 | 100,316 | 35.52 | 0.81 | 17.04 | 14.38 |
| | | | x3 | 109,564 | 36.84 | 0.85 | 18.05 | 15.06 |
| | | | x4 | 100,318 | 35.52 | 0.81 | 16.90 | 16.14 |




(a) FaceNet

(b) VGGFace

Figure 8: DET curves benchmarking the performance of the baseline method (without SR) and SR with x2, x3 and x4.

Table 2: Verification results for FaceNet and VGGFace. FNMR is given at FMR=1%.

| | FaceNet | | VGGFace | |
|---|---|---|---|---|
| **Method** | **EER(%)** | **FNMR** | **EER(%)** | **FNMR** |
| LR image | 8.70 | 18.01 | 10.05 | 31.01 |
| ESISR x2 | 9.21 | 20.02 | 9.94 | 29.12 |
| ESISR x3 | **8.90** | 19.12 | 9.92 | 27.04 |
| ESISR x4 | 9.52 | 24.01 | **9.90** | 24.05 |

In the feature extraction stage of our method, the structure of the CNN model extracts optimised features, which are subsequently sent to the reconstruction network. In this latter network, we only used a recursive convolutional block with pixel-shuffle to obtain a better reconstruction performance with reduced computational requirements. In addition, the model is designed to be capable of processing original size images. Using these techniques, our model can achieve state of the art performance with a fewer number of parameters (from the state of the art DSCN with 2 million parameters, we achieve a comparable quality with 27,000 parameters).

The perceptual loss function based on image sharpness that we propose allows us to keep the sharpness of iris images in the reconstructed images by x2, x3, and x4. This approach to improving the quality of the reconstruction and the SR in periocular recognition systems to be implemented in mobile devices.

Regarding to the verification system, FaceNet reached the best results in comparison to VGGFace. An EER of 8.7% without SR and 9.2% for x2, 8.9% for x3, and 9.5% for x4 respectively. Conversely, An small improvement in performance was reached when VGGFace was used. An EER of 10.05% without SR and 9.94% for x2, 9.92% for x3, and 9.90% for x4 respectively.

Overall, there are marginal improvements for verification systems when only the size of the images is considered in combination with SR images. The information extracted with an embedded vector from the periocular area with a pre-trained model has a high quality of information for verification because of the huge number of filters used during the training process.

The uncontrolled conditions such as sunlight, occlusions, rotations, or the number of people in an image when a remote selfie is capturing could be more challenging than the size of the image for RGB selfie images. The extension of this improvement to NIR iris images must be studied in a separate work. Those uncontrolled conditions need to be studying to improve the selfie periocular verification systems.

In future work, we are continue collected images in order to train a specific periocular verification system based on CNN from scratch or using transfer-domain techniques. In relation to the number of images, we believed if we used state of the art pre-train models, the machine learning-based methods can be replaced by the CNN models. This, the selection of the pre-trained models should be taken into account.

## 7. Acknowledgments

This research work has been partially funded by Fondecyt Iniciacion 11170189, the DFG-ANR RESPECT Project (406880674), and the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

## References

[1] S. Gonzalez, A. Valenzuela, J. Tapia, Hybrid two-stage architecture for tampering detection of chipless id cards, IEEE Transactions on Biometrics, Behavior, and Identity Science (2020) 1–1doi:10.1109/TBIOM.2020.3024263.

[2] F. Alonso-Fernandez, J. Bigun, A survey on periocular biometrics research, Pattern Recognition Letters 82 (2016) 92–105.

[3] F. Alonso-Fernandez, K. B. Raja, R., Raghavendra, C. Busch, J. Bigun, R. Vera-Rodriguez, J. Fierrez, Cross-sensor and cross-spectral periocular biometrics: A comparative benchmark including smartphone authentication, arXiv preprint arXiv:1902.08123.

[4] J. Tapia, C. Arellano, I. Viedma, Sex-classification from Cellphones Periocular Iris Images, Springer International Publishing, Cham, 2019, Ch. 10, pp. 227–242. doi:10.1007/978-3-030-26972-2_11.

[5] Z. Wang, J. Chen, S. C. H. Hoi, Deep learning for image super-resolution: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 1–1doi:10.1109/TPAMI.2020.2982166.

[6] P. Kumari, K. Seeja, Periocular biometrics: A survey, Journal of King Saud University - Computer and Information Sciencesdoi:https://doi.org/10.1016/j.jksuci.2019.06.003.
URL http://www.sciencedirect.com/science/article/pii/S1319157818313302

[7] R. Timofte, R. Rothe, L. V. Gool, Seven ways to improve example-based single image super resolution, CoRR abs/1511.02228. arXiv:1511.02228.
URL http://arxiv.org/abs/1511.02228

[8] Juan Tapia, Marta Gomez-Barrero, Christoph Busch , An efficient super-resolution single image network using sharpness loss metrics for iris, In press.

[9] J. Yamanaka, S. Kuwashima, T. Kurita, Fast and accurate image super resolution by deep CNN with skip connection and network in network, CoRR abs/1707.05425. arXiv:1707.05425.
URL http://arxiv.org/abs/1707.05425

[10] J. Yu, Y. Fan, J. Yang, N. Xu, Z. Wang, X. Wang, T. S. Huang, Wide activation for efficient and accurate image super-resolution, CoRR abs/1808.08718. arXiv:1808.08718.
URL http://arxiv.org/abs/1808.08718

[11] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, CoRR abs/1609.04802. arXiv:1609.04802.
URL http://arxiv.org/abs/1609.04802

[12] C. Dong, C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Transactions on Pattern Analysis & Machine Intelligence 38 (02) (2016) 295–307. doi:10.1109/TPAMI.2015.2439281.

[13] J. Kim, J. K. Lee, K. M. Lee, Deeply-recursive convolutional network for image super-resolution, CoRR abs/1511.04491. arXiv:1511.04491.
URL http://arxiv.org/abs/1511.04491

[14] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, Z. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, CoRR abs/1609.05158. arXiv:1609.05158.
URL http://arxiv.org/abs/1609.05158

[15] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, CoRR abs/1411.4038. arXiv:1411.4038.
URL http://arxiv.org/abs/1411.4038

[16] B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, Enhanced deep residual networks for single image super-resolution, CoRR abs/1707.02921.

arXiv:1707.02921.
URL http://arxiv.org/abs/1707.02921

[17] E. Ribeiro, A. Uhl, F. Alonso-Fernandez, Iris super-resolution using cnns: is photo-realism important to iris recognition?, IET Biometrics 8 (1) (2019) 69–78. doi:10.1049/iet-bmt.2018.5146.

[18] F. Alonso-Fernandez, R. A. Farrugia, J. Bigun, J. Fierrez, E. Gonzalez-Sosa, A survey of super-resolution in iris biometrics with evaluation of dictionary-learning, IEEE Access 7 (2019) 6519–6544. doi:10.1109/ACCESS.2018.2889395.

[19] K. B. Raja, R. Raghavendra, V. Vemuri, C. Busch, Smartphone based visible iris recognition using deep sparse filtering, Pattern Recognition Letters 57. doi:10.1016/j.patrec.2014.09.006.

[20] K. Ahuja, R. Islam, F. A. Barbhuiya, K. Dey, Convolutional neural networks for ocular smartphone-based biometrics, Pattern Recognition Letters 91 (2017) 17 – 26, mobile Iris CHallenge Evaluation (MICHE-II). doi:https://doi.org/10.1016/j.patrec.2017.04.002.
URL http://www.sciencedirect.com/science/article/pii/S0167865517301046

[21] C. N. Padole, H. Proença, Periocular recognition: Analysis of performance degradation factors, in: IAPR International Conf. on Biometrics - ICB, Vol. March, 2012, pp. 1–7.

[22] K. Raja, R. Raghavendra, M. Stokkenes, C. Busch, Multi-modal authentication system for smartphones using face, iris and periocular, in: 2015 International Conference on Biometrics (ICB), 2015, pp. 143–150. doi:10.1109/ICB.2015.7139044.

[23] K. Hernandez-Diaz, F. Alonso-Fernandez, J. Bigun, Periocular recognition using cnn features off-the-shelf, in: 2018 International Conference of the Biometrics Special Interest Group (BIOSIG), 2018, pp. 1–5. doi:10.23919/BIOSIG.2018.8553348.

[24] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, Z. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, CoRR abs/1609.05158.
URL http://arxiv.org/abs/1609.05158

[25] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: British Machine Vision Conference, 2015, p. 1.

[26] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823. doi:10.1109/CVPR.2015.7298682.

[27] G. Santos, E. Grancho, M. V. Bernardo, P. T. Fiadeiro, Fusing iris and periocular information for cross-sensor recognition, Pattern Recognition Letters 57 (2015) 52 – 59, mobile Iris CHallenge Evaluation part I (MICHE I). doi:https://doi.org/10.1016/j.patrec.2014.09.012.

## 8. Appendix: Examples of SR Images

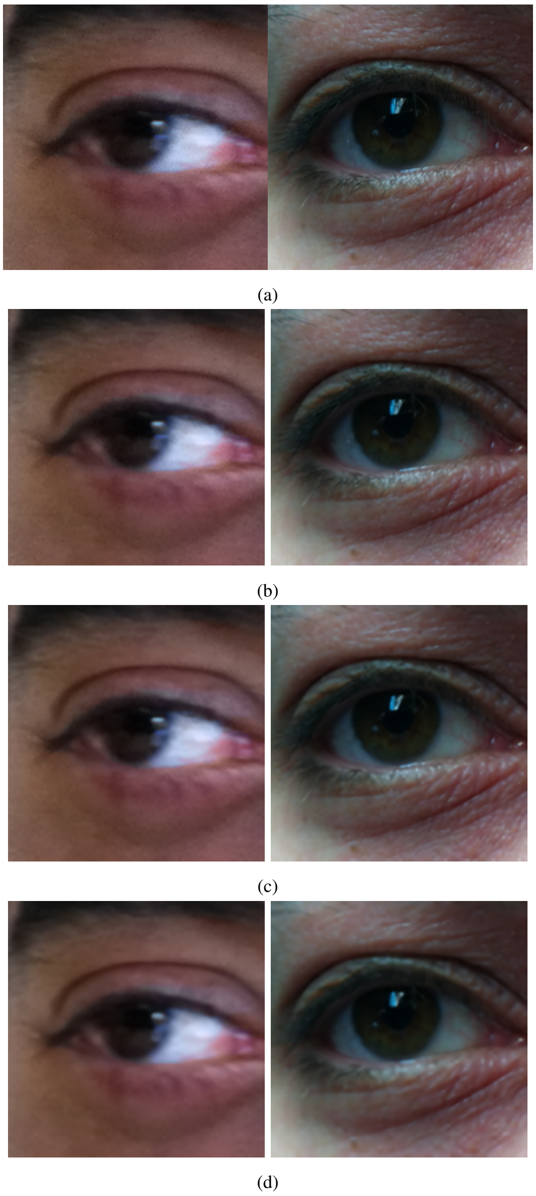



Figure 9: MOBIO examples (a) without SR, (b) with WSDR-A SR x2, (c) with WDSR-A SR x3, and (d) with WDSR-A SR x4.

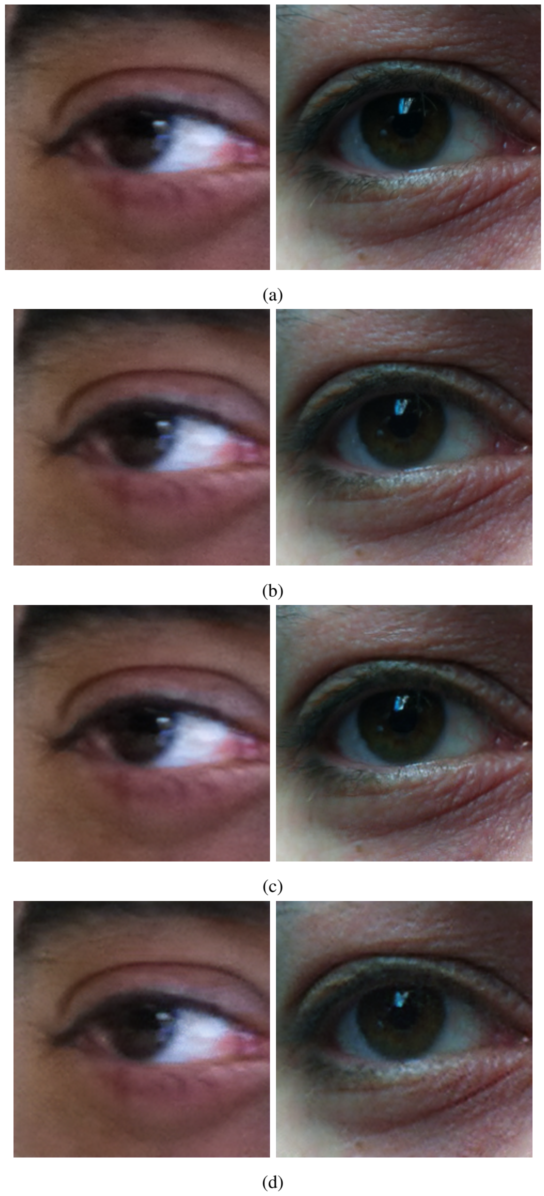
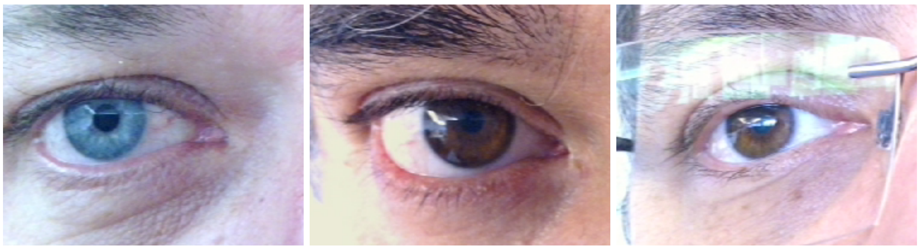
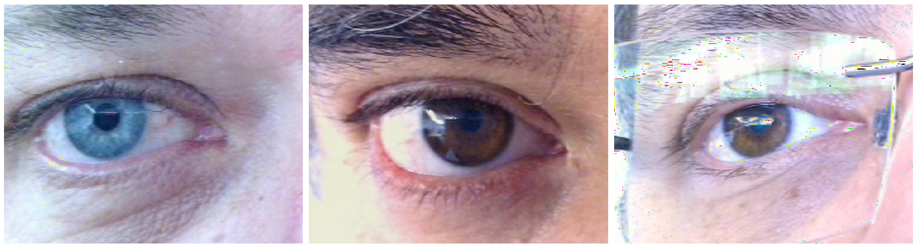


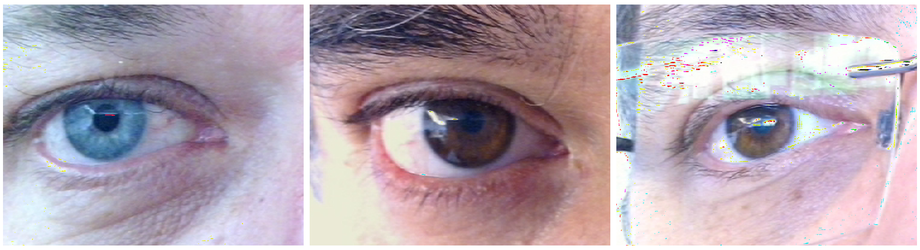

Figure 10: MOBIO examples (a) without SR, (b) with SRGAN SR x2, (c) with SRGAN SR x3, and (d) with SRGAN SR x4;



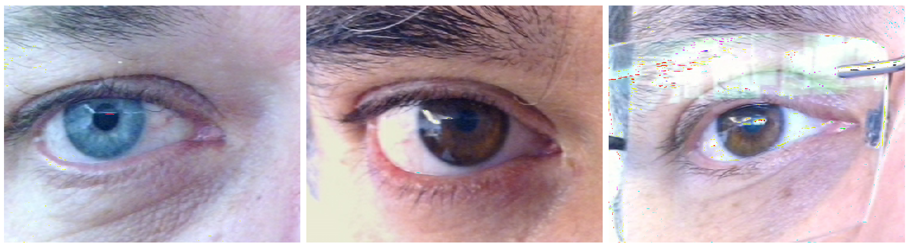
(a)



(b)



(c)



(d)

Figure 11: MOBIO examples (a) without SR, (b) with ESISR SR x2, (c) with ESISR SR x3, and (d) with ESISR SR x4.