redin Revista Facultad de Ingeniería, Universidad de Antioquia, No.112, pp. 98-110, Jul-Sep 2024

Enhancing facial recognition in surveillance systems through embedded super-resolution

Mejorando el reconocimiento facial en sistemas de vigilancia mediante super-resolución embebida

Andrés David Gómez-Bautista[1*] Francisco Carlos Calderón-Bocanegra[1]

[1]Departamento de Electrónica, Pontificia Universidad Javeriana. Carrera 7 # 40-62. C. P. 110231. Bogotá, Colombia.

CITE THIS ARTICLE AS:

A. D. Gómez-Bautista and F. C. Calderón-Bocanegra "Enhancing facial recognition in surveillance systems through embedded super-resolution", *Revista Facultad de Ingeniería Universidad de Antioquia*, no. 112, pp. 98-110, Jul-Sep 2024. [Online]. Available: https://www.doi.org/10.17533/udea.redin.20240203

ARTICLE INFO:

Received: June 15, 2023
Accepted: February 08, 2024
Available online: February 08, 2024

KEYWORDS:

Super-resolution; face enhancement; computer vision; surveillance; machine learning

Súper resolución; mejora de rostros; visión por computador; video vigilancia; aprendizaje supervisado

ABSTRACT: This document details the implementation of a sub-pixel convolutional neural network designed to enhance the resolution of face images. The model uses a series of filters to progressively increase the number of pixels, estimating the necessary information for new pixels from the original image and training derived from 22000 synthetic images produced by adversarial neural networks. Within the context of surveillance and related applications, the trained convolutional network exhibits beneficial characteristics. For instance, it can be deployed within a device to achieve higher-resolution images than those the physical camera can produce. This research underscores the feasibility of such a device through the implementation and evaluation of the network on the NVIDIA Jetson TX2 embedded system. The findings demonstrate the model's practicality for real-time surveillance applications and its ability to produce superior-quality images compared to several interpolation methods, as determined by an exhaustive testing process measuring various attributes of the generated images.

RESUMEN: Este documento detalla la implementación de una red neuronal convolucional de aplicación a nivel sub-pixel diseñada para mejorar la resolución de imágenes faciales. El modelo utiliza una serie de filtros para aumentar progresivamente el número de píxeles, estimando la información necesaria para los nuevos píxeles tanto de la imagen original como del entrenamiento derivado de 17,500 imágenes sintéticas producidas por redes neuronales adversarias. Dentro del contexto de la vigilancia y aplicaciones relacionadas, la red neuronal convolucional entrenada muestra características beneficiosas. Por ejemplo, se puede implementar dentro de un dispositivo para lograr imágenes de mayor resolución de las que la cámara física puede producir. Esta investigación subraya la viabilidad de dicho dispositivo a través de la implementación y evaluación de la red en el sistema embebido NVIDIA Jetson TX2. Los hallazgos demuestran la practicidad del modelo para aplicaciones de vigilancia en tiempo real y su capacidad para producir imágenes de calidad superior en comparación con varios métodos de interpolación, según lo determinado por un proceso de prueba exhaustivo que mide varios atributos de las imágenes generadas.

1. Introduction

The rapid progress of digital imaging technology in the past few decades has led to a significant increase in the application of image processing techniques. Super-resolution is one of these techniques, and it has received considerable attention due to its potential to enhance low-resolution images and produce high-resolution counterparts. This is particularly critical in surveillance applications, where image clarity and detail are paramount.

This paper aims to explore the application of an existing super-resolution method, the Efficient Sub-Pixel Convolutional Neural Network (ESPCN), in the context of video surveillance. The main contribution of this paper lies not in the proposal of a new method, but rather in the novel application and performance comparison of the ESPCN, with traditional interpolation techniques, in the specific scenario of video surveillance.

We further enhance the ESPCN implementation by

* Corresponding author: Andrés David Gómez-Bautista
E-mail: gomezan@javeriana.edu.co
ISSN 0120-6230
e-ISSN 2422-2844

DOI: 10.17533/udea.redin.20240203

applying it to all color components of an image rather than just the luma component, as traditionally done. This alteration allows us to better understand the potential effects of the model on the entire image, not just a single-color space.

Additionally, we leverage a unique dataset, comprised of generated faces instead of real-life human beings, to avoid ethical dilemmas. To our knowledge, the use of this dataset for the training and validation of the ESPCN, is a unique approach in the field.

The hardware implementation on an embedded system like Jetson TX2 further distinguishes our work. We strive to demonstrate the ESPCN's practical feasibility and efficiency in a real-world, resource-constrained setting.

We aim to present a comprehensive study involving the ESPCN's training on a unique dataset, its comparison with classical methods, and its implementation on an embedded system. This study can provide valuable insights for researchers and practitioners interested in utilizing super-resolution techniques for surveillance applications.

Throughout the paper, we will introduce the applied method and its variations, describe the implementation in detail, discuss the experimentation results, and, finally, provide a conclusion summarizing the findings and potential implications of the study.

2. Literature review

As technology continues to advance, it increasingly influences different facets of our lives, including surveillance and security applications. This is evident in the plethora of cameras and video recording nodes deployed across cities, which continually generate vast amounts of data. Within surveillance, particular interest lies in the treatment of facial images in video footage. Multiple techniques and algorithms focus on tasks like face detection [1, 2], identity verification [3, 4], face capturing [5], and face hallucination [6, 7]. However, this type of image processing presents several challenges, including multiple noise sources, motion blur [8], environmental disturbances due to illumination changes and contrast [9], and insufficient sensor density [10]. All these contribute to image degradation, significantly affecting capture quality [11]. The resolution of the images obtained, which can be understood as the degree of closeness between objects within an image that are distinguishable from each other [12], is a significant concern. The higher the resolution, the more detailed the information representation. The common approach to address this in literature involves incrementing the spatial resolution, i.e., increasing the number of pixels per unit area [10, 11, 13]. However, the resolution level of image acquisition hardware often restricts the number of pixels available to represent an image. This is due to the high cost and physical limitations of high-precision optical devices and sensors [10, 11]. Therefore, current efforts aim at enhancing the resolution of captured images through post-acquisition software methods.

2.1 Super-resolution methods

Super-resolution has become a hotspot of investigation in the last decades. It comprises a group of techniques that seek to produce higher-resolution images from one or more lower-resolution images [11]. This innovative approach serves as a more recent alternative to the classic interpolation algorithms. Although both are used to increase the size of a single image, the difference between super-resolution and classic methods lies in the ability to recover lost high-frequency components. The interpolation techniques have difficulty recreating this type of information, so they have not considered super-resolution methods [10].

These super-resolution techniques are divided into reconstruction-based and learning-based methods [11, 13]. The reconstruction-based methods, in turn, are divided into two major approaches: the frequency domain approach and the spatial domain approach. Early super-resolution techniques were frequency domain based, using the Fourier transform to find the representation of an image in the frequency domain and then eliminate the spectral aliasing [13–15]. Meanwhile, spatial domain methods can create a higher-resolution image of a scene based on a set of lower-resolution images of that same scene. Each of these images must be slightly different from one another to add new information. There are also combined methods [16]. These could represent the scene from different angles or be taken at different time lapses. The higher-resolution image reconstruction is possible by combining the information obtained from each lower-resolution image [10]. Reconstruction-based methods usually obtain moderate resolution gains because these need to estimate the effect of displacement, blur, and rotation. It is an arduous task with many complications [15].

Learning-based methods are the techniques that lead the current investigation on super-resolution. This approach consists of training a model with a large amount of information; therefore, it can learn the spatial structural relationship between the higher-resolution and lower-resolution images [13]. These methods successfully recreate high-frequency components without increasing the number of images needed [11, 13]. Nowadays, multiple

surveillance proposals of reconstruction-based [17, 18] and learning-based methods exist [19–21].

Recent literature has emphasized the potential of deep learning methodologies in addressing the complexities associated with super-resolution reconstruction. The proposal [22] introduced an innovative deep-learning model rooted in convolutional neural networks to generate high-resolution spatial representations of surface albedo from coarse-resolution remote sensing-based data. While their study primarily focuses on the downscaling of surface albedo for bifacial solar photovoltaic panels, the underlying principles, and methodologies offer valuable insights into the application of deep learning for image super-resolution in surveillance contexts.

Complementing these findings, [23] demonstrated the effectiveness of an integrated approach combining Efficient Sub-Pixel Convolutional Neural Network (ESPCN) and Convolutional Neural Network (CNN) for enhancing super-low-resolution facial images. Their method showcased notable improvements in image resolution and recognition accuracy, underscoring the efficacy of deep learning techniques in the realm of facial recognition within surveillance systems.

The project [24] compared the capabilities of the newly introduced Xilinx Versal ACAP platform against the conventional MPSoC FPGA, particularly for Deep Learning applications. Using the Vitis AI inference framework, they explored various convolutional and fully-connected models. A custom architecture for an image super-resolution model (ESPCN) on Versal ACAP yielded a 4.5x latency improvement over the standard Vitis AI framework implementation, highlighting the potential advantages of this new platform in addition to the now classical GPU implementations.

2.2 Aplied method and variations

The foundation of our super-resolution processing is built upon the Efficient Sub-pixel Convolutional Network (ESPCN) [25]. ESPCN is a convolutional network designed to perform interpolation on individual images or videos, with a central design goal being the minimization of runtime processing. This is achieved while also ensuring low consumption of hardware resources, facilitated primarily through a novel placement of the sub-pixel convolution layer within its architecture.

Original ESPCN Approach in YCbCr Space

Traditionally, the ESPCN implementation operates specifically on the YCbCr color space, more precisely on its luma (Y) component. The primary reasoning behind this design choice is that human vision is more sensitive to changes and details in luminance (brightness) than in chrominance (color). Thus, by enhancing the resolution of the Y channel, which captures luminance, a significant perceived improvement in image quality is attained. The chrominance channels, Cb and Cr, are then typically upscaled using standard interpolation methods, considering that the human eye is less sensitive to high-frequency changes in color.

Our RGB-centric approach

In our study, we introduce a variation to the conventional method. Instead of relying solely on the YCbCr color space and applying super-resolution exclusively to the Y channel, we propose applying ESPCN directly to each channel in the RGB (Red, Green, Blue) additive color model. This decision allows the model to predict and enhance details across all color components of the image. By doing so, we aim to understand the potential outcomes and benefits of extending super-resolution processing across the entirety of the image information, which may provide richer details, especially in contexts where color details are crucial.

While the traditional ESPCN approach in the YCbCr space is optimized considering the nuances of human vision, our ESPCN_RGB variation seeks a broader perspective. The essence of our variation is not solely rooted in human perceptual experience but also anchored in the objective metrics we deploy for evaluation. By applying the super-resolution process across all RGB channels, we hope to achieve a comprehensive enhancement that can be robustly evaluated through metrics such as PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index), FID (Fréchet Inception Distance), Blur, and Face Distance [26].

These metrics provide a holistic view of the image quality, capturing both perceptual and mathematical aspects of the super-resolution output. This dual emphasis -on human perception and objective quantification-ensures that our RGB-centric method caters to a wider range of real-world surveillance and recognition scenarios, where color fidelity, sharpness, and accurate face representations to be used in face detection-identification scenarios are paramount.

3. Implementation

3.1 Dataset and training

The dataset was created from 22000 images downloaded from the website [27]. These face pictures do not correspond to real-life human beings, and the faces are generated by a generative adversarial network (GAN) based on [28] work to avoid ethical dilemmas [29].

The ground truth and the decimated images are the labels that compose the dataset. The ground truth is the resolution ideal; the images within this group have 1,024x1,024 resolution. Meanwhile, the decimated images are the homologous set, where each image from the ground truth diminishes its resolution. The decimation process is carefully made to avoid spectral aliasing phenomena and thus evade future adverse effects on the training.

The dataset is divided into two subsets: the training and validation sets. The distribution between both sets is 75% (16500 Images) and 25% (5500 Images) of the total amount of images, respectively. Our ESPCN model was trained using the Pytorch deep learning library. We utilized the Adam optimizer, paired with a mini-batch strategy, to refine our model. The primary cost function was Mean Squared Error (MSE), and the learning rate was set to 1e-3. This approach allowed for the nuanced adjustment of the model's weights with each iteration, incrementally improving the quality of the super-resolved images.

Minor modifications were made to the original ESPCN implementation to enhance training efficiency and fully exploit our available hardware resources. The batch size was substantially increased to 1000 to maximize the computational power of an 8-GPU cluster equipped with GTX1080Ti graphics cards that we used in our training and model evaluation. Default values from the Pytorch library were used for other hyperparameters, with the learning rate maintained at 1e-3, and the Adam optimizer's Beta1, Beta2, and Epsilon parameters were set to 0.9, 0.999, and 1e-8, respectively. The effectiveness of these parameter settings is evident in Figure 1, which illustrates the steady improvement of MSE, PSNR, and SSIM metrics over successive training epochs; in all training process, we attain similar results in the convergence of the algorithm around 400 epochs.

3.2 Hardware

Once the model was appropriately trained, it was implemented on a Jetson TX2. This board is an embedded system manufactured by Nvidia, and designed for artificial intelligence applications. The TX2 has a camera module of 5 Mega Pixels, allowing the Jetson TX2 capabilities in artificial vision.

3.3 Software

The ESPCN is implemented by a piece of software responsible for coordinating and executing the duties of the embedded system. Figure 2 contains a flowchart that explains the functioning of the main program. Firstly, it is necessary to initialize essential objects like the ESPCN itself and a frontal face detector Haar cascade [30], then the main program enters the main loop, where the camera captures an initial image; this image is treated to search for any human face in it by the Haar cascade classifier. In case this search has a positive result, there is a human face in the image; the main program must estimate a region of interest (ROI) that preserves the face found correctly. This ROI is used as input for the model, so the ESPCN calculates a higher-resolution version. It is essential to clarify that

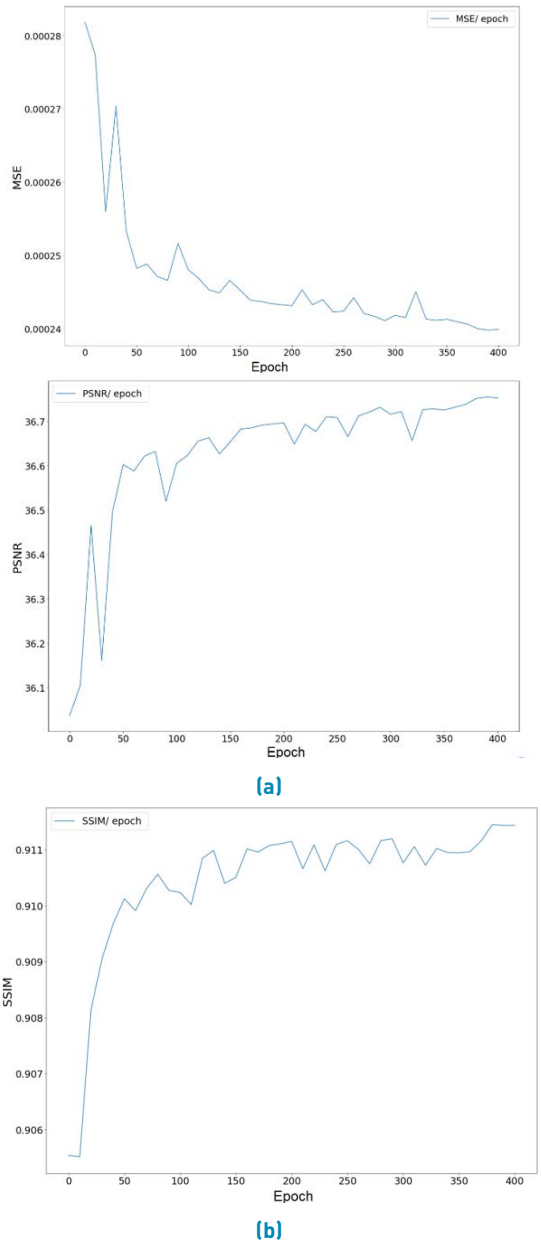


Figure 1 Training Performance Metrics for ESPCN: The progression of Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM) across epochs during the evaluation phase. From Top to Bottom

there are pre-process and post-process tasks related to the model operation. Finally, the output image is visualized on a monitor connected to the embedded system. If the face detector has a negative result, no human face is in the image; the initial image does not require any other treatment and is visualized on the monitor immediately. The main loop will run while the process supporting the main loop is running.

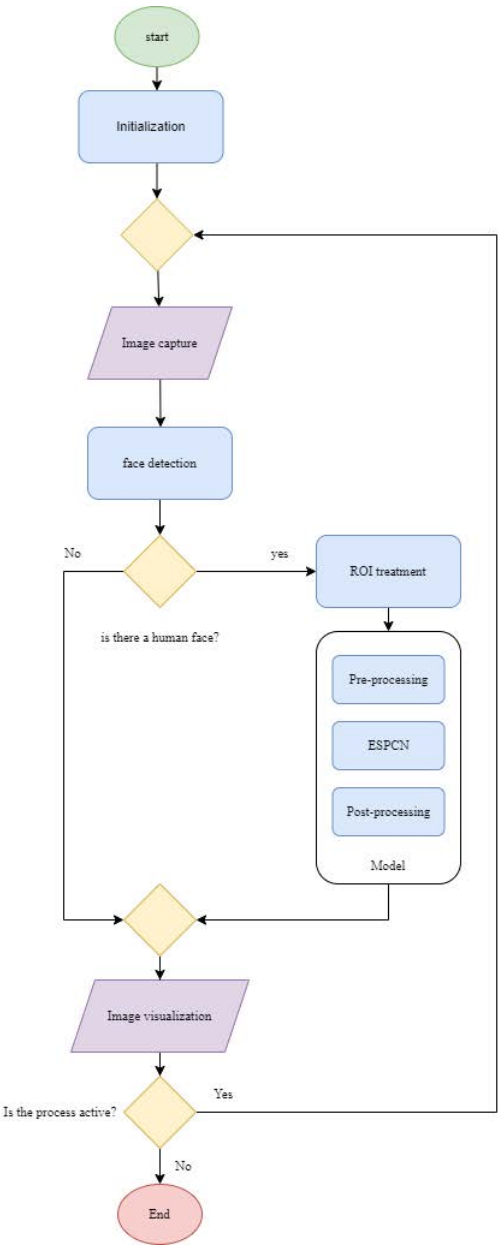


Figure 2 Flowchart of the main program

4. Experimentation

In the landscape of video surveillance, particularly where face identification is pivotal, the utility of a model is not just in its capability to enhance resolution but also in its ability to be deployed in real-time on embedded systems. The core objective of these experiments with ESPCN is to investigate its effectiveness in super-resolving facial images, making them suitable for video surveillance applications when deployed on embedded platforms like the Jetson TX2.

To demonstrate the potency of ESPCN, three distinct models were trained using images whose resolution was decreased by sub-sampling factors of 2, 4, and 8 from the original resolution (1,024x1,024). This translates to images with resolutions of 512x512, 256x256, and 128x128, respectively.

4.1 Model testing plan

The model is evaluated based on five metrics: peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), Frechet inception distance (FID), blur, and face distance. Each of these metrics measures different aspects of the result images from the ESPCN.

The classic interpolation methods: Nearest neighbor (NN), bilinear, bicubic, and Lanczos are used as performing references. These methods were chosen due to their extensive use in practical image processing applications, primarily because of their swift computational efficiency. Through a visual evaluation, we aim to illustrate the capabilities of ESPCN by juxtaposing its outputs with those derived from the classical methods on specific portions of a selected validation set image.

PSNR

Commonly employed in super-resolution research [25], the peak signal-to-noise ratio provides an invaluable perspective on image fidelity. As a logarithmic metric, PSNR gauges the pixel-wise similarity between two images. By comparing the ground truth to the images reconstructed using ESPCN, we can gauge how well the network is preserving original details during the upscaling process, a vital aspect in video surveillance for accurate face identification.

SSIM

The structural similarity index measure, akin to PSNR, evaluates image quality by comparing images on a pixel-by-pixel basis. Despite its popularity in super-resolution research, SSIM's sensitivity sets it apart.

Given its narrow metric range (0-1), it is less susceptible to variances than PSNR, making it a more robust metric for understanding structural deviations in facial features.

FID

The Frechet inception distance compares the distribution of two different groups of images, in this case, the ground truth and the reconstructed images.
Smaller FID is related to more similar distributions between the two sets of images. The method that increases image resolution is expected to not drastically affect the image distribution. This metric is very interesting since it is also sensitive to noise and other forms of disturbance on the images. A lower FID indicates that the reconstructed images maintain a distribution closely aligned with the ground truth, hinting at minimal disturbances or noise, an essential characteristic for reliable surveillance systems.

Blur

The blur is a measure of the level of "smoothing" in an image. High levels of blur are strongly related to a lack of high-frequency information. The range of this metric is between 0 and 1, where the higher the measure, the greater the blur of the image. For accurate face recognition, maintaining a low blur metric is crucial.

Face distance

Within the domain of video surveillance, accurate representation and preservation of facial features following super-resolution processes bear significant implications. Surveillance systems rely heavily on the clarity and fidelity of images to make precise identifications, and any discrepancies in enhanced images might lead to erroneous recognitions with potential consequences.

The face distance metric, elucidated in reference [26], is devised to gauge the proximity between the original and the reconstructed facial images. A value nearing zero indicates a high degree of congruence between the enhanced and original images, ensuring the reliability of the super-resolved image for identification purposes. Conversely, values approaching one signal notable discrepancies, which could challenge the utility of the super-resolved image in surveillance contexts.

The imperative nature of this metric accentuates the potential contributions of ESPCN to the realm of surveillance, highlighting the significance of consistent facial feature representation.

4.2 Embedded system implementation

The NVIDIA Jetson TX2 was selected for its optimized architecture for neural network computations, primarily driven by its GPU cores. This device, coupled with NVIDIA's CUDA technology, ensures not only efficient parallel processing but also offers adaptability across different platforms, facilitating the potential scalability of our solution.

The choice of resolutions, specifically 640x480, 320x240, and 160x120, aligns with the prevalent VGA standard used in video surveillance systems. These resolutions were chosen to assess the model's capability to enhance commonly encountered surveillance footage to a more detailed 1,280x960 resolution. In the context of surveillance, increased resolution aids in enhancing discernibility, thereby making details more perceptible.

Embedded systems, including the Jetson TX2, come with their set of challenges, such as memory constraints, power consumption considerations, and heat dissipation issues. These challenges were addressed through model optimization and leveraging the efficiencies inherent to the TX2. The resulting system exhibits commendable energy efficiency, which is essential for sustained surveillance operations.

In conclusion, the successful deployment of the ESPCN model on the Jetson TX2 highlights its potential for real-time surveillance applications, while also underscoring the broader applicability of sophisticated neural network models on embedded platforms.

5. Discussion

In general, with a few exceptions, the ESPCN outperformed other interpolation methods in all evaluated metrics. PSNR and SSIM data, presented in Table 1 and Table 2, respectively, demonstrate that ESPCN produced images that were closest to the ground truth. Additionally, the difference between the ground truth and ESPCN-generated images was minimal compared to most other interpolation methods, as evidenced by the FID results shown in Table 3, particularly when the sub-sampling factor is 2. The only exception was the nearest neighbor method, which was the only method capable of outperforming the ESPCN results with a sub-sampling factor of 8.

Table 4 shows that the blurring phenomenon is less noticeable in the ESPCN images and becomes more pronounced as the sub-sampling factor increases. The nearest neighbor method is again an exception, but this time due to its pixelated finish, which is not ideal, despite

not being blurred.

Finally, the face distance data reveals the ESPCN's potential as a support tool for an identity verification algorithm. Table 5 confirms that the ESPCN is more effective than classical interpolation methods in recreating key facial features for identity verification, particularly at higher sub-sampling factors. This is particularly useful when reconstructing super-resolution images from small images or when a significant increase in resolution is necessary.

Figure 3 shows a comparison of the whole image using all four interpolation methods and the two ESPCN methods. On the other hand, Figure 4, Figure 5, and Figure 6 focus on specific areas of the face: forehead, eye, crow's feet, wrinkles, and beard. These visual comparisons help to intuitively understand the resolution gain difference between the interpolation methods and the ESPCN. As the sub-sampling factor increases, the difference becomes more evident. However, the difference is not as obvious at first sight using the visual comparison, despite the larger performance gap according to the PSNR and SSIM metrics, especially when the factor is 2.

Figure 7 shows the lower-resolution images captured by the camera, while Figure 8 and Figure 9 show the results of the ESPCN implementation in the embedded system. These images are the super-resolution reconstruction of lower-resolution images like Figure 7. As the sub-sampling factor increases, the images quickly degrade, which is consistent with the metrics results. Higher super-resolution reconstruction requires more information from the lower-resolution image and demands higher processing power.

Comparing the ESPCN that reconstructs the RGB color channels with the ESPCN that just reconstructs the Y luma component, it is visible that the results are very similar, in both the metrics and the visual comparison images. The RGB results are often slightly superior, but the gap between them is minuscule

5.1 Model testing plan results

Our experiments demonstrate the superior performance of the Efficient Sub-Pixel Convolutional Network (ESPCN) model over classical interpolation methods across multiple metrics, as can be seen in Tables 1 to 5. For both the PSNR and SSIM results, the ESPCN model, in both Y and RGB modes, consistently outperforms Nearest Neighbor, Bilinear, Bicubic, and Lanczos interpolation methods across scale factors of 2, 4, and 8. This suggests that ESPCN provides better image quality with higher similarity to the original high-resolution image. Similarly, for the FID metric, ESPCN shows significantly lower values than the other methods, implying that the distribution of the generated images is closer to the distribution of real images. In the Blur metric results, ESPCN presents lower mean values, which indicates reduced blurriness in the produced images. Finally, for the Face Distance metric, ESPCN achieves lower mean values than most other methods, suggesting its potential superiority for facial image processing applications. This collection of results underlines the potential of the ESPCN model as a robust solution for enhancing image resolution, particularly in the context of video surveillance.

5.2 Embedded system implementation results

Computational complexity

The Efficient Sub-Pixel Convolutional Neural Network (ESPCN) uses a learnable upscaling process at the end of the network. The upscale factor in ESPCN is achieved by the PixelShuffle operation, which rearranges the elements in the tensor from the depth dimension to the spatial dimensions.

When the scale factor increases, the number of MACs decreases because, as can be seen in Table 6, in ESPCN, the images are upscaled in the feature map space rather than in the pixel space. The convolution operations are performed in low-resolution space, which requires fewer MACs as the scale factor increases. This is because the amount of detail the model needs to generate decreases as the scale factor increases, and the overall workload of the network decreases with the increase in the scale factor, which leads to the reduction in MACs. This is one of the benefits of using ESPCN over other super-resolution methods, as it can perform the upscaling operation more efficiently, especially for large upscale factors.

Our hardware implementation yielded similar average execution times for both the ESPCN_Y and ESPCN_RGB models, roughly around 5 seconds more than 5 times more on average than the worst case with Lanczos. This similarity may be attributed to specific characteristics of the compiler and the architecture of the hardware on which the models were run - the Jetson TX2 board. In this case, we did not utilize the GPU of the board to optimize the algorithm, just as we had not optimized the classic methods such as Nearest Neighbor, Bilinear, Bicubic, or Lanczos for the processor. Regardless, all models produced an output image with a resolution of 1280x960. The computational complexity of interpolation methods such as Nearest Neighbor, Bilinear, Bicubic, and Lanczos can be broadly estimated in terms of Multiply-Accumulate operations (MACs) in terms of a scale factor sf and a number of pixels n. Nearest Neighbor, performing one

Table 1 PSNR metric results

		N.N	Bilinear	Bicubic	Lanczos	ESPCN_Y	ESPCN_RGB
Factor 2	Mean	37.16	38.21	40.07	40.38	40.70	41.39
	Stand.dev.	1.84	2.11	2.20	2.23	1.73	1.96
Factor 4	Mean	32.36	33.50	34.34	34.43	34.97	35.15
	Stand.dev.	1.68	1.92	2.03	2.05	1.95	2.02
Factor 8	Mean	29.10	30.27	30.97	31.05	31.43	31.60
	Stand.dev.	1.50	1.70	1.79	1.80	1.76	1.82

Table 2 SSIM metric results

		N.N	Bilinear	Bicubic	Lanczos	ESPCN_Y	ESPCN_RGB
Factor 2	Mean	0.954	0.955	0.969	0.971	0.975	0.977
	Stand.dev.	0.010	0.012	0.008	0.008	0.006	0.006
Factor 4	Mean	0.863	0.880	0.894	0.895	0.908	0.909
	Stand.dev.	0.027	0.029	0.026	0.026	0.022	0.022
Factor 8	Mean	0.756	0.807	0.816	0.817	0.827	0.829
	Stand.dev.	0.043	0.042	0.040	0.040	0.038	0.038

Table 3 FID metric results

	N.N	Bilinear	Bicubic	Lanczos	ESPCN_Y	ESPCN_RGB
Factor 2	450.48	2,523.69	621.10	442.13	51.53	33.96
Factor 4	2,647.25	9,295.63	5,495.17	5,193.59	2,372.12	2,121.91
Factor 8	3,600.99	19,123.293	15,120.29	14,413.11	8,893.36	8,443.40

Table 4 Blur metric results

		N.N	Bilinear	Bicubic	Lanczos	ESPCN_Y	ESPCN_RGB
Factor 2	Mean	0.433	0.512	0.450	0.438	0.408	0.408
	Stand.dev.	0.043	0.043	0.042	0.042	0.042	0.042
Factor 4	Mean	0.365	0.664	0.613	0.594	0.523	0.522
	Stand.dev.	0.022	0.034	0.035	0.038	0.039	0.039
Factor 8	Mean	0.182	0.784	0.788	0.808	0.726	0.729
	Stand.dev.	5.78e-16	0.018	0.014	0.016	0.022	0.021

Table 5 Face distance metric results

		N.N	Bilinear	Bicubic	Lanczos	ESPCN_Y	ESPCN_RGB
Factor 2	Mean	0.0346	0.0338	0.0331	0.0333	0.0336	0.0337
	Stand.dev.	0.0177	0.0173	0.0176	0.0178	0.0183	0.0176
Factor 4	Mean	0.0411	0.0457	0.0347	0.0358	0.0348	0.0346
	Stand.dev.	0.0164	0.0148	0.0172	0.0173	0.0176	0.0176
Factor 8	Mean	0.0843	0.1090	0.0694	0.0625	0.0498	0.0495
	Stand.dev.	0.0160	0.0176	0.0160	0.0154	0.0168	0.0166

Table 6 Variation of ESPCN in the number of parameters and MACs vs. the scale factor

Scale_factor	Params	MACs
2	21284	504365056000
4	24752	66584576000
8	38624	10092544000

operation per output pixel, has complexity proportional to n * sf * sf. Bilinear, operating on a 2x2 neighborhood, multiplies this by four, leading to 4 * n * sf * sf. Bicubic, considering a 4x4 pixel grid, increases the complexity to 16 * n * sf * sf. Lastly, Lanczos typically operates on a larger neighborhood (e.g., 6x6 for Lanczos3), giving a complexity of 36 * n * sf * sf. The worst case is Lanczos with n=1280*960 and sf=8, the MACs will be 2831155200, which

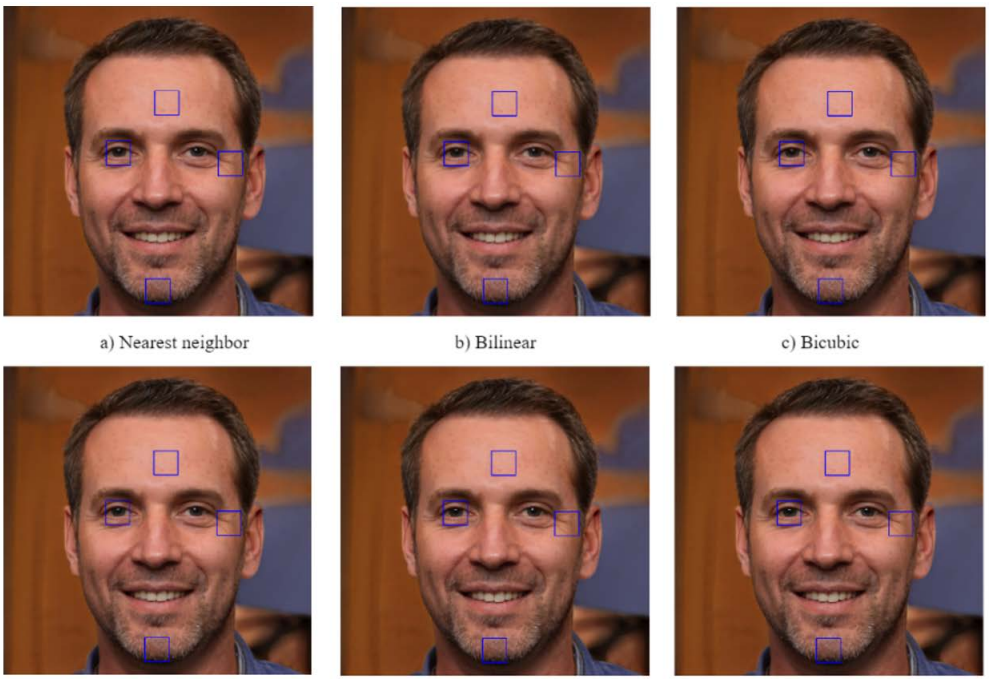


Figure 3 Factor 2 image comparison results

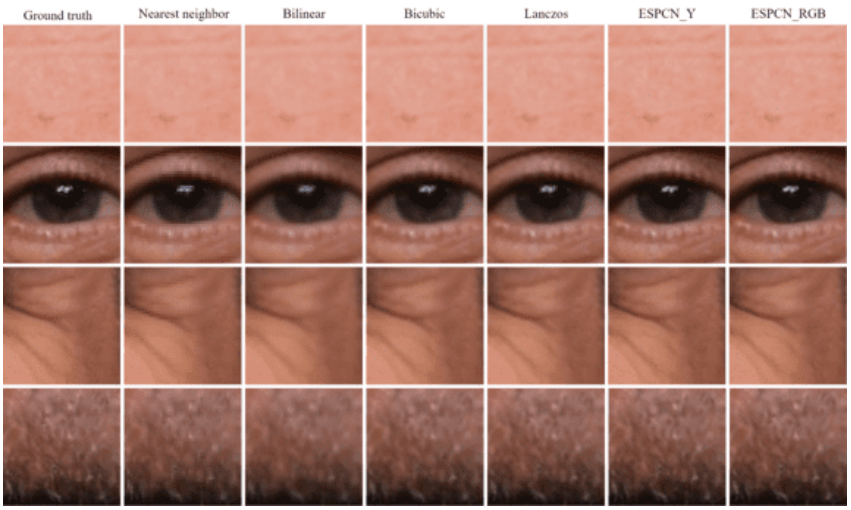


Figure 4 Factor 2 area image comparison results

is 3.6 less than the ESPCN_Y with the same resolution.

6. Conclusions

Considering the results, this study affirms the potential of implementing learning-based methods, particularly the Efficient Sub-pixel Convolutional Network (ESPCN), in video surveillance applications. The enhancements provided by the ESPCN, as demonstrated in our data, position it as a promising solution for improving image resolution in surveillance contexts, particularly frontal-face identification.

Importantly, though this paper does not explore specific training costs such as time and energy, the relative efficiency of ESPCN is demonstrated through the decreasing requirement for Multiply-Accumulate Operations (MACs) as the scale factor increases, as shown in our presented results. This efficiency allows ESPCN to outperform classical interpolation methods by reducing the reliance on high-precision, often expensive optical devices, and camera sensors.

Our results from the face distance metric indicate

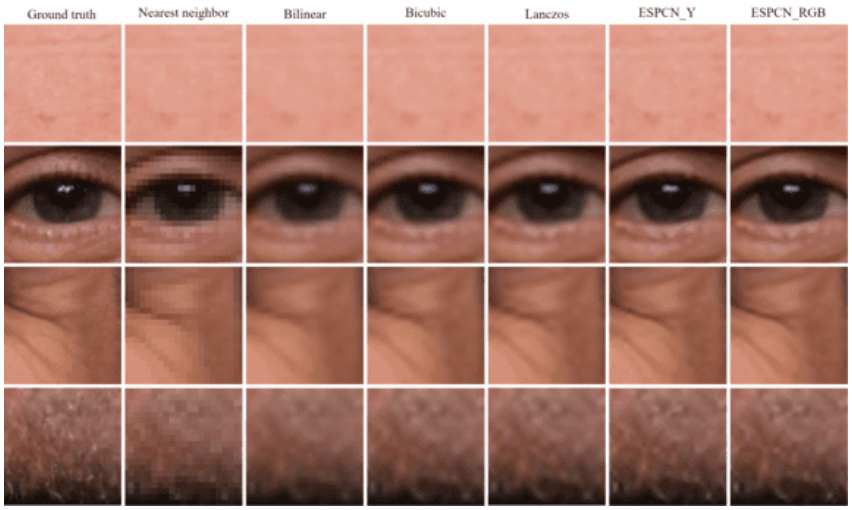


Figure 5 Factor 4 area image comparison results

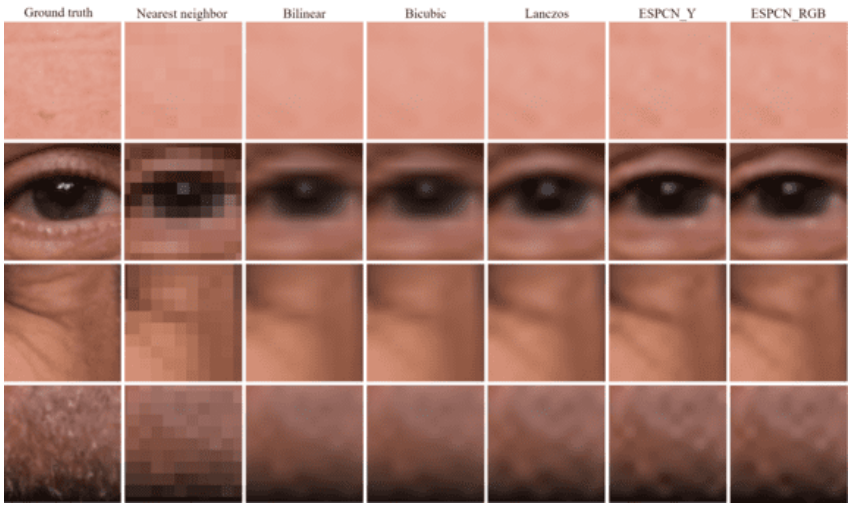


Figure 6 Factor 8 area image comparison results

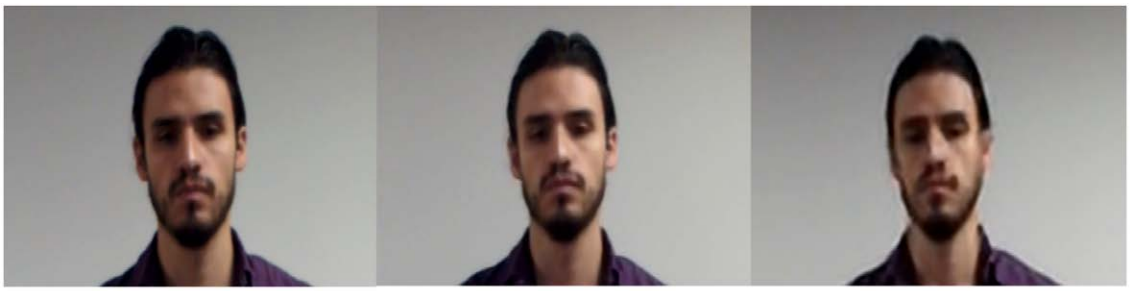
Figure 7 Initial images captured by the camera with 640x480 (a), 320x240 (b), and 160x120 (c) resolution

the potential for ESPCN to bolster other forms of face image processing algorithms. However, more research is necessary to fully understand the extent of the model's potential, its limitations, and how it can be fine-tuned for optimal results.

The results suggest that training based on artificial face images is an excellent approach for building a dataset with a large number of images, which allows us to achieve high-performance results. More importantly, it avoids

a) Factor 2 b) Factor 4 c) Factor 8

Figure 8 Images generated by the RGB model trained with 2 (a), 4 (b), and 8 (c) factors



a) Factor 2 b) Factor 4 c) Factor 8

Figure 9 Images generated by the YCbCr model trained with 2 (a), 4 (b), and 8 (c) factors

violating the intellectual property of anyone. Each person has the right to control their own image and personal information.

There are several factors to consider in further research to transition from our findings to a fully realized surveillance application prototype. These include operation range, and operation time in a particular hardware implementation, behavior under disturbances such as noise sources or luminous contamination, and face image perspective. For this paper, we utilized a frontal face perspective, due to its preference in surveillance applications. Still, future iterations of the model should incorporate training to recognize face images from various angles and perspectives.

As we navigate an ever-evolving world, the technification of surveillance continues to shape our lives. It gives governments and communities more control and capacity to preemptively detect threats, reduce risks, and protect individuals and societies. With time, we anticipate that advancements such as the ones proposed in this study will contribute significantly to these efforts.

In conclusion, the role of super-resolution represented by ESPCN is extremely applicable in modern video surveillance systems. Video surveillance applications continuously collect large amounts of redundant data, but specific moments of interest caused, for example, by movement, boundary violations, or in our case study, facial recognition, require higher levels of image sharpness. Using ESPCN at such critical moments ensures higher-resolution images to be saved, allowing for more detailed scrutiny. This not only increases the fidelity of stored images, but also strengthens the overall goal of increasing security measures and supporting forensic analysis with unprecedented accuracy.

7. Declaration of competing interest

We declare that we have no significant competing interests, including financial or non-financial, professional, or personal, interfering with the full and objective presentation of the work described in this manuscript.

8. Acknowledgments

The author(s) received no financial support for the research, authorship, and/or publication of this article.

9. Author contributions

SA. G. Collected the data, developed the training, executed the test plan, and implemented the model in the embedded

system. F. C. designed the initial data collection methodology and the testing plan. Both authors participate in the writing process of this document.

10. Data availability statement

The code associated with the paper and related information is available in the GitHub repository: gomezan/SRrostros; meanwhile, the original dataset created to perform this project validation and training is available on the OSF project page: https://osf.io/g4v5e/. Both can be accessed and used under the conditions of the third version of the GNU general public license.

References

[1] Z. Bojkovic and A. Samcovic, "Face detection approach in neural network based method for video surveillance," in *2006 8th Seminar on Neural Network Applications in Electrical Engineering*, Belgrade, Serbia, 2006. [Online]. Available: https://doi.org/10.1109/NEUREL.2006.341172

[2] H. Qezavati, B. Majidi, and M. Manzuri, "Partially covered face detection in presence of headscarf for surveillance applications," in *4th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, Tehran, Iran, 2019. [Online]. Available: https://doi.org/10.1109/PRIA.2019.8786004

[3] J. Harikrishnan, A. Sudarsan, A. Sadashiv, and R. Ajai, "Vision-face recognition attendance monitoring system for surveillance using deep learning technology and computer vision," in *International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN),*, Vellore, India, 2019. [Online]. Available: https://doi.org/10.1109/ViTECoN.2019.8899418

[4] Y. Wang, T. Bao, C. Ding, and M. Zhu, "Face recognition in real-world surveillance videos with deep learning method," in *2nd International Conference on Image, Vision and Computing (ICIVC)*, Chengdú, China, 2017. [Online]. Available: https://doi.org/10.1109/ICIVC.2017.7984553

[5] Q. Zhao and S. Wang, "Real-time face tracking in surveillance videos on chips for valuable face capturing," in *International Conference on Artificial Intelligence and Computer Engineering (ICAICE),*, Beijing, China, 2020. [Online]. Available: https://doi.org/10.1109/ICAICE51518.2020.00060

[6] A. Makhfoudi, S. Almaadeed, A. Bouridane, G. Sexton, and R. Jiang, "Visualization of faces from surveillance videos via face hallucination," in *International Conference on Control, Decision and Information Technologies (CoDIT)*, Metz, France, 2014. [Online]. Available: https://doi.org/10.1109/CoDIT.2014.6996982

[7] Z. Chen, Q. He, W. Pang, and Y. Li, "Frontal face generation from multiple pose-variant faces with cgan in real-world surveillance scene," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8462648

[8] A. Khan, M. Khan, F. Obaid, S. Jadoon, M. Khan, and M. Sikandar, "A novel multi-frame super resolution algorithm for surveillance camera image reconstruction," in *First International Conference on Anti-Cybercrime (ICACC)*, Riyadh, Saudi Arabia, 2015. [Online]. Available: https://doi.org/10.1109/Anti-Cybercrime.2015.7351950

[9] F. Mokhayeri, E. Granger, and G. Bilodeau, "Synthetic face generation under various operational conditions in video surveillance," in *IEEE International Conference on Image Processing (ICIP)*, Quebec City, QC, Canada, 2015. [Online]. Available: https://doi.org/10.1109/ICIP.2015.7351567

[10] S. Park, M. Park, and M. Kang, "Super-resolution image reconstruction: a technical overview," in *presented at IEEE Aerospace Conference*, 2003. [Online]. Available: https://doi.org/10.1109/MSP.2003.1203207

[11] L. Ziwei, W. Chengdong, C. Dongyue, Q. Yuanchen, and W. Chunping, "Overview on image super resolution reconstruction," in *26th Chinese Control and Decision Conference (2014 CCDC)*, Changsha, China, 2014. [Online]. Available: https://doi.org/10.1109/CCDC.2014.6852498

[12] IEEE, "Ieee standard computer dictionary: A compilation of ieee standard computer glossaries," in *presented at IEEE Std 610*, 1991. [Online]. Available: https://doi.org/10.1109/IEEESTD.1991.106963

[13] P. Shamsolmoali, M. E. Celebi, and R. Wang, "Deep learning approaches for real-time image super-resolution," *Neural Computing and Applications*, vol. 32, Jul. 15, 2020. [Online]. Available: https://doi.org/10.1007/s00521-020-05176-z

[14] X. Niu, "An overview of image super-resolution reconstruction algorithm," in *11th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, China, 2018. [Online]. Available: https://doi.org/10.1109/ISCID.2018.10105

[15] T. Cui, L. Tang, J. Nan, and Z. Li, "Space target super-resolution based on low-complex convolutional networks," in *IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*, Chongqing, China, 2019. [Online]. Available: https://doi.org/10.1109/ICSIDP47821.2019.9173516

[16] H. Huang, X. Fan, C. Qi, and S. Zhu, "A learning-based pocs algorithm for face image super-resolution reconstruction," in *International Conference on Machine Learning and Cybernetics*, Guangzhou, China, 2005. [Online]. Available: https://doi.org/10.1109/ICMLC.2005.1527837

[17] X. Yang, W. Wu, K. Liu, P. W. Kim, A. K. Sangaiah, and *et al.*, "Long-distance object recognition with image super resolution: A comparative study," in *presented at IEEE Access*, 2018. [Online]. Available: https://doi.org/10.1109/ACCESS.2018.2799861

[18] X. Hu, X. Liu, Z. Wang, X. Li, W. Peng, and G. Cheng, "Rtsrgan: Real-time super-resolution generative adversarial networks," in *Seventh International Conference on Advanced Cloud and Big Data (CBD)*, Suzhou, China, 2019. [Online]. Available: https://doi.org/10.1109/CBD.2019.00064

[19] Y. Lee, J. Yun, Y. Hong, J. Lee, and M. Jeon, "Accurate license plate recognition and super-resolution using a generative adversarial networks on traffic surveillance video," in *IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*, JeJu, Korea (South), 2018. [Online]. Available: https://doi.org/10.1109/ICCE-ASIA.2018.8552121

[20] S. Kim and P. Bindu, "Realizing real-time deep learning-based super-resolution applications on integrated gpus," in *15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Anaheim, CA, USA, 2016. [Online]. Available: https://doi.org/10.1109/ICMLA.2016.0122

[21] W. Shi, J. Caballero, F. Huszár, J. Totz, A. Aitken, R. Bishop, and *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Shi_Real-Time_Single_Image_CVPR_2016_paper.html

[22] AI Free. This person does not exist. Accessed Oct. 18, 2021. [Online]. Available: https://thispersondoesnotexist.com/

[23] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," *Arxiv*, vol. 2, Mar. 23, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.1912.04958

[24] *Cascade Classifier*, Open Source Computer Vision 3.4.20, 2022.

[25] Face recognition. GitHub. Accessed Nov. 13, 2013. [Online]. Available: https://github.com/ageitgey/face_recognition

[26] J. E. Sanabria-Moyano, M. D. P. Roa-Avella, and O. I. Lee-Pérez, "Tecnología de reconocimiento facial y sus riesgos en los derechos humanos," *Revista Criminalidad*, vol. 64, no. 3, 2022. [Online]. Available: https://doi.org/10.47741/17943108.366

[27] M. Márquez, C. A. Vargas, and H. Arguello, "Compact spatio-spectral algorithm for single image super-resolution in hyperspectral

imaging," *Ingeniería e Investigación*, vol. 36, no. 3, Sep-Dec. 2016 2016. [Online]. Available: https://doi.org/10.15446/ing.investig.v36n3.54267

[28] S. Karalasingham, R. C. Deo, D. Casillas-Pérez, N. Raj, and S. Salcedo-Sanz, "Downscaling surface albedo to higher spatial resolutions with an image super-resolution approach and proba-v satellite images," in *IEEE Access*, 2023. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3236253

[29] M. A. Talab, S. Awang, and S. A. d. M. Najim, "Super-low resolution face recognition using integrated efficient sub-pixel convolutional neural network (espcn) and convolutional neural network (cnn)," in *IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, Selangor, Malaysia, 2019. [Online]. Available: https://doi.org/10.1109/I2CACIS.2019.8825083

[30] A. Leftheriotis, A. Tzomaka, D. Danopoulos, G. Lentaris, G. Theodoridis, and D. Soudris, "Evaluating versal acap and conventional fpga platforms for ai inference," in *12th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, Athens, Greece, 2023. [Online]. Available: https://doi.org/10.1109/MOCAST57943.2023.10176615