UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MIARFID Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

Departament de Sistemes informàtics i Computació
Universitat Politècnica de València

Super resolución facial basado en Deep Learning

TRABAJO FIN DE MÁSTER

Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

Autor: Jorge Alcañiz Villanueva

Tutor: Roberto Paredes Palacios
Juan Maroñas Molano

Curso 2020-2021

Resum

Amb l'auge de la capacitat computacional dels ordinadors en les últimes dècades, a poc a poc s'ha pogut permetre introduir l'ús d'intel·ligència artificial en el món digital, acabant fins i tot a ser un instrument útil en el dia a dia. Per exemple, els sistemes biomètrics que permeten la verificació facial de l'usuari és una cosa bastant accessible arribant a ser una eina que s'usa en una cosa tan comuna i que tenim tot com són els dispositius mòbils. No obstant això, aquest tipus d'utensilis no solen tindre un correcte funcionament quan es parla d'imatges a molt baixa resolució. Per això, una de les solucions que es proposen per a aquesta mena de casos, és la millora de qualitat mitjançant la inferència d'un model especialitzat en la súper resolució facial.

En aquest treball s'ha detallat l'estat de l'art de la súper resolució centrat en el rostre facial, s'ha buscat una base de dades pública que complisca amb les limitacions i s'ha dut a terme una sèrie d'experiments amb un model de súper resolució facial proposat que ens permeta millorar el rostre facial de les imatges a una qualitat superior.

Paraules clau: super resolució facial, aprenentatge profund, xarxes neuronals convolucionals, reconeixement facial, super resolució

Resumen

Con el auge de la capacidad computacional de los ordenadores en las últimas décadas, poco a poco se ha podido permitir introducir el uso de inteligencia artificial en el mundo digital, acabando incluso a ser un instrumento útil en el día a día. Por ejemplo, los sistemas biométricos que permiten la verificación facial del usuario es algo bastante accesible llegando a ser una herramienta que se usa en algo tan común y que tenemos todo como son los dispositivos móviles. Sin embargo, este tipo de utensilios no suelen tener un correcto funcionamiento cuando se habla de imágenes a muy baja resolución. Por ello, una de las soluciones que se proponen para este tipo de casos, es la mejora de calidad mediante la inferencia de un modelo especializado en la súper resolución facial.

En este trabajo se ha detallado el estado del arte de la súper resolución centrado en el rostro facial, se ha buscado una base de datos pública que cumpla con las limitaciones y se ha llevado a cabo una serie de experimentos con un modelo de súper resolución facial propuesto que nos permita mejorar el rostro facial de las imágenes a una calidad superior.

Palabras clave: super resolución facial, Aprendizaje profundo, redes neuronales convolucionales, reconocimiento facial, super resolución

Abstract

With the rise of the computational capacity of computers in recent decades, it has gradually been possible to introduce the use of artificial intelligence in the digital world, even becoming a useful tool in everyday life. For example, biometric systems that allow facial verification of the user is something quite accessible, becoming a tool that is used in something so common and that we have everything like mobile devices. However, this type of tools do not usually have a correct operation when talking about very low resolution images. Therefore, one of the solutions proposed for this type of cases is the improvement of quality through the inference of a model specialized in facial super resolution.

In this work we have detailed the state of the art of super resolution focused on the facial face, searched for a public database that meets the constraints and carried out a series of experiments with a proposed facial super resolution model that allows us to improve the facial face of the images to a higher quality.

Key words: Face super-resolution, Deep learning, convolutional neural networks, facial recognition, super-resolution

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

En las últimas décadas, debido al gran incremento en la capacidad de cómputo de los ordenadores, el aumento de la cantidad de información existente en internet y el avance de numerosas técnicas algorítmicas relacionadas con el aprendizaje a partir de datos, se ha podido llegar a obtener grandes avances en áreas como el procesamiento del lenguaje natural o visión por computador.

De hecho, gracias a ese aumento de la capacidad computacional mencionada previamente, el uso de la inteligencia artificial hoy en día es tan común que ha pasado a ser una herramienta muy necesaria y útil en bastantes contextos del ámbito digital. Implicado en distintas áreas correspondientes a la visión por computador, marco en el que se centrará el proyecto, aplicaciones como sistemas de reconocimiento o verificación facial para el uso restringido de zonas de acceso, detección e identificación de objetos o *tracking* de objetos para muchos otros casos de uso. Otras de las áreas en las que se suelen aplicar Inteligencia Artificial, podría ser el sistema de reconocimiento del habla, véase los asistentes de voz como "Cortana" el cual hoy en día es bastante frecuente su uso.

Asimismo, gracias a estos avances en visión por computador, el análisis de imágenes y vídeos centrados en el ser humano ha recibido una atención cada vez mayor tanto en el ámbito académico como en el industrial en todo el mundo. Dado que el rostro es un carácter clave del ser humano, el análisis facial asistido por máquinas se convierte en un tema popular en diversas aplicaciones. Un ejemplo relacionado al trabajo es la aplicación de la súper resolución facial (SR) en el campo de análisis de imágenes faciales, los fabricantes de dispositivos móviles están muy interesados en desarrollar sistemas tanto de hardware como de software para la recogida de rostros faciales.

En la primera parte de este capítulo se presentará cuál es la motivación de nuestro trabajo, a continuación se explicará cuáles son los objetivos, posteriormente se comentará la hipótesis a la que nos sostenemos para realizar este proyecto y finalmente se mencionará la estructura del trabajo.

1.1 Motivación

Aparte de las huellas dactilares, las imágenes faciales son la característica más importante dentro del campo de la biometría, ya que cada uno tiene un rostro facial único y ello permite la verificación de los individuos. Así pues, La súper resolución facial es una tarea importante que permite mejorar la resolución de la imagen del usuario en caso de tener una resolución baja y transformarlo a una de mayor resolución. Se puede aprovechar en casos, en los que el sistema reconoce un rostro facial pero es un rostro que ocupa unos pocos píxeles de la imagen, por lo tanto, si la calidad del rostro es mala, su verificación facial finalmente será errónea.

Además, dado que generalmente los sistemas biométricos han sido entrenados con imágenes de alta calidad para aplicaciones como la verificación del individuo, en caso de que al sistema se le proporcionase una imagen el cual difiera un poco de la calidad en la que haya sido entrenado, su inferencia sería probablemente errónea. Bajo esta premisa, uno de los motivos por los que se ha querido realizar este proyecto ha sido aprender a aplicar un sistema que permita mitigar el problema expuesto.

Para ello, en este proyecto se presentará una experimentación con "embeddings", que representarían las características faciales de la imagen, extraídos de un modelo pre-entrenado con un dataset centrado en reconocimiento facial, junto a un modelo de súper resolución que mejore la calidad de la imagen. Para ello, los experimentos que se aplicarán consistirá en buscar la distribución de pesos más óptima entre la información que provee el modelo pre-entrenado y el modelo de súper resolución, buscando un equilibrio que permita asemejarse a las imágenes originales y mejorar su *baseline*.

En cuanto a los motivos más personales, surge de las ganas y curiosidad de poner en práctica los conocimientos aprendidos a lo largo del máster en la implementación de un sistema basado en redes neuronales convolucionales y aplicando en un ámbito que no se ha llegado a dar con mucha profundidad en el máster.

1.2 Hipótesis

El proyecto está basado en la siguiente hipótesis:

- Cuanto menor es la calidad de la imagen, peor se observará el rostro facial en la imagen llegando incluso a ser imperceptible, por lo que se empleará una red neuronal adicional pre-entrenada usada en verificación de rostros faciales para tratar de preservar la identidad de la persona.

1.3 Objetivos

El objetivo principal de este trabajo consiste en la construcción de un sistema basado en aprendizaje automático que permita mejorar la imagen que se provee a muy baja resolución aplicando una serie de experimentaciones basados en la hipótesis de la sección 1.2.

Los subobjetivos son los siguientes:

- Encontrar un *dataset* que se asemeje a las especificaciones propuestas.
- Diseñar la arquitectura del modelo y aplicar las modificaciones convenientes para la función de pérdida múltiple.
- Analizar los resultados obtenidos, proponiendo mejoras y trabajos futuros.

1.4 Estructura de la memoria

El presente documento se estructura en un total de 5 capítulos. A continuación se detalla cada uno de ellos.

- En el **capítulo 1**, contiene una descripción inicial del contenido de la memoria, la motivación del proyecto y la estructuración del contenido de la memoria.
- En el **capítulo 2**, se habla del estado del arte de súper resolución facial. Se comenta algunos de las técnicas más importantes que se han utilizado para este campo.
- En el **capítulo 3**, se expondrán los recursos utilizados, tanto los referidos a las imágenes usadas para nuestros experimentos como los recursos computacionales utilizados para realizar dichos experimentos. Posteriormente, se comentará la propuesta del modelo junto con la selección de base de datos para este trabajo y sus respectivas métricas.
- En el **capítulo 4**, se expondrán los resultados de la experimentación realizada y se profundizará en el desarrollo de la solución realizando un análisis.
- En el **capítulo 5**, se comentarán las conclusiones obtenidas, se detallará la relación del presente trabajo con los estudios cursados en el máster y finalmente se expondrán los futuros trabajos que se pueden realizar en base a este proyecto.

CAPÍTULO 2
Estado del arte

El objetivo de este capítulo consiste en realizar una revisión del estado del arte en la utilización de tecnologías de súper resolución facial. Se comenzará con una introducción de súper resolución y posteriormente se comentarán las distintas ramas de investigación que se han ido formando dentro del campo.

La idea básica que se presenta en [15] de súper resolución trata el concepto como una fusión de secuencias de imágenes borrosas y de baja resolución para finalmente producir una imagen o secuencia de imágenes de alta resolución mediante un proceso de inferencia.

Para la obtención de las imágenes borrosas y de baja resolución generalmente se realiza un preprocesado a la imagen de alta resolución, en el cual mediante un *downsampling* y ruido, se obtendría el input para los modelos, en la figura 2.1 se puede observar el proceso que se realiza para obtener la imagen de baja resolución junto a la formulación matemática en la ecuación 2.1.

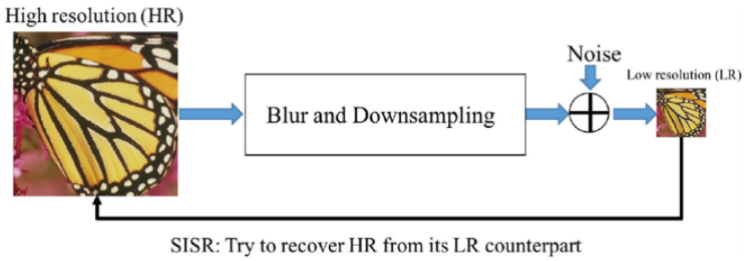

Figura 2.1: Estructura para obtener una imagen LR dada una imagen HR

$$I_{LR} = (I_{HR} \otimes k) \downarrow_s + n \tag{2.1}$$

donde $(I_{HR} \otimes k)$ es la convolución entre el kernel borroso k y la imagen de alta resolución I_{HR}, $\downarrow_s$ es la operación de *downsampling* y n ruido.

Cabe destacar que al inicio de la súper resolución se aplicaba la ecuación mostrada anteriormente 2.1, pero con el paso del tiempo se han ido haciendo pruebas

con distintos tratamientos de la imagen, concluyendo que el ruido no es del todo necesario aplicarlo a la imagen, quedando la fórmula como $I_{LR} = (I_{HR} \otimes k) \downarrow_s$, siendo la fórmula que se aplicará de ahora en adelante para todos los modelos que se presenten en este apartado.

Dentro de la súper resolución, existen distintos dominios a los que se aplica, uno de ellos serían las imágenes médicas, donde se busca tanto la mejora de las imágenes como la eliminación del posible ruido que se podría generar en las imágenes de resonancia magnética (MRI) y las tomografías computarizadas (CT), como se puede observar en la figura 2.2, en la imagen de la izquierda se puede ver un pequeño matiz de ruido comparado al de la derecha, que sería la imagen resultante de la súper resolución.

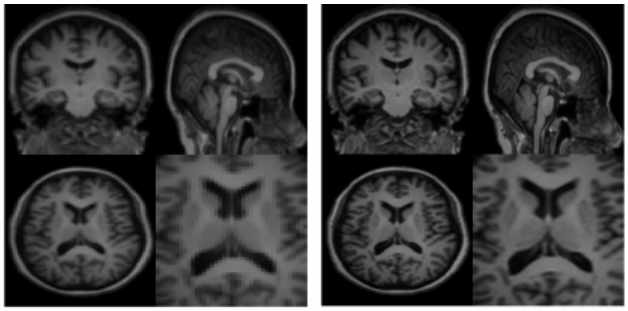

Figura 2.2: Resonancia magnética a baja calidad (izquierda) y la resultante a mayor calidad (derecha). Imágenes obtenidas de [16]

Otro de los dominios más comunes y más utilizados son las imágenes de satélite, ya que existe una gran dificultad para obtener información visual debido a las limitaciones existentes de los sensores de imagen actuales y de las complejas condiciones atmosféricas en los que se pueden encontrar en el momento de realizar la foto. Por lo que, una de las soluciones que se proponen a este problema es el uso de técnicas que permitan la mejora de las imágenes, como se puede llegar a ver en la figura 2.3, pudiendo escalar la imagen a una mayor resolución prácticamente sin perder información.

Figura 2.3: Imagen satelital a baja calidad (izquierda) y la resultante a mayor calidad (derecha). Imágenes obtenidas de [23]

Finalmente, el dominio asociado a este trabajo, es el de súper resolución facial, el cual a lo largo del capítulo se desarrollarán las distintas posibles soluciones para poder llegar a obtener una imagen facial de mayor resolución.

2.1 Súper resolución Facial

En muchos escenarios del mundo real, limitados por los sistemas físicos de obtención de imágenes y las condiciones posibles existentes que pueden provocar la mala calidad de la imagen, como los fenómenos meteorológicos, provocan que parte de las imágenes faciales estén a mala calidad. Por ello, una de las posibles soluciones para este tipo de problemas es la aplicación de súper resolución facial, pues ha sido un tema candente desde sus inicios en el campo de procesamiento de imágenes y visión por computador para aplicaciones del mundo real como los sistemas de seguridad y vigilancia por vídeo, reconocimiento y verificación facial.

La aparición del dominio específico se propuso en el año 2000 por Simon Baker y Takeo Kanade [1], argumentando la necesidad de realizar técnicas de super resolución en imágenes faciales extraídas de modelos de reconocimiento facial, en los cuales la cara resultante tenía baja resolución y debido a esto el modelo era incapaz de verificar qué persona era. Finalmente acababa proponiendo una técnica a lo que hoy en día llamaríamos, técnicas clásicas o basadas en estadística que se comentarán en la próxima sección.

2.1.1. Técnicas clásicas

Los que plantearon este nuevo campo, Kanade y Baker, propusieron un algoritmo el cual aprendiese la función de mejora de la resolución para imágenes de caras frontales. Usaron un algoritmo piramidal que aprendiese a priori sobre la derivada de las imágenes de alta resolución en función de la ubicación espacial en la imagen y la información de los niveles superiores de la pirámide. Un ejemplo claro de lo que se quiere mostrar como pirámide sería la figura 2.4, donde las características de la imagen de alta frecuencia se infieren de la *parent structure*, imágenes anteriores en la pirámide, mediante la búsqueda del vecino más cercano. La imagen final de nivel de gris se obtiene mediante descenso por gradiente para ajustar las restricciones mediante las características locales inferidas.

En la imagen, las figuras más bajas de G puede ser estimado desde un nivel más alto G_2 aplicando *sub-sampling* y añadiendo los niveles *laplacianos* de la imagen más bajos L_0 y L_1. Uno de los problemas de este modelo era que, no modelaba directamente la cara a priori y los píxeles se predicen individualmente, sin tener en cuenta el contorno global de la cara, causando discontinuidades y artefactos.

Otras de las técnicas que se aplicaban inicialmente son los métodos basado en subespacios. En [12], se emplea un enfoque de dos pasos para la mejora de caras, compuesto por un método local que utiliza una red de markov no paramétrica basada en parches para aprender la relación estadística entre la imagen global del rostro y las características locales, y un método global el cual supone una distribución *gaussiana* aprendida por *Principal Component Analysis* (PCA), técnica que nos permite reducir la dimensionalidad de los datos, transformando un gran conjunto de variables en otro más pequeño, pudiendo contener la mayor parte de la información del conjunto grande.

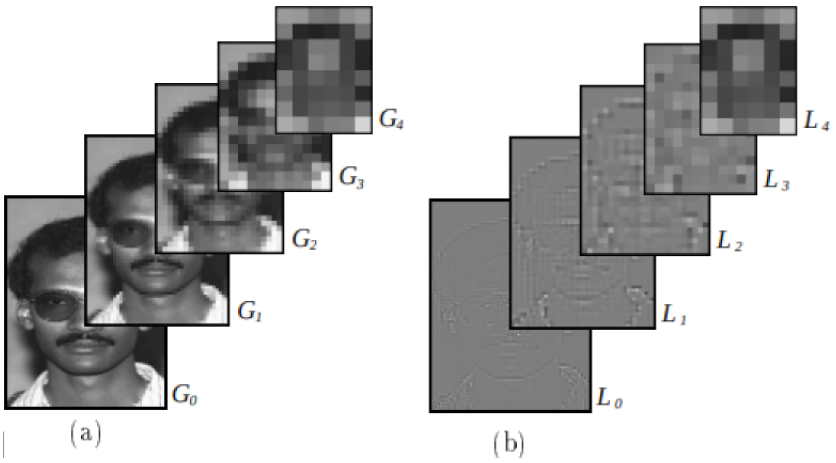

Figura 2.4: (a) Pirámide de Alta resolución a baja resolución y (b) Pirámide *laplaciano*. Imágenes obtenidas de [1]

Gracias a este método, se obtiene un modelo de apariencia generalizada al dataset de entrenamiento y posteriormente se aplica el modelo no paramétrico local para mejorar los detalles del modelo de apariencia generado por la PCA, un ejemplo de lo comentado se puede detallar en la figura 2.5, donde I_H^g es la imagen reconstruida con los componentes de baja frecuencia o resolución I_L, I_H^l es la imagen con los componentes de alta frecuencia o resolución obtenidos del modelo local no paramétrico (*Markov Random Field*) y finalmente se concatena ambas imágenes. Un detalle interesante de la figura 2.5 es que todas las imágenes son frontales y uno de los problemas de estos métodos es la necesidad de tener referencias de imágenes alineadas con la misma pose y la misma expresión facial, reduciendo así los posibles *datasets* en las que se pueda aplicar este tipo de técnicas.

Debido a las limitaciones existentes en los modelos anteriores, se desarrollaron múltiples técnicas que superaban esta barrera, entre ellas, Yang y compañía [22] presentaron un método estructurado de resolución facial que mejora la calidad de la imagen en base a la información estática de las imágenes faciales con la ayuda de técnicas de análisis Facial.

Con este algoritmo se consigue representar una cara mediante tres categorías locales que incluye componentes faciales, bordes faciales y regiones suaves, y aprovecha toda esta información para la mejora de la cara bajo diversas poses y expresiones. Como se observa en la figura 2.6, a la imagen oficial se le extrae los *landmarks* y con esos puntos del rostro estadísticamente se extraen los componentes en la imagen de buena resolución, los bordes y se aplica descenso por gradiente de las tres partes, integrando en una sola imagen los datos extraidos. El problema de este tipo de algoritmos es la dificultad de extraer los *landmarks* cuando el factor de *downsampling* es muy grande.

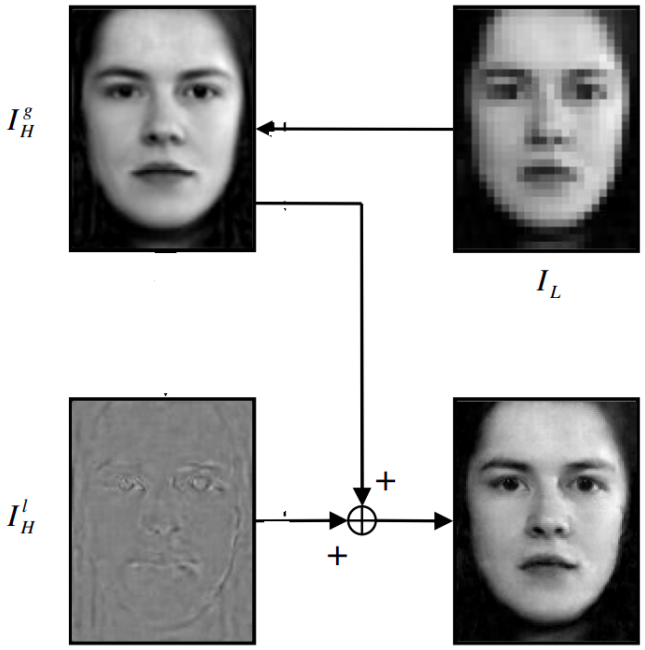

Figura 2.5: Ejemplo de aplicación super resolución con PCA + modelo no paramétrico local (Random Markov Field). Imágenes obtenidas de [12]

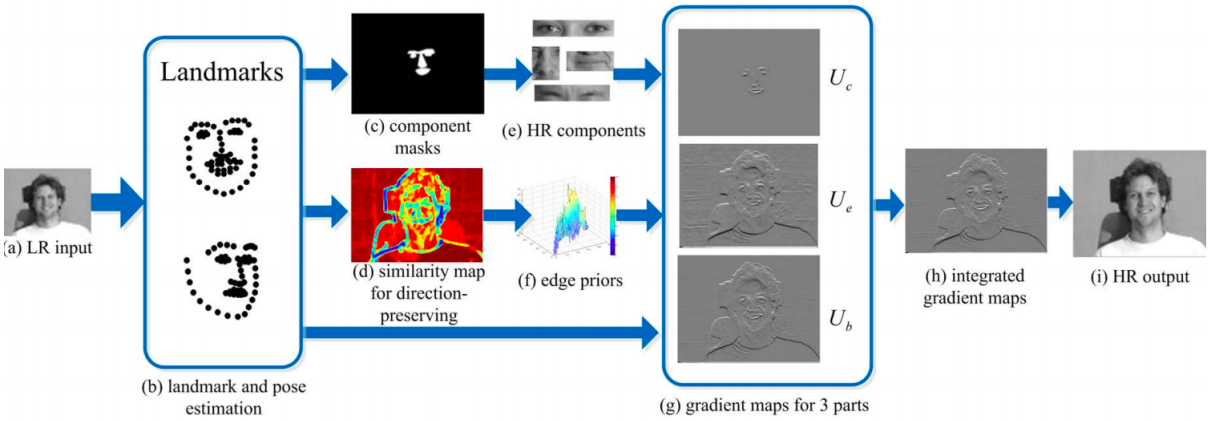

Figura 2.6: Ejemplo de aplicación super resolución extrayendo la información en base a los puntos de referencias faciales. Imágenes obtenidas de [22]

Con el rápido desarrollo de las técnicas de *Deep Learning*, se han obtenido grandes resultados comparado a intentos previos de súper resolución y se han aplicado tanto de imagen como vídeo centrado en la mejora de la resolución facial. En la siguiente sección se presentarán distintos aspectos de los mayores avances en súper resolución facial en *Deep Learning*.

2.1.2. Técnicas Deep Learning

Dentro de los métodos de súper resolución basados en *Deep Learning*, existen dos categorías, mostrados en la figura 2.7, en la primera se trata de explorar el potencial de redes eficientes para súper resolución facial pero dejando a un lado la particularidad de la cara, es decir, desarrollar una red neuronal convolucional (CNN) o una *Generative Adversarial Network* (GAN) para la reconstrucción sin tener en cuenta si la imagen contiene información facial o no.

Por otra parte, otros de los métodos están centrados en la utilización de información específica del rostro, denominado información previa o *prior*, donde se aprovecha la información que se puede extraer de las imágenes, por ejemplo, la estructura del rostro facial, los componentes de la cara, detalles adicionales como gafas, colgantes... Para finalmente generar una súper resolución mas limpia centrado en el rostro.

Existen además otros métodos que nos proveen información en base a secuencias de imágenes o audio, el cual puede usarse para la restauración facial, pero este tipo de métodos no se mencionará más a lo largo del proyecto.

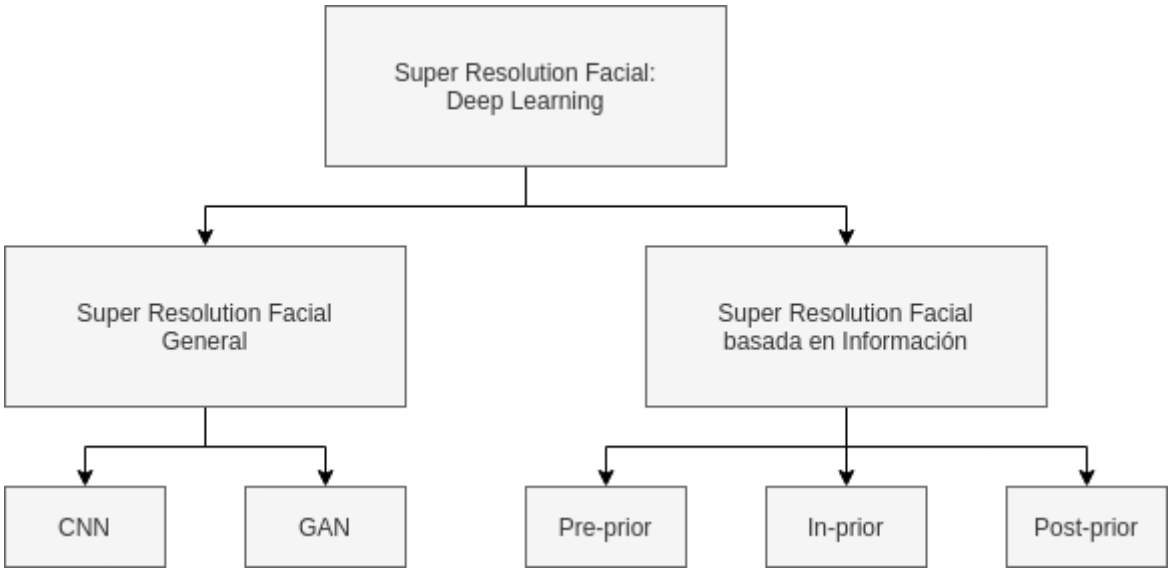

Figura 2.7: Estructuras de las técnicas *Deep Learning* para Súper resolución facial

Finalmente, se dividirá la sección en un total de dos categorías y se comentará dentro de cada categoría sus ramas, mostrando además un ejemplo en cada una de ellas.

Súper resolución facial General

En los métodos de súper resolución generalmente se diseña una red neuronal eficiente para imágenes de rostros sin tener en cuenta los aspectos específicos que aportan información del usuario. Estos métodos aprovechan el potencial de las redes neuronales desentendiéndose de los rostros, tratando de mejorar la imagen sin importar si contiene información facial o no. Estos son los denominados métodos de súper resolución facial general.

Se dividirá el apartado en dos categorías, aquellos métodos basados en redes neuronales convolucionales (CNN) y el resto de métodos en *Generative Adversarial Networks* (GAN). Existen otros tipos de métodos, los de aprendizaje por refuerzo y los basados en aprendizaje por ensamblaje pero en este proyecto no se mencionarán.

Métodos basados en CNN: Dentro de la categoría CNN existen distintos métodos, aquellos en los que alimentan la imagen entera a la red y posteriormente recuperan el rostro de las imágenes globalmente, métodos locales que recortan las imágenes faciales para alimentarlo a la red y luego recupera las imágenes faciales localmente y finalmente tenemos los métodos híbridos, es decir, recuperan la imagen local y globalmente. Además, existen diferentes topologías ya que como nuestro objetivo es convertir una imagen de baja resolución a alta resolución, dependiendo de en qué momento se quiere hacer el *upsampling*, se realiza una topología u otra. Un ejemplo claro es el siguiente, en la figura 2.8, en la cual el *upsampling* se puede realizar al inicio, a lo largo del recorrido, o al final del todo.

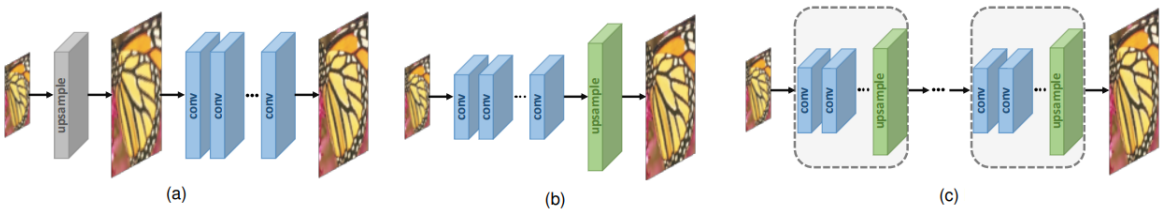

Figura 2.8: Ejemplo de posibles topologías para SR, (a) se aplica el *upsampling* al inicio (b) se aplican al final y (c) se aplica a por fases. Imágenes obtenidas de [20]

A continuación, se explicarán los distintos métodos existentes dentro de la categoría CNN para súper resolución.

- **Métodos globales**: el procedimiento general que se aplica en este tipo de métodos es el de aprender a mapear desde una imagen de baja resolución a una imagen a alta resolución. Zhou y compañía [25] propuso una red neuronal convolucional de doble canal (BCCNN) el cual integra interpretaciones faciales de la imagen extraídos por un modelo de extracción facial y la imagen original interpolada. Acabando con una combinación lineal entre los dos canales para obtener la predicción de la imagen resolutiva. En la figura 2.9 se puede observar lo que se ha descrito. Posteriormente, se propuso la misma súper resolución pero añadiendo más capas para la reconstrucción y extracción facial, denominado SRCNN [4], la estructura se puede observar en la figura 2.10, en la cual está dividido en dos, por una parte tenemos la extracción de características y por la otra se trata de reconstruir la imagen con las características obtenidas.

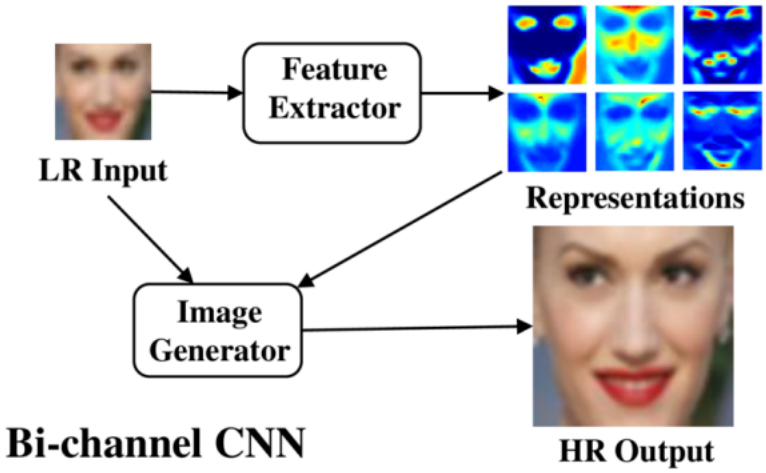

Figura 2.9: Estructura de la BCCNN. Imágenes obtenidas de [25]

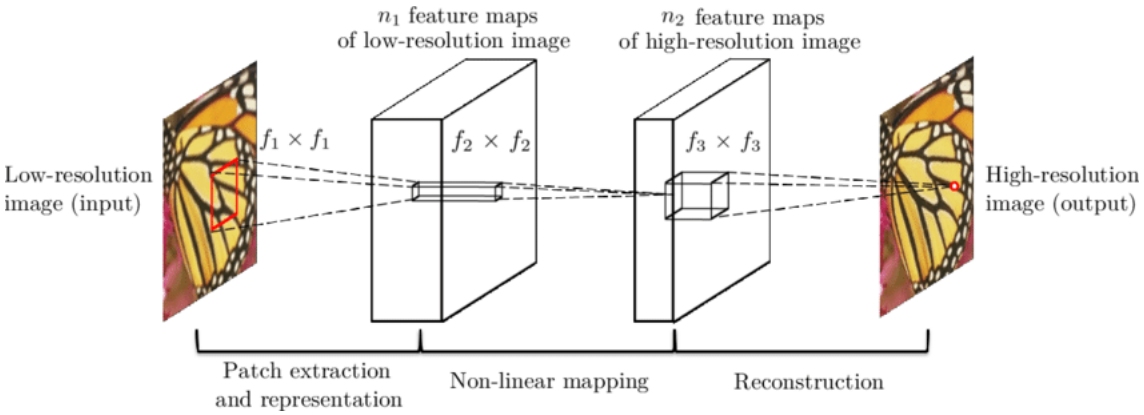

Figura 2.10: Estructura de la SRCNN. Imágenes obtenidas de [4]

- **Métodos locales**: los métodos globales pueden capturar información global de la imagen, pero ignoran las diferencias entre las regiones faciales, por lo tanto, el funcionamiento del modelo empeora cuando se realiza una inferencia en imágenes con rostros faciales. Uno de los ejemplos para el apartado de reconstrucción facial local es el propuesto por Xiao y compañía [7], *definition-scalable inference* (SRDSI), observando en la figura 2.11, los rostros etiquetados de alta resolución se descomponen primero en rostros básicos y rostros mejorados para entrenar un modelo de inferencia facial básico y un modelo de inferencia de rostros mejorados, posteriormente se usan los modelos para mejorar los rostros básicos de baja resolución (LR) y los rostros mejorados de alta frecuencia (HR). Finalmente la cara básica se fusiona con su cara mejorado y se mejora la misma cara con una red neuronal convolucional muy profunda de super resolución (VDSR) [10].

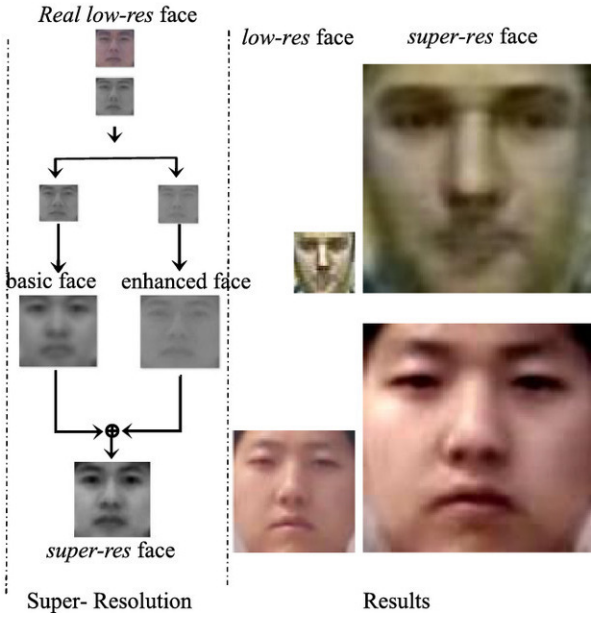

Figura 2.11: Estructura de la SRDSI. Imágenes obtenidas de [7]

- **Métodos híbridos**: se combina lo importante de un método y de otro, es decir, se captura la estructura global y se recuperan los detalles locales faciales. Un ejemplo reciente de estos métodos, sería el propuesto por Jiang y compañía [8], el cual se construyen dos ramas individuales, una subred de memoria global (GMS), caracteriza la forma facial holística empleando aprendizaje recurrente residual denso (*recurrent dense residual learning*) para excavar un contexto de amplio alcance a través de series espaciales. La otra subred, denominada subred de refuerzo local (LRM), se usa para el aprendizaje de componentes faciales locales, centrado en las relaciones de mapeo por parches entre el espacio de baja resolución (LR) y el de alta resolución (HR) en las regiones locales y no en toda la imagen. Finalmente pasa por un modelo de fusión y reconstrucción (FRM) que permite generar la correspondiente imagen de buena calidad. El proceso del método DPDFN se puede observar en la figura siguiente 2.12.

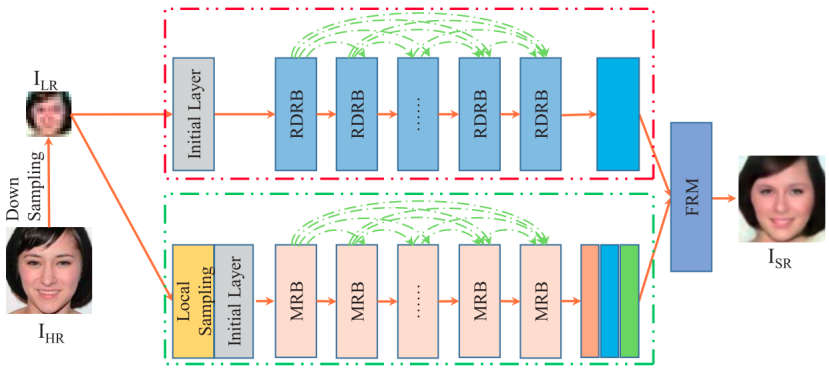

Figura 2.12: En rojo el extractor del contorno global de la cara y en verde detecta los detalles locales de los componentes faciales, al final se fusiona los dos resultados. Imágenes obtenidas de [8]

Métodos basados en GAN: Se trata de modelos que pueden ser entrenados con datos con y sin *labels* y pueden generar imágenes reales con detalles, el cual inspira a los investigadores a utilizar los GAN para recuperar imágenes faciales con detalles adicionales. El funcionamiento de los GAN's es el siguiente, se utilizan dos redes, un modelo discriminatorio que te distinga la imagen real a alta resolución de la salida de la imagen artificial y un modelo generativo que produce las imágenes faciales SR para engañar al modelo discriminatorio y minimizar el error existente entre la imagen generada SR y la de alta resolución HR. Por lo tanto, la competición entre ambas subredes permite al modelo generativo generar imágenes con una mejor calidad perceptual.

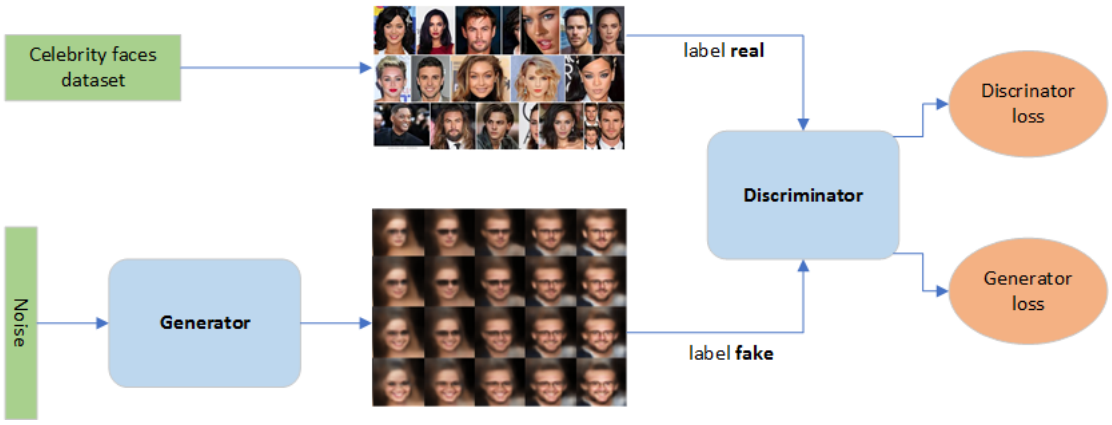

Figura 2.13: Ejemplo de aplicación GAN con base de datos Celeb_A. Imágenes obtenidas de [13]

Súper resolución facial basada en información previa

Muchas de las imágenes faciales tienen información específica que permite distinguir de otras caras, por ejemplo, la información previa que algunos datasets proveen, como pueden ser los puntos de referencia de la cara, también denominados *facial landmarks*, mapas de análisis faciales y mapas de calor faciales, que nos indican las ubicaciones de las distintas partes de la cara, como pueden ser, nariz, ojos, boca ...

Los métodos de súper resolución facial general ignoran esta información que se provee, generando imágenes faciales con una estructura facial difusa. Para poder recuperar imágenes faciales con una estructura facial más clara, se utilizan los métodos de súper resolución facial basada en información previa. Existen distintas formas de extraer la información previa facial, desde la imagen de baja resolución (*pre-prior*), en una fase intermedia de la súper resolución (*In-prior*), después de haber generado la imagen de súper resolución (*post-prior*) o en paralelo con la generación de la imagen de súper resolución (*Parallel-prior*), aunque esta forma no se mencionará en el proyecto. Cada categoría se explicará en los siguientes párrafos, añadiendo un modelo de cada tipo.

- **Pre-prior:** Se extrae la información previa desde la imagen LR mediante una red extractora el cual puede ser un modelo pre-entrenado o una subred en el modelo, posteriormente aprovecha esa información para facilitar la súper resolución facial. En [18], la imagen de entrada LR se divide en cinco componentes faciales mediante un modelo pre-entrenado para extraer los *facial landmarks*. Se escoge cada uno de ellos y se remuestrea utilizando una CNN correspondiente para generar un componente facial profundo. Las estructuras de grano fino pueden extraerse de las imágenes de entrenamiento de RH. Transferimos sus detalles para mejorar el componente facial profundo y generar un resultado. El *pipeline* del modelo se puede observar en la figura 2.14.

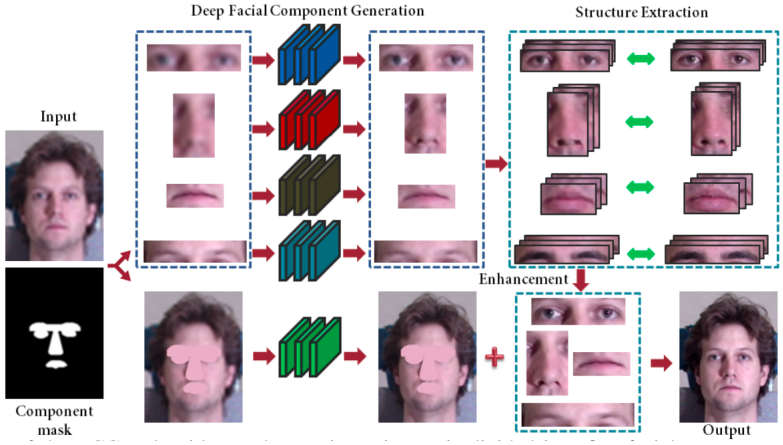

Figura 2.14: Estructura del modelo LCGE. Imágenes obtenidas de [18]

- **In-prior:** Al tratar de extraer la información previa facial, debido a la baja resolución de las imágenes, generalmente puede llegar a ser compleja e incluso en ocasiones imposible la obtención de la información. Por lo tanto, se plantea este tipo de métodos en los cuales la extracción de la información previa se lleva a cabo entre medias de la inferencia del modelo de súper resolución. El primer modelo que se llevó a cabo con este tipo de métodos es la FSRNet [3], donde al inicio a la imagen se le aplica una súper resolución robusta para mejorar la calidad, posteriormente, de cada una de las imágenes se extraen los *facial landmarks* y *heatmaps* y finalmente se concatenan y se aplica un *decoder* que obtendría la imagen mejorada en base a la imagen resolutiva y la información extraída. Una imagen clara de todo lo descrito se puede observar en la figura 2.15.

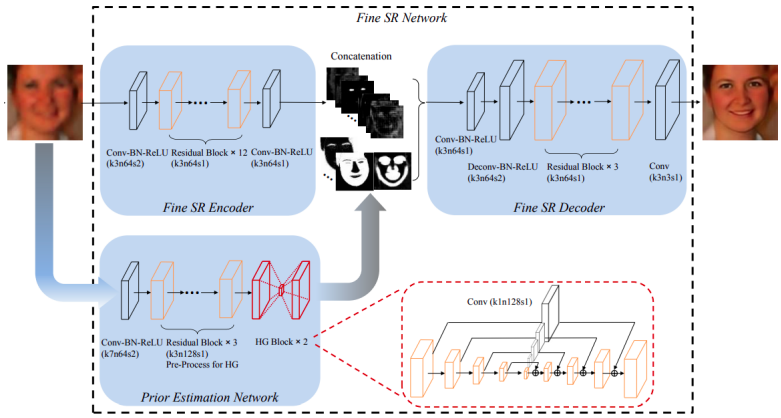

Figura 2.15: Ejemplo de aplicación super resolución FSRNet. Imágenes obtenidas de [3]

- **Post-prior:** A diferencia de los métodos *pre-* e *in-prior*, los métodos *post-prior* localizan la información una vez se tienen los resultados de la inferencia del modelo de súper resolución. Uno de los modelos que se han mencionado ya, FSRNet, pero modificando que la obtención de la información es una vez extraída la imagen súper resolutiva por completo [9].

CAPÍTULO 3
Metodología

En este apartado se van a describir con detalle los pasos, técnicas y herramientas que se han aplicado para el desarrollo tanto del modelo como del *pipeline* propuesto. Desde el conjunto de datos, pasando por su preprocesado, el diseño del modelo presentado hasta la experimentación que se ha aplicado.

Nuestro objetivo del trabajo es realizar súper resolución sobre una imagen y compararla con su representación original. Para ello, dentro de las técnicas que se han mostrado en el capítulo anterior, súper resolución facial general y súper resolución basada en información previa, se ha optado por la segunda opción, ya que nuestro método obtiene información previa de la cara mediante un extractor de características y se representa en forma de "embeddings". Como se comenta en el capítulo 2, dentro de la súper resolución basada en información previa hay distintas formas de extraer esa información, al inicio, durante el trascurso o al final, en este caso, se aplicará al principio ya que ésta información se usará como función de pérdida. Para el modelo principal, dentro de los posibles alternativas existentes, redes neuronales convolucionales (CNN), o redes generativas adversarias (GAN). Se escogerán las redes neuronales convolucionales, esto se debe a que se busca una representación real de la cara para posteriormente realizar una verificación facial, pues los modelos GAN lo que tratan de hacer es generar una cara similar a la de la imagen, aunque no sea la misma, como se puede observar en la imagen 3.1, es preferible la calidad perceptual que te proporciona los CNN que suele ser peor que las GAN a generar un rostro facial distinto al original, que puede llevar a confusión en la verificación facial.

Figura 3.1: Ejemplo de aplicación GAN para súper resolución facial

Nuestra implementación se basa en utilizar un modelo de súper resolución facial general, para la súper resolución de la imagen facial y un modelo pre-entrenado con VGGFace2 [2] , el cual es un dataset que cuenta con un total de 3.31 millones de imágenes de un total de 9131 personas, famosas generalmente, ya que los datos se han extraído de internet. El modelo pre-entrenado nos servirá para extraer las características faciales del usuario. De esta forma la estructura sería una similar a los métodos globales en CNN pero usando información previa *pre-prior* en la cual se usa un modelo global para la súper resolución facial y un modelo local que nos extrae las características del rostro para mejorar la calidad perceptual de la imagen.

3.1 Herramientas

Todas las librerías y *scripts* han sido implementados en *Python*, ya que contiene una gran cantidad de librerías y que, dada su comodidad y su rapidez a la hora de realizar prototipados, nos permite realizar nuevos experimentos. *Python* es el lenguaje más popular en el uso del *Deep Learning* gracias a la gran comunidad que hay detrás, pues gran parte de las API's necesarias para este proyecto están enfocados en este lenguaje de programación. En este caso, se ha utilizado la librería *Keras*, gracias a su facilidad para el entrenamiento e inferencia de modelos y *TensorFlow*, que nos permite utilizar métodos eficientes para el uso de procesamiento de imágenes.

Además, se han utilizado otras librerías que han sido también relevantes para el proyecto, *OpenCV*, para el tratamiento de imágenes, aumentando o disminuyendo su tamaño y *NumPy*, para la manipulación de matrices a la hora de convertir las imágenes en Arrays.

En cuanto al hardware empleado para la experimentación, se ha utilizado Google Colab y el kernel de Kaggle, con tarjetas gráficas Nvidia K80 o Nvidia T4 y CPU de Intel(R) Xeon(R) de 2 núcleos con 26 GB de RAM, aunque, a pesar de tener acceso a tales gráficas, el límite de uso es de 24 horas seguidas y/o 30 horas semanales.

3.2 Dataset

Muchos conjuntos de datos de imágenes faciales se utilizan para súper resolución facial, que difieren en muchos aspectos (cantidad de imágenes por usuario, información específica de la cara, componentes adicionales faciales ...). Estos datasets proveen una imagen facial de alta resolución, en el cual de ahí se genera su correspondiente imagen en baja resolución mediante un proceso de degradación (*downsampling*).

Muchos de los *datasets* no nos sirven, lo que se busca es tener una base de datos de un total de entre 20 y 30 personas, en la cual cada una de las etiquetas contenga un número grande de fotos con distintas poses y gestos en la imagen. Se quiere

escoger un cierto número de personas con muchas imágenes para poder realizar el apartado de verificación facial con un modelo entrenado con esas imágenes.

Ya con las restricciones dadas, muy pocos lo cumplen, generalmente las imágenes son extraídos de Internet y muchas de las etiquetas contienen dos o tres imágenes mientras que otros pueden contener cien, hay bastante irregularidad. Estos tipos de datasets, suelen ser usados para entrenar una GAN sin importar las etiquetas, ya que este tipo de arquitecturas permiten el entrenamiento sin importar el número de etiquetas por usuario.

3.2.1. Dataset utilizado

El conjunto de datos del proyecto se ha extraído en la página web oficial de la Universidad de San Diego, California [5]. Contiene un total de 16128 imágenes, con un total de 28 personas, de los cuales cada persona se le ha realizado imágenes con 9 poses distintas y un total de 64 condiciones de iluminación, por lo que, para cada persona hay un total de 576 imágenes. Se puede observar una serie de imágenes del dataset en la figura 3.2.

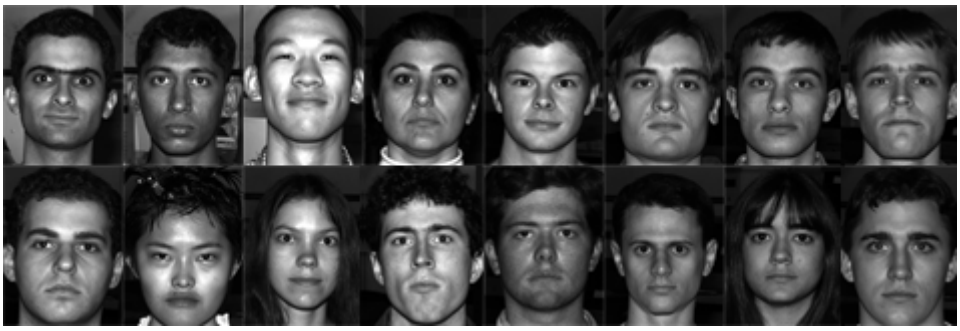

Figura 3.2: Imágenes dataset Yale-B Extended. Imágenes obtenidas de [5]

Las imágenes tienen un tamaño de 640 de alto y 520 de ancho, generalmente bastante grande para posteriormente pasarlo por una red neuronal artificial, por lo tanto lo que se ha decidido es reducir el tamaño de las imágenes a 256 tanto de ancho como de alto.

De forma que, se realizará un preprocesado de las imágenes para agilizar el proceso de aprendizaje e inferencia del modelo.

Preprocesado

Las imágenes están compuestas por una cara y prácticamente el mismo fondo para todas las imágenes, como en este proyecto queremos centrarnos todo lo relacionado al rostro, rasgos, poses del usuario en la imagen, se ha decidido seguir los siguientes pasos:

1. **Extracción facial**: Para ello se utilizará una red neuronal adicional para poder reducir el tamaño de las imágenes, en este caso, la imagen tendrá una dimensión de 640x520 y lo reduciremos a 256x256, para ello, se utiliza *Multi*

Task Cascade Convolutional Network (MTCNN), modelo que detecta el rostro facial y deja un *bounding box* alrededor del rostro facial.

2. **Reescalado**: En base al *bounding box* obtenido del paso anterior se aumenta las dimensiones hasta poder extraer las imágenes a 256x256.

3. **Normalizado de los datos**: Los valores de los píxeles se normalizan en un rango de 0 a 1 en cada uno de los componentes de la imagen. De forma que aquellos píxeles cercanos a 255 serán cercanos a 1 y se reescalaran proporcionalmente.

La distribución para este dataset se puede observar en la tabla 3.1.

	Partición		
	Entrenamiento	Test	Total
Yale-B	9323	1645	10968

Tabla 3.1: Distribución de las imágenes del dataset Yale-B Extended, en los subconjuntos de entrenamiento y test

3.3 Modelo

El objetivo de este apartado consiste en detallar la propuesta de diseño del sistema para la súper resolución, posteriormente se detallará las métricas que se han empleado para comparar los resultados y finalmente se hará hincapié en la función de pérdida propuesta.

3.3.1. Arquitectura de los modelos utilizados

Modelo Súper resolución

Para el diseño de los modelos se ha optado por utilizar redes neuronales convolucionales (CNN),debido a que son el tipo de redes más potentes para trabajar con imágenes gracias a su capacidad de captar características y patrones de las imágenes mediante sus filtros y, además por el motivo que se ha comentado al inicio de Metodología, capítulo 3, pues los *Generative Adversarial Networks* a pesar de generar una mejor calidad perceptual en la imagen, a la hora de realizar la verificación facial no podría reconocer correctamente a la persona.

El bloque básico, proveniente de este proyecto [24], está compuesto por una capa convolucional con los pesos normalizados, un kernel de 3x3 con función de activación ReLU, acompañado de otra capa convolucional más pequeña a la anterior, también con los pesos normalizados y acabando con una capa residual con el inicio del bloque. La estructura se puede observar en la figura 3.3, comparando al resto de bloques residuales que se han aplicado en los proyectos que se han llevado a cabo con Súper Resolución en los últimos años.

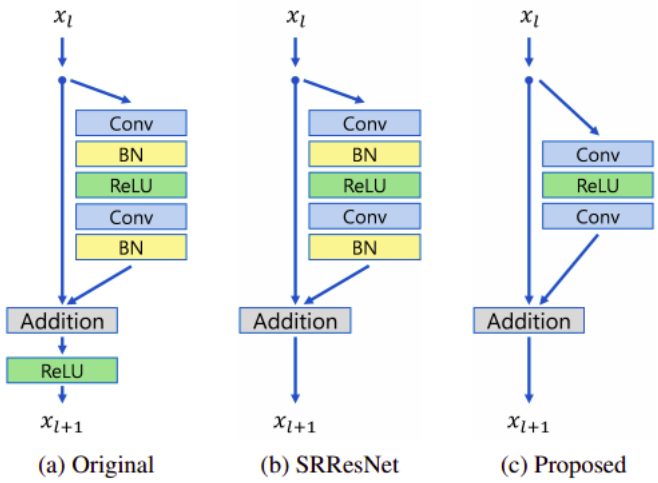

Figura 3.3: Arquitectura bloques residuales, añadiendo el propuesto para este proyecto

La topología del modelo propuesto está basado en la U-net, compuesto por 8 bloques convolucionales de: los 3 primeras capas acompañados por un *Maxpooling* de 2x2, nos permite reducir las dimensiones a la mitad y una capa *dropout* con la propbabilidad $p = 0{,}3$ de desactivar las neuronas para añadir regularización y evitar el *overfitting*. Posteriormente, se aplica una capa convolucional que representaría la información que se ha extraído en las capas anteriores y finalmente se usan otros 3 bloques acompañados por un *UpSampling* de 2x2 que nos permite aumentar las dimensiones al doble y se aplica a cada bloque una capa residual que suma la capa encoder y decoder. Acabando con una capa convolucional de 1 canal como en el inicio ya que la imagen será *grayscale*, como queremos que salga en el *output*. La estructura se puede comprender mejor con la siguiente imagen.

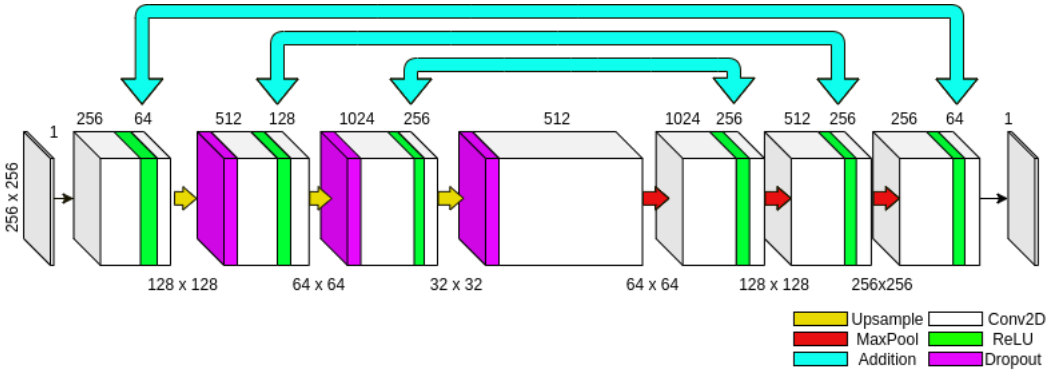

Figura 3.4: Arquitectura del modelo de super resolución

En base a las pruebas que se han realizado en este proyecto, se observaba un problema de *overfitting*, en el cual en el *training* el modelo entrenaba pero sin embargo en el *test* no avanzaba, por lo tanto se aplicó ese *dropout* de probabilidad 0.3. Además, como se puede observar en la figura 3.4 en azul tenemos *skip connections*, ya que en los proyectos más recientes de *Single Image Super resolution* (SISR) [24],

se recomendaba su uso ya que ello permitía una mejora bastante considerable y se sugería desistir en el uso de *BatchNorms* ya que empeoraban los resultados. Otra de las cosas que se han aplicado en el modelo para sustituir los *batchnorms* es la normalización de los pesos, ello permitía un entrenamiento mucho más estable con el paso de las épocas.

Modelo preservación de identidad

Una vez explicado con detalle el modelo que se usará para súper resolución facial, es turno de mostrar el modelo que se usará para realizar la función de pérdida propuesta. Para ello, lo que se usará un modelo pre-entrenado con el dataset VGGFace2 [2], con el objetivo de extraer características más profundas de las imágenes de entrada, centrándose en las características faciales. La topología del modelo es una *resnet*, arquitectura en la cual hace 6 años ganó en una competición en clasificación de imágenes, aplicando una novedad en su momento, para evitar el desvanecimiento de gradiente que ocurría cuando se usaban redes neuronales planas tan grandes. Se diseñó un tipo de bloque en el cual se entrena con la suma entre la operación de la capa y la entrada de la capa anterior, esto se denomina bloque residual, mostrado en la figura 3.5. Dentro de las posibles variantes que se pueden observar en la figura 3.6, se escogerá el intermedio, la *resnet50* con los pesos ya pre-entrenados con VGGFace2.

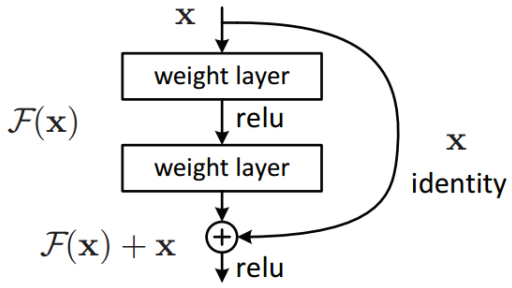

Figura 3.5: Esquema de la arquitectura de un bloque residual. Sacado de [6]

layer name	34-layer	50-layer	101-layer
conv1	7×7,64, stride 2		
	3×3 max pool, stride 2		
conv2_x	$\begin{bmatrix} 3 \times 3,64 \\ 3 \times 3,64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1,64 \\ 3 \times 3,64 \\ 1 \times 1,256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1,64 \\ 3 \times 3,64 \\ 1 \times 1,256 \end{bmatrix} \times 3$
conv3_x	$\begin{bmatrix} 3 \times 3,128 \\ 3 \times 3,128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1,128 \\ 3 \times 3,128 \\ 1 \times 1,512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1,128 \\ 3 \times 3,128 \\ 1 \times 1,512 \end{bmatrix} \times 4$
conv4_x	$\begin{bmatrix} 3 \times 3,256 \\ 3 \times 3,256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1,256 \\ 3 \times 3,256 \\ 1 \times 1,1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1,256 \\ 3 \times 3,256 \\ 1 \times 1,1024 \end{bmatrix} \times 23$
conv5_x	$\begin{bmatrix} 3 \times 3,512 \\ 3 \times 3,512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1,512 \\ 3 \times 3,512 \\ 1 \times 1,2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1,512 \\ 3 \times 3,512 \\ 1 \times 1,2048 \end{bmatrix} \times 3$
	average pool,2048-d fc		

Figura 3.6: Arquitecturas de las variantes del modelo ResNet. Extraído de [6]

El *dataset* que se utiliza para esta red neuronal es VGGFace2 [2], consiste en una base de datos con un total de 3.31 millones de caras de un total de 9131 sujetos, con una media de 320 imágenes por persona. Este tipo de *datasets* se utilizan para el reconocimiento de caras teniendo en cuenta la edad y la pose de cada usuario. Como este dataset contiene muchas imágenes y es bastante pesado, lo que se realizó fue entrenar con el modelo de este apartado y guardar los pesos, permitiendo así ahorrarnos mucho espacio y en caso de querer realizar un pre-entrenamiento, solo haría falta extraer los pesos.

Para este apartado se aplica *Transfer learning* [11], consiste en preparar un modelo que ha sido entrenado previamente con un dataset y ahora se entrena con otro dataset con una distribución de clases nueva. En este caso, se entrenará con el dataset *Yale-B*, añadiendo una capa densa de 4096 para la extracción de características para el modelo final y posteriormente una *softmax* del número de clases, en este caso, 28.

Modelo final

Una vez presentados los modelos que se van a utilizar para llevar a cabo este proyecto, quedaría únicamente como va estar estructurado, se puede observar en la figura 3.7, los dos bloques grandes representan los modelos descritos y los cuadrados son los *embeddings* resultantes obtenidos con la extracción de características del modelo pre-entrenado de 4096 neuronas. La línea roja representa la imagen de alta resolución y la azul la imagen obtenida en la inferencia de súper resolución. En el modelo de súper resolución se aplica la función de pérdida *Mean Squared Error* y para el modelo pre-entrenado se aplicará la función de pérdida preservación de identidad. Finalmente, se tratará de utilizar una función de pérdida múltiple atribuyendo un peso a cada una, todo esto se comentará en la sección correspondiente.

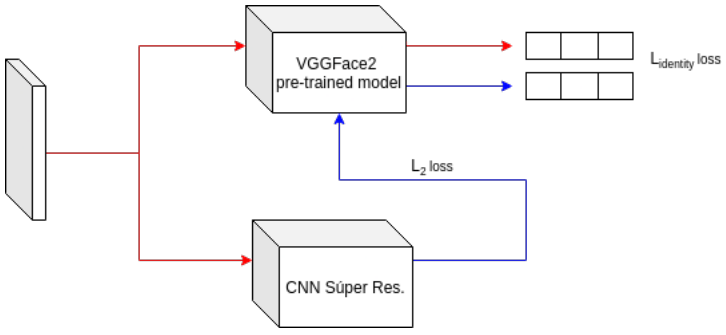

Figura 3.7: Estructura del modelo final

3.3.2. Métricas empleadas

Para la evaluación de los sistemas implementados se emplean unas métricas cualitativas determinadas que nos permitan medir la calidad de los resultados obtenidos y compararlos. Para observar el nivel de ruido existente en la imagen se ha usado la métrica PSNR y para observar la similaridad de las imágenes se ha utilizado la métrica SSIM [21].

Peak Signal Noise Ratio (PSNR)

Dada una imagen facial de referencia como *ground truth* I_{HR} y el resultado de las imágenes que han pasado por la fase de inferencia una vez entrenado el modelo I_{SR}, el primer paso que se realiza es calcular el *mean squared error* (MSE) entre ambas imágenes. Una vez se tiene el MSE y el número máximo de valor de cada píxel, L, se obtiene la métrica PSNR, el cual es definido de la siguiente forma:

$$PSNR = 10log_{10}(\frac{L^2}{MSE}),$$

donde L es el valor máximo obtenido en un píxel y suele ser 255, aunque nosotros al normalizar entre [0-1], el máximo valor de nuestro píxel será de 1, el cálculo de la *Mean Squared Error* es el siguiente:

$$MSE = \frac{1}{hwc}\sum_{i,j,k}(I_{SR}^{i,j,k} - I_{HR}^{i,j,k})^2,$$

donde h,w,c son la altura, el ancho y el número de canales respectivamente.

La función que tiene el PSNR es centrarse en la distancia entre cada pixel de ambas imágenes, el problema de esta métrica es que no es consistente con la percepción humana, a pesar de que cuanto mayor sea el valor, mejor será el modelo, su rendimiento es bastante pobre en casos en los que se tenga en cuenta la percepción humana, como va a ser nuestro caso.

Structural Similarity Index (SSIM)

Distinto a PSNR en el cual se mide únicamente la diferencia entre píxeles, el SSIM se propone para medir la similaridad estructural entre la información estructural entre ambas imágenes. Para ello, la métrica SSIM mide la similaridad desde tres aspectos, luminancia, contraste y estructura. Dada la imagen de referencia y la de súper resolución, I_{HR} y I_{SR}, la luminancia y el contraste se estiman como la media y la desviación típica de la imagen:

$$\mu_{I_{HR}} = \frac{1}{hwc}\sum_{i,j,k} I_{HR}^{i,j,k},$$

$$\sigma_{I_{HR}} = (\frac{1}{hwc-1}\sum_{i,j,k}(I_{HR}^{i,j,k} - \mu_{I_{HR}})^2)\frac{1}{2},$$

donde $\mu_{I_{HR}}$ es la media y $\sigma_{I_{HR}}$ la desviación típica de I_{HR}. La similaridad entre la luminancia y el contraste se ha definido de la siguiente manera:

$$L(I_{HR}, I_{SR}) = \frac{2\mu_{I_{HR}}\mu_{I_{SR}} + C_1}{\mu_{I_{HR}}^2 + \mu_{I_{SR}}^2 + C_1},$$

$$C(I_{HR}, I_{SR}) = \frac{2\sigma_{I_{HR}}\sigma_{I_{SR}} + C_1}{\sigma_{I_{HR}}^2 + \sigma_{I_{SR}}^2 + C_1},$$

donde C_1 es una constante para evitar la división por 0. La similaridad estructural entre las imágenes es estimado como la correlaciones entre valores de píxeles normalizados, el cual se expresa como

$$\sigma_{I_{HR},I_{SR}} = (\frac{1}{hwc-1}\sum_{i,j,k}(I_{HR}^{i,j,k} - \mu_{I_{HR}})(I_{SR}^{i,j,k} - \mu_{I_{SR}})),$$

$$S(I_{HR}, I_{SR}) = \frac{\sigma_{I_{HR},I_{SR}} + C_2}{\sigma_{I_{HR}}\sigma_{I_{SR}} + C_2},$$

quedando finalmente la fórmula de la métrica como

$$SSIM(I_{HR}, I_{SR}) = L(I_{HR}, I_{SR}) * C(I_{HR}, I_{SR}) * S(I_{HR}, I_{SR}).$$

La métrica varía entre 0 y 1, cuanto más cercano es a 1, mayor similaridad hay entre la imagen de súper resolución a la de referencia (alta resolución).

3.3.3. Función de pérdida propuesta

Función de pérdida píxel

La función de pérdida atribuida como *pixel loss* mide la distancia entre las dos imágenes a nivel de píxel, existen dos funciones de pérdidas, $\mathcal{L}_1$, calcula el error medio absoluto y L2 calcula el medio error cuadrático, para este proyecto se usará el $\mathcal{L}_2$, por lo tanto esta función de pérdida se expresará de la siguiente manera:

$$\mathcal{L}_2(I_{HR}, I_{SR}) = || \ I_{HR} - I_{SR} \ ||_2 = \frac{1}{hwc}\sum_{i,j,k}(I_{HR}^{i,j,k} - I_{SR}^{i,j,k})^2,$$

donde h,w y c denotan la altura, anchura y el número de canales de la imagen y $I^{i,j,k}$ es el píxel de la localización (i, j, k). Lo que se pretende con esta función de pérdida es tratar de tener los píxeles de la súper resolución más cercanos a los de alta resolución. La función de pérdida $\mathcal{L}_2$ da mayor peso a los errores más grandes y mucho menor peso a los de menor error, mientras que el $\mathcal{L}_1$ es indiferente cuan de grande sea el error.

Función de pérdida preservación de identidad

La función de pérdida de preservación de identidad se centra en mantener la coherencia de la identidad entre la imagen súper resolutiva y la imagen de alta resolución. Generalmente se utiliza un modelo pre-entrenado de reconocimiento facial (FRN) para mantener la identidad de la persona. Al modelo de súper resolución se le provee una imagen de baja resolución (I_{LR}) obteniendo una imagen súper resolutiva (I_{SR}), con esta imagen se la proporcionamos al modelo de reconocimiento facial para obtener las características de su identidad, . En el mismo momento, se le provee la imagen de alta resolución I_{HR} para obtener sus características correspondientes. La función de pérdida basada en la preservación de identidad se calcula de la siguiente forma:

$$\mathcal{L}_{Identidad} = || \, FRN(I_{HR}) - FRN(I_{SR}) \, ||_2 = \sum_i (FRN(I_{HR})^i - FRN(I_{SR})^i)^2,$$

es decir, se aplica la misma función de pérdida que la de píxel pero en vez de por píxel se aplica por cada componente de los *embeddings*.

Función de pérdida total

Una vez explicadas las funciones de pérdida propuestos, es el turno de darle un peso α a cada función, de forma que la función de pérdida final se calcula de la siguiente forma:

$$\mathcal{L}_T = \alpha \mathcal{L}_2 + (1 - \alpha) \mathcal{L}_{Identidad},$$

donde α es el peso que se le atribuye a cada función de pérdida.

CAPÍTULO 4
Experimentación y Resultados

En este capítulo se expondrán los detalles de la experimentación que se ha llevado a cabo con el dataset, los resultados que se han obtenido y se analizarán los mismos.

La distribución de los datos se ha llevado a cabo de la siguiente manera, el 85 % de las imágenes se han usado como entrenamiento y el resto, el 15 %, como test.

Se ha entrenado con un total de 40 *epochs* para cada modelo con distintos hiperparámetros , el optimizador utilizado ha sido *Adam* al ser uno de los que mejores resultados suele proveer. Para el *learning rate* se ha utilizado un valor inicial de 0.0001 y un *batch size* de 16. El motivo por el cual se ha escogido un *learning rate* tan bajo es debido al *overfitting* que se genera cuando el ratio de aprendizaje es un poco más alto, pues el modelo enseguida deja de aprender.

Las métricas usadas para comparar los resultados son las comentadas en el capítulo 3.3.2, PSNR y SSIM. El primero para observar el nivel de ruido que se puede alcanzar en la imagen, cuanto más bajo el valor, mejor y SSIM para observar cuanta similaridad hay entre la super resolución obtenida y la de alta resolución. El resultado se observará si mejora en base a nuestro *Baseline*, en este caso, aquellas inferencias con una función de pérdida *Mean Squared Error*.

Para reescalar las imágenes se aplicará interpolación bicúbica el cual para cambiar el tamaño de las imágenes es bastante barato y obtiene resultados generalmente mejores, más liso y con menos *artifacts* comparado a otro tipo de interpolaciones como la lineal. El proceso es el siguiente, dada una imagen original, se *downsamplea* hasta una cierta escala y posteriormente se aplica interpolación bicúbica del mismo tamaño que la imagen original. Por ejemplo, en la primera imagen de la figura 4.1, se observa que el tamaño de las imágenes respectivamente son de 16x16, 256x256 y 256x256, pues se ha aplicado *downsampling* a la tercera imagen que es la original a una escala x16, obteniendo la primera imagen y posteriormente se ha aplicado *upsampling* con interpolación bicúbica, acabando como en la segunda imagen.

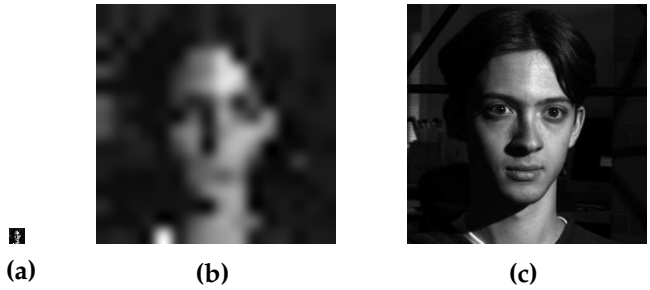

Figura 4.1: (a) Imagen original a 16x16, (b) Imagen a escala x16 aumentando su tamaño con interpolación bicúbica y (c) imagen original

En base a los resultados, se interpretarán dos aspectos, cuantitativamente y la calidad perceptual de las imágenes teniendo en cuenta el peso que se le atribuye a cada función de pérdida.

4.1 Resultados

Para seleccionar los mejores hiperparámetros, dada la limitación de las gráficas, lo que se ha hecho es realizar para cada modelo un entrenamiento de 40 épocas y observar tanto los resultados obtenidos con cada métrica como su *validation loss*, ya que, añadir una función de pérdida más puede modificar la suma de errores cuadráticos. Por lo que, se atribuirá cierto peso a cada función de pérdida, se probará con un valor de 0.1, 0.5 y 0.9 y nuestro *baseline* será el resultado obtenido con la función de pérdida MSE. Además, se quiere observar cualitativamente la calidad de las imágenes dependiendo de la distribución de pesos aplicados en este proyecto.

4.1.1. Comparación Cuantitativa

Escala	**Tam. Imagen**	**Resultados**			
		$\alpha = 0{,}1$	$\alpha = 0{,}5$	$\alpha = 0{,}9$	L_2
x2	128x128	39.73/98.35	**42.10/98.82**	41.61/98.75	40.43/98.56
x4	64x64	32.94/93.32	35.10/95.08	**35.79/95.47**	34.70/94.93
x8	32x32	30.15/87.65	**31.03/89.23**	30.84/89.13	29.93/88.50
x16	16x16	24.90/71.78	**25.53**/74.65	25.17/74.20	24.36/**75.51**
x32	8x8	20.53/52.96	20.82/61.32	**21.45**/63.01	21.20/**64.27**

Tabla 4.1: Tabla de resultados PSNR/SSIM modificando los pesos de cada función de pérdida

Los resultados obtenidos se pueden observar en la tabla 4.1, los valores de las alphas son de las métricas PSNR y SSIM. Se ve que el α con mejores resultados para escalas bajas, son con escalas x2, x4 y x8. Esto se debe a que se sigue teniendo información visual de la cara a pesar de haber reducido el tamaño de la imagen

y haberlo aumentado de nuevo con interpolación bicúbica. La distribución de pesos con mejores resultados cuantitativamente son con un alpha de 0.5 y 0.9, mejorando tanto la similaridad como el nivel de ruido comparado al *Baseline*, modelo con función de pérdida $\mathcal{L}_2$.

Sin embargo, en escalas mas altas, se observa que el nivel de ruido es mejor en un tipo de hiperparámetro que en otro, por ejemplo en la escala x16 la función de pérdida múltiple con un alpha de 0.5 tiene un mejor PSNR que el *baseline*, en cambio, en cuanto a similaridad se refiere, el *baseline* obtiene resultados mejores. Este suceso se puede explicar fácilmente con la fórmula del PSNR mencionada en la fórmula 3.3.2 y la tabla 4.2, resultados obtenidos de la función de pérdida MSE sobre cada alpha. Se puede interpretar de manera clara que cuanto mayor peso se le de a la función de pérdida $\mathcal{L}_{Identidad}$ menor es el error medio cuadrático y por lo tanto, como en la fórmula del PSNR se depende del MSE y los valores de los píxeles, cuanto menor es el MSE, más alto es el valor y mejores resultados se obtiene. Por lo tanto, cuando se aplica la métrica PSNR con la función de pérdida múltiple, provoca que a pesar de tener un valor alto, no se corresponde con los resultados obtenidos en la similaridad de las imágenes.

Escala	**Tam. Imagen**	**MSE**			
		$\alpha = 0{,}1$	$\alpha = 0{,}5$	$\alpha = 0{,}9$	L_2
x2	128x128	**0.00001**	0.00004	0.00007	0.00011
x4	64x64	**0.00008**	0.00017	0.00025	0.00036
x8	32x32	**0.00016**	0.00044	0.00079	0.00107
x16	16x16	**0.00048**	0.00158	0.00296	0.00392
x32	8x8	**0.00196**	0.00674	0.00707	0.00817

Tabla 4.2: Resultados obtenidos sobre el conjunto de test para cada distribución de pesos en la función de pérdida

4.1.2. Comparación Cualitativa

Una vez hablado de los resultados cuantitativos obtenidos, es turno de comparar cualitativamente las imágenes obtenidas para cada distribución de pesos. Se puede observar en la figura 4.2, un ruido en las imágenes dónde la función de pérdida $\mathcal{L}_{Identidad}$ predomina en la distribución de pesos, por lo que podemos suponer que ese ruido proviene de los "embeddings". En la figura 4.2, se puede ver que las imágenes con menor ruido y, por ende, mayor calidad son aquellos donde el α tiene un valor alto.

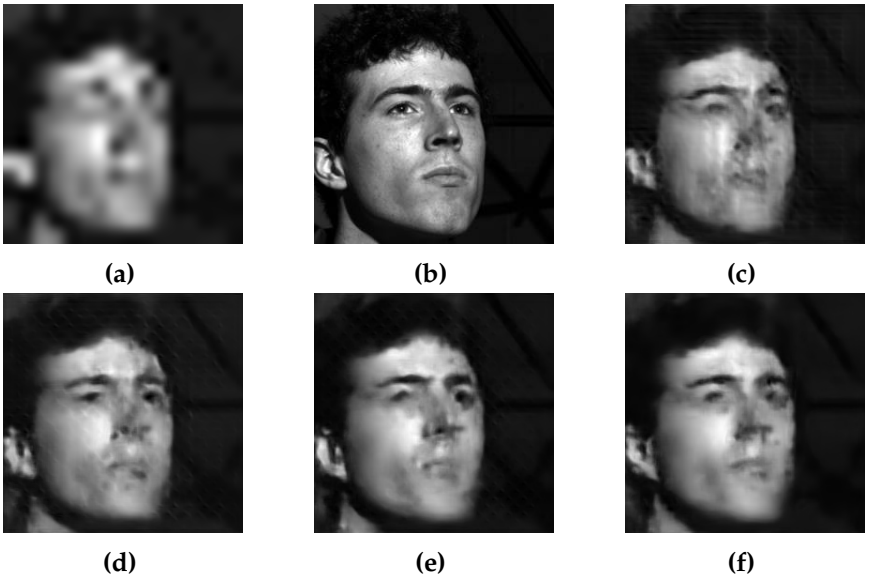

Figura 4.2: (a) Imagen a baja resolución, (b) imagen a alta resolución, (c) súper resolución con un α de 0.1, (d) resultado con α de 0.5, (e) resultado con un α de 0.9 y (f) *baseline*, sin usar función de pérdida múltiple

En cuanto hasta qué escala se puede reconstruir el rostro facial de las imágenes, se llevará a cabo apoyándonos en la tabla 4.1. Recordemos que se aplica en base a imágenes reescaladas, donde la escala más pequeña es x2, es decir, dada una imagen original de 256x256, se *downsamplea* a 128x128 y posteriormente se *upsamplea* al tamaño de la imagen original. Para este proyecto, como se comenta en la *hipótesis* del trabajo, nuestro objetivo consiste en tratar de reconstruir esas imágenes en las cuales no se pueda percibir mucho visualmente las caras y tratar de mejorarlas concatenando la información que se nos provee de los "embeddings" mediante una función de pérdida múltiple. Por lo tanto para este apartado, se experimentará hasta qué punto es visualmente perceptible el rostro facial. Para ello, primero de todo se establecerá el tamaño de la imagen inicial, que será de 32x32, ya que basándonos en la tabla 4.1 podemos ver que la similaridad entre la imagen original y la mejorada a partir de la escala x8 comienza a decrecer mucho.

Como se observa en la figura 4.3, con una escala x8 la imagen (a) se puede acatar correctamente la estructura de la cara, por lo tanto, en la imagen (d) y (g) no hay mucha diferencia visual entre ellas. Pasando a una escala mayor, x16, comenzamos a ver como nuestro modelo mejora un poco el rostro facial de la imagen entre (e) y (h), donde la nariz, el ojo derecho y la boca tienen una mejor reconstrucción. Esto se debe a que la imagen de baja resolución (b) se puede ver que no se tiene la misma información que la escala anterior (a), pues se trata de una imagen de 16x16 transformada a un tamaño de 256x256 aplicando interpolación bicúbica y como el modelo con la función de pérdida L_2 no tiene en cuenta los componentes faciales, no se reconstruyen bien. Sin embargo, con una imagen de 8x8 como la (c), ya no se tiene prácticamente información de dónde están los componentes faciales y ello provoca que ambos modelos con funciones de pérdida distintas no consigan reconstruir la cara en (f) y (i), concluyendo así que su límite es con escala x16.

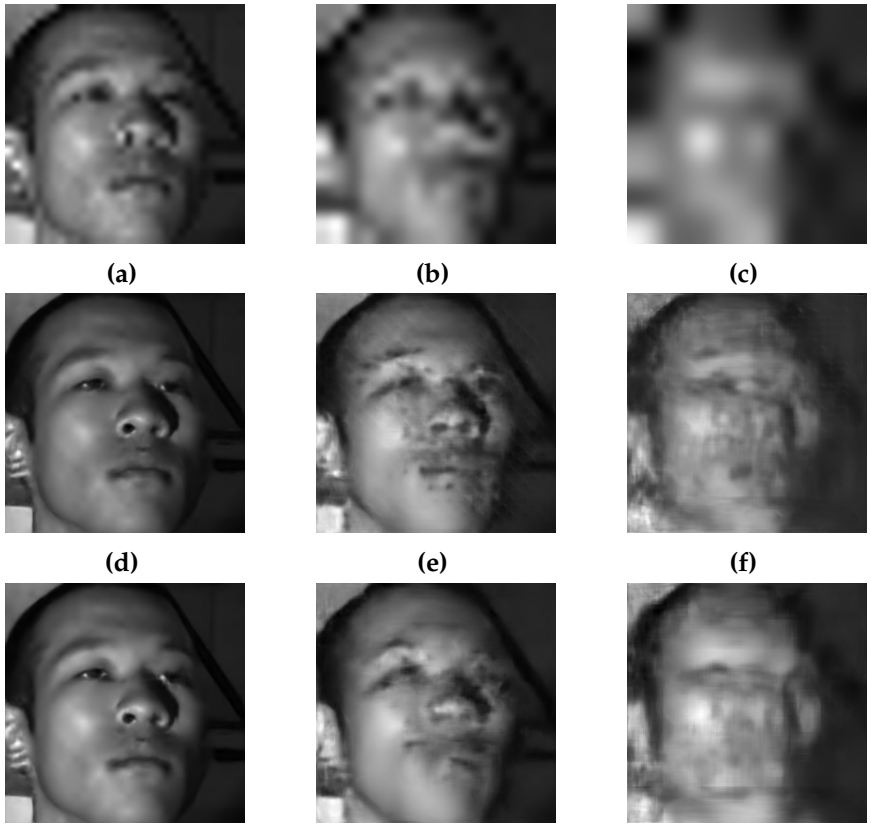

Figura 4.3: Por columnas de izquierda a derecha, imágenes con escala x8, x16 y x32 respectivamente, por filas, de arriba a abajo, imágenes de baja resolución, imagen resultante con un α de 0.9 y *baseline* con α de 1 o MSE de función de pérdida

CAPÍTULO 5
Conclusiones y trabajo futuro

5.1 Conclusiones

En este último capítulo se va resumir las conclusiones que se han llegado a extraer en el trabajo.

En el primer objetivo se ha revisado las bases de datos existentes más usadas en el ámbito de súper resolución y se ha concluido que no cumplían con los requisitos, por lo tanto se realizó una búsqueda exhaustiva de *datasets* de otros campos, llegando al dataset "Yale-B Extended" y modificándolo a nuestro antojo de forma que nos permitiese cumplir con las limitaciones expuestas.

A partir de esta base de datos quisimos realizar una serie de experimentos que pudiese implicar algo alejado a lo que se suele aplicar en el ámbito de súper resolución, con *embeddings* y reconocimiento facial, pudiendo así desarrollar con éxito los modelos y aplicando las elecciones de parámetros correctamente, cumpliendo nuestro segundo objetivo.

Sin embargo, a la hora de comparar cualitativamente se ha visto que las imágenes mejoraban cuando habían una escala baja con la función de pérdida múltiple con respecto al *baseline*, en cambio, cuanto mayor es la escala, peores resultados se obtienen con la función de pérdida múltiple, obteniendo una peor similaridad entre la imagen original y la imagen "mejorada" comparado a la función de pérdida MSE. En cuanto a la profundidad de las imágenes, con una escala baja, no hay prácticamente mucha diferencia entre la imagen original y la súper resolución obtenida, pero cuando alcanzamos escalas altas en los cuales la imagen prácticamente no se puede observar el rostro facial del humano, es cuando hace una pequeña diferencia con un modelo de súper resolución general, mejorando muy poca la calidad de los componentes faciales.

5.2 Relación del trabajo desarrollado con los estudios cursados

Para la realización de este trabajo han sido necesarios gran parte de los conocimientos impartidos en el máster, especialmente de las ramas relacionadas con el reconocimiento de formas.

Una de las asignaturas de la rama de reconocimiento de formas que ha sido realmente necesaria es la asignatura Redes Neuronales Artificiales (RNA), ya que aportaba los conocimientos teóricos y prácticos para desarrollar modelos de redes neuronales para clasificación, en este proyecto, modelos preentrenados para clasificación de imágenes.

Por otra parte, también ha sido de mucha utilidad la asignatura Visión por Computador (VPC), ha permitido el aprendizaje de conceptos y arquitecturas existentes en el ámbito de super resolución, permitiendo así una mayor comprensión a la hora de realizar una lectura de otro tipo de proyectos.

Finalmente, la asignatura de Biometría (BIOM), pues es la asignatura en la cual se introduce el papel del reconocimiento facial y los sistemas biométricos.

5.3 Trabajos futuros

Tras las conclusiones extraídas del trabajo realizado, las posibles líneas de trabajo futuro serían las siguientes:

- **Probar otro tipo de topologías**: al inicio del capítulo 2 en el apartado de *Deep learning*, se comenta los distintos tipos de arquitecturas existentes, únicamente se ha probado uno pero se podrían probar el resto para ver si el *pre-upsampling* que se ha aplicado en el proyecto afecta o no negativamente al resultado de la súper resolución.

- **Más profundidad al modelo**: Se ha visto que con el modelo que se ha llevado a cabo en el proyecto se han obtenido buenos resultados, pero sin embargo no se ha podido profundizar más en cuanto a mejoría se refiere modificando el modelo, realizando una serie de experimentos eliminando o añadiendo capas. Además, se ha observado que por culpa del uso de los "embeddings" en la función de pérdida múltiple existe un ruido en la imagen, esto se puede deber a que el número de características, 4096, se puede haber quedado corto o hay un exceso de número de neuronas.

- **Elección de métricas**: Está claro que las métricas cuantitativas no son las correctas para este tipo de proyectos, esto se puede deber a que parte de las imágenes no son rostros faciales y por lo tanto . Para ello, existe otro tipo de métricas que implican el reconocimiento facial y las métricas se miden en base al rostro facial y no a la imagen entera, son los denominados *FR-PSNR* y *FR-SSIM* [17].

- **Función de pérdida**: En este proyecto para la función de pérdida adicional se ha usado el error medio cuadrático (MSE), pero existen muchos otras funciones de pérdidas válidas para calcular la distancia existente entre dos *embeddings*, como podrían ser la de entropía cruzada, el error medio absoluto (MAE), distancias que se suelen usar como la euclídea [19], la similitud coseno [14] y otros muchas funciones de pérdida que podrían mejorar el resultado obtenido en este trabajo.

- **Verificación Facial**: Este proyecto se ha centrado en observar la calidad perceptual de las imágenes, pero también se puede probar si gracias a las nuevas imágenes mejoradas obtienen mejor o peor precisión según la distribución de pesos de cada función de pérdida.

Bibliografía

[1] Simon Baker and Takeo Kanade. Hallucinating faces. pages 83 – 88, 02 2000.

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. *CoRR*, abs/1710.08092, 2017.

[3] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. *CoRR*, abs/1711.10703, 2017.

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[5] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[7] Xiao Hu, Peirong Ma, Zhuohao Mai, Shaohu Peng, Zhao Yang, and Li Wang. Face hallucination from low quality images using definition-scalable inference. *Pattern Recognition*, 94:110–121, 2019.

[8] Kui Jiang, Zhongyuan Wang, Peng Yi, Tao Lu, Junjun Jiang, and Zixiang Xiong. Dual-path deep fusion network for face image hallucination. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2020.

[9] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Daeshik Kim. Progressive face super-resolution via attention to facial landmark. *CoRR*, abs/1908.08239, 2019.

[10] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016.

[11] Xiao Ling, Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Spectral domain-transfer learning. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 488–496, New York, NY, USA, 2008. Association for Computing Machinery.

[12] Ce Liu, Heung-Yeung Shum, and Chang-Shui Zhang. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001.

[13] Ph.D Michel Kana. Generative adversarial network (gan) for dummies - a step by step tutorial, Feb 2021.

[14] Hieu V. Nguyen and Li Bai. Cosine similarity metric learning for face verification. In Ron Kimmel, Reinhard Klette, and Akihiro Sugimoto, editors, *Computer Vision – ACCV 2010*, pages 709–720, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[15] Raymond Pearl and Lowell J Reed. On the rate of growth of the population of the united states since 1790 and its mathematical representation. *Proceedings of the National Academy of Sciences of the United States of America*, 6(6):275, 1920.

[16] Chi-Hieu Pham, Aurélien Ducournau, Ronan Fablet, and François Rousseau. Brain mri super-resolution using deep 3d convolutional networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 197–200, 2017.

[17] Abdul Rehman and Zhou Wang. Reduced-reference ssim estimation. In *2010 IEEE International Conference on Image Processing*, pages 289–292, 2010.

[18] Yibing Song, Jiawei Zhang, Shengfeng He, Linchao Bao, and Qingxiong Yang. Learning to hallucinate face images via component generation and enhancement. *CoRR*, abs/1708.00223, 2017.

[19] Liwei Wang, Yan Zhang, and Jufu Feng. On the euclidean distance of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1334–1339, 2005.

[20] Zhihao Wang, Jian Chen, and Steven C. H. Hoi. Deep learning for image super-resolution: A survey. *CoRR*, abs/1902.06068, 2019.

[21] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, Student Member, Eero P. Simoncelli, and Senior Member. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.

[22] Chih-Yuan Yang, Sifei Liu, and Ming-Hsuan Yang. Structured face hallucination. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1099–1106, 2013.

[23] Daiqin Yang, Zimeng Li, Yatong Xia, and Zhenzhong Chen. Remote sensing image super-resolution: Challenges and approaches. In *2015 IEEE International Conference on Digital Signal Processing (DSP)*, pages 196–200, 2015.

[24] Jiahui Yu, Yuchen Fan, Jianchao Yang, Ning Xu, Zhaowen Wang, Xinchao Wang, and Thomas S. Huang. Wide activation for efficient and accurate image super-resolution. *CoRR*, abs/1808.08718, 2018.

[25] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Learning face hallucination in the wild. In *AAAI*, 2015.