UNIVERSIDAD DE GRANADA

Departamento de Ciencias de la Computación e Inteligencia Artificial

Programa de Doctorado en Tecnologías de la información y la Comunicación

Aproximaciones disruptivas basadas en deep learning para identificación humana en antropología forense

Tesis Doctoral

Guillermo Gómez Trenado

Directores

Óscar Cordón García

Pablo Mesejo Santiago

Granada, 26 de julio de 2024

Editor: Universidad de Granada. Tesis Doctorales
Autor: Guillermo Gómez Trenado
ISBN: 978-84-1195-614-7
URI: https://hdl.handle.net/10481/97721

UNIVERSIDAD DE GRANADA

Aproximaciones disruptivas basadas en deep learning para identificación humana en antropología forense

MEMORIA PRESENTADA POR

Guillermo Gómez Trenado

PARA OPTAR AL GRADO DE DOCTOR

26 de julio de 2024

DIRECTORES

Óscar Cordón García

Pablo Mesejo Santiago

Departamento de Ciencias de la Computación e Inteligencia Artificial

Título en Español: Aproximaciones disruptivas basadas en deep learning para identificación humana en antropología forense

Título en Inglés: Disruptive Approaches Based on Deep Learning for Human Identification in Forensic Anthropology

Programa de doctorado: Programa de Doctorado en Tecnologías de la información y la Comunicación.

Doctorando: Guillermo Gómez Trenado

Directores: Óscar Cordón García y Pablo Mesejo Santiago.

"For to be poised against fatality, to meet adverse conditions gracefully, is more than simple endurance; it is an act of aggression, a positive triumph." - Thomas Mann

Agradecimientos

I wish to express my deepest gratitude to my directors, Pablo Mesejo Santiago and Óscar Cordón, for their invaluable guidance and for serving as exemplary role models throughout my PhD journey. I am also profoundly grateful to Stéphane Lathuilière for his mentorship during my research stay at Telecom Paris.

Additionally, I extend my heartfelt thanks to Óscar Ibañez, Rubén Martos, Rosario Guerra, María Alejandra Guativonza, Stefano De Luca, Guillermo Ramirez, Fernando Navarro, and the rest of the Panacea Cooperative Research team, as well as to the Physical Anthropology Lab at the University of Granada for their scientific support and crucial role on the data annotation and acquisition. This work would not have been possible without them.

Contents

List of Acronyms

List of Figures

List of Tables

Resumen

1. Introducción al problema

La ciencia forense es una ciencia multidisciplinar y constituye un pilar fundamental de la justicia moderna. Ésta aplica distintos métodos científicos de cara a resolver problemas legales mediante el análisis de pruebas físicas procedentes de escenas del crimen y otros contextos relevantes [HS09]. A lo largo del tiempo, este campo ha avanzado sustancialmente, integrando distintas tecnologías para mejorar la precisión y fiabilidad de las investigaciones. Una de las técnicas cruciales en la ciencia forense es la imaginería facial (*facial imaging* en inglés), que es particularmente importante para la identificación humana [SCGC19]. Esta técnica emplea procesamiento y análisis de imágenes de rostros humanos para identificar a sospechosos, víctimas y personas desaparecidas, desempeñando así un papel vital en la investigación criminal y los procedimientos legales. La continua evolución de la ciencia forense, apoyada por los nuevos avances tecnológicos, resalta su papel indispensable en el sistema de justicia.

Un subcampo de relevancia dentro de la ciencia forense es la Antropología Forense (AF), que aplica las teorías y métodos de la antropología biológica en contextos legales, particularmente aquellos que implican el análisis de restos óseos [Lar23]. Este campo es esencial para la identificación de individuos fallecidos, especialmente cuando los restos son irreconocibles debido a la descomposición o como resultado de una catástrofe. Técnicas como la comparación de registros dentales, el análisis de ADN, la aproximación facial (reconstrucción del rostro a partir de restos óseos), el fotomontaje molecular (reconstrucción del rostro a partir de restos genéticos) y la identificación craneofacial se emplean minuciosamente para determinar la identidad de los fallecidos [WR12]. Estos métodos son cruciales no solo para una muy necesitada respuesta a las familias, sino también para ayudar en las investigaciones legales determinando la identidad y causas de muerte, aspecto de una relevancia capital en los procedimientos legales.

El enfoque tradicional de la AF en el análisis de restos óseos incluye procedimientos como la estimación del perfil biológico (edad, sexo, afinidad poblacional y estatura), la aplicación de superposición craneofacial y radiografía comparativa con fines de identificación [DCI20, USK18]. Los antropólogos forenses son fundamentales en la búsqueda, recuperación y examen de restos, facilitando el proceso de identificación a través de un análisis meticuloso. Este campo no solo ayuda a resolver casos proporcionando información crucial sobre la identidad y las circunstancias de la muerte, sino que también contribuye a la comprensión de cuestiones

médico-legales mediante el estudio de características individualizantes y el análisis de traumas.

Tradicionalmente, los métodos de identificación forense se han basado en comparaciones manuales, las cuales, aunque valiosas, presentan un margen significativo de error debido a la subjetividad inherente del proceso, la propia pericia del experto e incluso a la fatiga mental y física [1]. Sin embargo, en los últimos años se ha observado un cambio hacia enfoques más avanzados, integrando tecnologías de Visión por Computador (VC) e Inteligencia Artificial (IA) para mejorar la precisión y la fiabilidad de las identificaciones. Estas tecnologías ofrecen una mayor robustez y objetividad, reduciendo los errores, acelerando los procesos y optimizando el flujo de trabajo de expertos forenses que puden dedicar su tiempo a aquellas áreas críticas donde su labor es más importante [MMI+20].

En los últimos años, el ámbito de aplicación y conocimiento de la AF se ha expandido para incluir la identificación de individuos vivos, reflejando su adaptabilidad a los cambios en los métodos de investigación criminal, las nuevas oportunidades y los nuevos desafíos, como la interpretación de grabaciones de vigilancia, la estimación de la edad en migrantes o la identificación de personas desaparecidas [GORT+16, SDR+16]. Técnicas como la modificación de la edad en fotos y la comparación facial forense se emplean ahora para emparejar individuos con grabaciones de vigilancia y otras muestras fotográficas, mejorando las capacidades del campo. La integración de IA, en contreto métodos basados en Aprendizaje Automático (AA), en los sistemas de reconocimiento facial ha mejorado aún más la precisión y eficiencia de la identificación de personas en videovigilancia tanto en tiempo real como en grabaciones [RN20, SO21, KKSK20]. Estos avances subrayan el potencial del campo para adaptarse a nuevas tecnologías, enfatizando la importancia de la innovación en la justicia y la seguridad pública.

A pesar de estos avances, los métodos de *facial imaging* forense, como la identificación facial y la modificación de la edad en imágenes, siguen limitados por prácticas manuales y laboriosas [MIM24, Wil15, DV15, WR12]. Estos métodos dependen del análisis visual realizado por los profesionales, lo que introduce posibles errores y una alta subjetividad asociada a la habilidad y experiencia individuales. La falta de fiabilidad, subjetividad, ausencia de metodologías estandarizadas y el tiempo significativo requerido por caso destacan la necesidad de innovación dentro de la AF [COG+17, USK18]. Abordar estas limitaciones es crucial para mejorar la utilidad y escalabilidad de estos métodos.

Asegurar la fiabilidad en la ciencia forense es esencial para que las pruebas tengan valor en los contextos legales. Esto requiere obtener resultados que sean consistentes, precisos y validados científicamente [BWW20]. El AA ofrece una vía prometedora para mejorar las prácticas forenses, incluida la *facial imaging*, al minimizar el error humano y la subjetividad. Los algoritmos pueden aprender de grandes conjuntos de datos, analizar características faciales con precisión, y automatizar y estandarizar el

[1] En este ejemplo, la antropóloga forense Josefina Lamas reconoció que se equivocó inicialmente al dictaminar que los huesos del caso Bretón eran de animales, rectificando meses después al confirmar que eran humanos tras el informe de otro experto forense. `https://www.europapress.es/epsocial/infancia/noticia-forense-dictamino-huesos-caso-breton-eran-animales-reconoce-equivoco-20130703153339.html`. Accedido por última vez el 11 de julio de 2024.

proceso de análisis para apoyar la toma de decisiones objetiva en las investigaciones forenses [CAIN+14b, MMI+20]. La imperativa global para una identificación precisa, impulsada por la necesidad de abordar la desaparición de personas, la identificación de víctimas de desastres y la resolución de crímenes, subraya la importancia de integrar técnicas de IA en la *facial imaging* forense.

2. Desarrollo realizado

El Deep Learning (DL) es una subrama del AA que utiliza redes neuronales profundas para modelar y resolver problemas complejos a partir de grandes cantidades de datos. La importancia del DL se debe a su éxito impulsando innovaciones y resolviendo problemas que anteriormente se consideraban irresolubles [GBC16, Chapter 1]. El trabajo desarrollado a lo largo de esta tesis doctoral se divide en tres bloques principales que abordan los objetivos planteados en la misma, utilizando métodos basados en DL:

1. El desarrollo de un modelo para la localización automática de puntos cefalométricos basado en técnicas de DL que asista en las tareas de identificación forense.
2. La creación de un método de envejecimiento facial controlado y preciso mediante DL, que permita modificar la edad de las personas en fotografías capturando la complejidad y variabilidad del fenómeno.
3. La propuesta de un marco versátil de edición automática de imágenes basado en descripciones textuales utilizando DL, que permitiría a los expertos forenses manipular imágenes según las necesidades descriptivas de la investigación.

Las siguientes secciones resumen nuestra contribución para cada uno de los bloques. Primero, se define el diseño e implementación de un sistema de DL que automatiza la localización de puntos cefalométricos, mejorando la eficiencia y precisión del análisis forense. Luego, se describe un modelo de DL para la edición de la edad en imágenes faciales preservando sus rasgos de identidad. Finalmente, se introduce un modelo de edición de imágenes basado en intrucciones textuales, que permite realizar modificaciones en imágenes a partir de sencillas descripciones de texto, tales como alterar atributos faciales, facilitando potencialmente la identificación y análisis forense.

Desarrollo de un modelo en cascada de redes convolucionales para la localización de puntos cefalométricos

La localización de puntos cefalométricos es una parte importante en la antropología física y forense [SCGC19, DCIn+11, CAIN+14a, CÁICÁ+15], permitiendo la caracterización precisa de la morfología de la cabeza mediante puntos concretos en la cara que guardan correspondencia con ubicaciones en el cráneo. Estos puntos son esenciales para tareas como la extracción de índices antropométricos [MVIA18],

la detección de patologías [Far94] y la identificación humana [HIWK15]. Tradicionalmente, este proceso se ha realizado de forma manual, lo que lo hace propenso a errores y consume mucho tiempo, especialmente en escenarios de identificación masiva de víctimas.

El problema de la localización de puntos cefalométricos comparte similitudes con la localización de puntos faciales estudiada en VC [WJ19], pero presenta desafíos únicos debido a las necesidades específicas de la AF. Los métodos existentes para puntos faciales no son aplicables directamente debido a la inconsistencia anatómica y la baja resolución de las anotaciones. Además, la variabilidad en las imágenes forenses, como la diferencia en la pose de la cabeza y la diversa calidad y origen de las fotos, añade complejidad al problema.

La automatización de este proceso utilizando técnicas avanzadas de DL, como las redes neuronales convolucionales, puede mejorar significativamente la precisión y eficiencia. Estos métodos, que incluyen enfoques basados en mapas de calor [WSC+20, SZJ+19] y modelos de máscara deformable en 3D [ZLL+16, FWS+18, GZY+20], pueden producir resultados más robustos y repetibles, acelerando el análisis y reduciendo la dependencia de la intervención manual. La implementación de estos métodos en contextos forenses promete una mejora considerable en la identificación y comparación facial, proporcionando herramientas poderosas para los expertos forenses.

El trabajo desarrollado se centra en la creación de un modelo en cascada de redes convolucionales, denominado *FSCNet*, para la localización automática de puntos cefalométricos en imágenes faciales. Este modelo está diseñado para mejorar la precisión y la eficiencia en el proceso de identificación forense. A continuación, se detallan los aspectos clave del desarrollo y evaluación del modelo.

FSCNet se compone de dos módulos principales. Primero, se utiliza un modelo de máscara 3D deformable preentrenado, *3DDFA v2* [GZY+20], para sugerir una localización inicial confiable de los puntos cefalométricos. Luego, se procesa una imagen recortada alrededor de cada punto sugerido a través de una red convolucional (ResNet-18) [HZRS16] entrenada para predecir el desplazamiento entre el centro de la imagen recortada y la localización del punto. Este enfoque permite aumentar significativamente la resolución del modelo y mejorar la precisión de la localización.

Se utilizan dos conjuntos de datos: un dataset de entrenamiento que incluye tanto imágenes de casos reales como fotografías tomadas en condiciones controladas, y un dataset de validación que contiene únicamente imágenes de casos forenses reales. El dataset de entrenamiento incluye 165 imágenes de diferentes sujetos, con hasta 30 puntos anotados por imagen. En total, se anotaron 3526 puntos individuales. Las imágenes varían en calidad y resolución, presentando desafíos adicionales para la precisión del modelo.

El proceso de localización se divide en varios pasos. En el primer paso, se utiliza el modelo *3DDFA v2* para obtener una máscara facial con aproximadamente 40,000 coordenadas 3D, identificando el mejor punto de malla para cada punto cefalométrico correspondiente. En el segundo paso, se optimizan los puntos ubicados fuera de la máscara a través de la optimización de transformaciones homogéneas mediante algoritmos de evolución diferencial [SP97]. En el tercer paso, se recortan imágenes de las regiones de interés alrededor de cada punto y se procesan con una

red convolucional para refinar la localización.

Se realiza un estudio de ablación para evaluar el impacto incremental de cada decisión arquitectónica en el rendimiento del modelo. Cada paso muestra un aumento significativo en el rendimiento, con la proyección de etiquetas [MK18] como el enfoque más efectivo para incorporar información de etiquetas en el modelo. Además, *FSCNet* se compara con tres métodos del estado del arte disponibles para localización de puntos faciales: *3FabRec* [BW20], *HRNET* [WSC+20] y *LUVLi* [KMM+20]. *FSCNet* supera significativamente a estos modelos en términos de error, mostrando una precisión superior en la localización de puntos cefalométricos. Los métodos basados en mapas de calor, como *HRNET* y *LUVLi*, son menos efectivos debido a sus limitaciones de resolución. Finalmente, se lleva a cabo un estudio con seis expertos forenses donde comparan las ubicaciones de los puntos cefalométricos generadas por *FSCNet* con las anotaciones de otros expertos. En la mitad o más de los casos, *FSCNet* obtiene un rendimiento igual o superior al de los expertos humanos. Esto demuestra la robustez y precisión del modelo en un entorno forense real.

De forma adicional a la localización, se resolvió la tarea de determinar la visibilidad de los puntos cefalométricos. Esta estimación de la visibilidad es un componente crucial del modelo. A pesar de no haber un criterio consistente en el conjunto de entrenamiento para la determinación de la visibilidad de un punto, se logró una precisión promedio del 83%, utilizando un criterio de visibilidad basado en la información geométrica de la malla 3D inicial producida por *3DDFA v2*.

En resumen, *FSCNet* ofrece una solución robusta y precisa para la localización automática de puntos cefalométricos, mejorando significativamente la eficiencia y exactitud en aplicaciones forenses.

Desarrollo de un modelo para la edición de la edad en imágenes faciales

La edición de la edad en imágenes faciales [FGH10, KSSS14, WCY+16] es una técnica crucial en varias aplicaciones, como la producción cinematográfica y la aproximación facial forense, permitiendo alterar automáticamente la edad en las imágenes faciales mientras se preserva la identidad. Los enfoques recientes de DL utilizan arquitecturas tipo codificador-decodificador para proyectar las imágenes a un espacio latente, manipular el contenido y decodificar la imagen alterada [ABD17, HKSC19, MHP21, OESF+20, WCY+16, WTLG18, YPN+21, ZSQ17]. Sin embargo, estos métodos muestran problemas al manejar diferencias significativas de edad y cambios en la forma facial, debido a que las transformaciones de edad para saltos significativos suelen fallar al no considerar adecuadamente las modificaciones en la estructura facial.

El trabajo desarrollado propone una solución novedosa que permite realizar cambios estructurales profundos en las transformaciones faciales, logrando una transformación realista de la imagen con diferencias de edad que implican cambios en la forma de la cabeza. Este marco permite al usuario ajustar el grado de preservación de la estructura en el momento de la inferencia, proporcionando diversas transformaciones donde la estructura se preserva en diferentes grados. Para este propósito,

se introduce el módulo *CUSP* (*CUstom Structure Preservation*), que identifica las regiones relevantes de la imagen de entrada que deben preservarse y aquellas donde los detalles no son relevantes para la tarea.

El método propuesto emplea una arquitectura codificador-decodificador basada en estilo [HLBK18, PZW+20, KLA19, KLA+20]. El proceso de transformación se basa en una arquitectura codificador-decodificador que separa el estilo y el contenido de la imagen de entrada. El decodificador o generador combina estas representaciones con la edad objetivo, mientras que el módulo *CUSP* permite ajustar el nivel de preservación de la estructura mediante máscaras de difuminado aplicadas a las Skip connection (SC) entre el codificador y el decodificador.

A nivel experimental, se evaluó el rendimiento de CUSP en múltiples tareas de edición de la edad, empleando métricas de preservación del contenido de la imagen y de precisión en el envejecimiento y rejuvenecimiento del rostro, sobre tres conjuntos de datos públicos: *FFHQ-RR* [KLA19, YPN+21], *FFHQ-LS* [OESF+20] y *CelebA-HQ* [KALL17, LLWT15]. Junto con ello, se realizaron estudios de ablación para evaluar el impacto de las decisiones arquitectónicas en el rendimiento del modelo. Los resultados mostraron que la utilización de un codificador independiente para el estilo y el uso del módulo *CUSP* mejoraron significativamente la precisión de la transformación de edad y la preservación de detalles. Además, el modelo se comparó con métodos del estado del arte, como *HRFAE* [YPN+21] y *LATS* [OESF+20]. *CUSP* demostró un rendimiento superior en la transformación de edad y la preservación de detalles, permitiendo transformaciones más profundas y realistas que los métodos existentes. Además, ofrece al usuario la capacidad de controlar el grado de preservación de la estructura. De forma adicional, se llevó a cabo un estudio con 80 usuarios que compararon *CUSP* con *HRFAE* y *LATS*. Los usuarios prefirieron las transformaciones generadas por *CUSP* en términos de precisión de edad, preservación de identidad, realismo y naturalidad del progreso de envejecimiento.

En resumen, *CUSP* ofrece una solución innovadora y flexible para la edición de edad facial, mejorando significativamente la precisión y realismo de las transformaciones, y proporcionando una herramienta adaptativa para aplicaciones forenses y de entretenimiento.

Desarrollo de un modelo de uso general para la edición de imágenes basado en texto

La edición de imágenes basada en texto es una técnica emergente en VC que modifica imágenes a partir de descripciones en lenguaje natural a través de redes neuronales avanzadas, generalmente modelos de difusión [HJA20, DN21]. Esta técnica permite a los usuarios realizar cambios en las imágenes simplemente describiéndolos en texto, como "hacer el cielo más azul" o "pintar el coche de rojo", ofreciendo un marco versátil para diversas manipulaciones guiadas por texto. Sin embargo, desafíos como el gran coste computacional del proceso de inversión en los modelos de difusión permanecen sin resolver.

El trabajo desarrollado propone una técnica novedosa denominada *SAGE* (*Self-Attention Guidance for Image Editing*), que equilibra la eficiencia computacional con la reconstrucción de alta fidelidad, permitiendo capacidades de edición versátiles.

De forma similar a otros métodos, nuestra aproximación utiliza la inversión *DDIM* (*Denoising Diffusion Implicit Models*) [SME21]. Sin embargo, nuestra contribución radica en la utilización de los mapas de auto-atención y atención-cruzada que el modelo de difusión computa internamente durante el proceso de inversión *DDIM*, permitiendo una reconstrucción precisa con un esfuerzo computacional menor.

SAGE emplea un modelo de difusión preentrenado para generación de imágenes basado en texto junto a nuestras contribuciones para edición. Este marco permite la reconstrucción alineada con la imagen de entrada mientras se logran modificaciones profundas basadas en indicaciones textuales. Introducimos un término novedoso como función de coste que guía la reconstrucción y edición modificando la dirección del gradiente descendente. Esta función de coste guía la salida de las capas de auto-atención de la red de difusión, asegurando una reconstrucción de alta fidelidad en las regiones no afectadas por el proceso de edición, sin aumentar significativamente las demandas computacionales.

La evaluación se basa en el conjunto de datos PieBench [JZB+23], que incluye 700 imágenes divididas equitativamente entre escenas naturales y artificiales, distribuidas en cuatro categorías: animales, humanos, interiores y exteriores. Estas imágenes, a su vez, se clasifican en diez tareas distintas: modificación de objetos, adición de objetos, eliminación de objetos, alteración de contenido, ajuste de pose, modificación de color, cambio de material, alteración de fondo y cambio de estilo. Además, utilizamos una colección de imágenes de alta resolución para evaluaciones cualitativas.

En este estudio, abordamos el desafío de la edición de imágenes basada en indicaciones de texto tal como se introdujo en [HMT+23, MHA+23]. Es decir, el usuario proporciona una imagen inicial junto con dos cadenas de texto: una descripción textual de entrada $\mathcal{P}^{in}$ de dicha imagen así como una indicación objetivo $\mathcal{P}^{out}$ que describe el resultado deseado después de la edición. El proceso de edición comienza con la inversión *DDIM* de la imagen de entrada utilizando su indicación asociada. Esta inversión genera el ruido latente estimado que sirve como punto de partida para el proceso de muestreo *DDIM* responsable de crear la imagen editada. Dentro de este marco, la U-Net [RFB15] procesa las indicaciones por separado. Para calcular el término que guía la generación, se comparan los mapas de auto-atención de la inversión *DDIM* y los mapas de auto-atención producidos en la nueva generación con $\mathcal{P}^{out}$.

Se realizaron estudios de ablación con el objetivo de evaluar los mecanismos utilizados para lograr la reconstrucción en regiones destinadas a ser preservadas durante el proceso de edición. Los resultados mostraron que el uso del gradiente sobre los mapas de atención supera consistentemente al reemplazo de estos mapas como mecanismo de guiado de la edición a lo largo de todas las métricas, demostrando la eficacia de nuestro mecanismo propuesto sobre los mapas de auto-atención. Además, *SAGE* se comparó con métodos de última generación como *Negative Prompt Inversion* [MIST23] y *Direct Inversion* [JZB+23]. Nuestro método mostró un rendimiento superior tanto en la preservación de la estructura como en la calidad de la edición, manteniendo una eficiencia computacional comparable o superior. Finalmente, se llevó a cabo un estudio con 22 participantes que compararon *SAGE* con otros métodos en tres aspectos clave: preservación de la estructura, preservación del fondo y adherencia a la indicación de texto, así como una valoración sobre la preferencia

general del usuario. Los resultados mostraron una preferencia consistente por nuestro método, destacando su ventaja significativa en la preservación del fondo y la estructura.

En conclusión, *SAGE* ofrece una solución innovadora y eficiente para la edición de imágenes guiada por texto, mejorando significativamente la precisión y el realismo de las transformaciones, y proporcionando una herramienta poderosa para aplicaciones en VC.

3. Conclusiones y trabajos futuros

Esta tesis ha presentado tres avances significativos en el campo de *facial imaging* a través del desarrollo y la aplicación de metodologías basadas en DL. Cada una de estas contribuciones aborda una solución de interés para el análisis forense, proporcionando soluciones innovadoras que pueden mejorar la precisión, eficiencia y fiabilidad en las investigaciones forenses. Primero, desarrollamos una herramienta robusta para la localización precisa de puntos cefalométricos en imágenes faciales, superando las limitaciones de un conjunto de datos pequeño mediante el uso de modelos preentrenados de detección de puntos faciales y refinando esta solución a través de una red convolucional condicional compartida por todos los puntos cefalométricos. Segundo, introdujimos una arquitectura novedosa para la edición de la edad en imágenes faciales capaz de producir modificaciones estructurales preservando detalles relevantes de la imagen original, validada contra el estado del arte en tres conjuntos de datos. Finalmente, revisamos la edición de imágenes basada en indicaciones de texto mediante modelos de difusión, demostrando que el proceso de inversión *DDIM* contiene información suficiente para la edición sin necesidad de reconstruir la imagen original, introduciendo y validando la utilización de mapas de auto-atención como un mecanismo superior para tareas de edición de imágenes.

El trabajo futuro para esta investigación implica varias áreas clave de desarrollo y mejora para aumentar el rendimiento y la aplicabilidad de los métodos propuestos, así como su validación en contextos forenses en una colaboración multidisciplinar. Para el método de localización de puntos cefalométricos, evaluaremos distintas soluciones basadas en AA para estimar la visibilidad de los puntos, para esto es esencial la creación de un conjunto de datos más amplio y robusto con criterios claros y consistentes para determinar la visibilidad. En el área de envejecimiento facial, se planea extender el módulo *CUSP* para aplicar los beneficios de la preservación estructural a otras tareas de edición de imágenes. Para la edición de imágenes, el trabajo futuro continuará explorando nuevas técnicas y metodologías para reducir la carga computacional del cálculo del gradiente que guía la generación sin sacrificar la calidad de las transformaciones.

Abstract

Forensic science, a multidisciplinary field crucial to modern justice, applies scientific methods to analyze physical evidence from crime scenes and other contexts. Among its vital techniques is facial imaging, essential for human identification. This technique aids in recognizing suspects, victims, and missing persons through advanced image processing and analysis, significantly impacting criminal investigations and legal procedures. Forensic Anthropology (FA), a subfield of biological anthropology, focuses on the analysis of skeletal remains to identify deceased individuals, especially when remains are unrecognizable due to decomposition or as a consequence of catastrophic events. Techniques like dental record comparison, DNA analysis, facial approximation, molecular photofitting, and craniofacial identification are meticulously used to determine the identity of the deceased. These methods are essential not only for providing closure to families but also for aiding legal investigations by confirming identities and causes of death. Recently, FA has expanded to include the identification of living individuals, adapting to changes in criminal investigation methods and new challenges, such as interpreting surveillance footage and identifying missing persons. The integration of Artificial Intelligence (AI) and Machine Learning (ML) in facial recognition systems has further enhanced the precision and efficiency in identifying individuals, underscoring the importance of innovation in justice and public safety.

Traditionally, forensic identification methods relied on manual comparisons, which, despite being valuable, had a significant margin of error due to the subjective nature of the process, the expert's skill, and even mental and physical fatigue. However, recent years have seen a shift towards more advanced approaches, integrating Computer Vision and AI technologies to improve the accuracy and reliability of identifications. These technologies enhance robustness and objectivity, reducing errors, speeding up processes, and optimizing forensic experts' workflow, allowing them to focus on critical areas where their work is most crucial.

Ensuring reliability in forensic science is essential for the evidentiary value of proofs in legal contexts, which depends on producing consistent, accurate, and scientifically validated results. Traditional forensic facial imaging methods, such as facial identification and age progression, remain limited by labor-intensive, manual practices that rely on the visual analysis of professionals, introducing potential human errors and high subjectivity. The lack of standardized methodologies and the significant time required per case highlight the need for innovation within FA. Artificial Intelligence (AI) offers a promising avenue for improving forensic practices by minimizing human error and subjectivity. Algorithms can learn from vast datasets, analyze facial features with precision, and automate and standardize the analysis process, supporting objective decision-making in forensic investigations. The global

imperative for precise identification, driven by the need to identify missing persons, disaster victims, and solve crimes, underscores the importance of integrating AI solutions in forensic facial imaging.

Deep Learning (DL) is a subfield of machine learning that uses deep neural networks to model and solve complex problems from large datasets. DL's current relevance and popularity stem from its success in driving innovations and solving previously intractable problems.This doctoral dissertation introduces three significant contributions based on DL that can be integrated in the field of forensic facial imaging. The first contribution is the development of FSCNet, a DL model designed for the automatic localization of cephalometric landmarks (precise locations on the head highly relevant for many FA tasks) in facial images, which can help enhancing the efficiency and accuracy of forensic identification processes. FSCNet employs a cascade of convolutional networks, starting with a pre-trained 3D deformable mask model to provide initial landmark location. This is followed by a convolutional neural network that refines these locations by predicting the displacement between the center of the cropped image and the landmark. Through rigorous testing and validation, FSCNet demonstrated superior performance compared to state-of-the-art methods, achieving higher precision in landmark localization and often outperforming human experts in real forensic scenarios. This development addresses the challenges posed by manual and labor-intensive landmark localization, offering a more reliable and efficient solution for forensic applications.

The second major contribution of this thesis is the introduction of a novel framework for facial age editing, the Custom Structure Preservation (CUSP) module. This framework leverages a style-based encoder-decoder architecture, inspired by advancements in image-to-image translation and unconditional image generation, to allow realistic age transformations in facial images while preserving key identity features. The CUSP module provides users with the ability to adjust the degree of structure preservation during the transformation process, enabling more profound changes in facial morphology such as head shape and hair growth, which are typically challenging for conventional methods. By the use of the CUSP module, which is able to differentiate which parts of the image should be edited and which should remain untouched, this framework achieves a higher level of realism and accuracy in age progression and regression tasks. Extensive evaluations demonstrated its superior performance in maintaining the balance between structural changes and identity preservation, making it a significant advancement in forensic facial approximation.

The third contribution is the development of SAGE (Self-Attention Guidance for Image Editing), an innovative technique for text-guided image editing that balances computational efficiency with high-fidelity reconstruction. SAGE utilizes a pre-trained diffusion model, specifically leveraging the intermediate Self-Attention (SA) and Cross-Attention (CA) maps computed during the reverse Denoising Diffusion Implicit Model (DDIM) process. This approach allows for precise image modifications based on textual descriptions without the need for explicit reconstruction of the input image. SAGE's unique SA guidance mechanism ensures faithful image editing and high-fidelity reconstruction in regions unaffected by the edits, providing an optimal balance between maintaining original image details and achieving the desired modifications. Comparative analyses against State-of-the-art (SOTA) methods showed that SAGE delivers comparable or superior editing quality with

minimal computational expense, making it a versatile and powerful tool for various forensic and general image editing applications.

In addition to the technological advancements presented, it is crucial to ensure that these methods are rigorously validated to guarantee their effectiveness and reliability in real-world applications. Each method introduced in this dissertation has undergone comprehensive user studies. FSCNet for cephalometric point localization was validated with the involvement of forensic experts, who assessed its accuracy, usability, and practical implications in a forensic scenario. For the CUSP module for facial age editing and the SAGE technique for text-guided image editing, diverse user groups were involved in the validation process to compare the methods' preferences and performance against existing techniques. These studies help ensure that the developed methods not only perform well in controlled environments but also meet the demands of actual forensic investigations and general image editing tasks. This user-centric approach underlines the commitment to advancing forensic science through methods that are both scientifically robust and practically applicable, ultimately enhancing the overall credibility and impact of forensic analyses.

In conclusion, this dissertation presents groundbreaking advancements in methodologies that can be used in forensic facial imaging through the development of FSCNet, the CUSP module for facial age editing, and the SAGE technique for text-guided image editing. These contributions could significantly enhance the accuracy, efficiency, and reliability of forensic analyses, addressing critical challenges in the field.

Chapter I

Introduction

"I first get a title and then I write a script for the title." — Aki Kaurismäki

Forensic science, a vital tool in the quest for justice, is the application of a broad spectrum of sciences to answer questions relevant to a legal system [HS09]. This multidisciplinary field plays an indispensable role in supporting the legal system through the meticulous analysis of physical evidence collected not only from crime scenes but also from various contexts requiring legal scrutiny. Over the years, forensic science has evolved dramatically, incorporating advanced technologies and methodologies to increase the accuracy and reliability of its results. Among these, facial imaging has emerged as a pivotal technique, especially in cases involving human identification [SCGC19]. Leveraging sophisticated image processing and analysis methods, forensic facial imaging assists in identifying suspects, victims, and missing persons, profoundly impacting criminal investigations and legal proceedings. This integration of technology not only enhances the capabilities of forensic experts but also underscores the evolving nature of forensic science in the modern justice system.

Forensic Anthropology (FA) is a subfield of biological anthropology [Lar23] that applies its theory and methods to the legal context, traditionally those related to the recovery, analysis and identification of the skeleton [CPB14]. When it comes to identifying the deceased, especially in cases of unrecognizable remains due to decomposition or as a consequence of catastrophic events, forensic experts rely on various techniques. These methods include dental records comparison, DNA analysis, facial approximation, molecular photofitting and Craniofacial Identification (CI) [WR12]. The process is meticulous, involving the careful gathering and analysis of evidence to reconstruct the identity of the deceased. This aspect of forensic science is crucial, not only for bringing closure to families of the missing but also for aiding legal investigations by confirming identities and causes of death. The accuracy and reliability of these identification methods have a significant impact on subsequent legal proceedings, making them an indispensable part of the forensic process.

The application of skeletal analysis aids in the identification of human remains, particularly in complex scenarios like those involving skeletonized, burned, or degraded conditions. Classic procedures within FA include biological profile estimation (determining age, sex, ancestry, and stature from skeletal remains) as well as the

application of craniofacial superimposition and comparative radiography for identification purposes [DCI20, USK18]. Forensic anthropologists play a crucial role in both the search and recovery of remains and the meticulous examination required to match ante-mortem and post-mortem materials, facilitating the identification process. This specialized field not only helps in solving cases by providing crucial information on the identity and circumstances of death but also contributes to the broader understanding of medicolegal issues through the study of individualizing features and trauma analysis.

In recent years, the scope of the field has notably broadened, incorporating the identification of living individuals alongside traditional practices [GORT+16]. This evolution responds to changing crime prosecution methods and the introduction of novel challenges, such as interpreting surveillance footage [SCGC19], estimating ages in migration contexts [SDR+16] or identifying missing people [USK18]. The field now employs a diverse array of techniques tailored to the living, including age progression (a method that applies forensic art and science to predict the current appearance of an individual who has been missing for a period of time) for locating long-term missing persons and Forensic Facial Comparison (FFC) to match individuals with surveillance footage and other photographic samples [SCGC19, WR12]. Furthermore, the integration of cutting-edge technologies such as Artificial Intelligence (AI) [RN20] -and more specifically Machine Learning (ML) [Bis06, Mur22]- into facial recognition systems has significantly enhanced the capability to identify persons of interest in real-time surveillance and post-event analysis outside the forensic context [SO21, KKSK20]. These developments reflect the field's potential for adaptation to the evolving landscape of new technologies, highlighting the importance of leveraging innovation for the purposes of justice and public safety. This adaptation into forensic living identification not only exemplifies FA's versatility but also its crucial role in addressing contemporary forensic challenges.

Despite the revolutionary changes witnessed in fields like medicine (with AI-driven diagnostics and treatments) or in forensic sciences (like genetics and crime scene investigation), forensic facial imaging methods like facial identification and age progression remain constrained by manual, time-intensive practices [MIM24, Wil15, DV15, WR12]. These tasks involve a practitioner-dependent visual analysis of comparative samples, a process full of potential errors and high subjectivity due to its reliance on individual skill and experience. The inherent limitations —lack of reliability, subjective nature of analyses, absence of methodological proposals for facial identification, and significant time requirements per case— severely restrict the utility and scalability of these methods. Consequently, these methods often serve only as secondary or negative identifiers, narrowing down possible matches without providing definitive identification, underscoring a critical area ripe for innovation and improvement within FA [COG+17, USK18].

Reliability within forensic science is paramount for ensuring the probative value of evidence presented in legal contexts. This critical aspect hinges on the ability to produce consistent, accurate, and scientifically validated results that can withstand legal scrutiny [BWW20]. ML methods offer a compelling avenue for enhancing the reliability of forensic practices, including facial imaging. By harnessing the power of algorithms to learn from vast datasets, ML can minimize human error and subjectivity, which traditionally plague manual identification techniques [CAIN+14b].

In forensic facial imaging, for instance, ML algorithms can analyze facial features with remarkable precision, identifying unique patterns and correlations that may be hardly noticeable to the human eye. These technologies can automate and standardize the analysis process, providing reproducible results that support objective decision-making in forensic investigations [MMI+20].

The identification of individuals, whether deceased or living, is increasingly becoming a global imperative, driven by the need to address issues such as missing persons, disaster victim identification, and crime solving. The integration of AI technologies into forensic facial imaging not only bolsters the reliability of identifications but also significantly contributes to the field's evolving body of knowledge, setting a new standard for evidence evaluation in the justice system. This point is underscored by the challenges faced in instances where facial recognition relies on subjective human interpretation, particularly in emotionally charged situations. For example, during the aftermath of the 2004 Tsunami and the Bali bombing on 12 October 2002, there were notable instances of misidentification, where 10% of the Tsunami victims and 50% of the Bali bombing victims were wrongly identified by their relatives through facial recognition [LGH03]. These examples highlight the limitations of relying solely on human recognition, further emphasizing the necessity for AI-enhanced methodologies that offer a higher degree of reliability and objectivity in the identification process, reducing the likelihood of such tragic errors [Rob18]. The risk of error and lack of reproducibility in manual identification methods also applies to the performance of forensic experts. For example, the task of annotating landmarks in photographs, a fundamental step in various forensic facial imaging applications [Far94], illustrates another layer of complexity and potential for error. Performed by forensic experts, this process is inherently susceptible to inter- and intra-expert variability. Such variability manifests not only in the positioning of the landmarks but also in determining their visibility [CAIN+14b]. This variability introduces a significant source of error and inconsistency, which can adversely affect the accuracy of subsequent analyses and identifications. In this context, bringing new AI developments into FA, particularly through advancements in identification methods, is positioned as a promising solution. The integration of innovative facial imaging technologies and methodologies offers unprecedented opportunities for accurate and efficient identification [MMI+20]. These ML-boosted technologies can process and analyze vast amounts of data with greater speed and accuracy than traditional manual methods, significantly improving the efficiency and effectiveness of the identification process. This capability is particularly crucial in situations requiring rapid identification, such as natural disasters or mass casualty events, where timely and accurate identification can aid in the swift resolution of cases and provide much-needed closure to affected families.

Moreover, the accessibility of these methods in developing countries highlights their practical value, offering cost-effective alternatives to more expensive techniques like DNA analysis [MMI+20]. CI and FFC require less financial investment and technical infrastructure, making them particularly appealing for regions with limited resources. These methods leverage photographs, which have become a ubiquitous resource over the last century, further underlining the importance of facial imaging techniques. The widespread availability of photographic evidence, from personal identification documents to social media, enhances the applicability of facial imaging methods in various contexts, from criminal investigations to humanitarian ef-

forts. Consequently, the evolution of FA to include these accessible and cost-efficient identification methods meets the growing global demand for identification, offering scalable, reliable solutions that support not only legal investigations but also global security and humanitarian action, especially in settings where resources are scarce.

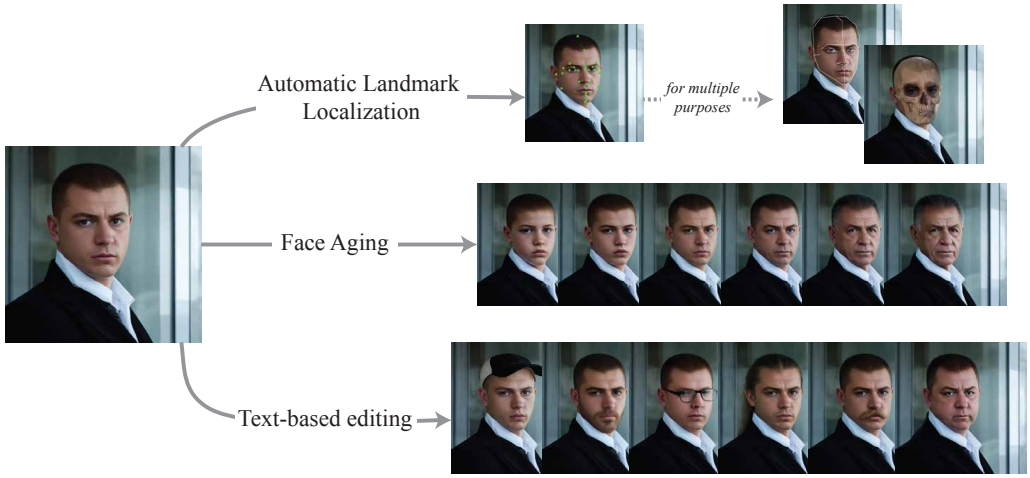


Figure 1: Illustration of the three main contributions of this PhD dissertation: automatic cephalometric landmark localization, face aging, and text-based image editing. These contributions aim to enhance the accuracy, efficiency, and reliability of forensic facial imaging through the development and application of advanced DL technologies.

Deep Learning (DL) is a subfield of machine learning that leverages deep neural networks to model and solve complex problems using large datasets. The significance of DL lies in its ability to drive innovations and tackle problems that were previously considered unsolvable [GBC16, Chapter 1]. This dissertation contributes significantly to the field of **forensic facial imaging** through the development and application of advanced **DL technologies** in this area. This contributions are illustrated in Fig. 1 and include the following:

- The first major contribution is the creation of a DL model for the **automatic localization of cephalometric landmarks** (points located on the face and head that enable forensic experts to describe the morphology of the head [Geo07]). This model represents a crucial advancement in facial imaging, as it provides an efficient and accurate method for identifying specific anthropometric points on the face that are vital for various applications, including CI and FFC [Far94]. By automating this process, the model facilitates quicker and more reliable analyses, thereby enhancing the overall efficacy of facial imaging methods used in forensic investigations.

- The second contribution addresses the challenge of **aging in facial photographs**. An area of particular interest and complexity within FA. The developed DL method enables the modification of a person's age in photographs with an unprecedented level of control and flexibility. This method allows users to adjust the degree of structural preservation during the aging process, acknowledging the fact that individuals age differently. Such flexibility is crucial for creating age-progressed images that accurately reflect the possible

appearance of missing persons or wanted individuals, improving the chances of identification in real-world scenarios.

- Lastly, the thesis dissertation extends beyond age modification to introduce a versatile DL approach for **editing arbitrary elements of pictures**, including but not limited to a wide range of facial features. This method capitalizes on the potential of foundational models, enabling not just the modification of age but also other attributes in face images, such as facial expressions, hair color, and even the presence of accessories. The utility of this approach in forensic contexts is vast. It offers unparalleled advantages in improving the accuracy and realism of suspects features modification, eyewitness composites, and facial depiction [SCGC19]. By enabling forensic experts to modify and fine-tune facial features with a high degree of precision, this technology significantly strengthen the reliability of identifications. Remarkably, the technology's adaptability allows it to edit features in any photograph, not limited to human subjects, demonstrating a broader application beyond forensic contexts.

The progressive increase in the complexity and scope of these methods mirrors the rapid and extraordinary advancements in DL over the past four years, highlighting the transformative impact of these technologies on forensic facial imaging and beyond. This development not only showcases the practical applications of DL in enhancing forensic methodologies but also sets the stage for future innovations that could further revolutionize the field.

I.1 Justification

This justification section explores three pivotal aspects underlying the need for advanced forensic facial imaging methods. Firstly, the global demand and market potential for human identification and facial recognition technologies underscore their growing social and economic importance. Secondly, the integration of cutting-edge technological advancements, particularly in AI and ML, promises to revolutionize forensic facial imaging by automating and enhancing identification processes. Lastly, the vast availability and accessibility of photographs provide a unique advantage for identification purposes, facilitating a more efficient and comprehensive approach to FA and law enforcement.

I.1.1 Global Demand and Market Potential

The global demand for human identification methods, particularly forensic facial imaging, has experienced a significant rise due to its increasing social and economic importance (see Fig. 2). The human identification market reached an estimated size of USD 1.2 billion in 2023 and is projected to grow to USD 3 billion by 2033, reflecting a compound annual growth rate (CAGR) of 9.6% [1] . This substantial

[1]Human Identification Market. `https://www.futuremarketinsights.com/reports/human-identification-market`, last accesed July 26, 2024

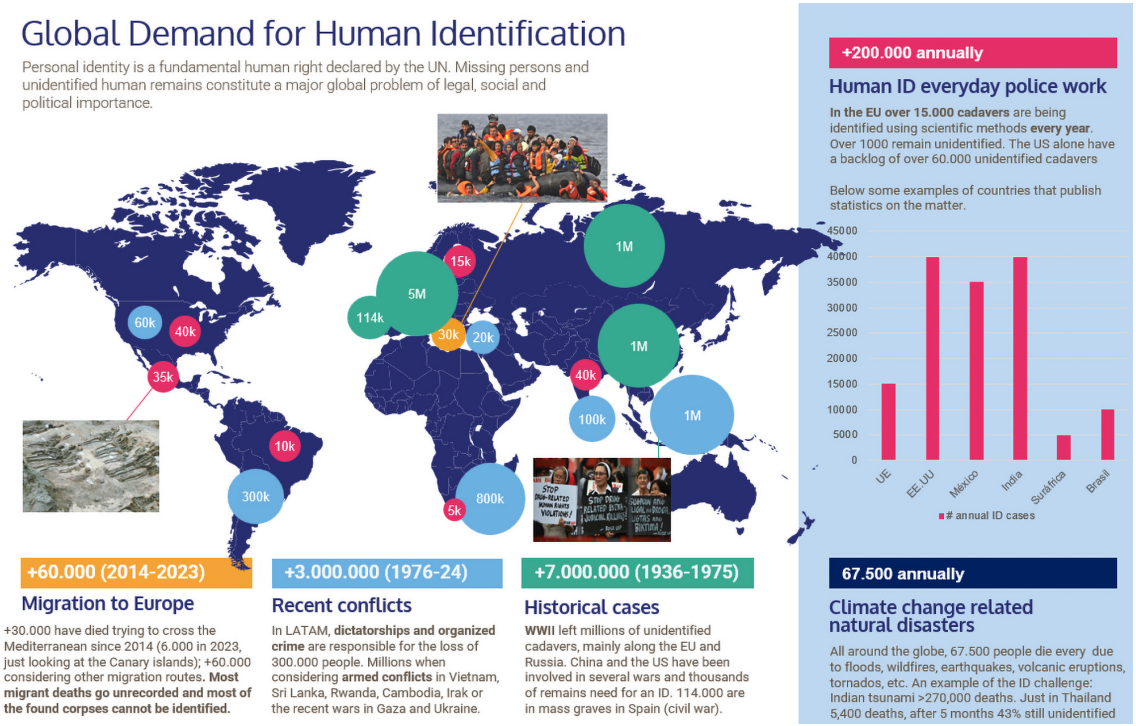


Figure 2: This figure illustrates the urgent need for human identification worldwide, driven by factors such as migration to Europe, recent conflicts, historical cases, and climate change-related natural disasters. Highlighted regions and associated data emphasize the scale of unidentified human remains, with notable numbers in Europe, Asia, Latin America, and Africa. Additionally, the figure presents annual human identification statistics, showcasing the extensive backlog of unidentified cadavers in the US and the ongoing efforts in the EU. These challenges underscore the critical role of forensic facial imaging in addressing the growing demand for human identification.

growth highlights the critical need for effective identification solutions. Moreover, the facial recognition market, estimated at USD 6.61 billion in 2024, is expected to reach USD 14 billion by 2029, growing at a CAGR of 16.20% [2] . This rapid expansion indicates a robust demand for facial imaging technologies, driven by the proliferation of surveillance cameras, mobile devices, and social media platforms, all of which contribute to an unprecedented volume of accessible facial photographs.

Furthermore, this heightened demand underscores the necessity for advanced forensic facial imaging methods to meet various law enforcement and public safety needs. The potential applications of facial identification extend beyond identifying deceased individuals; they play a crucial role in apprehending criminals, locating missing persons, and exposing identity theft. As more governments and agencies adopt facial recognition systems, the volume of leads generated for manual examination by experts continues to grow. For instance, the Facial Identification Section at NYPD received 9,850 requests for comparison in 2019, identifying 2,510 possible

[2]Facial Recognition Market. `https://www.mordorintelligence.com/industry-reports/facial-recognition-market`, last accesed July 26, 2024

matches [3]. Although a facial recognition match alone does not establish probable cause, it serves as a valuable investigative lead. This increasing reliance on facial imaging technologies for public safety and justice initiatives demonstrates the essential role of forensic facial imaging in contemporary society.

I.1.2 Technological Advancements and Artificial Intelligence Integration

The integration of advanced technological advancements, particularly AI, into forensic facial imaging represents a significant leap forward in the field. AI has revolutionized various domains by automating repetitive or complex tasks, often outperforming human capabilities [4]. In recent decades, DL [GBC16, Chapter 1], a subset of ML, has driven breakthroughs in image recognition, generation, and processing [BB23, Chapter 1]. This is particularly relevant for FA, which has traditionally been a manual and technologically limited discipline. By leveraging AI, forensic facial imaging can transition from labor-intensive methods to automated, accurate, and scalable processes, enhancing both efficiency and reliability.

Moreover, the application of AI in forensic facial imaging involves multiple advanced techniques. Convolutional Neural Networks (CNNs) are instrumental in Computer Vision (CV) tasks such as image classification and object recognition, making them powerful tools for identifying patterns in digital images [GBC16, Chapter 9]. In addition, Deep Generative models, including Generative Adversarial Networks (GANs) [GPAM+14] and Diffusion models [HJA20], are pivotal for image synthesis and processing tasks. While GANs are known for high-fidelity results [KLA19, KLA+20], diffusion models offer richer photorealistic outputs [DN21]. These technologies enable the processing of vast amounts of data, uncovering hidden patterns, and enhancing the overall quality of facial recognition systems. Despite the advancements, FA has yet to fully embrace these technologies, presenting an opportunity to significantly improve identification methods. By incorporating these AI-driven techniques, forensic facial imaging can achieve greater accuracy, explainability, and accountability, ultimately enhancing its role in modern law enforcement and forensic practices.

I.1.3 Advantages and Relevance of Facial Imaging

Facial imaging has become increasingly relevant in today's forensic science landscape due to the growing global demand for efficient and accurate identification methods. As highlighted earlier, the human identification market is expanding rapidly, driven by technological advancements and the need for robust forensic tools. Facial imaging, encompassing techniques such as facial identification and age progression, addresses this demand by leveraging the vast availability of photographs and the power of AI-driven analysis. These methods provide a non-invasive, easily

[3] Facial Recognition - NYPD. `https://www.nyc.gov/site/nypd/about/about-nypd/equipment-tech/facial-recognition.page`, last accesed July 26, 2024

[4] Stanford's 2024 AI INDEX ANNUAL REPORT. `https://aiindex.stanford.edu/report/`, last accesed July 26, 2024

accessible, and highly reliable means of identifying individuals, which is crucial for timely and effective forensic investigations.

Moreover, facial imaging techniques could greatly benefit from the integration of AI and ML, enhancing their accuracy and reliability. Methods such as facial identification could use similar sophisticated algorithms developed in CV to match facial features from photographs with those in databases, improving the chances of correctly identifying individuals [KKSK20]. Additionally, age progression technologies could benefit from similar advancements [YPN+21, OESF+20], enabling forensic experts to create accurate representations of how individuals might age over time. This is particularly valuable in long-term missing person cases, where updated images can greatly assist in the search and recovery efforts. The integration of these advanced technologies ensures that facial imaging remains at the forefront of forensic science innovation.

Furthermore, the practical applications of facial imaging extend beyond traditional identification purposes, playing a vital role in broader law enforcement and public safety initiatives. For example, facial identification can aid in tracking and apprehending suspects [VD15, Chapters 9 and 13], while advanced techniques such as age progression and the automatic editing of facial attributes provide crucial leads in various investigative scenarios [Mul12]. These technologies can simulate changes in a person's appearance over time, including aging, beard growth, hairstyle alterations, and color modifications. Such capabilities are invaluable in cases involving missing children or adults who may have altered their appearance to avoid detection [5]. The efficiency and scalability of modern facial imaging techniques enable forensic departments to process large volumes of data swiftly, enhancing their ability to solve cases and protect communities [6]. As technology continues to evolve, the relevance and importance of facial imaging in forensic science will only grow, solidifying its role as an essential tool for modern law enforcement and forensic investigations.

I.2 Objectives

The primary objective of this dissertation is to design innovative and disruptive DL-based solutions that can be effectively utilized in forensic facial imaging. By leveraging the power of DL, the aim is to enhance the accuracy, efficiency, and reliability of forensic facial analyses, thereby significantly improving the capabilities of forensic investigations. This objective encompasses the research and development of three subobjectives: developing and validating advanced methodologies for automatic cephalometric landmarks localization, face aging, and text-based image

[5] The National Center for Missing & Exploited Children (NCMEC) released a new age progression image of Tabitha Tuders, who disappeared in 2003 at age 13. The image, presented at CrimeCon in Nashville, shows what Tabitha might look like today at 34. The release included insights from forensic artist Joe Mullins and participation from Tabitha's family. `https://www.forensicmag.com/613385-New-Age-Progression-Image-of-Tabitha-Tuders-Missing-Since-2003/`. Last accessed on July 26, 2024.

[6] The FBI is questioned about its 640 million photographs facial recognition database. `https://www.forbes.com/sites/monicamelton/2019/06/04/government-watchdog-questions-fbi-on-its-640-million-photo-facial-recognition-database/`. Last acceses on July 26, 2024.

Figure 3: Some examples of text-based image editing, including adding facial features such as a beard, mustache, and long hair, as well as altering attributes like age and body weight. The unedited image has been created with Stable Diffusion 3 Medium (`https://huggingface.co/stabilityai/stable-diffusion-3-medium`), the edited images are generated with the methodology proposed in Chapter V.

editing. Each methodology developed in this dissertation has to be rigorously validated through user studies to ensure their practical applicability and effectiveness.

1. The first subobjective is to develop a DL-based system for the automatic localization of cephalometric landmarks on facial images. Cephalometric landmarks are critical points on the human face extensively used in FA and other areas. The goal is to create a robust and precise DL model that can accurately identify and annotate these landmarks, facilitating quicker and more reliable forensic analyses. This system aims to automate the traditionally manual process, thus reducing human error and increasing the scalability of forensic facial imaging techniques.

2. The second subobjective focuses on creating a DL model for face aging. This model aims to generate accurate age-progressed images of individuals, which is crucial in long-term missing person cases and criminal investigations where suspects may have aged over time. The method should allow the user to

determine the desired degree of structural preservation at inference time, acknowledging that individuals age in diverse ways. The model should be able to generate realistic and plausible age-progressed images that retain the key identity features of the individuals. By achieving this, the model will improve the chances of identification and provide valuable leads in forensic investigations.

3. The third subobjective is to design a versatile DL framework for text-based image editing. This approach allows forensic experts to manipulate images based on textual descriptions, enabling modifications such as adding, removing, or altering facial features (see Fig. 3). This capability is particularly useful in scenarios where visual alterations of suspects or victims are required based on witness descriptions or other investigative needs. The system should maintain a high level of realism and accuracy, aiding in the effective identification and analysis of individuals in forensic contexts.

By achieving these subobjectives, this dissertation aims to significantly advance the field of forensic facial imaging, providing powerful tools that enhance the accuracy, efficiency, and reliability of forensic investigations.

I.3 Structure of the dissertation

This dissertation is divided into three main parts besides the introduction: *Fundamentals*, *Proposal*, and *Final Remarks*. Each part contains several chapters that explore DL-based solutions for forensic facial imaging.

Part I, *Fundamentals*, covers the essential theoretical background. This section discusses cephalometric landmarks, how they are identified, and their significance. Furthermore, this part delves into DL, explaining how it uses neural networks to model complex patterns in data. Key topics include CNNs for image recognition, regression and classification tasks; GANs for creating realistic images; and diffusion models and attention mechanisms that enhance the quality of generated images

Part II, *Proposal*, presents the core research contributions of the dissertation. The first chapter focuses on developing a DL system to automatically locate cephalometric landmarks on facial images. This advancement aims to make forensic analysis quicker and more reliable by automating a process traditionally done manually. The second chapter introduces a DL model for face aging, which generates age-progressed images that could, among other purposes, help identify long-missing individuals. This model ensures that key identity features are preserved even as the person ages. The third chapter describes a DL framework for text-based image editing, allowing forensic experts to modify facial images based on textual descriptions. This capability is particularly useful for updating or creating facial composites based on witness descriptions.

Part III, *Final Remarks*, provides conclusions and future directions for the research. It summarizes the key findings and contributions, emphasizing the advancements made in cephalometric landmark localization, face aging, and image editing. This section also discusses potential future research paths and improvements for each area, ensuring ongoing development and refinement of these technologies. A

list of the scientific outcomes obtained from the current dissertation, in terms of journal and conference papers, is also included. This part finishes with the pertinent acknowledgements to the related projects and contracts which have supported the research developed.

The dissertation concludes with a comprehensive bibliography, listing all the references and sources cited throughout the research. This structure ensures a thorough and systematic exploration of DL-based solutions for forensic facial imaging, offering valuable insights and advancements in the field.

Part I

Fundamentals

Chapter II

Theoretical framework

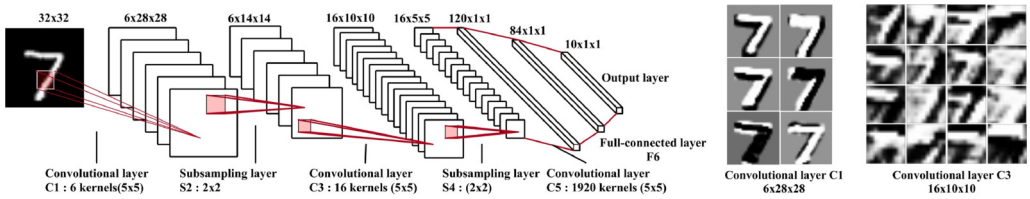
"¡Si uno conociera lo que tiene, con tanta claridad como conoce lo que le falta!" — Mario Benedetti

II.1 Deep Learning

II.1.1 Introduction

DL is a subfield of ML that focuses on learning complex concepts from simpler ones through a hierarchical structure, usually using artificial neural networks. This hierarchy enables a computer to learn intricate ideas by combining basic concepts in multiple layers, hence the name "deep learning". This methodologies allow AI systems to automatically understand data and make decisions based on it in increasingly sophisticated ways [GBC16, Chapter 1]. It has become a cornerstone of modern AI applications due to its ability to achieve State-of-the-art (SOTA) performance in various tasks such as image recognition, natural language processing, and game playing. The importance of DL in AI is underscored by its success in driving innovations and solving problems that were previously intractable [BB23, Chapter 1].

Neural networks, the fundamental building blocks of DL, consist of interconnected neurons organized into layers. These networks are trained through optimization processes that adjust the weights of the connections to minimize error. However, training deep networks can lead to overfitting, where the model performs well on training data but poorly on unseen data, especially in small data samples [GBC16, Chapter 7]. To address this, regularization techniques such as dropout [HSK+12] and weight decay [KH91] are employed. Additionally, evaluating the performance of DL models requires robust metrics and validation strategies to ensure that the models generalize well to new data. These core concepts are crucial for understanding how DL models are built, trained, and assessed.



(a) LeNet-5 network (b) Learned features

Figure 4: Illustration of a CNN and learned features. (a) The architecture of the network, which includes convolutional layers (C1, C3, C5), subsampling (or pooling) layers (S2, S4), and fully connected layers (F6 and the output layer). The network starts with a 32x32 input image and progressively reduces its dimensions through convolution and pooling operations, eventually producing a 1x1x10 output corresponding to class predictions. (b) Visualization of learned features at different layers, showing how initial layers capture simple features like edges and textures, while deeper layers capture more complex patterns. These features illustrate the hierarchical nature of feature learning in CNNs. Figures taken from [GWK+18].

II.1.2 Convolutional Neural Networks

CNNs are a specialized type of artificial neural network designed for processing structured grid data, such as images. They have revolutionized the field of CV and image processing by significantly improving the accuracy and efficiency of tasks like image classification [KSH12], object detection, and segmentation. The importance of CNNs in image processing lies in their ability to automatically and adaptively learn spatial hierarchies of features from input images, making them more effective than traditional neural networks for these tasks. Unlike traditional feedforward neural networks, which use fully connected layers where each neuron is connected to every neuron in the previous layer, CNNs use a series of convolutional layers that apply convolution operations to the input as a sliding window operation, reducing greatly the number of parameters to learn. This architecture allows CNNs to focus on local features and build complexity by stacking multiple layers. The result is a network that can recognize patterns, shapes, and objects in images with high accuracy [GBC16, Chapter 9].

An example of a CNN can be seen in Fig. 4. At the heart of CNNs are convolutional layers, which apply convolution operations using filters (or kernels) that slide over the input image to produce feature maps. These layers are followed by pooling layers that downsample the feature maps, reducing their dimensionality and computational load while retaining essential information. Fully connected layers are typically used towards the end of the network to make final predictions. Activation functions like ReLU (Rectified Linear Unit) [GBB11] introduce non-linearity into the model, allowing it to learn more complex patterns. These architectural components and their operations are key to the functionality and success of CNNs [HZRS16, SZ14].

II.1.3 Regression and classification

Regression and classification are fundamental tasks in supervised learning, where the goal is to learn a mapping from input data to output labels or values based on

example input-output pairs. Regression involves predicting a continuous output variable, such as temperature, price, age, or coordinates from input features. In contrast, classification entails predicting one or more discrete labels, such as spam or not spam, based on input data. These tasks are pivotal in supervised learning because they enable models to make informed predictions and decisions based on previously seen data, thereby solving a wide range of real-world problems.

CNNs have achieved significant milestones in both regression and classification tasks, particularly in the field of image processing. For regression tasks, CNNs have been employed to predict continuous variables such as age estimation from facial images [RTVG15] or depth estimation from 2D images [RLH+20]. In classification, CNNs have demonstrated outstanding performance in categorizing images into various classes [HZRS16], revolutionizing fields like medical imaging [RFB15], and facial recognition [RPC17]. Key milestones include the development of architectures like AlexNet [SZ14], which significantly improved image classification accuracy, and the application of CNNs to regression tasks, showing their versatility and robustness in handling different types of prediction problems.

II.1.4 Generative Adversarial Networks

GANs are a class of ML frameworks designed for generative modeling, where the goal is to generate new data samples that resemble a given dataset or data distribution. Introduced by Ian Goodfellow and his colleagues in 2014 [GPAM+14], GANs consist of two main components: the generator and the discriminator. These two networks are trained simultaneously in a game-theoretic framework where the generator tries to produce realistic data, and the discriminator attempts to distinguish between real and generated data. This adversarial process results in the generator learning to produce highly realistic data, making GANs extremely powerful for various generative tasks [KLA19, KLA+20].

As depicted in Fig. 5, the generator in a GAN aims to create data that is indistinguishable from real data, while the discriminator evaluates the data and determines whether it is real or generated. The importance of GANs in generative modeling lies in their ability to create high-quality, realistic data, which is invaluable in applications such as image synthesis, video generation, and data augmentation [SK19]. The architecture of GANs generator typically maps a latent space representation to the data space [GPAM+14, KLA19]. The latent space acts as a compressed representation of the data, usually following a known random distribution, enabling the generator to explore diverse data samples efficiently.

The generator network in GANs is responsible for generating new data samples from random noise or a latent space, while the discriminator network evaluates these samples to determine their authenticity. The core component of training GANs is the adversarial loss [GPAM+14], which drives the generator to produce data that can fool the discriminator. This adversarial training process involves alternating between optimizing the generator and the discriminator, with each network trying to outsmart the other. This delicate balance can lead to instability, resulting in issues such as mode collapse [GPAM+14], where the generator produces limited varieties of samples, and training oscillations. Ensuring stable training requires careful design of loss functions and training protocols, with adversarial loss playing a pivotal role

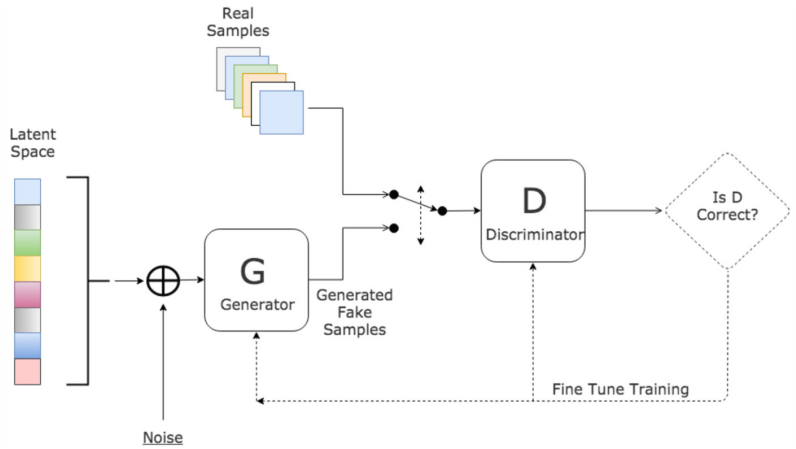


Figure 5: Overview of the GAN architecture. The generator (G) creates data samples from a latent space, typically consisting of random noise. These generated samples are then evaluated by the discriminator (D), which attempts to distinguish between real samples and those generated by the generator. The adversarial training process involves the generator learning to produce increasingly realistic data to fool the discriminator, while the discriminator simultaneously improves its ability to detect fake samples. This iterative process is guided by the adversarial loss, fine-tuning both networks to achieve high-quality data generation. Figure taken from [GZ19].

in guiding the generator towards producing realistic outputs. Solutions to these stability issues include techniques like Wasserstein GAN (WGAN) [ACB17], which introduces a different loss function to improve training stability and sample diversity. Variations and improvements in GANs have also explored different architectural adjustments and training strategies to enhance performance and stability [HRU+17, KALL17].

Several variations of GANs have been developed to address specific challenges and enhance the capabilities of the basic framework. Conditional GANs incorporate additional information, such as class labels or text descriptions, into the generator and discriminator [CUYH20], enabling controlled data generation. Style-based GAN architectures, exemplified by StyleGAN [KLA19] and StyleGAN2 [KLA+20], represent a significant advancement in the field of generative adversarial networks, specifically designed for high-quality image synthesis. Unlike traditional GANs, which directly map a latent space to the data space, style-based GANs introduce an intermediate latent space and manipulate features at different layers of the generator. This architecture utilizes adaptive instance normalization (AdaIN) [HB17] to control the style of generated images, allowing for fine-grained adjustments in features such as texture, color, and overall structure. AdaIN works by normalizing the mean and variance of the features at each layer and then modulating them with learned affine transformations derived from the style vector. This process effectively separates content from style, enabling the generator to apply complex, high-level attributes independently from the basic structure of the generated images. This design results in improved control over image attributes, leading to more realistic and diverse image generation. StyleGAN2 further refines this approach by addressing artifacts in the generated images and introducing weight demodulation, a more

stable solution, as a substitute for AdaIN.

II.1.5 Encoder-Decoder architectures

Encoder-decoder architectures form a fundamental building block in many advanced generative models, particularly within the realm of Image-to-Image (I2I) translation and other tasks requiring transformation of input data into a different form. The encoder-decoder framework consists of two primary components: the encoder, which compresses the input data into a compact latent representation, and the decoder, which reconstructs the output data from this latent space. This architecture is especially beneficial for tasks such as image synthesis [YPN+21], super-resolution [HNW+19], and semantic segmentation [guo].

The encoder typically employs a series of convolutional layers to progressively reduce the spatial dimensions of the input while increasing the depth of feature maps, capturing essential information and abstract features. The latent representation produced by the encoder serves as a compressed summary of the input data. The decoder, conversely, uses transposed convolutions or upsampling techniques to expand this representation back to the original data dimensions, synthesizing the final output. A notable variant of the encoder-decoder architecture is the U-Net [RFB15], which includes Skip connection (SC) between corresponding layers in the encoder and decoder. These connections help retain spatial information lost during downsampling [CCK+18, BW20], significantly improving the quality of the reconstructed output, especially in tasks requiring precise localization, such as medical image segmentation. Encoder-decoder architectures are also integral to conditional GANs. This approach is used in models like Pix2Pix [IZZE17] and CycleGAN [ZPIE17], where the goal is to learn mappings between two visual domains.

The versatility and effectiveness of encoder-decoder architectures make them a cornerstone in modern generative modeling, driving advances in various applications like artistic style transfer [ZPIE17] or realistic data augmentation [SK19].

II.1.6 Diffusion models

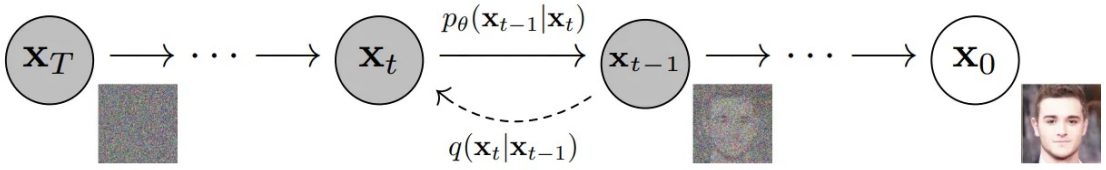


Figure 6: Illustration of the Denoising Diffusion Probabilistic Model (DDPM). The process starts with a noisy image $\mathbf{x}_T$ and gradually denoises it through a series of steps until the final image $\mathbf{x}_0$ is obtained. At each step t, the model predicts the previous step $\mathbf{x}_{t-1}$ using the conditional probability $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$. The reverse process (dashed arrow) approximates the forward process $q(\mathbf{x}_t|\mathbf{x}_{t-1})$. The images at each step illustrate the gradual denoising process, transitioning from a noisy image to a clear facial image. Figure taken from [HJA20].

Diffusion models are a class of generative models that define a data generation process through a series of steps involving the addition and subsequent removal of noise [HJA20]. Their primary purpose is to generate high-quality data samples by reversing a forward diffusion process that corrupts the data with noise. The basic principles of diffusion in the context of ML involve progressively adding noise to the data in small steps until it becomes unrecognizable, and then learning to reverse this process to recover the original data. The architecture and mechanism of diffusion models are designed to handle this forward and reverse process efficiently, making them powerful tools for various generative tasks [DN21].

To understand the mechanics of diffusion models, it is essential to grasp the forward and reverse diffusion processes. In the forward process, noise is gradually added to the data at each step, effectively transforming it into pure noise over many iterations (generally a random normal distribution) [DN21]. The reverse process, which the model learns, involves removing the noise step-by-step to reconstruct the original data. Key equations and algorithms, such as Diffusion Probabilistic Models [HJA20] (see Fig. 6) and Denoising Diffusion Implicit Model (DDIM) [SME21], govern these processes. These algorithms provide the mathematical framework for noise addition and removal, ensuring the model can generate high-fidelity samples for any intractable data distribution starting from a known random distribution. Diffusion models find applications in various generative tasks, including image generation and other areas where data synthesis is required.

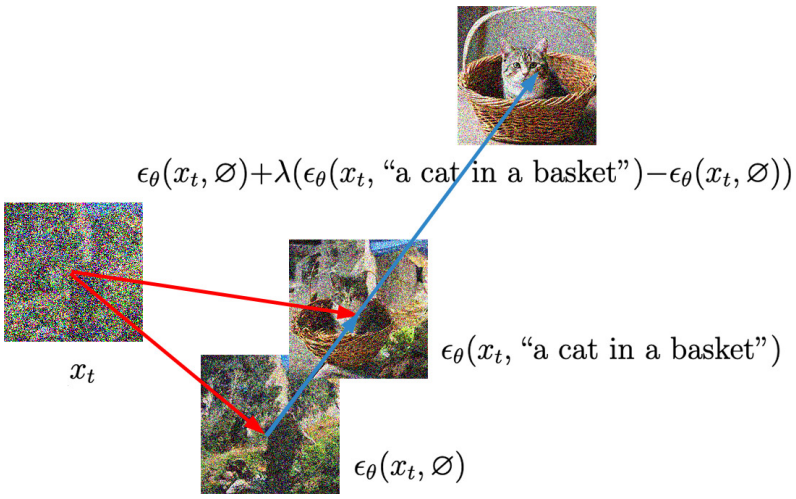


Figure 7: Illustration of Classifier-Free Guidance (CFG) in diffusion models. Starting from a noisy image $\mathbf{x}_t$, the model produces two intermediate outputs: $\epsilon_\theta(\mathbf{x}_t, \varnothing)$ which is the unconditional prediction, and $\epsilon_\theta(\mathbf{x}_t, \text{“a cat in a basket”})$ which is the conditional prediction guided by the prompt “a cat in a basket”. The final output is then computed by combining these predictions with a guidance scale λ, resulting in $\epsilon_\theta(\mathbf{x}_t, \varnothing) + \lambda(\epsilon_\theta(\mathbf{x}_t, \text{“a cat in a basket”}) - \epsilon_\theta(\mathbf{x}_t, \varnothing))$.

Moreover, conditional diffusion models extend the basic diffusion framework by incorporating additional information into the generation process, usually with the use of CA modules [RBL+22]. Techniques such as classifier guidance [DN21] and Classifier-Free Guidance (CFG) [HS21] are employed to condition the data generation on specific attributes or labels. Classifier guidance involves using a pre-trained classifier to influence the diffusion process, guiding the model towards generating

data that conforms to the specified conditions. In contrast, CFG simplifies this process by integrating the guidance directly into the diffusion model, eliminating the need for a separate classifier. CFG achieves this by using null-text prompts and regular prompts during training. The null-text prompt represents an empty or baseline condition, while the regular prompt contains the specific attributes or conditions desired in the output. By magnifying the difference between the outputs generated from these two prompts, CFG effectively guides the model towards the desired conditioned outputs (see Fig. 7). This approach enhances the model's ability to produce specific and controlled data, making conditional diffusion models highly effective for tasks that require precise and targeted data generation. These advancements significantly enhance the applicability of diffusion models in various targeted domains, enabling more accurate and contextually relevant data synthesis.

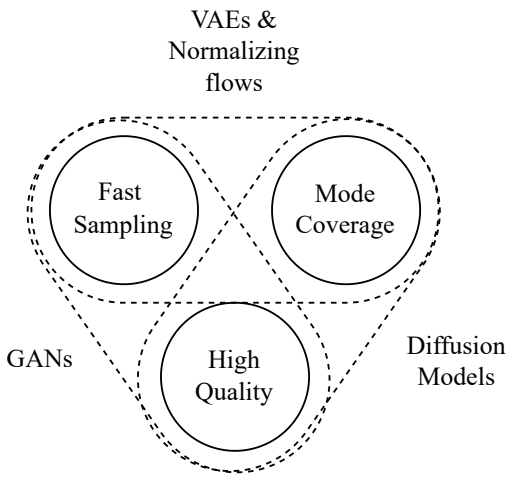


Figure 8: Trade-offs in generative models: GANs provide high-quality samples but struggle with mode coverage and diversity. Likelihood-based models such as VAEs and Normalizing Flows offer good mode coverage but may compromise on sample quality and speed. Denoising Diffusion Models achieve mode coverage and diversity with high-quality samples, but they often require thousands of network evaluations, making them slower in sampling.

As depicted in Fig. 8, there is a trade-off between different generative modeling approaches, each excelling in certain areas while compromising in others. GANs are renowned for producing high-quality samples but often face challenges in mode coverage, resulting in less diversity in the generated outputs. In contrast, likelihood-based models, such as Variational Autoencoders (VAEs) [KW13] and Normalizing Flows [PNR$^+$21], prioritize achieving comprehensive mode coverage and diversity, ensuring a wide range of possible outputs at the potential cost of sample quality and speed [XKV21]. Denoising Diffusion Models offer a balanced approach by providing robust mode coverage and diversity alongside high-quality samples. However, they require substantial computational resources due to the need for numerous network evaluations during the generation process unlike both GANs and likelihood-based models. These trade-offs highlight the necessity of selecting the appropriate generative model based on the specific requirements of the task, balancing factors like sample quality, diversity, and computational efficiency.

II.1.7 Attention in the context of diffusion models

The attention mechanism [VSP+17], described by the equation

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{II.1}$$

$$\begin{aligned} Q &= W_Q X_Q \quad \text{(Query matrix)} \\ K &= W_K X_K \quad \text{(Key matrix)} \\ V &= W_V X_V \quad \text{(Value matrix)} \\ d_k &= \text{dimension of the key vectors} \end{aligned}$$

is a fundamental component in modern deep learning architectures, particularly in transformer models. In this equation, Q, K, and V stand for the query, key, and value matrices, respectively, which are derived from the input data through learned weight matrices W_Q, W_K, and W_V. The term $\sqrt{d_k}$ represents the scaling factor, where d_k is the dimension of the key vectors, ensuring numerical stability during the softmax operation. Attention mechanisms allow models to focus on different parts of the input data dynamically as well as to mix different kinds of inputs like images and text. In the context of diffusion models, attention mechanisms can replace traditional convolutional layers, offering more flexibility in capturing dependencies across different parts of the data. This replacement enables the model to better understand and manipulate complex structures within the data, leading to improved performance in generative tasks.

Self-Attention (SA), is the mechanism where the queries, keys, and values all come from the same source. This means that, as an example, each word in a sentence is compared with every other word in the same sentence to understand their relationships and dependencies. In the context of the equation, the input matrix X_{QKV} for Q, K, and V is the same. This allows the model to weigh the importance of each word in relation to every other word, capturing context effectively within a single sequence. Self-attention is particularly powerful in capturing long-range dependencies, making it crucial for tasks such as natural language processing and machine translation.

Cross-Attention (CA), on the other hand, involves the queries coming from one source while the keys and values come from another. This mechanism is used to align and integrate information from different sources. For instance, in a sequence-to-sequence model for machine translation, the query Q might come from the decoder's current state while the keys K and values V come from the encoder's outputs. Cross-attention is essential in tasks where there is a need to combine information from multiple sources, such as in image generation or image captioning where the model needs to align textual descriptions with visual features (see Fig 9).

In Hertz et al. [HMT+23], they propose a method that leverages CA manipulation to facilitate precise and intuitive image editing using text prompts. Their work focuses on the key role of CA layers in linking the textual tokens of the prompt with the spatial layout of the generated image. The core idea behind this method is that

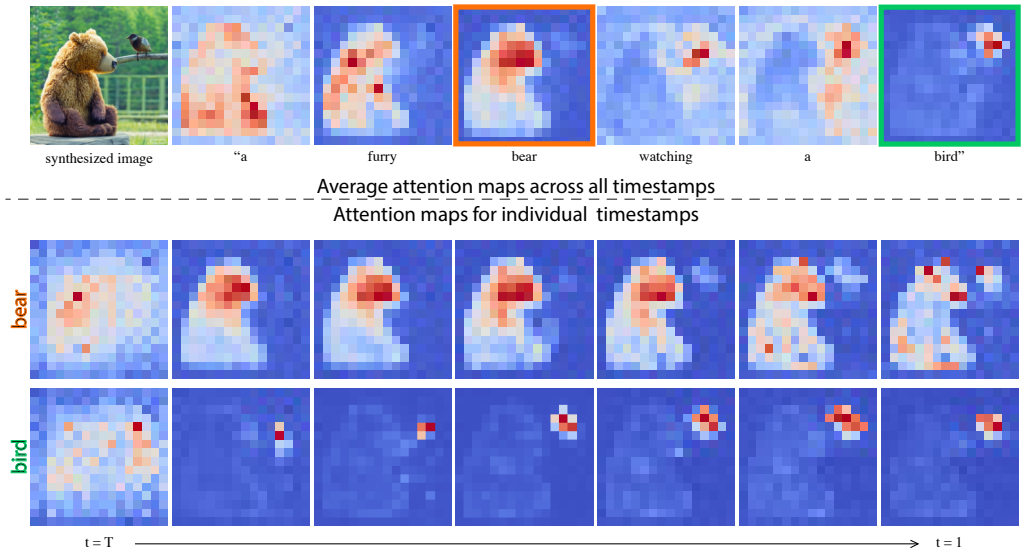


Figure 9: Cross-attention maps of a text-conditioned diffusion image generation. The top row displays the average attention masks for each word in the prompt that synthesized the image on the left. The bottom rows display the attention maps from different diffusion steps with respect to the words "bear" and "bird". Figure taken from [HMT+23].

the CA maps, which are high-dimensional tensors produced during the noise prediction process in text-conditioned diffusion models, contain rich semantic relations that significantly influence the generated image. By manipulating these CA maps, the model can control which pixels in the image correspond to specific words in the prompt at various diffusion steps, enabling detailed and controlled edits without needing additional input like spatial masks.

The method involves injecting CA maps from the original image generation process into a new generation process with a modified prompt. This ensures that the structural and compositional details of the original image are preserved while adapting to the new textual instructions. As seen in Fig. 10, this technique can be used for multiple tasks:

- Word Swap: For localized edits, where specific words in the prompt are replaced (*e.g.*, changing "dog" to "cat"), the CA maps from the original image are injected to maintain the overall composition while adapting the new word's content.
- Adding New Phrases: For global edits, new phrases are added to the prompt (*e.g.*, changing the style or adding new attributes), and the attention maps for the unchanged part of the prompt are retained to preserve the original image structure while allowing new features to be integrated.
- Attention Re-weighting: To fine-tune the influence of certain words, the attention maps are scaled, thereby amplifying or attenuating the visual effect of specific tokens.

In [MHA+23], they present a solution to allow this method to be also applied to real images through a computationally intensive inversion process, where a real

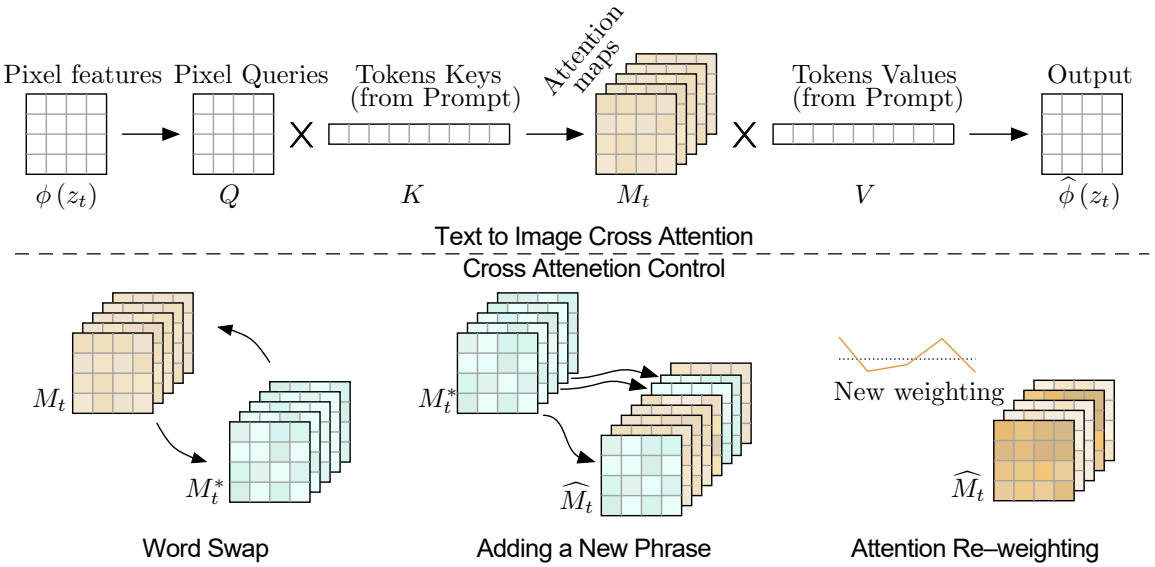


Figure 10: [HMT+23] overview. Top: visual and textual embedding are fused using cross-attention layers that produce spatial attention maps for each textual token. Bottom: they control the spatial layout and geometry of the generated image using the attention maps of a source image. This enables various editing tasks through editing the textual prompt only. When swapping a word in the prompt, they inject the source image maps M_t, overriding the target image maps M_t^*, to preserve the spatial layout. Where in the case of adding a new phrase, they inject only the maps that correspond to the unchanged part of the prompt. Amplify or attenuate the semantic effect of a word achieved by re-weighting the corresponding attention map. Figure taken from [HMT+23].

image is transformed into a latent noise vector and then edited using the same prompt-to-prompt techniques. The process consists of two key components: pivotal inversion and null-text optimization. Pivotal inversion starts with DDIM inversion to obtain a noise vector sequence, which serves as a pivot for optimization, improving reconstruction fidelity. Null-text optimization refines the unconditional textual embedding used in CFG, allowing for high-fidelity reconstruction while maintaining the model's weights and conditional embedding intact. This method enables intuitive, high-quality edits on real images using text prompts, without the need for cumbersome fine-tuning of the model's parameters.

II.2 Cephalometric landmark localization

II.2.1 Cephalometric landmarks

Cephalometric landmarks are points located on the face and head that allow forensic experts to characterize the morphology of the head [Geo07]. Each of them has a corresponding craniometric landmark, *i.e.*, a corresponding point on the surface of the skull, as seen in Figure 11 [1]. Within physical and FA, cephalometric

[1]The distinction between cephalometric and craniometric landmarks is based on the criteria presented in [Geo07]. Other areas, such as orthodontics, refer to landmarks annotated on X-ray skull images as cephalometric landmarks [PHM+19, RSC98]. The latter taxonomy is different from ours.

Figure 11: Comparison between some corresponding craniometric landmarks (in black) and cephalometric landmarks (in red) on a lateral head view. The alignment, depicted with blue lines, depends heavily on the characteristics of the soft tissue and varies among different head regions.

landmarks are highly relevant in tasks such as anthropometric proportion index extraction, pathology detection and classification [Far94], forensic facial comparison [SCGC19], and forensic human identification [HIWK15]. Specifically, concerning the latter, there are automatic craniofacial superimposition techniques that seek to optimally align cephalometric and craniometric landmarks taking into consideration the existence of soft tissues [DCIn$^+$11, CAIN$^+$14a, CÁICÁ$^+$15]. Regarding forensic facial comparison, cephalometric landmarks are relevant, for instance, in photoanthropometry and in estimating 3D proportionality indices from 2D measurements (*i.e.*, given a photograph of an unidentified subject, calculating the range of 3D dimensions and proportionality indices of that person in the photograph) [MVIA18]. The 30 cephalometric landmarks that will be studied in this dissertation are listed in Tab. 1 and shown in Fig. 12.

Despite the significant need and potential benefits of an accurate, robust, and efficient method for cephalometric landmark annotation on facial images, this task has largely remained manual. Developing a method capable of reliably and automatically locating landmarks in the types of images commonly used in FA would yield more robust, faster, and repeatable results. This would enable the processing of a large number of photographs in a reduced timeframe. Additionally, such a tool could be used as a final solution or as an initial landmark placement that forensic experts could later refine if necessary. Consequently, this automation would save substantial time that could be allocated to other forensic tasks. Moreover, these resources are critically important in mass disaster victim identification scenarios, where the high volume of identifications required makes manual landmarking impractical.

In forensic settings, frontal high-quality consistent photographs cannot always be expected to be available. Instead, different poses and data sources are considered. Besides the difficulties derived from the presence of occlusions and the diverse image acquisition processes, an additional obstacle is the absence of large and reliable datasets. Figure 13 displays some examples that reflect the complexity of the data.

Current methods for identifying cephalometric landmarks in forensic analysis are

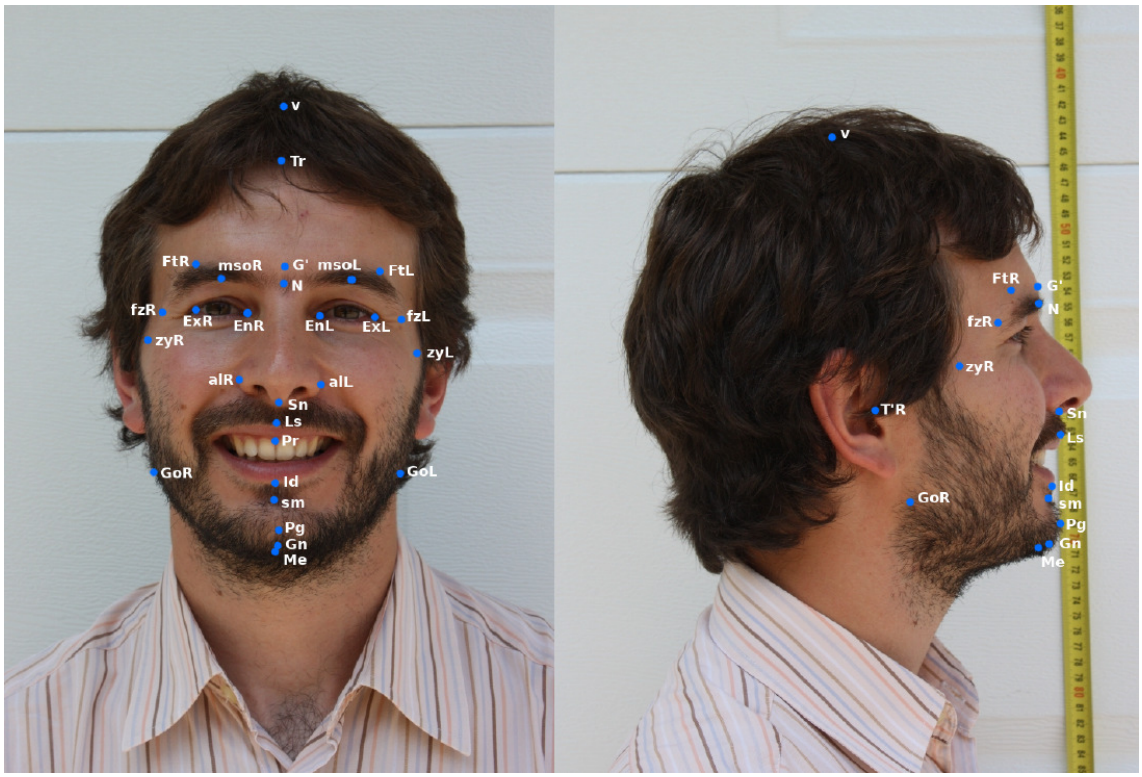


Figure 12: Facial image (frontal and lateral views) annotated with the cephalometric landmarks employed in this dissertation. Abbreviations can be found in Table 1.

Figure 13: Some examples of the kind of images they work with in forensic settings. As can be easily noticed, it presents a high variability in terms of image resolution, pose, illumination, background, occlusions, and facial expression. Unfortunately, most images cannot be published due to privacy issues.

primarily manual, which can lead to a number of issues, including subjectivity, a lack of speed, and a dependence on the expertise, knowledge, and experience of the forensic expert. Additionally, the repetitive nature of manual landmark identification can contribute to errors resulting from the expert's fatigue. These factors collectively demonstrate the necessity for automating the process of landmark identification.

The process of annotating landmarks in photographs is subject to inter- and intra-expert variability, both in terms of the position of the landmarks and in determining their visibility [CAIN+14b]. Automating this process must consider the high degree of variability present in images, including differences in physical appearance and head pose, as well as variations in image origin (such as images from digital cameras, identity card scans, and older grayscale photos) and resolution. Additionally, it must account for the possibility that faces may not be prominently featured in the foreground or that photographs may be out of focus or noisy. We should also

Table 1: Cephalometric landmarks considered in this PhD dissertation.

	Landmarks	**Abbreviation**
1	Menton	Me
2	Gnathion	Gn
3	Pogonion	Pg
4	Prosthion	Pr
5	Labiale Superius	Ls
6	Subnasale	Sn
7	Nasion	N
8	Glabella	G'
9	Vertex	v
10/11	Left/Right Gonion	GoL/GoR
12/13	Left/Right Zygion	zyL/zyR
14/15	Left/Right Alare	alL/alR
16/17	Left/Right Endocanthion	EnL/EnR
18/19	Left/Right Exocanthion	ExL/ExR
20/21	Left/Right Tragion	T'L/T'R
22	Infradentale	Id
23	Trichion	Tr
24	Supramentale	sm
25/26	Left/Right Frontotemporale	FtL/FtR
27/28	Left/Right Frontozygomaticus	fzL/fzR
29/30	Left/Right Midsupraorbital	msoL/msoR

take into account two factors to establish a landmark as visible: the confidence of the expert to annotate it accurately and the absence of occlusions of any kind (such as glasses, hair, and hands covering parts of the face). In short, the ability to locate visible landmarks in all types of photographs is crucial to ensure that this method is helpful for forensic experts in their professional practice.

To the best of our knowledge, only three existing studies address this task directly. However, they do not consider the *in-the-wild* nature of the images and instead focus on a simplified problem involving photographs taken under controlled and stable conditions. The approach presented in [AIA+14] only addresses two specific landmarks, the *endocanthion* and the *exocanthion*, using Haar-like features [VJ01], with both landmarks located in highly contrasted areas. Meanwhile, [RSIM19] employs a DL model to detect the face and then utilizes active shape models [CT04] to locate the position of the landmarks. Similarly, [PLF+19] uses Haar-like features to locate the face and morphable face models [HFC+15] trained on a dataset of 1000 frontal images with 28 labeled landmarks to predict the location of the landmarks. These methods are not able to make predictions for different face rotations or to deal with diverse and uncontrolled image conditions.

II.2.2 Facial landmarks and its relation to cephalometric landmarks

There is abundant literature in the field of *facial* landmark detection [WJ19], a problem in appearance similar to ours. However, unlike for *cephalometric* landmarks, existing methods and data present some limitations for our specific application domain:

1. The facial landmarks studied in such works have no interest from the FA point of view. They are inconsistent and not anatomically oriented (*e.g.*, in usual benchmarks [STZP13, WQY+18], the same landmarks are used for describing the outline of the face and the outline of the jaw depending on the head rotation).
2. Low-resolution annotations pose a limit to potential model performance as shown in [CBGB20].
3. Available methods have not been validated in a forensic environment.
4. Facial landmarks have not been evaluated in forensic tasks.

For these reasons, there is a strong need for developing tools in this research area and reliably assisting forensic anthropologists in their daily work in an automatic or semi-automatic fashion. Furthermore, these kinds of solutions can have a high impact at a human, legal and economic level, including the fulfillment of the admissibility criteria for expert evidence [CC09, Fra10, MG15].

The literature on facial landmark localization primarily employs three different families of methods, all based on DL approaches due to the availability of large datasets and the ability of CNNs to learn from data: *1)* Initial methods based on coordinates regression like Hyperface [RPC17] have been surpassed in the literature by *2)* heatmap-based and *3)* deformable 3D mask methods, which require more parameters and greater computational resources.

On the one hand, regarding heatmap-based methods, HRNET [WSC+20, SZJ+19] trains a U-shaped convolutional network [RFB15] to generate a heatmap for each landmark. A similar approach with sub-pixel accuracy is used in [WLL+20]. 3FabRec [BW20] addresses the problem in a few-shot scenario with a large dataset of unlabeled images and a small number of images with annotated landmarks. First, an unsupervised autoencoder is trained, and then the intermediate convolutional blocks are frozen to learn the prediction heatmaps. LUVLi [KMM+20] learns a convolutional network that predicts the position of the landmark using heatmaps with sub-pixel accuracy while also outputting an uncertainty and visibility estimation. ADNet [HYL+21] analyses the error bias on both human and predicted annotations and suggests that the error does not follow an isotropic distribution. A new loss function and a model tailored for anisotropic error bias are proposed to improve landmark prediction performance. The proposals in [CBGB20] and [WBH+21] use high-resolution, accurately-annotated additional data to improve their performance. In addition, the method proposed in [WBH+21] uses augmented computer-generated high-quality diverse images with a low domain-distribution gap, whereas that from

[CBGB20] uses a cascade two-stage approach for high-resolution prediction trained on labeled private images of up to 4k pixels in size.

On the other hand, for deformable mask methods, 3DDFA [ZLL+16] solves the landmark localization problem in a 3D space through deformable 3D masks trained on an augmented dataset. A similar solution with improved efficiency is presented in [FWS+18]. Furthermore, 3DDFA v2 [GZY+20] employs a meta-joint optimization strategy to regress a smaller set of parameters for the deformable 3D mask, which improves both speed and accuracy.

The most frequent metric to evaluate facial landmark localization methods is the *Normalized Mean Error (NME)*. It is a measure of the Euclidean distance between the predicted position $\mathbf{y}_{ij}$ and the ground truth position $\mathbf{y'}_{ij}$ for the landmark j normalized by the largest side of the bounding box c_i for the i^{th} image. The NME is computed as follows

$$NME_j = \frac{1}{N} \sum_{i=1}^{N} \frac{||\mathbf{y}_{ij} - \mathbf{y'}_{ij}||_2}{c_i} \tag{II.2}$$

and is usually presented as $NME(\%) = NME * 100$.

A popular benchmark for facial landmark localization is the Annotated Facial Landmarks in the Wild (AFLW) dataset [KWRB11]. This large-scale dataset contains 25,000 in-the-wild face images, with up to 21 landmarks per image, accounting for self-occluded and not-visible landmarks, and with no restrictions regarding head pose. The performance of these methods on the subset of the AFLW dataset containing only frontal images is referred to as $\text{NME}_{\text{frontal}}$. In contrast, the measure of the error on all AFLW images is referred to as NME_{full}. The latter metric better capture the kind of difficulties we find in our problem (*i.e.*, cephalometric landmarks localization), and will be used for comparing the available methods.

Table 2: Comparison of NME_{full} (%) reported performance on the AFLW dataset for several SOTA methods belonging to the three families described in Sec. II.2.2. We include the number of training samples (N) and predicted landmarks (*#Landmarks*). (*) It should be noticed that 3DDFA v2 uses the 300W-LP dataset for training (a custom 3D augmented dataset extending several datasets that adopted the 68 landmarks annotation convention from 300W [STZP13]). Many newer methods like [CBGB20, HYL+21, WBH+21] do not report results for AFLW, they only report results on datasets where all landmarks are always visible and are therefore not included in this table. Every heatmap-based method only reports results on the 19 central landmarks, leaving out of the evaluation the 2 most complex ones located near the ears. This might account for a portion of the NME_{full} (%) improvement.

Method	**Method family**	**N**	**#Landmarks**	**NME_{full} (%)**
Hyperface [RPC17]	Coordinate regression	24,993	21	2.93
HRNET [WSC+20]	Heatmap-based	20,000	19	1.57
3FabRec [BW20] *all data*	Heatmap-based	20,000	19	1.87
3FabRec [BW20] *1% of data*	Heatmap-based	200	19	2.38
LUVLi [KMM+20]	Heatmap-based	20,000	19	2.28
3DDFA v2 [GZY+20]	Deformable 3D mask	61,225*	21	4.43

Table 2 shows that a top-performing coordinate regression method (*i.e.*, Hyperface with Resnet-101 as the backbone network [RPC17]) was eventually surpassed

by many heatmap-based methods. Even 3FabRec trained on only 1% of the data, performs better. This may be due to the fact that the inductive bias of U-shaped networks used for generating heatmaps aligns better with the landmark localization problem than convolutional coordinate regression methods. On the other hand, deformable 3D mask methods, like 3DDFA v2 [GZY+20], have worse NME_{full} (%) values, but they offer advantages such as the ability to infer self-occlusions and perform 3D operations on the predicted mask.

II.3 Face aging

Figure 14: Age progression of a person's face, generated using the OpenAI ChatGPT image generation model (`https://openai.com/index/dall-e-3/`). This illustration highlights significant changes across different ages that should be addressed in a face aging tool that can serve in forensic anthropology for the identification and aging of individuals.

Face age editing [FGH10, KSSS14, WCY+16], or aging, consists in automatically modifying an input face image to alter the age of the depicted person while preserving identity (see Fig. 14). Over the last few years, this problem has attracted a growing interest because of its numerous applications. In particular, it is used in the movie production industry to edit actors' faces or in forensic facial approximation to reconstruct the faces of missing people [DCD+23]. The advances in DL methods unlock the development of fully automatic edition algorithms that avoid hours of makeup and post-production retouching.

Recent DL approaches adopt an encoder-decoder architecture [ABD17, HKSC19, MHP21, OESF+20, WCY+16, WTLG18, YPN+21, ZSQ17]. The image is encoded in a latent space that can be modified depending on the target age and fed to a decoder that generates the output image. The overall network is usually trained using a combination of losses that assess image quality, identity preservation, and age matching. However, despite the success of all these approaches, face editing remains challenging, and current methods usually fail when faced with sizeable differences

between the age of the person displayed in the input image and the target age. Indeed, most approaches [ABD17, HKSC19, WCY+16, WTLG18, YPN+21, ZSQ17] only superficially modify the skin's texture while the face's shape is kept unchanged. These approaches fail with significant age gaps since face shape can change significantly during a lifetime. Few methods try to go beyond some limited age gaps, but they either consider only a tightly cropped face region [KSSS14, WCY+16] or require specific pre-processing involving an image segmentation step [OESF+20]. Furthermore, some methods [WTLG18, APCO21] add an identity term to the total loss to better ensure the preservation of the identity during the translation process. All these methods principally differ in the choice of the network architecture and the manner the latent representation is manipulated. For instance, Wang *et al.* [WCY+16] introduce a recurrent neural network to iteratively alter the image, while in [YPN+21], the latent image representation is modified using a simple affine transformation. Re-AgingGAN [MHP21] employs an age modulator that outputs transformations that are applied then to the decoder, and Or-El *et al.*[OESF+20] adopt a multi-domain translation formulation, showing that segmentation information can be leveraged to improve aging.

A close research area is I2I translation. This consists in learning a mapping between two visual domains. In the pioneering work of Isola *et al.*[IZZE17], an encoder-decoder network is trained using a dataset composed of image pairs from the two domains. Later, many works addressed I2I translation in an unpaired setting, assuming two independent sets of images of each domain [FGW+19, LBK17, ZPIE17]. These works, of which cycleGAN [ZPIE17] is a paradigmatic example, mainly focus on introducing regularization mechanisms when training the I2I translation models. Another research direction is designing more advanced architectures to improve image quality or obtain several possible outputs for a given input [HLBK18, LTH+18, ZZP+17]. Disentangling style and content information has led to both higher image quality and diversity [HLBK18, PZW+20].

Style-based architectures recently attracted much attention for the problem of unconditional image generation. In particular, StyleGAN2 [KLA+20] is now used in many face manipulation tasks [RAP+21, YNGH21]. In the case of face aging, [APCO21] uses a pretrained StyleGAN2 model equipped with a pSp encoder [RAP+21], and an age classifier [RTG18] to tailor an age editing model with unlabeled data. In StyleGAN2, a network maps a Gaussian latent space onto style vectors. These vectors are later combined via a convolutional network to produce the output image. Finally, the synthesis network aggregates the style vectors through modulation operations.

A prominent new research path in the general image editing problem consists of employing masking mechanisms or attention maps to preserve relevant parts in the input image [ALTK19, KKC21, PAM+18, TXSY19]. For instance, mask consistency is employed in [KKC21] to improve multi-domain translations where masks are estimated using the Guided Backpropagation (GB) algorithm [SDBR15]. In the case of facial images, a mask is employed in GANimation [PAM+18] to different regions that should be preserved and those that should be modified to change the facial expression. In GANimation, masks are predicted by the main network, while an auxiliary network and GB are used to obtain the mask.

II.4 Text-guided image editing

Figure 15: Prompt-based image editing: the user can add, omit, change, or enhance elements in an image by providing a descriptive prompt of the original image and indicating the words that must be removed (in **red**) or added (in **blue**).

Text-guided image editing is an emerging technique in CV that allows for the modification of images based on natural language descriptions. This method leverages advanced neural networks, such as GANs or Diffusion Models, to interpret textual instructions and apply corresponding visual changes to images. This paradigm enables users to input textual instructions, such as "make the sky bluer" or "turn the car red", which are then processed by advanced neural networks to produce the desired modifications in the image [MHA+23]. This capability extends beyond traditional DL applications, such as face aging, where a model is trained to predict and render age-progressed facial images [OESF+20]. While DL face aging focuses on a specific transformation related to the passage of time and its effects on facial features, text-guided image editing encompasses a broader range of modifications dictated by diverse textual inputs. Therefore, text-guided image editing can be viewed as supertool that encompasses face aging among many other tasks, providing a flexible framework for various image manipulations based on descriptive text, rather than being limited to age-related changes.

II.4.1 Text-guided image synthesis

Prior to discussing text-guided image editing, we should talk about the advancements in text-guided image synthesis through diffusion models, which have gathered considerable attention due to their ability to achieve remarkable realism and diversity [SCS+22, RBL+22]. These large-scale models enable image generation from text prompts and have unlocked a new level of creativity. As a result, research is intensifying around the use of these models to manipulate images for editing purposes.

One of the most striking innovations is the possibility of editing images through intuitive text prompts, offering users the power to modify images without professional editing skills. We focus on the prompt-based image editing task as formulated in [MHA+23]: a user provides an image alongside its textual description. Then, by simply indicating changes in the sentence, the user can instruct the model to add, omit, change, or enhance elements (see Fig 15 for example). The models implicitly determine which areas of the input image are irrelevant to the target task and should be reconstructed, and which areas require being altered while preserving the relevant identity and geometry.

The SOTA methods for the prompt-guided editing task need the inversion of the target image (later showcased on Fig. 32). Although inversion processes have greatly improved within GANs, they remain a significant hurdle in diffusion models due to their iterative sampling process. Current techniques [MHA+23] require repetitive optimization steps, resulting in excessive computational demands with even moderately-sized images (512×512), taking upwards of a minute to process per image. Alternatives that reduce computational workload [MHA+23, MIST23, PKSZ+23] often compromise on reconstruction quality, which results in unsatisfactory alterations of the input image.

With the impressive advancements in text-to-image diffusion models [RBL+22], there has been a growing interest in exploring image editing using pre-trained diffusion models. These studies have presented several editing solutions where the user can guide the generated image through various inputs. For instance, SDEdit [MHS+22] allows users to apply brush strokes to areas they wish to edit. The model then injects random noise into these targeted areas and uses the diffusion process for denoising. To create new images from examples, techniques like Textual Inversion [GAA+23] and Dream-Booth [RLJ+23] employ gradient-descent-based optimization to learn personalized concepts. Text-based editing, in particular, has garnered considerable interest due to its intuitive and user-friendly interaction style. In this domain, DiffusionCLIP [KKY22] uses DDIM inversion [SME21] to reverse the diffusion process and applies fine-tuning. This approach guides the generation with a CLIP-based loss to align the generated image more closely with the intended edit. Another method, as demonstrated in ControlNet [ZRA23], involves conditioning the generation process on the edges or pose information extracted from the input image. This technique aims to generate an image retaining the original spatial structure but styled (in a broad sense) according to the given prompt.

In the prompt-based editing task [MHA+23] (see Fig. 15), a user provides an image along with a textual prompt $\mathcal{P}^{in}$ which describes the input image. The user can then instruct the model to add, remove, change, or enhance elements in the image by providing a target prompt $\mathcal{P}^{out}$ corresponding to the desired image (also called positive prompt). This problem formulation has inspired several subsequent studies [MHA+23, PKSZ+23, TGBD23, JZB+23].

Part II

Proposals

Chapter III

Cascade of convolutional models for cephalometric landmarks localization

"We all give ourselves a lot of leeway, but we want consistency from other people." — Richard Linklater

III.1 Introduction

In this section, we present a tool for the automatic location of cephalometric landmarks on in-the-wild images as described in II.2.1, which can be later integrated into a forensic human identification procedure. This work focuses on 30 landmarks in the head's frontal and lateral views. The complete list is reported in Table 1 and is also displayed in Figure 12. We automate the annotation/location of cephalometric landmarks on photographs, determining their position as well as a primitive visibility estimation. We employ ML techniques [Bis06] and, in particular, DL approaches [GBC16, Chapter 1] to acquire robust knowledge about the complex nature of the landmarks to be located.

III.2 Materials

Given the challenges discussed in Sec. II.2.1, we focus on the well-established problem of facial landmark localization in CV. To this end, we take the AFLW dataset [KWRB11] as a related problem to our research. Given the similarities between this problem and our research goal, methods developed to solve this problem may be either a potential solution or a starting point for ours.

III.2.1 Available datasets

For this work, we have access to two different datasets available: a training dataset, which comprises both real-cases images and pictures taken under controlled

conditions, as well as a user-study dataset, which only includes images from actual forensic cases. The training dataset has been compiled over time for various forensic tasks and by multiple forensic experts, while the user-study dataset was built following a uniform protocol. Both datasets were annotated by forensic experts from the *Physical Anthropology Lab at the University of Granada* and the *Panacea Cooperative Research* company in a joint effort.

Training dataset

- 165 images from different subjects, ranging from 1 to 7 photographs per subject. The images correspond to 104 males (of European ancestry) and 61 females (primarily European).
- Resolutions range from 156 to 4350 pixels on the wider side.
- Up to 30 annotated landmarks per image, with a total of 3526 individual landmarks annotations. See Figure 16 for landmark frequency distribution.
- The images broadly differ in quality: we can find good quality frontal and lateral portraits, *in-the-wild* images, and scanned historical pictures.
- As this dataset was heterogeneously built over time for different forensic tasks and by different forensic experts, the absence of a landmark annotation in the dataset does not necessarily mean the landmark is not visible in the corresponding image (this will be further discussed in Section III.3.1).

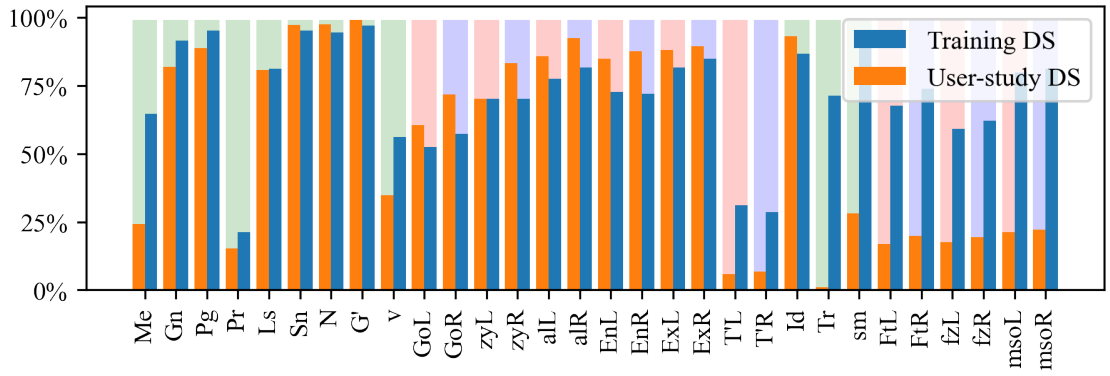


Figure 16: Proportion of landmark presence by image depending on the dataset. The background of the bars is color-coded: in green, those landmarks in the sagittal plane; in red, the left side of the face; in blue, the right side of the face.

A model successfully trained on such a diverse dataset should be robust enough to face most of the challenges of day-to-day work in a forensic environment. However, the small size of this dataset could also create problems with uncommon poses or underrepresented landmarks. For example, some landmarks, such as the *Prosthion* or the *Tragion*, have only 25 and 40 examples, respectively. The most frequent landmarks are those located on the sagittal plane, but the landmarks located on both sides of the face are also necessary to estimate a more robust model of the head that can be required in some tasks, such as craniofacial superimposition.

User-study dataset

We conducted a user study on a different set of 46 images. All of them come from real forensic cases, and each one is annotated by up to seven different forensic experts (5408 individual landmark annotations in total). The distribution of landmarks is uneven (see Figure 16). Some landmarks are similarly represented as in the training dataset, whereas others appear at a lower frequency (*e.g.*, *Frontotemporale* and *Frontozygomaticus*). The *Trichion* is not annotated, resulting in 29 landmarks in total. The dataset is similarly distributed regarding gender, but the population is older than in the training dataset, and the pose is more frequently frontal (see Table 3).

Table 3: Dataset statistics extracted from the automatic analysis performed with *Face++ Face Detection* API [1].

		Gender		**Age**					**Head Yaw**
	N	Male	Female	Min	P_{25}	Median	P_{25}	Max	$\|r\| > 30°$
Training dataset	165	106	62	14	35	44	61	89	22%
User-study dataset	46	26	20	33	58	66	71	86	7%

III.3 Methods

III.3.1 Design considerations

The design choices for our solution are informed by two main factors: the available training data and the characteristics of the task at hand.

Firstly, we have a relatively small dataset (165 images) compared to AFLW (25,000 images) with no restrictions regarding pose, lighting, or other aspects of the scene or photographic medium. For these reasons, it is unlikely to train an entire high-resolution, ad-hoc convolutional network due to the large amount of data required by these methods. Additionally, the heterogeneous data collection process with no consistent criteria determining the visibility of a landmark makes it challenging to automatically learn a visibility-prediction model as in [KMM+20]. For example, a non-annotated landmark could be differently motivated [CAIN+14b]: *1)* The landmark is occluded by an object or not visible because of the pose; *2)* The forensic expert would not be able to annotate it confidently because of the low quality of the image or the presence of a significant amount of soft tissue (i.e., skin or fat); or *3)* The landmark was not required for the intended task, and the forensic expert did not annotate it. As a result, we cannot rely on the available data to train a classifier on landmark visibility.

Secondly, the proposed solution should provide a robust and accurate landmark localization system that can detect and localize high-resolution landmarks and consider a heterogeneous source of images (see Fig. 13). A solution that relies solely on low-resolution heatmaps may not be sufficient to meet this requirement. Additionally, the inter- and intra-subject dispersion observed in the landmark annotation

[CAIN+14b] varies among landmarks, which should be considered during the design, training, and validation of the method. Finally, the proposed solution should be evaluated against other methods and human experts to assess its performance.

III.3.2 Method description

Figure 17: Full process step by step as described in Section III.3.2 (zoom in for details). Notice the different crop sizes depending on the landmark as well as the significant improvement in accuracy.

Our proposed method FSCNet (Few-Shot Cephalometric landmarks localization NETwork) involves a cascade of several steps (see Fig. 17 for a detailed graphical description of the whole procedure) involving two modules. Firstly, a pre-trained deformable 3D mask model [GZY+20] is used to suggest a reliable initial landmark location. Secondly, a cropped image around each suggested landmark is run through a residual network [HZRS16] (*i.e.*, ResNet-18) trained to predict the displacement between the center of the input cropped image and the landmark localization. In this way, we highly increase the resolution of the model. As with many other few-shot approaches, we use a model trained on a similar task and apply *transfer learning* to reduce the size of the hypothesis space and augment the supervised experience [WYKN20]. Every step is optimized or trained on the same training dataset.

A description of the 4 steps of the method is provided as follows:

Step 1 **Deformable 3D mask.** A pre-trained 3DDFA v2 model [GZY+20] outputs a face mask consisting of a mesh with approximately 40,000 3D coordinates.

During the training process, we can identify the best-fitting mesh point for each corresponding landmark and use it as the initial landmark location. There are two main advantages to this approach:

(a) We can use a robust network that has been pre-trained on the 300W-LP dataset [ZLL+16], which consists of 122,450 augmented images captured in real-world conditions with a wide range of poses.

(b) We can use the mesh point's normal vector to estimate the landmark's visibility based on whether it is pointing toward or away from the camera. However, we assume that this visibility estimation may be imprecise in both external occlusions (*e.g.*, an object blocking the view of the landmark) and facial self-occlusions (*e.g.*, a point on the cheek being occluded by the nose, even though its normal vector is pointing towards the camera).

Step 2 **Landmarks outside the mask.** Some facial landmarks, such as the *Vertex*, *Trichion*, and *Tragion*, are located outside the region captured by the 3DDFA v2 model, such as on the top of the head or ears (as shown in Figures 17 and 18). To improve the accuracy in the location of these landmarks, we first determine the closest mesh point as described in *Step 1*. Then, we apply a two-step process that involves the composition of two 12-parameter affine matrices, N_i and T_j, for each image i and each landmark j.

The matrix N_i is obtained by solving a system of equations that normalizes 3DDFA's output mask to common rotation, scale, and position. The affine transformations T_j are obtained through differential evolution optimization [SP97] to minimize the mean distance between the transformed prediction and the ground truth for the corresponding landmark j. We use the differential evolution algorithm for each of the 30 landmarks to optimize and evaluate a unique matrix T_j. This algorithm is trained and validated only on the training subset for each cross-validation fold. We keep only those T_j matrices that successfully reduce the mean distance. In practice, we only modify the *Vertex* and *Trichion* landmarks using this process, as shown in Figure 17. This process is illustrated in Fig. 18.

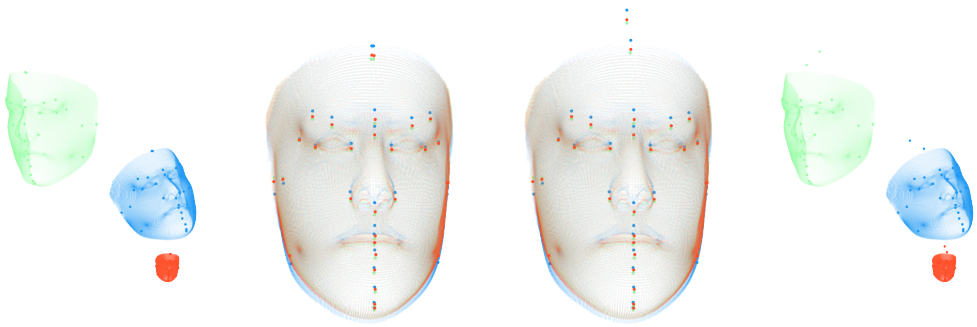


Figure 18: Out-of-the-mesh landmark optimization steps. Every landmark coordinate vector $\mathbf{y}_{ij}$ (for the i^{th} image and the j^{th} landmark) is transformed by the described affine transformations (zoom in for details). In this example, each image mesh is represented in a different color (*i.e.*, green, blue, and red), normalized by N_i, individual landmarks are transformed by T_j, and lastly *denormalized* back by N_i^{-1} to obtain the final landmark coordinates.

Step 3 **Region-of-interest (ROI) cropping.** We use the landmarks obtained in the former step for image cropping. A square crop will be made for every landmark position (see *Step 3* in Figure 17 where a different crop is made around every single landmark), and the side of this crop will be determined by the landmark error distribution obtained after *Step 2*. More specifically, the crop depends on two values: the scaling factor s_j for the landmark j (see Eq. III.1) and the widest side of the bounding-box c_i that encloses face i. The resulting formula is $\text{side}_{ij} = \lambda s_j c_i$, where $\lambda = 2.5$ is a constant obtained in preliminary experiments to optimize both the context (wider crop) and resolution (tighter crop) needed for the prediction, whereas P_{90} is the 90^{th} *percentile* of the error distribution.

$$s_j = P_{90}\{||\mathbf{y}_{ij} - \mathbf{y'}_{ij}||_\infty, \text{ for every image } i\} \tag{III.1}$$

Step 4 **Residual network and label projection.** Finally, every landmark crop is run through a reduced version of ResNet-18 [HZRS16] pre-trained on ImageNet [RDS+15] with three residual blocks instead of four to take as input 32×32 pixels images instead of 224×224 as it was originally trained for ImageNet classification. In addition, the final classifier fully-connected layer has been replaced with a 2-neuron fully connected layer that predicts the distance from the crop center to the landmark.

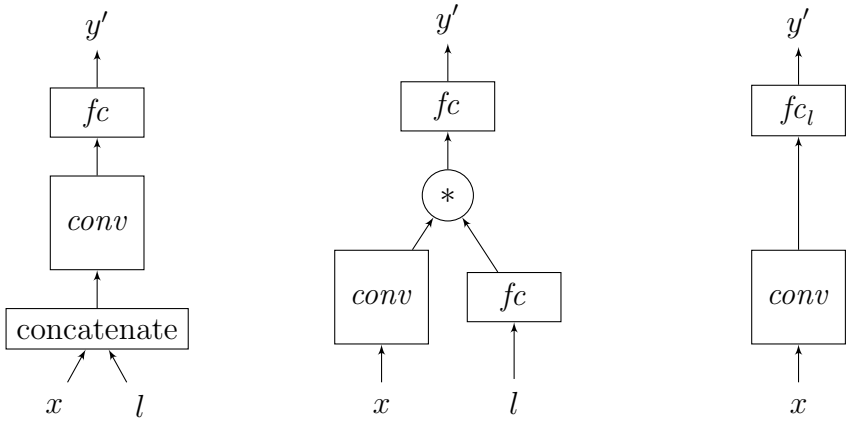
We evaluated three different alternatives for incorporating label information, represented by the landmark $l \in \{1 \dots 30\}$ and the image tensor $x \in \mathbb{R}^{W \times H \times 3}$, into a model:

(a) In the first alternative, depicted in Fig. 19a and similar to [MO14], the label l is encoded as a $W \times H \times L$ tensor, where L is the number of landmarks. Every channel in the L-dimensional label tensor is set to 0 except for the channel corresponding to the label l, which is set to 1. It is then concatenated with the image tensor x to form a $W \times H \times (L + 3)$ tensor, *i.e.*, L in addition to the 3 RGB channels.

(b) In the second alternative, depicted in Fig. 19b and similar to the conditional discriminator of [MK18, KAH+20], the label is projected to a 256-D embedding (the number of output channels of the convolutional network), and an element-wise product is performed with the convolutional $256 \times 1 \times 1$ output before the fully connected layer.

(c) In the third alternative, depicted in Figure 19c, a different fully-connected layer is trained for each label l after a common residual convolutional network.

A single convolutional ResNet-based regressor is trained for all landmarks instead of different networks for each one. By training a single network on multiple landmarks we increase non-artificially the number of data samples the model is trained on (besides the artificial data augmentation described in Section III.3.3.2). When training on full faces or training a model for each landmark, only one training sample is available per labeled face image. However, cropping a different training sample around each landmark and sharing the same conditional network, results in up to 30 training samples per labeled face image, reducing the negative impact that small datasets have in DL solutions.



(a) Concatenation (b) Projection (c) Multi-head

Figure 19: Different label conditioning alternatives for the landmark l and the image tensor $x \in \mathbb{R}^{W \times H \times 3}$.

III.3.3 Evaluation protocol

III.3.3.1 Validation

Two kinds of validations are developed:

1. A 5-fold cross-validation is performed on the training dataset. In each of the five experiments, the model is trained on four of the five equally-sized splits of the dataset and tested on the remaining one. The results of the five experiments are aggregated to obtain an unbiased prediction for every sample. This validation is used for the ablation study and benchmarking against SOTA methods.

2. A user study is performed on a different dataset, where forensic experts are repeatedly presented with two annotations for the same landmark (one from the human annotators and one from our automatic model trained on the training dataset) and asked to select the most accurate one.

III.3.3.2 Metrics

Four metrics are used for the evaluation of our method's performance:

1. *Normalized Mean Error (NME)* as described in Eq. II.2. To avoid potential biases, the bounding box used during training (the same as in [GZY+20]) and the one used for the metrics calculation (obtained through the Face++ Face Detection API [1]) are different.

2. *Mean relative performance order (ROrder).* Not every landmark error should be compared against each other because the inter/intra-subject variation depends on each landmark [CAIN+14b]. For example, in the user-study dataset,

[1] Face++ Face Detection API documentation: `https://www.faceplusplus.com/face-detection/` *Last accessed on July 26, 2024.*

the inter-subject variation is up to 9 times higher depending on the specific landmark. This metric is used as one of the global key performance indicators for ablation and SOTA comparison. It is defined as the mean relative position o rank of the NME_j for each landmark j among every compared method.

For example, if we compare three methods for landmark detection on three different landmarks, and their NME values are as follows:

- Method A: 0.05, 0.10, 0.08
- Method B: 0.04, 0.11, 0.09
- Method C: 0.06, 0.09, 0.07

The ROrder for each method is calculated by ranking the NME values for each landmark and then averaging these ranks. Method A might have ranks [1, 1, 2], Method B [2, 3, 3], and Method C [3, 2, 1]. The mean rank gives us the ROrder, showing the overall performance across landmarks. Therefore, the ROrder would be 1.3, 2.7, and 2 respectively.

3. *Mean relative NME (RNME).* Equal weighting of landmark prediction losses is a common assumption, but research suggests that annotator dispersion varies by landmark [CAIN+14b, HYL+21]. To aggregate the NME values for each landmark and to prevent the error to become mostly determined by the landmark with the most significant variation, every NME_j value is divided by the minimum NME_j value for landmark j among comparing methods before the mean is calculated.

4. *Wilcoxon signed-rank test.* It is a non-parametric statistical hypothesis test that compares the location or shift of the distribution of two matched populations [Wil45]. Its role in the validation is to assess statistical differences between competing methods [DGMH11].

The motivation behind these metrics is to capture the complexity of cephalometric landmark validation, which is generally overlooked in facial landmark comparisons. NME accounts for the Euclidean distance between predicted and ground truth positions, normalized by the bounding box size to avoid biases. ROrder ranks each method based on their NME for different landmarks, reflecting overall performance despite inter/intra-subject variations. RNME adjusts NME values by their minimum to prevent dominance by landmarks with significant variation. The Wilcoxon signed-rank test statistically assesses differences between methods, providing a robust validation framework.

Every experiment is run on a single *NVIDIA GeForce RTX 3090* GPU for 850 epochs with a batch size of 60, a learning rate of $4e^{-4}$, and the following data augmentation scheme: random image rotation between -5 and 5 degrees and random horizontal and vertical shifts of up to 4 pixels. For *Step 4* the whole ResNet model is fine-tuned as we do not freeze any layer.

III.4 Experiments

III.4.1 Ablation study

Table 4: Ablation study of each incremental contribution. *Step 3* is presented after the rest because it is based on the best-performing method for *Step 4*, the projection approach. The Wilcoxon test result is omitted as every *p-value* is lower than $1e^{-40}$.

	Configuration	*NME (%)*	*ROrder*	*RNME*
Step 1	Best 3DDFA v2 [GZY+20] landmark	3.49	6.06	3.05
Step 2	Out-of-the-mask optimization	2.87	6.78	2.74
Step 4	Convolutional refinement (Fig. 19)			
	Multi-head	2.56	5.16	2.30
	Concatenation	1.34	3.53	1.22
	Projection	**1.10**	**1.22**	**1.02**
Step 3	Best w/o landmark-error crop adjustment	1.24	2.69	1.19
∗	Best w/o augmentation	1.26	2.56	1.16

Table 4 demonstrates the incremental impact of each architectural decision on the model performance. Each step results in a significant increase in performance. The only minor change seems to be from *Step 1* to *Step 2*, but this is because it affects only two of the thirty landmarks: the *Vertex* and *Trichion*. For these two landmarks, the NME (%) decreases from 21.25 to 2.50 and 7.78 to 1.99, respectively.

Every convolutional refinement network (illustrated in Fig. 19) improves the performance of *Step 2*. However, the label projection approach [MK18] (Fig. 19b) outperforms the other two by a significant margin. Furthermore, we also found that using data augmentation improved the quality of the results significantly, as the best configuration without data augmentation resulted in a 14% decrease in performance as measured by the RNME metric. Additionally, using a single global scale factor instead of a by-landmark error-based scale factor s_j in step *Step 3* resulted in a 17% decrease in performance.

III.4.2 Comparison with State-of-the-Art Methods

In Table 5, our method FSCNet is compared with the three SOTA methods with code and pre-trained networks available: 3FabRec[2] [BW20], HRNET[3] [WSC+20], and LUVLi[4] [KMM+20]. In every case, we start from the pre-trained weights provided by the authors, replace the last layer to predict 30 landmarks instead of 19, and fine-tune the model on our cephalometric dataset. The used hyperparameters are taken from each original work and every model is trained until convergence.

[2]Code and weights are available at: `https://github.com/browatbn2/3FabRec`. *Last accessed on July 26, 2024.*

[3]Code and weights are available at: `https://github.com/leoxiaobin/deep-high-resolution-net.pytorch`. *Last accessed on July 26, 2024.*

[4]Code and weights are available upon request at: `https://www.merl.com/research/license/LUVLi`. *Last accessed on July 26, 2024.*

[5]3DDFA v2 is not fine-tuned on our data, its result correspond to the *Step 1* in Table 4.

Table 5: Comparison of the performance with the SOTA in facial landmark detection, the Wilcoxon test result is omitted as every *p-value* is lower than $1e^{-300}$.

Model	*NME (%)*	*ROrder*	*RNME*
3FabRec	3.71	4.88	3.35
3DDFA v2 [5]	3.49	4.12	2.97
HRNET	2.21	2.59	2.04
LUVLi	2.37	2.41	2.03
FSCNet (Ours)	**1.10**	**1.00**	**1.00**

HRNET and LUVLi are high-performing methods for annotating facial landmarks, but their results in our problem are significantly worse. Their RNME is two times worse than ours. Moreover, none of them is designed to perform on a few-shot scenario as our problem requires because of the few training data (165 images compared to AFLW's 25,000 examples). On the other hand, 3FabRec, a SOTA few-shot method, is surprisingly the worst performer. Even though several fine-tuning approaches were evaluated for 3FabRec (fine-tuning both the unsupervised and supervised training stage or only the supervised one as in the original paper), it performs 3.2 times worse than our method and approximately 1.6 times worse than LUVLi or HRNET.

We believe that the poor performance of analyzed SOTA models in our problem is because they are not designed for high-resolution localization. All of the heatmap-based methods that were compared are limited by the resolution of both the heatmap output and the input image fed to the convolutional network, as they are essentially U-shaped networks (*e.g.*, HRNET and LUVLi use 64×64 heatmaps). It can be observed in Table 5 that HRNET and LUVLi NME (%) performance (2.21 and 2.37 respectively) is close to their NME (%) values on AFLW (1.57 and 2.28 respectively). Nevertheless, cascade high-resolution approaches might have not been widely adopted in facial landmark literature as available datasets lack sufficient image resolution and landmark precision [CBGB20]. Our analysis supports the conclusions of [CBGB20], where they were able to perform better than SOTA on their high-resolution dataset by using attention-driven cropping, but their performance on the low-resolution 300W dataset was comparable to SOTA results.

Another difference in our data is the stability of the positions. Recent works, such as [CBGB20, HYL+21, WBH+21], are only validated on datasets [STZP13, WQY+18] that feature diverse poses but assume that all landmarks are always visible. For example, [WBH+21] even refines its initial anatomically correct output to account for this anatomical ambiguity. Furthermore, the usual benchmark metric assumes that all landmark prediction losses should be given equal weight. However, research such as [CAIN+14b] has shown that annotators' dispersion varies depending on the specific cephalometric landmark, and [HYL+21] has also shown this to be true for facial landmarks.

III.4.3 User study

For this study, six forensic experts were presented with two landmark annotation locations for comparison in a blind, two-way evaluation. Each expert was shown a

single image twice, displayed side-by-side, with one version depicting the predicted landmark location and the other showing the annotation of another expert. The experts could zoom in on both high-resolution images simultaneously to examine the annotation locations in greater detail. The experts were asked to select from one of three options: "Left better", "Right better", or "Both equally good". The left and right positions were randomly assigned for each image to prevent bias. A total of 2600 comparisons were conducted among all six experts, and the results are displayed in Table 6.

Table 6: User study results comparing the performance of human experts against our model.

	G'	**Gn**	**Id**	**Ls**	**alL/R**	**EnL/R**	**ExL/R**	**FtL/R**	**fzL/R**	**GoL/R**
Human better	73	63	**77**	**60**	118	107	112	**43**	28	98
FSCNet equal or better	**77**	63	63	53	**144**	**142**	**144**	10	**35**	**111**

	msoL/R	**T'L/R**	**zyL/R**	**Me**	**N**	**Pg**	**Pr**	**Sn**	**sm**	**v**
Human better	**38**	**7**	**141**	**22**	72	72	**19**	69	**22**	**32**
FSCNet equal or better	37	4	120	15	**103**	**73**	3	**87**	21	22

Even though we do not achieve expert-like performance, our model seems to perform as well as a human expert in half or more cases. Our results for *equal or better* are higher than *human better* in ten out of the twenty landmarks (or 14 out of 29 if we distinguish left from right). In some landmarks, this difference was more significant (*e.g.*, *Alare*, *Endocanthion*, or *Exocanthion*), whereas in others there is no appreciable difference (*e.g.*, *Gnathion*, *Midsupraorbital*, or *Pogonion*). The most challenging landmarks seem to be the *Prosthion*, probably because of the low number of examples, and the *Frontotemporale*, because of the difficulty of its accurate annotation. These are the only two landmarks where human annotations are distinctly preferred.

III.4.4 Visibility estimation

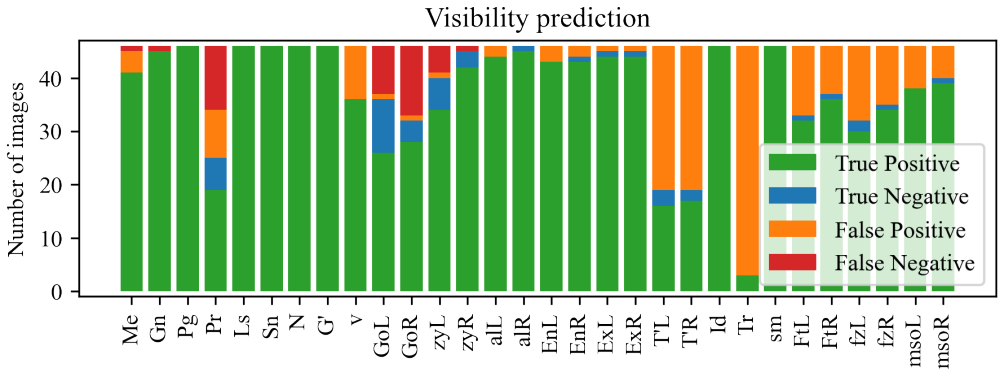


Figure 20: Visibility estimation results on the user study dataset. A landmark is defined as visible if at least one of the annotators marked it as visible.

The visibility prediction analysis is complex due to the need for consistent criteria for identifying visible landmarks. For example, when examining agreement in the user study dataset, we found a mean Cohen Kappa score [McH12] of 0.26, with 15

out of 29 landmarks receiving scores lower than 0.1, indicating a lack of agreement. Therefore, we have established that a landmark is visible in the user study dataset if at least one annotator marked it as visible. Using this criterion, we have achieved an average accuracy of 83%. However, most landmarks were predicted as visible as seen in the confusion matrix (see Fig. 20). This result could be due to two factors: *1)* The majority of faces in the dataset are frontal (as shown in Table 3), and *2)* The use of the normal vector to determine visibility results in high sensitivity but low specificity.

Chapter IV

Custom Structure Preservation in Face Aging

"I'm always pushing back against the last thing I did in some way, and some of that is restlessness and a sense of limited time." — Alex Garland

IV.1 Introduction

As discussed in Sec. II.3, face aging focuses on modifying facial images to alter age and other attributes like beard, hair, and color while preserving identity. Recent DL approaches use encoder-decoder architectures to encode images into a latent space, manipulate the content, and decode the altered image. These methods often employ a combination of loss functions to ensure image quality, identity preservation, and accurate age transformation. However, challenges remain, particularly in handling significant age gaps and changes in facial shape. To enhance these transformations, some methods incorporate segmentation information [OESF+20]. Additionally, style-based architectures like StyleGAN2 [KLA+20] and attention-based techniques further improve the precision and realism of facial editing.

This chapter proposes a novel framework that allows profound structural changes in facial transformations. Our work achieves realistic image transformations with age gaps that imply changes in head shape or hair. In addition, we argue that the face editing task is an ill-posed problem because every person gets older in a different and non-deterministic way: some people drastically change, while others are easily recognizable in old photographs. In this sense, we propose a methodology that allows the user to adjust, at inference time, the degree of structure preservation. Thus, the user can provide an image and obtain different transformations where the structure (*i.e.*, face shape or hair growth) is preserved at different levels. Fig. 21 shows some qualitative results obtained with our method. Furthermore, the user can choose different degrees of structure preservation: with high preservation, the model only changes the texture, while with lower preservation, the shape of the face is also modified.

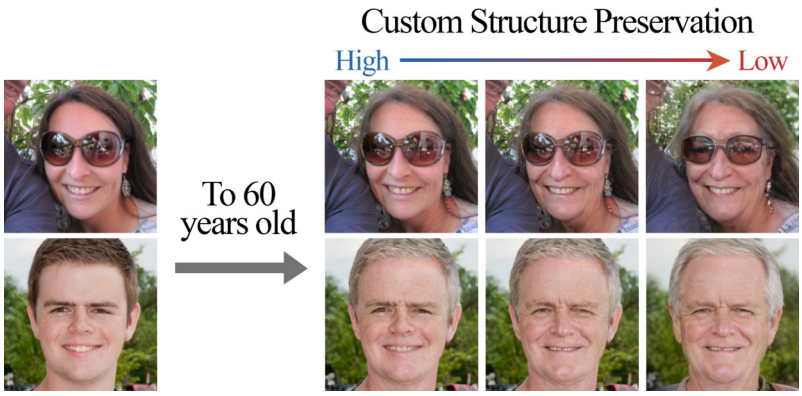


Figure 21: The user can choose the degree of structure preservation at inference time. Facial morphology transformations are more profound as we move to the right (lower structure preservation).

We adopt an encoder-decoder framework similar to [HKSC19, YPN+21]. However, our approach goes beyond existing methods that generate a single image for a given image-target age pair. Indeed, we offer the user the possibility to adjust the degree of structure preservation during translation, and, in this way, we can output a set of plausible resulting facial images. Our method also leverages recent advances from the **I2I translation** research area. We take inspiration from the StyleGAN2 generator to design a novel decoder that combines the input style and the target age with the content representation via weight demodulation [KLA+20]. We disentangle style and content as in [HLBK18, PZW+20] in order to allow custom structure preservation. Thanks to this strategy, our CUstom Structure Preservation (CUSP) module can act on the spatial information passing through the content branch while preserving style information. Regarding the more general **image editing** problem, our method shares similarities with several approaches employing masking mechanisms or attention maps to preserve relevant parts in the input image [ALTK19, KKC21, PAM+18, TXSY19]. As in [KKC21] our masks are estimated using the GB algorithm [SDBR15].

The contribution of this research can be summarized as follows:

- We propose a novel architecture for face age editing that can produce structural modifications in the input image while maintaining relevant details present in the original image. We take advantage of recent advances in I2I translation [HLBK18, LTH+18] and unconditional image generation [KLA+20] to design our architecture. We disentangle the style and content of the input image, and we propose a new decoder network that adopts a style-based strategy to combine the style and content representations of the input image while conditioning the output on the target age.
- We go beyond existing aging methods allowing the user to adjust the degree of structure preservation in the input image at inference time. To this aim, we introduce a masking mechanism, through a so-called CUSP module, that identifies the relevant regions in the input image that should be preserved and those where details are irrelevant to the task. Importantly, our mechanism for adjustable structural preservation does not require additional training supervision.
- Experimentally, we show that our method outperforms existing approaches

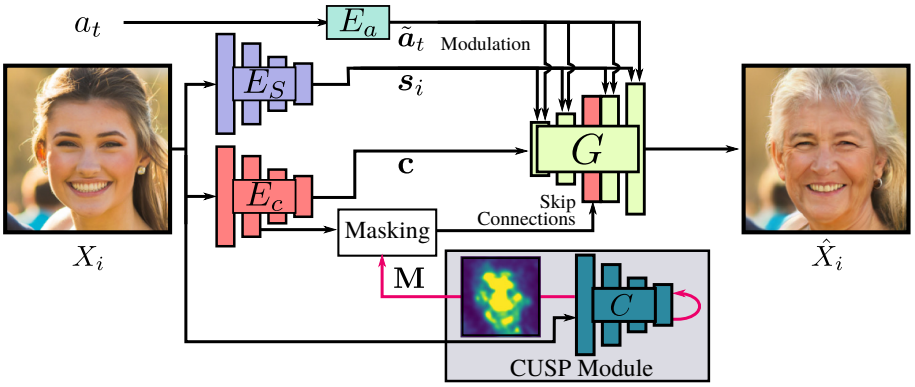


Figure 22: Illustration of the proposed approach. A style encoder E_s extracts a style representation of the input image $\mathbf{X}_i$. A content encoder E_c encodes spatial information. Target age a_t is embedded using a multi-layer perceptron E_a. Our generator G outputs the image $\hat{\mathbf{X}}_i$ by combining the input style and content representations conditioned on the target age. Our CUSP module predicts a blurring mask $\mathbf{M}$ applied to the SCs to allow the user to choose a CUstom level of Structure Preservation.

in three publicly available high-resolution datasets and demonstrate the effectiveness of our mechanism for adjusting structure preservation.[1]

IV.2 Methods

In this work, we address the face age editing problem. Therefore, our goal is to train a network able to transform an input image $\mathbf{X}$, such that the person depicted looks like being of the target age a_t. At training time, we assume that we have at our disposal a dataset composed of I face images of resolution $H \times W$, such that $\mathbf{X}_i \in \mathbb{R}^{H \times W \times 3}, i = 1, ..., I$ with their corresponding age label $a_i \in \{1, ..N\}$ (*i.e.*, an age annotation for each image). Note that the age labels are automatically obtained using a pre-trained age classifier. Similar to previous approaches [YPN+21, ZSQ17], we employ the DEX classifier [RTVG15].

One of the main difficulties lies in modifying the relevant details in the input image while preserving non-age-related regions. To this aim, we introduce a style-based architecture detailed in Sec. IV.2.1. In contrast to previous works, the CUSP module allows the user to indicate the desired level of structure preservation through two parameters: $\sigma_m > 0$ and $\sigma_g > 0$. These parameters act locally and globally, respectively, as later detailed in Eq. IV.2.

IV.2.1 Style-based Encoder-decoder

As illustrated in Fig. 22, our architecture employs five different networks: (1) A style encoder E_s extracts a style representation $\boldsymbol{s}_i$ of the input image $\mathbf{X}_i$. E_s discards any spatial information via global-average-pooling at the last layer. The use of a

[1]Code and pretrained models are available at `https://github.com/guillermogotre/CUSP`.

style encoder allows global information to be used at any location in the decoder. (2) A content encoder E_c outputs a tensor $\mathbf{c}$ describing the content of the input image. Contrary to E_s, the content encoder preserves spatial and local information. Both E_s and E_c share almost the same architecture but for the last layer. In our case, the use of separated style and content encoders is justified by the fact that our CUSP module should not affect the image style $\boldsymbol{s}_i$ but only the structure of the image. (3) An 8-layers fully connected network, E_a, embeds the target age a_t: $\tilde{\boldsymbol{a}}_t = E_a(a_t)$. (4) An image generator G estimates the output image $\hat{\mathbf{X}}_i$ by combining the style and content representations with the target age embedding $\tilde{\boldsymbol{a}}_t$. (5) Finally, our CUSP module allows the user to choose the level of structure preservation. This module predicts a mask $\mathbf{M}$ used to act on the SCs between the content encoder and the decoder. More precisely, we blur the regions indicated by the mask $\mathbf{M}$ to propagate only the non-age-related structural information to the decoder through the SCs.

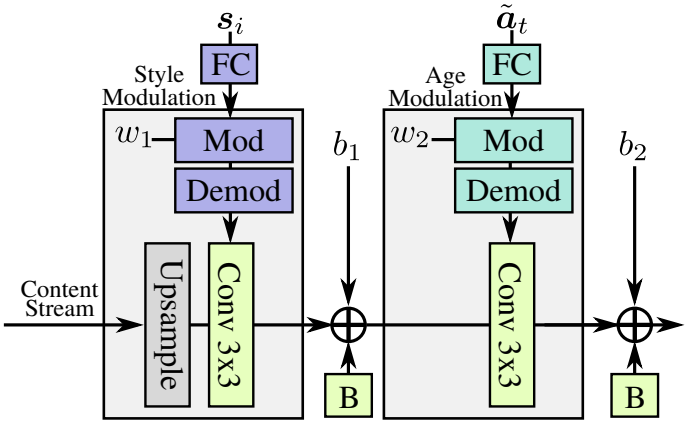


Figure 23: Illustration of the decoder blocks used in G . B denotes the addition of random noise, and FC denotes a fully-connected layer. w_1 and w_2 are two learned scaling parameters, while b_1 and b_2 are learned biases.

Our image generator G is designed to combine the outputs of the style and content encoders with the target age embedding. Its architecture is inspired by StyleGAN2 [KLA+20], which achieves SOTA performance in unconditional image generation. However, we provide several modifications to tailor the architecture to the aging task. G comprises a sequence of elementary blocks (see Fig. 23). Unlike [KLA+20], our decoder block takes three inputs: the former block output, the style embedding, and the age embedding. Each decoder block outputs an image twice the size of its input and is composed of two consecutive sub-blocks: the *style sub-block* and the *age block*. In the *style sub-block*, the input is upscaled through bilinear interpolation. Then the upscaled input is transformed through weight demodulation (w_1) based on a linear projection of the style embedding (s_i). In the second sub-block, the age embedding $\tilde{a}_t$ and w_2 are used for transforming the former sub-block output. Both s_i and $\tilde{a}_t$ are shared by every block. After each step, 0-centered random noise B is added to the output.

Note that all blocks are combined following the *input skips* architecture of StyleGAN2, where a layer named *tRGB* is introduced. Such layer predicts intermediate images at every resolution scaled and added to generate the final image. tRGB is also conditioned on the age embedding. Contrarily to U-Net [RFB15] that includes SCs in every layer, in our work we perform only a small set of SCs.

IV.2.2 CUSP Module

SC [RFB15] are efficient tools to provide high-frequency information from the input to the decoder allowing accurate reconstruction [IZZE17]. High frequencies carry accurate spatial information that favors pixel-to-pixel alignment between inputs and outputs, as, for instance, needed in segmentation. However, previous works [SSLS18] show they are not suited for tasks where the input and output images are not pixel-to-pixel aligned. For example, input and output images are aligned when the age gap is small in the aging task. However, this assumption does not hold in every image region with significant gaps. This misalignment is particularly predominant in areas other than the background since facial morphology or hairstyle may change.

Therefore, we propose to control the amount of structural information that flows through the SC. This control is obtained by blurring the feature maps going through them. Nevertheless, every region should not be treated in the same way. For instance, depending on the task, the user may prefer to preserve the background while blurring the foreground to loosen conditioning on the input image in this region. Therefore, we propose a specific mechanism to identify relevant image regions for the translation.

Mask Estimation. We employ an additional classification network C, pretrained to recognize the age of the person depicted on an image. We use the DEX classifier [RTVG15] again. Since DEX is pretrained on 224×224, the input image is rescaled to this resolution. Then, we apply the GB algorithm [SDBR15] to obtain a tensor $\mathbf{B} \in \mathbb{R}^{224\times224\times3}$, where locations with higher norm correspond to regions predominantly used by DEX for classification. In other words, $\mathbf{B}$ pinpoints relevant regions for the age classification task. GB points out the key areas to recognize the age and should, therefore, be modified by the aging network. Importantly, GB is usually used to visualize the regions that influence one specific network output (*i.e.*, one specific class) [SDBR15]. In our case, we apply GB to the sum of the classification layer before softmax normalization to obtain class-independent masks (this decision is later ablated in Sec. IV.3.1.2). We select GB over other approaches [SCD+17, MY20, SF19] since it is a fast, simple, and strongly supported method for visualization.

We need to transform $\mathbf{B}$ to obtain a mask $\mathbf{M} \in [0,1]^{224\times224}$. We proceed in several steps. First, we average $\mathbf{B}$ over the RGB channels, take the absolute value, and apply Gaussian blur to get smoother maps. In this way, we obtain a tensor $\tilde{\mathbf{B}} \in {\mathbb{R}_{\geq0}}^{224\times224}$ that indicates relevant regions. To obtain values in $[0,1]$, we need to normalize $\tilde{\mathbf{B}}$. Our preliminary experiments showed that after normalizing by twice the variance σ of $\tilde{\mathbf{B}}$ (over the locations), relevant areas for the aging task are close to 1 or above. We apply clipping to bring down all those important regions to 1 (See Fig. 24). Formally the mask values are computed as follows:

$$\mathbf{M} = \min\left(\frac{\tilde{\mathbf{B}}}{2\times\sigma}, 1\right) \tag{IV.1}$$

where *min* denotes the element-wise minimum. Next, we detail how this mask is employed in our encoder-decoder architecture.

Skip connection blurring. Assuming a feature map $\mathbf{F}_c \in \mathbb{R}^{H'\times W'\times C}$ provided by the content encoder E_c, we resize $\mathbf{M}$ to the dimension of $\mathbf{F}_c$ obtaining a mask

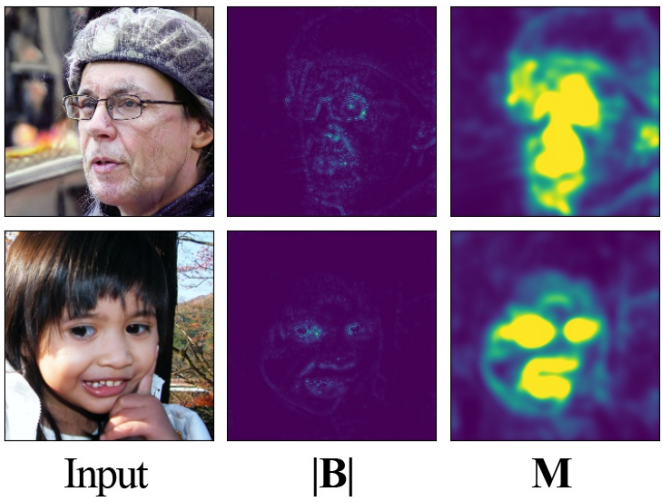


Figure 24: Example outputs of the CUSP module. From left to right: *1)* Input image; *2)* Matrix $|\mathbf{B}|$, the absolute value of the guided backpropagation output averaged over the RGB dimension; and *3)* Mask $\mathbf{M}$ predicted by the CUSP module. We see that $\mathbf{B}$ is very sparse. Therefore, we apply blur before normalization and clipping to enlarge the activated regions. In this way, we obtain the mask in the last column. We see that the high values of the masks are primarily located in the eye and mouth regions, while the background is associated with very low values. This visualization shows that our CUSP module can act only on the relevant regions in the foreground.

$\tilde{\mathbf{M}} \in [0,1]^{H' \times W'}$. We then blur $\mathbf{F}_c$ using two different Gaussian kernels with variance $\sigma_m > 0$ and $\sigma_g > 0$. The variance σ_m is applied in the region indicated by $\mathbf{M}$, while σ_g is used over the whole feature map. The motivation for this choice is that the user can choose to alter structure preservation locally, globally, or both. At training time, σ_m and σ_g are sampled randomly to force the generator G to perform well for any blur parameter. At test time, both values might be provided by the user. Formally, the blurred feature map is computed as follows:

$$\tilde{\mathbf{F}}_c = \tilde{\mathbf{M}} \circ (\mathbf{F}_c * \mathbf{k}_m) + (1 - \tilde{\mathbf{M}}) \circ (\mathbf{F}_c * \mathbf{k}_g) \tag{IV.2}$$

where $*$ denotes the convolution operation, $\circ$ is the Hadamard product, and $\mathbf{k}_m$ and $\mathbf{k}_g$ are the Gaussian kernels of variances σ_m and σ_g.

IV.2.3 Overall Training Procedure

Training facial age editing models is particularly challenging since a big enough dataset of paired images is generally unavailable. Therefore, similarly to [MHP21, OESF+20, YPN+21], our training strategy is either focused on reconstruction (when the target age matches the input age) or I2I translation (when the target age is different). Also, similar to [MHP21, OESF+20, YPN+21], training is performed using a set of complementary losses described below.

Reconstruction loss ($\mathcal{L}_r$). When the target age a_t is equal to the image age a_i, we expect to reconstruct the input image. We, therefore, adopt an L1 reconstruction

loss:

$$\mathcal{L}_r = \|T(\mathbf{X}_i, a_i) - \mathbf{X}_i\|_1 \tag{IV.3}$$

where T denotes the whole aging network, which output is the scaled addition of every *tRGB* block.

Age fidelity losses ($\mathcal{L}_D, \mathcal{L}_C$). Following [CCK+18], we use a conditional discriminator D to asses that generated images correspond to the target age a_t. More precisely, we employ the discriminator architecture of StyleGAN2 equipped with a multiclass prediction head, together with the training loss $\mathcal{L}_D$ defined in [MK18].

We employ a loss $\mathcal{L}_C$ that assesses age matching using the same pretrained classifier C used in the CUSP module to complement the adversarial loss. Furthermore, $\mathcal{L}_C$ is implemented using the Mean-Variance loss [PHSC18], a classification loss tailored for age estimation.

Cycle-Consistency loss ($\mathcal{L}_{cy}$). Following [ZPIE17], we adopt a cycle consistency $\mathcal{L}_{cy}$ to force the network to preserve details that are not specific to the age (*e.g.*, background or face identity). $\mathcal{L}_{cy}$ is given by:

$$\mathcal{L}_{cy} = \|\mathbf{X}_i - T(T(\mathbf{X}_i, a_t), a_i)\|_1 \tag{IV.4}$$

Full objective. Finally, the total cost function can be written

$$\min_M \max_D \lambda_r \mathcal{L}_r + \lambda_C \mathcal{L}_C + \lambda_D \mathcal{L}_D + \lambda_{cy} \mathcal{L}_{cy} \tag{IV.5}$$

where $\lambda_r, \lambda_C, \lambda_D$, and λ_{cy} are constant weights.

IV.2.4 Evaluation protocol

Every paper employs different metrics, datasets, and tasks in the aging literature. Therefore, we include a large set of metrics, datasets, and tasks in our experiments to allow comparison with most existing methods.

IV.2.4.1 Datasets

In this chapter, we employ three widely-used, publicly available high-resolution datasets for face aging and analysis:

- *FFHQ-RR*: Initially proposed in [YPN+21], this aging dataset based on FFHQ [KLA19] comprises of 48K images depicting people from 20 to 69 years old. Because of this *Restricted age Range*, we refer to this dataset as *FFHQ-RR*. Images are downsampled to 224×224.
- *FFHQ-LS*: This aging dataset, introduced in [OESF+20], is composed of the 70K images from FFHQ [KLA19], manually labeled in 10 age clusters that try to capture both geometric and appearance changes throughout a person's life: 0-2, 3-6, 7-9, 10-14, 15-19, 20-29, 30-39, 40-49, 50-69 and 70+ years old. Consequently, this dataset is referred to as *FFHQ-LS* because of its *LifeSpan* age range. The resolution of these images is 256×256 pixels.

- *CelebA-HQ* [KALL17, LLWT15]: It consists of 30K images at 1024×1024 resolution, which we downsample to 224×224 pixels. The only age-related label in the dataset is *young*, which can be either true or false.

The use of *FFHQ-RR* and *FFHQ-LS* may seem redundant since they are both based on the FFHQ dataset, but we perform distinct experiments on both datasets to allow comparison with existing SOTA methods (which report results on at least one of them).

IV.2.4.2 Tasks

We employ two tasks to evaluate the performance:

- *Young* → *Old*: as in [YPN+21], we sample 1000 images belonging to the "young" category and translate them to a target age of 60. This task is only performed on CelebA-HQ.

- *Age group comparison*: similarly to [MHP21], we consider different age groups: (20-29), (30-39), (40-49), and (50-69) on *FFHQ-RR* and additionally (0-2), (3-6), (7,9), (15,19) on *FFHQ-LS*. We again sample the first 1000 test images and translate every one of them into the central age of each of the four different age groups (25, 35, 45, and 55, respectively).

IV.2.4.3 Metrics

We choose metrics to evaluate the two main aspects of the aging task. Firstly, the translated/generated images must preserve the content of the input image in terms of identity, facial expression, and background. Secondly, the age translation might be accurate. In particular, we adopt the following metrics:

- *LPIPS* [ZIE+18] measures the perceptual similarity when the target age coincides with the input image age.

  This metric is designed to evaluate perceptual similarity between image patches by leveraging deep network features. It outperforms traditional metrics like PSNR [HZ10] and SSIM [WBSS04] by using learned deep features that better align with human perception. In our experimental setting, it utilizes deep features extracted from a pre-trained VGG network [SZ14]. The feature stacks from L layers are denoted as $\hat{y}^l, \hat{y}'^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ for the reference and edited patches, respectively.

  The LPIPS distance between a reference patch x and a distorted patch x' is calculated as follows:

$$d(x, x') = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\hat{y}^l_{hw} - \hat{y}'^l_{hw} \right) \right\|_2^2 \tag{IV.6}$$

  Where:

  - H_l and W_l are the height and width of the feature map at layer l.
  - $w_l \in \mathbb{R}^{C_l}$ is a vector of learned weights for each channel. This involves learning a small number of parameters on top of the fixed network features, which essentially calibrates the perceptual space of the pre-trained network on a perceptual judgment dataset.
  - And $\odot$ denotes element-wise multiplication.

  Lower LPIPS values indicate higher perceptual similarity between the two images, as perceived by human observers

- *Age Mean Absolute Error (MAE).* We employ a pretrained and independent age estimation network to compare the predicted age with the target age given an input image. As we already use the DEX pretrained classifier [RTVG15] at training time, we utilize Face++ API for this metric [2]. Experiments show that DEX is more biased towards younger age predictions than Face++. Therefore, reporting the MAE to the input target age a_t would be biased. To compensate for this DEX-Face++ misalignment, we estimate the age of the original images with Face++ and compute the mean for each group. We then report the distance between the mean group predicted age and the transformed image predicted age. The DEX-Face++ discrepancy may bias evaluation since an aging method that fails in generating images corresponding to the target age could be favored if the method is biased in the same direction as the Face++ classifier.

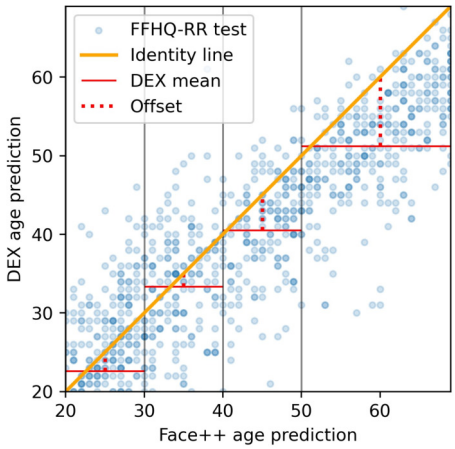


Figure 25: DEX classifier and Face++ distribution discrepancy by age group on *FFHQ-RR* test set. Color intensity denotes distribution density. The red horizontal lines represent the mean age of each age group according to DEX.

To visualize this discrepancy, we plot in Fig 25 the distribution of the DEX-Face++ predictions on the FFHQ-RR dataset. In the case of perfect agreement, all the blue points would be located on the orange identity line. We also report the mean age of each age group according to DEX (red horizontal lines). A vertical dotted line represents the amplitude of the discrepancy. In this case, the discrepancy is especially noticeable in older groups.

[2]Face++ Face detection API: `https://www.faceplusplus.com/` (last visited on July 26, 2024).

Therefore, in our evaluation protocol, we estimate the age of the original images with Face++ and compute the mean for each group. Age MAE is then computed as the distance between the mean group predicted age and the transformed image predicted age.

- *Kernel-Inception Distance (KID) [BSAG18]* assesses that the generated images are similar to real ones for similar ages. While FID [HRU+17] is adopted in [MHP21], we adopted KID as it is better suited for smaller datasets. We report the KID between original and generated images within the same age groups.

  KID is a metric designed to evaluate the quality of images generated by generative models, such as GANs. It measures the similarity between the distributions of features extracted from real and generated images using a pre-trained Inception network [SLJ+15]. KID is based on the Maximum Mean Discrepancy (MMD), which is a distance between two distributions defined in a reproducing kernel Hilbert space (RKHS) [GBR+12].

  KID values are non-negative and lower values indicate that the generated images are more similar to the real images, implying better generative performance.

- *Gender*, *Smile*, and *Face expression* preservation and *Blurriness*: Face++ provides these metrics to evaluate input image preservation and quality. *Gender*, *Smile*, and *Face expression* preservation are reported in percentages as in [YPN+21].

IV.2.4.4 Implementation details

We use the same training settings as StyleGAN2-ADA [KAH+20] with $\lambda_r = 10$, $\lambda_C = 0.06$, $\lambda_D = 1$, $\lambda_{cy} = 10$. The optimizer used is Adam with $lr = 0.0025$ and $\beta_1 = 0$, $\beta_2 = 0.99$.

FFHQ-RR and CelebA-HQ models are trained for 65 epochs with a batch size of 18. FFHQ-LS is trained for 140 epochs with a batch size of 16. All experiments are run on a single Nvidia A100 GPU.

IV.3 Experiments

IV.3.1 Ablation study

IV.3.1.1 Architecture ablation

We consider four variants of our approach where we ablate the SCs and the style encoder. In (i) variant, the style encoder is not used; an *Average Pooling layer* replaces E_s on top of the output from E_c. Variant *(ii)* employs a style encoder but no SCs, while variant *(iii)* employs SCs in every layer. Finally, variant *(iv)* follows the proposed architecture employing SCs in the second-to-last layer only. In order to make an unbiased evaluation of the architecture and not the masking operation

performed by CUSP, we report the performance of CUSP with high preservation $(\sigma_m, \sigma_g) = (0.0, 0.0)$, as variant *(ii)* applies no masking.

Table 7: Ablation study: impact of the SCs and the style encoder.

	LPIPS	**Age MAE**	**Mean KID**
(i) No style encoder	0.84	6.21	0.0163
(ii) No SC	1.70	**6.17**	0.0109
(iii) SCs at every layer	1.85	6.34	0.0175
(iv) Full	**0.78**	6.29	**0.0089**

Results shown in Table 7 suggest that a separate style encoder, as in our *Full* model (variant *(iv)*), yields better reconstruction (lower LPIPS) and similar aging performance (Age MAE and Mean KID) than using a single encoder for both content and style as in variant *(i)*. Regarding SCs, not using them leads to an important reconstruction error (see high LPIPS) since the network cannot reconstruct the image details. However, SCs in every layer also results in low reconstruction performance. We hypothesize that the model faces optimization issues. More specifically, adding SCs on every layer dramatically increases the decoder's complexity (approximately doubling its number of parameters), making the network slower and harder to train.

IV.3.1.2 CUSP module analysis

In Fig. 26 we qualitatively evaluate the impact of the kernel values used in CUSP. We compare images obtained with High, Custom, and Low structure preservation (referred to as HP, CP, and LP), where we use kernel values ranging from $\sigma = 0$ to $\sigma = 9$. We also display the mask $\mathbf{M}$ estimated by the CUSP module. We observe that when the user provides low kernel values (*i.e.*, higher preservation), the shape of the face is kept, while with higher kernel values, the network has the freedom to change its shape. The impact is clearly visible on the neck and chin of the women in the second and last row.

The visualization of the mask shows that our approach identifies those regions that change with age (chin, mouth, and forehead). We also quantitatively measure the impact of each kernel parameter. In Fig. 27, we report the Age MAE and LPIPS while changing the local and global blur parameters. By increasing the local blur, we can see that CUSP achieves a significantly lower age error while keeping a small reconstruction error. On the contrary, using global blur to improve the age performance (*i.e.*, reduce the age MAE) implies a substantial increase in the LPIPS metric, reflecting some loss of details. Overall, these experiments demonstrate the conflicting nature of aging and reconstruction performances. These observations further justify our motivation to offer the user the possibility of controlling this trade-off, thereby demonstrating the value of CUSP and its masking strategy. The ability to modify both σ_m and σ_g with different values allows us to achieve the same age-accurate transformation results while minimizing the reconstruction performance drop.

We complete this analysis with an ablation study regarding the GB-based computation of the CUSP masks. More precisely, two strategies are compared: in *Top-1*

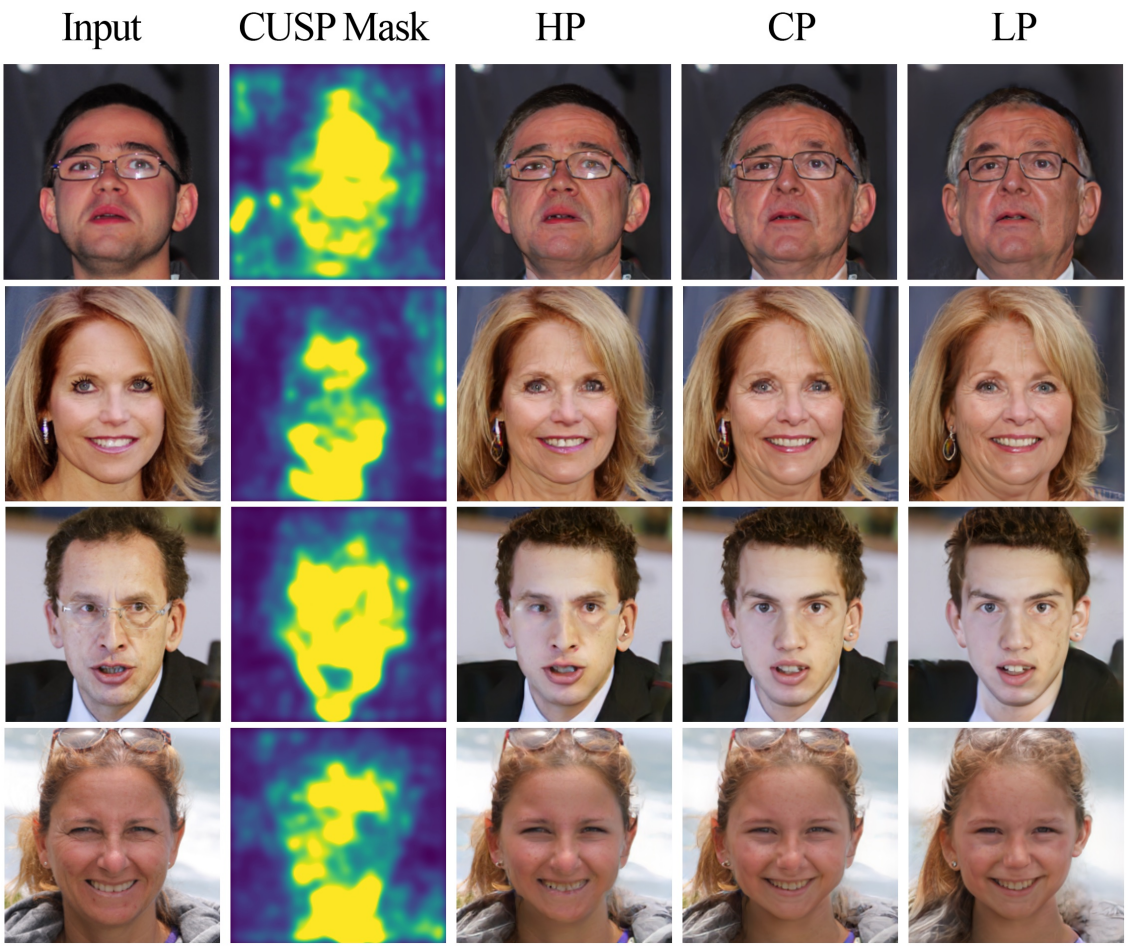


Figure 26: Impact of the kernel value: images obtained with High, Custom, and Low structure preservation (HP, CP, and LP). HP:$(\sigma_m, \sigma_g) = (0, 0)$; CP:$(\sigma_m, \sigma_g) = (9, 0)$; LP:$(\sigma_m, \sigma_g) = (9, 9)$. The second column shows the mask estimated by CUSP.

class, we apply GB on the most-activated class, while in *class-independent*, we adopt the proposed strategy of taking the sum of the classification layer before softmax. Results reported in Tab. 8 demonstrate that the class-independent strategy performs best. Indeed, using every class output from the age classifier might benefit the masking, as every age-related feature is relevant for the translation, not only those involving its current age.

IV.3.2 Comparison with State-of-the-Art

From our literature review (Sec. II.3), we identify HRFAE [YPN+21] and LATS [OESF+20] as the two main competing methods. Indeed, Re-aging GAN [MHP21] cannot be included in the comparison since neither the code nor the age classifier used for evaluation are publicly available. Since HRFAE and LATS report experiments on different datasets and follow different protocols, we perform experiments using the two tasks previously described. First, we follow HRFAE, which employs the *Young* $\rightarrow$ *Old* task on *CelebA-HQ*. In this case, the performance of FaderNet [LZU+17], PAG-GAN [YHWJ18], IPC-GAN [WTLG18], and HRFAE (on 1024×1024 resolution images) is reported in [YPN+21] and is included in our experi-

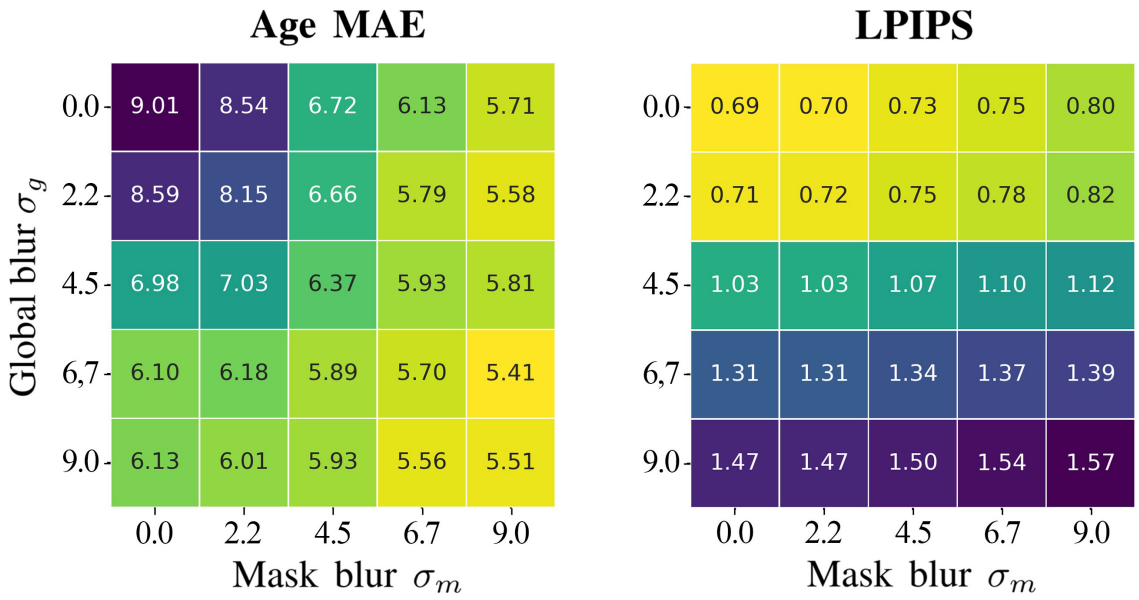


Figure 27: CUSP parameters and impact on Age MAE (left) and LPIPS×10 (right).

Table 8: Ablation study: impact of the masking strategy used in CUSP.

	LPIPS	**Age MAE**	**Mean KID**
Top-class GB	1.25	6.19	0.0145
Class-indep. (Ours)	**0.78**	**6.29**	**0.0089**

mental comparison. Second, we employ the *age group comparison* task to allow better comparison with LATS on the most challenging *FFHQ-LS* dataset. Indeed, since no automatic quantitative evaluation is reported on the *FFHQ-LS* in [OESF+20], we chose the *age group comparison* task that provides richer analysis than the *Young* → *Old* task.

IV.3.2.1 Qualitative comparison

In Fig. 28, we show a qualitative comparison with the state-of-the-art evaluated on the *celebA-HQ* dataset, where we transform the input image to the age of 60 years old. First, we observe that Fader, PAGGAN, and IPCGAN generate images with important artifacts. On the contrary, HRFAE, LATS, and our approach generate consistent images with only minor artifacts. However, only CUSP produces images that correspond to the correct target age. Other methods generate images where people look younger than expected since they are unable to make suitable structural changes. Furthermore, LATS operates only in the foreground, requiring a previous masking procedure. For this reason, in Fig. 28, the outputs related to LATS display a constant gray background. In addition, CUSP can preserve identity and non-age-related details.

We also perform a qualitative comparison with the two main competitors: HRFAE on *FFHQ-RR* in Fig. 29 and with LATS on *FFHQ-LS* in Fig. 30. We show that CUSP achieves more profound facial structure modifications (*e.g.*, thin face shapes that grow wider and wrinkled skin) and hair color transformation. The age pro-

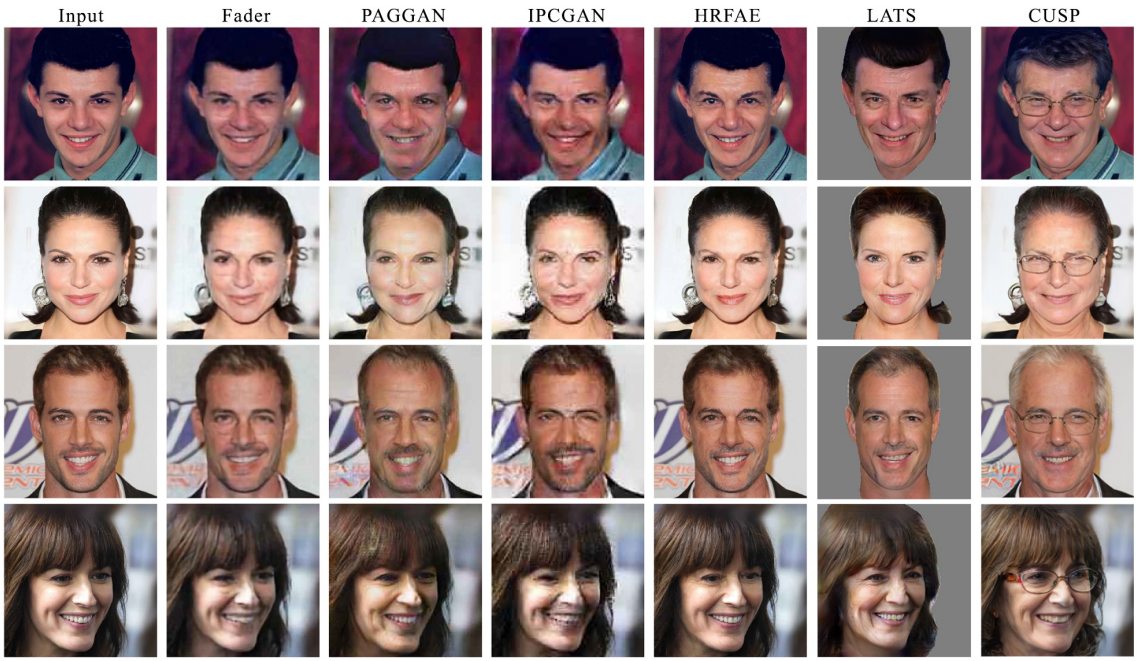


Figure 28: Comparison with SOTA methods on CelebA-HQ for the *Young* $\rightarrow$ *Old* task employing a target age of 60 years old.

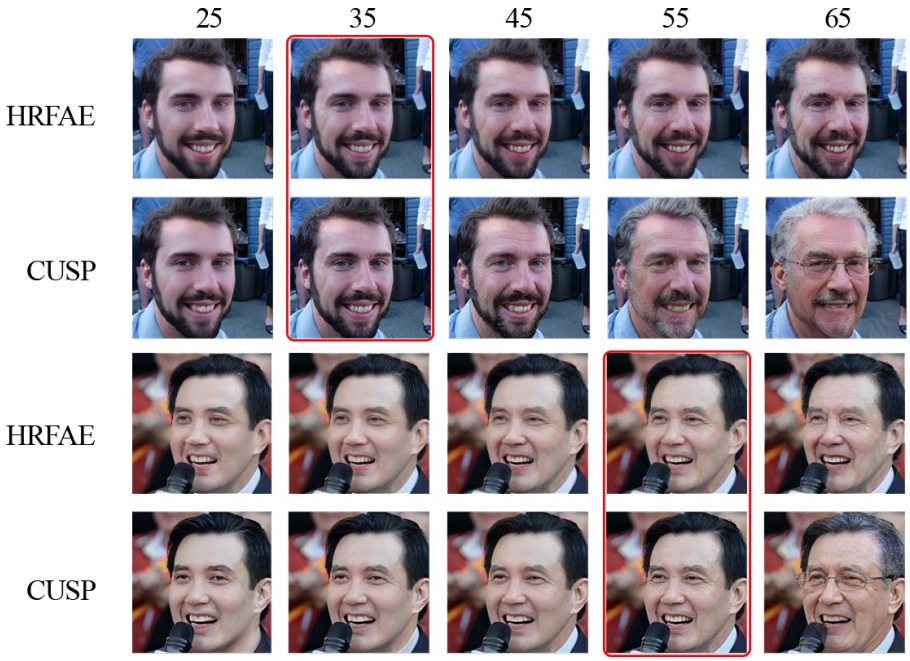


Figure 29: Qualitative comparison with HRFAE. The images corresponding to the input ages are highlighted with red frames.

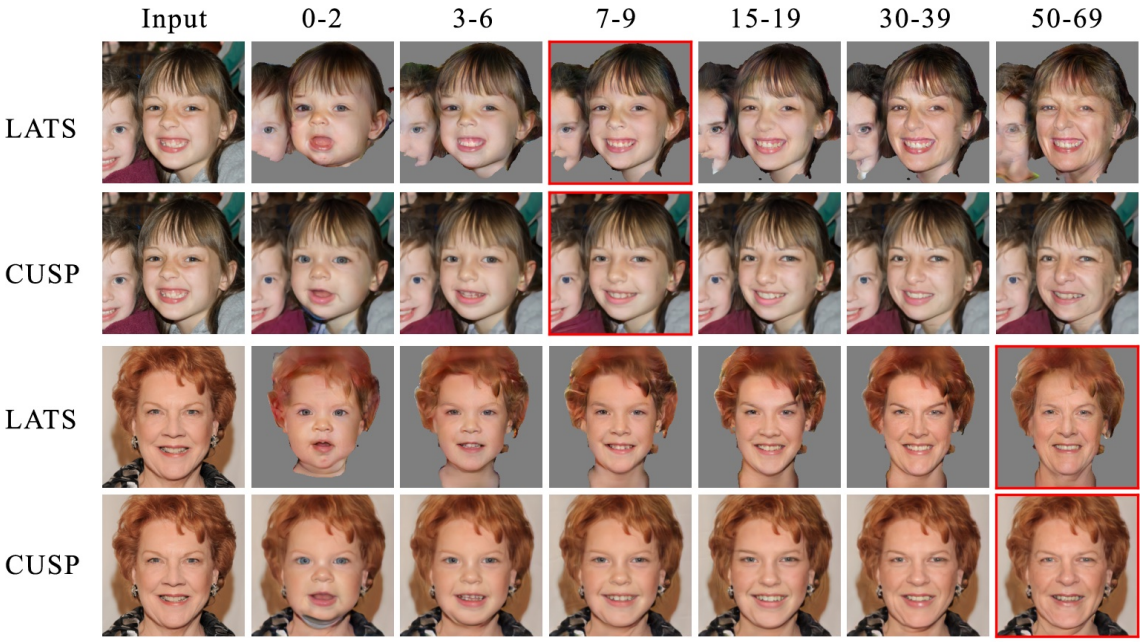


Figure 30: LATS comparison for different age targets. The images corresponding to the input ages are highlighted with red frames.

gression is smooth. Close ages produce almost identical pictures, but global age progression seems realistic and natural. Regarding LATS (Fig. 30), we see that we obtain similar performance while our method has four major advantages: (1) it operates directly on the entire image and deals with backgrounds and clothing; (2) it does not require an externally trained image segmentation network; (3) CUSP employs a single network while LATS needs to use a separately trained network for each gender; and (4) it offers user control as shown in our ablation study (see Sec. IV.3.1).

Table 9: Quantitative comparison on *CelebA-HQ* for the *Young* $\rightarrow$ *Old* task employing a target age of 60. CUSP HP (High preservation) is run with $\sigma_m = \sigma_g = 1.8$.

Method	**Predicted Age**	**Blur**	**Gender**	**Smiling**	**Neutral**	**Happy**
Real images	*68.23 ± 6.54*	2.40	-	-	-	-
FaderNet	44.34 ± 11.40	9.15	97.60	95.20	90.60	92.40
PAGGAN	49.07 ± 11.22	3.68	95.10	93.10	90.20	91.70
IPCGAN	49.72 ± 10.95	9.73	96.70	93.60	89.50	91.10
HRFAE	54.77 ± 8.40	**2.15**	97.10	**96.30**	**91.30**	**92.70**
HRFAE-224	51.87 ± 9.59	5.49	**97.30**	95.50	88.30	92.50
LATS	55.33 ± 9.33	4.77	96.55	92.70	83.77	88.64
CUSP HP	**67.76 ± 5.38**	2.53	93.20	88.70	79.80	84.60

IV.3.2.2 Quantitative comparison

In Table 9, we report a quantitative comparison evaluated on the *CelebA-HQ* dataset employing the *Young* $\rightarrow$ *Old* task. Every model has been trained on *FFHQ-RR*. Regarding HRFAE, we report the performance obtained with models trained

and tested at 224×224 and 1024×1024 resolutions (referred to as HRFAE-224 and HRFAE, respectively). We used the available code for LATS to train a model on this dataset. We also report (first row) the mean age predicted by the Face++ classifier when feeding the images of the age class 60 according to the DEX classifier used at training time. We observe an 8.23-year discrepancy. In other words, to generate images that look similar to those labeled as 60 at training time, we need to predict images that the Face++ classifier will perceive on average as 68.23 years old. These experiments confirm that CUSP outperforms other methods, being the only method that substantially modifies the image to adjust the person's target age.

In addition, CUSP ranks second in terms of Blur, quantifying the good quality of our images. For instance, the performance of HRFAE-224 worsens the predicted age with respect to its 1024×1024 counterpart and deteriorates noticeably in the Blur metric, suggesting a severe drop in the generated image quality. Interestingly, the more profound and realistic transformations yielded by CUSP and LATS imply slightly worse scores according to the preservation metrics. Indeed, preservation metrics suffer from the increased ability to make structural changes to pictures. However, this drop in quantitative fidelity is not manifested in the user study or qualitative results (Figs. 29 and 30). Two hypotheses can explain this discrepancy between qualitative and quantitative results. First, several biases can impact the results (*e.g.*, sports clothing is replaced for formal clothes at higher ages, and glasses appear in older targets as well). In addition, there may also be some expression-related biases in different age groups. Second, the CUSP module more frequently targets the image's mouth and eye areas. Those areas are the most related to facial expression detection, and their blurring might negatively affect facial expression preservation.

Table 10: Quantitative comparison with LATS on the FFHQ-LS dataset for the *age group comparison* task. CUSP CP (Custom preservation) and LP (Low preservation) are run with $(\sigma_m, \sigma_g) = (8, 4.5)$ and $(\sigma_m, \sigma_g) = (8, 8)$ respectively.

	Age MAE							**Gender Preservation (%)**						
	0-2	**3-6**	**7-9**	**15-19**	**30-39**	**50-69**	**Mean**	**0-2**	**3-6**	**7-9**	**15-19**	**30-39**	**50-69**	**Mean**
LATS	7.68	8.91	6.59	5.19	**8.23**	**5.73**	7.05	72.2	70.6	74.2	**93.7**	**93.9**	**93.9**	**83.1**
CUSP CP	6.89	**8.26**	7.67	6.70	10.67	10.86	8.51	**74.5**	69.3	**78.1**	88.3	92.1	85.9	81.4
CUSP LP	**6.49**	9.29	**5.59**	**4.99**	8.36	5.74	**6.74**	69.0	**76.0**	**78.1**	87.4	86.1	80.1	79.4

We report in Table 10 a comparison with LATS, both trained and evaluated on the *FFHQ-LS* dataset. The results support the qualitative analysis performed regarding Figs. 29 and 30. Our proposed method is on par with LATS performance concerning the aging task and achieves those results while preserving numerous image details. CUSP with low preservation even outperforms LATS in terms of Mean Age-MAE. Surprisingly, we also notice that our approach obtains similar performance in terms of gender preservation while not using gender annotations and employing a single model, while LATS uses two diffrent models for each gender.

IV.3.3 User Study

Additionally, to validate our results and provide a more comprehensive evaluation, we conducted a study on 80 different users comparing CUSP with HRFAE and

Table 11: User study on four different aspects of image aging comparing CUSP.

	Age accuracy			*Identity preservation*			*Overall quality*			*Natural progression*
	20-29	*50-69*	*Added*	*20-29*	*50-69*	*Added*	*20-29*	*50-69*	*Added*	-
CUSP	**60.2**	**72.9**	**66.6**	**50.8**	**63.7**	**57.3**	**55.8**	**67.7**	**61.8**	**60.6**
HRFAE	17.5	15.6	16.6	24.4	24.0	24.2	21.7	20.6	21.1	24.9
LATS	22.3	11.5	16.9	24.8	12.3	18.5	22.5	11.7	17.1	14.5

LATS on the young-to-old and old-to-young tasks on FFHQ-RR. Each test consisted of 48 random questions on four different topics. Similarly to [OESF+20], we asked about user preferences regarding identity preservation, target age accuracy, realism, the naturalness of the age transition, and overall preference (See Fig 31).

In this study, you will be presented with several sets of images to choose from. We will compare several AI solutions to transform a person's age in an image, similar to widely known apps like FaceApp. There are four kinds of questions, you'll have to click on your chosen image, there are no correct answers:

1. **Age accuracy:** From the images displayed, which one better depicts a person from the *target age group*? An actual person's picture (not shown) has been transformed to a target age with different mechanisms. We want to know which one you think is more accurate.
2. **Identity preservation:** From the images displayed, which one better transforms the shown original picture to the target age group while *reasonably maintaining the person's identity*? You'll have to judge which result seems more reasonable, attending to age transformation and identity preservation.
3. **Overall better:** From the images displayed, which one is *overall better* transforming the age of the person depicted in the picture? Which one do you prefer? Which image seems more pleasing?
4. **Whole age progression:** From the different shown *age progressions*, which seems *more natural and reasonable*?

In case of doubt, choose the image you subjectively prefer.

Figure 31: Description of the user study displayed to the experimental subjects.

From FFHQ-RR, 50 images were selected for each group (20-29, 30-39,40-49,50-69) and transformed to target ages 25, 35, 45, 60 with each comparing method (HRFAE, LATS, and ours), resulting in 200 original images and 2400 transformed images. Age translations were done from 20-29 and 30-39 to 50-69 (young to old) and from 40-49 and 50-69 to 20-29 (old to young).

In Question-Kind 1 (QK 1) and QK 3 (See Fig 31), three randomly ordered transformed images were presented next to a target age group. In QK 2, the original image is included. Finally, in QK 4, besides the original image, four images showing age progression (25, 35, 45, and 60) are presented for each method.

As seen in Table 11, CUSP outperforms HRFAE and LATS in every single category by a large margin (CUSP was selected globally in 62% of cases, compared to 22% and 17%, respectively). Furthermore, CUSP's results depict people of the tar-

get age with greater accuracy while maintaining the source image identity. On top of that, it outputs higher quality images, and the progression seems more natural and realistic.

Chapter V

Self-Attention Guidance for Image Editing

"Everybody is always in the middle of their own opera." — Greta Gerwig

V.1 Introduction

As discussed in Sec. II.4, text-guided image editing is an emerging CV technique that modifies images based on natural language descriptions through advanced neural networks like GANs or Diffusion Models. This approach enables users to make changes to images by simply describing them in text, such as "make the sky bluer" or "turn the car red." It goes beyond traditional DL applications, offering a versatile framework for various image manipulations guided by textual inputs. Despite its potential, challenges such as the computational intensity of the inversion process in diffusion models remain.

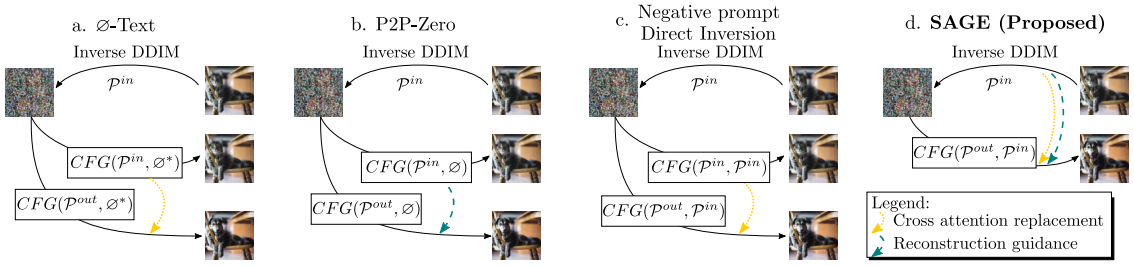


Figure 32: Comparative Analysis of Diffusion-Based Image Editing Techniques. Our review contrasts existing methodologies, which utilize CFG [HS21] with various combinations, including the pretrained null-prompt $\varnothing$, an optimized latent representation $\varnothing^*$, the descriptive prompt of the input image $\mathcal{P}^{in}$, and the target editing prompt $\mathcal{P}^{out}$.

In Fig. 32 we showcase a comparative analysis of existing diffusion-based image editing techniques. The four approaches to prompt-based image editing —∅-Text

Inversion (NT), P2P-Zero (P2P-Zero), Negative Prompt Inversion (NPI), and Direct Inversion (DI)— each employ distinct methodologies to achieve image modifications, leveraging different aspects of attention mechanisms and optimization strategies.

In the ∅-Text Inversion (NT) method (illustrated in Fig 32a), Mokady *et al.* [MHA+23] optimize the null prompt embedding for reconstructing the original image. Editing is then performed using a CA map mechanism as previously described in [HMT+23]. In P2P-Zero [PKSZ+23] (Fig 32b), the approach avoids the computationally demanding inversion process by incorporating a guidance term at each diffusion step, guiding the model towards precise reconstruction. On the other hand, Negative Prompt Inversion (NPI) [MIST23] (Fig 32c) replaces the traditional null prompt in CFG with a negative prompt, utilizing CA manipulation for editing, similar to [HMT+23, MHA+23]. Direct Inversion (DI) [JZB+23] (also shown in Fig 32c) introduces a direct inversion technique that guides the reconstruction process as well as establishes an editing benchmark. All these methods require additional steps to explicitly reconstruct the input image, which is computationally intensive. These methods differ in their inversion techniques and CFG use.

To tackle this limitation, we present Self-Attention Guidance for Editing (SAGE), a novel technique that balances computational efficiency with high-fidelity reconstruction, while also enabling versatile image editing capabilities. Similar to other methods [MHA+23, MIST23, PKSZ+23], our approach uses DDIM [SME21] inversion. Nevertheless, our unique contribution lies in the utilization of the intermediate SA and CA maps that the diffusion model internally computes during the reverse DDIM process, which allows for precise details reconstruction with minimal computational effort. During the sampling phase we implement a combined use of CFG (refer to Sec. V.2.2) and a novel SA reconstruction guidance mechanism. This mechanism stores and exploits the SA maps within the diffusion U-Net [RFB15], providing an optimal balance between editing and maintaining the original image details.

Our research (illustrated in Fig 32d) differs from the compared methods in two significant ways: (i) We enable effective editing without the need for explicit reconstruction of the input image. (ii) We use intermediate latent vectors calculated during the inverse DDIM stage, which allows for editing while preserving content in regions not affected by the edits. This approach leads to a method that is simpler and more computationally efficient.

In summary, our contributions are threefold:

- We propose an innovative editing framework that employs a pre-trained diffusion model, utilizing intermediate noise vectors from the inverse DDIM process. This framework allows for image editing that aligns to the input image while achieving modifications based on textual prompts.

- We introduce a novel reconstruction guidance loss term that functions within the SA layers of the diffusion network. This ensures high-fidelity reconstruction in regions not impacted by the editing process, without significantly increasing computational demands.

- Through rigorous experimental validation, we compare our method with recent

approaches in the field, showing that SAGE delivers comparable or superior editing quality with minimal computational expense.

Figure 15: Prompt-based image editing: the user can add, omit, change, or enhance elements in an image by providing a descriptive prompt of the original image and indicating the words that must be removed (in **red**) or added (in **blue**). *Repeated from page* 46.

V.2 Methods

V.2.1 Problem Formulation and Preliminaries

In this study, we tackle the challenge of prompt-based image editing as introduced in [HMT+23, MHA+23] (as illustrated next in Fig. 15): the user supplies an initial image $\boldsymbol{i}$ along with an input textual description $\mathcal{P}^{in}$ of that image. Additionally, the user provides a target prompt $\mathcal{P}^{out}$ that describes the desired outcome after editing. To accomplish this task, we introduce SAGE, a Self-Attention Guidance-based method for image Editing, illustrated in Fig. 33.

Our method assumes the use of a pre-trained text-to-image diffusion model [HJA20], specifically a latent diffusion model operating within the latent space of a pre-trained autoencoder [RBL+22]. Diffusion models are generative techniques that use a neural network to predict noise, $\varepsilon_\theta^t(\boldsymbol{z}_t, \mathcal{P})$, which aims to denoise data points incrementally at different time steps $t \in [0, T]$. In this context, $\boldsymbol{z}_t$ denotes the noise-perturbed version of the original sample $\boldsymbol{z}_0$, formulated as $\boldsymbol{z}_t = \sqrt{\alpha_t}\boldsymbol{z}_0 + \sqrt{1-\alpha_t}\varepsilon$, where ε represents Gaussian noise. The variable α_t controls the noise level, ranging from near 1 (minimal noise) to almost 0 (complete Gaussian noise) over time from 1 to T. The variable $\mathcal{P}$ is an optional conditioning parameter, in this case, a textual prompt. The neural network $\varepsilon_\theta^t(\boldsymbol{z}_t, \mathcal{P})$ is implemented using a U-Net architecture that incorporates both SA and CA layers [VSP+17] to process the conditioning data.

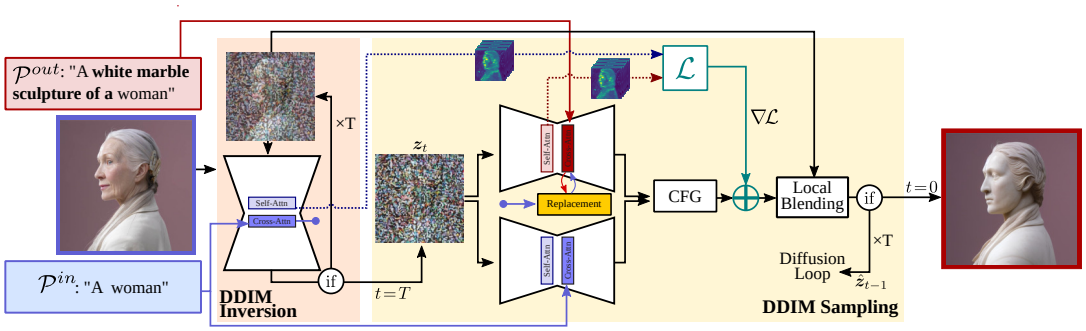


Figure 33: Pipeline of SAGE: Initially, DDIM inversion is performed on the input image using its associated prompt $\mathcal{P}^{in}$. This inversion yields the estimated noise $\boldsymbol{z}_T$, which serves as the starting point for the DDIM sampling process responsible for creating the edited image. Within this framework, the U-Net processes the editing and initial prompts ($\mathcal{P}^{out}$ and $\mathcal{P}^{in}$, respectively) separately, utilizing CFG. To compute a guidance term, a comparison is made between the SA from the DDIM inversion and the SA derived in the U-Net when it receives the target prompt $\mathcal{P}^{out}$. The latent representation obtained from the DDIM inversion is then integrated with the latent representation produced after applying the guidance through a process called local blending. For the sake of simplicity, the mask computation step is not illustrated here.

Following prior studies [PCWS22, MHA+23, MIST23], we employ DDIM for faster sampling [SME21]. We utilize a pre-trained encoder $Enc(\cdot)$ to map the input image $\boldsymbol{i}$ into the latent space, resulting in $\boldsymbol{z}_0^{in} = Enc(\boldsymbol{i})$. Deterministic DDIM inversion [SME21] is then applied to reverse the diffusion process. With the input prompt $\mathcal{P}^{in}$, the DDIM inversion yields a sequence of noisy latent variables $\boldsymbol{z}_t^{in}$, with t increasing from 0 to T:

$$\begin{aligned}\boldsymbol{z}_{t+1}^{in} =& \frac{\sqrt{\alpha_{t+1}}}{\sqrt{\alpha_t}}(\boldsymbol{z}_t^{in} - \sqrt{1-\alpha_t}\varepsilon_\theta^t(\boldsymbol{z}_t^{in}, \mathcal{P})) \\ &+ \sqrt{1-\alpha_{t+1}}\varepsilon_\theta^t(\boldsymbol{z}_t^{in}, \mathcal{P}^{in})\end{aligned} \tag{V.1}$$

V.2.2 Reconstruction with Guidance

The estimated $\boldsymbol{z}_T^{in}$ serves as the starting point for generating the edited image using the DDIM sampling procedure. To address the dual objectives of image editing —modifying the input image according to the prompt $\mathcal{P}^{out}$ (the *editing* goal) and preserving unchanged regions (the *reconstruction* goal)— we introduce two distinct guidance mechanisms.

V.2.2.1 Classifier-Free guidance and Negative-prompt

In its original form, CFG [HS21] facilitates conditional generation by combining unconditional and conditional noise estimates through a linear interpolation. In this work, we utilize the CFG variant proposed in [MIST23], which incorporates negative prompts for improved image editing performance:

$$\begin{aligned}\tilde{\varepsilon}_\theta^t(\boldsymbol{z}_t, \mathcal{P}^{in}, \mathcal{P}^{out}) &= \varepsilon_\theta^t(\boldsymbol{z}_t, \mathcal{P}^{in}) \\ &+ w \cdot \left(\varepsilon_\theta^t(\boldsymbol{z}_t, \mathcal{P}^{out}) - \varepsilon_\theta^t(\boldsymbol{z}_t, \mathcal{P}^{in})\right)\end{aligned} \tag{V.2}$$

where $w > 0$ adjusts the strength of the conditional denoising process. The noise estimated by CFG is then integrated into the DDIM sampling equation to estimate $\boldsymbol{z}_{t-1}$:

$$\begin{aligned}\boldsymbol{z}_{t-1} = &\frac{\sqrt{\alpha_{t-1}}}{\sqrt{\alpha_t}} \left(\boldsymbol{z}_t - \sqrt{1-\alpha_t}\tilde{\varepsilon}_\theta^t(\boldsymbol{z}_t, \mathcal{P}^{in}, \mathcal{P}^{out})\right) \\ &+ \sqrt{1-\alpha_{t-1}}\tilde{\varepsilon}_\theta^t(\boldsymbol{z}_t, \mathcal{P}^{in}, \mathcal{P}^{out})\end{aligned} \tag{V.3}$$

In our method, the parameter w in Eq. (V.2) is crucial. It guides the transformation from the input prompt $\mathcal{P}^{in}$ to the target prompt $\mathcal{P}^{out}$, thus controlling the editing strength during generation. When using CFG starting from $\boldsymbol{z}_T^{in}$, there is a delicate balance between achieving accurate reconstruction and fulfilling the editing goals. A higher value of w enhances adherence to $\mathcal{P}^{out}$, but at the cost of reducing reconstruction quality, as it increasingly emphasizes the editing prompt over the original image structure.

Figure 34: Estimation of $\hat{\boldsymbol{z}}_0$ for positive and negative prompt over different timesteps. In this figure, we display the estimated $\hat{\boldsymbol{z}}_0$ for positive prompt $\mathcal{P}^{out}$, negative prompt $\mathcal{P}^{in}$, and CFG (incorporating both $\mathcal{P}^{out}$ and $\mathcal{P}^{in}$) during the DDIM sampling process with CFG.

The trade-off between editing and reconstruction depending on w is illustrated in Fig. 34. Initially, the noise predictions for both $\mathcal{P}^{in}$ and $\mathcal{P}^{out}$ guide $\hat{\boldsymbol{z}}_t$ towards the original image. However, as shown by the steps within the **red** square in Fig 34, a divergence occurs for middle inference steps ts; $\varepsilon_\theta^t(\boldsymbol{z}_t, \mathcal{P}^{in})$ leans towards reconstruction due to inherent ambiguity, while $\varepsilon_\theta^t(\boldsymbol{z}_t, \mathcal{P}^{out})$ leans towards editing. As $\hat{\boldsymbol{z}}_t$ advances closer to $t = 1$, moving towards the final edited image, both noise estimations begin to align (final columns of Fig 34). Setting w too low results in a generation biased towards reconstructing the original image, while setting it too high causes the generation to disregard the input image, resulting in an output heavily influenced by $\mathcal{P}^{out}$.

V.2.2.2 Reconstruction Guidance with Self-Attention

Starting the DDIM sampling process with $\boldsymbol{z}_T^{in}$ and $\mathcal{P}^{in}$ generates outputs that are similar to the input image. However, relying solely on $\boldsymbol{z}_T^{in}$ does not suffice for accurate editing or even precise reconstruction [HMT+23, MIST23]. While other methods often employ specialized techniques such as optimization [MHA+23, MIST23] or guidance [PKSZ+23] to reconstruct the input image explicitly, our approach enables editing without requiring the reconstruction of the input image.

To ensure that the unaffected parts of the edited image are reconstructed accurately, we constrain the latent $\boldsymbol{z}_t$ during the editing process to closely match the $\boldsymbol{z}_t^{in}$ from the DDIM inversion. Instead of directly comparing $\boldsymbol{z}_t^{in}$ with $\boldsymbol{z}_t$ (*i.e.*, the noisy latent image projection), which can lead to diffusion instability (see Sec V.3.1), we propose using SA maps within the U-Net architecture ε_θ^t to assess the similarity between $\boldsymbol{z}_t^{in}$ and $\boldsymbol{z}_t$. While CA maps have been utilized previously for guidance in reconstruction [HMT+23, HWC+24] as they are easier to deal with and manipulate, we favor SA due to its richer encoding of semantics in the image. In SA, each image token (or image latent coordinate) is evaluated in the context of all other tokens, which offers a more thorough comprehension compared to CA, which only connects word and image tokens. This capability of SA allows for the reconstruction of image details not specified in the caption, whereas CA primarily focuses on the regions mentioned in $\mathcal{P}^{in}$. Our experimental results further support this preference (see Sec V.3.1), demonstrating that SA maps provide superior performance in both reconstruction and editing.

Specifically, during the inverse DDIM process, we capture SA maps $S_{i,t}^{in}$ at each time step t for each transformer block i, at a resolution tailored to the input image. For example, with an input image resolution of 512×512, we use 32×32 maps. Simultaneously, during the image synthesis phase, we gather SA maps $S_{i,t}^{out}$, which are generated during the estimation of $\varepsilon_\theta^t(\boldsymbol{z}_t, \mathcal{P}^{out})$. We then compute a loss $\mathcal{L}_t^{\text{self}}$ at each time step t to guide the editing process, defined as follows for all the N transformer blocks in the U-Net used for guidance:

$$\mathcal{L}_t^{\text{self}} = \sum_i^N \left\| S_{t,i}^{in} - S_{t,i}^{out} \right\|_1 \tag{V.4}$$

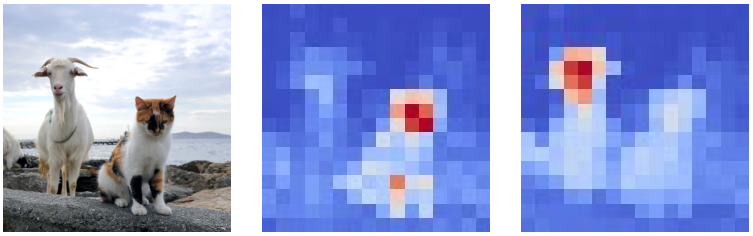
As in the classifier guidance approach [DN21], $\mathcal{L}_t^{\text{self}}$ loss is subsequently employed to guide both the editing and reconstruction by adjusting the noise estimate. Specifically, we incorporate the gradient of $\mathcal{L}_t^{\text{self}}$ with respect to $\boldsymbol{z}_t$, scaled by λ, as an extra term. The goal is to steer the diffusion process toward effective denoising while also minimizing the loss. Adding this gradient component to the CFG noise estimate from Eq (V.2) results in our final noise estimate:

$$\hat{\boldsymbol{z}}_{t-1} = \boldsymbol{z}_{t-1} - \lambda \nabla_{\boldsymbol{z}_t} \mathcal{L}_t^{\text{self}} \tag{V.5}$$

Inspired by the approach in [CCC+23], we gradually reduce λ at each time step t. This way, we acknowledge the different purposes of the diffusion steps: the initial steps mainly focus on guiding the editing according to $\mathcal{P}^{out}$, whereas the later steps concentrate on refining the image through denoising, so the guidance is less relevant. This progressive difference in roles is illustrated in Fig. 34.

V.2.3 Cross-Attention Manipulation

The CA maps within the U-Net connect the image tokens or latent space coordinates with prompt tokens. These maps have proven valuable for structurally guiding the editing process in image editing tasks like ours [HMT+23, MHA+23, PKSZ+23, HWC+24, JZB+23, MIST23] due to their ability to link image locations with each word in the input prompt. This capability is demonstrated in Fig. 35. To further enhance our image editing framework, we incorporate three mechanisms inspired by the contribution of [HMT+23] that was described in Sec. II.1.7 , leveraging the CA layers in the diffusion U-Net.



Input *"cat"* *"goat"*

Figure 35: 16×16 averaged CA maps for "*cat*" and "*goat*" given the input phrase "*A cat and a goat*".

V.2.3.1 Local Blending

While the SA reconstruction guidance succesfully reconstructs the general structure of the original image, it often falls short in accurately preserving colors and fine details. This shortfall is likely due to the low resolution of the SA maps and the loss of texture information across the multiple layers of the U-Net. We denote $C^{in}_{t,i}$ and $C^{out}_{t,i}$ for each CA layer i in $\varepsilon^t_\theta(\boldsymbol{z}^{in}_t, \mathcal{P}^{in})$ and $\varepsilon^t_\theta(\boldsymbol{z}^{in}_t, \mathcal{P}^{out})$ at each resolution. The CA map resolution used for local blending is experimentally chosen based on the input image resolution (*e.g.*16×16 for a 512×512 image). By averaging the CA maps $C^{in}_{t,i}$ and $C^{out}_{t,i}$, which correspond only to a subset of tokens from $\mathcal{P}^{in}$ and $\mathcal{P}^{out}$ defined by the user, over t and i into a single mask, normalizing, thresholding, and upscaling to the image size, we create a binary mask M as described in [HMT+23] used to blend the edited latent $\boldsymbol{z}_t$ with the original $\boldsymbol{z}^{in}_t$ at each step t:

$$\hat{\boldsymbol{z}}_{t-1} = M \odot \boldsymbol{z}_{t-1} + (1 - M) \odot \boldsymbol{z}^{in}_{t-1} \tag{V.6}$$

Our findings show that this approach is more computationally efficient than the method in [HMT+23], as it takes advantage of both $\mathcal{P}^{in}$ and $\mathcal{P}^{out}$ branches of a single CFG noise estimation, instead of requiring two separate noise estimations and a explicit input recontruction.

V.2.3.2 Cross-Attention Replacement

When the edit prompt $\mathcal{P}^{out}$ involves replacing a word in $\mathcal{P}^{in}$, it is often desirable to preserve the shape of the modified object (see Fig 36). Following the methodology of [HMT+23], we achieve this by swapping the CA maps $C^{out}_{t,i}[k]$ corresponding

to the new word token k with the original maps $C^{in}_{t,i}[k]$. This substitution is performed exclusively during the initial diffusion steps as in [HMT+23] (specifically, the first 20%). Although extending this substitution to more steps could improve the retention of the edited object's shape, it risks degrading the overall image quality.

V.2.3.3 Cross-Attention reweighting

Adopting the technique of [HMT+23], users have the option to adjust the importance of a word in $\mathcal{P}^{in}$ by modifying the weights of the associated CA maps $C^{out}_{t,i}[k]$, where k is the index of the target word. This allows users to fine-tune the impact of specific tokens on the final image, either increasing or decreasing their influence as desired.

V.2.4 Evaluation

V.2.4.1 Data

The evaluation is based on PieBench [JZB+23], a dataset comprising 700 images equally divided between natural and artificial scenes across four categories: animal, human, indoor, and outdoor. These images are sorted into ten distinct tasks: object modification, object addition, object deletion, content alteration, pose adjustment, color modification, material change, background alteration, style change, and a random task chosen from the aforementioned categories. Each image includes a source prompt, a target prompt, the subjects for editing, and a manually annotated mask to assess background preservation (only for certain tasks). Additionally, we use a collection of high-resolution images from Pexels[1] to conduct qualitative evaluations.

V.2.4.2 Metrics

Adopting the comprehensive evaluation strategy outlined in [JZB+23], we assess and compare the performance of SAGE. This analysis is divided into three main benchmark categories:

1. *Structure distance.* As described in [TBTBD22], the structure distance between two images is based on their deep feature representations extracted from a pre-trained Vision Transformer model [DBK+21], specifically DINO-ViT [CTM+21].

   An image I is processed by the Vision Transformer, which divides it into patches, linearly embeds each patch into a d-dimensional vector. The set of patches are passed through multiple Transformer layers, each consisting of normalization, multi-head self-attention (MSA) modules, and MLP blocks.

   In each MSA block, tokens are linearly projected into queries Q, keys K, and values V:

$$Q^l = T^{l-1}W^l_Q, \quad K^l = T^{l-1}W^l_K, \quad V^l = T^{l-1}W^l_V \tag{V.7}$$

[1] Images from `https://www.pexels.com/` are free for commercial use.

The self-similarity matrix $S_L(I)$ at the deepest layer L is defined using the cosine similarity between the keys:

$$S_L(I)_{ij} = \text{cos-sim}(K_i^L(I), K_j^L(I)) \tag{V.8}$$

The Structure Distance $L_{structure}$ between two images I_s and an output image I_o is given by the Frobenius norm of the difference between their self-similarity matrices:

$$L_{structure} = \|S_L(I_s) - S_L(I_o)\|_F \tag{V.9}$$

This distance quantifies the difference in spatial structure between two images by comparing the self-similarity of their deep features extracted from a pre-trained Vision Transformer. This spatial structure should be independent to image changes if the edited attributes are spatially distributed in the same way as the original attributes. Lower values indicate higher similarity, meaning the structures of the images are more alike

2. *Background preservation*: We evaluate the masked area using two metrics: LPIPS [ZIE+18] (previously described in Sec. IV.2.4.3) and SSIM [WBSS04].

   The Structural Similarity Index (SSIM) is a method for measuring the similarity between two images. Unlike the former two, it doesn't use the latent space of a pre-trained network, instead it considers changes in structural information, luminance, and contrast. The SSIM index is defined as follows:

$$\text{SSIM}(I_s, I_o) = \frac{(2\mu_{I_s}\mu_{I_o} + C_1)(2\sigma_{I_sI_o} + C_2)}{(\mu_{I_s}^2 + \mu_{I_o}^2 + C_1)(\sigma_{I_s}^2 + \sigma_{I_o}^2 + C_2)}, \tag{V.10}$$

   where:

   - I_s and I_o are the source and output images being compared.
   - μ_{I_s} and μ_{I_o} are the mean intensities of I_s and I_o, respectively.
   - $\sigma_{I_s}^2$ and $\sigma_{I_o}^2$ are the variances of I_s and I_o, respectively.
   - $\sigma_{I_sI_o}$ is the covariance of I_s and I_o.
   - $C_1 = (K_1L)^2$ and $C_2 = (K_2L)^2$ are constants to stabilize the division with weak denominator, where L is the dynamic range of the pixel values (typically 255 for 8-bit grayscale images), $K_1 = 0.01$, and $K_2 = 0.03$.

   The SSIM index can be separated into three components: luminance, contrast, and structure comparison.

   - The luminance comparison function is defined as:

$$l(I_s, I_o) = \frac{2\mu_{I_s}\mu_{I_o} + C_1}{\mu_{I_s}^2 + \mu_{I_o}^2 + C_1} \tag{V.11}$$

   - The contrast comparison function is defined as:

$$c(I_s, I_o) = \frac{2\sigma_{I_s}\sigma_{I_o} + C_2}{\sigma_{I_s}^2 + \sigma_{I_o}^2 + C_2} \tag{V.12}$$

- The structure comparison function is defined as:

$$s(I_s, I_o) = \frac{\sigma_{I_s I_o} + C_2/2}{\sigma_{I_s}\sigma_{I_o} + C_2/2} \tag{V.13}$$

Combining these components, we get the overall SSIM index:

$$\text{SSIM}(I_s, I_o) = l(I_s, I_o) \cdot c(I_s, I_o) \cdot s(I_s, I_o) \tag{V.14}$$

Higher SSIM values indicate higher similarity, with a maximum value of 1 indicating perfect similarity

3. *Target-prompt fidelity*: We measure using CLIP Similarity (SIM) [WHZ+21], applied both to the entire image and specifically to the edited sections. Given an image I and a textual description T, SIM is computed using the CLIP model's image and text embeddings [RKH+21]. The similarity score is defined as:

$$\text{SIM}(I, T) = \frac{\mathbf{f}_{\text{img}}(I) \cdot \mathbf{f}_{\text{text}}(T)}{\|\mathbf{f}_{\text{img}}(I)\|\|\mathbf{f}_{\text{text}}(T)\|} \tag{V.15}$$

where:

- $\mathbf{f}_{\text{img}}(I)$ is the image embedding vector of the image I.
- $\mathbf{f}_{\text{text}}(T)$ is the text embedding vector of the textual description T.
- $\cdot$ denotes the dot product between the image and text embeddings.
- $\|\cdot\|$ represents the Euclidean norm (or L2 norm).

The result is a similarity score between -1 and 1, where a higher score indicates greater similarity between the image and the textual description.

V.3 Experiments

Each experiment with SAGE was conducted on a single NVIDIA A100-40GB from a DGX A100 server. The images for the other methods used in the quantitative comparison (Sec V.3.2) and the user study (Sec V.3.3) were obtained from the PieBench experimental results [JZB+23]. The images for the qualitative comparison (Table 38 and Sec V.3.2) corresponding to NT, NPI, and ProxNPI were taken from [HWC+24], while the others were generated using the code available [JZB+23], adjusting the hyperparameters to achieve the best result.

As highlighted by [MHA+23, HWC+24], both the diffusion models and the methods under comparison [HS21] exhibit sensitivity to hyperparameter settings. To ensure a fair comparison, the SAGE results reported in Tables 12, 13, and 15 were generated using fixed hyperparameters: 50 DDIM steps, a CFG scale of 7.5, local blending during the initial 40 steps, CA replacement within the first 5 steps, a SA guidance scale of 200, and without CA reweighting.

Each method employs 512×512 images and uses Stable Diffusion 1.4 as diffusion model, except for Plug-n-Play, which utilizes Stable Diffusion 1.5. For the images presented in Fig 15, 768×768 images were generated using Stable Diffusion 2.1 as the core of SAGE, no further comparison with SOTA using Stable Diffusion 2.1 was performed as none of the discussed methods report results on this diffusion architecture [2].

Our initial analysis indicated that the optimal results for 512×512 images were obtained using 32×32 SA maps and 16×16 CA maps from the second and third encoder blocks of the U-Net respectively, as well as the corresponding upsampling block, in a manner similar to [HMT+23]. For 768×768 images, the optimal results were achieved using 24×24 SA and CA maps from the third block. Additionally, our implementation supports FP16 computation, including gradient calculations, which substantially reduces both time and memory requirements. To avoid zero gradients in half-precision floating-point computations, the loss term is scaled by 500 before calculating the gradients and the value λ is applied afterward (see Sec. V.2.2).

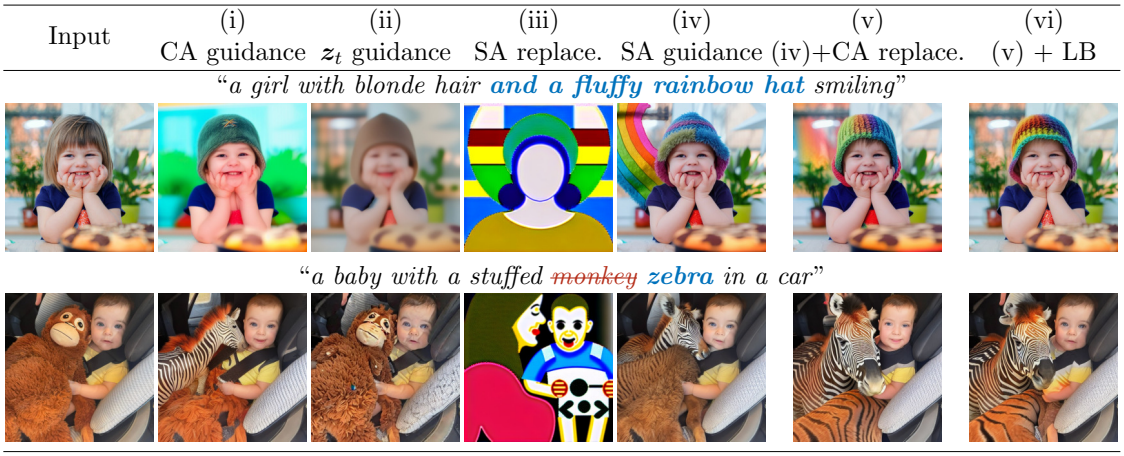


Figure 36: Table 12 ablation study.

V.3.1 Ablation Study

V.3.1.1 Reconstruction Mechanism

In this ablation study, we begin by evaluating the mechanisms used to achieve reconstruction in regions that are meant to be preserved through the editing process. We consider four baseline approaches, utilizing either guidance or replacement, applied to CA or SA layers. These baselines include: (i) employing CA map reconstruction guidance as outlined in [PKSZ+23], and (ii) computing the guidance term within the noisy latent space of the diffusion model, denoted as z_t guidance. For the SA layers, we investigate: (iii) SA replacement as done in [HMT+23, MHA+23] for CA maps, and (iv) SA guidance. Our proposed method enhances (iv) by sequentially incorporating: (v) CA replacement, and (vi) Local blending, elaborated in Sec. V.2.3.

[2] The different Stable Diffusion models can be found in `https://stability.ai/stable-image` and `https://huggingface.co/models?other=stable-diffusion`. Last accesed July 26, 2024.

Reconstruction	CA replace.	LB	Struct.↓	LPIPS ↓	CLIP ↑
(i) CA guidance	-	-	15.7	58	21.9
(ii) z_t guidance	-	-	40.0	111.3	21.5
(iii) SA replace	-	-	178.4	376.0	18.8
(iv) SA guidance	-	-	15.7	42.0	**22.0**
(v) SA guidance	✓	-	14.7	49.5	21.9
(vi) SA guidance	✓	✓	**11.0**	**39.6**	**22.0**

Table 12: Quantitative analysis was carried out using PieBench. Our evaluation includes four different strategies (i-iv) for directing reconstruction, utilizing either guidance or replacement, applied to both CA and SA layers. Additionally, we assess (v) the replacement of CA and (vi) Local Blending (denoted as LB) alongside the most effective reconstruction method.

The quantitative results are presented in Table 12, where we measure Structure distance, LPIPS, and CLIP Similarity. The latter is computed only according to the provided segmentation maps, corresponding to the areas that should be edited. In addition to these quantitative metrics, we provide qualitative examples in Fig 36, illustrating the results achieved with the same baselines compared in Table 12.

Among the different reconstruction techniques evaluated, guidance-based methods (i, ii, and iv) consistently outperform the replacement approach (iii) across all metrics. This is also visually evident in Fig. 36, where the images produced using SA replacement appear notably unrealistic and differ significantly from the originals. Although CA guidance (i) produces satisfactory results, it is outperformed by our proposed SA guidance method by a margin of 16 points in the LPIPS metric. This qualitative discrepancy is clearly visible in the first row of Fig 36. Moreover, $\boldsymbol{z}_t$ guidance (ii) is inferior both quantitatively and qualitatively compared to the other guidance methods.

Regarding CA replacement (v), our findings indicate a slight increase in LPIPS, which is offset by better structure preservation metrics. This qualitative improvement is particularly apparent in the first row of Fig. 36, where the background is better preserved. In the case of (vi), Local Blending (LB) enhances both the structure metric and LPIPS, without sacrificing the CLIP metric. This demonstrates the effectiveness of our LB mechanism in maintaining editability while preserving structural integrity, as it only impacts areas unrelated to the editing.

V.3.1.2 Classifier-Free Guidance and Self-attention Guidance

In Fig. 37, we evaluate how CFG (w) and SA guidance (λ) influence the generation process. It is evident that increasing λ enhances attention to the input structure, while higher w values improve the attention to $\mathcal{P}^{out}$ on the generated image but lead to more saturated colors, as previously noted in [HS21]. Greater SA guidance (λ) not only maintains better structure preservation but also improves color preservation, counteracting the adverse effects of high CFG on natural-looking colors. Striking an optimal balance is crucial to produce images that retain the input structure while performing profound, natural transformations reflective of $\mathcal{P}^{out}$.

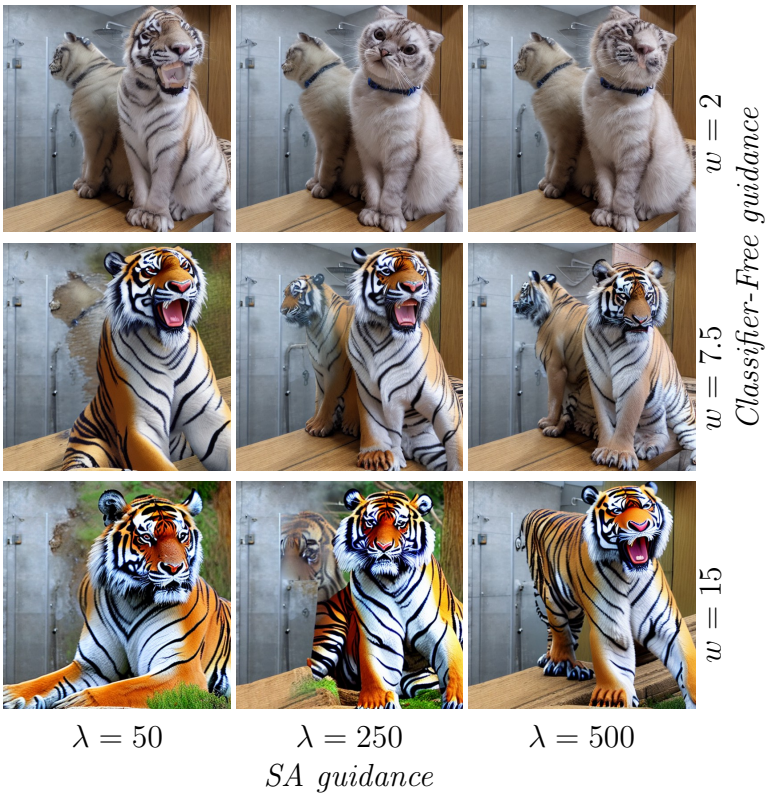


Figure 37: This matrix demonstrates the interaction between the SA guidance scale λ and CFG w, underlining their impact on image creation. It reveals how different values of λ and w influence the image to lean towards reconstruction (when λ is high and w is low) or editing (when λ is low and w is high). The images are generated using the prompt "'*a ~~cat~~ **tiger** sitting next to a mirror*".

V.3.2 Comparison with State-of-the-Art

V.3.2.1 Quantitative Comparison

Table 13 presents our quantitative comparison against existing methods. Initially, it is clear that P2P-Zero consistently underperforms compared to other methods, showing the lowest results across all metrics. Among the remaining methods, those achieving the highest CLIP similarity, such as Plug-n-Play, tend to perform poorly in terms of structure and background preservation. This trade-off is due to the inherent challenge of effectively editing images for a specific task while maintaining the original content and structure not related to the task as argued in Sec. IV.3.1.2 of the previous chapter. Conversely, methods like Proximal NPI display the opposite trend. It remains uncertain whether the superior Edited CLIP performance of Plug-n-Play [TGBD23] is due to its use of StableDiffusion 1.5 or the less restricted editing, which compromise structure preservation. Our approach excels in background preservation and achieves the second-best structure distance while still maintaining competitive CLIP similarity (ranking first in Whole CLIP similarity and third in the masked measure). Overall, these results highlight the effectiveness of SAGE, showing that SOTA performance is attainable without explicit inversion of the input image.

	Struct.	BG		CLIP Similariy	
Method	**Dist.** ↓	**LPIPS** ↓	**SSIM** ↑	**Whole** ↑	**Edited** ↑
∅-Text Inversion [MHA+23]	13.4	60.7	84.1	24.8	21.9
Negative prompt [MIST23]	16.2	69.0	83.4	24.6	21.9
Proximal NPI [MIST23]	**7.4**	42.0	**86.0**	24.3	21.4
Plug-n-Play [TGBD23]	28.2	113.5	79.0	25.4	**22.6**
Direct Inversion [JZB+23]	11.7	54.6	84.8	25.0	22.1
P2P-Zero [PKSZ+23]	61.7	172.2	74.7	22.8	20.5
SAGE (ours)	11.0	**39.6**	**86.0**	**25.5**	22.0

Table 13: A quantitative analysis utilizing PieBench has been conducted. The data for all entries except for our own originates from [JZB+23]. BG denotes Background.

V.3.2.2 Qualitative Comparison

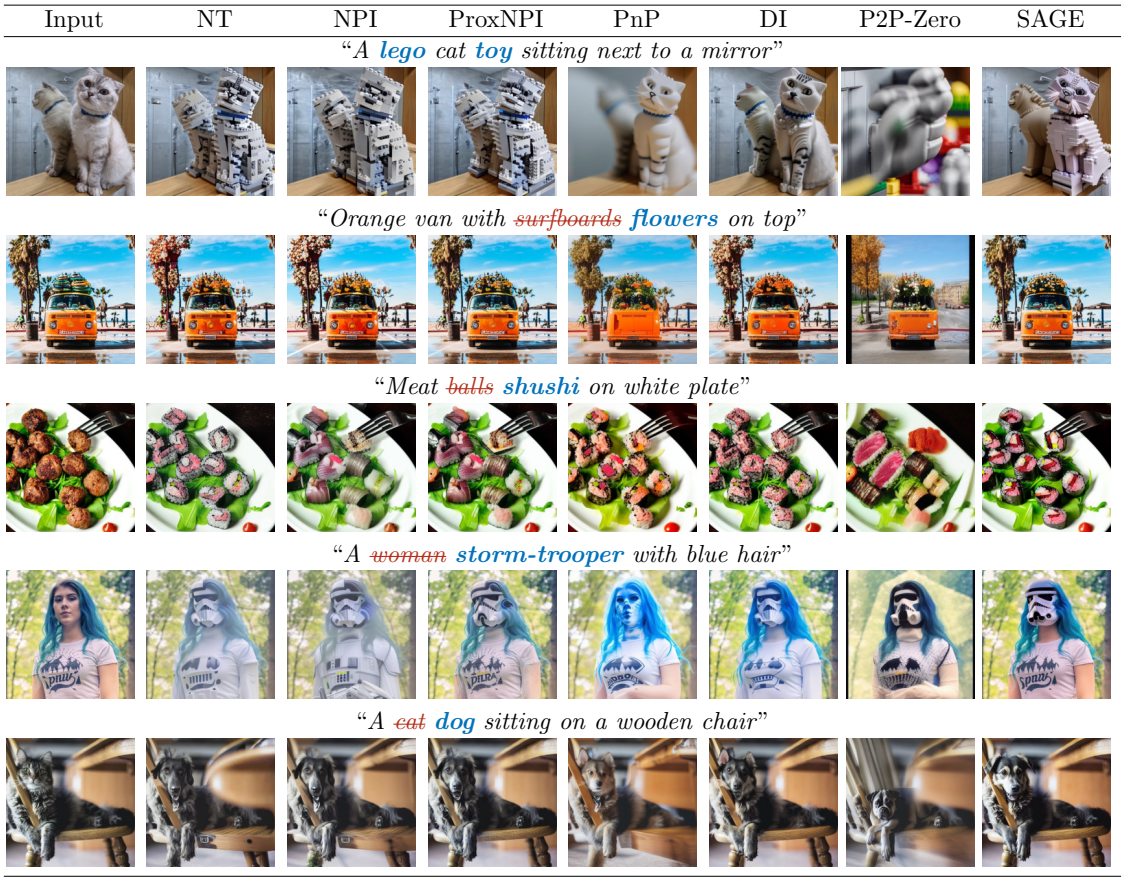


Figure 38: Qualitative analysis using SAGE. Here, we demonstrate examples of edits, including inserting words and swapping words.

Fig. 38 illustrates that our qualitative analysis aligns with the quantitative evaluation. Our method successfully retains the original image structure and content while providing robust editing performance. For instance, only SAGE and ProxNPI preserve the tree's appearance in the second row. Similarly, SAGE uniquely maintains the t-shirt sleeve, hair, and background trees' details and colors in the fourth row.

Additionally, our method generates more natural-looking images, especially ev-

ident in the dog and sushi examples. In the dog example, not only are the dog and chair preserved, but the dog's face and lighting appear more natural. In the sushi example, our method is the only one that maintains all image details and colors while producing a deeper red meat color and natural-looking *nori* algae. In contrast, other methods generate shapeless and unnatural sushi pieces. This advantage is attributed to our method's ability to reduce the guidance term over time, as discussed in Sec. V.2.2.

In Figure 39, we present additional qualitative examples illustrating the outcomes produced by SAGE. These examples were used to calculate the metrics shown in Tables 12, 13, and 15. We argue that because SA maps primarily encode low-resolution semantic features rather than precise shapes, SAGE excels at generating images that can edit the style and content of input images while keeping its structure. For instance, in the second row of Fig 39, SAGE modifies the shape of a fox, and in the third row, it adjusts the tulips' forms. Additionally, SAGE can remove elements from the original image and seamlessly fill the void with plausible content, as seen with the camera in the last row of Fig. 39. Notably, SAGE is the only method that applies the pixel-art style, demonstrated by the clown in Fig 39. This supports our hypothesis that SA maps capture low-resolution semantic features, evidenced by SAGE's ability to not only change the style but also significantly alter the type of chair (from modern to classic) to better match the new prompt in the last row of Fig 39.

V.3.2.3 Inference time and memory comparison

Method	**Time** (s) ↓	**Memory** (GB) ↓
∅-Text Inversion [MHA+23]	115.3	21.1
Negative prompt [MIST23]	26.6	38.9
Proximal NPI [MIST23]	23.9	38.9
Plug-n-Play [TGBD23]	**12.5***	38.9
Direct Inversion [JZB+23]	33.4	**12.4***
Direct Inversion *FP16* [JZB+23]	**12.5**	**7.4**
P2P-Zero[PKSZ+23]	52.7	25.0
SAGE (ours)	12.6	**7.4**
SAGE *FP32* (ours)	27.6	29.3

Table 14: Performance evaluation for both temporal and memory usage was conducted on a NVIDIA A100-40GB. The Time column illustrates the duration needed to create two 512×512 pixel images, subtracting the duration taken to generate a single image. This method effectively isolates the image generation time, excluding the model and data loading overhead. The Memory column indicates the peak memory allocation for a single image generation as reported by `nvidia-smi`.
The best results for FP32 precision methods are highlighted with an asterisk (*).

In Table 14, we display the memory usage and speed performance of the FP16 version of SAGE, which has the smallest memory footprint among the compared methods and is the second fastest by just a tenth of a second. The FP32 version remains competitive with the other top methods. The results for NT, PnP, P2P-Zero, and DI were obtained using DI's source code [JZB+23][3], while the results for

[3] `https://github.com/cure-lab/DirectInversion`, last accesed July 26, 2024

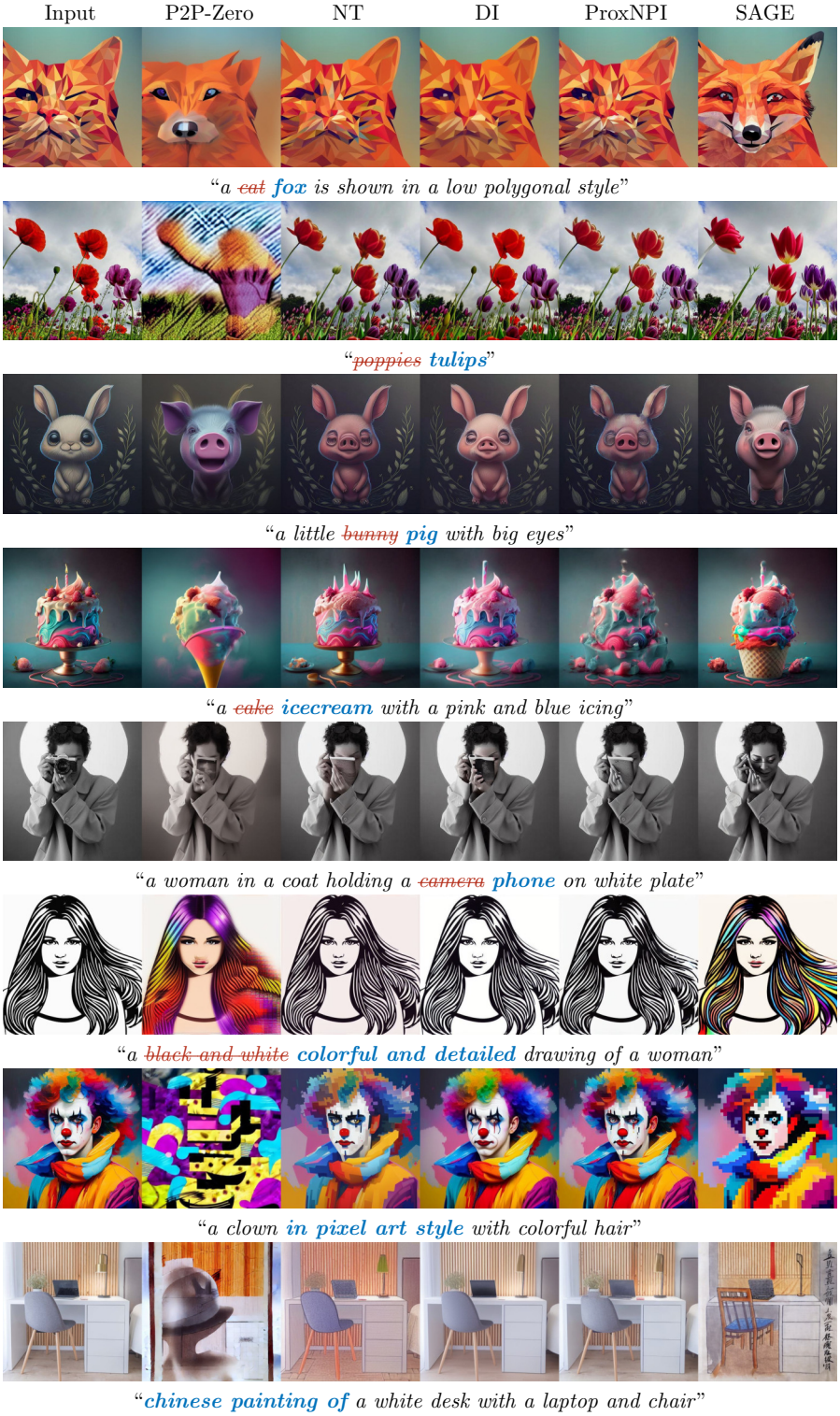


Figure 39: Samples from the PieBench dataset, as mentioned in Tables 12, 13, and 15. These images exemplify the variety of scenarios utilized in our evaluations and analyses.

NPI and ProxNPI were derived from the source code provided by [HWC+24][4]. All methods were executed in the same conda environment.

[4] `https://github.com/phymhan/prompt-to-prompt`, last accesed July 26, 2024

The efficiency of SAGE in both processing speed and memory usage, despite the intensive gradient-based guidance operation, can be attributed to two main factors. First, the computation of guidance is restricted to a subset of SA maps, which significantly decreases the overall computational burden. Second, and most importantly, SAGE omits the input image reconstruction step entirely, which not only simplifies the process but also reduces memory consumption. These factors collectively highlight SAGE's capability to deliver high performance while managing resource constraints effectively.

V.3.3 User Study

To reinforce the comparison with existing approaches detailed in Sec V.3.2, we conducted a user study. This study evaluates our method against others based on three key aspects: structure preservation, background preservation, and adherence to the prompt, along with overall user preference. We engaged 22 participants in *one versus one* comparisons using images from the PieBench *random editing* task. The methods compared were Negative Prompt Inversion [MIST23], Direct Inversion [MIST23], Proximal NPI [HWC+24], P2P-Zero [PKSZ+23], and our method.

Participants were shown different sets of images depending on the evaluation criteria. For structure preservation, the original image and two edited versions were presented. For background preservation, the input image was masked to highlight relevant areas. For prompt fidelity and overall user preference, only the target prompt and the edited versions were displayed. Images were presented in a randomized order to ensure unbiased judgments, as participants were unaware of which methods were used.

SAGE vs	**Structure**	**Background**	**Prompt**	**Global**
P2P-Zero [PKSZ+23]	93.8%	92.0%	83.0%	75.9%
∅-Text Inversion [MHA+23]	70.5%	69.6%	52.7%	54.5%
Direct Inversion [JZB+23]	55.4%	62.5%	52.7%	52.7%
Proximal NPI [HWC+24]	58.9%	58.0%	62.5%	59.8%
Average	69.6%	70.5%	62.7%	60.7%

Table 15: Outcomes of the user study, showing how often our method, SAGE, was preferred over other methods. A total of 1792 questions were answered by the participants.

The findings from this study, summarized in Table 15, align with our quantitative evaluations, showing a consistent preference for our method across all evaluation criteria. While the preference for our method is slightly narrow compared to Direct Inversion in terms of prompt fidelity and overall preference, a notable difference is observed in background preservation. This significant advantage in preserving the background underscores the effectiveness of our approach. Overall, the user study substantiates the strong performance of our method, affirming its strengths in both quantitative metrics and subjective user assessments.

Part III

Final remarks

Chapter VI

Final remarks

"No matter how many times you do it, you don't get used to the sadness –for me at least– of coming to the end of a film." — Paul Thomas Anderson

VI.1 Conclusions

This PhD dissertation has presented three significant advancements in the field of forensic facial imaging through the development and application of DL-based methodologies. Each of these contributions addresses a critical aspect of forensic analysis, providing innovative solutions that could enhance accuracy, efficiency, and reliability in forensic investigations.

Firstly, we developed a robust tool for accurately locating cephalometric landmarks on facial images. This tool overcomes the limitations of a small dataset by utilizing pre-trained facial landmark detection models and optimizing data usage through a shared conditional residual network across different landmarks. By incorporating a pre-trained 3D facial landmark detection model, we achieved reliable visibility estimation even with limited training data. Our systematic evaluation of each model component demonstrated that our method is three times more accurate than approaches based solely on pre-trained deformable 3D masks. Moreover, our method outperformed SOTA techniques in facial landmark localization on our cephalometric landmark dataset, showing a performance improvement of two times over the closest competitor. A user study with forensic anthropologists further validated our method, achieving human-comparable accuracy in 50% of cases. These strong performance results have led to the integration of our method's predictions as initial estimations in Skeleton-ID[1], a commercial AI-assisted forensic identification solution used when DNA or fingerprint analysis is not feasible.

Secondly, we introduced a novel architecture for face age editing capable of producing structural modifications while preserving relevant details of the original image. Our approach has two key contributions: a style-based strategy that combines style and content representations of the input image, conditioned on the target age, and a CUSP module that allows users to adjust the degree of structure preservation

[1]Skeleton-ID: `https://skeleton-id.com/`. *Last accessed on July 26, 2024.*

at inference time. Validation against six SOTA solutions on three different datasets showed that our method generates more natural-looking, age-accurate transformed images. It allows for more profound facial changes while preserving identity and modifying only age-related aspects. An extensive user study confirmed these findings, underscoring the effectiveness and usability of our approach.

Lastly, we revisited prompt-based image editing within diffusion models, challenging the conventional need for explicit input image reconstruction. Our investigation revealed that the DDIM inversion process alone contains sufficient information for effective editing, reconstruction of the original image is not necesary, thus simplifying the process by applying guidance exclusively to the $\mathcal{P}^{in}$ branch. This streamlined approach not only simplifies the editing process but also yields superior results, as demonstrated by our extensive comparative analyses. Our most significant contribution in this area is the introduction and validation of SA guidance as a superior mechanism for image editing tasks. Through quantitative analyses, ablation studies, and user feedback, we established that SA guidance, which captures a broader contextual understanding within images, facilitates better edits compared to traditional CA techniques. This method maintains closer fidelity to the original image content while accurately implementing the desired edits. The superiority of SA guidance was further supported by an extensive user study, where our method, SAGE, was preferred by 60.7% of participants over competing approaches, highlighting its potential to redefine standard practices in image editing within diffusion models.

Overall, these three works collectively advance the field of forensic facial imaging, providing powerful tools that enhance the accuracy, efficiency, and reliability of forensic investigations. Besides, each method developed in this dissertation has been rigorously validated through comprehensive user studies to ensure their practical applicability and effectiveness. This thorough validation process underscores the relevance and robustness of the proposed methods, ensuring they meet the stringent demands of forensic investigations and general image editing tasks, thereby enhancing the credibility and impact of its potential use in forensic analyses.

VI.2 Future Work

The future work for this research involves several key areas of development and improvement to enhance the performance and applicability of the proposed methods.

For the cephalometric landmark localization method, further evaluation of landmark visibility is essential. A more robust dataset is needed, including clear and consistent guidelines for determining visibility. This dataset should differentiate between landmarks occluded by posture or external objects and those where the position cannot be accurately determined, similar to previous studies [KMM+20]. To compare the effectiveness of our vertex normal-based visibility estimation with other methods, we propose training an ad-hoc classification layer at the end of a convolutional neural network on reliable data, as previously done in studies [RPC17] and [KMM+20]. Additionally, increasing the number of samples in the dataset could significantly improve model quality, narrowing the performance gap between human and machine, similar to other biomedical applications [BDGB+19, EKN+17,

LZL+17, GPC+16]. Acquiring and utilizing a bigger better-quality dataset will be a priority for future work. Furthermore, to enhance the few-shot performance of our model, incorporating different sources of knowledge, such as intermediate activation maps from pre-trained models along with raw RGB values, will be explored.

In the domain of face aging, we aim to extend the CUSP module to encompass a broader range of image editing tasks. This extension will leverage the advantages of structural preservation more comprehensively, thereby enhancing the versatility and efficacy of the face aging methodology. Another prospective research direction involves incorporating expert knowledge in face aging [BAPJ10] into the workflow, along with integrating identity preservation constraints into the model [WTLG18, APCO21]. These enhancements will improve the utility of age-edited images in forensic analyses by ensuring the key identity features of individuals are accurately maintained.

For image editing, future research will focus on refining and advancing the SA guidance mechanism. One potential improvement area is the development of an alternative guidance term that does not rely on gradient calculation, which could significantly accelerate the image generation process. Although the PieBench dataset serves as an excellent benchmark for various SOTA methods in this domain, it remains necessary to evaluate our method's performance in tasks specific to facial imaging, such as face aging (analyzing potential age biases), facial expression editing, and feature manipulation.

In both face aging and image editing, the development of more sophisticated user interfaces will be pursued. These interfaces will enable users to interact more effectively with the models, providing enhanced control over the editing process and improving the overall user experience. Additionally, the validation and integration of the proposed methodologies into commercial forensic imaging tools will be a pivotal area of future work, in collaboration with the Panacea Cooperative Research team. This integration will ensure that the developed solutions are accessible to forensic experts and can be effectively employed in real-world forensic investigations.

VI.3 Publications

Works published in JCR-indexed journals

Gomez-Trenado, G., Mesejo, P., and Cordon, O. (2023). **Cascade of convolutional models for few-shot automatic cephalometric landmarks localization**. *Engineering Applications of Artificial Intelligence*, Volume 123, Part B. `https://doi.org/10.1016/j.engappai.2023.106391` (JCR 2023. Journal Impact Factor: 7.5; Cat.: COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE; Pos.: 24/197; Q1). Related to Chapter III.

Gomez-Trenado, G., Mesejo, P., and Cordon, O. (2024). **Don't Forget your Inverse DDIM for Image Editing**. Submitted to *IEEE Transactions on Multimedia*. Related to Chapter V.

Works published in other journals

Ibáñez, Ó., Alemán, I., Bermejo, E., Corbal, I., Cordón, Ó., Damas, S., Gomez-Trenado, G., Gómez, I., Gómez, Ó., González, A., Macías, M., Martos, R., Mesejo, P., Panizo, M., Prada, K., and Valsecchi, A., (2020). **El proyecto Skeleton-ID: hacia una identificación humana más rápida, objetiva y precisa**. *Revista Internacional de Antropología y Odontología Forense / International Journal of Forensic Anthropology and Odontology*, Volume 3, Number 2, pages 71-88.

Works published in international conferences

Gomez-Trenado, G., Lathuilière, S., Mesejo, P., and Cordón, Ó. (2022). **Custom structure preservation in face aging**. *European Conference on Computer Vision* (ECCV'22), in Tel-Aviv, Israel. Related to Chapter IV.

Gomez-Trenado, G., Mesejo, P., Ibáñez, Ó., Valsecchi, A., and Cordón, Ó. (2019). **Automatic localization of cephalometric landmarks using convolutional networks**. *11th International Scientific Meeting of the Spanish Association of Forensic Anthropology and Odontology* (AEAOF'19), in Pastrana, Spain.

Gomez-Trenado, G., Mesejo, P., Ibáñez, Ó., and Valsecchi, A. (2019). **Automatic Cephalometric Landmarks Localization using Deep Convolutional Neural Networks**. *18th Biennial Meeting of the International Association of Craniofacial Identification* (IACI'19), in Baton Rouge, Louisiana, USA.

VI.4 Acknowledgements

This work was supported by: *i)* the Spanish Ministry of Science, Innovation, and Universities (MICIU) under grant FPU19/00591; *ii)* MICIU and European Regional Development Funds (ERDF) under grant EXASOCO (PGC2018-101216-B-I00); *iii)* by MICIU/AEI/10.13039/501100011033 and by "ERDF A way of making Europe" under grant CONFIA (PID2021-122916NB-I00); and *iv)* by the Regional Government of Andalusia and the University of Granada under grants EXAISFI (P18-FR-4262) and FORAGE (B-TIC-456-UGR20), including European Regional Development Funds (ERDF). Funding for Open Access publication was provided by the University of Granada: CBUA, Spain.

Chapter VII

Bibliography

[ABD17] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *IEEE International Conference on Image Processing*, 2017.

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 2017.

[AIA+14] Salina Mohd Asi, Nor Hidayah Ismail, Roshahida Ahmad, Effirul Ikhwan Ramlan, and Zainal Arif Abdul Rahman. Automatic craniofacial anthropometry landmarks detection and measurements for the orbital region. *Procedia Computer Science*, 42:372–377, 2014.

[ALTK19] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Attribute manipulation generative adversarial networks for fashion images. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[APCO21] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. *ACM Transactions on Graphics*, 40(4):1–12, 2021.

[BAPJ10] S. Black, A. Aggrawal, and J. Payne-James. *Age Estimation in the Living: The Practitioner's Guide*. Wiley, 2010.

[BB23] Christopher Michael Bishop and Hugh Bishop. *Deep Learning - Foundations and Concepts*. 1 edition, 2023.

[BDGB+19] Antoine Buetti-Dinh, Vanni Galli, Sören Bellenberg, Olga Ilie, Malte Herold, Stephan Christel, Mariia Boretska, Igor V Pivkin, Paul Wilmes, Wolfgang Sand, et al. Deep neural networks outperform human expert's capacity in characterizing bioleaching bacterial biofilm composition. *Biotechnology Reports*, 22:e00321, 2019.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[BSAG18] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[BW20] Bjorn Browatzki and Christian Wallraven. 3FabRec: Fast few-shot face alignment by reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6110–6120, 2020.

[BWW20] Kaye N Ballantyne and Linzi Wilson-Wilde. Assessing the reliability and validity of forensic science–an industry perspective. *Australian Journal of Forensic Sciences*, 52(3):275–281, 2020.

[CÁICÁ+15] B Rosario Campomanes-Álvarez, Oscar Ibánez, Carmen Campomanes-Álvarez, Sergio Damas, and Oscar Cordón. Modeling facial soft tissue thickness for automatic skull-face overlay. *IEEE Transactions on Information Forensics and Security*, 10(10):2057–2070, 2015.

[CAIN+14a] B Rosario Campomanes-Alvarez, O Ibáñez, F Navarro, I Alemán, M Botella, S Damas, and O Cordón. Computer vision and soft computing for automatic skull–face overlay in craniofacial superimposition. *Forensic Science International*, 245:77–86, 2014.

[CAIN+14b] Blanca Rosario Campomanes-Alvarez, Oscar Ibáñez, Fernando Navarro, Inmaculada Alemán, Oscar Cordón, and Sergio Damas. Dispersion assessment in the location of facial landmarks on photographs. *International Journal of Legal Medicine*, 129:227–236, 2014.

[CBGB20] Prashanth Chandran, Derek Bradley, Markus Gross, and Thabo Beeler. Attention-driven cropping for very high resolution facial landmark detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5861–5870, 2020.

[CC09] Angi M. Christensen and Christian M. Crowder. Evidentiary standards for forensic anthropology. *Journal of Forensic Sciences*, 54(6):1211–1216, 2009.

[CCC+23] Guillaume Couairon, Marlène Careil, Matthieu Cord, Stéphane Lathuilière, and Jakob Verbeek. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision*, 2023.

[CCK+18] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[COG+17] Zuzana Caplova, Zuzana Obertova, Daniele M. Gibelli, Debora Mazzarelli, Tony Fracasso, Peter Vanezis, Chiarella Sforza, and Cristina Cattaneo. The reliability of facial recognition of deceased persons on photographs. *Journal of Forensic Sciences*, 62(5):1286–1291, 2017.

[CPB14] Angi M. Christensen, Nicholas V. Passalacqua, and Eric J. Bartelink. Chapter 1 - introduction. In Angi M. Christensen, Nicholas V. Passalacqua, and Eric J. Bartelink, editors, *Forensic Anthropology*, pages 1–17. Academic Press, San Diego, 2014.

[CT04] Timothy F Cootes and Chris J Taylor. Statistical models of appearance for computer vision. Technical report, University of Manchester, 2004.

[CTM+21] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[CUYH20] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[DBK+21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[DCD+23] Laura Donato, Rossana Cecchi, Sara Dagoli, Michele Treglia, Margherita Pallocci, Claudia Zanovello, Douglas H Ubelaker, and Luigi Marsella. Facial age progression: Review of scientific literature and value for missing person identification in forensic medicine. *Journal of Forensic and Legal Medicine*, page 102614, 2023.

[DCI20] Sergio Damas, Oscar Cordón, and Oscar Ibáñez. *Handbook on craniofacial superimposition: The MEPROCS project*. Springer Nature, 2020.

[DCIn+11] Sergio Damas, Oscar Cordón, Oscar Ibáñez, Jose Santamaría, Inmaculada Alemán, Miguel Botella, and Fernando Navarro. Forensic identification by computer-aided craniofacial superimposition: A survey. *ACM Computer Surveys*, 43(4):27:1–27:27, 2011.

[DGMH11] Joaquín Derrac, Salvador García, Daniel Molina, and Francisco Herrera. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, 1(1):3–18, 2011.

[DN21] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 2021.

[DV15] Josh P Davis and Tim Valentine. Human verification of identity from photographic images. In *Forensic facial identification: Theory and*

practice of identification from eyewitnesses, composites and CCTV, chapter 9, pages 211–238. John Wiley & Sons, 2015.

[EKN+17] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[Far94] L.G. Farkas. *Anthropometry of the Head and Face*. Raven Press, 1994.

[FGH10] Yun Fu, Guodong Guo, and Thomas S Huang. Age synthesis and estimation via faces: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1955–1976, 2010.

[FGW+19] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[Fra10] Daniel Franklin. Forensic age estimation in human skeletal remains: Current concepts and future directions. *Legal Medicine*, 12(1):1 – 7, 2010.

[FWS+18] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3D face reconstruction and dense alignment with position map regression network. In *IEEE/CVF European Conference on Computer Vision*, pages 534–551, 2018.

[GAA+23] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023.

[GBB11] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[GBR+12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[Geo07] R.M. George. *Facial Geometry: Graphic Facial Analysis for Forensic Artists*. Charles C. Thomas, 2007.

[GORT+16] D. Gibelli, Z. Obertová, S. Ritz-Timme, P. Gabriel, T. Arent, M. Ratnayake, D. De Angelis, and C. Cattaneo. The identification of living persons on images: A literature review. *Legal Medicine*, 19:52–60, 2016.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

[GPC+16] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Journal of the American Medical Association*, 316(22):2402–2410, 2016.

[guo] gan-based virtual-to-real image translation for urban scene semantic segmentation.

[GWK+18] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.

[GZ19] Liang Gonog and Yimin Zhou. A review: Generative adversarial networks. In *IEEE Conference on Industrial Electronics and Applications*, pages 505–510. IEEE, 2019.

[GZY+20] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3D dense face alignment. In *IEEE/CVF European Conference on Computer Vision*, 2020.

[HB17] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE/CVF International Conference on Computer Vision*, 2017.

[HFC+15] Patrik Huber, Zhen-Hua Feng, William Christmas, Josef Kittler, and Matthias Rätsch. Fitting 3D morphable face models using local features. In *IEEE International Conference on Image Processing*, pages 1195–1199. IEEE, 2015.

[HIWK15] María Isabel Huete, Oscar Ibáñez, Caroline Wilkinson, and Tzipi Kahana. Past, present, and future of craniofacial superimposition: Literature and international surveys. *Legal Medicine*, 17(4):267–278, 2015.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[HKSC19] Zhenliang He, Meina Kan, S. Shan, and Xilin Chen. S2gan: Share aging factors across ages and share aging trends among individuals. *IEEE/CVF International Conference on Computer Vision*, 2019.

[HLBK18] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *IEEE/CVF European Conference on Computer Vision*, 2018.

[HMT+23] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

[HNW+19] Xiaodan Hu, Mohamed A Naiel, Alexander Wong, Mark Lamm, and Paul Fieguth. Runet: A robust unet architecture for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[HRU+17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

[HS09] Max M Houck and Jay A Siegel. *Fundamentals of forensic science*. Academic Press, 2009.

[HS21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[HSK+12] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[HWC+24] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, Di Liu, Qilong Zhangli, Jindong Jiang, Zhaoyang Xia, Akash Srivastava, and Dimitris Metaxas. Improving tuning-free real image editing with proximal guidance. In *Winter Conference on Applications of Computer Vision*, 2024.

[HYL+21] Yangyu Huang, Hao Yang, Chong Li, Jongyoo Kim, and Fangyun Wei. Adnet: Leveraging error-bias towards normal direction in face alignment. In *IEEE/CVF International Conference on Computer Vision*, pages 3080–3090, 2021.

[HZ10] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *International Conference on Pattern Recognition*, 2010.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[IZZE17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[JZB+23] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2304.04269*, 2023.

[KAH+20] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020.

[KALL17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *International Conference on Learning Representations*, 2017.

[KH91] Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in Neural Information Processing Systems*, 4, 1991.

[KKC21] Daejin Kim, Mohammad Azam Khan, and Jaegul Choo. Not just compete, but collaborate: Local image-to-image translation via cooperative mask prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[KKSK20] Paramjit Kaur, Kewal Krishan, Suresh K Sharma, and Tanuj Kanchan. Facial-recognition algorithms: A literature review. *Medicine, Science and the Law*, 60(2):131–139, 2020.

[KKY22] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[KLA19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[KLA+20] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[KMM+20] Abhinav Kumar, Tim K. Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[KSSS14] Ira Kemelmacher-Shlizerman, Supasorn Suwajanakorn, and Steven M Seitz. Illumination-aware age progression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014.

[KW13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[KWRB11] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *IEEE/CVF International Conference on Computer Vision*, pages 2144–2151, 2011.

[Lar23] Clark Spencer Larsen. *A companion to biological anthropology*. John Wiley & Sons, 2023.

[LBK17] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, 2017.

[LGH03] Russell Lain, Chris Griffiths, and John MN Hilton. Forensic dental and medical response to the bali bombing. *Medical Journal of Australia*, 179(7):362–365, 2003.

[LLWT15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE/CVF International Conference on Computer Vision*, 2015.

[LTH+18] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *IEEE/CVF European Conference on Computer Vision*, 2018.

[LZL+17] Kisuk Lee, Jonathan Zung, Peter Li, Viren Jain, and H Sebastian Seung. Superhuman accuracy on the snemi3D connectomics challenge. *arXiv preprint arXiv:1706.00120*, 2017.

[LZU+17] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic DENOYER, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, 2017.

[McH12] Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22:276 – 282, 2012.

[MG15] Nicholas Márquez-Grant. An overview of age estimation in forensic anthropology: perspectives and practical considerations. *Annals of Human Biology*, 42(4):308–322, 2015.

[MHA+23] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[MHP21] Farkhod Makhmudkhujaev, Sungeun Hong, and In Kyu Park. Re-aging gan: Toward personalized face age transformation. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[MHS+22] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

[MIM24] Rubén Martos, Oscar Ibáñez, and Pablo Mesejo. Artificial Intelligence in Forensic Anthropology: State of the Art and Skeleton-ID Project. In Ann H. Ross and Jason H. Byrd, editors, *Methodological and Technological Advances in Death Investigations*, chapter 3, pages 83–153. Academic Press, 2024.

[MIST23] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023.

[MK18] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations*, 2018.

[MMI+20] Pablo Mesejo, Rubén Martos, Óscar Ibáñez, Jorge Novo, and Marcos Ortega. A survey on artificial intelligence techniques for biomedical image analysis in skeleton-based forensic human identification. *Applied Sciences*, 10(14):4703, 2020.

[MO14] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[Mul12] Joe Mullins. Age progression and regression. In *Craniofacial identification*, chapter 6. Cambridge University Press, Cambridge, 2012.

[Mur22] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

[MVIA18] Rubén Martos, Andrea Valsecchi, Oscar Ibáñez, and Inmaculada Alemán. Estimation of 2D to 3D dimensions and proportionality indices for facial examination. *Forensic Science International*, 287:142–152, 2018.

[MY20] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *International Joint Conference on Neural Networks*, 2020.

[OESF+20] Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *IEEE/CVF European Conference on Computer Vision*, 2020.

[PAM+18] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *IEEE/CVF European Conference on Computer Vision*, 2018.

[PCWS22] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[PHM+19] Ji-Hoon Park, Hye-Won Hwang, Jun-Ho Moon, Youngsung Yu, Hansuk Kim, Soo-Bok Her, Girish Srinivasan, Mohammed Noori A Aljanabi, Richard E Donatelli, and Shin-Jae Lee. Automated identification of cephalometric landmarks: Part 1 - comparisons between the latest deep-learning methods YOLOV3 and SSD. *The Angle Orthodontist*, 89(6):903–909, 2019.

[PHSC18] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[PKSZ+23] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH*, pages 1–11, 2023.

[PLF+19] Lucas Faria Porto, Laise Nascimento Correia Lima, Marta Regina Pinheiro Flores, Andrea Valsecchi, Oscar Ibañez, Carlos Eduardo Machado Palhares, and Flavio de Barros Vidal. Automatic cephalometric landmarks detection on frontal faces: An approach based on supervised learning techniques. *Digital Investigation*, 30:108–116, 2019.

[PNR+21] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[PZW+20] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 2020.

[RAP+21] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a StyleGAN encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[RBL+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Internation Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[RLH+20] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2020.

[RLJ+23] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[RN20] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson, 2020.

[Rob18] Simon Robins. Missing in migration: From research to practice. *Practicing Anthropology*, 40(2):24–27, 2018.

[RPC17] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, 2017.

[RSC98] DJ Rudolph, PM Sinclair, and JM Coggins. Automatic computerized radiographic identification of cephalometric landmarks. *American Journal of Orthodontics and Dentofacial Orthopedics*, 113(2):173–179, 1998.

[RSIM19] Gururajaprasad Kaggal Lakshmana Rao, Arvind Channarayapatna Srinivasa, Yulita Hanum P Iskandar, and Norehan Mokhtar. Identification and analysis of photometric points on 2D facial images: a machine learning approach in orthodontics. *Health and Technology*, 9:715–724, 2019.

[RTG18] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018.

[RTVG15] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *IEEE/CVF International Conference on Computer Vision Workshops*, 2015.

[SCD+17] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE/CVF International Conference on Computer Vision*, 2017.

[SCGC19] Carl N Stephan, Jodi M Caple, Pierre Guyomarc'h, and Peter Claes. An overview of the latest developments in facial imaging. *Forensic Sciences Research*, 4(1):10–28, 2019.

[SCS+22] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022.

[SDBR15] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations*, 2015.

[SDR+16] Andreas Schmeling, Reinhard Dettmeyer, Ernst Rudolf, Volker Vieth, and Gunther Geserick. Forensic age estimation: methods, certainty, and the law. *Deutsches Ärzteblatt International*, 113(4):44, 2016.

[SF19] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. *Advances in Neural Information Processing Systems*, 32, 2019.

[SK19] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[SLJ+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[SME21] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[SO21] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, 2021.

[SP97] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, 1997.

[SSLS18] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[STZP13] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE/CVF International Conference on Computer Vision Workshops*, 2013.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[SZJ+19] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019.

[TBTBD22] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[TGBD23] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[TXSY19] Hao Tang, Dan Xu, Nicu Sebe, and Yan Yan. Attention-guided generative adversarial networks for unsupervised image-to-image translation. In *International Joint Conference on Neural Networks*, 2019.

[USK18] Douglas H. Ubelaker, Austin Shamlou, and Amanda Kunkle. Contributions of Forensic Anthropology to Positive Scientific Identification: A Critical Review. *Forensic Sciences Research*, 4(1):45–50, 10 2018.

[VD15] Tim Valentine and Josh P Davis. Forensic facial identification: A practical guide to best practice. In *Forensic facial identification: Theory and practice of identification from eyewitnesses, composites and CCTV*, chapter 13, pages 323–347. John Wiley & Sons, 2015.

[VJ01] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages 511–518, 2001.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[WBH+21] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *IEEE/CVF International Conference on Computer Vision*, pages 3681–3691, 2021.

[WBSS04] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[WCY+16] Wei Wang, Zhen Cui, Yan Yan, Jiashi Feng, Shuicheng Yan, Xiangbo Shu, and Nicu Sebe. Recurrent face aging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[WHZ+21] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021.

[Wil45] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[Wil15] Caroline Wilkinson. Craniofacial analysis and identification. In *Forensic facial identification: Theory and practice of identification from eyewitnesses, composites and CCTV*, chapter 5, pages 93–126. John Wiley & Sons, 2015.

[WJ19] Yue Wu and Qiang Ji. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127(2):115–142, 2019.

[WLL+20] Jun Wan, Zhihui Lai, Jun Liu, Jie Zhou, and Can Gao. Robust face alignment by multi-order high-precision hourglass network. *IEEE Transactions on Image Processing*, 30:121–133, 2020.

[WQY+18] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[WR12] Caroline Wilkinson and Christopher Rynn. *Craniofacial identification*. Cambridge University Press, Cambridge, 2012.

[WSC+20] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2020.

[WTLG18] Zongwei Wang, Xu Tang, Weixin Luo, and Shenghua Gao. Face aging with identity-preserved conditional generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[WYKN20] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3):1–34, 2020.

[XKV21] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021.

[YHWJ18] Hongyu Yang, Di Huang, Yunhong Wang, and Anil K Jain. Learning face age progression: A pyramid architecture of gans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[YNGH21] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A latent transformer for disentangled face editing in images and videos. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[YPN+21] Xu Yao, Gilles Puy, Alasdair Newson, Yann Gousseau, and Pierre Hellier. High resolution face age editing. In *International Conference on Pattern Recognition*, 2021.

[ZIE+18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[ZLL+16] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3D solution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2016.

[ZPIE17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE/CVF International Conference on Computer Vision*, 2017.

[ZRA23] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision*, 2023.

[ZSQ17] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[ZZP+17] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Multimodal image-to-image translation by enforcing bi-cycle consistency. In *Advances in Neural Information Processing Systems*, 2017.