ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dipartimento di Informatica - Scienza e Ingegneria

Corso di Laurea Triennale in informatica

INVECCHIAMENTO FACCIALE NELLE IMMAGINI: ANALISI E CONFRONTO DI MODELLI DI AI

Relatrice:
Prof.
FRANCO ANNALISA

Correlatori:
Dott.
NICOLÒ DI DOMENICO
Prof.
GUIDO BORGHI

Presentata da:
Giuseppe Spathis

Sessione Ottobre 2024
Anno Accademico 2023/2024

Indice

Capitolo 1

Introduzione

Il presente lavoro indaga l'ambito delle tecniche di trasformazione delle immagini facciali. Prendendo le mosse dalla tecnica di face morphing e, in particolare, dai suoi possibili usi criminosi per eludere i controlli visivi in luoghi con accesso riservato quali aeroporti e banche, si elabora in questa sede una metodologia atta a rendere più efficaci i sistemi di riconoscimento facciale trainandoli con un dataset di immagini invecchiate.

Il face morphing consiste nella creazione di un'immagine composita che combina le caratteristiche facciali di due soggetti, unendo sia informazioni di texture che di geometria [1, 2]. La proporzione tra queste due componenti può essere variata per creare un'immagine che somigli maggiormente a uno dei due contributori. Tale processo avviene attraverso due tecniche principali: blending per la texture e warping per la geometria. Il blending si occupa di fondere le caratteristiche superficiali della pelle dei due volti, come colore e dettagli cutanei, creando una miscela fluida delle strutture del derma in base alla proporzione scelta (ad esempio 50:50 o 70:30, come si può osservare in Figura 1.1). Il warping,

invece, riguarda l'organizzazione geometrica del volto, ossia la forma e la disposizione dei tratti facciali come occhi, naso e bocca. Durante il warping, i punti chiave del viso vengono deformati per combinare gradualmente le geometrie di entrambe le facce, ottenendo un risultato realistico. Una volta creata l'immagine morphed, è possibile post-processarla per rimuovere eventuali tracce, evidenti o impercettibili, di manipolazione.

Figura 1.1: I due contributori (a sinistra e a destra) e i tre morph intermedi con proporzioni di contributo di texture e geometria rispettivamente di 70:30, 50:50 e 30:70.

Il morphing è stato utilizzato con finalità creative e pratiche in diversi settori. Nel mondo del cinema, ad esempio, è stato impiegato nel film *Terminator 2, Il giorno del giudizio* (1991, regista James Cameron), in cui il T-1000, un robot terminator in metallo liquido, usa il face morphing per trasformarsi in altre persone.

Un'altra applicazione del face morphing si trova nell'analisi genetica predittiva. Alcuni software, come BabyMaker, permettono alle coppie di vedere una rappresentazione di come potrebbero apparire i loro figli, combinando i tratti facciali di entrambi i genitori. Utilizzando tecniche di morphing, il software mescola la texture e la geometria dei volti parentali per creare un'immagine

che rappresenta una possibile combinazione dei loro tratti genetici. Questo tipo di applicazione, sebbene sia principalmente per curiosità e intrattenimento, sfrutta il potenziale del morphing per rappresentare scenari futuri basati su dati reali.

Tuttavia, le stesse tecniche utilizzate per scopi ludici o scientifici possono avere applicazioni malevole, soprattutto nel settore della sicurezza. Negli ultimi anni, i sistemi di verifica dell'identità basati sul riconoscimento facciale hanno assunto un ruolo di crescente rilevanza, in particolare all'interno di contesti ad alta sicurezza, come gli *Automated Border Control (ABC) gate* negli aeroporti. Tuttavia, tali sistemi sono sempre più vulnerabili a tecniche avanzate di manipolazione delle immagini, tra cui i *Face Morphing Attack* [3, 4]. Questi attacchi si basano sulla fusione di due identità distinte in un'unica immagine, consentendo a entrambi gli individui coinvolti di superare i controlli di verifica. Tale minaccia rappresenta un rischio concreto per le applicazioni di sicurezza, in quanto, con questo stratagemma, si riesce spesso a eludere sia gli operatori umani, sia i sistemi automatici [5, 6].

Un esempio concreto di Face Morphing Attack riguarda l'azione congiunta di un criminale e un complice. Il criminale, che non potrebbe superare i controlli perché schedato, collabora con il complice creando una foto morphed che combina i volti di entrambi e viene inserita nel passaporto del complice. Quando il malfattore presenta il documento, i sistemi automatici riconoscono l'immagine morphed come associata al complice, che può quindi oltrepassare i varchi automatici permettendo allo stesso tempo il passaggio del criminale.

I sistemi di *Morphing Attack Detection (MAD)* [7] sono stati

sviluppati per contrastare tali minacce, ma presentano ancora limiti significativi, in parte dovuti alla scarsità di protocolli di validazione comuni e di dataset pubblicamente disponibili. Inoltre, uno degli aspetti meno indagati della questione è il modo in cui le variazioni biometriche naturali, come l'invecchiamento, influenzano la capacità di rilevamento di questi sistemi. Poiché i documenti di viaggio leggibili elettronici (*electronic Machine Readable Travel Documents*, eMRTD) hanno una durata standard di dieci anni, le variazioni fisiologiche nei volti legate al passare del tempo devono essere gestite adeguatamente dai sistemi di riconoscimento e verifica.

L'obiettivo di questa tesi è partire dal dataset ONOT [8] — dataset sintetico di immagini di fototessere ad alta qualità che rispettano gli standard ISO/IEC — e utilizzare modelli di face aging per generare *AMONOT (Aged Morphed ONOT)*, un nuovo dataset che integri l'invecchiamento dei volti al fine di migliorare la robustezza dei sistemi MAD nei confronti delle variazioni biometriche indotte dall'età.

Questo dataset, interamente composto da immagini sintetiche generate mediante tecniche di age progression, è stato progettato per simulare in maniera accurata il processo di invecchiamento e per consentire una valutazione più completa dei sistemi MAD in condizioni che rispecchino scenari reali. La scelta di creare volti invecchiati con intervalli di dieci anni deriva da esigenze pratiche dei sistemi di controllo automatico, in cui le immagini vengono confrontate con foto scattate in momenti molto distanti nel tempo.

I risultati attesi dal presente lavoro consistono nell'individua-

zione di un modello che permetta di ottenere un invecchiamento realistico delle immagini, preservando le caratteristiche distintive dell'originale.

Capitolo 2

Analisi della letteratura

Il *face aging*, ovvero la manipolazione dell'età di volti nelle immagini, rappresenta un fenomeno di grande interesse, sia nel mondo accademico che in ambito industriale, trovando applicazione in numerosi settori della società contemporanea, fra i quali l'intrattenimento, la medicina legale e la sanità. In campo forense, il face aging si rivela di fondamentale importanza per la ricostruzione dei volti di persone scomparse o latitanti. Emblematico è il caso di Denise Pipitone, bambina di 3 anni scomparsa nel 2004 a Mazara del Vallo, per la quale sono state generate immagini del possibile aspetto attuale mediante avanzati algoritmi di progressione dell'età. In campo sanitario la manipolazione dell'età viene impiegata per lo studio dei cambiamenti fisionomici legati all'invecchiamento, nonché per la diagnosi e il monitoraggio di malattie genetiche rare, come la progeria, che provoca una senescenza accelerata. Nell'industria cinematografica, la modifica dell'età degli attori è una pratica molto diffusa, realizzata spesso attraverso il trucco o effetti visivi speciali. Sebbene tali tecniche possano produrre risultati impressionanti, come nel film *The Curious Case of Be-*

njamin Button (2008) di David Fincher, in cui il protagonista, per esigenze di trama, è sottoposto a processi di invecchiamento/-ringiovanimento, questi meccanismi sono estremamente lunghi e complessi. Di conseguenza, lo sviluppo di algoritmi automatici, robusti e di alta qualità per la modifica dell'età è fortemente auspicabile. Tuttavia, modificare l'età di un volto è un compito intrinsecamente difficile, poiché anche piccoli *artefatti*, ovvero errori o distorsioni o anomalie che emergono nei dati generati da un modello, e tali da non riflettere correttamente la realtà o il contesto desiderato, possono essere facilmente percepiti dall'occhio umano e compromettere la qualità del risultato finale. Per questa ragione, il focus di molte ricerche è stato rivolto alla produzione di immagini nitide, fotorealistiche e senza artefatti.

Il face aging non consiste soltanto nel modificare superficialmente la texture della pelle, ma comporta cambiamenti significativi delle caratteristiche biometriche del viso, cercando comunque di mantenere l'identità della persona ritratta [9]. Quest'ultima esigenza rappresenta una delle principali sfide nel campo: modificare l'età di un volto senza alterare tratti identificativi fondamentali. Tale problema è particolarmente complesso quando si richiedono grandi variazioni d'età, poiché le modifiche devono apparire convincenti, pur preservando l'identità dell'individuo.

Tra gli approcci più tradizionali, possiamo annoverare modelli basati su architetture *autoencoder*, cioè reti neurali costituite da un encoder che riduce l'input a una rappresentazione compressa e un decoder che cerca di ricostruire l'immagine originale partendo da tale rappresentazione. Questo processo di compressione e ricostruzione può essere adattato per la generazione di nuove

immagini, e quindi per la manipolazione dell'età usando varianti di autoencoder, come ad esempio i *Variational Autoencoder (VAE)* [10] che adottano un approccio probabilistico. A differenza dei normali autoencoder, i VAE non mappano l'input direttamente a una rappresentazione codificata fissa, ma apprendono una distribuzione probabilistica. L'encoder genera una distribuzione, da cui viene campionato un vettore latente, mentre il decoder ricostruisce l'immagine a partire da questa distribuzione. Ciò permette di generare nuove immagini variando casualmente i campioni nello spazio latente e rendendo i VAE utili per creare prodotti realistici e coerenti.

I modelli basati su *Generative Adversarial Networks (GAN)* [11] hanno poi rappresentato un punto di svolta nella generazione di immagini di alta qualità. Le GAN operano tramite l'interazione tra due reti neurali: un generatore, che tenta di creare immagini realistiche, e un discriminatore, che ha il compito di distinguere le immagini reali e quelle generate. Questo processo iterativo permette al generatore di migliorare progressivamente, cercando di ingannare il discriminatore. Tuttavia, sebbene le GAN abbiano fissato nuovi standard per la sintesi di immagini naturali, esse sono caratterizzate da due grandi difetti: la presenza di piccoli artefatti [11] e l'instabilità del processo di addestramento [11]. Le limitazioni suddette si riflettono anche nei modelli di face aging basati su GAN, i quali spesso producono sfondi sfocati e, piú in generale, non riescono a garantire una trasformazione dell'età fine e dettagliata. Successivamente, é stata proposta un'evoluzione delle GAN, chiamata *StyleGAN* [12], che introduce un nuovo approccio alla generazione delle immagini. A differenza delle

GAN tradizionali, StyleGAN separa i vari fattori di variazione delle immagini, come colore, forma e caratteristiche di alto e basso livello, attraverso uno spazio latente intermedio, il cosiddetto "W space". Ciò consente un controllo più preciso sugli attributi delle immagini, permettendo di manipolare lo stile senza alterare altri aspetti importanti, quali l'identità. Uno degli elementi distintivi di StyleGAN è l'uso dell'*Adaptive Instance Normalization (AdaIN)*, che separa chiaramente il contenuto dallo stile, rendendo possibile applicare diversi stili allo stesso volto, mantenendo in ogni caso invariata la struttura del viso.

Negli ultimi anni, sono stati introdotti i *diffusion model* [13, 14], che hanno guadagnato notevole popolarità nel campo della generazione di immagini. Essi operano partendo da una distribuzione casuale di rumore e trasformando progressivamente il rumore in un'immagine di alta qualità. Il processo di diffusione include una fase iniziale in cui un'immagine viene degradata aggiungendo rumore, seguita da una fase di denoising, nella quale il modello ripristina l'immagine rimuovendo il rumore gradualmente. I modelli di diffusione possono essere utilizzati sia in modo incondizionato, per cui il modello genera immagini senza alcuna guida esterna, sia in modo condizionato, quando l'immagine generata è guidata da informazioni, come ad esempio un testo descrittivo. Uno dei moduli di condizionamento più popolari per guidare la generazione di immagini a partire da testo è *CLIP* [15] di OpenAI, che è in grado di riconoscere e classificare immagini in base a prompt testuali. Integrando CLIP con i modelli di diffusione, è possibile migliorare la qualità e la coerenza delle immagini generate in applicazioni di text-to-image, come il face aging. Dall'esame della produzione

accademica si può evincere che la ricerca sul face aging ha subito un'evoluzione significativa nel corso del tempo, dando luogo a una pluralità di modelli, ognuno dei quali presenta peculiarità distintive in termini di approccio, architettura e metodologia.

Si ritiene necessario preliminarmente fornire una panoramica riassuntiva dei diversi modelli, presentati in ordine cronologico. Fra i primi si può menzionare *HRFAE*[1] [16], che ha utilizzato un'architettura encoder-decoder per l'editing dell'età dei volti in immagini ad alta risoluzione, mantenendo una qualità visiva elevata. Per risolvere problematiche come la tendenza a generare artefatti visivi e l'instabilità durante l'addestramento, l'architettura è mantenuta il più semplice possibile, utilizzando una singola rete sia per l'invecchiamento che per il ringiovanimento dei volti. Successivamente, *SAM* [9] ha introdotto l'uso di StyleGAN, sfruttando le capacità di controllo stilistico avanzato di questa architettura. Gli approcci più recenti, infine, sono tutti basati su modelli di diffusione, come *FADING*[2] [17], che è stata la prima architettura ad usare un modello di diffusione pre-addrestato, in particolare il Latent Diffusion Model (LDMs) [18], sottoposto al fine-tuning per risolvere il singolo task dell'invecchiamento facciale[3]. Si può inoltre citare *Identity-Preserving Aging of Face Images via Latent Diffusion Models* [19], che ha combinato tecniche di contrastive loss e biometric loss per mantenere l'identità facciale durante la manipolazione dell'età. *CUSP* [20], invece, si distingue dagli altri modelli recenti in quanto non utilizza un diffusion mo-

[1] `https://github.com/InterDigitalInc/HRFAE`

[2] `https://github.com/MunchkinChen/FADING`

[3] Il fine-tuning è una tecnica di machine learning che consiste nell'adattare un modello pre-addestrato su un nuovo dataset specifico, affinando i suoi parametri per migliorare le prestazioni su un compito particolare, senza addestrarlo da zero.

del ma una nuova rete chiamata *Style-based Encoder-decoder*, che si basa su un'architettura encoder-decoder e adotta una strategia focalizzata sulla combinazione delle caratteristiche stilistiche e del contenuto dell'immagine di input, condizionando l'output sull'età target. Inoltre, in fase di inferenza, è possibile scegliere diversi gradi di preservazione della struttura: con un alto grado di preservazione, il modello modifica solo la texture dell'immagine mentre, con un grado di preservazione più basso, è in grado di modificare anche la forma del volto.

2.1 Dataset Sintetico ONOT

ONOT è un dataset sintetico di volti ad alta qualità conforme agli standard ISO/IEC 39794-5 [8], che regolano il formato delle immagini nei documenti di viaggio elettronici (eMRTD). Seguendo le linee guida dell'ICAO, il processo di creazione è volto a produrre immagini ben controllate e precise, con caratteristiche specifiche: posa frontale, sfondo uniforme, illuminazione bilanciata, espressione neutra e assenza di ombre.

La procedura di creazione del dataset ONOT si articola in quattro fasi principali, illustrate nella Figura 2.1.

In primo luogo, le immagini facciali vengono generate utilizzando una versione ottimizzata di Stable Diffusion 1.5, denominata Realistic Vision 5.1, tramite l'interfaccia Stable Diffusion Web UI. Il processo di generazione delle immagini parte dalla definizione di pseudo-classi, che non corrispondono necessariamente a identità reali ma sono utilizzate come rappresentazioni fittizie di individui. In particolare, vengono definite 15.000 pseudo-classi, ciascuna

Figura 2.1: Processo di creazione del dataset ONOT. La generazione delle immagini avviene tramite un modello di diffusione generativa basato su attributi specifici. Le immagini vengono verificate rispetto alla conformità con lo standard ICAO e successivamente sottoposte a controlli di coerenza intra-classe e inter-classe, nei quali vengono analizzate l'uniformità all'interno della stessa identità e le dissimilarità tra identità differenti.

caratterizzata da un seed casuale che contiene informazioni sull'identità virtuale. Per ogni pseudo-classe, vengono generate 64 immagini. Le immagini, con una risoluzione di 512x512 pixel, sono prodotte tramite il campionatore DPM++ SDE Karras in 25 passaggi.

Successivamente, la conformità con gli standard ISO/IEC 39794-5 viene verificata tramite un SDK commerciale [4] che controlla i vincoli della scena (come pose ed espressioni), le proprietà fotografiche (illuminazione, posizionamento e messa a fuoco) e le caratteristiche digitali dell'immagine (risoluzione e dimensioni).

Nel terzo passaggio, si verifica la coerenza intra-classe, ovvero si controlla che le immagini raggruppate in ogni pseudo-classe appartengano alla stessa identità. Infine, la coerenza inter-classe assicura che le pseudo-classi contengano volti abbastanza dissimili tra loro, evitando somiglianze eccessive tra le identità.

La Figura 2.2 illustra la vasta gamma di diversità presente nel

[4] `https://www.correlance.com/cms/en/home`

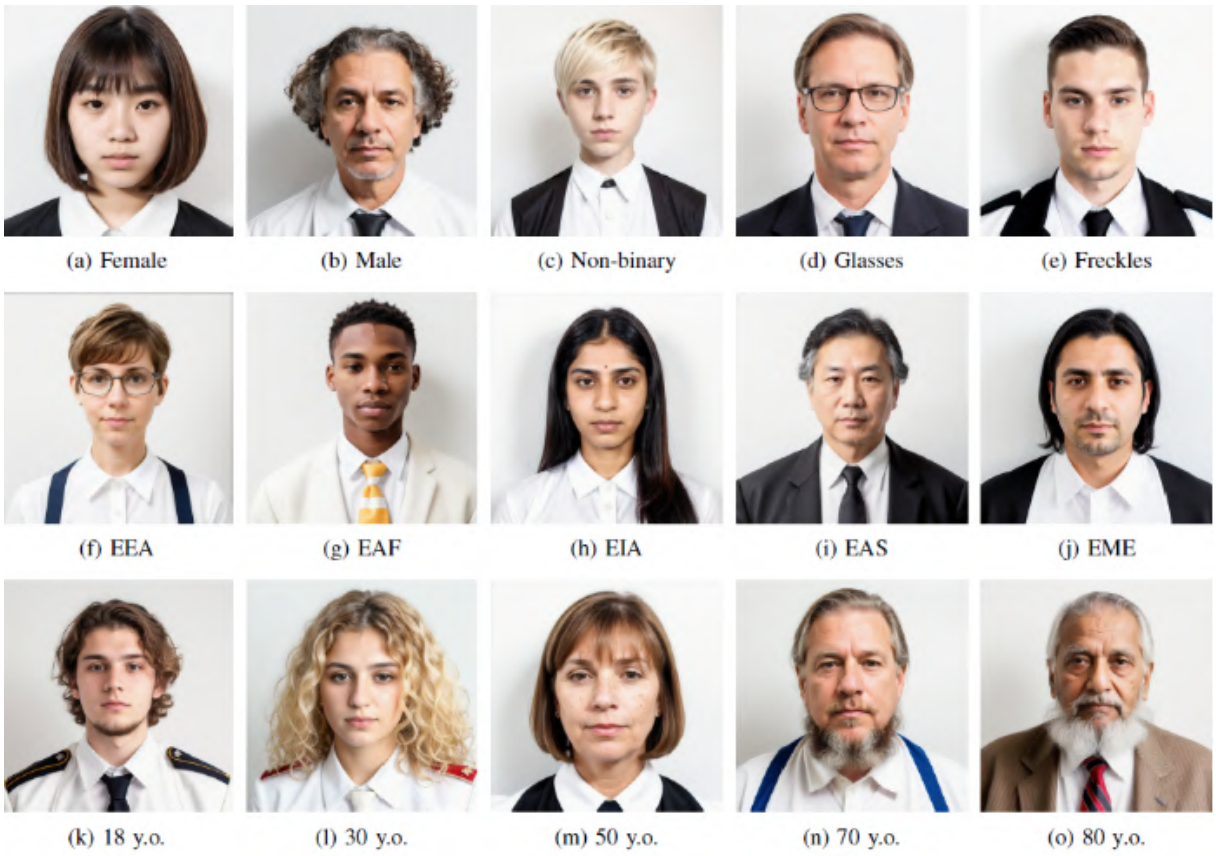


Figura 2.2: Insieme rappresentativo di soggetti inclusi nel dataset ONOT. I volti sono suddivisi in base a diverse caratteristiche: genere (donna, uomo, non binario), presenza di occhiali o lentiggini, e altre variabili etniche.

dataset. ONOT include una varietà di generi, etnie, età e tratti facciali, raggruppati in cinque categorie principali: Europei/Americani (EEA), Africani (EAF), Indiani/Asiatici (EIA), Est Asiatici (EAS) e Medio Orientali (EME). Le immagini mostrano inoltre una notevole differenziazione riguardo a stili, acconciature e accessori, riflettendo la ricchezza visiva del dataset. Ciò garantisce un'ampia copertura di volti e tratti distintivi, fondamentali per l'uso in applicazioni come il riconoscimento facciale e la verifica dell'identità.

Capitolo 3

Ricerca e selezione dei modelli

Dopo aver esaminato la letteratura esistente, è stato avviato un processo di selezione per individuare modelli che siano in grado di generare immagini realistiche di persone invecchiate o ringiovanite, mantenendo al contempo i lineamenti del viso dell'immagine originale. L'obiettivo della presente ricerca consiste dunque nell'identificazione di quei modelli che offrano le migliori prestazioni in termini di realismo dell'output e coerenza con l'immagine di input.

Nel corso di questo lavoro, sono stati testati diversi modelli che si erano dimostrati promettenti nella fase di analisi teorica, allo scopo di individuare quello più efficace per ottenere risultati visivamente realistici e accurati. Tra i modelli esplorati, figurano *HRFAE, SAM, FADING, CUSP, e ID-Preserving-Facial-Aging.* Sono emerse alcune difficoltà durante il percorso: ad esempio, il modello ID-Preserving-Facial-Aging ha evidenziato svariate problematiche tecniche che ne hanno compromesso il corretto

funzionamento. Nonostante i tentativi iniziali di risolvere tali criticità, la complessità della questione ha portato infine alla decisione di non proseguire ulteriormente con questo approccio. Un'esperienza simile si è verificata con CUSP, la cui documentazione, non sempre chiara, ha reso difficoltoso il processo di comprensione e conseguentemente il testing.

SAM è stato relativamente semplice da testare, ma i risultati visivi erano deludenti: le immagini generate, come si può osservare in Figura 3.1, apparivano poco realistiche e con facce a tal punto allungate che i volti in output non sembravano talvolta nemmeno persone, ma forme generiche.

Figura 3.1: Inferenza su SAM. Ogni riga inizia con un'immagine di input (la prima a sinistra) e, attraverso l'inferenza del modello, il volto viene progressivamente invecchiato o ringiovanito, generando una sequenza di volti che coprono un intervallo di età che va dai 18 ai 100 anni. Le immagini a destra dell'input rappresentano versioni modificate della figura originale, con variazioni di età progressive.

Molto probabilmente, questi risultati insoddisfacenti sono dovuti alle limitazioni già evidenziate nel contributo relativo a SAM [9]; in particolare, l'uso di un generatore StyleGAN pre-addestrato, pur semplificando il processo di training e consentendo la generazione di immagini di alta qualità, rende difficile modellare pose del viso che si discostino significativamente da una posizione

frontale standard, come inclinazioni della testa molto accentuate o espressioni complesse e accessori; come si può osservare in Figura 3.2, nella fase di aging il cappello indossato dalla persona nella terza immagine viene completamente stravolto. Inoltre, il metodo utilizzato da SAM è limitato alle immagini che possono essere correttamente rappresentate nello spazio latente di StyleGAN, e ciò può aver contribuito alla scarsa qualità dei risultati prodotti. Anche la difficoltà nel catturare cambiamenti naturali legati all'età e la possibile limitazione del predittore di età per immagini di bambini piccoli potrebbero aver influenzato negativamente la qualità degli output. Ne è derivata la necessità di scartare anche questo modello.

Figura 3.2: Limitazioni di SAM. Il modello potrebbe avere difficoltà quando si trova di fronte a espressioni estreme o ad input fuori dal dominio del dataset di training.

Infine, la mia attenzione si è concentrata su due modelli che si sono dimostrati di maggiore efficacia: HRFAE e FADING. Essi si sono distinti per la loro capacità di generare immagini di alta qualità, rispettando l'identità del soggetto e mostrando un invecchiamento realistico.

3.1 Descrizione dell'architettura di HR-FAE

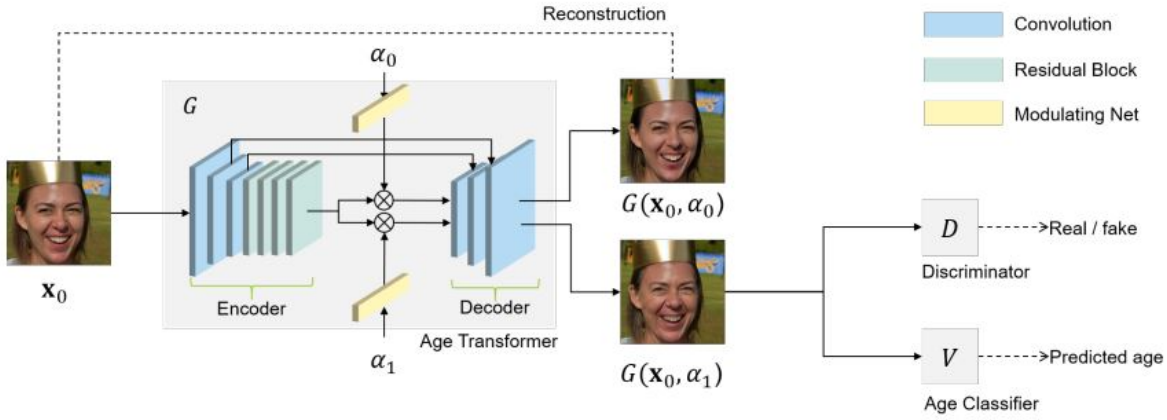


Figura 3.3: L'architettura di HRFAE è costituita da un modello encoder-decoder. L'immagine originale viene codificata, trasformata in base all'età target, e poi decodificata per generare una nuova immagine. Un discriminatore verifica l'autenticità dell'immagine generata, mentre un classificatore stima l'età risultante.

Come si può osservare nella Figura 3.3, HRFAE è costruito su un'architettura di tipo *encoder-decoder*, progettata per modificare l'età del volto in immagini ad alta risoluzione, mantenendo un'alta qualità visiva: l'encoder ha il compito di ridurre l'immagine originale in una rappresentazione compatta, che contiene tutte le informazioni essenziali del volto, escludendo però dettagli specifici come l'età; il decoder utilizza tali informazioni per ricostruire l'immagine, applicando le modifiche necessarie per riflettere l'età desiderata.

Una delle caratteristiche distintive di HRFAE è l'uso di un *feature modulation layer*, componente che agisce direttamente sullo *spazio latente*[1], permettendo transizioni di età continue e fluide,

[1]Lo spazio latente è un concetto fondamentale nel machine learning e si

senza la necessità di reti separate per invecchiare e ringiovanire. In pratica, lo stesso modello può essere utilizzato sia per aumentare l'età di un volto, sia per ridurla, grazie alla capacità di mantenere intatte le informazioni cruciali come l'identità e l'espressione del volto.

3.1.1 Fase di Training

Il modello HRFAE è stato addestrato utilizzando il dataset FFHQ ad alta risoluzione [21], che contiene 70.000 immagini facciali a 1024×1024 pixel, con una varietà di età, etnie, pose, illuminazione e sfondi. Tuttavia, poiché il dataset originale non include etichette di età, è stato utilizzato il classificatore di età DEX, preaddestrato sul dataset IMDB-WIKI [22], per assegnare etichette di età alle immagini. Durante l'addestramento, è stato notato un forte sbilanciamento nel dataset a favore di volti giovani rispetto a quelli anziani. Questo squilibrio può rappresentare un problema, poiché il modello, così trainato, tenderebbe a concentrarsi più sull'invecchiamento che sul ringiovanimento, portando a risultati insoddisfacenti per quest'ultimo. Per compensare tale disparità, è stato utilizzato il modello StyleGAN, generando 300.000 immagini sintetiche per riequilibrare la distribuzione delle età. Il dataset di training utilizzato per l'addestramento contiene volti appartenenti alla fascia d'età compresa tra i 20 e i 69 anni. Di conseguenza, il processo di invecchiamento facciale viene eseguito esclusivamente

riferisce a uno spazio astratto e multidimensionale in cui vengono rappresentate le caratteristiche dei dati in modo compresso e spesso non direttamente interpretabile. Esso contiene variabili latenti che non sono osservabili direttamente, ma che catturano i pattern o le strutture fondamentali dei dati. L'idea è che le variazioni complesse presenti nei dati in oggetto possano essere ridotte a rappresentazioni più semplici e gestibili, facilitando la comprensione e l'elaborazione dei dati da parte dei modelli.

all'interno di questo intervallo di età, limitando la capacità del modello di generare volti per età inferiori ai 20 anni o superiori ai 69 anni. Un elemento chiave della fase di training di HRFAE è il *discriminatore*, che generalmente rientra nel contesto delle *GAN (Generative Adversarial Networks)*, costituendone una componente essenziale. Nei modelli tradizionali di GAN, il discriminatore ha il compito di distinguere tra immagini reali e immagini generate, cercando di far coincidere la distribuzione delle immagini sintetiche con quella delle immagini reali. In questo contesto, il discriminatore è spesso condizionato su specifiche caratteristiche, come l'età target delle immagini generate. Nel caso di HRFAE, tuttavia, il discriminatore utizzato si differenzia da quello presente nei modelli tradizionali di face aging; esso infatti non è condizionato dall'età target ma possiede come unico obiettivo assicurarsi che l'immagine sembri fotorealistica, ovvero che appaia il più possibile naturale e priva di difetti o artefatti visivi. Tale approccio permette al modello di concentrarsi sulla qualità visiva complessiva, riducendo così la possibilità che l'immagine invecchiata presenti distorsioni o imperfezioni dovute a vincoli eccessivi imposti dal discriminatore. Inoltre, per evitare artefatti significativi, è stato scelto un intervallo di età minimo tra l'età di partenza e quella target; ciò ha ulteriormente aiutato il discriminatore a ridurre gli artefatti nelle trasformazioni che comportano grandi cambiamenti di età. Il modello è stato addestrato per 20 epoche: le prime 10 epoche su immagini a 512×512 pixel con un batch size di 4, e le successive 10 epoche su immagini a 1024×1024 pixel, riducendo il batch size a 2 e abbassando il learning rate da 10^{-4} a 10^{-5}.

3.2 Descrizione dell'architettura di FADING

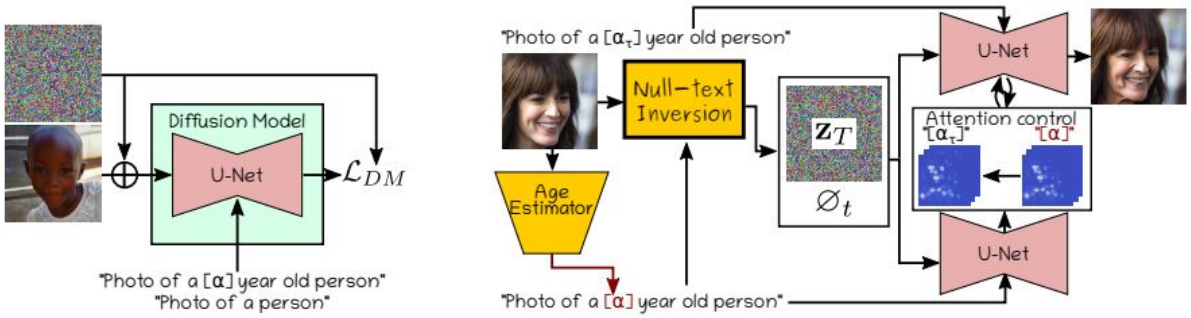


Figura 3.4: Architettura del modello FADING. A sinistra viene eseguito il fine-tuning di un modello Stable Diffusion pre-addestrato [13, 14] per specializzarlo nel compito di invecchiamento; invece a destra, dato un volto in input, il processo di diffusione viene invertito e l'immagine viene modificata sostituendo l'età stimata con l'età target

L'architettura di FADING è progettata per modificare l'età di una persona in un'immagine articolando il processo in due fasi principali: specializzazione ed editing.

Nella fase di specializzazione, si parte da un modello *text-to-image Diffusion Model* pre-addestrato chiamato Latent Diffusion Model (LDMs) [18], modello generativo avanzato che crea immagini a partire da descrizioni testuali[2]. Esso opera nello spazio latente — rappresentazione compressa dell'immagine in cui è più facile applicare trasformazioni complesse — come l'invecchiamento del volto. In questa fase, il modello viene adattato per il compito specifico di modificare l'età, utilizzando coppie di immagini di volti e descrizioni testuali che specificano l'età della persona, come "Foto di una persona di [α] anni". Questo processo di fine-tuning, che

[2]Per ulteriori dettagli tecnici sui Diffusion Dodels, si rimanda all'Appendice 6.5

consiste nell'ulteriore addestramento di un modello pre-addestrato su un nuovo set di dati specifico, permette di separare meglio le informazioni relative all'età da altre caratteristiche non pertinenti, come l'identità o lo sfondo dell'immagine.

La fase di editing avviene in due passaggi: *inversion* e *Localized Age Editing with Attention Control*. Nell'inversione si utilizza un algoritmo chiamato *null-text inversion*, che permette di invertire il processo di diffusione dell'immagine originale nello spazio latente, dove il modello può meglio manipolare l'immagine. Tale procedimento è necessario: nel processo di inversione, infatti l'immagine in input deve essere riportata a uno stato di rumore in modo che il modello di diffusione (DM) possa manipolarla piú efficacemente; sarà quindi possibile, in un momento successivo, generare l'immagine invecchiata o ringiovanita a partire dal rumore ottenuto. A tal fine, viene usata un'architettura chiamata *DDIM (Denoising Diffusion Implicit Models)*, che consente di invertire l'immagine generata fino a raggiungere lo spazio di rumore di partenza.

Durante questa fase di ritorno al rumore si accumulano errori, compromettendo la fedeltà dell'immagine ricostruita. Per risolvere questo problema, viene utilizzato l'algoritmo *null-text inversion* [23], che ottimizza il *null-text embedding*, ovvero l'embedding associato alla stringa vuota, utilizzata durante la generazione dell'immagine per fare previsioni generiche, non influenzate da nessun prompt testuale specifico. Il processo di ottimizzazione avviene progressivamente, e l'embedding ottimizzato funge da "guida invisibile", aiutando il modello a mantenere la fedeltà all'immagine originale nel corso del processo di inversione. Nel secondo passaggio, si applica l'editing localizzato dell'età utilizzando le

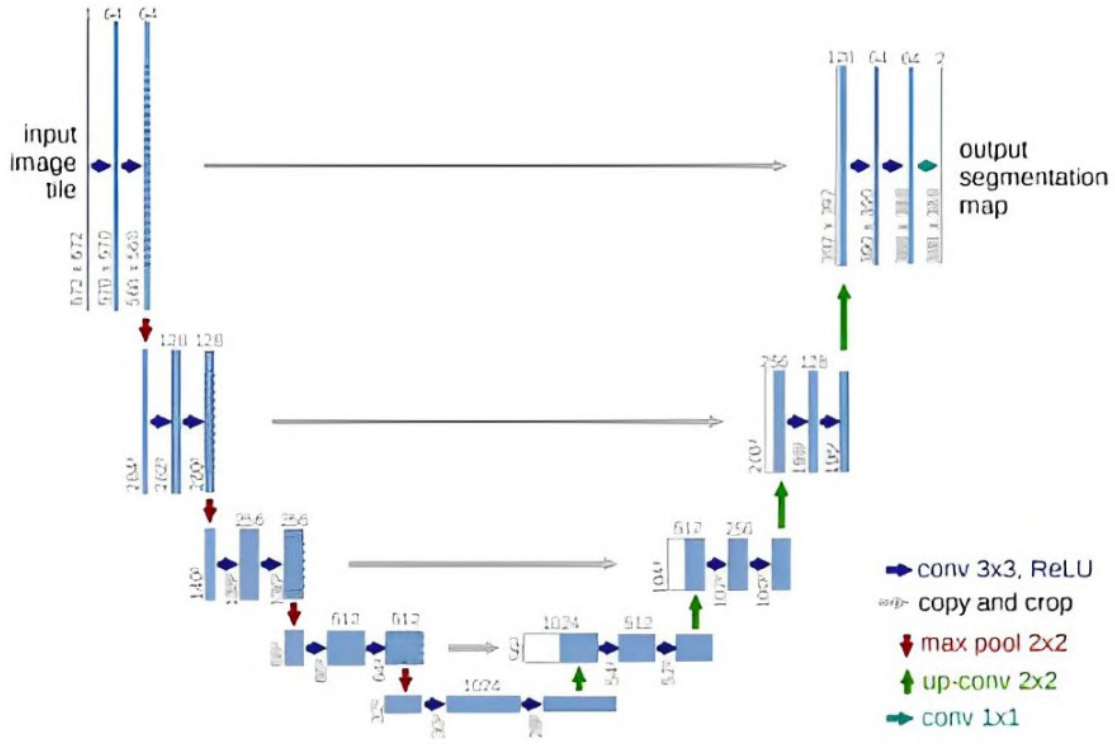


Figura 3.5: Architettura U-Net, i box blu rappresentano le feature maps a più canali, mentre i box bianchi rappresentano le feature maps copiate. Le frecce di colori diversi rappresentano diverse operazioni.

cross-attention maps, rappresentazioni che indicano quali parti dell'immagine sono più rilevanti rispetto a determinate parole nel prompt testuale, come l'età specificata. In sintesi, queste mappe collegano elementi testuali (ad esempio, "50 anni") con le aree corrispondenti dell'immagine, come le rughe o il colore dei capelli. Durante l'editing, il modello usa queste mappe per modificare solo le aree rilevanti all'età, mantenendo intatte le altre parti dell'immagine.

Nell'architettura mostrata nella Figura 3.4, si puó notare che questa fase di attenzione e modifica utilizza due reti neurali chiamate *U-Net* [24] con lo scopo di applicare in modo preciso modifiche localizzate, basate sulle cross-attention maps.

Esaminando nello specifico la configurazione della rete neurale U-net [24], si osserva che essa rappresenta un'architettura ampiamente utilizzata per la segmentazione semantica delle immagini, particolarmente nel campo della visione artificiale applicata alla medicina. La struttura di U-net è costituita da due parti principali: un percorso di contrazione (*downsampling*) e un percorso di espansione (*upsampling*), come si può evincere dalla Figura 3.5.

Nella parte sinistra dell'immagine, si trova il percorso di contrazione, che segue la tipica struttura di una rete convoluzionale. Esso è caratterizzato da una serie di operazioni ripetute: due convoluzioni 3x3 (senza padding), seguite da una funzione di attivazione ReLU, e un'operazione di max pooling 2x2 con stride 2 per ridurre la dimensione spaziale. Ad ogni passo di downsampling, il numero di canali di feature raddoppia, come indicato dai numeri sopra i blocchi, passando, ad esempio, da 64 a 128. Questo processo di contrazione consente alla rete di catturare le feature globali dell'immagine.

Nella parte destra dell'immagine si trova invece il percorso di espansione, che serve a recuperare le informazioni spaziali perse durante il downsampling. Ogni step presente in questa sezione consiste in un upsampling della mappa delle feature — cioè la rappresentazione intermedia di un'immagine generata da una rete neurale convoluzionale (CNN) durante l'elaborazione dei dati di input — seguito da una convoluzione 2x2 (detta "up-convolution") che dimezza il numero di canali di feature, come si puó notare dalle frecce verdi. In seguito, vengono applicate altre due convoluzioni 3x3, seguite da funzioni di attivazione ReLU.

Nello stato finale, viene utilizzata una convoluzione 1x1 per

mappare ogni vettore di feature, composto da 64 componenti, al numero desiderato di classi, dando luogo in tal modo alla mappa di segmentazione finale. Le frecce grigie rappresentano invece le *skip connections* che collegano direttamente il layer specifico di downsampling con quello di upsampling; tali connessioni servono per trasferire informazioni importanti ai layer di upsampling, che potrebbero invece risultare perse durante la fase di downsampling. In totale, la rete U-Net ha 23 strati convoluzionali. Questa complessa struttura consente alla rete di combinare efficacemente il contesto globale catturato durante la fase di contrazione con i dettagli locali ripristinati nella fase di espansione, risultando molto efficace nella segmentazione di immagini complesse, come ad esempio quelle utilizzate in ambito medico.

3.2.1 Fase di Training

Nella fase di training sono state selezionate 150 immagini dal dataset FFHQ-Aging [21] per eseguire il fine-tuning del modello di diffusione pre-addestrato, per un totale di 150 passi con un batch size di 2. Il processo di ottimizzazione — tecnica stocastica utilizzata frequentemente nel deep learning per aggiornare iterativamente i pesi del modello in base ai gradienti calcolati — è stato gestito tramite *l'ottimizzatore Adam* [25]. Gli ottimizzatori, in generale, hanno il compito di minimizzare la loss function del modello, migliorando così l'accuratezza delle sue predizioni. L'ottimizzatore Adam, nello specifico, integra i benefici del metodo di momentum e dell'Adaptive Learning Rate, risultando particolarmente adatto a modelli complessi e dati rumorosi. Inoltre per il fine-tuning è stato utilizzato un learning rate di 5×10^{-6} e

parametri $\beta_1 = 0.9$ e $\beta_2 = 0.999$, che regolano rispettivamente il decrescere del rate di apprendimento e il controllo della varianza. L'intero processo è stato eseguito su una singola GPU A100, richiedendo un minuto per il fine-tuning, un minuto per l'inversione e solo 5 secondi per la modifica dell'età.

Capitolo 4

Utilizzo e valutazione dei modelli

L'impiego dei modelli FADING e HRFAE ha permesso la generazione di immagini invecchiate partendo da volti con età variabile. Questi modelli sono stati utilizzati per produrre una serie di immagini in output che rappresentano lo stesso soggetto in diverse fasi della vita, da 18 a 100 anni, con intervalli di 10 anni. Il processo sfrutta le capacità dei modelli di simulare con precisione l'invecchiamento, mantenendo le caratteristiche distintive del volto originale, ma adattandole alle diverse età richieste. Tale approccio consente di visualizzare in modo realistico l'evoluzione di un volto nel tempo, indipendentemente dall'età iniziale dell'immagine in input. I risultati ottenuti sono stati inizialmente sottoposti a un'analisi qualitativa, mirata a valutare visivamente la coerenza dei volti generati rispetto ai criteri di invecchiamento desiderati, così da verificare l'assenza di anomalie evidenti o deformità rispetto alle aspettative. Solo successivamente, si è passati alla valutazione quantitativa attraverso due analisi comunemente adottate in que-

sto ambito. Per quanto riguarda la prima, si sono utilizzati degli *age estimator*, come *DeepFace* [26] e *DEX* [27], modelli di deep learning addestrati per stimare l'età di una persona basandosi su un'immagine del volto. Essi forniscono una stima numerica dell'età del soggetto nell'immagine, permettendo di valutare l'accuratezza del face aging generato dai modelli, verificando quindi se l'età target è stata effettivamente raggiunta. In particolare, è stato calcolato l'errore assoluto tra l'età target specificata per la generazione dell'immagine e quella stimata dall'age estimator. Ad esempio, se il modello ha generato l'immagine di un volto di 20 anni e l'age estimator rileva un'età di 18 anni, l'errore sarà di 2 anni. Per quanto riguarda la seconda analisi, sono stati utilizzati AdaFace [28] e ArcFace [29], due modelli avanzati di riconoscimento facciale noti per la loro elevata accuratezza nell'estrazione delle caratteristiche facciali. ArcFace sfrutta una loss function angolare per migliorare la discriminabilità delle rappresentazioni facciali, mentre AdaFace si adatta dinamicamente a diverse condizioni di input, come bassa risoluzione, cambiamenti di illuminazione, variazioni di posa o espressione facciale, risultando quindi particolarmente robusto anche in situazioni difficili. Inizialmente, i volti sono stati allineati per assicurare che tutte le immagini fossero nella stessa posa, facilitando, in tal modo, un confronto accurato. In un secondo momento, è stato estratto l'embedding dei volti — cioè una rappresentazione vettoriale delle feature dell'immagine (vedi Appendice 6.3 per ulteriori dettagli) — utilizzando AdaFace e ArcFace. Infine, è stata calcolata la distanza coseno tra questi embedding, un passaggio cruciale per valutare la dissimilarità tra i volti. La distanza coseno, infatti, misura tale dissimilarità

analizzando l'angolo tra due vettori nello spazio multidimensionale. A differenza di altre metriche di distanza, come la distanza euclidea, che considerano le differenze nelle grandezze dei vettori, la distanza coseno si concentra esclusivamente sulla direzione, risultando particolarmente utile quando l'ampiezza dei vettori è irrilevante per l'analisi. Dal punto di vista geometrico, la distanza coseno deriva dal coseno dell'angolo compreso tra i due vettori, il quale fornisce una misura della loro similarità. Se l'angolo tra i vettori è pari a zero (ossia i vettori sono orientati nella stessa direzione), il coseno sarà 1, mentre per vettori ortogonali il coseno sarà 0. La distanza coseno può essere formalmente definita come:

$$D_C(\mathbf{A}, \mathbf{B}) = 1 - \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

dove:

- $\mathbf{A} \cdot \mathbf{B}$ rappresenta il prodotto scalare tra i vettori $\mathbf{A}$ e $\mathbf{B}$
- $\|\mathbf{A}\|$ denota la norma del vettore $\mathbf{A}$
- $\|\mathbf{B}\|$ denota la norma del vettore $\mathbf{B}$

Il valore risultante, che varia tra 0 e 2, riflette il grado di dissimilarità tra i vettori: valori vicini a 0 indicano una forte somiglianza direzionale, mentre valori più elevati segnalano una maggiore divergenza.

Questa distanza è stata utilizzata per analizzare la somiglianza tra l'immagine in input e le varie immagini in output, al fine di verificare se l'immagine generata rappresenti la stessa persona dell'immagine originale. Inoltre, è stata calcolata la distanza coseno tra le immagini generate a intervalli di 10 anni (ad esempio,

tra l'immagine a 10 anni e quella a 20 anni, e così via fino a 100 anni). Ciò ha permesso di verificare se il processo di invecchiamento, col progredire dell'età, sia stato eseguito in modo ottimale, mantenendo la coerenza visiva e l'identità del volto lungo tutto l'arco temporale simulato.

La metodologia adottata e le scelte tecniche compiute sono illustrate in due sezioni distinte, ciascuna dedicata a uno dei modelli utilizzati, seguite dall'esame dettagliato dei risultati ottenuti attraverso le suddette metriche. In primo luogo, è stata condotta un'analisi qualitativa, seguita dalle analisi quantitative, al fine di fornire una valutazione completa dei risultati.

4.1 Confronto dei modelli dal punto di vista qualitativo

L'analisi qualitativa è stata condotta confrontando i tre modelli di face aging a cui è stata applicata inferenza: SAM, FADING e HRFAE. L'obiettivo di questa analisi è valutare la capacità di ciascun modello di generare volti invecchiati mantenendo integra l'identità del soggetto, evitando di distorcere eccessivamente i lineamenti originali. Nella Figura 4.1 è presente un confronto visivo tra i tre modelli: la prima immagine di ogni riga rappresenta il volto originale dato in input mentre, progredendo verso destra, vi sono le versioni invecchiate generate progressivamente dal modello corrispondente.

La prima riga della Figura 4.1 presenta i risultati ottenuti dal modello SAM. Come si può notare, SAM fallisce nell'esecuzione di un invecchiamento realistico, distorcendo pesantemente i

Figura 4.1: L'immagine mostra un confronto visivo tra tre diversi modelli di invecchiamento facciale.

lineamenti del viso. L'allungamento sproporzionato del volto è particolarmente evidente già nelle prime fasi di invecchiamento, con un effetto che peggiora progressivamente facendo assumere al soggetto un aspetto teratomorfo. Ciò porta a una perdita quasi totale delle caratteristiche identificative del viso originale, che appare deformato piuttosto che invecchiato, con un risultato che va ben oltre un normale processo di aging, dando l'impressione di un'alterazione artificiale piuttosto che di un naturale processo di decadimento fisico. Questo evidente fallimento qualitativo ha portato alla decisione di non proseguire con analisi quantitative, poiché sarebbe stato poco produttivo, dato che il modello non riesce a fornire una base accettabile per ulteriori valutazioni.

La seconda riga della Figura 4.1 rappresenta i volti generati dal modello HRFAE. Sebbene i risultati siano decisamente migliori rispetto a SAM, HRFAE presenta comunque delle evidenti criticità.Come già menzionato nel capitolo 3.1.1, questo modello fatica a generare volti per età estreme, in particolare sotto i 20 anni e sopra i 70. La difficoltà di HRFAE a generalizzare al di fuori del suddetto intervallo compromette quindi la qualità dei risultati all' estremità del range di età. Tuttavia, all'interno di esso, i volti

generati appaiono in qualche modo accettabili, mantenendo una somiglianza discreta con l'immagine di partenza, pur con delle imperfezioni nella transizione.

La terza riga della Figura 4.1 mostra i risultati ottenuti con il modello FADING, che rappresenta il miglior equilibrio tra realismo e accuratezza nel preservare l'identità del soggetto. Il processo di invecchiamento applicato da FADING, a differenza degli altri due modelli, mantiene intatti i lineamenti fondamentali del viso, dando luogo a una trasformazione più dettagliata e affidabile. Le transizioni tra età successive risultano molto più naturali rispetto agli altri modelli, e anche per le età più avanzate (oltre i 70 anni) i tratti del volto non vengono distorti in maniera eccessiva. Ciò suggerisce che FADING riesca a preservare, seppur con qualche difficoltà per età molto estreme, una coerenza accettabile nella rappresentazione del volto invecchiato. È importate poi rilevare la continuità delle immagini generate da FADING: la coerenza visiva lungo il processo di invecchiamento è molto più fluida rispetto a SAM e HRFAE; mentre gli altri modelli tendono a presentare cambiamenti bruschi o irrealistici, FADING riesce a mantenere un'armonia nel passaggio da un'età all'altra, anche se, per età estremamente avanzate si può osservare una lieve perdita di precisione nei dettagli del volto. Ciò può essere comunque considerato accettabile, dato che le immagini di una fascia d'età estrema sono meno presenti nei dataset di training.

4.2 Utilizzo del modello FADING

Per generare le immagini invecchiate utilizzando il modello FADING è stato utilizzato uno script Python che automatizza l'intero processo di trasformazione del viso originario[1]. Nel processo di generazione delle immagini in output, è stato necessario fornire come input sia l'età effettiva del soggetto, sia il genere, affinché il modello potesse applicare correttamente l'invecchiamento. Durante la valutazione dei parametri di input, sono state fatte scelte progettuali specifiche. Per la previsione dell'età dell'immagine di input si è utilizzato il modello DEX, selezionato per la sua capacità di fornire stime accurate dell'età. Allo stesso modo, per la determinazione del genere, è stato scelto DeepFace, modello che ha rivelato buone performance in questo specifico task. La combinazione dei due strumenti ha permesso di ottenere un processo di invecchiamento delle immagini accurato e affidabile.

4.2.1 Predizione dell'età e calcolo dell'errore

Inizialmente, per la predizione dell'età all'interno del modello FADING, è stato impiegato il classificatore dell'età DeepFace. Tuttavia, analizzando la Figura 4.2 si nota che vi è una concentrazione elevata di predizioni nelle fasce d'età tra i 20 e i 30 anni, con una significativa sovrarappresentazione dell'età suddette. Si ipotizza quindi che DeepFace tenda a sottostimare l'età per i soggetti più anziani, attribuendo età inferiori rispetto a quelle

[1]Per ulteriori dettagli tecnici sull'utilizzo di FADING, si rimanda all'Appendice 4.2

reali. Il modello presenta quindi un evidente bias verso le fasce d'età più giovani.

Age Range	**Average Error**
Overall	22.907
10-19	3.718
20-29	2.800
30-39	7.970
40-49	15.301
50-59	22.624
60-69	29.671
70-79	34.622
80-89	41.645
90-99	42.576

Tabella 4.1: Errore assoluto medio per ogni fascia di età di immagini generate con FADING usando il classificatore dell'età DeepFace.

L'elevato errore medio complessivo di 22.9 osservabile nella Tabella 4.1 è un chiaro indicatore della difficoltà del modello nel generalizzare su tutto lo spettro di età compreso tra 18 e 100 anni. In particolare, come risulta evidente dalla Figura 4.2, per le fasce d'età più avanzate (50 anni e oltre), si registra un incremento notevole dell'errore, con valori che raggiungono anche 42 anni di scarto per la fascia 90-99 anni.

Considerata questa evidente limitazione di DeepFace, si è deciso di adottare un altro modello per la stima dell'età, DEX, che ha dimostrato di essere più affidabile, robusto e accurato. L'utilizzo di DEX ha portato a una distribuzione delle predizioni più omogenea su tutto il range di età da 18 a 100 anni, come si può osservare in Figura 4.3. Analizzando la Tabella 4.2 si può notare una riduzione significativa dell'errore medio totale, che è sceso a 5.7 anni, valore molto più accettabile rispetto a quello ottenuto con DeepFace.

Risulta inoltre interessante e particolarmente utile considerare

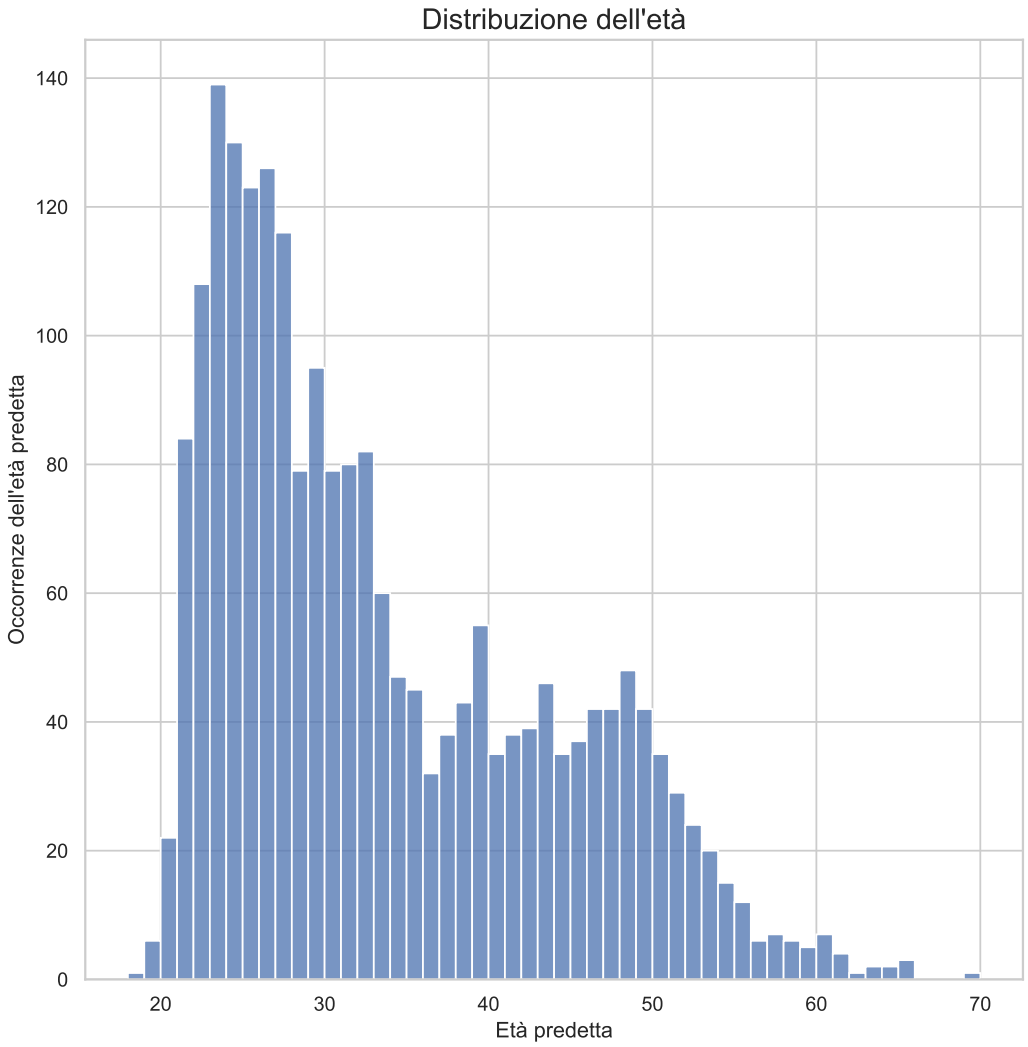


Figura 4.2: L'immagine mostra la distribuzione delle età stimate utilizzando il modello DeepFace come predittore dell'età all'interno del modello FADING. Le occorrenze predette si concentrano maggiormente nelle fasce giovanili, con un picco intorno ai 22 anni. Questo squilibrio nella distribuzione delle predizioni porta a un errore medio totale elevato, pari a 22.9 anni, con errori specifici particolarmente pronunciati nelle fasce d'età più avanzate.

un caso ipotetico, in cui sia il classificatore dell'età sia il modello di face aging siano perfettamente accurati. In questa condizione ideale la distribuzione delle predizioni risulta omogenea e accurata per tutte le fasce d'età. Ogni soggetto verrebbe classificato esattamente nella sua fascia anagrafica di appartenenza, senza alcun tipo di bias verso differenti gruppi di età. Questa rappresentazione ideale è un utile punto di riferimento per evidenziare quanto i modelli reali si discostino dalla perfezione. Mentre nel

Age Range	Average Error
Overall	5.740
10-19	3.015
20-29	3.790
30-39	4.014
40-49	4.334
50-59	4.236
60-69	6.209
70-79	6.564
80-89	9.714
90-99	12.655

Tabella 4.2: Errore assoluto medio per ogni fascia di età di immagini generate con FADING usando il classificatore dell'età DEX.

caso ipotetico si potrebbe riscontrare un errore medio pari a zero, le deviazioni osservate con DeepFace e DEX dimostrano i limiti degli approcci tradizionali utilizzati nella predizione dell'età.

4.2.2 Analisi della similarità dei volti

Per effettuare l'analisi della similarità dei volti generati dal modello FADING, si è proceduto in primo luogo all'allineamento delle immagini. Questa fase di pre-processing è stata effettuata per garantire che tutte le immagini prodotte avessero la medesima posa, facilitando così in condizioni uniformi il confronto successivo tra volti. Tale uniformità è risultata cruciale per ottenere misurazioni di similarità più accurate e attendibili.

Successivamente, sono stati impiegati i modelli ArcFace e AdaFace per estrarre dai volti generati gli embedding che costituiscono la base per il calcolo della similarità. La misurazione relativa è stata ottenuta attraverso la distanza coseno tra l'immagine di input e i vari prodotti generati dal modello di invecchiamento

facciale. In particolare, è stata calcolata la distanza tra l'immagine di input e ciascuna delle immagini a cui è stata applicata la trasformazione di aging (ad esempio, tra l'originale e il suo corrispettivo a 20 anni, a 30 anni, e così via). Osservando la Figura 4.4, è possibile notare un comportamento caratteristico nella variazione della distanza coseno tra l'immagine di input e quelle generate attraverso il modello di invecchiamento facciale. Nel caso specifico di un volto con età reale pari a 40 anni, si osserva che la distanza coseno tra l'immagine di input e le versioni invecchiate tende a diminuire progressivamente all'avvicinarsi della fascia d'età corrispondente (da 18 anni fino ai 40). Tale riduzione indica un aumento della somiglianza tra la figura originale e le trasformazioni progressivamente più vicine all'età reale del soggetto. Superata la soglia dei 40 anni, il trend si inverte: la distanza coseno inizia nuovamente ad aumentare, descrivendo un andamento a forma di "V". Questo comportamento è atteso, in quanto le immagini trasformate a età superiori si discostano progressivamente dall'età anagrafica del soggetto, determinando un incremento della dissimilarità rispetto alla figura originaria. La forma a "V" della curva evidenzia chiaramente che l'età reale di 40 anni rappresenta il punto di minima distanza coseno, a cui corrisponde la massima somiglianza tra l'immagine originale e le versioni invecchiate.

In aggiunta, è stata eseguita una seconda analisi dei prodotti generati, in cui si è calcolata la distanza coseno tra immagini corrispondenti a intervalli di 10 anni di età. Ad esempio, è stata misurata la distanza tra l'immagine generata corrispondente a 18 anni e quella a 28 anni, proseguendo con l'intervallo 28-38 e così via.

I risultati ottenuti, rappresentati graficamente nella Figura 4.5, evidenziano un comportamento significativo in riferimento alle fasce di età più giovani. In particolare, nelle coppie dei seguenti intervalli anagrafici (10-19)-(20-29) e (20-29)-(30-39), si osservano distanze coseno maggiori rispetto a quelle riscontrate per fasce di età più avanzate. Questo risultato è atteso perchè coerente con la naturale evoluzione dei tratti somatici. Durante le prime fasi della vita, ad esempio tra i 15 e i 25 anni, i cambiamenti nei lineamenti facciali sono infatti più marcati, pronunciati e rapidi rispetto alle trasformazioni che si verificano in età più avanzata, come tra i 65 e i 75 anni. Con l'avanzare dell'età, la progressione dei cambiamenti somatici tende a rallentare, determinando una minore variazione nella somiglianza facciale e, conseguentemente, distanze coseno più basse. Confrontando invece i risultati ottenuti con AdaFace e ArcFace nei vari grafici, si nota una differenza significativa nella distribuzione delle distanze coseno lungo le diverse fasce d'età. In particolare, AdaFace presenta meno picchi pronunciati, e variazioni di ampiezza minore rispetto ad ArcFace. Questo comportamento può essere attribuito alla capacità di AdaFace di adattarsi dinamicamente a diverse condizioni di input, come menzionato all'inizio del capitolo, suggerendo una maggiore stabilità del modello rispetto alle variazioni legate all'età.

4.3 Utilizzo del modello HRFAE

Il secondo modello impiegato è HRFAE, che offre un approccio differente per la generazione delle immagini invecchiate. L'utilizzo di HRFAE si è rivelato particolarmente agevole grazie alla sua

prerogativa di gestire autonomamente i parametri di età e genere, rendendo il processo di invecchiamento delle immagini non solo più rapido ma anche più accessibile[2].

4.3.1 Predizione dell'età e calcolo dell'errore

Nel modello HRFAE, passando dall'utilizzo di DeepFace a DEX, si è riscontrato un decremento dell'errore totale come già si è evidenziato nel caso di FADING (si veda a riguardo il capitolo 4.2.1), quando sono stati esaminati i benefici derivanti dall'adozione di un diverso classificatore dell'età. Tuttavia, sebbene l'errore complessivo sia diminuito da 31 a 24.99, come riportato nella Tabella 4.7, resta evidente che rimane ancora significativamente alto. Si può osservare anche che la distribuzione dell'età è leggermente più omogenea, come si può notare in Figura 4.8

Un'analisi più approfondita ha mostrato che questo errore elevato è principalmente dovuto alla difficoltà del modello HRFAE di generare correttamente l'aspetto di un soggetto quando l'età target assegnata è inferiore ai 20 anni o superiore ai 70 anni. Tale limite è chiaramente osservabile nella Figura 4.8, dove si nota come l'errore di predizione tenda a crescere in modo considerevole per questi due estremi. La ragione principale di tale difficoltà risiede nel fatto che il modello è stato allenato su un dataset limitato, comprendente esclusivamente immagini di individui che rientrano nella fascia d'età compresa tra i 20 e i 70 anni, come è stato già accennato nel capitolo 3.1.1. Di conseguenza, la capacità del modello di generalizzare al di fuori di questo intervallo risulta

[2]Per ulteriori dettagli tecnici sull'utilizzo di HRFAE, si rimanda all'Appendice 4.3

compromessa. A supporto di questa affermazione, è utile fare riferimento al file di configurazione del training nel modello HRFAE presente in Appendice 6.4, in cui si può notare chiaramente la definizione dei parametri di input, tra cui la specifica dei limiti dell'età minima e massima utilizzata durante il training, fissati rispettivamente a 20 e 70 anni. In conclusione, sebbene il passaggio da DeepFace a DEX abbia ridotto l'errore complessivo, restano delle criticità legate alla difficoltà di generazione di soggetti molto giovani o molto anziani, a causa della distribuzione non uniforme delle immagini utilizzate nel processo di training.

4.3.2 Analisi della similarità dei volti

L'analisi della similarità dei volti nel modello HRFAE è stata condotta seguendo una procedura analoga a quella descritta per FADING (si veda a riguardo il capitolo 4.2.2). Anche in questo caso l'allineamento delle immagini è stato eseguito per garantire una posa univoca e facilitare il confronto tra i volti generati e quelli originali.

Per estrarre gli embedding dai volti prodotti dal modello di invecchiamento sono stati nuovamente utilizzati ArcFace e AdaFace. La misurazione della similarità è stata calcolata tramite la distanza coseno tra l'immagine di input e le immagini generate, con l'obiettivo di valutare quanto l'identità del soggetto fosse preservata nel processo di invecchiamento.. Osservando la Figura 4.9 si può notare che HRFAE clusterizza il processo di aging in vari gruppi di età. In particolare, si possono distinguere tre macro-categorie: Giovani (10-19 anni), Mezza età (30-59 anni), Anziani (70-99 anni).

Le immagini originali appartenenti alle fasce d'età 10-19 mostrano distanze coseno relativamente basse e poco variabili, indipendentemente dall'età dell'immagine in input. Questo comportamento indica come HRFAE non riesca a applicare il process di aging a tale fascia di età.

Riguardo alla mezza età si osserva che le distanze coseno risultano generalmente più alte rispetto alle fasce più giovani, ma la variabilità interna è ridotta, suggerendo una certa uniformità nel modo in cui HRFAE gestisce i soggetti maturi. Nonostante le distanze aumentino lievemente, l'analisi evidenzia come il modello tenda a trattare le trasformazioni invecchiate per questa fascia anagrafica in modo omogeneo, senza marcate differenze nei tratti somatici.

Infine, nel caso degli anziani, si osserva nuovamente una riduzione delle distanze coseno, come è avvenuto per i volti giovani. Ciò indica che, per i soggetti di età avanzata, il modello tende a produrre immagini invecchiate con variazioni minime rispetto all'originale, evidenziando una maggiore omogeneità e una minore capacità di rappresentare differenze somatiche rilevanti tra soggetti attempati.

In sintesi, HRFAE mostra una tendenza a raggruppare i soggetti in tre grandi categorie anagrafiche: giovani, individui maturi e anziani. All'interno di ciascun gruppo, le variazioni di distanza coseno sono contenute e uniformi, mentre differenze più significative emergono tra le fasce, evidenziando come il modello percepisca le differenze somatiche in modo più pronunciato tra questi tre intervalli d'età, rispetto a quanto accade all'interno di ciascuna fascia.

Le analisi includono anche la misurazione della distanza coseno tra immagini corrispondenti a età distanziate di 10 anni, al fine di osservare l'evoluzione della similarità facciale nel tempo. Analizzando la Figura 4.10 che rappresenta i risultati ottenuti, si osserva un fenomeno che suggerisce nuovamente una suddivisione in cluster omogenei. Le distanze coseno risultano elevate principalmente tra le seguenti coppie di fasce d'età: (10-19, 20-29), (30-39, 40-49) e (60-69, 70-79). Questo comportamento indica che le transizioni tra i vari cluster mostrano distanze coseno più alte, segnalando che il modello modifica in modo più marcato le caratteristiche facciali quando si passa da una fascia d'età a un'altra. Ciò è particolarmente evidente, ad esempio quando si passa da soggetti di mezza età a soggetti anziani nella transizione (60-69,70-79).

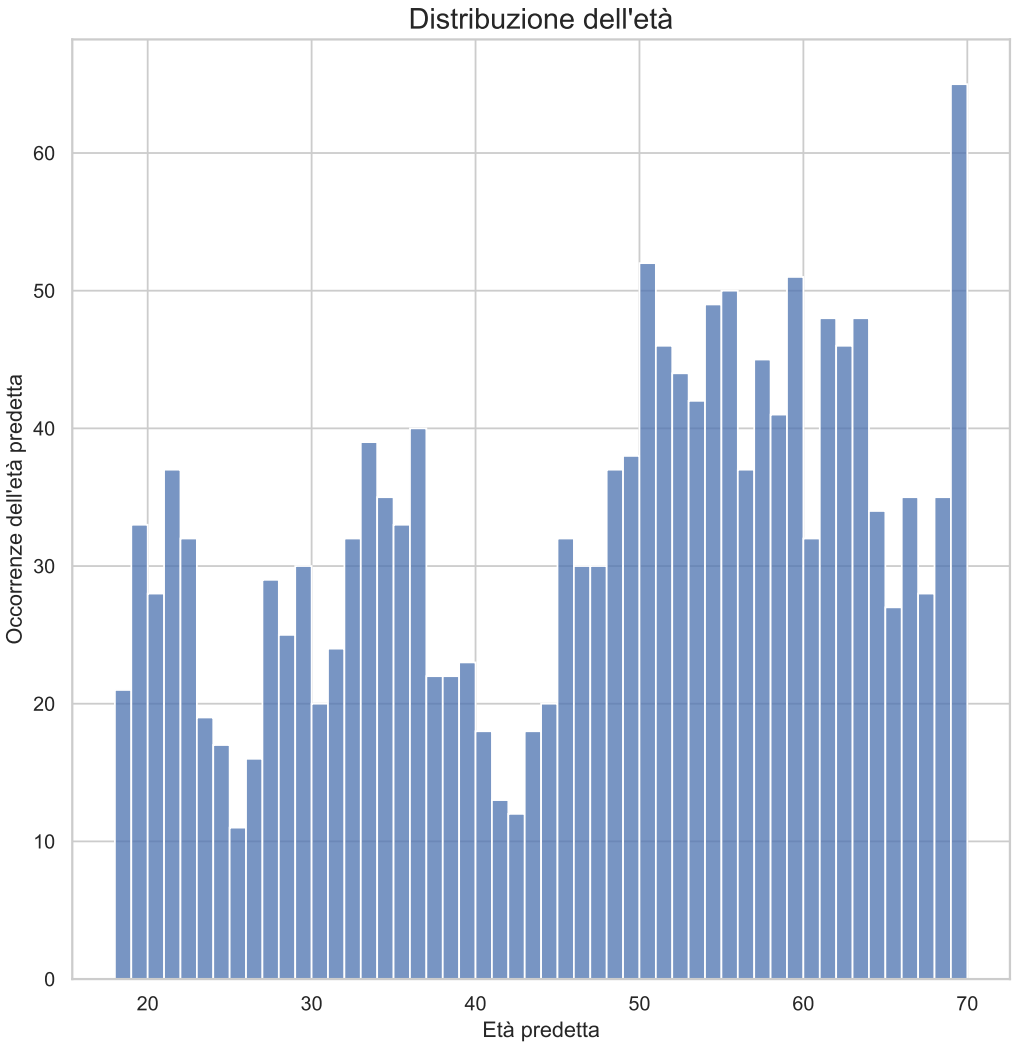
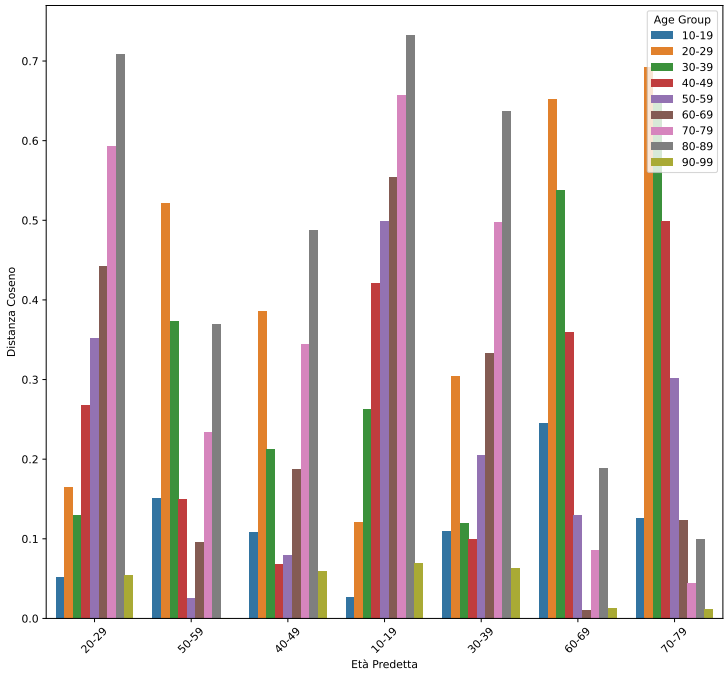
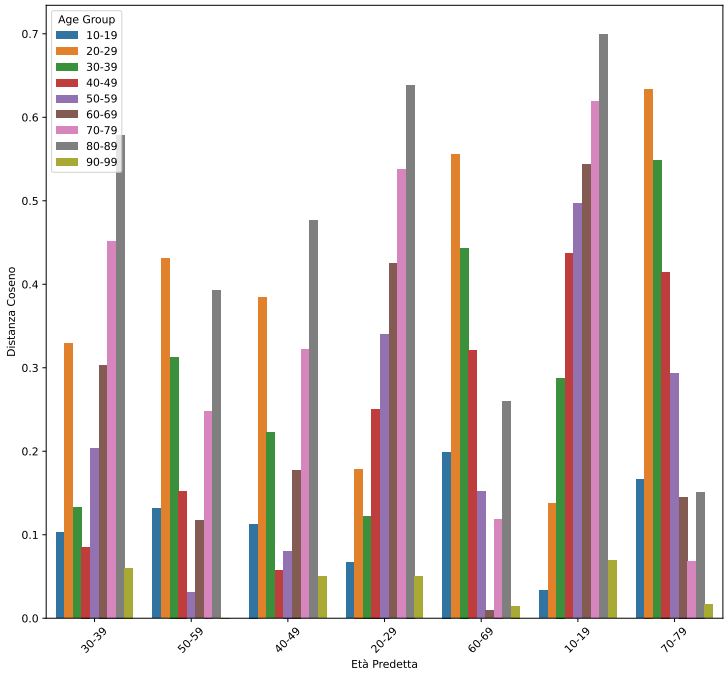


Figura 4.3: L'immagine rappresenta la distribuzione delle età stimate utilizzando il modello DEX come predittore all'interno del modello FADING. A differenza della distribuzione ottenuta con DeepFace, qui si osserva una maggiore omogeneità nelle occorrenze predette lungo tutto il range di età, da 18 a 80 anni. Questa distribuzione più equilibrata ha portato a una riduzione significativa dell'errore medio totale, che si attesta a 5,7 anni, con una predizione più accurata anche nelle fasce d'età avanzate

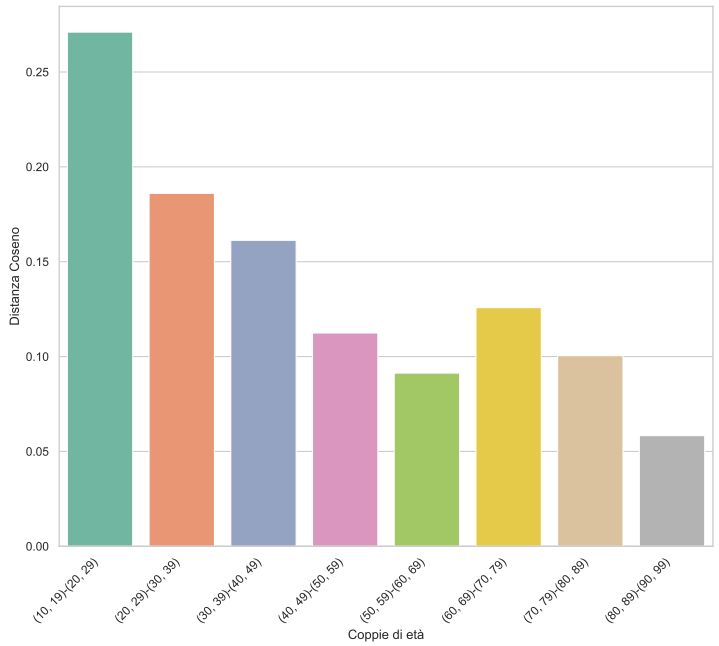


(a) distanza coseno tra l'immagine in input e ciascuna delle immagini generate da un processo di aging utilizzando ArcFace per estrarre l'embedding

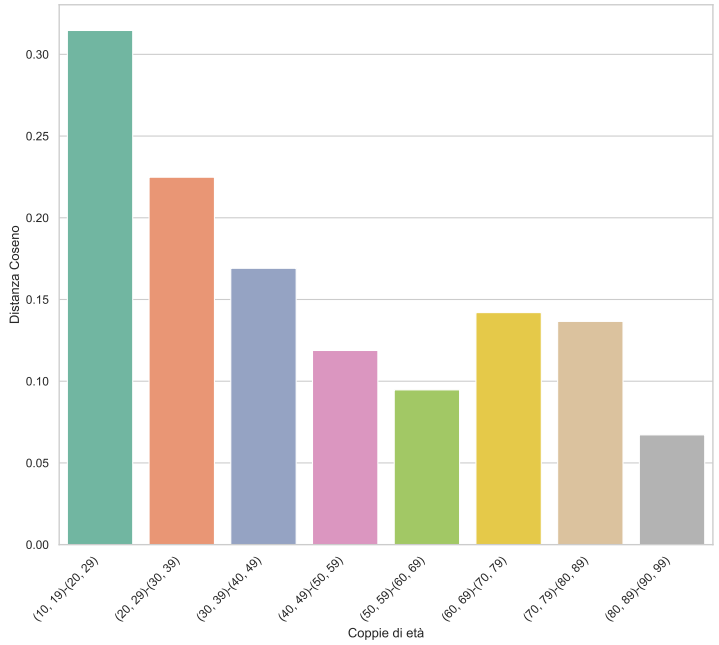


(b) distanza coseno tra l'immagine in input e ciascuna delle immagini generate da un processo di aging utilizzando AdaFace per estrarre l'embedding

Figura 4.4: confronto delle distanze coseno tra immagini generate con FADING il cui embedding è stato estratto con AdaFace e ArcFace



(a) distanza coseno tra immagini generate a intervalli di 10 anni di età usando il modello ArcFace per estrarre l'embedding



(b) distanza coseno tra immagini generate a intervalli di 10 anni di età usando il modello AdaFace per estrarre l'embedding

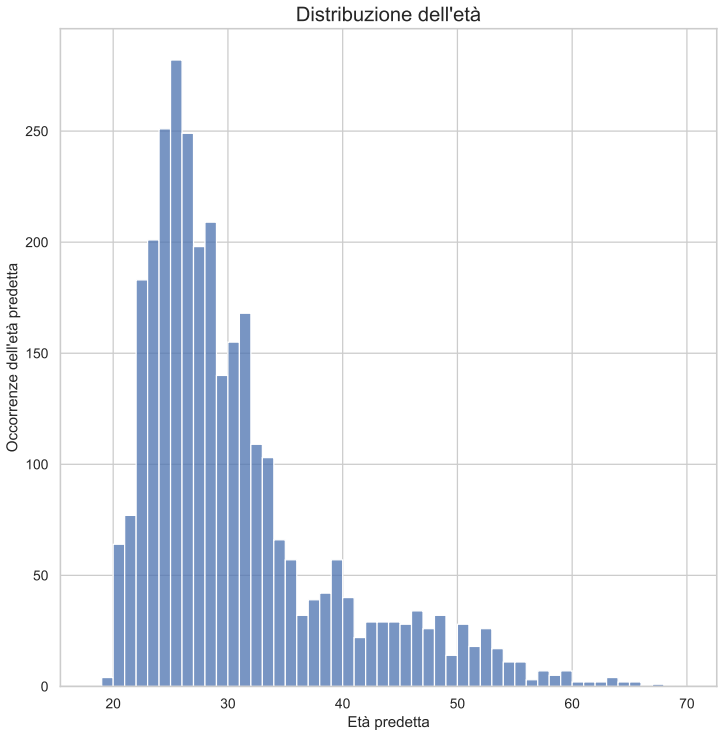
Figura 4.5: confronto delle distanze coseno tra immagini generate dal modello FADING il cui embedding è stato estratto con AdaFace e ArcFace

Age Range	Average Error
Overall	31.047
10-19	11.678
20-29	4.146
30-39	10.744
40-49	18.123
50-59	26.112
60-69	32.845
70-79	48.356
80-89	58.356
90-99	68.356

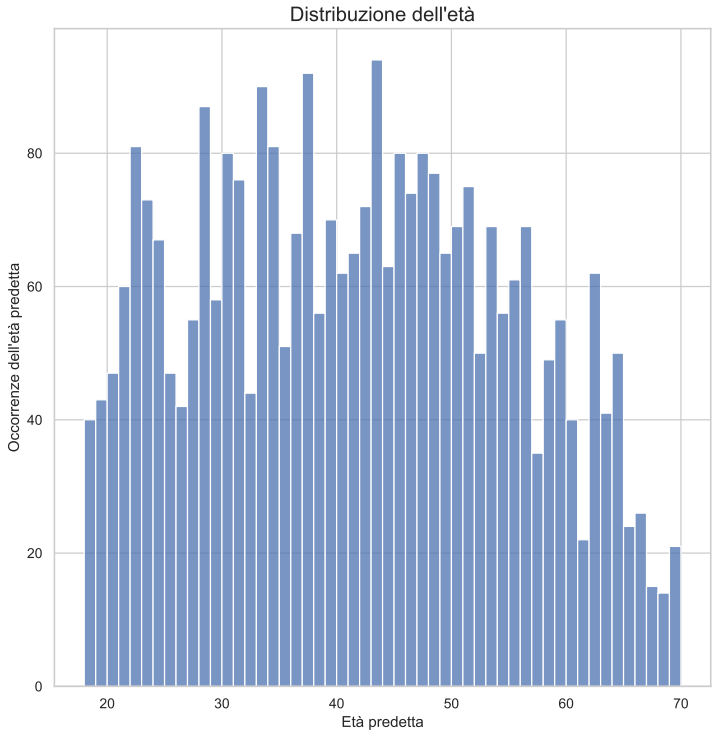
Figura 4.6: Errore assoluto medio per ogni fascia di età di immagini generate con HRFAE usando il classificatore dell'età DeepFace.

Age Range	Average Error
Overall	24.993
10-19	22.739
20-29	8.882
30-39	10.613
40-49	11.163
50-59	13.312
60-69	16.312
70-79	37.306
80-89	47.306
90-99	57.306

Figura 4.7: Errore assoluto medio per ogni fascia di età di immagini generate con HRFAE usando il classificatore dell'età DEX.

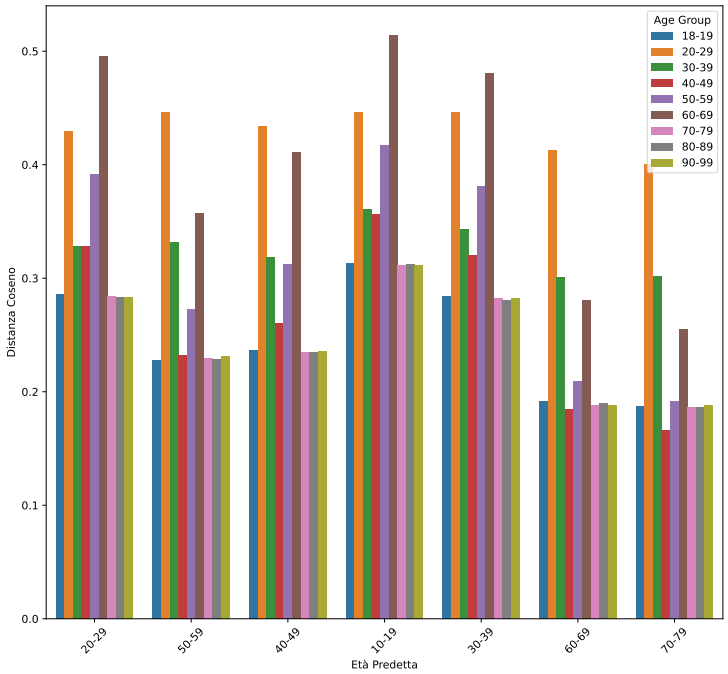


(a) Distribuzione dell'età usando il classificatore dell'età DeepFace

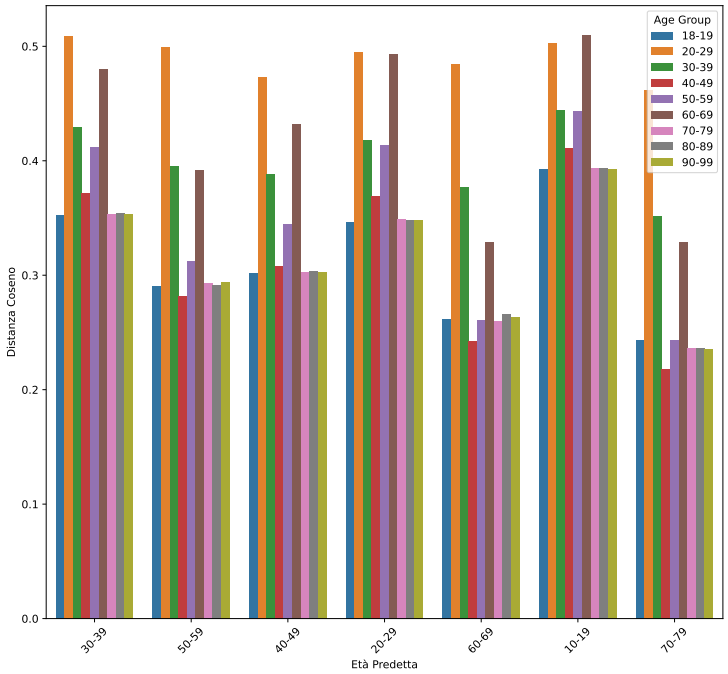


(b) Distribuzione dell'età usando il classificatore dell'età DEX

Figura 4.8: Confronto delle distribuzioni dell'età delle immagini generate da HRFAE usando i classificatori DeepFace e DEX

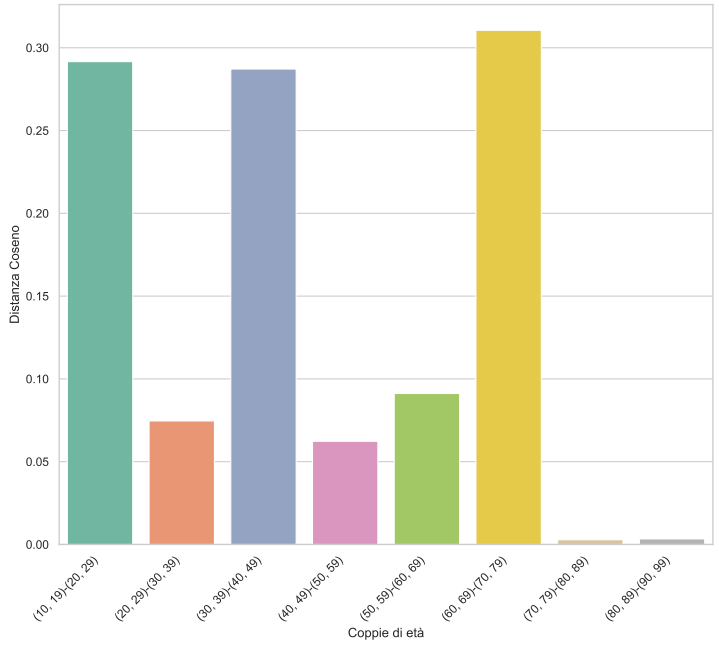


(a) distanza coseno tra l'immagine in input e ciascuna delle immagini generate da un processo di aging utilizzando ArcFace per estrarre l'embedding

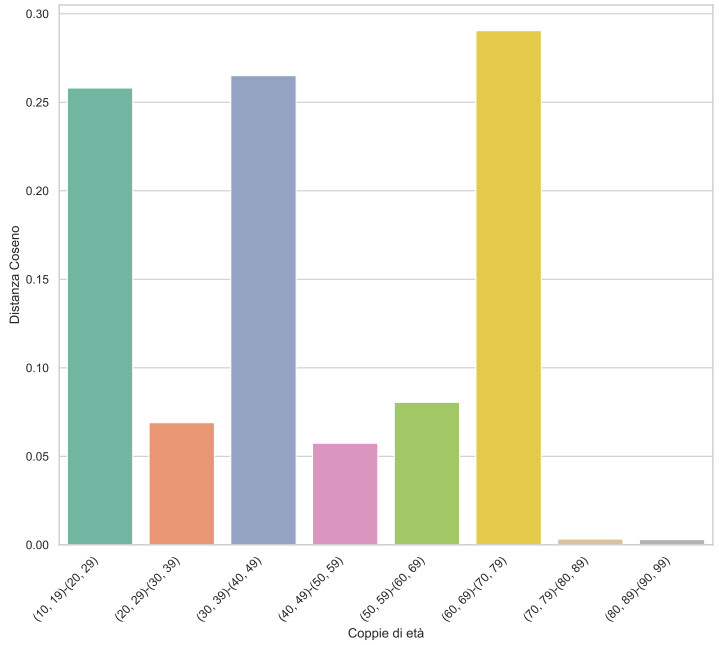


(b) distanza coseno tra l'immagine in input e ciascuna delle immagini generate da un processo di aging utilizzando AdaFace per estrarre l'embedding

Figura 4.9: confronto delle distanze coseno tra immagini generate con HRFAE, il cui embedding è stato estratto con AdaFace e ArcFace



(a) distanza coseno tra immagini generate a intervalli di 10 anni di età usando il modello ArcFace per estrarre l'embedding



(b) distanza coseno tra immagini generate a intervalli di 10 anni di età usando il modello AdaFace per estrarre l'embedding

Figura 4.10: confronto delle distanze coseno tra immagini generate dal modello HRFAE il cui embedding è stato estratto con AdaFace e ArcFace

Capitolo 5

Conclusioni

In questo lavoro sono stati confrontati due modelli di face aging: HRFAE e FADING. Entrambi i modelli sono stati valutati sia qualitativamente che quantitativamente, rivelando differenze significative nelle loro performance. HRFAE ha dimostrato delle limitazioni importanti: in particolare il modello tende a clusterizzare i volti in gruppi specifici, determinando a una scarsa variabilità somatica tra immagini generate all'interno delle stesse fasce d'età. Inoltre il modello non riesce ad applicare un efficace processo di aging per fasce di età estreme, nello specifico per soggetti di età inferiore ai 20 anni e superiore ai 70. Ciò ha comportato un errore medio più elevato e una minore distanza coseno tra i volti generati e quelli di partenza. Il modello FADING si è invece rivelato più produttivo, come si evidenzia in Figura 5.1 che mostra il confronto fra soggetti di varie etnie, maschi e femmine, alcuni dei quali indossano accessori come gli occhiali. Si puó osservare che il modello riesce a mantenere fedelmente non solo i tratti somatici principali, ma anche dettagli specifici come la presenza di barba o accessori, preservando l'identità del sogget-

to per tutte le fasce d'età. Questo risultato è particolarmente significativo perché esprime la capacità del modello di adattarsi a una grande varietà di caratteristiche fisiche e stilistiche, fornendo una simulazione dell'invecchiamento naturale accurata e convincente, indipendentemente dal genere o dalle peculiarità individuali. Inoltre le transizioni tra diverse età sono risultate più fluide e coerenti, e il modello è stato in grado di gestire anche le fasce anagrafiche estreme con maggiore precisione. Sebbene anche FADING presenti qualche difficoltà nell'invecchiamento di volti estremamente anziani, le sue prestazioni complessive risultano nettamente superiori rispetto a HRFAE. Alla luce di questi risultati, è stato adottato FADING come modello principale per la generazione di volti invecchiati. Questo approccio bilancia in modo ottimale il realismo e la conservazione dell'identità, risultando particolarmente adatto a contesti di sicurezza in cui sia fondamentale il riconoscimento facciale. I volti invecchiati generati tramite FADING costituiscono il dataset AMONOT (Aged Morphed ONOT), che sarà utilizzato per la valutazione della robustezza dei sistemi di rilevamento degli attacchi di morphing in scenari reali caratterizzati da variazioni biometriche legate all'età.

Figura 5.1: Immagini generate con il modello di face aging FADING, ogni riga rappresenta un individuo diverso, con progressiva modifica dell'età da sinistra a destra: da un ringiovanimento di 20 anni (-20) a un invecchiamento di 20 anni (+20)

Capitolo 6

Appendice

6.1 Dettagli Tecnici del Modello FADING

Per generare le immagini invecchiate utilizzando il modello FADING è stato sviluppato e utilizzato uno script Python che automatizza l'intero processo di trasformazione delle immagini. Il comando per eseguire questo script è il seguente:

```
python age_editing.py \
--image_path <PATH_TO_INPUT_IMAGE> \
--age_init <INITIAL_AGE> \
--gender <female|male> \
--save_aged_dir <OUTPUT_DIR> \
--specialized_path <PATH_TO_SPECIALIZED_MODEL> \
--target_ages 10 20 30 40 50 60 70 80 90 100
```

Nel suddetto comando, i seguenti parametri influenzano il funzionamento del modello e il risultato finale:

- `image_path` indica il percorso dell'immagine in input che deve essere elaborata. Tale immagine, che può rappresentare un individuo di qualsiasi età, è il punto di partenza per il processo di invecchiamento.

- `age_init` rappresenta l'età effettiva del soggetto nell'immagine di input. Questo parametro è essenziale per il modello, in quanto consente di calcolare l'evoluzione delle caratteristiche facciali in funzione del tempo.

- `gender` specifica il genere del soggetto nell'immagine (`female` o `male`), distinzione necessaria per garantire che il modello applicato utilizzi correttamente le caratteristiche associate a ciascun genere.

- `save_aged_dir` è il percorso della cartella in cui verranno salvate le immagini generate in output.

- `specialized_path` indica il percorso del modello pre-addestrato che verrà utilizzato per il processo di invecchiamento. Nel caso specifico, è stato impiegato il modello pre-addestrato fornito nella repository al seguente link.

- `target_ages` definisce l'insieme delle età target per le immagini in output. All'interno del presente lavoro, sono stati scelti intervalli di 10 anni, con target di età compresi tra 10 e 100 anni, per monitorare l'invecchiamento progressivo.

6.2 Dettagli Tecnici del Modello HRFAE

Per l'esecuzione di tale modello, è stato utilizzato lo script `test.py`, il cui comando di esecuzione è il seguente:

```
python test.py --config 001 --target_age 65
```

In questo caso, lo script è stato eseguito per ciascuno dei target di età precedentemente definiti (da 18 a 100 anni, con intervalli di 10 anni).

Le immagini di input devono essere salvate nel percorso `/test/input`, da cui lo script le preleverà per eseguire il processo di invecchiamento. Le immagini generate in output, a loro volta, verranno automaticamente salvate nella directory `/test/output`, mantenendo così un'organizzazione chiara e strutturata dei file generati. A differenza di FADING, HRFAE integra al suo interno un classificatore per determinare l'età dell'immagine di input, eliminando così la necessità di fornire manualmente un parametro di età iniziale, o di genere. Ciò rappresenta quindi un significativo vantaggio in termini di semplicità d'uso.

6.3 Embedding

In ambito di Computer Vision, l'embedding di immagini è una rappresentazione numerica che consente di catturare le caratteristiche distintive di un'immagine sotto forma di un vettore di feature in uno spazio a più dimensioni. Tale vettore, noto come feature vector, racchiude informazioni rilevanti che descrivono l'immagi-

ne, come contorni, colori, texture e altre proprietà visive, in una forma compatta e interpretabile dai modelli di apprendimento automatico.

L'obiettivo principale degli embedding consiste nel mappare immagini simili vicine nello spazio vettoriale e immagini dissimili lontane. È possibile in tal modo misurare la somiglianza tra due immagini calcolando la distanza tra i loro vettori di feature nello spazio vettoriale. Tecniche come la distanza coseno o la distanza euclidea sono spesso utilizzate per quantificare la somiglianza suddetta.

Gli embedding di immagini sono fondamentali per molte applicazioni di computer vision, come il riconoscimento facciale, il recupero di immagini, la classificazione e i sistemi di raccomandazione visiva. Ad esempio, nel riconoscimento facciale, l'embedding di un volto viene confrontato con embedding di volti presenti in un database per identificare o verificare l'identità della persona.

In sintesi, l'embedding di immagini è una tecnica chiave per comprendere e misurare le somiglianze tra immagini, supportando una vasta gamma di applicazioni che richiedono la capacità di distinguere e classificare contenuti visivi in modo efficiente e accurato.

6.4 File di configurazione per la fase di training HRFAE

Di seguito viene riportato il codice del file di configurazione per la fase di training di HRFAE, che evidenzia i principali iperparametri impiegati nel processo:

```yaml
# Input data
input_w: 1024
input_h: 1024
age_min: 20
age_max: 70
# Training hyperparameters
batch_size: 2
epochs: 20
# Optimizer parameters
lr: 0.0001
beta_1: 0.9
beta_2: 0.999
weight_decay: 0.0005
# Learning rate scheduler
step_size: 10
gamma: 0.1
# Tensorboard log options
image_save_iter: 3000
image_log_iter: 1000
log_iter: 10
# Loss weight
w:
  recon: 10
  class: 0.1
  adver: 1
  dis: 1
  gp: 1
  tv: 0
```

6.5 Diffusion Models

I Diffusion Models sono una classe di modelli generativi utilizzati per creare immagini o altri tipi di dati a partire da una distribuzione casuale di rumore; essi operano attraverso un processo iterativo che trasforma progressivamente il rumore in un'immagine ad alta qualità. Tale processo si articola in due fasi principali, diffusione e denoising: nella prima fase, un'immagine chiara e dettagliata viene gradualmente degradata con l'aggiunta di rumore fino a diventare interamente rumorosa, perdendo del tutto le proprie caratteristiche visive. La fase di denoising costituisce il processo inverso: il modello prende in ingresso un'immagine rumorosa e, attraverso una serie di passaggi, rimuove progressivamente il rumore, recuperando l'immagine originale o generandone una nuova. I modelli di diffusione possono generare immagini in due modalità: incondizionata e condizionata. Nella generazione incondizionata, il modello parte da rumore casuale e crea un'immagine senza alcuna guida esterna; nella generazione condizionata si utilizzano informazioni aggiuntive, come una descrizione testuale o un'etichetta di classe, per guidare il processo di generazione e ottenere un'immagine coerente con le specifiche fornite. Questa capacità di controllare la generazione delle immagini rende i modelli di diffusione particolarmente potenti e flessibili, portando a risultati di alta qualità in diversi campi di applicazione, dalla creazione artistica alla sintesi di immagini per applicazioni scientifiche e mediche.

6.6 Ringraziamenti per il supporto tecnico

Si ringrazia il CINECA per l'assegnazione delle risorse di calcolo ad alte prestazioni (HPC) e per il supporto fornito nell'ambito dell'iniziativa ISCRA.

Bibliografia

[1] I. Batskos, F. F. de Wit, L. J. Spreeuwers, and R. J. Veldhuis, "Preventing face morphing attacks by using legacy face images," *IET biometrics*, vol. 10, no. 4, pp. 430–440, 2021.

[2] G. Borghi, N. Di Domenico, A. Franco, M. Ferrara, and D. Maltoni, "Revelio: A modular and effective framework for reproducible training and evaluation of morphing attack detectors," *IEEE Access*, 2023.

[3] S. Venkatesh, R. Ramachandra, K. Raja, and C. Busch, "Face morphing attack generation and detection: A comprehensive survey," *IEEE transactions on technology and society*, vol. 2, no. 3, pp. 128–145, 2021.

[4] F. Matteo, F. Annalisa, and M. Davide, "The magic passport," in *IEEE International Joint Conference on Biometrics (IJCB'14)*, 2014, pp. 1–7.

[5] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, "Face recognition systems under morphing attacks: A survey," *IEEE Access*, vol. 7, pp. 23 012–23 026, 2019.

[6] U. Scherhag, A. Nautsch, C. Rathgeb, M. Gomez-Barrero, R. N. Veldhuis, L. Spreeuwers, M. Schils, D. Maltoni, P. Grother, S. Marcel *et al.*, "Biometric systems under morphing

attacks: Assessment of morphing techniques and vulnerability reporting," in *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2017, pp. 1–7.

[7] C. Seibold, W. Samek, A. Hilsmann, and P. Eisert, "Detection of face morphing attacks by deep learning," in *Digital Forensics and Watermarking: 16th International Workshop, IWDW 2017, Magdeburg, Germany, August 23-25, 2017, Proceedings 16*. Springer, 2017, pp. 107–120.

[8] N. Di Domenico, G. Borghi, A. Franco, and D. Maltoni, "Onot: a high-quality icao-compliant synthetic mugshot dataset," *arXiv preprint arXiv:2404.11236*, 2024.

[9] Y. Alaluf, O. Patashnik, and D. Cohen-Or, "Only a matter of style: Age transformation using a style-based regression model," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–12, 2021.

[10] L. Pinheiro Cinelli, M. Araújo Marins, E. A. Barros da Silva, and S. Lima Netto, "Variational autoencoder," in *Variational Methods for Machine Learning with Applications to Deep Networks*. Springer, 2021, pp. 111–149.

[11] A. Aggarwal, M. Mittal, and G. Battineni, "Generative adversarial network: An overview of theory and applications," *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100004, 2021.

[12] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of

stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[14] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[15] F. Carlsson, P. Eisen, F. Rekathati, and M. Sahlgren, "Cross-lingual and multilingual clip," in *Proceedings of the thirteenth language resources and evaluation conference*, 2022, pp. 6848–6854.

[16] X. Yao, G. Puy, A. Newson, Y. Gousseau, and P. Hellier, "High resolution face age editing," in *2020 25th International conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 8624–8631.

[17] X. Chen and S. Lathuilière, "Face aging via diffusion-based editing," *arXiv preprint arXiv:2309.11321*, 2023.

[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[19] S. Banerjee, G. Mittal, A. Joshi, C. Hegde, and N. Memon, "Identity-preserving aging of face images via latent diffusion models," in *2023 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2023, pp. 1–10.

[20] G. Gomez-Trenado, S. Lathuilière, P. Mesejo, and Ó. Cordón, "Custom structure preservation in face aging," in *European Conference on Computer Vision*. Springer, 2022, pp. 565–580.

[21] R. Or-El, S. Sengupta, O. Fried, E. Shechtman, and I. Kemelmacher-Shlizerman, "Lifespan age transformation synthesis," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 739–755.

[22] N. Pavlichenko and D. Ustalov, "Imdb-wiki-sbs: An evaluation dataset for crowdsourced pairwise comparisons," *arXiv preprint arXiv:2110.14990*, 2021.

[23] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6038–6047.

[24] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE access*, vol. 9, pp. 82 031–82 057, 2021.

[25] Z. Zhang, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*. Ieee, 2018, pp. 1–2.

[26] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC 2015-Proceedings of the British Machine*

Vision Conference 2015. British Machine Vision Association, 2015.

[27] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 10–15.

[28] M. Kim, A. K. Jain, and X. Liu, "Adaface: Quality adaptive margin for face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 750–18 759.

[29] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.