

Corso di Laurea in Ingegneria e Scienze Informatiche

Inversione del processo di face morphing mediante algoritmi di deep learning e tecniche di preprocessing delle immagini biometriche.

Tesi di laurea in:
VISIONE ARTIFICIALE

Relatore

Prof. Annalisa Franco

Candidato

Elena Montalti

Correlatori

Prof. Matteo Ferrara

Dott. Nicolò Di Domenico

Sommario

L'analisi automatica di volti tramite tecniche di visione artificiale è oggi un elemento centrale nei sistemi di identificazione biometrica. Nell'ambito dei documenti di identità elettronici, la corretta verifica dell'identità è fondamentale per garantire sicurezza e affidabilità. Diventa quindi necessario accertare che le immagini utilizzate nei controlli non siano state alterate e che il volto presente sul documento corrisponda effettivamente al suo proprietario.

In questo contesto emerge il problema del *face morphing*, una manipolazione che permette di ottenere immagini apparentemente autentiche ma create mediante la fusione di più identità. Tale alterazione costituisce una vulnerabilità significativa, poiché l'immagine morphed può risultare compatibile con più soggetti, compromettendo i processi di verifica e aumentando il rischio di frodi documentali. Per questo motivo, l'analisi automatica del morphing richiede strumenti in grado di rilevare e valutare quantitativamente tali manipolazioni.

Il lavoro di tesi è incentrato sull'analisi delle tecniche di morphing facciale e sulla progettazione di possibili contromisure per contrastarne l'impatto sui sistemi biometrici. In particolare, l'attività comprende lo studio dei processi di generazione delle immagini morphed, lo sviluppo e l'applicazione di tecniche di post-processing e di inversione del morphing (*demorphing*), con l'obiettivo di comprenderne il funzionamento, riprodurre il processo di generazione e verificarne sperimentalmente la possibilità di inversione attraverso l'analisi dei risultati ottenuti.

Ai miei genitori, per l'amore e il supporto.

Indice

Sommario	iii
1 Introduzione	1
1.1 Face morphing come vulnerabilità nei sistemi biometrici	1
1.2 Scenario applicativo: Automated Border Control (ABC) gate	2
1.3 Obiettivi e contributi della tesi	3
2 Face recognition	5
2.1 Face recognition basato su deep-learning	5
2.2 Metriche di similarità/distanza	7
3 Face morphing con Arc2Face	9
3.1 Definizione matematica del face morphing	9
3.2 Rimozione del background per ICAO compliance	13
3.2.1 Descrizione dei diversi metodi testati	15
3.2.2 Confronto visivo e commenti	18
4 Face demorphing con Arc2Face	25
4.1 Definizione del problema inverso: il demorphing	25
4.2 Inversione approssimata della SLERP	27
5 Prove sperimentali	31
5.1 Dataset e modelli utilizzati	31
5.1.1 Dataset utilizzati	31
5.1.2 Modelli della pipeline generativa	32
5.2 Metriche di valutazione	34
5.2.1 Scenario "ideale" di morphing attack	35
5.2.2 Scenario "reale" di morphing attack	36
5.2.3 Scenario documento bona fide	37
5.3 Protocollo	38
5.4 Analisi dei risultati	40

INDICE

6 Conclusioni	45
6.1 Sintesi dei risultati	45
6.2 Limiti e sviluppi futuri	46
	49
Bibliografia	49

Elenco delle figure

3.1	Confronto geometrico tra LERP e SLERP	12
3.2	Esempio di segmentazione ottenuta con U ² -Net	19
3.3	Esempio di segmentazione ottenuta con Sam2	20
3.4	Esempio di segmentazione ottenuta con SegFormer	21
3.5	Esempio di segmentazione ottenuta con BiSeNet	22
3.6	Esempio di segmentazione ottenuta con BEN2	23
3.7	Esempio di segmentazione ottenuta con ModNet	24
5.1	Sintesi della pipeline generativa [DDFFM26]	34

Capitolo 1

Introduzione

1.1 Face morphing come vulnerabilità nei sistemi biometrici

Il face morphing rappresenta una tecnica di manipolazione che combina due volti differenti in un'unica immagine sintetica, mantenendo caratteristiche biometriche di entrambi i soggetti coinvolti [FFM14].

Il processo di generazione di un'immagine morphed consiste nella combinazione controllata delle due immagini facciali distinte, in cui le informazioni caratteristiche di ciascun volto vengono integrate per produrre un volto sintetico che risulti plausibile e compatibile con entrambe le identità di partenza.

Il risultato è un'immagine visivamente realistica che incorpora tratti di entrambe le identità coinvolte. Questa proprietà rende il morphing particolarmente critico in ambito biometrico, poiché l'immagine ottenuta può essere riconosciuta come sufficientemente simile a ciascuno dei soggetti originali dai sistemi automatici di verifica dell'identità. In altre parole, un volto morphed può produrre una rappresentazione biometrica intermedia in grado di soddisfare le soglie decisionali adottate dai sistemi di riconoscimento facciale.

Di conseguenza, un documento di identità contenente una fotografia morphed può rappresentare una vulnerabilità significativa. Più individui potrebbero potenzialmente utilizzare il medesimo documento per superare i controlli automatici, compromettendo il legame univoco tra identità e documento. Per questo motivo,

lo studio di tecniche in grado di rilevare e analizzare automaticamente immagini morphed risulta essenziale per migliorare la robustezza dei sistemi biometrici moderni.

1.2 Scenario applicativo: Automated Border Control (ABC) gate

Negli ultimi anni i documenti elettronici di identità (*electronic Machine Readable Travel Documents*, eMRTD) hanno progressivamente sostituito i tradizionali documenti cartacei, includendo al loro interno informazioni biometriche utili alla verifica automatica dell'identità. In particolare, l'*International Civil Aviation Organization* (ICAO) ha stabilito il volto come principale tratto biometrico nei documenti di identità elettronici, definendone formati e requisiti in modo da renderlo uno standard condiviso a livello internazionale per i controlli di identità.

In questo scenario si collocano i sistemi di *Automated Border Control* (ABC), ovvero infrastrutture automatizzate progettate per supportare l'identificazione dei viaggiatori mediante un processo di verifica biometrica. Il funzionamento tipico di un ABC gate prevede che l'utente presenti il documento elettronico, dal quale viene acquisita l'immagine del volto memorizzata nel chip, e successivamente venga acquisita un'immagine live tramite videocamera o sensore. Le immagini vengono elaborate da un sistema di verifica del volto che utilizza algoritmi proprietari per confrontare l'immagine acquisita in tempo reale con quella memorizzata nel documento elettronico. Tali sistemi si basano su algoritmi di riconoscimento facciale che analizzano le caratteristiche del volto e producono una misura di similarità tra le due immagini. Il processo di verifica consiste nel valutare tale misura rispetto a una soglia decisionale prestabilita.

L'adozione di tali sistemi consente di automatizzare i controlli, migliorare l'efficienza e ridurre l'intervento umano; tuttavia, rende i processi di verifica particolarmente sensibili alla presenza di alterazioni intenzionali nelle immagini biometriche. In particolare, la presenza di una fotografia morphed all'interno di un documento di identità può compromettere l'affidabilità del sistema, se la fusione tra i volti è sufficientemente realistica e produce punteggi di similarità superiori alla soglia

decisionale, l'immagine sintetica può risultare compatibile con entrambi i soggetti coinvolti.

La diffusione di questa fragilità è stata favorita dalla crescente automazione dei controlli e dall'utilizzo di fotografie fornite dall'utente, spesso acquisite in condizioni non completamente controllate. Tali fattori, possono ingannare e risultano particolarmente difficili da rilevare anche da un operatore umano.

Alla luce di queste vulnerabilità, diventa necessario sviluppare metodi automatici in grado di analizzare immagini sospette e supportare la verifica dell'identità nei sistemi biometrici, riducendo il rischio di accettazione di documenti contenenti fotografie morphed.

1.3 Obiettivi e contributi della tesi

L'obiettivo del progetto è contribuire allo sviluppo di un approccio orientato all'identificazione e alla valutazione di immagini biometriche manipolate tramite tecniche di face morphing. Una possibile strategia di analisi consiste nel tentare di ricostruire l'identità nascosta a partire dall'immagine del documento e da una fotografia live scattata durante il controllo. Se la ricostruzione produce un volto compatibile con una seconda identità diversa dal soggetto presente al controllo, l'immagine del documento può essere potenzialmente generata tramite morphing.

Il contributo principale di questa tesi consiste, da un lato, nell'introduzione di una fase di postprocessing, basata su modelli di rimozione dello sfondo, per garantire la conformità delle immagini morphed ai requisiti ISO/ICAO previsti per le foto dei documenti d'identità, e dall'altro, nello sviluppo e nella sperimentazione di un algoritmo di demorphing in grado di invertire il processo di generazione delle immagini morphed.

Il resto del lavoro è strutturato come segue.

Il Capitolo 2 fornisce una panoramica dei concetti fondamentali di face recognition, utile a comprendere il funzionamento di base degli algoritmi e metriche di valutazione adottati.

Il Capitolo 3 descrive il processo di face morphing nello spazio degli embedding, analizzando le tecniche di interpolazione lineare e sferica, e mostra l'adattamento

delle immagini generate agli standard ICAO mediante tecniche di rimozione e normalizzazione dello sfondo.

Il Capitolo 4 presenta il problema inverso del demorphing realizzato attraverso la formula di inversione della SLERP e descrive la strategia proposta per la ricostruzione dell'identità complementare.

Il Capitolo 5 mostra le prove sperimentali effettuate, i dati e i modelli impiegati nella pipeline generativa e di verifica, mettendo a confronto i risultati attesi con quelli ottenuti nelle diverse configurazioni analizzate.

Infine, nel Capitolo 6 sono riportate le conclusioni del lavoro, con una sintesi dei risultati raggiunti, la discussione dei limiti emersi e le possibili direzioni di sviluppo futuro.

Capitolo 2

Face recognition

2.1 Face recognition basato su deep-learning

Il *face recognition* è una tecnologia biometrica progettata per identificare o verificare l'identità di un individuo analizzando le caratteristiche fisiologiche del suo volto. I moderni sistemi di riconoscimento facciale si basano principalmente su modelli di deep learning e seguono una pipeline di elaborazione composta da una sequenza di fasi consecutive, ciascuna funzionale alla successiva.

Il processo inizia con la fase di rilevamento, durante la quale il sistema individua la presenza di uno o più volti all'interno dell'immagine e ne delimita la posizione. Una volta isolato il volto, si procede con l'allineamento, che consiste nell'identificazione di punti di riferimento facciali al fine di correggere eventuali variazioni di posa, inclinazione o rotazione. Successivamente, l'immagine viene sottoposta a normalizzazione. Il volto viene ridimensionato e i valori dei pixel vengono standardizzati per garantire che l'immagine in ingresso sia compatibile con quelle su cui la rete è stata addestrata.

L'immagine processata viene fornita alla rete neurale che, attraverso l'uso di *Deep Convolutional Neural Network* (DCNN), è in grado di tradurre i pixel di un'immagine in un vettore numerico compatto, detto *embedding*.

Questa rappresentazione contiene informazioni più astratte legate all'identità della persona. I modelli di Deep Learning sono addestrati per essere invarianti rispetto ai cambiamenti che non modificano l'identità, come illuminazione, contra-

sto, rumore o colore della pelle, che possono alterare molti pixel senza influenzare il riconoscimento umano.

Lo spazio degli embedding è uno spazio vettoriale, di dimensione fissa, in cui ogni volto viene rappresentato come un vettore numerico.

$$x \in \mathbb{R}^d$$

Durante l'addestramento, il modello viene ottimizzato tramite funzioni di "perdita" (*loss*) che impone vincoli di separabilità tra classi. Il modello posiziona i vettori all'interno dello spazio in modo tale che la distanza tra di essi rifletta direttamente la somiglianza dell'identità: volti della stessa persona devono avere una distanza intra-classe piccola, mentre volti di persone diverse devono avere una distanza inter-classe grande [WD21].

In questo contesto, sono stati proposti numerosi approcci per il face recognition, che differiscono per architettura della rete, funzione di perdita e modalità di strutturazione dello spazio degli embedding. Alcuni metodi si basano su schemi di classificazione tradizionale con softmax, altri adottano perdite metriche come la triplet loss, mentre approcci più recenti introducono vincoli geometrici espliciti per rafforzare la separabilità tra identità [SO24].

In particolare, alcuni modelli riformulano il problema della separabilità in termini puramente angolari, imponendo un vincolo di normalizzazione sugli embedding e introducendo un margine esplicito tra le classi.

Il modello di riconoscimento facciale ArcFace (*Additive Angular Margin Loss*) [DGXZ19] affina questo processo proiettando gli embedding su una ipersfera unitaria. Questo significa che i vettori non sono più distribuiti liberamente in \mathbb{R} , ma sono vincolati a stare sulla superficie della sfera S^{d-1} . In questo spazio la somiglianza tra due vettori non dipende dalla lunghezza, ma solo dall'angolo tra di essi.

In uno spazio geometrico normalizzato la somiglianza tra due volti corrisponde direttamente all'ampiezza dell'angolo tra i vettori corrispondenti. ArcFace introduce un vincolo aggiuntivo imponendo un margine angolare m sull'angolo associato alla classe corretta. Questo significa che la distanza tra due volti simili risulti

maggiore, poiché

$$\cos(\theta + m) < \cos(\theta)$$

dove θ è l'angolo tra i due embedding.

Questo significa che, anche se il volto è già abbastanza vicino alla classe corretta, il modello, durante l'addestramento, lo considera comunque “non sufficientemente simile”. Per ridurre l'errore e ottenere una predizione corretta, la rete è quindi costretta a produrre embedding ancora più vicini al centro della classe, creando confini netti e distanti tra una persona e l'altra. Questo migliora la capacità del modello di distinguere identità simili [DGY⁺22].

2.2 Metriche di similarità/distanza

Gli embedding, visti singolarmente, non hanno significato immediato, poiché rappresentano una codifica numerica astratta del volto. La loro utilità deriva dal fatto che sono costruiti in modo da poter essere confrontati tra loro all'interno di uno spazio geometrico. Questo permette di misurare la somiglianza tra due volti applicando specifiche metriche di distanza o similarità.

Nei sistemi di face recognition che operano su embedding normalizzati, la somiglianza viene misurata tramite la *cosine similarity*. Definita come:

$$\text{sim}(z_1, z_2) = \frac{z_1 \cdot z_2}{\|z_1\| \|z_2\|}$$

dove $z_1 \cdot z_2$ rappresenta il prodotto scalare e $\|z\|$ la norma euclidea. Questa misura rappresenta il coseno dell'angolo compreso tra gli embedding e quantifica quanto essi siano orientati nella stessa direzione. Per definizione, la cosine similarity assume valori nell'intervallo $[-1, 1]$.

- $\text{sim} \approx 1$, i vettori sono fortemente allineati e quindi i volti associati risultano molto simili.
- $\text{sim} \approx 0$, i vettori hanno scarsa correlazione e quindi bassa somiglianza.
- $\text{sim} \approx -1$, i vettori hanno direzione opposta (generalmente raro).

Viceversa, la dissimilarità tra due embedding può essere espressa tramite la *cosine distance*. Definita come:

$$\text{dist}(z_1, z_2) = 1 - \text{sim}(z_1, z_2)$$

La distanza coseno assume valori nell'intervallo $[0, 2]$, dove valori prossimi a 0 indicano alta somiglianza tra i volti, mentre valori più elevati indicano una maggiore differenza. In altre parole la distanza coseno fornisce una misura complementare rispetto alla similarità coseno.

La decisione di verifica biometrica viene effettuata confrontando il valore ottenuto con una soglia τ , scelta empiricamente in base al contesto.

Capitolo 3

Face morphing con Arc2Face

3.1 Definizione matematica del face morphing

Un *morphing attack* consiste nella generazione di un'immagine facciale artificiale ottenuta combinando due identità distinte, con l'obiettivo di produrre un volto che venga accettato dal sistema biometrico come appartenente a entrambe.

La generazione può avvenire mediante trasformazioni geometriche basate sull'individuazione dei facial landmark, ossia punti di riferimento anatomici del volto (come occhi, naso e bocca), che vengono allineati e interpolati tra le due identità prima di essere fusi nello spazio dei pixel con l'obiettivo di produrre un'immagine visivamente coerente e plausibile.

Più recentemente, l'introduzione del deep learning ha portato allo sviluppo di approcci deep learning-based, che non si limitano alla fusione geometrica delle immagini, ma sfruttano modelli neurali per generare o ottimizzare direttamente la rappresentazione facciale. In questi sistemi l'immagine viene trasformata in una rappresentazione numerica astratta appresa dal modello durante l'addestramento.

Questo approccio risulta notevolmente efficace rispetto alle tecniche basate esclusivamente su manipolazioni geometriche o pixel-level, poiché permettono di agire direttamente sulla rappresentazione utilizzata dal sistema di riconoscimento che prende la decisione. Infatti, i modelli di face recognition basati su deep learning non confrontano le immagini a livello di pixel, ma valutano la somiglianza tra vettori all'interno dello spazio appreso dal modello.

In altre parole, un attacco guidato dagli embedding non mira unicamente a produrre un'immagine visivamente plausibile, ma a generare un volto che si collochi in una regione dello spazio delle feature compatibile con più identità. In questo modo, l'immagine morphed può risultare sufficientemente simile, secondo la metrica adottata dal modello, sia all'identità di partenza sia a quella con cui viene combinata, aumentando la probabilità di superare con successo la fase di verifica.

Siano z_C e z_A gli embedding normalizzati associati ai due volti sorgenti (ad esempio *criminal* e *accomplice*). L'obiettivo dell'attacco è ottenere un embedding intermedio z_M , che rappresenti una combinazione delle due identità e risulti sufficientemente vicino a entrambi nello spazio utilizzato dal sistema di face recognition.

Tale combinazione può essere espressa tramite una funzione di interpolazione parametrizzata da $t \in [0, 1]$, che controlla la posizione del punto generato tra le identità C e A .

In letteratura vengono comunemente utilizzati due approcci:

- **Interpolazione Lineare (LERP)**

L'interpolazione lineare calcola un nuovo punto che si trova lungo la linea retta che congiunge due punti noti nello spazio vettoriale. Formalmente il punto interpolato tra due vettori sorgenti C e A viene calcolato come:

$$\text{Lerp}(z_C, z_A; t) = (1 - t)z_C + tz_A$$

dove $t \in [0, 1]$ determina quanto l'embedding risultante sia più vicino a una delle due identità. In particolare:

- Se $t = 0$, allora $z_M = z_C$
- Se $t = 1$, allora $z_M = z_A$
- Se $t = 0.5$, si ottiene un embedding intermedio tra z_C e z_A

Il principale limite dell'interpolazione lineare è che funziona correttamente solo in uno spazio lineare. Tuttavia, nei moderni sistemi di face recognition come ArcFace, gli embedding vengono solitamente normalizzati e distribuiti

sulla superficie di un'ipersfera unitaria. Applicare una combinazione lineare tra due embedding produce in genere un punto che cade all'interno della sfera, cioè fuori dalla superficie su cui si distribuiscono normalmente le rappresentazioni facciali.

Geometricamente, la LERP non segue la curva della sfera 3.1, ma si muove lungo la corda che collega i due punti. Questo introduce un errore geometrico, perché l'interpolazione non rispetta la struttura dello spazio in cui il sistema misura la somiglianza tra identità [SR22].

Per ottenere una rappresentazione coerente con lo spazio geometrico utilizzato dai modelli di face recognition, risulta più appropriato utilizzare la funzione di SLERP che mantiene l'interpolazione sulla superficie dell'ipersfera.

- **Interpolazione Lineare Sferica (SLERP)**

La funzione SLERP tra due embedding normalizzati z_C e z_A è definita come:

$$\text{Slerp}(z_C, z_A, t) = \frac{\sin((1-t)\theta)}{\sin \theta} z_C + \frac{\sin(t\theta)}{\sin \theta} z_A$$

dove $t \in [0, 1]$ è il parametro di interpolazione e vale $\|A\| = 1$ e $\|C\| = 1$. L'angolo θ tra i due vettori è calcolato tramite:

$$\theta = \arccos(z_A \cdot z_C)$$

Questa formula è costruita in modo che il punto interpolato sia una combinazione pesata dei due vettori, ma con pesi tali da farlo restare sulla sfera. In particolare, i coefficienti dipendono dall'angolo θ e rappresentano quanto il vettore risultante si avvicina a ciascuna entità di partenza.

Questa formulazione garantisce una proprietà fondamentale: l'interpolazione avviene preservando la norma del vettore risultante. Questo significa che se gli embedding di partenza sono normalizzati, anche il vettore ottenuto tramite SLERP risulta normalizzato. Di conseguenza, l'embedding interpolato appartiene ancora alla superficie dell'ipersfera unitaria, mantenendo coerenza con la geometria dello spazio delle feature.

La SLERP è definita in modo tale che il parametro di interpolazione t controlli direttamente la posizione del punto lungo l'arco. Al crescere di t , l'angolo tra il vettore iniziale e il vettore interpolato aumenta in maniera proporzionale. In altre parole, se θ rappresenta l'angolo complessivo tra i due vettori estremi, SLERP costruisce il punto interpolato ad una distanza angolare pari a $t\theta$ dal primo vettore.

Queste caratteristiche sono particolarmente rilevanti in contesti come face recognition, in cui la similarità tra embedding viene valutata tramite cosine similarity e quindi dipende dall'angolo tra i due vettori.

Sebbene SLERP sia teoricamente più corretta per interpolare vettori sulla superficie di una sfera, presenta un limite legato alla stabilità numerica. In particolare, la formula di SLERP contiene il termine $\frac{1}{\sin(\theta)}$ e, quando θ assume un valore molto piccolo, il calcolo può diventare instabile. Infatti, se l'angolo θ tende a zero, di conseguenza anche $\sin(\theta)$ tende a zero e la divisione per un valore prossimo allo zero amplifica gli errori numerici [Gus06].

In questa situazione, SLERP tende a comportarsi in modo molto simile a una normale interpolazione lineare, e per questo motivo si utilizza LERP come approssimazione, perché più stabile e computazionalmente meno costosa.

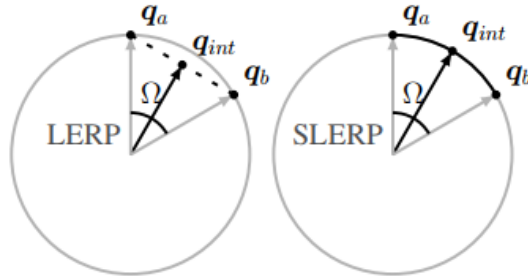


Figura 3.1: Confronto geometrico tra LERP e SLERP

Una volta ottenuto l'embedding interpolato M , per completare la pipeline di morphing è necessario convertirlo in un'immagine facciale. I modelli di face recognition, come ArcFace, sono progettati per estrarre embedding a partire da immagini, ma non permettono di ricostruire direttamente il volto nello spazio dei pixel, poiché questa trasformazione non è invertibile.

Per effettuare questo passaggio vengono impiegati modelli generativi, cioè architetture in grado di sintetizzare immagini realistiche a partire da una rappresentazione latente. In particolare, Arc2Face utilizza come input gli embedding prodotti da ArcFace, che rappresentano una descrizione compatta dell'identità facciale. Tali embedding vengono impiegati come informazione guida per il modello generativo, che sintetizza un volto realistico tale che, una volta analizzato nuovamente da ArcFace, produca un embedding coerente con quello fornito in input.

In questo modo, l'embedding M , ottenuto dall'interpolazione tra due identità, può essere trasformato in un'immagine morphed che combina caratteristiche biometriche di entrambi i soggetti.

3.2 Rimozione del background per ICAO compliance

L'*International Civil Aviation Organization* (ICAO) considera la foto facciale un elemento essenziale del documento, poiché collega fisicamente e biometricamente il documento al suo titolare.

Per questo motivo, la fotografia del volto presente sui documenti di identità elettronici deve essere una rappresentazione standardizzata, coerente e confrontabile, tale da garantire interoperabilità internazionale. ICAO definisce quindi una serie di requisiti che le immagini facciali devono rispettare, non solo per consentire l'identificazione visiva da parte di un operatore umano, ma soprattutto per assicurare il corretto riconoscimento biometrico da parte dei sistemi di face recognition.

La conformità delle immagini viene definita attraverso vincoli descritti nello standard ISO 39794-5 e nelle linee guida ICAO, formalizzati in una serie di test che verificano aspetti geometrici e fotografici dell'immagine [FFMM12a].

- **Vincoli geometrici:** requisiti relativi alla *dimensione* del volto e alla sua *posizione* all'interno dell'immagine. Tali vincoli non sono espressi in modo qualitativo, ma vengono definiti attraverso misure geometriche standardizza-

te, in modo da garantire un corretto inquadramento e una rappresentazione coerente del soggetto.

- **Vincoli fotometrici:** requisiti che riguardano la qualità fotografica dell'immagine e la corretta postura del soggetto durante l'acquisizione. L'obiettivo è garantire che il volto sia chiaramente visibile e riconoscibile, evitando condizioni che possano compromettere l'identificazione biometrica.

La seconda categoria risulta particolarmente complessa da valutare poiché comprende requisiti che non possono essere descritti unicamente tramite misure geometriche oggettive, ma dipendono anche da condizioni qualitative e percettive difficili da formalizzare. Essa include infatti caratteristiche che un osservatore umano è in grado di valutare con relativa facilità, mentre un sistema automatico deve necessariamente tradurre tali valutazioni in parametri numerici misurabili. Inoltre, la qualità fotografica intesa in senso tradizionale (ad esempio nitidezza o assenza di rumore) non coincide sempre con la qualità richiesta dallo standard ISO/ICAO, su cui invece influiscono altri aspetti come illuminazione, ombre, riflessi, sfondo e visibilità dei tratti facciali.

Uno dei requisiti di questa categoria è la valutazione dello sfondo. Lo sfondo non è solo un elemento estetico, ma ha un impatto diretto sulla qualità biometrica della fotografia.

Lo standard richiede che lo sfondo sia uniforme e privo di elementi che possano interferire con la visibilità del volto, poiché la presenza di dettagli o variazioni possono ridurre la separazione tra soggetto e ambiente e compromettere le procedure automatiche di analisi. La qualità dello sfondo risulta particolarmente rilevante non solo per garantire il corretto rilevamento e la corretta segmentazione del volto, ma anche perché influisce direttamente nella valutazione di altri requisiti fotografici. Uno sfondo non uniforme può rendere difficile distinguere correttamente ombre, difetti e artefatti, che possono essere interpretati e valutati erroneamente dai sistemi di verifica automatica.

Di conseguenza, uno sfondo controllato e omogeneo rappresenta un prerequisito importante per ridurre ambiguità e migliorare la robustezza complessiva dei test di conformità.

3.2.1 Descrizione dei diversi metodi testati

Per ottenere immagini conformi ai requisiti relativi allo sfondo previsti dallo standard ISO/ICAO, è stato necessario introdurre una fase di post-processing finalizzata alla normalizzazione del background. A tal fine, sono stati adottati modelli di *background removal*, ovvero sistemi basati su tecniche di segmentazione automatica in grado di separare il soggetto dallo sfondo e generare una maschera utile alla sua sostituzione con uno sfondo uniforme.

Questo passaggio risulta particolarmente delicato, poiché una rimozione non accurata può introdurre artefatti visivi, ad esempio bordi irregolari lungo il profilo del volto oppure perdita di dettagli come capelli o orecchie, che possono compromettere la naturalezza dell'immagine e, di conseguenza, la sua conformità agli standard. Per questo motivo è stata posta particolare attenzione nella scelta di modelli in grado di garantire una segmentazione precisa e stabile, riducendo al minimo errori lungo i contorni e possibili ambiguità, senza però eliminare informazioni essenziali per l'identificazione biometrica.

Nei moderni sistemi di segmentazione basati su deep learning, la maggior parte delle architetture segue una struttura generale di tipo *encoder-decoder*.

L'immagine di input viene inizialmente elaborata da una rete di codifica, detta anche *backbone* o *encoder*, che ha il compito di estrarre una rappresentazione semantica significativa della scena. L'immagine viene analizzata mediante una sequenza di livelli convoluzionali che apprendono feature gerarchiche: nei primi strati vengono rilevati pattern semplici e generici, mentre negli strati più profondi la rete apprende rappresentazioni sempre più complesse e semantiche. Durante questo processo, l'encoder tende progressivamente a ridurre la risoluzione spaziale dell'immagine (*downsampling*), questo permette di diminuire il costo computazionale e, soprattutto, di aumentare il *receptive field* dei neuroni, consentendo alla rete di catturare contesti più ampi e informazioni globali utili per riconoscere correttamente gli oggetti presenti nella scena.

Successivamente, una rete di decodifica proietta tali informazioni nello spazio originale dei pixel, riportando la rappresentazione alla risoluzione dell'immagine di partenza (*upsampling*). Il risultato finale prodotto dal decoder è generalmente una mappa di valori reali, detta mappa di *logits*, che costituisce una rappresentazione

intermedia non ancora direttamente interpretabile.

Infine, tali valori vengono trasformati in output significativo, sotto forma di probabilità P , da cui si ottiene la maschera di segmentazione utilizzata per separare il soggetto dallo sfondo.

In questo contesto, la natura della mappa prodotta dipende dal tipo di modello utilizzato. Nei modelli di *semantic segmentation* i logits rappresentano punteggi associati alle classi e vengono convertiti in una maschera discreta. Nei modelli di *alpha matting*, invece, l'output corrisponde a una stima continua della trasparenza (*alpha matte*), che descrive la porzione di foreground presente in ciascun pixel, consentendo transizioni più graduali tra soggetto e sfondo.

Sulla base di tale distinzione, sono stati analizzati diversi modelli appartenenti a entrambe le categorie.

Modelli di semantic segmentation

Nei modelli di *semantic segmentation*, la rimozione dello sfondo viene trattata come un problema di classificazione a livello di pixel. Il modello assegna a ogni pixel p una classe e produce quindi una maschera discreta, in cui ogni pixel appartiene a una sola categoria. Il valore assegnato a ogni pixel può assumere soltanto due possibili stati:

$$P(p) \in \{0, 1\}$$

In base al numero di categorie previste nel processo di classificazione, si distinguono due principali paradigmi: binary segmentation e multi-class segmentation.

- **Binary segmentation**

La *binary segmentation* è una forma di segmentazione in cui l'obiettivo è distinguere soltanto due categorie: *foreground* e *background*. In questo caso, a ciascun pixel dell'immagine viene assegnata una probabilità di appartenenza al soggetto, da cui si ottiene una maschera binaria finale.

Nel progetto sono stati testati modelli come **U²-Net** [QZH⁺20], utilizzato nella variante addestrata per la segmentazione di soggetti umani, noto per la buona capacità di estrarre dettagli fini grazie a una struttura encoder-decoder annidata, **SAM2** [RGH⁺24], basato su un approccio di segmenta-

zione *prompt-based*, in cui la predizione viene guidata tramite input quali punti o bounding box, permettendo di ottenere maschere accurate anche su categorie non viste durante l'addestramento. Nel lavoro, i prompt forniti a SAM2 sono generati automaticamente a partire dalle bounding box prodotte dal modello di object detection YOLO. In questo modo, il sistema combina le capacità di localizzazione di YOLO con la precisione di segmentazione di SAM2, realizzando una pipeline integrata di rilevamento e segmentazione.

- **Multi-class segmentation**

La *multi-class segmentation* estende l'approccio utilizzato nella binary segmentation introducendo più classi possibili. In questo caso ogni pixel viene assegnato a una tra C categorie differenti, producendo una segmentazione più dettagliata rispetto alla semplice distinzione foreground-background.

Sebbene l'obiettivo finale del progetto sia l'isolamento del soggetto dallo sfondo, l'impiego di un modello multi-classe, con classi specifiche per le componenti facciali, permette di preservare le informazioni strutturali del volto, evitando la perdita di dettagli che potrebbe verificarsi adottando direttamente una segmentazione binaria. In fase successiva all'identificazione le diverse classi del volto vengono unite in un'unica classe di foreground, riconducendo il problema a una rappresentazione binaria funzionale all'obiettivo.

Tra i modelli testati rientrano **SegFormer** [XWY⁺21], basato su un backbone Transformer e reso specializzato sul volto tramite addestramento su un dataset facciale, e **BiSeNet** [YWP⁺18], con backbone ResNet [GGEO⁺20], progettato per combinare informazioni di dettaglio spaziale e contesto globale, utilizzando ResNet come estrattore di caratteristiche per produrre una mappa di segmentazione multi-classe a livello di pixel.

Modelli di alpha matting

Nei modelli di *alpha matting* la separazione non viene vista come una classificazione, ma come una stima continua. Il modello produce una mappa di trasparenza (*alpha matte*), in cui ogni pixel p assume un valore $P(p)$ compreso tra 0 e 1, che

rappresenta la porzione di foreground presente in quel pixel:

$$P(p) \in [0, 1]$$

In questo contesto sono stati valutati modelli come **MODNet** [HG], progettato specificamente per il portrait matting in tempo reale, e **BEN2** [MS25], utilizzato per ottenere una separazione più precisa nelle zone di transizione tra soggetto e sfondo

3.2.2 Confronto visivo e commenti

Definite le basi teoriche e le principali differenze tra gli approcci considerati, analizziamo i risultati ottenuti dai modelli selezionati, con l'obiettivo di valutarne in maniera concreta l'efficacia rispetto agli obiettivi del progetto.

L'analisi è guidata da una doppia esigenza. Da un lato si vuole garantire un'elevata accuratezza visiva della segmentazione. Dall'altro lato la segmentazione non deve alterare tratti distintivi del volto al punto da modificare il comportamento dei sistemi di riconoscimento facciale.

Questo significa che la maschera generata deve isolare correttamente il soggetto, preservandone struttura, contorni e dettagli significativi, evitando sia l'inclusione di porzioni di sfondo (*false positive*), sia l'eliminazione di parti del volto (*false negative*), che comporterebbero una perdita di qualità e di informazioni rilevanti.

Per verificare che la segmentazione non incida sul comportamento dei sistemi di riconoscimento facciale per ogni modello è stata calcolata la relativa *Morph Acceptance Probability (MAP) matrix*, utilizzata per misurare la probabilità che un'immagine morphed venga accettata dai sistemi biometrici considerati. La MAP rappresenta una matrice di probabilità che descrive la percentuale di immagini morph in grado di superare le soglie di accettazione dei modelli di face recognition.

In particolare, le righe della matrice corrispondono al numero di immagini disponibili per ciascun soggetto coinvolto nel morph, mentre le colonne rappresentano il numero di sistemi biometrici che devono simultaneamente accettare l'immagine affinché l'attacco sia considerato riuscito.

Nel contesto di questa analisi, la MAP viene calcolata utilizzando quattro modelli di face recognition: MagFace, AdaFace, CurricularFace e SwinFace.

Il confronto tra i valori ottenuti sulle immagini originali e quelli calcolati sulle immagini segmentate consente di verificare se l'operazione di rimozione dello sfondo alteri la probabilità di accettazione dei morph.

Ogni matrice di probabilità può essere così interpretata:

	≥ 1	≥ 2	≥ 3	4 Sistemi
1 Immagine	$P(1, 1)$	$P(1, 2)$	$P(1, 3)$	$P(1, 4)$
2 Immagini	$P(2, 1)$	$P(2, 2)$	$P(2, 3)$	$P(2, 4)$

Tabella 3.1: Interpretazione della Morph Acceptance Probability (MAP)

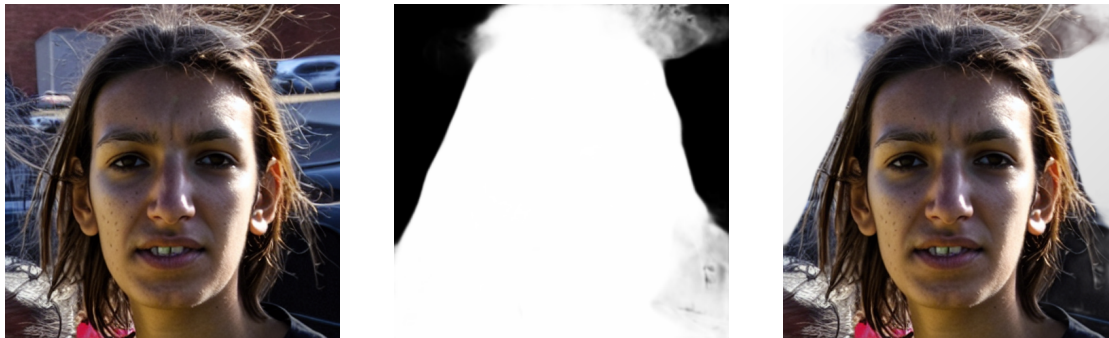
Dove ogni elemento $P(a, f)$ rappresenta la probabilità che un'immagine morph venga accettata da almeno f sistemi biometrici con a immagini disponibili per soggetto.

La matrice 3.2 rappresenta la MAP delle immagini morphed originali e viene utilizzata come riferimento di base per il confronto con i risultati ottenuti dai diversi modelli analizzati.

	1	2	3	4
1	1.0	1.0	0.998	0.928
2	0.995	0.984	0.978	0.797

Tabella 3.2: MAP immagini morphed originali

U²-Net



(a) Immagine originale

(b) Maschera binaria

(c) Immagine segmentata

Figura 3.2: Esempio di segmentazione ottenuta con U²-Net

	1	2	3	4
1	0.998	0.998	0.996	0.933
2	0.994	0.981	0.974	0.796

Tabella 3.3: MAP per U²-Net

Rispetto alla MAP di riferimento, i valori ottenuti con U²-Net mostrano scostamenti minimi. La probabilità di accettazione rimane sostanzialmente invariata, in particolare nella cella più restrittiva (2,4), indicando che la segmentazione non altera in modo significativo il comportamento biometrico dei morph.

Dal punto di vista qualitativo, trattandosi di un modello di segmentazione generica, il rilevamento del soggetto risulta complessivamente corretto ma poco preciso nei dettagli fini come contorni e capelli.

Sam2

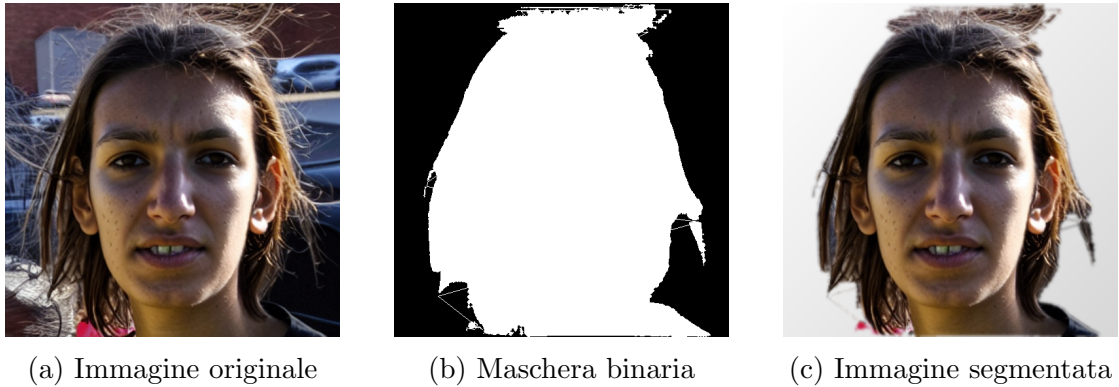


Figura 3.3: Esempio di segmentazione ottenuta con Sam2

	1	2	3	4
1	0.982	0.981	0.98	0.916
2	0.979	0.967	0.962	0.798

Tabella 3.4: MAP per Sam2

Nel caso di Sam2 si osserva una lieve riduzione dei valori rispetto alla MAP di riferimento, più evidente nella prima riga, ma comunque contenuta.

Qualitativamente, la segmentazione risulta più coerente nella separazione foreground/background rispetto a U²-Net, ma permane una difficoltà nella gestione dei dettagli sottili, come le ciocche di capelli.

SegFormer



Figura 3.4: Esempio di segmentazione ottenuta con SegFormer

	1	2	3	4
1	1.0	0.999	0.998	0.935
2	0.998	0.984	0.977	0.809

Tabella 3.5: MAP per SegFormer

Con SegFormer, modello addestrato per il face parsing, si osserva una segmentazione più precisa nell'individuazione delle regioni facciali. I valori della MAP risultano molto vicini a quella di riferimento, con leggere oscillazioni che non evidenziano una riduzione significativa della probabilità di accettazione.

Notiamo un miglioramento qualitativo rispetto ai modelli precedenti, tuttavia anche in questo caso la segmentazione dei capelli rimane poco accurata, con contorni netti o semplificati.

BiSeNet

Figura 3.5: Esempio di segmentazione ottenuta con BiSeNet

	1	2	3	4
1	1.0	0.999	0.997	0.926
2	0.993	0.984	0.978	0.814

Tabella 3.6: MAP per BiSeNet

Analogamente a SegFormer, BiSeNet mostra una buona capacità di identificazione delle regioni del volto, confermando l'efficacia dei modelli di face parsing rispetto a quelli generici. I valori della MAP restano molto simili a quelli delle immagini originali.

Nonostante la migliore localizzazione del volto, la gestione delle aree complesse come i capelli non risulta ancora sufficientemente dettagliata.

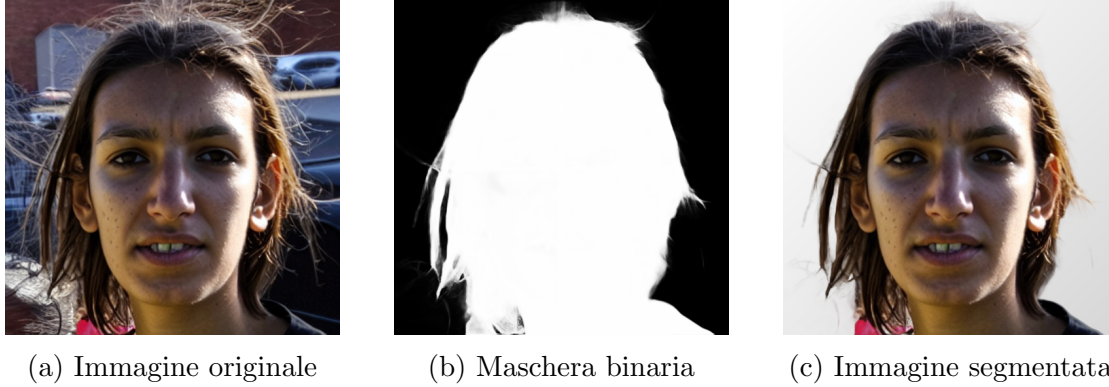
BEN2

Figura 3.6: Esempio di segmentazione ottenuta con BEN2

	1	2	3	4
1	1.0	0.999	0.997	0.934
2	0.995	0.984	0.976	0.809

Tabella 3.7: MAP per BEN2

Con il passaggio a BEN2, appartenente alla categoria dei modelli di alpha matting, si osserva un netto miglioramento qualitativo nella resa visiva. La segmentazione risulta più naturale e precisa, in particolare nei dettagli complessi come i capelli e i contorni sottili.

Dal punto di vista quantitativo, la MAP rimane sostanzialmente allineata alla baseline. Si osserva persino un lieve incremento nella cella più restrittiva (2,4).

ModNet

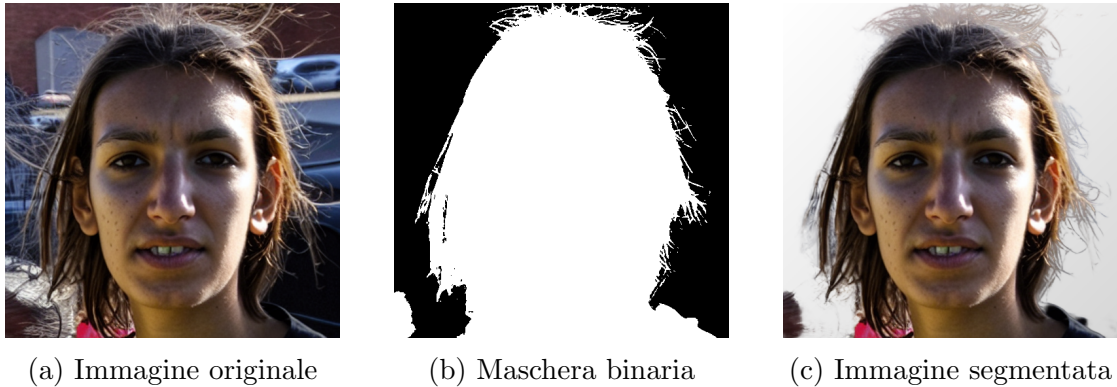


Figura 3.7: Esempio di segmentazione ottenuta con ModNet

	1	2	3	4
1	1.0	0.999	0.998	0.941
2	0.998	0.985	0.978	0.804

Tabella 3.8: MAP per ModNet

ModNet conferma ulteriormente i vantaggi dell'approccio di alpha matting. La transizione tra soggetto e sfondo appare più graduale e i dettagli fini, in particolare i capelli, risultano meglio preservati rispetto ai modelli di segmentazione binaria.

Anche in questo caso, i valori della MAP rimangono molto vicini a quelli delle immagini originali.

Capitolo 4

Face demorphing con Arc2Face

4.1 Definizione del problema inverso: il demorphing

Nei contesti di controllo automatico ai confini (*Automated Border Control*) un attacco biometrico può compromettere la validità del sistema. Un'immagine facciale memorizzata nel documento può essere stata generata combinando due identità distinte, producendo un volto artificiale che risulta compatibile con entrambi i soggetti coinvolti.

Un approccio per affrontare questo problema consiste nel sfruttare la disponibilità simultanea di due immagini durante la fase di verifica: la fotografia presente nel documento M e una fotografia acquisita dal vivo \tilde{C} . In questo scenario, è possibile tentare di invertire il processo di morphing, con l'obiettivo di separare le identità combinate nell'immagine del documento e individuare eventuali anomalie riconducibili a un morphing attack.

In una prima approssimazione lineare, l'immagine memorizzata sul documento può essere rappresentata come combinazione delle due identità sorgenti:

$$M \approx A + C$$

mentre nel caso di un documento genuino, l'immagine può essere interpretata come

una combinazione di due copie della stessa identità:

$$M \approx C + C$$

Durante la verifica, si calcola quindi un'immagine *demorphed* D sottraendo dall'immagine del documento la fotografia live acquisita al controllo:

$$D = M - \tilde{C}$$

A questo punto, il comportamento del sistema dipende dal fatto che il documento sia genuino o morphed.

Se il documento è genuino, allora la sottrazione rimuove una componente compatibile con l'identità reale, lasciando un'immagine ancora simile al soggetto presente al controllo ($D \approx C$). In questo caso, il confronto biometrico tra D e \tilde{C} produce un punteggio di similarità elevato.

Al contrario, se l'immagine del documento è un morph tra le due identità, la sottrazione dell'immagine live elimina principalmente la componente corrispondente a C , lasciando una rappresentazione più vicina all'accomplice ($D \approx A$). Di conseguenza, il confronto tra D e \tilde{C} produce un punteggio di similarità basso, fornendo un'indicazione della presenza di un morphing attack.

Questo approccio consente quindi di rilevare il morphing sfruttando l'informazione fornita dall'immagine live acquisita durante il controllo [FFM17].

Tale modello, pur essendo intuitivo, non riflette il comportamento reale dei moderni sistemi biometrici. Nei sistemi di face recognition basati su deep learning, infatti, il morphing non avviene nello spazio dei pixel, ma nello spazio degli embedding. L'immagine morphed è il risultato di un'interpolazione tra vettori identificativi normalizzati, e la decisione biometrica viene presa confrontando tali rappresentazioni nello spazio metrico appreso. Ricostruire le identità sorgenti a partire dall'immagine morphed non equivale quindi a eseguire una semplice sottrazione diretta, ma a ricostruire un embedding complementare nello spazio delle feature.

L'obiettivo è stimare l'embedding dell'identità mancante, ovvero il volto sorgente C , assumendo che l'immagine morphed sia stata generata tramite un'interpolazione.

zione tra due identità. In altre parole, a partire dall'embedding della fotografia del documento z_M e dall'embedding della fotografia acquisita al controllo $z_{\tilde{C}}$, si vuole produrre un vettore che rappresenti l'identità complementare rispetto a quella osservata al controllo, cioè una possibile ricostruzione dell'altro soggetto coinvolto nel morphing.

Una volta stimato tale embedding, è necessario ricondurlo nuovamente nello spazio delle immagini per poter effettuare un confronto visivo o biometrico coerente. Per questo passaggio viene impiegato un modello generativo compatibile con lo spazio degli embedding di ArcFace, nello specifico Arc2Face. Questo modello è progettato per sintetizzare un'immagine facciale a partire da un embedding identitario, preservandone le caratteristiche biometriche. In tal modo, l'embedding ricostruito tramite inversione può essere trasformato in un volto sintetico coerente, il cui embedding, ricalcolato mediante il modello di face recognition, può essere confrontato con quello della fotografia live.

Equivalentemente al caso lineare semplificato, un documento può essere considerato genuino se l'embedding ricostruito risulta simile all'embedding associato alla fotografia \tilde{C} scattata ai controlli. Al contrario, in presenza di un morphing attack, l'operazione di inversione tende a produrre un embedding più vicino alla seconda identità sorgente A , cioè all'*accomplice*.

4.2 Inversione approssimata della SLERP

Assumendo che l'immagine morphed sia il risultato esatto di un'interpolazione sferica tra due vettori normalizzati, il demorphing può essere visto come un problema inverso: dato il vettore interpolato e uno dei due estremi, si vuole ricavare l'altro vettore sorgente.

Tuttavia la funzione di interpolazione sferica non è una trasformazione invertibile nel senso classico. In generale, un'operazione è invertibile quando esiste una funzione inversa unica che, a partire dall'output, permette di ricostruire esattamente gli input originali. Nel caso del morphing, invece, la combinazione di due identità produce un risultato intermedio che non conserva tutta l'informazione necessaria a distinguere in modo univoco le due componenti.

Avendo a disposizione l'embedding associato a una delle due identità coinvolte nel processo di morphing, è possibile formulare una stima dell'eventuale seconda identità che ha contribuito alla generazione dell'immagine morphed.

Assumiamo che l'embedding morphed z_M sia stato generato esattamente tramite interpolazione sferica tra due vettori normalizzati z_C e z_A , secondo un parametro $t \in [0, 1]$:

$$z_M = \text{Slerp}(z_A, z_C, t)$$

Ipotizziamo che l'identità osservata al controllo corrisponda a C , e che quindi possiamo risalire attraverso modelli di face recognition all'embedding z_C , estratto dalla fotografia live. L'obiettivo è quindi identificare la seconda entità coinvolta A , e quindi il suo corrispondente vettore z_A .

Dal punto di vista geometrico, SLERP genera un punto lungo l'arco di cerchio massimo che collega z_C e z_A sull'ipersfera unitaria. Se ω rappresenta l'angolo tra z_C e z_A , il punto z_M si trova ad una distanza angolare pari a $(1 - t)\omega$ da z_C . Di conseguenza è possibile ricavare ω a partire dall'angolo tra z_C e z_M :

$$\omega = \frac{\arccos(z_C \cdot z_M)}{(1 - t)}$$

Sostituendo tale valore nella definizione della SLERP 3.1 e isolando z_C si ottiene una stima dell'embedding complementare:

$$z_A = \frac{z_M - az_C}{b}$$

dove

$$a = \frac{\sin((1 - t)\omega)}{\sin(\omega)}, b = \frac{\sin(t\omega)}{\sin(\omega)}$$

Questo equivale a prolungare l'arco che unisce z_M e z_C fino a ricostruire il punto finale z_A .

Nel caso in cui si fosse a conoscenza dell'identità A , il problema è del tutto analogo, ricavando z_C a partire da z_A e z_M e dal medesimo parametro di interpolazione t , semplicemente scambiando i ruoli dei due vettori estremi nella relazione di SLERP.

La stima dell'embedding mancante dipende in modo diretto dal parametro t

che controlla la posizione di z_M lungo l'arco che unisce z_C e z_A , e quindi determina quanto l'arco debba essere prolungato per raggiungere l'estremo mancante. In uno scenario reale, t non è noto e non può essere dedotto univocamente dai soli embedding disponibili, poiché dipende dal processo con cui il morph è stato generato. Per questo motivo, nelle analisi successive si assume $t = 0.5$, corrispondente a un contributo bilanciato tra le due identità.

Capitolo 5

Prove sperimentali

5.1 Dataset e modelli utilizzati

5.1.1 Dataset utilizzati

Chicago Face Database (CFD)

Il *Chicago Face Database* (CFD) è un database di immagini facciali ad alta risoluzione sviluppato da Ma, Correll e Wittenbrink (2015), comprendente fotografie controllate e un insieme di valutazioni soggettive e oggettive delle caratteristiche morfologiche del volto. Il dataset dispone di centinaia di volti maschili e femminili appartenenti a differenti gruppi etnici.

Tutte le fotografie sono state scattate in ambiente controllato: con distanza fissa tra soggetto e fotocamera, altezza della camera allineata al livello degli occhi, sfondo bianco uniforme e sistema di illuminazione a tre punti per ridurre al minimo ombre e variazioni luminose. I soggetti sono ritratti in posa frontale, con testa verticale e sguardo diretto verso l'obiettivo. Inoltre, le fotografie sono state successivamente ritagliate e normalizzate in modo da uniformare le proporzioni dei principali tratti facciali [MCW15].

Queste caratteristiche risultano coerenti con i requisiti internazionali ISO/IEC per l'acquisizione di immagini facciali destinate a sistemi biometrici e di conseguenza adatte alla simulazione della generazione di immagini morphed.

FEI Face Database

Il *FEI Face Database* è un dataset di immagini facciali sviluppato presso l'Artificial Intelligence Laboratory della FEI University, comprendente fotografie relative a 200 individui distinti maschili e femminili.

Tutte le immagini sono state acquisite in ambiente controllato: sfondo bianco uniforme, illuminazione relativamente costante, risoluzione standard.

Per ogni soggetto sono disponibili 14 immagini che lo ritraggono con diverse pose del capo e differenti espressioni facciali. Questa caratteristica consente di simulare in modo realistico uno scenario di Automatic Border Control (ABC). Nei sistemi ABC, l'identità del viaggiatore viene verificata confrontando un'immagine facciale scattata in tempo reale al gate con l'immagine digitale memorizzata nel documento di viaggio elettronico (eMRTD).

Questo scenario viene riprodotto sfruttando la struttura del dataset:

- L'immagine del documento viene generata a partire dalle fotografie frontali del database, conformi agli standard ISO/IEC per immagini facciali biometriche.

Nel caso di immagini bona fide, l'immagine del documento coincide con una fotografia frontale genuina del soggetto. Invece, nel caso di attacco morphing, l'immagine del documento viene creata mediante algoritmi di morphing a partire da due soggetti distinti.

- L'immagine acquisita in tempo reale viene rappresentata da un'altra fotografia dello stesso soggetto, oppure, nel caso di attacco, di uno dei due soggetti utilizzati per generare il morph, ma caratterizzata da una diversa posa o espressione facciale.

5.1.2 Modelli della pipeline generativa

InsightFace

InsightFace è un framework per il riconoscimento facciale che fornisce modelli pre-addestrati per face detection ed estrazione degli embedding biometrici. In particolare, è stata utilizzata la configurazione "antelopev2", per la rilevazione del volto delle immagini di partenza e per la successiva estrazione del vettore

che ne rappresenta l'identità. Gli embedding estratti costituiscono la base per l'interpolazione dell'identità morphed. Inoltre, tale rappresentazione permette di effettuare confronti quantitativi tra identità, basati su metriche numeriche, al fine di verificarne la corrispondenza.

Arc2Face

Arc2Face è un modello generativo basato su *Stable Diffusion*, adattato per generare immagini facciali sintetiche a partire esclusivamente da embedding biometrici di identità, come quelli prodotti da ArcFace.

Il modello di generazione Stable Diffusion permette di ricostruire gradualmente l'immagine partendo da una descrizione testuale. Tale descrizione viene elaborata da un text encoder basato su CLIP (*Contrastive Language-Image Pretraining*), che trasforma la sequenza di token testuali in una rappresentazione vettoriale densa nello spazio latente del modello. Questa rappresentazione fornisce al generatore informazioni su cosa deve essere prodotto. Arc2Face modifica questo meccanismo di condizionamento. Invece di utilizzare una descrizione testuale, impiega l'embedding biometrico che rappresenta l'identità del soggetto. Arc2Face proietta quindi l'embedding nello stesso spazio numerico in cui si trovano gli embedding testuali di CLIP, permettendo al modello di diffusione di interpretarlo come segnale guida per la generazione.

EMOCA

EMOCA è un modello di ricostruzione facciale 3D che, a partire da un'immagine 2D, stima una rappresentazione parametrica del volto. Partendo da un'immagine in input il modello ricostruisce la geometria tridimensionale del viso e genera una *normal map*, ovvero una rappresentazione che codifica la struttura e la posa del volto in modo indipendente dall'identità. In questo modo, il sistema dispone delle informazioni geometriche necessarie per garantire che la posa e l'inclinazione del volto rimangano coerenti con le immagini di input, conformi agli standard ISO/IEC, anche durante il processo di generazione.

ControlNet

ControlNet è un'estensione dei modelli di diffusione che permette di aggiungere un controllo sulla struttura dell'immagine durante la generazione. In particolare, la generazione dell'immagine finale è guidata da due tipologie di informazioni. Da un lato, intervengono le informazioni semantiche, rappresentate dagli embedding biometrici estratti mediante InsightFace. Dall'altro lato, la generazione è vincolata da informazioni geometriche, fornite dalla normal map prodotta da EMOCA. La combinazione di queste due componenti permette al sistema di gestire separatamente l'identità e la struttura del volto, ottenendo un'immagine finale coerente sia nell'identità sia nella struttura.

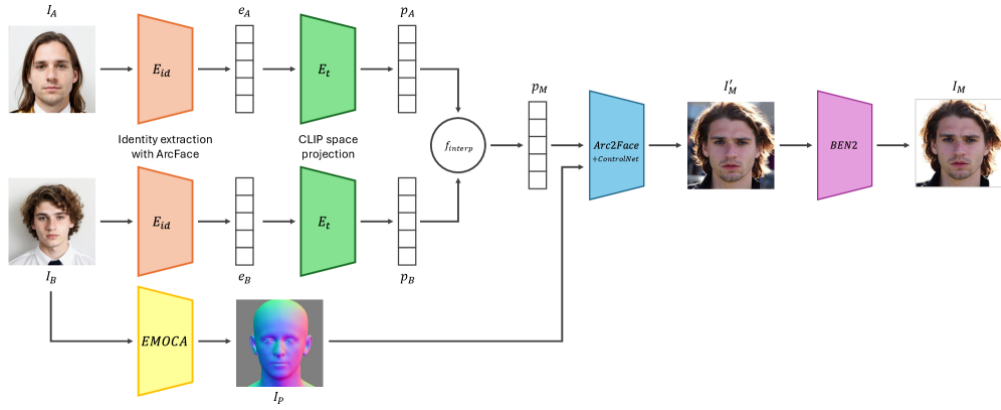


Figura 5.1: Sintesi della pipeline generativa [DDFFM26]

5.2 Metriche di valutazione

La similarità tra le immagini è stata valutata mediante il calcolo della distanza coseno 2.2 tra gli embedding biometrici estratti mediante il modello InsightFace. In questo spazio gli embedding possono essere valutati in base alla loro distanza sulla superficie della sfera unitaria su cui sono definiti. In altre parole, la somiglianza tra i due volti confrontati diminuisce all'aumentare dell'angolo misurato tra i vettori corrispondenti.

Poiché gli embedding sono L2-normalizzati, la distanza coseno $dist(v_0, v_1)$ assume valori compresi nell'intervallo $[0, 2]$, dove:

- $dist(v_0, v_1) \approx 0$, indica un'elevata similarità
- $dist(v_0, v_1) \approx 2$, indica un'elevata dissimilarità

5.2.1 Scenario "ideale" di morphing attack

Siano:

- z_A : embedding dell'accomplice
- z_C : embedding del criminale
- t : parametro di interpolazione uguale a 0.5
- z_M : embedding dell'immagine morphed del documento, ottenuto tramite interpolazione sferica

$$m = slerp(z_A, z_C, t)$$

- z_D : embedding dell'immagine demorphed, ottenuto tramite interpolazione inversa

$$d = inverse_slerp(z_C, z_M, t)$$

Nello scenario ideale si assume che il confronto venga effettuato utilizzando la stessa immagine di partenza z_A , ovvero lo stesso embedding impiegato nel processo di interpolazione. In questo caso, il vettore di riferimento coincide esattamente con quello utilizzato per generare il morph, eliminando qualsiasi variabilità dovuta a posa, illuminazione o condizioni di acquisizione. Di conseguenza, eventuali differenze sono dovute unicamente al comportamento delle operazioni di interpolazione e inversione nello spazio degli embedding.

In questo scenario ci aspetta:

- $dist(z_A, z_C) \approx 1$, indicando identità distinte poiché i due vettori sono ben separati nello spazio latente.
- $dist(z_A, z_M) \approx dist(z_C, z_M) \approx t \cdot dist(z_A, z_C)$, indicando una somiglianza parziale a entrambe le identità, poiché il vettore del morph si colloca geometricamente in posizione intermedia tra z_A e z_C .

- $dist(z_D, z_A) \approx 0$, indicando la stessa identità e quindi la corretta ricostruzione dell'identità dell'accomplice, poiché l'operazione di interpolazione inversa riporta il vettore nella posizione originaria di z_A .
- $dist(z_D, z_C) \approx 1$, indicando identità distinte e confermando che il vettore demorphed rimane separato dall'identità del criminale.

5.2.2 Scenario "reale" di morphing attack

Siano:

- z_A , embedding dell'accomplice
- z_C , embedding del criminale
- t , parametro di interpolazione uguale a 0.5
- z_M , embedding dell'immagine morphed del documento, ottenuto tramite interpolazione sferica

$$z_M = \text{slerp}(z_A, z_C, t)$$

- z_L , embedding dell'immagine live dell'accomplice, acquisita in condizioni operative reali
- z_D , embedding dell'immagine demorphed, ottenuto tramite interpolazione inversa

$$z_D = \text{inverse_slerp}(z_L, z_M, t)$$

Nello scenario reale il confronto avviene con un'immagine live L , che rappresenta lo stesso soggetto A , ma acquisita in condizioni differenti. L'embedding z_L rappresenta infatti la medesima identità di z_A , ma può differire a causa di variazioni di posa, illuminazione, espressione e condizioni di acquisizione. Di conseguenza, $z_L \neq z_A$, pur rappresentando la stessa identità. In questo caso, eventuali differenze sono dovute sia al comportamento delle operazioni di interpolazione e inversione, sia alla naturale variabilità dovuta alla presenza di artefatti.

In questo scenario ci si aspetta:

- $dist(z_A, z_C) \approx 1$, indicando identità distinte poiché i due vettori rimangono ben separati nello spazio latente.
- $dist(z_A, z_L)$ sia contenuta ma non nulla, indicando la stessa identità ma con variabilità dovuta alle differenti condizioni di acquisizione.
- $dist(z_D, z_A)$ sia bassa ma maggiore di zero, indicando una ricostruzione corretta dell'identità dell'accomplice, pur in presenza di variabilità reale. In particolare, si assume che valga la relazione $dist(z_D, z_A) < dist(z_D, z_L)$; tuttavia, la differenza tra i due valori è attesa essere contenuta, poiché z_A e z_L rappresentano la stessa identità, pur in presenza di variazioni dovute all'acquisizione live.
- $dist(z_D, z_C)$ rimanga elevata, indicando identità distinte e confermando la separazione rispetto all'identità del criminale.

5.2.3 Scenario documento bona fide

Siano:

- z_C , embedding del soggetto presente nel documento
- z_L , embedding dell'immagine live dello stesso soggetto presente nel documento
- z_D , embedding dell'immagine demorphed, ottenuto tramite interpolazione inversa

$$z_D = \text{inverse_slerp}(z_L, z_C, t)$$

In questo scenario il documento non contiene informazioni biometriche composite, poiché non è stato generato tramite morphing. Pertanto, non esiste un'identità da recuperare attraverso il processo di demorphing.

L'obiettivo dell'analisi è verificare che, in assenza di morphing, il metodo non produca una ricostruzione riconducibile a un'identità alternativa.

Ci si aspetta che:

- $dist(z_C, z_L)$ sia prossima a zero, indicando che i due embedding rappresentano la medesima identità. Un eventuale scostamento dallo zero è attribuibile alla variabilità intra-soggetto, dovuta a differenze di posa, espressione facciale o condizioni di illuminazione.

5.3 Protocollo

Per gli esperimenti è stato utilizzato il dataset FEI Face Database. Il dataset fornisce, per ciascun soggetto, immagini acquisite in condizioni controllate (cartella `doc`) conformi allo standard ISO/IEC per fotografie da documento, caratterizzate da espressione neutra, illuminazione uniforme e sfondo omogeneo.

Per gli stessi soggetti sono disponibili due immagini **live**, acquisite in condizioni meno controllate, che includono variazioni di posa, espressione facciale, etc.

Le coppie di soggetti utilizzate per il morphing sono specificate nel file `pairs.txt`, che definisce le combinazioni tra identità da fondere.

Il morphing è stato realizzato nello spazio degli embedding tramite interpolazione sferica (SLERP). Data una coppia di embedding normalizzati v_0 e v_1 , l'embedding morphed c è stato ottenuto come:

$$c = \text{slerp}(v_0, v_1, t)$$

dove il parametro di interpolazione è stato fissato a $t = 0.5$, corrispondente a una fusione simmetrica delle due identità.

L'interpolazione sferica garantisce che il vettore risultante rimanga sulla superficie dell'ipersfera unitaria, preservando la geometria dello spazio degli embedding.

Le immagini morphed generate seguono la convenzione di naming:

`M_<subj1>_<subj2>_<algorithm>_<factor>_<factor>_<postproc>`

dove `subj1` e `subj2` identificano i soggetti coinvolti nel morphing [DDBFM23].

Una volta generata l'immagine morphed tra due soggetti ID_1 e ID_2 , vengono selezionate le immagini **live** delle identità coinvolte.

Poiché per ciascun soggetto sono disponibili due immagini **live**, il processo di demorphing viene eseguito quattro volte per ogni morph:

- due volte utilizzando le immagini live di ID_1 ,
- due volte utilizzando le immagini live di ID_2 .

Il demorphing viene calcolato tramite l'inversione dell'interpolazione sferica. Dato:

$$c = \text{slerp}(v_0, v_1, t)$$

la procedura inversa consente di ricostruire v_1 a partire da v_0 , c e t , assumendo nuovamente $t = 0.5$.

Formalmente:

$$v_1 = \text{inverse_slerp}(v_0, c, t)$$

Il vettore ricostruito viene successivamente normalizzato per garantire la permanenza sulla superficie dell'ipersfera unitaria.

La qualità della ricostruzione viene valutata calcolando la distanza coseno tra l'embedding demorphed e gli embedding delle identità originali.

La distanza coseno 2.2 è definita come:

$$d(v_0, v_1) = 1 - \text{sim}(v_0, v_1)$$

Valori prossimi a zero indicano elevata similarità tra gli embedding, mentre valori prossimi a uno indicano identità differenti.

5.4 Analisi dei risultati

Scenario ideale



Tabella 5.1: Simulazione caso ideale

Sub1	Sub2	distance
ID1	ID2	0.8322289437055588
MORPH	ID1	0.23592734336853027
MORPH	ID2	0.2358245849609375
DEMORPH1	ID1	0.0000001192092896
DEMORPH2	ID2	5.960464477539063e-08

Tabella 5.2: Distanza coseno nel caso ideale

La tabella 5.2 riporta le distanze coseno calcolate tra gli embedding facciali dei soggetti coinvolti nel processo di morphing e demorphing. In particolare, vengono confrontati gli embedding delle identità originali (ID1 e ID2), dell'immagine morphed (MORPH) e delle immagini ricostruite tramite demorphing (DEMORPH).

Si osserva innanzitutto che la distanza coseno tra ID1 e ID2 è pari a 0.8322, valore elevato che conferma la significativa dissimilarità tra le due identità di partenza. Al contrario, le distanze tra MORPH e le due identità originali risultano pari a circa 0.236, indicando che l'immagine morphed si colloca, come atteso, in una posizione intermedia nello spazio degli embedding.

Le distanze coseno tra DEMORPH e l'identità corretta di riferimento sono dell'ordine di 10^{-7} , quindi prossime a zero. Questo indica che l'identità ricostruita coincide con quella originale. In particolare, nel caso di DEMORPH1 il processo di

demorphing è stato eseguito utilizzando la coppia (MORPH, ID2), con l'obiettivo di ricostruire ID1. La distanza prossima a zero tra DEMORPH1 e ID1 conferma la corretta ricostruzione dell'identità. Analogamente, DEMORPH2 è stato ottenuto a partire da (MORPH, ID1) e ricostruisce correttamente ID2, come dimostrato dalla distanza coseno anch'essa trascurabile.

Simulazione caso reale di morphing attack



Tabella 5.3: Simulazione di un caso reale assumendo ID1 come soggetto criminale

Sub1	Sub2	distance
DEMORPH	ID1	0.8329989165067673
DEMORPH	ID2	0.12894326448440552

Tabella 5.4: Distanza coseno nel caso reale simulato con il soggetto ID1

La Tabella 5.4 riporta i risultati ottenuti in uno scenario più realistico rispetto al caso ideale precedentemente analizzato. In questo esperimento, il processo di demorphing è stato eseguito utilizzando la coppia (DOC, LIVE), dove:

- DOC rappresenta l'immagine del documento generata tramite morphing a partire dalle identità ID1 e ID2;
- LIVE è un'immagine del soggetto ID1 acquisita in condizioni differenti (soriso)

L'obiettivo del demorphing, in questo contesto, è ricostruire l'identità complementare a LIVE che ha contribuito alla generazione di DOC, ovvero ID2. I risultati

mostrano che la distanza coseno tra l'embedding dell'immagine demorphed e ID2 è pari a circa 0.129. Questo indica che l'embedding ricostruito si colloca nello spazio delle caratteristiche in prossimità di ID2, confermando la corretta ricostruzione dell'identità attesa. A scopo di verifica, è stata inoltre calcolata la distanza coseno tra l'immagine demorphed e ID1. In questo caso si osserva un valore pari a circa 0.833, sostanzialmente equivalente alla distanza originaria tra ID1 e ID2 nel caso di confronto diretto.



Tabella 5.5: Simulazione di un caso reale assumendo ID2 come soggetto criminale

Sub1	Sub2	distance
DEMORPH	ID1	0.10461181402206421
DEMORPH	ID2	0.8428909927606583

Tabella 5.6: Distanza coseno nel caso reale simulato con il soggetto ID2

È stato inoltre analizzato il caso inverso, in cui l'immagine LIVE rappresenta il soggetto ID2 acquisito con un'espressione differente a quella utilizzata per la generazione del morph. Anche in questo scenario, il processo di demorphing è stato applicato alla coppia (DOC, LIVE), con l'obiettivo di ricostruire l'identità complementare presente nell'immagine morphed (ID1).

I risultati riportati in Tabella 5.6 risultano coerenti con quanto osservato nel caso precedente. In particolare, la distanza coseno tra l'embedding dell'immagine demorphed e ID1 è pari a circa 0.105, indicando una corretta ricostruzione dell'identità ID1.

Simulazione caso reale di documento bona fide


Tabella 5.7: Simulazione di un caso reale con documento bona fide del soggetto ID1



Tabella 5.8: Simulazione di un caso reale con documento bona fide del soggetto ID2

Sub1	Sub2	distance
DEMORPH1	ID1	0.14877784252
DEMORPH1	ID2	0.9234048128

Tabella 5.9: Distanza coseno nel caso reale bona fide col soggetto ID1

Sub1	Sub2	distance
DEMORPH2	ID1	0.855810821
DEMORPH2	ID2	0.10475689172

Tabella 5.10: Distanza coseno nel caso reale bona fide col soggetto ID2

In questo caso l'immagine del documento non è generato tramite morphing, ma rappresenta un caso bona fide. In tale scenario, si assume che:

- DOC rappresenta il documento appartenente realmente al soggetto ID1 5.7 / ID2 5.8
- LIVE è l'immagine dello stesso soggetto presente nel documento con espressione differente rispetto al documento

In questo contesto non è presente alcuna identità latente da separare. I risultati riportati nelle Tabelle 5.9 e 5.10 mostrano che l'embedding ottenuto dal demorphing risulta significativamente più vicino all'identità presente nel documento.

Capitolo 6

Conclusioni

6.1 Sintesi dei risultati

Il lavoro svolto è incentrato sull'analisi delle tecniche di face morphing, che rappresentano una vulnerabilità significativa nei moderni sistemi biometrici di controllo dell'identità, in particolare nei contesti di Automated Border Control.

La presenza di immagini morphed elude infatti i sistemi di controllo perché compromette il principio di univocità tra identità e documento.

Il lavoro svolto è incentrato sull'analisi della generazione di immagini morphed nello spazio degli embedding biometrici. Il morphing viene realizzato attraverso un'interpolazione tra i vettori identitari associati ai soggetti di partenza, producendo un embedding intermedio che rappresenta un'identità compatibile con entrambe le entità originali, ma che non può essere ricondotta in modo univoco a nessuna delle due.

Partendo da questa configurazione, ci si è interrogati sulla possibilità di ricostruire, almeno in maniera approssimata, una delle identità sorgenti a partire dall'immagine morphed del documento e dall'embedding della fotografia live acquisita al controllo. L'obiettivo è quindi verificare se, attraverso un'operazione di demorphing nello spazio delle feature, sia possibile stimare l'identità complementare e valutarne la compatibilità biometrica.

Alla luce delle analisi teoriche, è stato possibile osservare che l'inversione approssimata del processo di interpolazione nello spazio degli embedding produce

risultati coerenti con le attese. In particolare, l'operazione di demorphing consente di stimare un embedding riconducibile all'identità che ha contribuito alla generazione dell'immagine morphed.

Le metriche di valutazione utilizzate mostrano che l'embedding stimato presenta valori di distanza molto bassi rispetto alle immagini appartenenti al soggetto sorgente, indicando un'elevata similarità biometrica. La validazione del processo è stata estesa ad ulteriori immagini del soggetto messe a disposizione dal dataset FEI, caratterizzate da variazioni di posa, espressione facciale e condizioni di illuminazione. Anche in presenza di tali variazioni, le metriche di distanza restituiscono valori generalmente contenuti, confermando la coerenza della ricostruzione e la stabilità dell'identità stimata nello spazio latente.

Tali risultati suggeriscono che, in uno scenario controllato in cui siano noti il metodo di interpolazione e il parametro utilizzato per la generazione del morph, l'inversione del processo risulta effettivamente in grado di recuperare informazioni significative sull'identità complementare.

6.2 Limiti e sviluppi futuri

Tuttavia, è importante sottolineare che la coerenza dei risultati è strettamente legata alle condizioni controllate della sperimentazione.

Il problema inverso è stato ricostruito assumendo nota a priori l'intera pipeline di generazione dell'immagine morphed, inclusi i modelli utilizzati per l'estrazione degli embedding, l'algoritmo di interpolazione impiegato e l'ordine delle operazioni nella fase generativa. In particolare, si è ipotizzato che il morphing fosse realizzato tramite interpolazione sferica (SLERP) nello stesso spazio latente in cui viene successivamente effettuata l'inversione. Questa coerenza architetturale riduce le fonti di variabilità e garantisce una forte dipendenza tra fase di generazione e fase di inversione.

In uno scenario reale, tuttavia, il metodo e la pipeline generativa con cui un'immagine morphed è stata prodotta potrebbero essere differenti.

In questa prospettiva, un naturale sviluppo futuro consiste nell'estendere l'analisi a tecniche di morphing eterogenee, valutando la robustezza dell'approccio

rispetto a differenti tecniche e la capacità del metodo di adattarsi a pipeline generative non note a priori.

Anche assumendo che l'immagine del documento sia stata generata tramite interpolazione sferica, emerge un ulteriore vincolo teorico. La funzione SLERP non è invertibile nel senso matematico classico e la sua ricostruzione dipende strettamente dal parametro di interpolazione t , che controlla la posizione del punto lungo l'arco sull'ipersfera.

Nel corso dell'esperimento, tale parametro è stato assunto pari a 0.5, corrispondente a un contributo bilanciato tra le due identità e rappresentativo di uno scenario tipico di morphing. In un contesto reale, tuttavia, il valore di t non è noto né direttamente stimabile a partire dagli embedding disponibili, rendendo l'assunzione adottata una semplificazione teorica. Questo aspetto apre a ulteriori sviluppi, come l'analisi della sensibilità del processo di demorphing rispetto a diversi valori del parametro t .

Bibliografia

- [DDBFM23] Nicolò Di Domenico, Guido Borghi, Annalisa Franco, and Davide Maltoni. Combining identity features and artifact analysis for differential morphing attack detection. In *International Conference on Image Analysis and Processing*, pages 100–111. Springer, 2023.
- [DDFFM26] Nicolò Di Domenico, Annalisa Franco, Matteo Ferrara, and Davide Maltoni. Arc2morph: Identity-preserving facial morphing with arc2face. *arXiv preprint arXiv:2602.16569*, 2026.
- [DGXZ19] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.
- [DGY⁺22] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, October 2022.
- [FFM14] Massimo Ferrara, Alessandro Franco, and Davide Maltoni. The magic passport. In *Proceedings of the International Joint Conference on Biometrics (IJCB)*, 2014.
- [FFM17] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. Face demorphing. *IEEE Transactions on Information Forensics and Security*, 13(4):1008–1017, 2017.

- [FFMM12a] Matteo Ferrara, Annalisa Franco, Dario Maio, and Davide Maltoni. Face image conformance to iso/icao standards in machine readable travel documents. *IEEE Transactions on Information Forensics and Security*, 7(4):1204–1213, 2012.
- [FFMM12b] Matteo Ferrara, Annalisa Franco, Dario Maio, and Davide Maltoni. Face image conformance to iso/icao standards in machine readable travel documents. *IEEE Transactions on Information Forensics and Security*, 7(4):1204–1213, 2012.
- [GGOE⁺20] Alberto Garcia-Garcia, S Orts-Escolano, S Oprea, V Villena-Martinez, and J Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. arxiv 2017. *arXiv preprint arXiv:1704.06857*, 2020.
- [Gus06] Henrik Gustavsson. Sigrad 2006. 2006.
- [HG] ISHMFLFM HAtt and BSHM Our GT. Modnet: Real-time trimap-free portrait matting via objective decomposition.
- [MCW15] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. The chicao face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47(4):1122–1135, 2015.
- [MS25] Maxwell Meyer and Jack Spruyt. Ben: Using confidence-guided matting for dichotomous image segmentation, 2025.
- [QZH⁺20] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106:107404, 2020.
- [RGH⁺24] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr

- Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024.
- [SO24] Sefik Serengil and Alper Ozpinar. A benchmark of facial recognition pipelines and co-usability performances of modules. *Journal of Information Technologies*, 17(2):95–107, 2024.
- [SR22] Jag Mohan Singh and Raghavendra Ramachandra. Reliable face morphing attack detection in on-the-fly border control scenario with variation in image resolution and capture distance. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2022.
- [VRRB20] Sushma Venkatesh, Raghavendra Ramachandra, Kiran Raja, and Christoph Busch. Face morphing attack generation & detection: A comprehensive survey. *arXiv preprint arXiv:2011.02045*, 2020.
- [WD21] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.
- [XWY⁺21] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.
- [YWP⁺18] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation, 2018.