Università degli Studi di Camerino

Scuola di Scienze e Tecnologie

Corso di Laurea in Informatica (Classe L-31)

Deepfake: Analisi su tecnologia, prestazioni ed impatto sociale

Laureando
Leonardo Olivieri

Matricola 099580

Relatore
Marcantoni Fausto

A.A. 2020/2021

Indice

Elenco delle figure

Elenco delle tabelle

1. Introduzione

Alla base di questa tesi vi è lo studio del DeepFake, dall'aspetto tecnologico a quello sociale e legale. Inoltre verrà esaminato il software più popolare per la creazione di questi contenuti. Verranno analizzate in modo approfondito le prestazioni con diverse impostazioni e configurazioni. Oltre a ciò verranno realizzati diversi video demo descrivendo le varie difficoltà e problematiche riscontrate.

Nel prossimo capitolo, ossia il Capitolo 2 "La tecnologia DeepFake", verrà descritta la tecnologia DeepFake. Più nello specifico di che branca dell'intelligenza artificiale fa parte e come riconoscere un filmato alterato con il DeepFake. Nel terzo capitolo "Aspetto Legale e Sociale" si illustreranno le diverse applicazioni del DeepFake e cosa prevede, e non prevede, la legge italiana e quella estera. Nel Capitolo 4 "Il software: DeepFaceLab" si esaminerà il software Open-Source DeepFaceLab scelto per la realizzazione di questi contenuti. Nel quinto capitolo "Analisi Performance" verranno analizzate approfonditamente le performance di tale software in diverse configurazioni ed impostazioni. Verranno confrontati i modelli messi a disposizione e le diverse versioni del software. Nel capitolo 6 "Realizzazione video" si commenteranno i risultati ottenuti dalla realizzazione di tre video differenti con il DeepFake. In particolar modo verranno fatte presenti le problematiche ed i relativi difetti dei video nonché dei suggerimenti per migliorare il tutto. Infine i capitoli 7 e 8 sono conclusivi per cui ci sarà una conclusione e verranno menzionati eventuali sviluppi futuri.

1.1 Motivazione

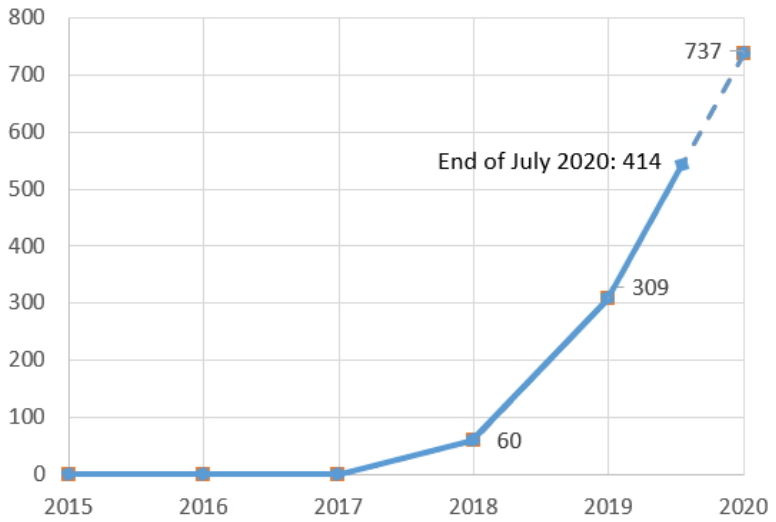

Figura 1.1: Numero di documentazioni relative al DeepFake fra il 2015 ed il 2020[TTN]

I video DeepFake, nel corso degli ultimi anni, sono diventati sempre più popolari in rete. Di conseguenza anche il numero di post, notizie e documentazioni relative a ciò è aumentato esponenzialmente (figura 1.1). Ciò che mi ha motivato nella realizzazione di questo scritto è il conoscere i requisiti minimi per la realizzazione di tali contenuti, analizzare le performance con hardware differenti per capire quale sia il budget minimo di cui si ha bisogno ed infine studiare la tecnologia alla base del DeepFake per avere dei fondamenti per capire come si realizzino certi contenuti.

1.2 Obiettivi

Lo scopo di questa tesi è studiare la tecnologia DeepFake con l'obbiettivo di comprendere le sue basi e riuscire a realizzare dei contenuti essendo in grado, fin da subito, di intuire se ci possono essere o meno delle difficoltà nella creazione del video. La conoscenza dell'hardware è fondamentale poiché si deve essere in grado di capire se la macchina che abbiamo a disposizione, o quella con cui andremo a lavorare, è sufficientemente "potente" per la realizzazione di tali contenuti. Inoltre, per la personalizzazione dei modelli, è fondamentale conoscere le specifiche della macchina onde evitare di incappare in errori di vario genere. Infine bisogna essere in grado, in condizioni ottimali, di realizzare un video realistico (hardware permettendo).

2. La tecnologia Deepfake

Quando si parla di Deepfake si intende una tecnica di manipolazione di una o più immagini (come ad esempio i frame di un video) che ha come obiettivo la creazione di un file modificato senza l'intervento manuale. Questo è possibile grazie a delle tecniche di Machine Learning (più nello specifico il Deep Learning) che fanno uso di reti neurali create appositamente per la specifica operazione. Ad esempio una rete allenata per falsificare immagini non è la stessa che si usa per falsificare degli audio, più nello specifico una rete allenata per creare video falsi di una specifica persona, andrà ri-allenata se si vuole creare video falsi di un'altra persona e comunque questa stessa rete non andrà mai bene per falsificare degli audio. La parola deriva da un neologismo inglese che incrocia la locuzione Deep Learning con il sostantivo fake "falso".

2.1 Deep Learning

Il Deep Learning è un campo di ricerca del Machine Learning e dell'intelligenza artificiale e descrive una classe di algoritmi di apprendimento automatico. Spesso si confondono "Deep Learning" con "Machine Learning" o addirittura si pensa che sia qualcosa al di fuori dell'intelligenza artificiale. La figura 2.1 mira a chiarire la differenza fra questi termini: con la parola Intelligenza Artificiale (IA) comprendiamo sia il Machine Learning sia il Deep learning. Tuttavia con l'intelligenza artificiale includiamo anche l'aspetto psicologico e morale che una macchina potrebbe o non potrebbe avere: come per esempio le "3 leggi della robotica" scritte da Isaac Asimov[1]. In sostanza potremmo dire che, quindi, il Deep Learning è un tipo particolare di Machine Learning.

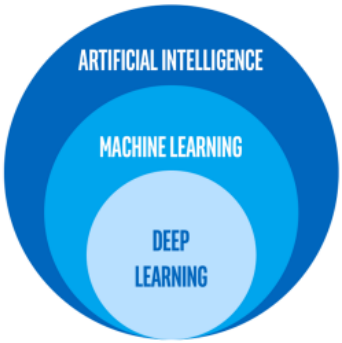

Figura 2.1: Distinzione fra Intelligenza artificiale (IA), Machine Learning e Deep Learning[Rob]

[1] https://it.wikipedia.org/wiki/Tre_leggi_della_robotica

2.1.1 Machine Learning e Deep Learning

Il Machine Learning raccoglie un insieme di metodi statistici che utilizzano parametri da dati esistenti noti e quindi prevede i risultati su nuovi dati. Per esempio, se si sapessero quante automobili (modello, prezzo ecc) siano state vendute in provincia di Macerata, sarebbe possibile, con il Machine Learning, creare un modello in grado di predire quanto il prezzo di un'automobile possa cambiare da provincia a provincia. Di norma, il Machine Learning si basa su un set di "features" che considera importanti nel dataset. Nel caso delle automobili queste features possono essere il colore, la marca, il modello, il numero di cavalli del mezzo, i vari accessori, se si tratta di o meno di una versione speciale ecc. Il processo che permette di introdurre queste features in un algoritmo di Machine Learning viene chiamato "Feature Engineering". Questo richiede una vasta conoscenza di un determinato argomento (in questo caso le automobili) e può rivelarsi un lavoro piuttosto lungo per un data scientist. Come detto il Deep Learning è un tipo di Machine Learning particolare che si è sviluppato molto nel corso degli ultimi anni. Con il Deep Learning l'algoritmo non ha bisogno di sapere quali siano le features importanti poiché è in grado di scoprirlo da solo utilizzando una rete neurale. Parlando sempre di automobili possiamo vedere in figura 2.2 come lavorano il Machine Learning ed il Deep Learning nel riconoscimento di un'immagine di un'automobile. Nel

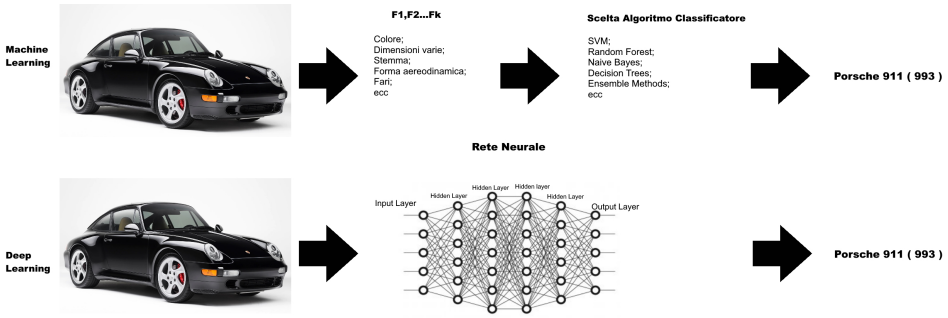

Figura 2.2: Come operano Machine Learning e Deep Learning

Machine learning, un data scientist, deve identificare per prima cosa tutte le features importanti (F1,F2...Fk) per poi utilizzare un algoritmo classificatore che consente di trovare un pattern tra le features e quindi restituire una sorta di identità unica con un certo grado di precisione. La difficoltà di questo approccio è che spesso non si sanno quali siano le features più importanti e, pur sapendole, potrebbe risultare complicato calcolarle. Per esempio possiamo immaginare di calcolare la distanza fra i due fari di un auto per aiutarci a capirne le dimensioni e/o la forma. L'immagine in figura 2.3 aiuta a chiarire il concetto mostrando un automobile con fari a scomparsa, in foto retratti rendendo complicato calcolare la distanza fra i due. Il Deep Learning invece

Figura 2.3: Honda NSX con fari a scomparsa [Pel]

ci consente di ignorare il "Feature Engineering". Infatti con una grossa mole di dati a disposizione (per esempio, in questo caso, immagini di automobili) e con un training adeguato, un modello Deep Learning riesce da solo ad identificare tutte le features rilevanti restituendo poi un risultato con una certa precisione. Nonostante la teoria dietro il Deep Learning sia relativamente vecchia, come scritto, in precedenza il Deep Learning si è sviluppato molto negli ultimi anni. Questo per 3 motivi:

- Dataset più grandi. Per esempio possiamo pensare al numero di video presenti su Youtube o il numero di immagini presenti in rete che continuano ad aumentare di giorno in giorno.
- Hardware migliore. La tecnologia va avanti e le aziende rilasciano componenti sempre migliori.
- Algoritmi più efficienti. Basti pensare a quanto aziende come Google ed Amazon abbiano investito nel IA ed il risultato sono nuovi algoritmi più efficienti. Algoritmi che spesso sono anche gratuiti.

Gli algoritmi del Deep Learning utilizzano più livelli per estrarre progressivamente caratteristiche di livello superiore dall'input non elaborato (grezzo). Ad esempio, nell'elaborazione delle immagini, i livelli superiori possono identificare i concetti rilevanti per un essere umano come cifre, lettere o volti, mentre i livelli inferiori possono identificare altri aspetti non rilevanti per l'uomo. Per esempio, tale funzionalità, è una delle features del nuovo sistema operativo per iPhone: iOS 15. In questa versione del sistema operativo di Apple lo smartphone è in grado di riconoscere, in base alle foto scattate e/o presenti in galleria, scritte o numeri. Inoltre, fornendo un certo grado di precisione, è in grado di riconoscere determinate razze animali (visibile in figura 2.4). Tuttavia, al momento della stesura di questo documento, è ancora in fase sperimentale

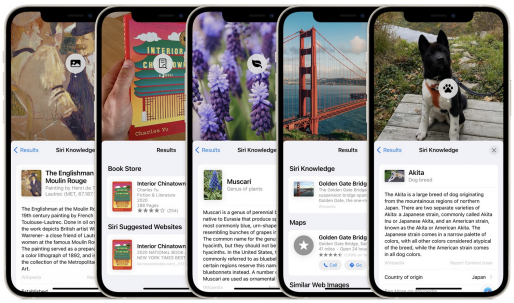

Figura 2.4: "Visual Lookup" di iOS 15 [Pel]

e con funzioni limitate specialmente se paragonata alla piattaforma Google: Google Lens (rilasciata nel 2017). Inoltre, il Deep Learning, è stato utilizzato dall'azienda britannica DeepMind per la realizzazione di AlphaGO. Si tratta di un software creato per giocare a "Go" ed è stato il primo software in grado di battere un maestro umano senza nessun handicap ed in condizioni standard. AlphaGO è considerato una vera e propria pietra miliare sulla ricerca dell'intelligenza artificiale[2].

[2] https://www.independent.co.uk/life-style/gadgets-and-tech/news/google-alphago-computer-beats-professional-world-s-most-complex-board-game-go-a6837506.html

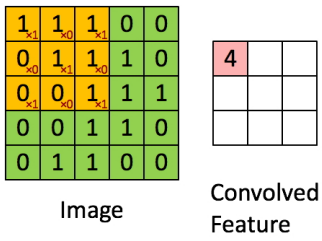

Figura 2.5: Prima convoluzione (filtro in giallo)[Det]

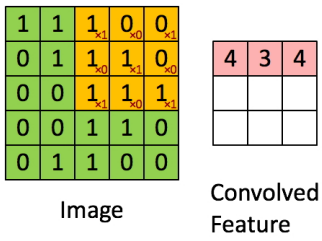

Figura 2.6: Terza convoluzione (filtro in giallo)[Det]

2.1.2 Reti Convoluzionali (CNN)

L'architettura Deep Learning più utilizzata per applicazioni che lavorano su immagini sono le reti neurali convoluzionali, anche dette CNN (Convolutional Neural Networks)[3] [4] [5].

Una CNN utilizza molti strati convoluzionali, ovvero strati dove viene impiegata un'operazione matematica chiamata convoluzione che filtra gli input per trovare le informazioni più utili (In figura 2.5 e 2.6 possiamo vedere due convoluzioni). L'architettura CNN presuppone esplicitamente che gli input siano immagini. Le reti CNN si ispirano alla corteccia visiva. La corteccia visiva ha piccole aree che sono sensibili a specifiche regioni del campo visivo. Questo fu scoperto da Hiber e Wiesel. Il loro lavoro mostrò come la corteccia visiva di gatti e scimmie contenesse neuroni che individualmente rispondono a piccole regioni del campo visivo. Supposto che gli occhi non si muovano, la regione del campo visivo entro cui lo stimolo influenzi l'innesco di un singolo neurone è noto come il suo "campo recettivo"[6]. La rete CNN è costituita da un blocco di input, uno o più blocchi nascosti che effettuano calcoli tramite funzioni di attivazione (esempio ReLU acronimo di Rectified Linear Units che consente di annullare i valori non utili nei livelli precedenti. Il ReLU è parte integrante del processo di convoluzione) ed un blocco di output che effettua la classificazione. La figura 2.7 mira a rendere il processo tutto più chiaro.

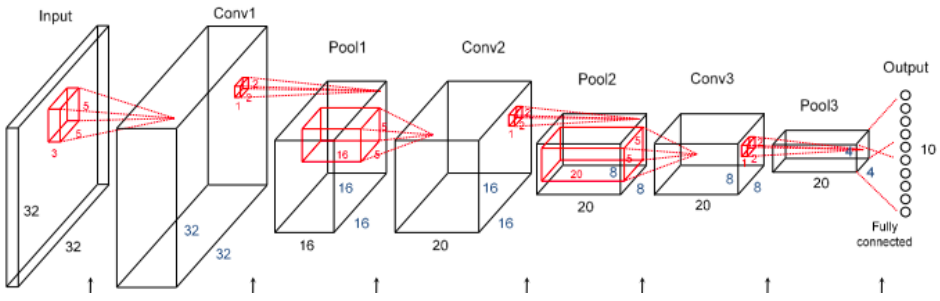

Figura 2.7: I vari livelli del CNN[Sor]

[3] https://en.wikipedia.org/wiki/Deep_learning#Deep_neural_networks

[4] https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

[5] https://developer.nvidia.com/blog/deep-learning-nutshell-core-concepts

[6] https://it.wikipedia.org/wiki/Rete_neurale_convoluzionale

Nella figura 2.8 il filtro viene rappresentato da una matrice 3x3 quindi si prenderà in esame solo una porzione (appunto 3x3) dell'immagine d'ingresso a cui si dovrà fare il prodotto con il filtro di convoluzione scelto.

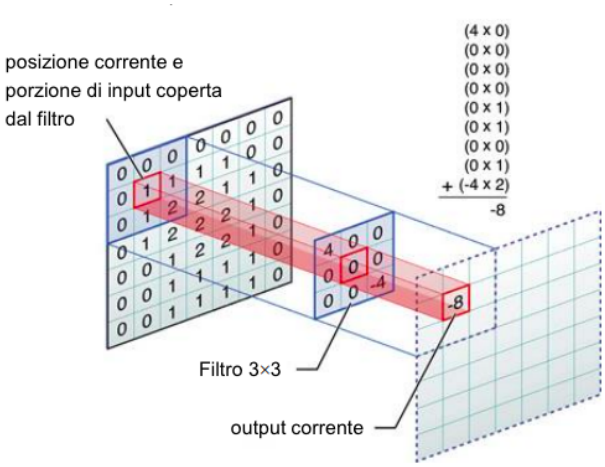

Figura 2.8: Filtro di Convoluzione[Sor]

In figura 2.9 possiamo vedere il risultato con un filtro casuale. In sostanza l'immagine

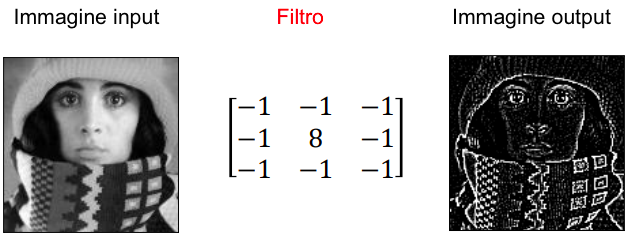

Figura 2.9: Esempio visivo del risultato[Sor]

viene scansionata pezzo per pezzo ottenendo una matrice di valori più piccola che rappresenta l'immagine "caratterizzata". Il Livello Pooling permette di identificare se la caratteristica di studio è presente nel livello precedente e rende più grezza l'immagine, mantenendo la caratteristica utilizzata dal livello convoluzionale. In altre parole il livello di pooling esegue un'aggregazione delle informazioni, generando feature map di dimensione inferiore. Livello FC (acronimo di Fully connected) è il livello che esegue di fatto la classificazione delle immagini. Nei capitolo successivi verrà menzionato il training/allenamento. Con training/allenamento si intende la modifica del kernel/filtro.

In passato, per il machine learning, venivano utilizzate solo le CPU ma a seguito di un articolo del 2005 [7] svariate pubblicazioni [8] [9] descrissero come l'utilizzo della GPU possa migliorare notevolmente le performance e l'efficienza.

[7] https://www.computer.org/csdl/proceedings-article/icdar/2005/24201115/12OmNylKAVX
[8] https://hal.inria.fr/inria-00112631/document
[9] http://yann.lecun.com/exdb/publis/pdf/ranzato-06.pdf

2.2 Come riconoscere un DeepFake

Nei capitoli successivi si parlerà delle problematiche e difficoltà relative alla creazione di un video Deepfake realistico. Molti contenuti presenti online, infatti, per via delle limitazioni menzionate in seguito, sono di bassa qualità e facilmente riconoscibili se si sa cosa guardare (Figura 2.10)

Figura 2.10: Screen tratto da un video chiaramente falso

Per riconoscere un Deepfake bisognerebbe fare attenzione:

- Agli occhi. In base ai test effettuati gli occhi vengono curati solo dopo un numero elevato di iterazioni. In un video di bassa qualità, appaiono senza vita, sono in bassa risoluzione, non si muovono e le palpebre non sbattono.

- Alle espressioni facciali. Poiché si va ad "attaccare" il volto di una persona su quello di un'altra, con un basso numero di iterazioni, le espressioni facciali non sembrano realistiche. Tendono infatti ad essere innaturali e limitate (Per esempio un sorriso potrebbe apparire come un leggero ghigno). Infine va anche notata, nel complesso, la forma del volto che può risultare irrealistica (Es. occhi enormi e volto piccolo o il contrario).

- All'illuminazione. Successivamente verranno menzionate le condizioni di luce ideali per un video ma in generale è possibile riconoscere un Deepfake in base a come la luce reagisce sul volto. Infatti è possibile notare come nella parte "originale" del volto la luce rifletta in un certo modo mentre nella parte falsa il riflesso sia diverso (Figura 2.11). Idem per le ombre o per il colore della pelle dell'individuo a cui abbiamo "preso" il volto: in determinati casi può apparire più scura o più chiara del normale.

- Alla bocca. Ci si ricollega in parte alle espressioni facciali tuttavia va fatto notare che in determinati video la bocca può assumere forme strane con poche iterazioni ed i denti non compaiono (o se lo fanno, non sono realistici).

Figura 2.11: Notare la quantità di luce dell'orecchio destro rispetto a quella del volto

- Al blur. Nella fase di merging, di cui è scritto in seguito, è possibile applicare degli effetti come il blur. Tuttavia, con un basso numero di iterazioni, anche senza applicarlo nel filmato sarà presente una quantità eccessiva di blur. Per esempio un soggetto apparirà con una pelle estremamente liscia, senza rughe o lentiggini.
- Al flickering. Con flickering si intende uno sfarfallio che avviene quando il soggetto tende a muoversi e l'attaccamento del volto non avviene correttamente causando uno sfarfallio nel video nella parte del viso. E' forse la cosa più semplice da notare.

Tutti questi "metodi", però, aiutano a riconoscere un Deepfake solo se di "bassa qualità". Nei contenuti estremamente realistici si fa molta fatica a capire se ciò che stiamo vedendo è vero o falso. In questo caso, se non è stata alterata anche la parte audio (argomento non trattato nella tesi), è possibile riconoscere il video conoscendo la voce del soggetto originale. Tuttavia, come vedremo, con un buon imitatore è possibile risolvere questo problema. Per questo grandi aziende, come Facebook[10][11], hanno sviluppato o stanno sviluppando delle "IA" in grado di riconoscere i contenuti falsi da quelli veri.

[10] https://ai.facebook.com/blog/reverse-engineering-generative-model-from-a-single-deepfake-image/

[11] https://www.npr.org/2021/06/17/1007472092/facebook-researchers-say-they-can-detect-deepfakes-and-where-they-came-from?t=1628261061147

3. Aspetto Legale e Sociale

I video prodotti con il DeepFake sono qualcosa di nuovo e mai creato in precedenza e, avendo a disposizione tempo e potenza hardware, è possibile creare filmati falsi indistinguibili da un filmato vero e proprio. La tecnologia in se e per se non è pericolosa, dipende tutto dall'uso che se ne fa. In questo capitolo, per quanto riguarda l'aspetto legale, vedremo come alcuni stati si sono mossi per regolamentare l'utilizzo di questa tecnologia e la situazione in Italia. Mentre, per l'aspetto sociale, vedremo dove e come è stata utilizzata.

3.1 Aspetto Legale

Come scritto in precedenza la tecnologia, presa a se, non è pericolosa: è pericoloso l'eventuale l'uso che se ne fa. Da una ricerca condotta nel 2019 dalla società di cybersecurity Sensity (ex Deeptrace), è emerso che i video DeepFake presenti in rete fossero quasi 15'000 di cui il 96% era costituito da video pornografici[1]. Nel loro report del 2020[2], riguardante i bot di Telegram, è emerso come, in un sondaggio anonimo, il 63& fosse interessato a creare contenuti per adulti di ragazze che conoscono nella "vita vera". Solo il 16% vorrebbe invece creare contenuti erotici con personaggi famosi (attrici, cantanti ecc). Contenuti come quello in figura 3.1, nonostante fossero stati

Figura 3.1: Frame tratto da un video deepfake con l'attrice britannica Daisy Ridley[All]

[1] http://regmedia.co.uk/2019/10/08/deepfake_report.pdf

[2] https://www.medianama.com/wp-content/uploads/Sensity-AutomatingImageAbuse.pdf

bannati dai social più famosi come Reddit[3], Facebook [4], TikTok[5] e Twitter [6]o da piattaforme con contenuti per gli adulti come Pornhub[7], sono facilmente reperibili e rappresentano solo la punta dell'iceberg.

Attualmente in Italia non esiste una legge specifica riguardo l'uso della tecnologia DeepFake. In base a quanto scritto in precedenza, nel caso di contenuti amatoriali con persone non famose, si potrebbe pensare di applicare la legge riguardante il Revenge porn. Il revenge porn è un'espressione inglese che indica la condivisione di immagini o video intimi senza il consenso di uno dei protagonisti. Il reato di revenge porn è stato introdotto nel 2019 con la legge del *19 luglio 2019 n.69* nota come Codice Rosso[8]. Nello specifico l'articolo 10 della suddetta legge inserisce nel codice penale *l'articolo 612 ter* riguardante, appunto il delitto di diffusione illecita di immagini o video sessualmente espliciti. Il problema, però, è che la legge non fa alcun riferimento a contenuti multimediali non reali. Quindi, per via del principio di tassatività della procedura penale, che implica che la norma penale deve individuare con precisione gli estremi del fatto reato in essa contenutili e della già descritta mancanza di giurisprudenza in materia, potrebbe risultare piuttosto complesso far rientrare la pornografia DeepFake nella fattispecie incriminatrice del revenge porn.

Quindi dal punto di vista legale come ci si può tutelare? In primis sporgendo una querela per diffamazione[9] in quanto tali contenuti, seppur falsi, sono in grado di offendere la reputazione del soggetto ritratto. Nell'ipotesi che questi contenuti vengano realizzati al fine di minacciare o estorcere qualcosa dalla vittima (per esempio denaro in cambio della non pubblicazione o eliminazione) può configurarsi un tentativo di estorsione[10]. Inoltre, nel tentativo di recuperare quanto più materiale possibile (foto e/o video) dalla vittima potrebbe verificarsi il reato di stalking[11] ed infine il trattamento illecito dei dati ed il furto d'identità personali[12][13].

[3] https://www.bbc.com/news/technology-42984127
[4] https://www.bbc.com/news/technology-51018758
[5] https://www.theverge.com/2020/8/5/21354829/tiktok-deepfakes-ban-misinformation-us-2020-election-interference
[6] https://www.vox.com/recode/2020/2/4/21122653/twitter-policy-deepfakes-nancy-pelosi-biden-trump
[7] https://www.theverge.com/2018/2/6/16980920/pornhub-bans-deepfakes-fake-ai-celebrity-porn-video
[8] https://it.wikipedia.org/wiki/Codice_Rosso_(legge)
[9] https://www.brocardi.it/codice-penale/libro-secondo/titolo-xii/capo-ii/art595.html
[10] https://www.brocardi.it/codice-penale/libro-secondo/titolo-xiii/capo-i/art629.html
[11] https://www.brocardi.it/codice-penale/libro-secondo/titolo-xii/capo-iii/sezione-iii/art612bis.html
[12] https://www.brocardi.it/codice-della-privacy/parte-iii/titolo-iii/capo-ii/art167.html
[13] https://www.altalex.com/documents/news/2021/01/28/furto-identita-digitale-e-illecito-utilizzo-dati-raccolti

3.1.1 Situazione all'estero

Nel 2018, negli Stati Uniti, è stato introdotto il *Malicius Deep Fake Prohibition*[14] e nel 2019 è stato introdotto il *DEEPFAKES Accountability Act*[15]. Nel Novembre del 2019 la Cina ha annunciato che, a partire dal 2020, i DeepFake dovranno avere un watermark o comunque un qualcosa per rendere subito chiaro che si ha di fronte un contenuto falso altrimenti verrà considerato un crimine. Inoltre, il governo Cinese, si riserva il diritto di perseguire sia le piattaforme online sia gli utenti che non rispetteranno la legge[16].

3.1.2 Come proteggersi

Purtroppo non esiste un metodo per proteggersi dai malintenzionati che vogliano utilizzare il DeepFake contro di noi. L'unica cosa che si può fare è cercare di fornire meno contenuti multimediali possibili. Quindi limitare la pubblicazione di foto ma soprattutto video nei social media e fare in modo che, questi contenuti, siano accessibili solo a persone che conosciamo realmente. Infine il non utilizzare applicazioni dubbia provenienza che richiedono l'accesso alla galleria del nostro smartphone o applicazioni come FaceAPP che utilizzano il nostro volto per applicare determinati effetti senza farci sapere dove i video sorgente vengano inviati.

3.2 Aspetto Sociale

Per quanto riguarda il restante 4% la tecnologia DeepFake viene utilizzata nei più svariati modi: dai meme alla recitazione.

3.2.1 Recitazione

L'utilizzo di tale tecnologia nel cinema si è rivelata estremamente utile e pratica nel sostituire un attore scomparso o troppo vecchio per interpretare un determinato ruolo. Nell'industria cinematografica, inoltre, questa tecnologia può anche essere utilizzata per rendere meno visibile l'utilizzo della controfigura.

Fast and Furious 7

Nel settimo film della saga di "Fast and Furious"[17] si dovette utilizzare la tecnologia DeepFake per via della morte, in un incidente stradale, dell'attore Paul Walker che interpretava uno dei protagonisti della serie: "Brian O'Conner". La morte dell'attore avvenne durante le riprese del film quindi dovettero sospendere la produzione. Dopo aver modificato il copione decisero di adottare tale tecnologia per far apparire nuovamente Paul Walker nello schermo. Per farlo utilizzarono tutti i filmati (anche quelli d'archivio) dei precedenti film e, per la persona "fisica" a cui "attaccare" il volto dell'attore americano, ingaggiarono i due fratelli: Cody e Caleb. Così facendo, come visibile in figura 3.2, per via dei molti filmati a disposizione potettero creare un modello

[14] https://www.congress.gov/bill/115th-congress/senate-bill/3805/text?format=txt

[15] https://www.congress.gov/bill/116th-congress/house-bill/3230/text

[16] https://www.theverge.com/2019/11/29/20988363/china-deepfakes-ban-internet-rules-fake-news-disclosure-virtual-reality

[17] https://it.wikipedia.org/wiki/Fast_%26_Furious_7

Figura 3.2: A sinistra il fratello Cody Walker, a destra il risultato con il Deepfake[GUR]

di Paul Walker estremamente realistico che sarebbe stato sovrapposto ad un soggetto già somigliante all'attore scomparso[18].

Figura 3.3: Scene del film con Paul Walker "vero" (sinistra) e falso (destra)[Why]

Il risultato è sorprendente: si fa fatica a distinguere quali scene siano state girate con l'attore ancora in vita ed in quali sia stata utilizzata la tecnologia DeepFake. La figura 3.3 mette a confronto due frame da due scene del film da una parte con l'attore ancora in vita e dall'altra con il fratello più il DeepFake.

Rogue One

In Rogue One[19] la situazione fu differente. Tale tecnologia non venne utilizzata per concludere un film nel miglior modo a causa della scomparsa di uno dei protagonisti ma per far ritornare due personaggi: Tarkin e la principessa Leila. L'attrice di quest'ultima morì nell'anno in cui uscì Rogue One mentre l'attore che interpretava Tarkin era deceduto nel 1994. Quindi il Deepfake venne utilizzato per mostrare a tutti le potenzialità di questa tecnologia. Per il personaggio di Tarkin, originariamente interpretato da Peter Cushing, venne scelto l'attore britannico Guy Henry. Uno dei motivi della scelta fu la somiglianza con l'attore scomparso. Anche qui utilizzarono i filmati dei film precedenti più quelli di archivio per creare un modello realistico. Henry, al fine

[18] https://www.youtube.com/watch?v=yQC6AJ6x8SM&ab_channel=VFXGURU
[19] https://it.wikipedia.org/wiki/Rogue_One:_A_Star_Wars_Story

di rendere il tutto più realistico, studiò e copiò il modo di muoversi e di parlare di Cushing[20] [21] [22].

Figura 3.4: Guy Henry ed i modelli creati[Wir]

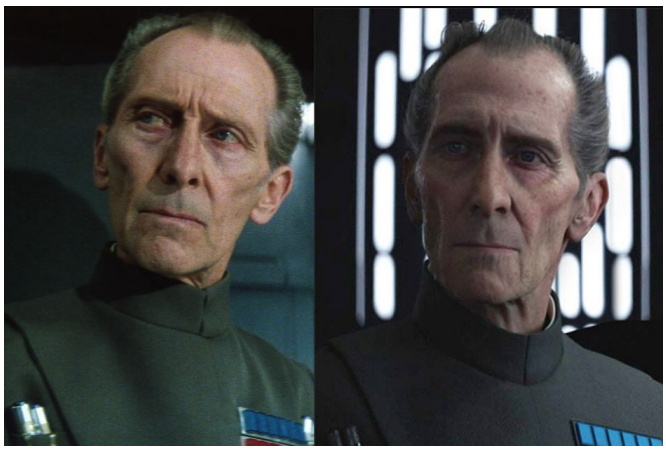

Figura 3.5: Confronto fra il Tarkin originale e quello creato con il deepfake[Lin]

Nelle figure 3.4 e 3.5 è possibile vedere i modelli realizzati con l'attore ed i filmati d'archivio più un confronto fra l'attore originale che interpretava Tarkin ed il risultato in Rogue One.

Per la principessa Leila venne scelta l'attrice norvegese Ingvild Deila. Leila, in Rogue One, appare per circa 15 secondi ma la particolarità è che venne utilizzata la voce dell'attrice originale: Carrie Fisher[23]. Venne usato un audio d'archivio in cui la Fisher pronuncia la parola "Hope"[24]. In figura 3.6 possiamo vedere il risultato confrontato con la Leila "reale".

[20] https://www.youtube.com/watch?v=vlSn50_BePU&t=41s&ab_channel=ILMVFX
[21] https://variety.com/2016/film/news/rogue-one-peter-cushing-digital-resurrection-cgi-1201943759/
[22] https://www.youtube.com/watch?v=WRCgKXZJ7Ks&ab_channel=GrandMoffTarkin
[23] https://www.technologyreview.com/2018/10/16/139739/how-acting-as-carrie-fishers-puppet-made-a-career-for-rogue-ones-princess-leia/
[24] https://www.yahoo.com/entertainment/rogue-one-sound-editors-reveal-how-they-found-princess-leias-hope-and-more-production-secrets-132116513.html

Figura 3.6: A sinistra la Leila falsa in Rogue One a destra la Leila vera in "Star Wars IV - Una nuova speranza"[Bon]

Figura 3.7: Al Pacino in Taxi Driver[FACb]

Figura 3.8: Home Stallone[FACa]

3.2.2 Meme e produzioni amatoriali

La creazione di video DeepFake riguardanti il cinema non è esclusiva delle sole case cinematografiche. Con l'aumentare della potenza hardware a disposizione e l'ottimizzazione dei software esistenti è possibile creare in casa degli ottimi contenuti multimediali.

Contenuti come posizionare il volto di Al Pacino al posto di quello di De Niro in "Taxi Driver"[25](figura 3.7) oppure creare un cortometraggio basandosi su "Mamma ho perso l'aereo" riscrivendo la storia e posizionando il volto di Stallone sopra quello di ogni uomo[26] (figura 3.8)

Ma è nella creazione dei meme che questa tecnologia sta spopolando. Poiché spesso la qualità non è importante si riescono a creare velocemente brevi video riguardanti dei trend popolari. I più popolari di solito sono quelli riguardanti il karaoke: si passa dal far cantare "Video Killed The Radio Star" al far cantare "Baka Mitai" a persone o cose[27].

[25] https://www.youtube.com/watch?v=9NkKj0aNB0s
[26] https://www.youtube.com/watch?v=2svOtXaD3gg
[27] https://www.youtube.com/watch?v=jlGusWZM6Rk

3.2.3 Altri utilizzi

Medicina

Il Machine Learning può essere utilizzato dai medici come aiuto per l'individuazione delle neoplasie[28] o per l'individuazione delle malattie della cornea[29]. Riguardo possibili malattie dell'occhio, nel 2018, la precedentemente nominata azienda DeepMind in collaborazione con l'Ospedale Oculistico di Moorfield e l'University College di Londra ha rilasciato un software che promette di riconoscere l'anatomia dell'occhio, la patologia ed in grado di consigliare ai medici una cura adeguata [30]. In pratica viene effettuata prima una tomografia all'occhio del paziente per acquisire le immagini in alta risoluzione e successivamente, queste immagini, vengono elaborate dall'algoritmo che le confronta con quelle di una banca dati clinica in modo da effettuare una diagnosi e consigliare una cura ottimale. Tuttavia la sicurezza non va trascurata poiché un gruppo di ricercatori ha dimostrato che possono essere creati dei malware in grado di manipolare i dati dei test ed alterare le scan ingannando i medici e diagnosticando erroneamente i pazienti[31]. Nella

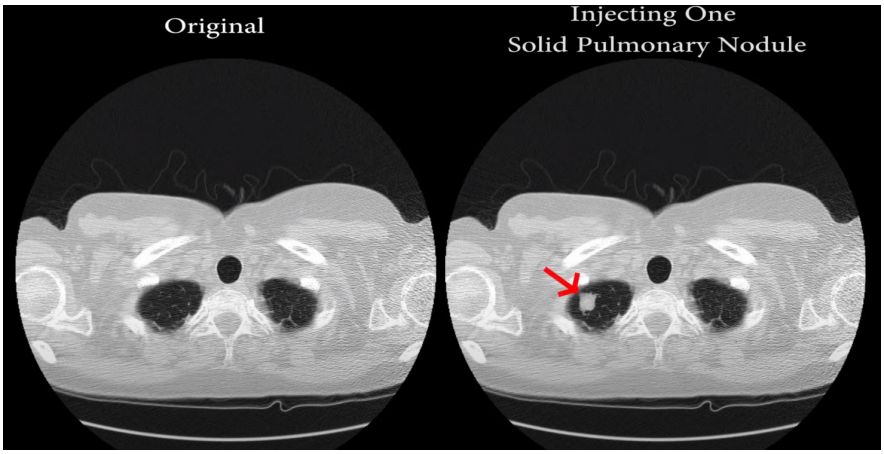

Figura 3.9: A sinistra polmoni sani a destra gli stessi polmoni con un nodulo falso[O'N]

figura 3.9 è possibile vedere uno screen radiografia. A sinistra i polmoni di un soggetto sano a destra i polmoni dello stesso soggetto a cui è stato inserito un carcinoma[32].

Arte

Il dipinto visibile in figura 3.10 è stato realizzato dalla scannerizzazione di altri 15 mila dipinti con l'utilizzo delle reti Gan (Generative Adversarial Network). Il dipinto è stato poi venduto per la somma di 432'500 dollari[33].

[28] https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0681-4
[29] https://tvst.arvojournals.org/article.aspx?articleid=2772646
[30] https://www.lastampa.it/tecnologia/news/2018/09/03/news/l-intelligenza-artificiale-di-deepmind-riconosce-piu-di-50-malattie-oculari-1.34042602
[31] https://gizmodo.com/researchers-demonstrate-malware-that-can-trick-doctors-1833786672
[32] https://www.youtube.com/watch?v=_mkRAArj-x0
[33] https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx

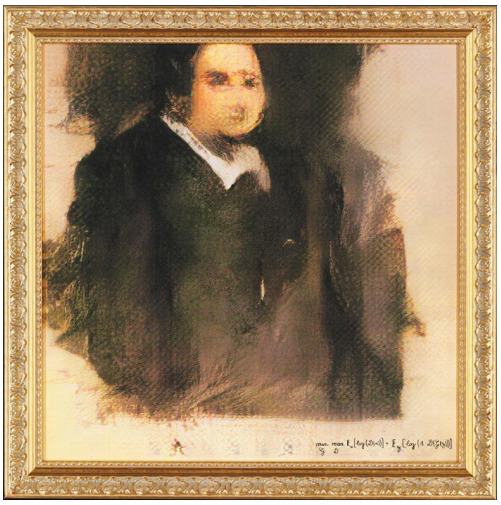

Figura 3.10: Portrait of Edmond Belamy, 2018, created by GAN[Chr]

Satira e parodie: Striscia la notizia

Figura 3.11: Deepfake di Andrea Pirlo (sx) ed Andrea Conte (dx) realizzato da Striscia la Notizia[Pia]

L'imitazione è sempre stata un cavallo di battaglia del programma televisivo "Striscia la Notizia". Per imitare una persona è necessario, oltre ad essere in grado a modulare la propria voce e ricreare l'accento di colui che vogliamo imitare, soprattutto cercare di assomigliare a tale individuo mediante del trucco, delle parrucche ecc. Con l'utilizzo della tecnologia DeepFake quest'ultima parte è diventata irrilevante. Sono stati realizzati dei servizi da Striscia la Notizia dove venivano imitate persone di ogni tipo: da politici ad allenatori di calcio come in figura 3.11. Per la realizzazione di questi contenuti hanno assunto sempre degli imitatori con il solo scopo di imitare la voce del soggetto. Per tutto il resto veniva utilizzato il DeepFake. Dovendo imitare figure pubbliche e famose non si hanno problemi a reperire contenuti multimediali per creare un modello e, trattandosi comunque di uno studio televisivo, non hanno alcuna difficoltà nel creare un video di destinazione adatto ad ogni scopo (risoluzione, qualità o luci non sono di certo un problema).

4. Il software: DeepFaceLab

In rete è possibile trovare molti programmi che sfruttano la tecnologia di base per la realizzazione dei Deepfake anche per altri scopi[1]; tuttavia il software che è stato utilizzato per questa tesi è DeepFaceLab.

DeepFaceLab[2] [3] è il programma più popolare per la creazione di deepfake. Si tratta di un software open-source con oltre 26 mila stelle e 6 mila "fork" su GitHub e si stima che la maggioranza dei video deepfake presenti in rete siano stati creati con questo software. Inoltre la presenza di molti tutorial dettagliati[4], l'enorme quantità di documentazione ed il supporto della community rende questo software il più gettonato per la creazione di contenuti.

DeepFaceLab fornisce un framework di deep learning di alto livello basato su TensorFlow puro, che mira ad evitare le restrizione non necessarie ed i costi aggiuntivi portati da alcuni framework comunemente utilizzati come Keras e plaidML. Iperov, ossia colui che ha scritto DeepFaceLab, l'ha chiamato Leras, abbreviazione di Lighter Keras. I principali vantaggi di Leras sono:

- Costruzione di modelli semplici e flessibili: uno stile simile a Python e a PyTorch (ovvero definizione di livelli, composizione di modelli neurali, scrittura di ottimizzatori), ma in graph mode invece che con eager execution.
- Implementazione focalizzata sulla performance: con l'utilizzo di Leras al posto di Keras, il tempo di addestramento si riduce in media di circa il 10-20%.
- Fine-granularity tensor management: la motivazione per passare a TensorFlow puro è che Keras e plaidML non sono abbastanza flessibili, inoltre sono in gran parte obsoleti e non danno il pieno controllo su come vengono elaborati i tensori.

DeepFaceLab è disponibile in tre versioni, adatte a diversi tipi di schede grafiche:

- DeepFaceLab_DirectX12_build
- DeepFaceLab_NVIDIA_up_to_RTX2080Ti_build
- DeepFaceLab_NVIDIA_RTX3000_series_build

La prima versione è adatta alle schede video AMD recenti ed a quelle Nvidia fino alle serie 900 (esclusa). La seconda versione è ottimizzata per schede video Nvidia

[1] https://github.com/enochkan/awesome-gans-and-deepfakes
[2] https://github.com/iperov/DeepFaceLab
[3] https://arxiv.org/pdf/2005.05535v4.pdf
[4] https://mrdeepfakes.com/forums/thread-guide-deepfacelab-2-0-guide

della serie 900, 1000 e 2000. Infine, la terza versione, è adatta solo alle schede video Nvidia RTX3000. E' possibile utilizzare una scheda video della serie 1000 con la prima build tuttavia l'utilizzo della memoria video sarà limitato e, come vedremo in seguito, verranno limitate le performance.

Come visibile in figura 4.1 il software fornisce una lista di file eseguibili con interfaccia a riga di comando. Per la realizzazione di un video senza pretese basta utilizzare le impostazioni di default degli eseguibili sottolineati.

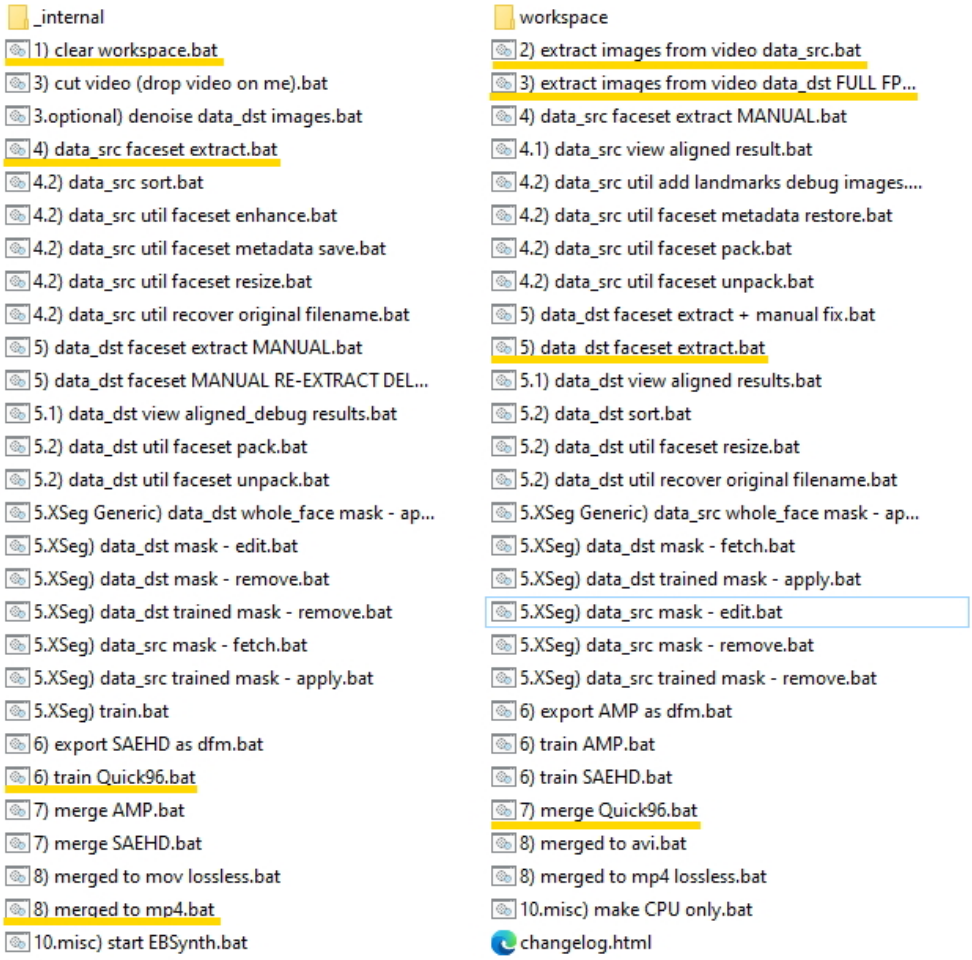

Figura 4.1: File che compongono DeepFaceLab, per realizzare un video basta eseguire i file evidenziati nell'ordine in cui sono numerati

Ogni file compie un passo del processo comune ad ogni software per la realizzazione di un DeepFake. Il primo eseguibile (file 1) serve per cancellare tutto quel che abbiamo creato nella cartella "workspaces": se è la prima volta che si usa il programma non è necessario eseguirlo. In primis avviene l'estrazione dei frame dal video sorgente e dal video destinazione (file 2 e 3), successivamente avviene l'estrazione dei volti dai frame già estratti (file 4 e 5) in tre fasi:

- Riconoscimento del viso;
- Allineamento del viso;
- Segmentazione del viso;

Successivamente avviene la fase di training tramite autoencoders[5]. Un autoencoder è una rete neurale artificiale, con la condizione aggiuntiva che l'output sia lo stesso dell'input. Modificando l'architettura in modo appropriato affinché che la rete sia strozzata, l'autoencoder ottiene una rappresentazione dimensionale inferiore dell'input, di fatto "comprimendolo". Per il training il programma ci fornisce 3 opzioni differenti:

- Quick96;
- SAEHD;
- AMP (introdotto nell'ultima build);

Il Quick96 è consigliato per produzioni amatoriali dove non si ha a disposizione tanta potenza hardware e/o poca memoria video (2/4GB VRAM). Inoltre, a differenza degli altri due, non presenta molte opzioni personalizzabili e quindi è adatto anche a chi è poco esperto. Per la realizzazione di meme è il più adatto tuttavia, con un buon numero di frame di "qualità" a disposizione, molte iterazioni e degli accorgimenti (per esempio la scelta del video di destinazione è fondamentale) è possibile realizzare dei video di qualità.

Il SAEHD è destinato a chi possiede schede video potenti e con tanta memoria video. E' altamente configurabile quindi adattabile ad ogni sorta di GPU a disposizione. In base ai test effettuati con la GTX 1080 queste sono alcune impostazioni da settare tenendo in considerazione la memoria e la potenza della scheda video a disposizione:

- resolution 192 (più si sale più usa la memoria video)
- ae_dims 512 (più si sale più usa la memoria video)
- ed_ch_dims 21 (più si sale più usa la memoria video)
- batch_size 8 (più si sale più usa la memoria video)

Tuttavia questo è solo un settaggio indicativo. E' possibile aumentare la risoluzione e diminuire gli altri valori e non avere problemi con una scheda da 8GB di memoria video. Effettuare dei test con i vari settaggi con l'hardware che si ha a disposizione è fondamentale, se si esagera con le opzioni e si satura la memoria video il training non partirà o stamperà a console un errore. Se si dispone di una scheda video più potente ma sempre con 8 GB di memoria video queste opzioni restano valide, l'unica differenza è che le performance saranno migliori.

Infine l'AMP, introdotto nell'ultima build, si pone come alternativa al SAEHD. Infatti esso fornisce, al contrario del Quick96, e come il SAEHD, molte opzioni per adattarlo alla GPU che si ha a disposizione. Nel capitolo successivo vedremo le sue performance confrontate con gli altri due modelli. Va però considerato che si tratta di un modello ancora in fase sperimentale e che, attualmente, sembra fornire risultati qualitativamente inferiori rispetto al SAEHD[6] [7].

Terminata la fase di training avverrà la fase di merging. In questa fase verrà fuso il volto del file sorgente che abbiamo generato con quello del file destinazione. E' possibile applicare svariati effetti, come visibile in figura 4.2, per migliorare la resa della fusione.

[5] https://www.alanzucconi.com/2018/03/14/an-introduction-to-autoencoders/

[6] https://www.youtube.com/watch?v=KKhGbN7f7DQ

[7] https://www.youtube.com/watch?v=ryXihJ-SL1o

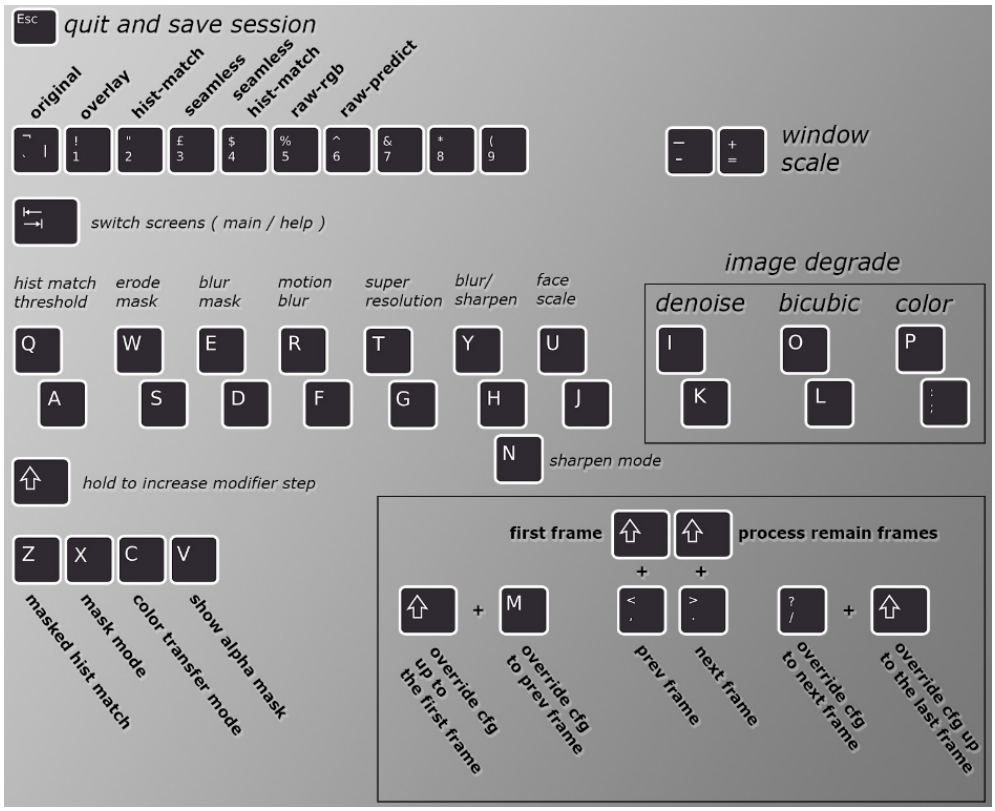

Figura 4.2: Le impostazioni disponibili nel merging

Tuttavia è necessario ricordare che, pur avendo a disposizione una miriade di effetti, non si può sopperire alla mancanza di frame utili per la creazione dei modelli dei volti o alla scelta di un video di destinazione non adatto. Il merging dipende da quale tipologia di training abbiamo scelto ed il programma ce lo mostra chiaramente dandoci 3 versioni differenti di merging.

Infine si ha l'ultimo eseguibile (file 8) che ci consente di esportare il tutto in un video mp4.

All'interno della cartella "workspace" dobbiamo inserire il video sorgente ed il video destinazione. Dovremmo rinominarli rispettivamente "data_src.mp4" ed "data_dst.mp4". Nel video sorgente ci deve essere il volto del soggetto per cui vogliamo creare il DeepFake mentre, nel video destinazione, dovremmo sovrapporre il volto ricavato dal video sorgente con quello già presente. Il programma ci fornisce due video per effettuare dei testi: nel video sorgente è presente "Tony Stark" mentre nel video sorgente è presente "Elon Musk". Così facendo possiamo "incollare" il volto del personaggio Marvel su quello dell'imprenditore sudafricano.

Come già ripetuto in precedenza la scelta del video di destinazione è fondamentale. Diversi elementi come barba, occhiali o cappelli rendono difficile la realizzazione di un video di qualità. Ma anche altri elementi come la forma del volto, le condizioni di luce o anche gli stessi movimenti del volto (specialmente se improvvisi) rischiano di minare seriamente la qualità del video risultate. Nel caso dovessimo lavorare con pochi frame, per la realizzazione di un video di qualità, bisogna scegliere un soggetto che non si muova molto, con condizioni di luci simili e con una forma del viso somigliante. In base ai test effettuati 3000 frames sono un buon punto di partenza per realizzare un video

di qualità e, più frame si hanno, più si riescono ad ignorare gli elementi fatti presenti in precedenza. Tuttavia è importante far notare che è meglio avere 1000 frames di qualità che 3000 fatti male. Per frames di qualità si intendono frame in alta risoluzione, presi da un video con ottime condizioni di luce e dove si riesce a riprendere il volto del soggetto in ogni angolazione e si riesce a catturare ogni espressione del viso. Per esempio un video che riprende il volto, in primo piano, del soggetto a cui destinare il deepfake è ideale e genererà frames di qualità mentre un video dove il soggetto è solo in lontananza, all'ombra o in scarse condizioni d'illuminazione genererà un frames scadenti.

Infine va fatto presente che DeepFaceLab consente di utilizzare, nella fase di training e nella fase di estrazioni dei volti, la CPU. Il vantaggio principale di utilizzare la CPU nella fase di training è la quantità di memoria a disposizione. Come vedremo in seguito la CPU è meno performante rispetto alla GPU nel training tuttavia il non avere limitazioni di memoria ci consente di cambiare le impostazioni del modello puntando ad una maggiore qualità. Infatti una RTX 3090 dispone di "soli" 24GB di memoria mentre CPU relativamente vecchie di fascia consumer riescono a supportare senza problemi anche 64GB di ram. Se quindi si ha molto tempo a disposizione e si ignora il consumo energetico per avere la massima qualità, il training con la CPU può essere un opzione. Più core si hanno meglio è tuttavia il programma, attualmente, non riesce a sfruttare al meglio le configurazioni multi socket come quella in figura 4.3. Determinate CPU

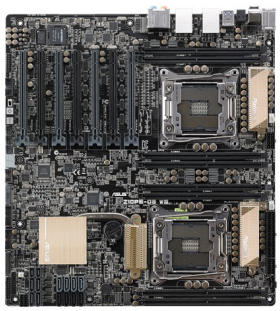

Figura 4.3: Una scheda madre con due socket[ASU]

possono risultare migliori di altre a prescindere dal numero di cores, dalla frequenza o dall'IPC. Le CPU Intel basate su architettura "Ice Lake"[8] supportano la funzionalità "Intel Deep Learning Boost"[9] in grado di aumentare le performance nel Machine Learning[10]. Quindi, se il budget non è un problema, sulla carta, la miglior CPU per il training è lo Xeon Platinum 8380[11] con 40 cores ed 80 threads che supporta fino a 6TB di memoria in 8 canali. Nel prossimo capitolo, nella sezione dedicata, vedremo in dettaglio le performance della CPU e come aumentarle.

[8] https://en.wikipedia.org/wiki/Ice_Lake_(microprocessor)

[9] https://www.intel.com/content/www/us/en/artificial-intelligence/deep-learning-boost.html

[10] Per le performance con Deep Learning vedere benchmark oneDDN https://www.phoronix.com/scan.php?page=article&item=intel-xeon-8380-linux&num=5

[11] https://ark.intel.com/content/www/us/en/ark/products/212287/intel-xeon-platinum-8380-processor-60m-cache-2-30-ghz.html

5. Analisi Performance

Nonostante il software ci consenta di realizzare un video senza pretese senza alcuna conoscenza pregressa necessitiamo comunque una certa potenza hardware per poter realizzare il suddetto video in tempistiche umane. L'analisi delle performance ci consentirà di vedere quale hardware risulta essere più adatto, eventuali problemi e/o limitazioni.

Per la fase di testing abbiamo utilizzato i due video che il programma ci fornisce per fare delle prove. Il primo video, in 1080p e lungo 27 secondi, mostra una scena presa da Iron Man 2 dove viene mostrato il monologo di Tony Stark (interpretato dall'attore Robert Downey Jr.). La scena ruota intorno al protagonista prendendo ogni punto del volto e da questo video verranno estratti 655 frames. Il secondo video, in 720p e lungo 54 secondi, mostra un intervento di Elon Musk. Anche qui la telecamera riesce a riprendere quasi ogni punto del volto e da questo video verranno estratti 1619 frames.

Configurazioni Hardware

Per i test sono stati utilizzate macchine differenti: 2 portatili(di cui uno di mia proprietà), una workstation Dell (fornita dall'università) ed il mio computer fisso personale.

Computer fisso

- CPU i7 7700k[1] @4.9ghz (NOTA: @ indica la frequenza a cui è stato overclockato);
- RAM 16GB;

Con questo computer sono state testate le seguenti schede video:

- MSI GT 1030 2GB GDDR5[2];
- MSI GTX 750Ti 2GB[3] / OC: +200mhz core +500mhz memory ;
- EVGA GTX 980 FTW 4GB[4];
- MSI GTX 1080 8GB[5];

Portatili

[1] https://ark.intel.com/content/www/it/it/ark/products/97129/intel-core-i7-7700k-processor-8m-cache-up-to-4-50-ghz.html

[2] https://it.msi.com/Graphics-Card/GeForce-GT-1030-2G-LP-OC/Overview

[3] https://it.msi.com/Graphics-Card/N750Ti-TF-2GD5OC

[4] https://www.evga.com/products/specs/gpu.aspx?pn=2a951cf6-ed22-467b-9b00-57b0926f87ee

[5] https://it.msi.com/Graphics-Card/GeForce-GTX-1080-GAMING-X-Plus-8G

- Dell XPS 15; CPU: i7 6700HQ, RAM 16GB, GPU: GTX 960M;
- HP OMEN; CPU: i7 8750h, RAM 16GB, GPU: RTX 2060 (mancheranno dei dati specifici per questa macchina);

Workstation Dell

- CPU 2x Xeon 6256[6] (12 core ciascuno);
- RAM 128GB;
- GPU RTX 3090[7];

Per ogni versione del programma verranno testate le performance nel riconoscimento dei volti nel video sorgente e destinazione (tempo impiegato in minuti e secondi) ed il numero di iterazioni eseguite dopo un lasso di tempo (1 ora).

5.1 DeepFaceLab_DirectX12

Come specificato nel precedente capitolo, DeepFaceLab, è diviso in 3 versioni differenti adatte alla tipologia di scheda video che possediamo. Poiché questa versione ci consente di utilizzare tutte le schede è stata scelta per confrontare tutte le GPU testate.

GPU	Tempo (min:sec)	VRAM MAX	Utilizzo medio GPU
GTX 960M	17:50	586MB	73%
GT 1030	15:35	607MB	91%
GTX 750 Ti	11:38	711MB	89%
GTX 750 Ti OC	10:15	700MB	90%
GTX 980	5:30	824MB	77%
RTX 3090	4:50	3228MB	8%
RTX 2060	4:37	N/A	N/A
GTX 1080	3:19	798MB	59%

Tabella 5.1: Data_src faceset Extract DeepFaceLab_DirectX12

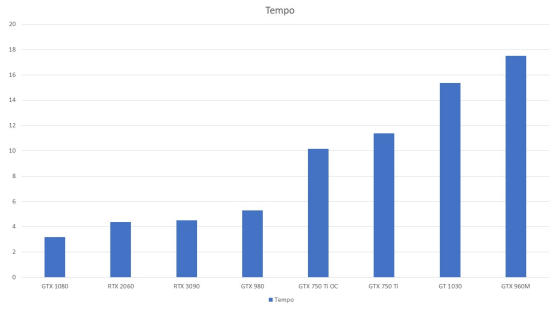

Figura 5.1: Grafico Data_src faceset Extract DeepFaceLab_DirectX12

[6] https://ark.intel.com/content/www/it/it/ark/products/198655/intel-xeon-gold-6256-processor-33m-cache-3-60-ghz.html
[7] https://www.nvidia.com/it-it/geforce/graphics-cards/30-series/rtx-3090/

Nella grafico presente in figura 5.1 possiamo osservare le tempistiche delle varie schede a confronto

GPU	**Tempo (min:sec)**	**VRAM MAX**	**Utilizzo medio GPU**
GTX 960M	46:51	566MB	71%
GT 1030	39:45	1199MB	91%
GTX 750 Ti	29:15	717MB	86%
GTX 750 Ti OC	25:58	701MB	95%
RTX 3090	14:54	3219MB	9%
GTX 980	14:07	843MB	74%
RTX 2060	12:23	N/A	N/A
GTX 1080	9:00	812MB	55%

Tabella 5.2: Data_dst faceset Extract DeepFaceLab_DirectX12

Poiché il video è più lungo e sono stati estratti più frame il tempo aumenta considerevolmente.

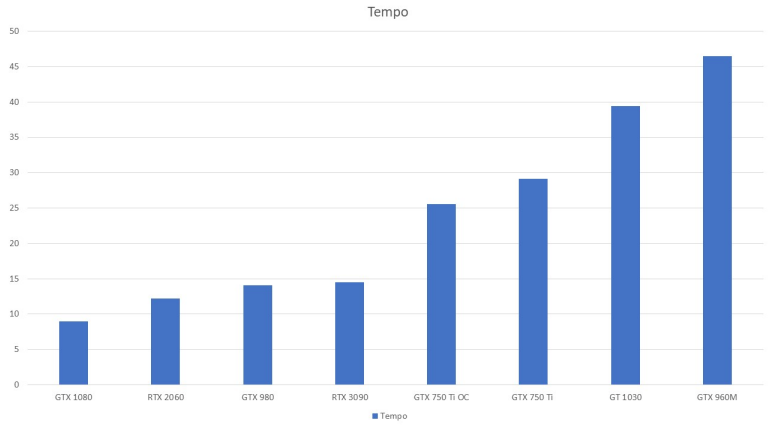

Figura 5.2: Grafico Data_dst faceset Extract DeepFaceLab_DirectX12

Il grafico presente in figura 5.2 ci mostra il confronto delle sole tempistiche con tutte le schede video testate.

Per il confronto delle iterazioni è stato scelto il Quick96 per i motivi spiegati nel capitolo 4.

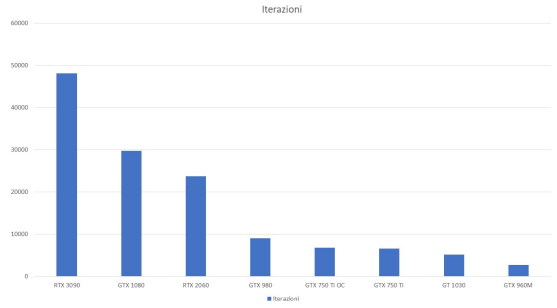

Figura 5.3: Grafico Iterazioni DeepFaceLab_DirectX12

GPU	Iterazioni	VRAM MAX	Utilizzo medio GPU	Utilizzo CPU
GTX 960M	2731	586MB	73%	100%
GT 1030	5143	1118MB	75%	35-40%
GTX 750 Ti	6570	717MB	68%	60-75%
GTX 750 Ti OC	6787	740MB	62%	60-75%
GTX 980	8981	876MB	36%	100%
RTX 2060	23697	N/A	N/A	N/A
GTX 1080	29802	2.45GB	95%	30-40%
RTX 3090	48174	3.06GB	58%	N/A

Tabella 5.3: Iterazioni dopo 1 ora DeepFaceLab_DirectX12

La tabella ed il grafico presente il figura 5.3 ci permette di vedere quante iterazioni in un'ora una determinata scheda riesca a compiere. Il numero delle iterazioni è più in linea con l'effettiva potenza delle schede rispetto a quelli delle estrazioni delle immagini.

Osservazioni

- La RTX 3090 è chiaramente la GPU più prestante fra quelle testate, tuttavia presenta dei problemi nella fase di estrazione dei volti dai frame. Come visibile nella sezione relativa all'utilizzo della scheda è possibile notare come venga riportato un utilizzo basso della GPU. Per risolvere questa problematica è necessario selezionare, nella fase di estrazione, la CPU anziché la GPU. Così facendo la GPU verrà utilizzata al 100% e le performance aumenteranno drasticamente (lo vedremo in seguito).

- Nella fase di training, nelle schede più vecchie, la CPU viene utilizzata al 100% mentre, in quelle più recenti, decisamente meno. Questo avviene poiché le GPU più vecchie non supportano l'Hardware Accelerated GPU Scheduling[8]. Questa funzionalità, supportata dalle schede video Nvidia dalla serie 1000 (compresa) in poi, consente di distribuire il carico di lavoro efficientemente aumentando le performance in determinati scenari. Questo è uno di quelli. Il "toggle" visibile

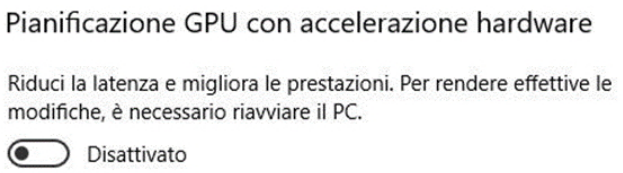

Figura 5.4: Il toggle da abilitare per attivare l'Hardware Accelerated GPU Scheduling

in figura 5.4 è visibile solo se si possiede una scheda grafica compatibile. Aver presente l'esistenza di questa features è fondamentale: con l'Hardware Accelerated GPU Scheduling è possibile lavorare pur avendo una CPU poco performante mentre, se si dispone di una GPU vecchia ma performante (ES. GTX 980 Ti) ed

[8] https://www.gamersnexus.net/guides/3599-windows-10-hardware-accelerated-gpu-scheduling-benchmarks

una CPU debole, si limiteranno notevolmente le performance della scheda nella fase di training (come è avvenuto con la GTX 980 nei test).

- La GTX 750 Ti è stata l'unica scheda testata sia a frequenze stock sia overcloccata. Lo scopo era dimostrare se effettivamente aumentare le frequenze operative potesse aumentare o meno le performance. Dai test effettuati un miglioramento c'è. Tuttavia, tale miglioramento, si verifica anche solo aumentando la frequenza delle memoria della scheda: l'aumentare la frequenza del core della GPU non apporta alcun beneficio tangibile.

5.2 DeepFaceLab_NVIDIA_up_to_RTX2080Ti

In questa sezione verranno testate alcune delle schede adatte a questa versione del software e verranno paragonati i risultati con la versione "adatta a tutte".

GPU	**Tempo (min:sec)**	**VRAM MAX**	**Utilizzo medio GPU**
GTX 960M	15:58	1622MB	57%
GT 1030	14:20	1.9GB	89%
GTX 980	5:07	3.56GB	63%
GTX 1080	3:24	5.16GB	65%

Tabella 5.4: Data_src faceset Extract DeepFaceLab_NVIDIA_up_to_RTX2080Ti

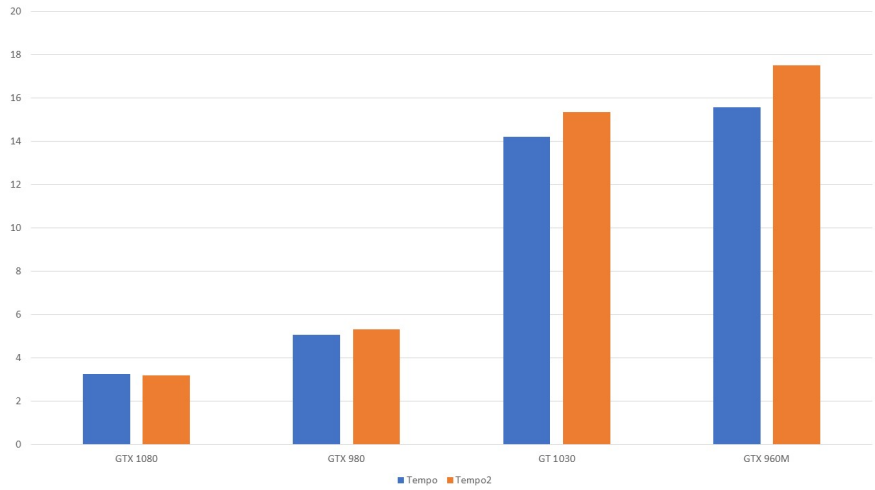

Figura 5.5: Confronto tempistiche fra le due build per l'estrazione da src

Nella figura 5.5 si possono osservare i risultati dell'estrazione con questa build (in azzurro) confrontati con i risultati della precedente build (in arancione).

Nella figura 5.6 è possibile vedere le tempistiche di questa build (in azzurro) confrontate con quelle della precedente build (in arancione).

Nella figura 5.7 è possibile notare le iterazioni effettuate in ora dalle schede presenti confrontate con l'altra build. E' possibile notare come ci sia un miglioramento delle performance con ogni GPU.

GPU	Tempo (min:sec)	VRAM MAX	Utilizzo medio GPU
GTX 960M	41:43	1622MB	51%
GT 1030	35:53	1.9GB	87%
GTX 980	13:38	3.56GB	54%
GTX 1080	9:02	5.18GB	61%

Tabella 5.5: Data_dst faceset Extract DeepFaceLab_NVIDIA_up_to_RTX2080Ti

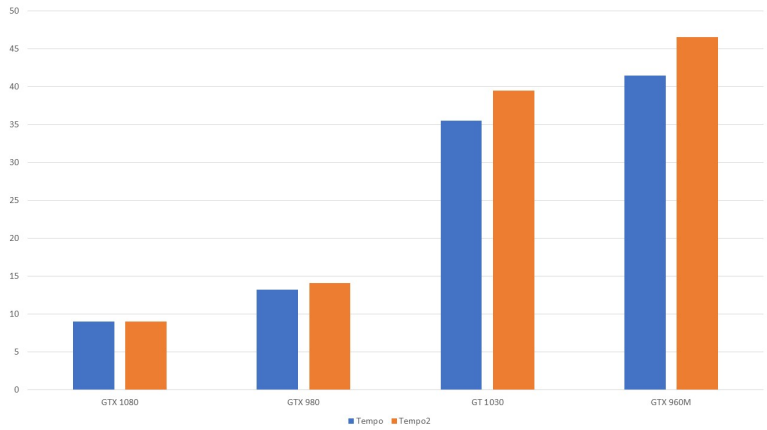

Figura 5.6: Confronto tempistiche fra le due build per l'estrazione da dst

GPU	Iterazioni	VRAM MAX	Utilizzo medio GPU	Utilizzo CPU
GTX 960M	3064	1.62GB	21%	100%
GT 1030	6557	1.94GB	96%	35-40%
GTX 980	10938	3.62GB	35%	100%
GTX 1080	32985	7.39GB	95%	20-35%

Tabella 5.6: Iterazioni dopo 1 ora DeepFaceLab_NVIDIA_up_to_RTX2080Ti

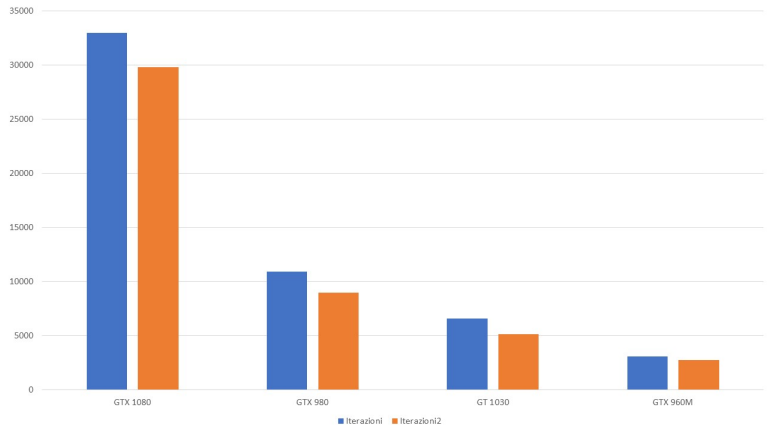

Figura 5.7: Confronto delle iterazioni effettuate dalle schede fra le due build

Figura 5.8: Versione DeepFaceLab_DirectX12

Figura 5.9: Versione DeepFaceLab_NVIDIA_up_to_RTX2080Ti

Osservazioni

- Dalle due build testate è possibile notare un enorme differenza di utilizzo della memoria video disponibile: nella prima build, in molte schede, viene utilizzato circa 1GB di memoria mentre, nella seconda build, viene quasi saturata tutta la memoria.
- Questa versione, come scritto nel capitolo precedente, è adatta dalle schede video Nvidia dalla serie 900 compresa in poi. Nella fase di estrazione dei volti dai frame è possibile osservare un leggero incremento delle performance. Tuttavia, più la scheda video a disposizione è potente, più il miglioramento diminuisce fino ad arrivare al margine di errore. Ciò non avviene per la fase di training dove il miglioramento è più marcato ed è presente in ogni scheda.
- Nonostante le differenze prestazionali e di utilizzo delle risorse hardware il video risultante, con le impostazioni di default ed il Quick96, ha differenze minime quindi la scelta della versione è solo relativa alle performance. Nelle figure 5.8 e 5.9 è stato preso un frame del video Tony Stark - Elon Musk creato con entrambe le versioni, entrambe con gli stessi settaggi ed il training di 10 mila iterazioni effettuato con il Quick96. Ad esclusione della forma del che è stato "attaccato" sono identici, come lo è la qualità.

5.3 DeepFaceLab_NVIDIA_RTX3000_series

Per questa versione del software verrà testata l'unica scheda video della serie RTX3000 ossia la RTX 3090.

GPU	**Tempo (min:sec)**	**VRAM MAX**	**Utilizzo medio GPU**
RTX 3090	7:51	6.77GB	18%

Tabella 5.7: Data_src faceset Extract DeepFaceLab_NVIDIA_RTX3000_series

Osservazioni

- Qui, come nei precedenti test, l'utilizzo della CPU per la RTX 3090 non è mostrato poiché era installata in un sistema con 24 cores e sarebbe stato irrilevante anche perché questa scheda supporta l'hardware accelerated GPU scheduling.

GPU	Tempo (min:sec)	VRAM MAX	Utilizzo medio GPU
RTX 3090	17:37	6.6GB	20%

Tabella 5.8: Data_dst faceset Extract DeepFaceLab_NVIDIA_RTX3000_series

GPU	Iterazioni	VRAM MAX	Utilizzo medio GPU	Utilizzo CPU
RTX 3090	39075	10.39GB	38%	NA

Tabella 5.9: Iterazioni dopo 1 ora DeepFaceLab_NVIDIA_RTX3000_series

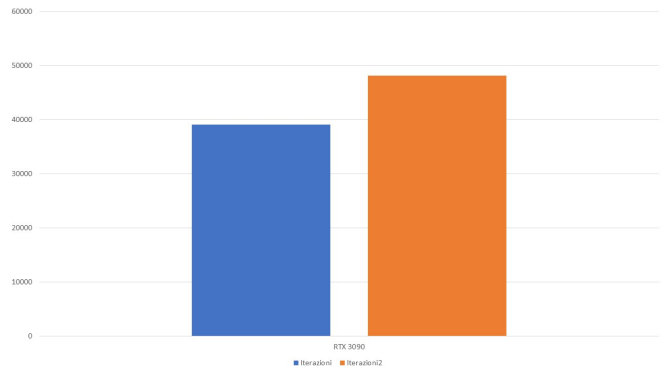

Figura 5.10: Confronto delle iterazioni con la build DirectX12 (in arancio) con l'attuale (in azzurro)

- Come nella build indicata per le schede video alla RTX 2080Ti qui si ha un aumento della memoria video utilizzata. La RTX 3090 ha 24GB di memoria quindi siamo ben lontani dal saturarla. Però, al contrario dell'altra versione, qui assistiamo ad un peggioramento complessivo delle performance: sia nella fase di estrazione sia nella fase di training come mostrato in figura 5.10. Va ricordato però che la serie di GPU Nvidia RTX 3000 è recente e lo sviluppatore gli ha dedicato una versione a parte quindi ci si aspettano dei miglioramenti nelle prossime release.

- Al contrario della build DirectX12, per la fase di estrazione, qui non è possibile impostare la CPU e far utilizzare al massimo la GPU. Qui la CPU, come indica il comando, sostituisce direttamente la GPU. Questo potrebbe far intuire un bug nella versione DirectX12 poiché qui, ma anche nell'altra versione, funziona correttamente.

5.4 Confronto modelli

Per confrontare i modelli verranno utilizzati due parametri: prestazioni e qualità. Per l'analisi prestazionale utilizzeremo una GTX 1080 per tutti e tre i modelli con le impostazioni di default.

Nella tabella possiamo vedere come il modello più performante è il Quick96 mentre quello meno performante è il più recente AMP. L'AMP, però, è molto personalizzabile e quindi, modificando le varie opzioni di default, è possibile aumentarne le performance.

GPU	**Iterazioni**	**VRAM MAX**	**Utilizzo medio GPU**	**Utilizzo CPU**
Quick96	32985	7.39GB	95%	20-40%
SAEHD	10396	7.71GB	96%	30-50%
AMP	6213	7.59GB	92%	20-40%

Tabella 5.10: Confronto prestazionale modelli

Ma il numero delle iterazioni non dice nulla se non si ha un riscontro visivo del risultato. Nelle figure 5.11, 5.12 e 5.13 abbiamo un confronto visivo tra i vari modelli con un numero simile di iterazioni.

Figura 5.11: Quick96 con circa 10k iterazioni

Figura 5.12: SAEHD con circa 10k iterazioni

Figura 5.13: AMP con circa 12k iterazioni

Nonostante il basso numero di iterazioni possiamo osservare come il SAEHD riesca a catturare maggiormente i dettagli ed i lineamenti del volto. Il Quick96 però, si difende bene soprattutto considerando che è 3 volte più veloce. Se si ignora il tempo e si

sale maggiormente con le iterazioni il SAEHD produrrà un video migliore tuttavia, il Quick96, per produzioni amatoriali con potenza hardware limitata (e video in condizioni ideali) si conferma essere un ottimo modello. Il nuovo modello "AMP", invece, è un disastro. Non solo è il più lento dei 3 ma il risultato, nonostante circa 2000 iterazioni in più, è pessimo. Probabilmente necessita di un numero maggiore di iterazioni e di un tuning manuale ma, al momento della scrittura della tesi, non è un modello che mi sento di raccomandare.

Nel caso non si abbia tempo e/o risorse per creare dei modelli da zero è possibile scaricare dei modelli dai vari forum dedicati a "DeepFaceLab". Tuttavia bisogna tener presente che la compatibilità rischia di essere un problema. Alcuni modelli sono stati allenati solo per determinate tipologie di persone o per una persona specifica e quindi non possono essere utilizzati per tutte le tipologie di video da realizzare. Inoltre, tutti i modelli pre-allenati, sono SAEHD. Questo vuol dire che possiamo trovarci di fronte un problema di compatibilità. Se infatti un determinato modello soddisfa i nostri requisiti per la realizzazione di un video (sesso, etnia ecc) ma è stato allenato, per esempio, con una GPU da 16GB di memoria video non sarà possibile utilizzarlo sulla nostra GPU da 8GB.

5.5 Costi

Avere maggiore potenza hardware è conveniente? Ne vale la pena? Nella tabella sono

GPU	**Presso MSRP $**	**Prezzo Ebay €**
GT 1030	80	78
GTX 750 Ti	149	84
GTX 980	549	235
GTX 1080	599 (FE 699)	458
RTX 3090	1499	1744
GTX 1060 6GB	249 (FE 299)	285

Tabella 5.11: Confronto costi GPU

state inserite tutte le GPU testate ad esclusione di quelle presenti nei PC portatili. Questo perché non si possono comprare singolarmente le schede video dei portatili essendo saldate ed è impossibile avere un prezzo indicativo per via delle differenti specifiche dei vari modelli di laptop presenti sul mercato con una specifica scheda video installata. E' stata aggiunta anche la GTX 1060 6GB. Questa scheda non è stata testata e quindi non si hanno informazioni sulle performance in questo programma, tuttavia è la scheda che raccomando di comprare nel caso si voglia iniziare a realizzare contenuti con questo software. I motivi per cui la raccomando sono tre:

- Supporta l'hardware accelerated GPU scheduling;
- Dispone di 6GB di memoria video;
- Bassi consumi (TDP 120W);

Il che vuol dire che se si ha una CPU poco performante ed un alimentatore con basso wattaggio si può installare questa scheda nel proprio PC ed iniziare a creare contenuti

con questo programma senza alcun problema. Inoltre, con 6GB di memoria video, ci lascia un poco di margine per creare modelli con il SAEHD. Al di fuori di questo programma, la GTX 1060 6GB, ha una potenza computazionale simile alla GTX 980 ma grazie al supporto all'hardware accelerated GPU scheduling garantisce maggiori performance. Da non confondere con la GTX 1060 3GB: quest'ultima non solo ha la metà della memoria ma è anche meno performante.

Nel grafico in figura 5.14 è possibile notare come sia aumentato il valore di determinate schede nonostante fossero usate. Questo a causa dell'attuale periodo di shortages causato dallo smart-working ma soprattutto a causa del mining. E' stato effettuato un cambio 1:1 per i prezzi relativi al nuovo.

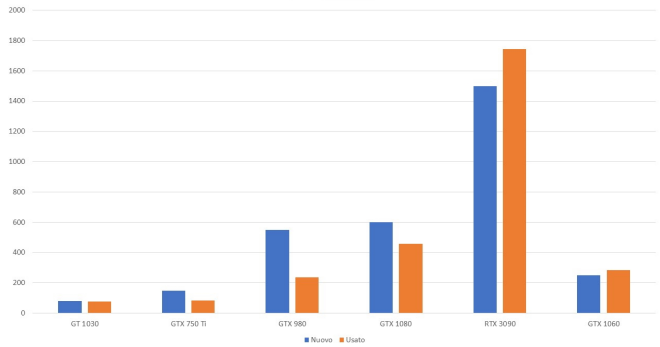

Figura 5.14: Confronto Prezzi GPU (blu nuovo, arancione usato)

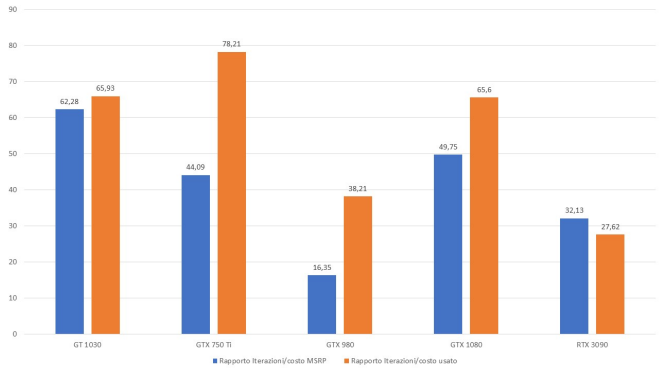

Figura 5.15: Rapporto Iterazioni/Costo (blu nuovo, arancione usato)

In base al numero delle iterazioni effettuate con la versione DirectX12 del programma, il grafico in figura 5.15 ci consente di vedere quale scheda si riveli la più conveniente se non si vogliono realizzare video di qualità (e quindi non si guarda la memoria video a disposizione) ma si vogliono realizzare meme o video di bassa qualità. Va ricordato che per aumentare le performance di determinate schede basta utilizzare la versione del programma adatta e/o overcloccare le memorie della GPU.

5.6 Performance CPU

In base alla versione del programma è possibile utilizzare la CPU nella fase di estrazione delle immagini. Nella versione DirectX12 va a ridurre il tempo di estrazione

ma non sostituisce la GPU. Nelle altre versioni, invece, sostituisce la GPU e quindi le performance dipenderanno unicamente dal tipo di CPU che si ha.

GPU	**Tempo (min:sec)**	**VRAM MAX**	**Utilizzo medio GPU**
GTX 980	4:20	798MB	100%
RTX 3090	2:27	3228MB	93%

Tabella 5.12: Data_src faceset Extract DeepFaceLab_DirectX12 con CPU

GPU	**Tempo (min:sec)**	**VRAM MAX**	**Utilizzo medio GPU**
GTX 980	10:41	812MB	99%
RTX 3090	5:57	3219MB	95%

Tabella 5.13: Data_dst faceset Extract DeepFaceLab_DirectX12 con CPU

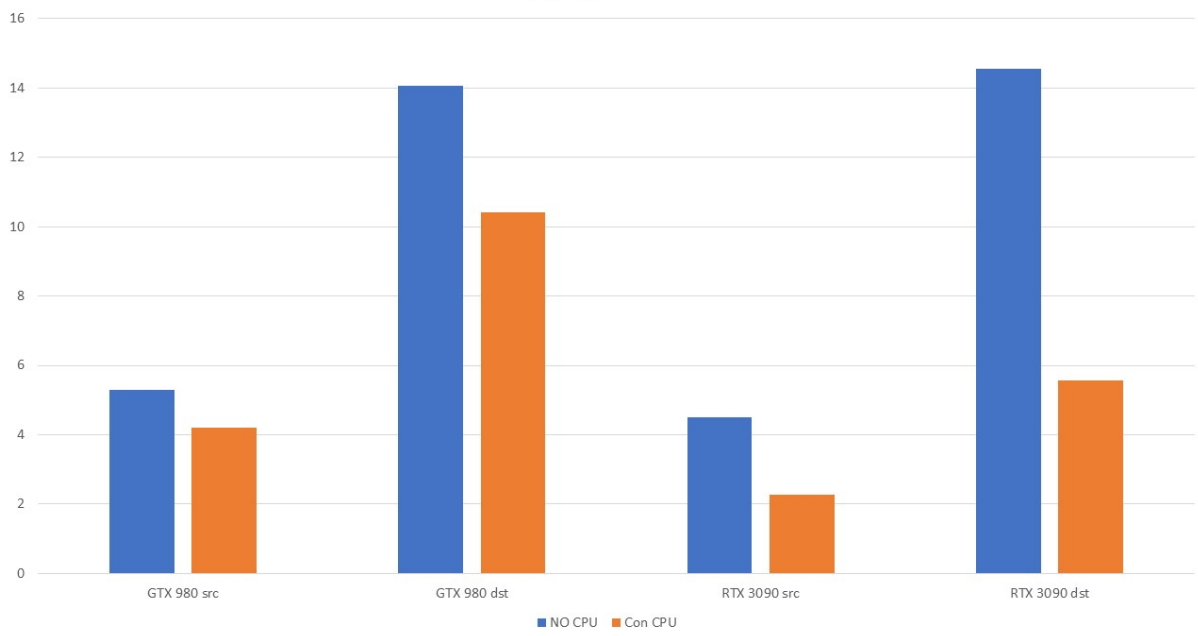

Figura 5.16: Tempi estrazione volti da src e dst con (arancione) e senza CPU (blu)

Come visibile nel grafico in figura 5.16 c'è un notevole aumento delle performance nel selezionare la CPU in questa versione del programma. Dalle tabelle precedenti si nota come la GPU venga utilizzata al massimo e, nel caso della RTX 3090, la differenza in fatto di performance è enorme. Va ricordato, però, che la CPU, pur selezionandola, non lavora al 100%: l'utilizzo medio è simile a quando viene selezionata la GPU. Molto probabilmente si tratta di un bug di questa versione ma consente di aumentare notevolmente le performance.

Come specificato nel precedente capitolo, utilizzare la CPU ha senso nel training solo se si punta alla massima qualità e si ha a disposizione tanta memoria a disposizione. Per i test verrà utilizzato un i7 7700k @4.9ghz con 16GB di RAM a velocità e latenze

Velocità Ram	**Latenza**	**Iterazioni dopo 1 ora**
1333MT/s	CL 9-9-9-28 2T	2032
1600MT/s	CL 9-9-9-28 2T	2284
1866MT/s	CL 10-10-10-30 2T	2452
2133MT/s	CL 11-11-11-35 2T	2532
2500MT/s	CL 11-12-12-35 2T	2676

Tabella 5.14: Training CPU con ram a velocità e latenze differenti

differenti. E' possibile apprendere dai dati presenti nella tabella 5.14 che, come per le GPU, maggiore è la velocità della memoria maggiori saranno le performance. Il che vuol dire che, se si vuol fare training con la CPU, bisogna cercare di avere si più RAM possibile ma bisogna far si che sia alla velocità massima che la CPU supporta. Il costo sarà maggiore ma se si vuol fare per forza il training con la CPU per ottenere il massimo della qualità probabilmente il budget non è un problema.

5.7 Merging

Il tempo di merging dipende da diversi fattori: lunghezza del video, risoluzione del video, effetti applicati. Per analizzare le performance nel merging è stato utilizzato il PC con CPU i7 7700k @4.9ghz, 16GB RAM e la GTX 980. Il video utilizzato è "Tony Stark - Elon Musk" con una rete allenata a 69k iterazioni.

Tipo	**Tempo (min:sec)**	**Utilizzo CPU**	**Utilizzo medio GPU**
Normale	1:21	85-100%	20%
Con effetto up-res 50%	3:29	35-50%	63%

Tabella 5.15: Tempo di merging

Per "normale" si intende l'applicazione di effetti base che si limitano ad "incollare" al meglio il nuovo volto (gli effetti disponibili nel merging sono visibili nel capitolo precedente). Va fatto notare come alcune effetti, come l'effetto "up-res" che aumenta artificialmente la risoluzione del volto, possano aumentare il tempo necessario per completare il merging. La particolarità è che, come visibile in tabella, nel primo caso basterebbe avere una CPU più performante per ridurre le tempistiche (notare l'alto utilizzo della CPU con il basso utilizzo della CPU) mentre, con l'effetto "up-res" non è chiaro quale componente eventualmente aggiornare: CPU e GPU vengono utilizzate, ma non al 100% e quindi o il collo di bottiglia è altrove o il software non è ottimizzato a dovere.

5.8 Conclusione analisi performance

In base ai test effettuati è possibile farsi un idea su quanto tempo necessiti la creazione di un video (che vedremo nel capitolo successivo). Visto l'hardware testato è possibile stilare una lista di componenti in base al budget ed a quello che si vuol fare. Se il budget non è un problema:

- CPU Xeon Platinum 8380 40 cores / 80 threads ;
- RAM il più possibile DDR4-3200 (velocità massima supportata dalla CPU);
- GPU RTX A6000[9];

Escludendo la ram, solo con CPU e GPU siamo a circa 12 mila euro di spesa. La RTX A6000, visibile in figura 5.17, è una RTX 3090 leggermente più performante con il doppio della memoria video (48GB) ad un costo di circa 5000 euro. Inoltre ricordo che le configurazioni multi-socket vanno evitate.

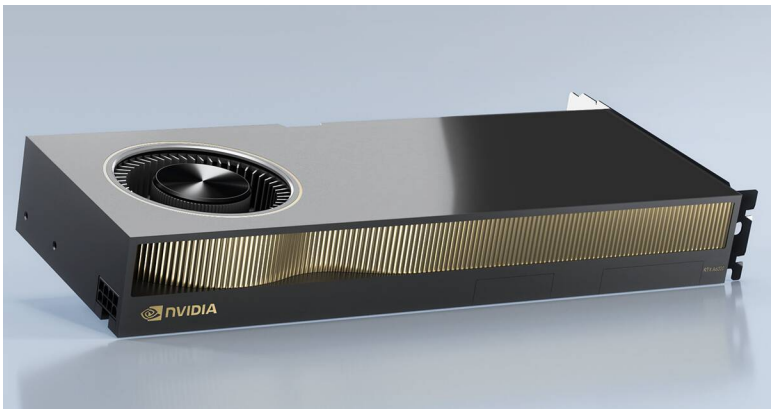

Figura 5.17: Nvidia RTX A6000: Attualmente la miglior GPU per questa tipologia di applicazioni [Buz]

Se si volesse aggiornare un sistema già esistente, come già detto consiglio la GTX 1060.

Nel caso, invece, si volesse creare un sistema da zero:

- CPU Intel Core i7 11700k 8 cores / 16 threads[10];
- RAM 16/32GB DDR4-3200 o più veloce;
- GPU RTX 3060 12GB[11];

La scelta della CPU è ricaduta sul Core i7 11700k per il supporto all'Intel Deep Learning Boost. Riguardo alla quantità di ram la scelta dipende dall'utente mentre per la velocità c'è un discorso più ampio da porre. Le CPU Intel di 11esima generazione dispongono di un nuovo controller della memoria il cui clock ha due modalità: sincronizzato con la frequenza della memoria (Gear 1) o in rapporto 1:2 con la memoria (Gear 2). In modalità "Gear 1", in base alla "lotteria del silicio"[12], è possibile arrivare alle frequenza di 1700-1800mhz e quindi supportare memorie con frequenza

[9] https://www.nvidia.com/it-it/design-visualization/rtx-a6000/
[10] https://ark.intel.com/content/www/us/en/ark/products/212047/intel-core-i7-11700k-processor-16m-cache-up-to-5-00-ghz.html
[11] https://www.nvidia.com/it-it/geforce/graphics-cards/30-series/rtx-3060-3060ti/
[12] https://en.wikipedia.org/wiki/Product_binning#Overclocking_and_core_unlocking

3400-3600MT/S (va ricordato che 3400-3600 è il doppio della frequenza reale trattandosi di memorie DDR). In modalità "Gear 2", per le stesse memorie, il controller lavorerebbe ad una frequenza di 850-900mhz penalizzando notevolmente le performance. La seconda modalità si rivela utile per utilizzare memorie ad altissima frequenza (5000mhz o più) sempre che la scheda madre non abbia problemi a supportarle. Vista l'impossibilità nel testare tali configurazioni consiglio memorie da almeno 3200MT/S che la CPU supporta senza problemi in "Gear 1". Infine, riguardo la GPU, la scelta è ricaduta sulla RTX 3060. La scheda garantisce performance di poco superiori alla GTX 1080 che abbiamo testato ma ha soprattutto 4GB di memoria video in più garantendoci maggiore flessibilità con i modelli SAEHD. Utilizzando il sito PCPartPicker[13] ho creato un'ipotetica configurazione di questo tipo con 32GB di ram, visibile in figura 5.18.

Component	Selection		Base		Price
CPU		Intel Core i7-11700K 3.6 GHz 8-Core Processor	€386.88	—	€386.88
CPU Cooler		Noctua NH-U12S redux 70.75 CFM CPU Cooler	€49.90	—	€49.90
Motherboard		ASRock Z590 Phantom Gaming 4 ATX LGA1200 Motherboard	€160.99	—	€160.99
Memory		G.Skill Ripjaws V 32 GB (2 x 16 GB) DDR4-3600 CL19 Memory	€187.58	—	€187.58
Storage		Crucial BX500 1 TB 2.5" Solid State Drive	€89.89	—	€89.89
Storage		Samsung 980 Pro 250 GB M.2-2280 NVME Solid State Drive	€79.78	—	€79.78
Video Card		Zotac GeForce RTX 3060 12 GB GAMING Twin Edge OC Video Card	€702.48	—	€702.48
Case		Cooler Master MasterBox K501L ATX Mid Tower Case	€42.99	—	€42.99
Power Supply		SHARKOON SilentStorm Cool Zero 650 W 80+ Gold Certified Fully Modular ATX Power Supply	€94.70	—	€94.70
				Total:	**€1795.19**

Figura 5.18: Ipotetica configurazione per assemblare un nuovo computer per questa tecnologia

Nella configurazione sono stati aggiunti gli altri componenti che possono essere sostituiti con alcuni più economici o inferiori (ssd più lento e/o meno capiente, alimentatore con wattaggio inferiore ecc) ma in generale stiamo parlando di una configurazione, al momento della scrittura di questa tesi, da 1500 a 2000 euro.

[13] https://it.pcpartpicker.com/list/dYPzQD

6. Realizzazione video

Analizzate le performance in ogni aspetto si passa alla realizzazione del video. In questo capitolo si parlerà di 3 video differenti realizzati descrivendo la scelta dei filmati (sorgente e destinazione) assieme alle problematiche riscontrate. Per ogni video verranno riportate le tempistiche approssimative senza tener conto di eventuali esperimenti effettuati[1].

6.1 Video 1: Tony Stark - Elon Musk

Nell'analisi delle performance, per standardizzare la metodologia di testing, abbiamo utilizzo i due video che il programma include. Nelle immagini 6.1 e 6.2 possiamo vedere due screen tratti dai due video.

Uno screen del video risultate lo abbiamo visto nel capitolo precedente, tuttavia era tratto da un filmato realizzato con un basso numero di iterazioni. Nella figura 6.3 possiamo vedere un frame del video realizzato con questo software combinando i due.

Il software ci mette a disposizione questi due filmati per fare degli esperimenti poiché si adattano molto bene alla realizzazione di video. Nel video sorgente il soggetto viene ripreso in ogni angolo e si riescono a catturare tutti i dettagli del volto. Non si avranno a disposizione moltissimi frame (solo 655) però sono tutti buoni ed utilizzabili. Il video "destinazione" si rivela anch'esso un'ottima scelta. Il soggetto è sopra un palco (come il soggetto del video sorgente), non si muove molto, le condizioni di luce sono ottimali e vagamente simili a quelle del video sorgente. Anche con poche iterazioni e pochi frame, quindi, si riesce a realizzare un buon video. L'unica problematica è quando, nel video destinazione, Elon Musk si mette quasi di profilo. In questo caso

[1] https://drive.google.com/drive/folders/1mQgsmCO2bLfBCPZ_PwPNlL3cO2upesTq?usp=sharing

Figura 6.1: VIDEO 1: Screen dal video sorgente tratto da "Iron Man 2"

Figura 6.2: VIDEO 1: Screen dal video destinazione con Elon Musk sul palco

Figura 6.3: Frame tratto dal video risultante con 69 mila iterazioni

Figura 6.4: VIDEO 2: Screen dal video sorgente

Figura 6.5: VIDEO 2: Screen dal video destinazione

abbiamo qualche problema ma ci sono moltissime soluzioni per ovviare a questo piccolo difetto come provare ad aumentare il numero delle iterazioni o cambiare direttamente modello. Per questo filmato è stato utilizzato il Quick96 e siamo sopra le 2 ore di lavoro con quasi 70 mila iterazioni.

6.2 Video 2: Marcantoni - Trump

Testato il programma con i video che ci fornisce sono stati realizzati ulteriori video. Uno di questi prevedeva il porre il volto del professor Marcantoni su quello dell'ex presidente degli Stati Uniti Donald Trump. In questo caso noi dovevamo ottenere un filmato utilizzabile come video sorgente e trovare un video di destinazione di Donald Trump adatto ai nostri scopi.

Il professor Marcantoni ha acconsentito a registrarsi per poco più di 20 secondi e dal video realizzato è stato possibile ricavare 780 frames utilizzabili. In figura 6.4 possiamo vedere uno screen dal video registrato. E' evidente come tutti i frame sono di ottima qualità: abbiamo un primo piano del soggetto che si muove in ogni direzione, parla e ci consente di catturare molte sue espressioni facciali. Inoltre le condizioni di luce sono ottimali e la risoluzione è alta.

Per il video di destinazione invece è stato scelto un comunicato effettuato da Trump presso la Casa Bianca. La scelta è ricaduta su questo video per due motivi: il soggetto non si muoveva molto, il suo sguardo rimaneva rivolto verso la camera e le condizioni d'illuminazione erano simili. Il video ha una durata di circa 1 minuto ed è stato possibile ricavare 1456 frames tutti utilizzabili. Come visibile in figura 6.5 anche qui è stato possibile ricavare degli ottimi frames.

Figura 6.6: VIDEO 2: Frame tratto dal video risultante con 270 mila iterazioni circa

Nella figura 6.6 è possibile vedere un frame del video finale: il risultato è eccellente. Con 270 mila iterazioni siamo a circa 9 ore di lavoro. Dai test effettuati salire ulteriormente con il numero delle iterazioni non migliorerebbe di molto la qualità. Seppur abbiamo a disposizione dei frame buoni, il basso numero a disposizione ci limita. Per migliorare la qualità si potrebbe provare ad utilizzare il SAEHD ma, come visto nel capitolo precedente, ci vorrebbe più tempo (considerando la differenza con le impostazioni di default verrebbe triplicato). Inoltre, secondo il mio modesto parere, sarebbe dell'ulteriore lavoro quasi inutile poiché già il risultato è più che buono. Questo video, però, dimostra che pur avendo pochi ma buoni frame, si possa realizzare un video realistico scegliendo accuratamente il video di destinazione e con conoscenze ed hardware "limitato" poiché stiamo utilizzando il Quick96.

6.3 Video 3: Marcantoni - Hartman

Per il terzo video sono stati utilizzati i frame del video sorgente del secondo video (Fig. 6.4) ma questa volta è stato è stato scelto come video di destinazione una piccola parte del monologo del sergente Hartman tratto da "Full Metal Jacket"(figura 6.7). Il video di destinazione è corto, dura poco più di 10 secondi ed in una parte non si vede neanche il sergente ma solo un soldato. In questo caso sono stati rimossi i frame del soldato per far si che il DeepFake venisse applicato solo al sergente. In questo modo, però, ci ritroviamo solo 109 frame utilizzabili.

Figura 6.7: VIDEO 3: Screen dal video destinazione con il Sergente Hartman tratto da "Full Metal Jacket"

E' stata scelta questa parte del video per cercare di aumentare le possibilità di riuscita del video poiché avevamo un primo piano del soggetto tuttavia ci sono diversi elementi che rendono difficile la realizzazione del video:

- Condizioni di luce non ottimali;
- Il soggetto indossa un cappello che copre parte del volto;
- Il soggetto muove la mano davanti al volto rendendo difficile un merging realistico;

Questo video vuole dimostrare come possa essere difficile realizzare un DeepFake se non si tengono in considerazione i fattori descritti in precedenza. Con l'attuale tecnologia la realizzazione dei DeepFake è possibile ed i risultati sono eccellenti ma bisogna ricordarsi che non tutti i video sono ideali. Nella figura 6.8 è possibile vedere un frame di uno

Figura 6.8: VIDEO 3: Frame dal video finale dopo 470 mila iterazioni

screen del video. Il risultato non è soddisfacente: il dito si sovrappone al viso, si intravede chiaramente dove il viso viene "attaccato", la forma del volto del soggetto nel video di destinazione non è ideale e le condizioni di luce inadatte rendono ancor più chiaro il fatto che ci ritroviamo di fronte un video falso e di bassa qualità. Aumentare il numero delle iterazioni, quindi, non basta. Siamo a poco più di 470 mila e quindi a poco più di 15 ore di lavoro con una GTX 1080 con il modello Quick96. Nelle impostazioni di merging sono stati provati diversi effetti, specialmente quelli relativi al colore con scarsi risultati. Questa è la conferma che salire con le iterazioni è superfluo se si hanno determinate problematiche.

Figura 6.9: VIDEO 3: Frame dal video finale dopo poco più di 97 mila iterazioni con SAEHD

Come visibile in figura 6.9 neppure passando al SAEHD si riesce a migliorare la situazione. Il video realizzato proviene da una rete con poco più di 97 mila iterazioni (più di 9 ore di lavoro). Rispetto al Quick96 l'immagine è meno definita (abbiamo

Figura 6.10: VIDEO 2: Frame tratto dal video creato con circa 10 mila iterazioni

Figura 6.11: VIDEO 2: Frame tratto dal video risultante con 270 mila iterazioni circa

comunque meno di un quarto delle iterazioni rispetto al Quick96) ma aumentando il numero di iterazioni questo potrà cambiare, i contorni sono delineati meglio e nel complesso la gestione della luce è migliore ma resta comunque un risultato insufficiente.

Conclusione

Come è stato possibile vedere dai tre video demo realizzati il processo di realizzazione di un DeepFake, seppur semplice all'apparenza, necessita di un minimo di competenza ed un accurata selezione del video di destinazione se si vuole puntare alla qualità. Non esiste un numero di iterazioni ideale da raggiungere: in base ai test effettuati 40-50 mila iterazioni sono un buon punto di partenza per la realizzazione di un video in condizioni "ideali" (descritte in precedenza). Se ci si trova, appunto, in queste condizioni "ideali" più iterazioni si fanno e meglio è ma, dopo un certo numero di iterazioni, noteremo che i guadagni in termini di qualità inizieranno a diminuire sempre di più. Sta quindi all'utente decidere quando fermarsi. Invece, se il video non è adatto per svariati motivi, salire con le iterazioni equivale semplicemente a perdere tempo. Fortunatamente è possibile capire se il video è adatto o meno alla realizzazione di un DeepFake anche con poche iterazioni. Nelle immagini in figura 6.10 e 6.11 mostrano due screen tratti da due filmati realizzati con un numero differente di iterazioni. Il singolo frame non rende giustizia: a 270 mila iterazioni la qualità è decisamente migliore, i movimenti sono più realistici, non è presente alcuno sfarfallio e vengono aggiunti particolari, come il battito delle palpebre, per rendere il tutto più realistico. Tuttavia, il risultato a 10 mila iterazioni, pur essendo qualitativamente insufficiente, mostra come il video abbia del "potenziale" per diventare realistico.

7. Conclusioni

In conclusione, dopo aver analizzato ogni aspetto del DeepFake, è possibile dire con assoluta certezza che, con il passare del tempo, questa tecnologia verrà utilizzata sempre di più. Abbiamo visto come già in passato questa tecnologia sia stata utilizzata per determinati scopi e come la teoria dietro il DeepFake sia vecchia di oltre un decennio. Tuttavia ci sono due motivi per cui, da qualche anno, sia aumentata a dismisura la quantità di contenuti DeepFake presenti in rete. Quantità che continua ad aumentare di giorno in giorno. Le motivazioni sono:

- Sviluppo tecnologico;
- Aumento dei contenuti multimediali (foto/video) presenti in rete;

Sviluppo Tecnologico

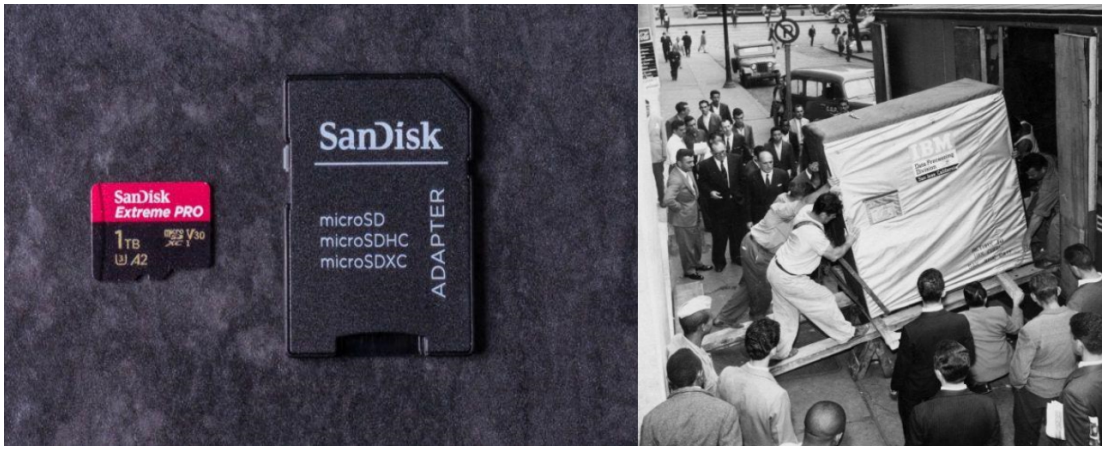

Figura 7.1: A Sinistra una Micro SD da 1TB a destra un HDD da 5MB prodotto dalla IBM (1956)

Quando si pensa allo sviluppo tecnologico, specialmente nell'informatica, si pensa ad un cambiamento radicale come quello mostrato in figura 7.1. Si è passati da Hard Disk così ingombranti e con capacità ridotte a delle Micro SD con capacità decisamente superiori e che possiamo tenere fisicamente con un dito. Non tutti i cambiamenti, però, sono così eclatanti. Una delle ragioni per cui i video "DeepFake" sono sempre più presenti riguarda lo sviluppo tecnologico di altri componenti: CPU, GPU e RAM. Queste tre componentistiche, presenti in ogni computer, nel corso degli anni, esteticamente non sono cambiate molto. Nella figura 7.2 possiamo vedere due CPU di generazioni differenti:

Figura 7.2: A Sinistra un Pentium 4 672 a destra un Core i9 10900k

un Intel Pentium 4 672[1] ed un Intel Core i9 10900k[2]. Escludendo la nomenclatura che, bene o male è entrata nel gergo comune, non ci sono grandi differenze estetiche fra i due. Anche le dimensioni sono simili. Eppure, fra i due, c'è un enorme differenza in fatto prestazionale[3][4]. Nelle RAM e GPU qualche piccola differenza estetica è possibile notarla nelle dimensioni e colore del dissipatore integrato. In sostanza, con lo sviluppo tecnologico, abbiamo a disposizione dei componenti molto simili, in fatto di dimensioni ed estetica, a quelli presenti in passato ma con una "potenza" decisamente superiore. Avere a disposizione tanta potenza hardware ha fatto si che chiunque potesse iniziare a sfruttare questa tecnologia per i suoi scopi. Nel capitolo riguardante l'analisi delle performance abbiamo anche visto come poter sfruttare al meglio l'hardware che si ha (overclocking, scelta modelli ecc). Questo ha fatto si che, quindi, la realizzazione di contenuti multimediali, anche di qualità, possa essere eseguita anche da privati e non più da grosse aziende (Es. Case cinematografiche).

Aumento dei contenuti multimediali

"La potenza è nulla senza controllo" recitava lo slogan della Pirelli per sponsorizzare i propri pneumatici. Qui il discorso è molto simile: non serve a nulla avere il miglior hardware in commercio se poi non si hanno a disposizione i contenuti multimediali. CPU performanti e GPU potenti con tanta memoria esistevano anche diversi anni fa, eppure i contenuti DeepFake presenti in rete erano pochi. Il secondo motivo per cui sono aumentati esponenzialmente i video DeepFake in rete è a causa dell'aumento dei contenuti multimediali di determinate persone presenti nei vari social. I Social

[1] https://ark.intel.com/content/www/it/it/ark/products/27488/intel-pentium-4-processor-672-supporting-ht-technology-2m-cache-3-80-ghz-800-mhz-fsb.html

[2] https://ark.intel.com/content/www/it/it/ark/products/199332/intel-core-i9-10900k-processor-20m-cache-up-to-5-30-ghz.html

[3] https://www.cpubenchmark.net/cpu.php?cpu=Intel+Pentium+4+3.80GHz&id=1081

[4] https://www.cpubenchmark.net/cpu.php?cpu=Intel+Core+i9-10900K+%40+3.70GHz&id=3730

Network esistono da anni ma, in passato, non si pubblicavano costantemente foto e video. Pensiamo a Twitter dove si pubblicavano solo dei post (o meglio dei Tweet) con un numero di caratteri pre-stabilito. Oppure a Facebook dove era comune postare semplicemente qualche riga ed al massimo una foto. Perfino su Instagram, social nato con lo scopo di far pubblicare all'utente solo delle immagini in un determinato formato, era raro trovare profili con tante foto. Oggi è tutto cambiato. Si pubblicano una quantità eccessiva di foto, ma soprattutto di video nei vari social. Social che sono diventati sempre più numerosi. L'aumento dei contenuti multimediali Deepfake presenti in rete è dipeso anche da questo.

In conclusione, con l'ulteriore aumento dei contenuti multimediali facilmente reperibili nei vari social media e l'ulteriore sviluppo tecnologico, i contenuti DeepFake presenti in rete aumenteranno a dismisura. La tecnologia ha sicuramente degli utilizzi interessanti come visto nel capitolo "Aspetto Sociale e Legale" tuttavia, riguardo la realizzazione di contenuti DeepFake, l'attuale zona grigia presente nella legge Italiana (ed anche di altri stati esteri) rischia di essere un problema. Considerando il ritmo con cui i contenuti vengono pubblicati in rete il rischio è che la realizzazione di certi contenuti venga regolamentata e punita (se necessario) quando sarà troppo tardi.

8. Sviluppi Futuri

- Utilizzo di software alternativi a DeepFaceLab;
- Testing con schede video più vecchie per scoprire quale sia l'effettivo minimo per utilizzare DeepFaceLab (con relativa configurazione hardware annessa);
- Testing con GPU AMD;
- Testare il programma con configurazioni Multi-GPU (SLI per Nvidia e Crossfire per AMD);
- Testing con CPU Intel basate su architetture Ice Lake per osservare quanto l'Intel Deep Learning Boost aumenti le performance con la nostra metodologia di testing;
- Testing con CPU AMD Ryzen;
- Osservare come la tecnologia possa essere utilizzata per anche per la parte audio;
- Lip Sync Deepfake: ossia la modifica delle sole labbra del soggetto per adattare un audio da inserire. Molto utile nel doppiaggio.

Bibliografia

[All] Saverio Alloggio. *FakeApp, intelligenza artificiale al servizio del porno fake.* URL: https://www.tomshw.it/altro/fakeapp-intelligenza-artificiale-al-servizio-del-porno-fake/.

[ASU] ASUS. *Z10PE-D8 WS.* URL: https://www.asus.com/it/Commercial-Servers-Workstations/Z10PED8_WS/overview/.

[Bon] Josefina Bonnefont. *Así se ve el "resucitado" Tarkin y la joven Leia en "Rogue One: una historia de Star Wars".* URL: http://www.upsocl.com/muvi/asi-se-ve-el-resucitado-tarkin-y-la-joven-leia-en-rogue-one-una-historia-de-star-wars/.

[Buz] Antonello Buzzi. *NVIDIA RTX A6000 è un vero mostro, ecco alcuni benchmark.* URL: https://www.nvidia.com/it-it/design-visualization/rtx-a6000/.

[Chr] Christie's. *Is artificial intelligence set to become art's next medium?* URL: https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx.

[Det] Tim Dettmers. *Deep Learning in a Nutshell: Core Concepts.* URL: https://developer.nvidia.com/blog/deep-learning-nutshell-core-concepts.

[FACa] CTRL SHIFT FACE. *Home Stallone [DeepFake].* URL: https://www.youtube.com/watch?v=2svOtXaD3gg.

[FACb] CTRL SHIFT FACE. *Taxi Driver starring Al Pacino [DeepFake].* URL: https://www.youtube.com/watch?v=9NkKj0aNB0s.

[GUR] VFX GURU. *Furious Seven - Paul Walker CGI - VFX Breakdown By "Weta Digital".* URL: https://www.youtube.com/watch?v=yQC6AJ6x8SM.

[Lin] Lucas Roos Lindgreen. *Blurring the Line Between Real and Fake: the Dangers of DeepFakes.* URL: https://medium.com/@lucasrooslindgreen/blurring-the-line-between-real-and-fake-the-dangers-of-deepfakes-14894effe295.

[O'N] Patrick Howell O'Neill. *Researchers Demonstrate Malware That Can Trick Doctors Into Misdiagnosing Cancer.* URL: https://gizmodo.com/researchers-demonstrate-malware-that-can-trick-doctors-1833786672.

[Pel] Augusto Pellucchi. *Honda NSX, la supercar nel segno di Ayrton.* URL: https://it.motor1.com/news/437998/honda-nsx-1990-storia/.

[Pia] Davide Piacenza. *Storia di come Striscia la notizia ha portato i deepfake in Italia*. URL: https://www.wired.it/play/televisione/2021/03/22/striscia-la-notizia-deepfake-intervista-creatore/.

[Rob] Mark Robins. *The Difference Between Artificial Intelligence, Machine Learning and Deep Learning*. URL: https://www.intel.it/content/www/it/it/artificial-intelligence/posts/difference-between-ai-machine-learning-deep-learning.html.

[Sor] Domenico Soriano. *Come funziona una rete neurale CNN (convolutional Neural Network)*. URL: https://www.domsoria.com/2019/10/come-funziona-una-rete-neurale-cnn-convolutional-neural-network/.

[TTN] Tien Dung Nguyen Thanh Thi Nguyen Cuong M. Nguyen. *Deep Learning for Deepfakes Creation and Detection: A Survey*. URL: https://www.researchgate.net/publication/336058980_Deep_Learning_for_Deepfakes_Creation_and_Detection_A_Survey.

[Why] The Why. *50 Facts You Didn't Know About the Fast and Furious*. URL: https://www.youtube.com/watch?v=nE7IpxRlOCU.

[Wir] Wired. *How 'Rogue One' Recreated Grand Moff Tarkin — Design FX — WIRED*. URL: https://www.youtube.com/watch?v=OUIHzanm5Mk.

Ringraziamenti

Per prima cosa volevo ringraziare il Prof. Marcantoni Fausto, relatore di questa tesi, per avermi guidato alla stesura di questo lavoro e per avermi fatto avvicinare, ed appassionare, a molti aspetti della materia che lui insegna.

Vorrei ringraziare, inoltre, tutte le persone che mi hanno supportato in questo percorso, familiari e non. In particolare ringrazio mio padre che, dopo aver perso la moglie con cui ha condiviso oltre 30 anni della sua vita, ha trovato la forza di spronarmi per farmi andare avanti, mi ha sostenuto e mi ha dato una mano a rialzarmi quando avevo toccato il fondo. Questa laurea la dedico a te mamma. Ovunque tu sia spero di averti reso orgogliosa di me, sia per la laurea sia per sia per questi anni senza di te dove sono dovuto crescere più in fretta del dovuto.