Fenomenologia dei *Deepfake*: aspetti teorici e operativi per la detection di volti umani "artificiali"

di LUCA GUARNERA[1,2], OLIVER GIUDICE[1], SEBASTIANO BATTIATO[1,2]

1. Multimedia Forensics e Deepfake

La Multimedia Forensics (MF) assume un ruolo sempre più rilevante anche grazie alla diffusione dei social network le cui stime attuali sul numero e quantità di dati multimediali (immagini, video e audio) condivisi sulle varie piattaforme hanno raggiunto livelli molto elevati. In generale, i dati multimediali contengono un'impronta digitale univoca, utile per ricostruire la storia dei dati stessi (Forensics Ballistics)[3]; per definire i modelli e le caratteristiche degli strumenti di acquisizione (scanner, fotocamera, ecc.); il social network sui quali sono stati caricati[4]; il software di editing multimediale (es. PhotoShop, GIMP, ecc.) utilizzato per la manipolazione (spesso per scopi dannosi) e molto altro.

L'emergere del fenomeno dei Deepfake ha aperto un nuovo fronte nel settore specifico della creazione di contenuti manipolati. Immagini, video e audio creati con questa nuova potente tecnologia, chiamati anche *media sintetici*, sono generati utilizzando algoritmi di intelligenza

[1] IPLAB – Università di Catania, Catania, Italia.
[2] iCTLab s.r.l. – Spinoff dell'Università di Catania, Catania, Italia.

[3] S. BATTIATO, O. GIUDICE, and A. PARATORE, *Multimedia Forensics: discovering the history of multimedia contents*. Proceedings of the 17th International Conference on Computer Systems and Technologies, pp. 5–16, 2016.

[4] O. GIUDICE, et al. *A classification engine for image ballistics of social data*. International Conference on Image Analysis and Processing. Springer, Cham, pp. 625–636, 2017.

artificiale basati su framework chiamati Autoencoder[5] e Generative Adversarial Networks (GAN)[6].

Gli algoritmi per la generazione di Deepfake sono in grado di risolvere compiti come: *image-to-image translation* (ad esempio trasformare un cavallo in una zebra, o trasformare una scena con sfumature invernali in sfumature estive), *sketch-to-image translation* (ad esempio, uno schizzo di una borsa in bianco e nero può essere trasformato in una borsa colorata e reale) e molto altro ancora. Queste tecniche si sono fortemente evolute negli ultimi anni e attualmente esistono algoritmi in grado di cambiare gli attributi dei volti delle persone come il colore dei capelli, il sesso o persino generare in maniera pseudocasuale il volto di una persona. La Figura 1 mostra una panoramica dei principali approcci Deepfake proposti in letteratura.

Molte industrie cinematografiche utilizzano già questi algoritmi per creare doppiaggi video di attori di altissima qualità. Sfortunatamente però, molti sono gli utilizzi per così dire malevoli che generano dei fenomeni deprecabili. Il primo utilizzo inappropriato di questa potente tecnologia si è palesato al grande pubblico nel 2017, dove un utente di Reddit ha sviluppato un algoritmo di Deep Learning in grado di trasferire volti di celebrità in video porno.

I tradizionali algoritmi di Multimedia Forensics presenti in letteratura riguardanti il rilevamento di *forgery* non sono in grado di distinguere i Deepfake dai contenuti multimediali reali con risultati di classificazione elevati. Per questo motivo è emersa la necessità di creare nuovi algoritmi di rilevamento dei Deepfake che fossero in grado di contrastare l'uso malevolo di queste nuove tecnologie.

Come già accennato, tutti i contenuti multimediali contengono un'impronta digitale univoca che può essere estratta per definire, ad esempio, il modello di telecamera utilizzato durante l'acquisizione. È stato possibile dimostrare che anche i Deepfake contengono una traccia unica lasciata dai modelli generativi che può essere utilizzata non solo per risolvere task del tipo *Real Vs. Deepfake* ma anche per definire quale algoritmo di Deep Learning è stato utilizzato per generare i corrispondenti dati sintetici.

[5] D.P. KINGMA, M. WELLING. *Auto-Encoding Variational Bayes*. arXiv preprint arXiv:1312.6114, 2013.

[6] I. GOODFELLOW et al., *Generative Adversarial Nets*. Advances in Neural Information Processing Systems, pp. 2672–2680, 2014 (27).

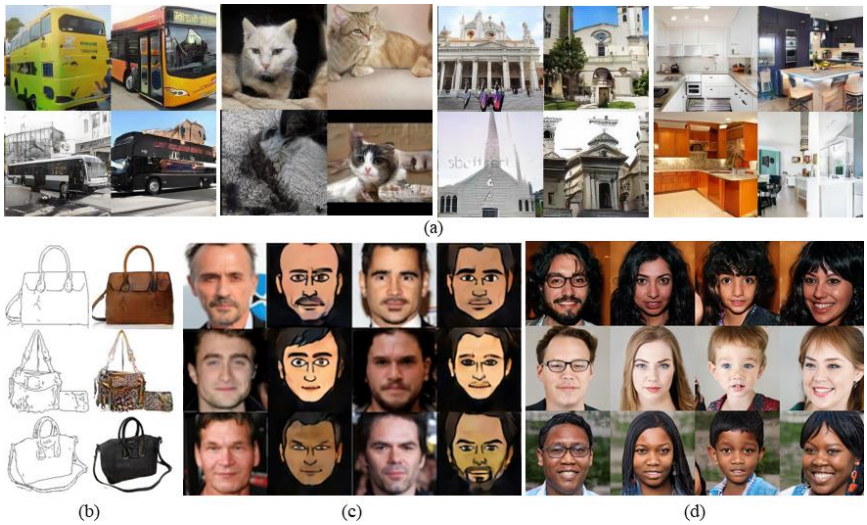

Figura 1. Esempi di immagini Deepfake generate da (a) architettura ProGAN[7]; (b) traduzione "sketch-to-image"[8]; (c) traduzione "Photos-to-Emojis"[9]; (b) architettura StyleGAN[10] che genera il cosiddetto "this person does not exist"[11].

2. Fenomenologia dei Deepfake

Un contenuto multimediale digitale può essere manipolato con molti strumenti e software. Chiunque abbia un pò di esperienza nell'uso di Photoshop o GIMP può "attaccare" le immagini per modificarne il contenuto, la semantica e, stravolgerne il messaggio in esso contenuto. Tuttavia, questo tipo di contraffazione è stato ampiamente studiato negli

[7] T. KARRAS et al., *Progressive Growing of GANs for Improved Quality, Stability, and Variation.* arXiv preprint arXiv:1710.10196, 2017.

[8] P. ISOLA et al., *Image-to-Image Translation with Conditional Adversarial Networks*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125-1134, 2017.

[9] Y. TAIGMAN, A. Polyak, and L. WOLF, *Unsupervised Cross-Domain Image Generation.* arXiv preprint arXiv:1611.02200, 2016.

[10] T. KARRAS, S. LAINE, AND T. AILA, *A Style-Based Generator Architecture for Generative Adversarial Networks.* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401-4410, 2019.

[11] https://thispersondoesnotexist.com/.

ultimi anni e sono disponibili anche strumenti commerciali con la capacità di rilevarli e descriverli[12,13].

I progressi nel campo dell'intelligenza artificiale e nell'elaborazione delle immagini stanno cambiando il modo in cui le persone interagiscono con immagini e video digitali. Nello specifico, l'avvento degli Autoencoder e delle Generative Adversarial Networks (GAN), ha permesso la creazione e la diffusione di tecniche estremamente raffinate in grado di *attaccare* i dati digitali, *alterarli* o *crearne i contenuti da zero*. Questi strumenti sono in grado di ottenere risultati sorprendentemente realistici portando alla nascita del fenomeno *Deepfake*.

Per Deepfake si intendono tutti quei contenuti multimediali sinteticamente alterati o creati sfruttando i *modelli generativi* nel settore del Deep Learning. In generale, un modello generativo è un algoritmo che tecnicamente si occupa di "imparare" la distribuzione dei dati di addestramento al fine di generare nuovi dati con caratteristiche simili rispetto al set di dati originale. In altre parole, nel caso in cui volessimo creare oggetti della categoria "automobili", allora dovremmo creare un algoritmo che sia in grado di "inferire" la distribuzione di dati che caratterizza varie tipologie di automobili, ovvero catturare gli elementi essenziali (come i colori, le forme, ecc.) che definiscono l'oggetto in questione al fine di riprodurlo nel miglior modo possibile. Non è possibile ottenere esattamente la stessa distribuzione dei dati di input, ma in generale si riesce a definire un modello che si avvicina il più possibile alla vera distribuzione dei dati di training. Per raggiungere questo obiettivo, negli ultimi anni i ricercatori hanno utilizzato la potenza delle *Deep Neural Networks* (*Reti Neurali Profonde)* per apprendere una funzione che può approssimare la distribuzione del modello di dati di addestramento.

Questi metodi generativi si sono evoluti molto rapidamente dando luogo a dati multimediali sintetici estremamente sorprendenti, soprattutto per quanto riguarda l'elevata qualità dei dettagli dei contenuti Deepfake creati.

Diversi utilizzi di tale tecnologia sono stati già sviluppati ed applicati in vari contesti come, ad esempio, il *Salvador Dalì Museum* in Florida

[12] L. VERDOLIVA, *Media Forensics and Deepfakes: an overview*. IEEE Journal of Selected Topics in Signal Processing 14.5 (2020): 910-932.

[13] A. PIVA, *An Overview on Image Forensics*. International Scholarly Research Notices, pp. 1-22, 2013.

(Figura 2(a)) che ha dato la possibilità ai visitatori del museo di interagire con l'artista[14]; l'applicazione che presenta il calciatore David Beckham[15] che parla nove lingue diverse (Figura 2(b)) per una campagna contro la *malaria*; approcci in grado di descrivere un dipinto o un paesaggio ai non vedenti rendendoli capaci di connettersi con l'ambiente circostante.

(a) (b)

Figura 2(a). Esempi di immagini di Salvador Dalì (b) Esempi di immagini di David Beckham in cui parla 3 di 9 lingue diverse.

Recentemente, volti di veline, politici, attori, presentatori televisivi e molti altri sono stati oggetto principale di attacchi mediante Deepfake: uno dei primi esempi è il famoso scambio di volti di Jim Carrey sul corpo di Alison Brie[16] (Figura 3(a)), o Mr. Bean è Charlize Theron nella versione Deepfake dello spot di J'adore[17] (Figura 3(b)), e ancora Jim Carrey invece di Jack Nicholson in Shining[18] (Figura 3(c)), o Tom Cruise che sostituisce Robert Downey Jr. in Iron Man[19] (Figura 3(d)).

[14] https://www.youtube.com/watch?v=mPtcU9VmIIE.
[15] https://www.youtube.com/watch?v=QiiSAvKJIHo.
[16] https://www.youtube.com/watch?v=SEar_6UtX9U.
[17] https://www.youtube.com/watch?v=tDATo EnJEY8.
[18] https://www.youtube.com/watch?v=JbzVhzNaTdI.
[19] https://www.youtube.com/watch?v=iDM69UEyM3w.

La tecnologia Deepfake è stata utilizzata anche nel mondo della politica, ad esempio quando il volto del presidente argentino Mauricio Macri è stato sostituito da quello di A. Hitler o il volto del cancelliere Angela Merkel sostituito dal volto di Donald Trump. Quest'ultimo nel 2019 è stato oggetto di un Deepfake in cui veniva deriso per il suo aspetto e il colore della pelle. Nel 2018, invece, è stato creato un Deepfake dal regista Jordan Peele utilizzando l'immagine di Barack Obama in cui pronunciava parole mai dette, ma il messaggio era semplice: mettere in guardia sul pericolo dell'uso illecito di questa potente tecnologia.

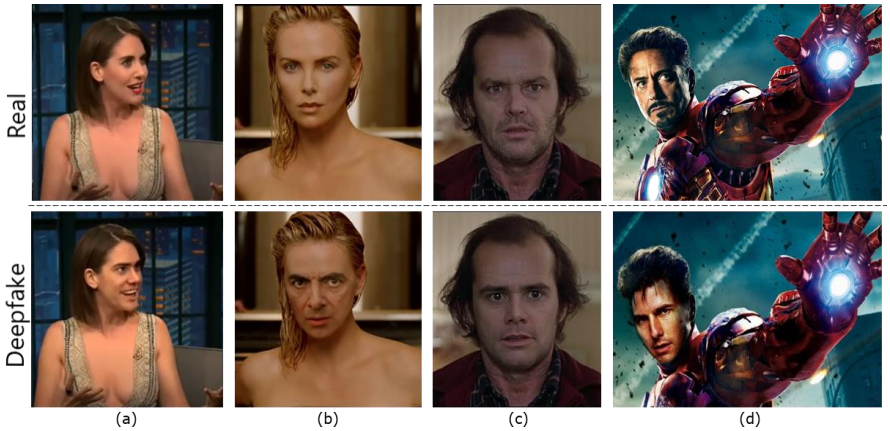

Figura 3. Esempi di Deepfake: (a) il volto di Jim Carrey è stato trasferito sul corpo di Alison Brie, (b) Mr. Bean è Charlize Theron in una versione Deepfake dello spot di J'adore, (c) Jim Carrey invece di Jack Nicholson in Shining e (d) Tom Cruise sostituisce Robert Downey Jr. in Iron Man.

Le manipolazioni sui volti delle persone riguardano quattro modalità principali:

– *intera sintesi del volto* (*entire face synthesis):* crea intere immagini di volti inesistenti[10,20,21]. Questa manipolazione potrebbe essere utilizzata ad esempio nel campo dei videogiochi per creare

[20] T. KARRAS et al. *Analyzing and Improving the Image Quality of StyleGAN*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110-8119, 2020.

[21] T. KARRAS et al. *Training Generative Adversarial Networks with Limited Data*. arXiv preprint arXiv:2006.06676, 2020.

nuovi protagonisti, ma potrebbe anche essere utilizzata per generare disinformazione (es. creazione di profili falsi nei social network);

- *manipolazione degli attributi (attribute manipulation):* modifica gli attributi del viso come il colore dei capelli, il sesso, l'età, ecc. Questo tipo di manipolazione è noto anche come *modifica del viso* o *ritocco del viso*. Attualmente, diverse tecnologie hanno realizzato questo tipo di manipolazione, come l'applicazione mobile chiamata FaceApp[31] o i potenti motori GAN come StarGAN[22], StarGAN-v2[23];
- *scambio di identità (identity swap):* sostituisce il volto di una persona in un video con il volto di un'altra persona. Questo tipo di manipolazione può essere ottenuto considerando tecniche di Deep Learning (come Deepfakes[24]) e approcci di computer grafica (come FaceSwap[25]). Come riportato in precedenza, queste manipolazioni possono essere sfruttate per scopi industriali (come all'interno dell'industria cinematografica) ma anche per scopi dannosi come nella creazione di video pornografici;
- *scambio di espressioni (expression swap)*: modifica l'espressione facciale della persona. Questo tipo di manipolazione è anche conosciuto come *reenactment del volto*. Face2Face[26] e NeuralTexture[27] sono esempi di questa tecnologia ed è possibile usarli per diffondere informazioni false come il video di Obama[28], creato da Buzzfeed in collaborazione con Monkeypaw Studios, o il video in cui Mark Zuckerberg[29] realizza una serie di affermazioni sulla capacità della piattaforma di rubare i dati dei suoi utenti.

[22] Y. Choi et al. *StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789-8797, 2018.

[23] CHOI, Yunjey, et al. *StarGAN v2: Diverse Image Synthesis for Multiple Domains.* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8188-8197, 2020.

[24] https://github.com/deepfakes/faceswap.

[25] https://github.com/MarekKowalski/FaceSwap.

[26] J. THIES et al., *Face2face: Real-Time Face Capture and Reenactment of RGB Videos.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2387-2395, 2016.

[27] J. THIES et al, *Deferred Neural Rendering: Image Synthesis Using Neural Textures.* ACM Transactions on Graphics (TOG) 38.4 (2019): 1-12.

[28] https://www.youtube.com/watch?v=cQ54GDm1eL0.

[29] https://www.youtube.com/watch?v=NbedWhzx1rs.

La Figura 4 mostra esempi di queste tipologie di manipolazioni di volti di persone.

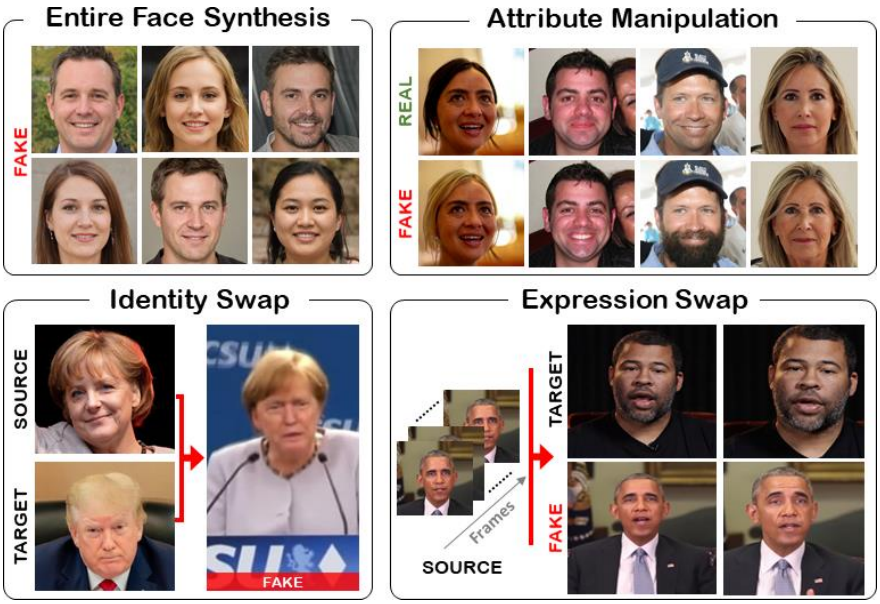

Figura 4. Manipolazioni sui volti delle persone. *Intera sintesi del volto*: immagine generata dalle architetture StyleGAN[10] e StyleGAN2[21]. *Manipolazione degli attributi*: esempi di immagini generate dal motore StarGAN[23]. *Scambio di identità*: il volto della cancelliera Angela Merkel sostituito dal volto di Donald Trump. *Scambio di espressioni*: esempi di fotogrammi video del Deepfake di Barack Obama.

Le architetture Deepfake si sono evolute considerevolmente negli ultimi anni, ottenendo contenuti multimediali sintetici con una risoluzione sempre più elevata, una qualità dei dettagli sempre più accurata e una quasi completa assenza di anomalie (Figura 5).
La manipolazione dei volti umani è nata ancor prima dell'avvento dei Deepfake. Ciò che oggi chiamiamo "reenactment del volto" si può in qualche modo risalire a una prima applicazione di un progetto accademico del 1997, con uno strumento chiamato *Video Rewrite* proposto da Bregler et al.[30] che consentiva la creazione automatica di un nuovo video di una persona in grado di mimare una sequenza di parole (data come traccia audio di riferimento) mai dette dal protagonista del video

[30] C. Bregler, M. Covell, and M. Slaney. *Video Rewrite: Driving Visual Speech with Audio*. Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, pp. 353-360, 1997.

originale. Queste tecnologie si sono sorprendentemente evolute negli anni, raggiungendo risultati sempre più sofisticati, come il lavoro proposto da Thies et al.[27] con il framework chiamato *Face2Face* in cui l'obiettivo di animare in tempo reale le espressioni facciali di un video di destinazione dalle espressioni facciali di una persona catturate da una fotocamera (webcam).

Figura 5. L'evoluzione dei modelli generativi (ad esempio tramite architetture GAN) per la creazione di contenuti multimediali sintetici (da Goodfellow et al.[6] (2014) a Karras et al.[10] (2018)).

La tecnologia Deepfake non è solo coinvolta in attività relative alla manipolazione di volti di persone, ma potrebbe essere utilizzata per scambiare o generare luoghi, animali, oggetti, ecc. in maniera sorprendentemente realistica. In effetti, questo potrebbe portare un'innovazione dirompente in molti ambiti lavorativi, come nell'industria automobilistica o in architettura, poiché è possibile generare un'auto o un appartamento attraverso *GAN* dedicati o nell'industria cinematografica dove è possibile, quando necessario, sostituire il volto di uno *stuntman* con un attore. Inoltre, la creazione di Deepfake sta diventando estremamente facile: applicazioni come FaceApp[31], sono in grado di produrre trasformazioni sui volti umani come il cambio di genere, l'invecchiamento, e tanto altro ancora. I risultati sono sorprendentemente realistici ed estremamente facili da produrre anche per utenti non esperti con pochi tocchi sul proprio smartphone. Sfortunatamente diversi video di Deepfake collegati alla pornografia sono stati utilizzati per screditare attrici famose come Emma Watson o Angelina Jolie, o in altri contesti sono stati utilizzati per diffondere disinformazione e fake news.

[31] https://www.faceapp.com/.

Risulta quindi assolutamente necessario creare nuove tecniche in grado di rilevare e contrastare questo nuovo potente fenomeno[12,32].

Sebbene il rilevamento di un Deepfake sia difficile per gli umani, lavori recenti hanno dimostrato che potrebbero essere rilevati in modo sorprendentemente semplice utilizzando le *Reti Neurali Convoluzionali* (CNN)[33,34,35,36,37] specificamente addestrate al compito. Tuttavia, le soluzioni basate su CNN presentate fino ad oggi, mancano di robustezza, capacità di generalizzazione e spiegabilità poiché risultano essere estremamente specifici per il contesto in cui sono stati formati e, inoltre, essendo molto profonde, le architetture neurali tendono ad estrarre la semantica sottostante dalle immagini senza dedurre alcuna impronta digitale univoca.

Diverse grandi aziende, da Facebook a Microsoft, hanno deciso di agire contro questo fenomeno. Rossler et al. hanno creato un database di video Deepfake[38], mentre Facebook e Microsoft hanno lanciato l'iniziativa *Deepfake Detection Challenge*[39] invitando i ricercatori di tutto il mondo a creare nuovi strumenti per rilevare Deepfake e media multimediali manipolati.

Al fine di ottenere soluzioni in grado di definire se un contenuto multimediale è un Deepfake, occorre comprendere come questi dati vengono creati in modo da identificare quelle tracce univoche lasciate dai vari processi generativi. Nella sezione successiva verranno descritte le principali tecnologie utilizzate per la creazione dei Deepfake.

[32] L. GUARNERA et al., *Preliminary Forensics Analysis of Deepfake Image*s, 2020 AEIT International Annual Conference (AEIT). IEEE, pp. 1-6, 2020.

[33] X. XUAN et al., *On the Generalization of GAN Image Forensics*. Chinese Conference on Biometric Recognition. Springer, Cham, pp. 134-141, 2019.

[34] H. CHIH-CHUNG, Y. ZHUANG, and C. LEE, *Deep Fake Image Detection Based on Pairwise Learning*. Applied Sciences 10.1, 2020: 370.

[35] A. GANDHI AND J. SHOMIK, *Adversarial Perturbations Fool Deepfake Detectors.* 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1-8, 2020.

[36] L. LI et al., *Face X-ray for More General Face Forgery Detection.* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5001-5010, 2020.

[37] SY. WANG et al. *CNN-generated images are surprisingly easy to spot... for now*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8695-8704, 2020.

[38] A. ROSSLER et al. *Faceforensics++: Learning to Detect Manipulated Facial Images*. Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1-11, 2019.

[39] https://deepfakedetectionchallenge.ai.

2.1. *Metodi di creazione di Deepfake*

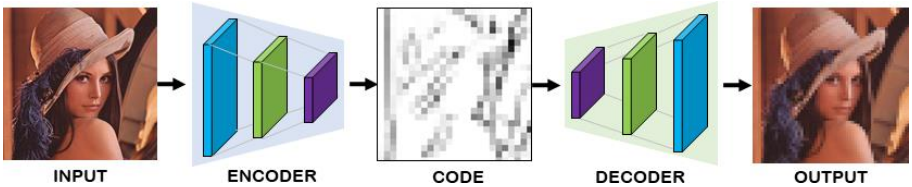

Figura 6. Flusso di lavoro del motore AE. Data un'immagine come input al codificatore (encoder), genererà una rappresentazione latente (code) che verrà utilizzata dal decodificatore (decoder) per generare la stessa immagine come output.

Una delle prime tecniche di Deep Learning utilizzate per creare Deepfake sono stati gli Autoencoder (AE)[40.]. In generale, un'architettura Autoencoder è composta da tre componenti: *codificatore (encoder), rappresentazione latente (code)* e *decodificatore (decoder)*.

La Figura 6 mostra uno schema generico di un'architettura AE. Il codificatore prende un'immagine come input e genera un *code* o la cosiddetta *rappresentazione latente* che è una rappresentazione intermedia di dimensione inferiore dei dati di input. L'obiettivo del decoder è ricostruire un dato multimediale con le stesse caratteristiche (dimensioni, ecc.) dell'input a partire dalla *rappresentazione latente*. Notare che il decoder non conosce i dati di input all'encoder. Nello specifico, il decoder sarà in grado di ricostruire un'immagine tanto più simile a quella in input all'encoder quando l'encoder genererà la migliore rappresentazione latente dell'input. Per fare ciò, l'architettura sarà addestrata in modo da minimizzare l'errore tra la differenza della struttura dei dati di input rispetto ai dati ricostruiti dal decoder. Questo processo termina quando la perdita di ricostruzione è ridotta al minimo, risultando in un output molto simile all'input.

[40] Una rete neurale di apprendimento auto-supervisionata progettata per ricostruire il suo input da una rappresentazione dimensionale inferiore.

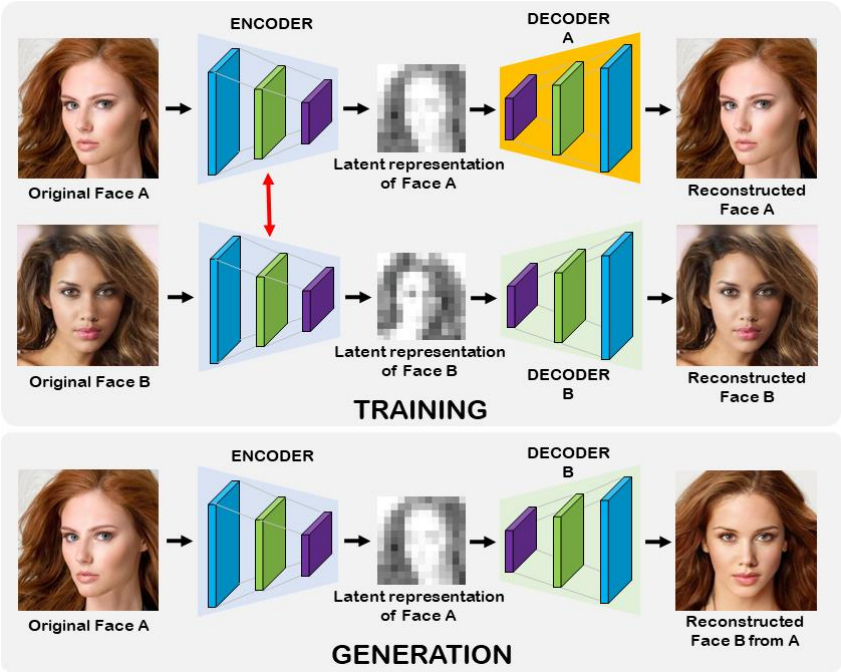

Figura 7. Scambio di volti utilizzando due coppie di codificatori e decodificatori. Questo approccio viene utilizzato per apprendere le caratteristiche latenti dei volti nella fase di allenamento. Successivamente, la coppia di decodificatori viene scambiata per generare la Faccia A con le caratteristiche della Faccia B.

Un'architettura basata su AE per la creazione di Deepfake è costituita dai cosiddetti *Deep Autoencoder*[41]. Come descritto in precedenza, l'obiettivo del decodificatore è provare a ricostruire l'immagine di input al codificatore. Consideriamo il caso in cui il compito di un framework Deepfake è lo scambio di volti tra due persone. Questo può essere fatto considerando due coppie encoder-decoder. Ogni coppia verrà addestrata su una serie di immagini, come due architetture standard. La caratteristica principale di questo framework è che i parametri dei due encoder sono condivisi al fine di trovare e apprendere la somiglianza tra due insiemi di immagini del viso (posizioni degli occhi, posizioni della bocca, ecc.). In altre parole, i due encoder vengono considerati come "gemelli" poiché condividono gli stessi parametri di rete neurale. In particolare, una volta completato l'addestramento delle due

[41] P. VINCENT et al., *Extracting and Composing Robust Features with Denoising Autoencoders*. Proceedings of the 25th International Conference on Machine learning, pp. 1096-1103, 2008.

coppie di codificatore e decodificatore, i due decodificatori vengono scambiati al fine di utilizzare il codificatore dell'immagine sorgente e il decodificatore dell'immagine bersaglio per generare l'immagine bersaglio con le caratteristiche dell'immagine sorgente. FakeApp[42], ZAO[43], FaceSwap[44] e REFACE[45] sono alcune delle prime applicazioni sviluppate per la creazione di Deepfake considerando questo approccio. Un esempio di manipolazione del face-swap che sfrutta questo principio è mostrato nella Figura 7 in cui l'encoder allenato con *Face A* è connesso al decoder di *Face B* per ricostruire la *Face B* dalla *Face A* originale.

Risultati sorprendentemente realistici sono ottenuti attraverso framework chiamati *Generative Adversarial Networks* (GAN) introdotte per la prima volta da Goodfellow et al.[6]. Gli autori hanno proposto un nuovo approccio in cui due modelli si addestrano contemporaneamente: un modello generativo *G*, che impara la distribuzione dei dati di input, e un modello discriminativo *D*, che stima la probabilità che un campione arrivi dai dati di addestramento anziché da *G*. La procedura di allenamento per *G* consiste nel massimizzare la probabilità che *D* commetta un errore, risultando così in una partita a due giocatori di tipo "min-max".
Matematicamente, il generatore accetta un input casuale *z* (il quale potrebbe essere un numero, una sequenza di numeri ecc.) con densità di probabilità p_z e restituisce un output $x = G(z, \Theta_g)$ secondo una certa distribuzione di probabilità p_g (Θ_g rappresenta i parametri del modello generativo). Il discriminatore, $D(x, \Theta_d)$ calcola la probabilità che *x* provenga dalla distribuzione dei dati di training p_{data} (Θ_d rappresenta i parametri del modello discriminativo). L'obiettivo è ottenere un generatore che sia un buon stimatore di p_{data}. Quando ciò accade, il discriminatore verrà "ingannato" e non sarà più in grado di distinguere immagini provenienti da p_{data} e p_g; quindi p_g seguirà la distribuzione di probabilità mirata, ovvero p_{data}.

In sintesi, il discriminatore viene addestrato in modo da massimizzare la probabilità di assegnare l'etichetta corretta sia agli esempi di addestramento che ai campioni di *G*. Simultaneamente, il generatore viene addestrato in modo da minimizzare la distanza tra le due distribuzioni.

[42] https://faceswap.dev/.
[43] https://apps.apple.com/cn/app/zao/id1465199127.
[44] https://www.malavida.com/en/soft/fakeapp/.
[45] https://reface.app/.

La Figura 8 mostra una struttura semplificata di un framework GAN. *G* può essere pensato come una squadra di falsari che cercano di produrre valuta falsa, mentre *D* può essere associato alla polizia, cercando di rilevare l'attività dannosa.

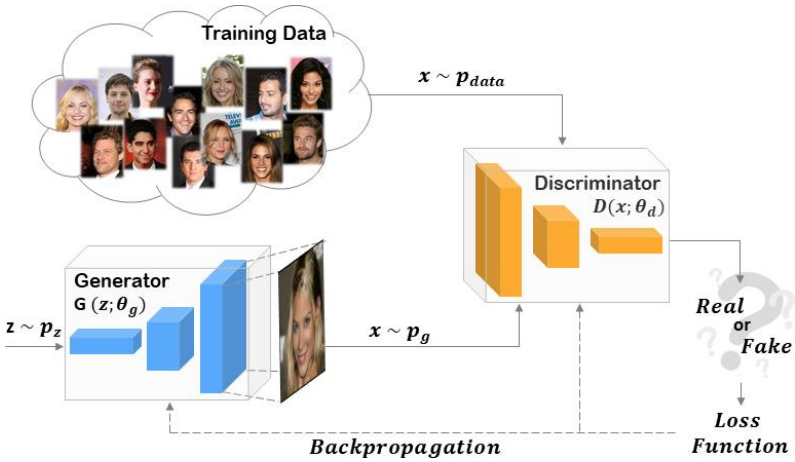

Figura 8. Schema generico di un framework GAN.

I progressi nelle architetture GAN portano a diversi lavori basati sulla manipolazione di volti umani. Come descritto precedentemente, le principali tecniche per manipolare i volti delle persone sono lo scambio di volti, la rievocazione del volto, la manipolazione degli attributi e la sintesi dell'intera faccia.

In generale, i metodi tradizionali basati sullo scambio di volti[46],[47] consistono nel:

- rilevamento del volto nelle immagini di origine e selezione di un'immagine del volto candidato nel set di dati di addestramento con caratteristiche simili all'input;
- sostituzione dei componenti del viso (come occhi, naso e bocca), regolazione della struttura dell'immagine (illuminazione, colori, ecc.) e fusione dei due volti;

[46] D. Bitouk et al., *Face Swapping: Automatically Replacing Faces in Photographs*. ACM SIGGRAPH 2008 papers. 2008, pp.1-8.

[47] B.M. Smith and Z. Li, *Joint Face Alignment with Non-parametric Shape Models*. European Conference on Computer Vision. Springer, Berlin, Heidelberg, pp. 43-56, 2012.

– classificare l'output calcolando una distanza sulla regione di sovrapposizione.

Con l'evoluzione dei metodi basati sul Deep Learning i principali limiti di queste tecniche (quali la perdita di espressione dell'immagine del volto in ingresso; la presenza di artefatti, posizioni innaturali del volto finale, ecc.) sono stati quasi completamente risolti ottenendo risultati più sofisticati e realistici.

In passato queste manipolazioni venivano eseguite con tecniche di Computer Grafica, ricostruendo il volto a partire da un modello 3D[48]. Queste tecniche si sono evolute con il sorprendente sviluppo dei modelli basati su GAN (ad esempio Nirkin et al.[49]).

Approcci recenti come quello proposto da Korshunova et al.[50] sono in grado di trasferire il contenuto semantico (ad esempio, postura facciale, espressione facciale e condizioni di illuminazione) dell'immagine di input nell'immagine di output.

Natsume et al.[51] hanno proposto l'implementazione di *RSGAN* tramite due Autoencoder per codificare rispettivamente la rappresentazione latente delle regioni del viso e dei capelli. Un limite di questo metodo è che non gestisce bene le occlusioni facciali (occhiali, viso parzialmente coperto da oggetti, ecc.) ottenendo immagini con artefatti. Questo problema è stato affrontato da Li et al.[52] con il framework chiamato *FaceShifter* ottenendo ottimi risultati nel task di face-swap.

Un'altra categoria di manipolazione di volti di persone riguarda la classe di algoritmi inerenti al *reenactment*, i quali consentono di manipolare le espressioni facciali di una persona (trasferimento di gesti facciali, movimenti oculari, ecc.). In generale, questa tecnica viene applicata ai video multimediali. Il lavoro di Thies et al.[53] è stato tra i primi a

[48] Y. NIRKIN et al., *On Face Segmentation, Face Swapping, and Face Perception.* 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, pp. 98-105, 2018.

[49] Y. NIRKIN et al., *FSGAN: Subject Agnostic Face Swapping and Reenactment.* Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7184-7193, 2019.

[50] I. KORSHUNOVA et al., *Fast Face-swap Using Convolutional Neural Networks.* Proceedings of the IEEE International Conference on Computer Vision, pp. 3677-3685, 2017.

[51] R. NATSUME, et al., *RSGAN: Face Swapping and Editing using Face and Hair Representation in Latent Spaces.* arXiv preprint arXiv:1804.03447, 2018.

[52] L. LI et al., *Faceshifter: Towards High Fidelity and Occlusion Aware Face Swapping*. arXiv preprint arXiv:1912.13457, 2019.

[53] J. THIES et al., *Real-time Expression Transfer for Facial Reenactment.* ACM Trans. Graph. 34.6, 2015, 183-1.

proporre tecniche di ricostruzione del volto sfruttando un sensore RGB-D per tracciare e ricostruire il modello 3D di un attore sorgente e target. Uno dei metodi più potenti basati su approcci matematici per il Face-Reenactment è il lavoro proposto da Thies et al.[27] con l'architettura chiamata *Face2Face*. Questo metodo è in grado di alterare i movimenti facciali dei flussi video RGB in tempo reale, utilizzando una webcam standard per catturare i movimenti da alterare nella sorgente.

Le tecniche di ricostruzione del volto si sono evolute molto negli ultimi anni portando non solo a modificare l'espressione facciale ma anche a rianimare i video di ritratto cambiando ad esempio la posa della testa, lo sguardo degli occhi e molto altro. Nel lavoro di Kim et al.[54], viene inizialmente calcolata una rappresentazione parametrica del volto per produrre un rendering sintetico dell'identità del bersaglio. Una rete di traduzione render-to-video basata su GAN condizionali (cGAN) elabora questa rappresentazione per generare un rendering sintetico in fotogrammi video fotorealistici. In generale, queste tecniche richiedono una grande quantità di dati di addestramento. Questa limitazione è stata recentemente superata proponendo approcci di rievocazione facciale "one-shot" utilizzando una o un numero limitato di immagini di origine[55].

Infine, le ultime due categorie che possiamo considerare per la manipolazione del viso riguardano la sintesi dell'intero volto e la manipolazione degli attributi di volti umani.

In generale, la manipolazione degli attributi del viso utilizza tecniche di *image-to-image translation* per affrontare la diversità di stile[56,57,58]. Il principale limite di questi metodi è la mappatura tra due domini che non li rende scalabili al numero crescente di domini. In altre parole, è possibile addestrare un modello per cambiare, ad esempio, solo il colore dei capelli, cambiare solo l'espressione facciale. Matematicamente, se considerassimo N domini, questi metodi richiederebbero

[54] K. HYEONGWOO et al., *Deep Video Portraits.* ACM Transactions on Graphics (TOG) 37.4. 2018, 1-14.

[55] O. WILES et al., *X2Face: A Network for Controlling Face Generation using Images, Audio, and Pose Codes*. Proceedings of the European Conference on Computer Vision (ECCV), pp. 670-686, 2018.

[56] A. ALMAHAIRI et al., *Augmented CycleGAN: Learning Many-to-Many Mappings from Unpaired Data*. International Conference on Machine Learning. PMLR, pp. 195-204, 2018.

[57] X. HUANG et al., *Multimodal Unsupervised Image-to-Image Translation*. Proceedings of the European conference on computer vision (ECCV), pp. 172-189, 2018.

[58] Q. MAO et al., *Mode Seeking Generative Adversarial Networks for Diverse Image Synthesis.* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1429-1437, 2019.

un addestramento di *N* x *(N-1)* generatori per gestire le *traduzioni* di ogni dominio. Questi approcci sembrano alquanto limitati sotto questo aspetto. Per cercare di creare un modello unico per risolvere il compito di *image-to-image translation* su più domini, diversi studi in letteratura sono stati proposti.

StarGAN[23] è un'architettura in grado di eseguire compiti di *image-to-image translation* su più domini utilizzando un unico modello. Data un'etichetta casuale come input, come il colore dei capelli, l'espressione facciale, ecc., il framework è in grado di eseguire un'operazione di *image-to-image translation*. I risultati hanno dimostrato come *StarGAN* riesce a generare immagini di qualità visiva superiore rispetto ai metodi esistenti in letteratura. Questa è una delle prime migliori soluzioni in questa categoria di manipolazione, ma *StarGAN* ha una limitazione importante: non cattura la natura multimodale della distribuzione dei dati. In altre parole, questa architettura apprende una mappatura deterministica di ciascun dominio, ovvero data un'immagine e un'etichetta fissa come input, il generatore produrrà lo stesso output per ciascun dominio. Ciò dipende principalmente dal fatto che ogni dominio è definito da un'etichetta predeterminata. Questa limitazione è stata risolta con la seconda versione di StarGAN chiamata *StarGAN-v2*[24].

He et al.[59], hanno proposto una nuova tecnica chiamata *AttGAN* in cui viene applicato un vincolo di classificazione degli attributi all'immagine generata, al fine di garantire solo le corrette modifiche degli attributi desiderati. I risultati ottenuti hanno mostrato come *AttGAN* supera lo stato dell'arte sulla modifica realistica degli attributi facciali.

Infine, le migliori tecniche di sintesi dell'intero volto sono StyleGAN[10] e StyleGAN2[21] con esempi disponibili nella pagina web *this person does not exist*[11]. Gli autori di StyleGAN hanno proposto un generatore diverso rispetto agli standard motori GAN mediante la mappatura di punti nello spazio latente ad uno spazio latente intermedio che controlla l'output di stile in ogni punto del processo di creazione. Pertanto, StyleGAN è in grado non solo di generare foto di volti incredibilmente realistiche e di alta qualità, ma offre anche parametri di controllo in termini di stile complessivo dell'immagine generata a diversi livelli di dettaglio. Piccole anomalie potrebbero rivelare la falsità delle immagini generate. Per correggere quelle imperfezioni, Karras et al.

[59] Z. He et al., *AttGAN: Facial Attribute Editing by Only Changing What you Want.* IEEE Transactions on Image Processing 28.11 (2019): 5464-5478.

hanno apportato alcuni miglioramenti al generatore di StyleGAN proponendo StyleGAN2[21].

2.2. *Metodi di Detection di Deepfake*

I metodi esistenti di rilevamento di Deepfake risolvono un compito di classificazione binaria, ovvero stabiliscono solo se un contenuto multimediale è *Reale* oppure *Deepfake*.

Le immagini generate dalle architetture GAN risultano essere molto complicate da classificare, poiché questi framework riescono a produrre immagini sintetiche molto realistiche e di alta qualità. Esistono diversi algoritmi di rilevamento dei Deepfake in letteratura. Tuttavia, la maggior parte dei lavori sono considerati privi di generalizzazione poiché funzionano molto bene solo con dati con cui sono stati allenati. Questa caratteristica risulta essere un limite cruciale di queste tecniche a causa dello sviluppo sempre più rapido di nuove architetture generative.

In generale, per le immagini sintetiche, le attuali tecniche di rilevamento mirano a identificare quelle anomalie introdotte nel processo generativo che sono spesso presenti sullo sfondo[60,61]. Altri metodi invece cercano di estrarre l'impronta lasciata dal modello generativo[62,63,64,65].

Per i video sintetici, invece, vengono analizzati anche i cosiddetti artefatti temporali. Ad esempio, le tecniche di rievocazione del volto potrebbero portare ad anomalie nella sincronizzazione dei fotogrammi video[66,67]non presenti comunemente in dati multimediali reali.

[60] S. McCloskey and M. Albright, *Detecting GAN-generated Imagery Using Color Cues*. arXiv preprint arXiv:1812.08247, 2018.

[61] X. Li et al., *Fighting Against Deepfake: Patch&Pair Convolutional Neural Networks* (PPCNN). Companion Proceedings of the Web Conference, pp. 88-89, 2020.

[62] N. Yu et al., *Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints.* Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7556-7566, 2019.

[63] L. Guarnera, O. Giudice, S. Battiato, *Deepfake Detection by Analyzing Convolutional Traces.* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 666-667, 2020.

[64] L. Guarnera, O. Giudice, S. Battiato, *Fighting Deepfake by Exposing the Convolutional Traces on Images*. IEEE Access 8 (2020): 165085-165098.

[65] O. Giudice, L. Guarnera, S. Battiato, *Fighting Deepfakes by Detecting GAN DCT Anomalies.* J. Imaging 2021, 7, 128. https://doi.org/10.3390/jimaging7080128.

[66] D. Güera, et al., *We Need No Pixels: Video Manipulation Detection Using Stream Descriptors*. arXiv preprint arXiv:1906.08743, 2019.

[67] I. Amerini et al., *Deepfake Video Detection through Optical Flow Based CNN*. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019.

Diversi metodi allo stato dell'arte utilizzano approcci basati sul Deep Learning per cercare di distinguere i dati sintetici dai dati reali.

Li et al.[68] hanno ottenuto buoni risultati di classificazione utilizzando diverse architetture neurali profonde addestrate tramite particolari punti di riferimento facciali estratti con DLIB[69]. Una tecnica molto interessante è stata proposta da Li et al.[70] in cui hanno utilizzato un'architettura di Deep Learning per analizzare la mancanza di battiti di ciglia nei fotogrammi video dei volti sintetizzati. Un limite di questa tecnica è che non riesce a rilevare i volti manipolati con gli occhi chiusi. Le tecniche di creazione di Deepfake si stanno evolvendo molto rapidamente creando video sintetici con battito di ciglia simili ai video reali. In quest'ultimo scenario, quest'ultimo metodo non sarà più in grado di distinguere i video reali dai Deepfake.

Approcci di Deep Learning sono stati impiegati per classificare le immagini Deepfake.

Un lavoro interessante è stato proposto da Agarwal et al.[71] in cui cercano di identificare le manipolazioni di scambio di volti tramite un framework in grado di calcolare due caratteristiche biometriche: la biometria facciale (estratta tramite l'architettura di deep learning) e la biometria comportamentale (estratta tramite una rete di codificatori-decodifica). La combinazione di queste caratteristiche ha permesso l'identificazione di manipolazioni di scambio di volti.

Ottimi risultati sono stati ottenuti da quelle tecniche in grado di estrarre informazioni riguardanti segnali biologici come nel lavoro proposto da Fernandes et al.[72] in cui è stata calcolata la frequenza cardiaca combinando la variazione del colore della pelle, l'intensità ottica media e l'ingrandimento euleriano del video. Gli autori hanno addestrato un modello di equazioni differenziali ordinarie neurali con questa funzione per il compito di classificazione Deepfake.

[68] Y. LI, S. LYU, *Exposing Deepfake Videos by Detecting Face Warping Artifacts.* arXiv preprint arXiv:1811.00656, 2018.

[69] D.E. KING, *Dlib-ml: A Machine Learning Toolkit.* The Journal of Machine Learning Research 10, 2009, pp.1755-1758.

[70] Y. LI, et al., *In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking*. 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018.

[71] S. AGARWAL et al., *Detecting Deep-Fake Videos from Appearance and Behavior*. 2020 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2020.

[72] S. FERNANDES et al., *Predicting Heart Rate Variations of Deepfake Videos Using Neural ODE.* Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019.

Alcuni modelli generativi non sono in grado di creare tutti gli elementi essenziali per ottenere immagini sintetiche che risultino autentiche all'occhio umano. Spesso, infatti, mancano i riflessi; sono presenti diversi colori degli occhi; sono presenti artefatti nei denti; orecchie di forma diversa e altro ancora. Sfruttando queste informazioni, Matern et al.[73] hanno addestrato due classificatori estraendo caratteristiche in grado di descrivere queste anomalie. Un limite di questa tecnica è che funziona solo nel caso in cui si considerino volti sintetici con occhi aperti o denti visibili.

Alcune tecniche di rilevamento di Deepfake utilizzano approcci "artigianali/tradizionali" per il task di classificazione, ovvero approcci non basati su architetture come le Reti Neurali. Ad esempio, nel lavoro proposto da Guarnera et al.[68,69], è stato dimostrato che la correlazione dei pixel nelle immagini Deepfake risulta essere differente rispetto a immagini reali acquisite da comuni dispositivi. Questa caratteristica ha consentito la creazione di un classificatore robusto in grado di distinguere immagini reali da immagini Deepfake con un alto livello di classificazione. Il metodo riesce anche a identificare le immagini manipolate con FaceApp[31] come Deepfake.

Le immagini Deepfake contengono delle statistiche frequenziali differenti rispetto a immagini reali[33]. Queste possono essere estratte e considerate come impronte digitali uniche lasciate dai singoli motori GAN. Recentemente, Giudice et al.[70] hanno proposto una nuova pipeline in grado di rilevare le cosiddette *GAN Specific Frequencies* (GSF), ovvero quelle frequenze che caratterizzando le immagini Deepfake le quali definiscono un'impronta digitale unica di diverse architetture generative. Utilizzando la Trasformata Discreta del Coseno (DCT – *Discrete Cosine Transform*), sono state rilevate queste frequenze anomale. Le statistiche desunte dalla distribuzione dei coefficienti estratti dalla DCT sono risultate essere fondamentali per riconoscere i dati generati dal motore GAN. I test di robustezza effettuati attaccando le immagini tramite operazioni di compressione, mirroring, rotazione, ridimensionamento e aggiunta di rettangoli di dimensioni e colori casuali hanno ulteriormente dimostrato l'efficacia del metodo proposto dagli autori.

[73] F. MATERN, C. RIESS, M. STAMMINGER, *Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations*. 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). IEEE, 2019.

Inoltre, non solo sono stati ottenuti dei risultati di classificazione superiori rispetto quanto presente nella letteratura, ma il metodo fornisce anche molti spunti in termini di spiegabilità.

Jain et al.[74] hanno proposto un lavoro noto come DAD-HCNN, basato su una pipeline di classificazione gerarchica composta da tre livelli per distinguere rispettivamente immagini reali Vs alterate tramite tool di editing (primo livello), alterate Vs immagini generate da GAN (secondo livello) e, infine, definire l'architettura GAN specifica (terzo livello). Il framework è in grado di rilevare ritocchi e immagini generate da GAN con elevata precisione.

Un set di dati di riferimento è stato introdotto da Rossler et al.[39] come punto di riferimento per il rilevamento di Deepfake. Si chiama FaceForensics++ e si basa principalmente su quattro metodi di manipolazione: due metodi basati sulla grafica computerizzata (Face2Face[27], FaceSwap[26]) e due approcci basati sul Deep Learning (DeepFakes[25], NeuralTextures[28]).
Considerando approssimativamente la letteratura sul campo, sembra che il rilevamento di Deepfake sia un compito facile, già risolto. Tuttavia, le tecniche analitiche basate sul dominio della frequenza mancano ancora di accuratezza e gli algoritmi basati su reti neurali riescono ad ottenere buoni risultati di classificazione ma tendono a discriminare la semantica più delle tracce specifiche delle varie GAN. Inoltre, queste tecniche sono computazionalmente onerose e difficili da comprendere o controllare[14].

3. Discussione e conclusione

La rapida diffusione e la facilità di utilizzo degli strumenti di editing dei media digitali hanno introdotto metodologie per alterare i file multimediali con tecniche sempre più complesse. D'altra parte, la comunità scientifica si è attrezzata di conseguenza sviluppando algoritmi e tecniche altrettanto efficaci in grado di produrre evidenze forensi della eventuale manipolazione. I rapidi progressi nel settore dell'Intelligenza Artificiale tramite l'utilizzo di tecniche di Deep Learning e del loro paradigma generativo hanno innalzato il livello dello scontro con l'introduzione dei cosiddetti Deepfake.

[74] A. JAIN et al., *Detecting GANs and Retouching based Digital Alterations via DAD-HCNN*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.

I principali motori utilizzati per la creazione di contenuti multimediali sintetici sono gli Autoencoder e Generative Adversarial Network. Queste tecnologie consentono di manipolare volti di persone modificando il colore dei capelli, il genere o persino creare una persona che non esiste nel mondo reale.

Gli attacchi Deepfake di tipo *Spear Phishing*, *CEO Fraud* e *Phishing* possono riguardare anche piccole e grandi realtà aziendali. Rispetto alle tradizionali tecniche di manipolazioni di immagini, un attaccante, attraverso le piattaforme oggi disponibili per la creazione di Deepfake è in grado di creare contenuti multimediali estremamente realistici al fine di perfezionare i tradizionali attacchi informatici come frodi ed estorsioni online, rubare l'identità digitale delle persone e tanto altro ancora. I Deepfake sono in grado di aggirare sistemi di intelligenza artificiale, come quelli di autenticazione: basti pensare ad esempio ad un sistema di autenticazione basato sul riconoscimento del volto che viene attaccato mostrando un'immagine generata tramite GAN per poter accedere in un'area riservata di un'azienda (es. per azioni tipiche di spionaggio industriale). Altre tipologie di attacchi effettuati tramite questa nuova potente tecnologia sono stati di tipo *Spear Phishing* e di *Whaling*. In particolare, in questo contesto, alcuni attaccanti hanno utilizzato immagini Deepfake per creare due profili falsi della stessa persona su due social network differenti quali *Twitter* e *LinkedIn* di una "presunta" giornalista del quotidiano *Bloomberg* chiamata *Maisy Kinsley*[75]. L'obiettivo dell'attacco era quello di realizzare un'azione di spionaggio industriale tramite l'acquisizione di informazioni riservate, entrando in contatto con 195 venditori della società Tesla. Fortunatamente, in questo caso la testata giornalistica stessa ha confermato su richiesta esplicita della controparte, che non esisteva nel suo team una giornalista di nome Maisy Kinsley rendendo di fatto nullo l'attacco. Un esempio invece in cui l'attacco è stato effettuato con successo riguarda una società inglese, in cui è stato autorizzato un bonifico bancario della cifra pari a 200.000 sterline verso un conto corrente ungherese. In quest'ultimo caso la tecnica Deepfake è stata utilizzata per manipolare in real-time il contenuto audio di una telefonata.

[75] https://www.theregister.com/2019/06/17/roundup_ai/.

Altre potenziali abusi riguardano gli aspetti geo-politici; si pensi ad esempio ad una possibile campagna elettorale, in cui dei contenuti video real-time o meno, anche su una testata giornalistica autorevole, potrebbero veicolare il risultato di un Deepfake.

Il settore di maggiore diffusione riguarda senza timore di smentita il mondo della pornografia. Si parla in questo caso di *DeepNude*, *DeepPorn* o *Porno Deepfake*. In questo contesto, il volto della vittima viene sostituito al volto di un attore in un video porno al fine di divulgare contenuti pornografici mai avvenuti. Sotto questo aspetto alcune applicazioni disponibili nei vari *store* sono state ritirate, poiché consentivano di denudare "virtualmente" l'immagine di una persona sotto attacco. In questi contesti si realizzano dei veri e propri attacchi di tipo *Revenge Porn* anche a partire da una quantità limitata di dati della vittima: basta anche una semplice foto del profilo scaricata dai vari social della vittima.

Giuridicamente risulta essere complicato definire e classificare i reati basati su tecnologie Deepfake. Nel dicembre del 2020 i Deepfake sono stati definiti, in un documento pubblicato dal *Garante privacy*[76], come una forma particolare di furto di identità. L'uso illecito di questa tecnologia può essere ricondotto a due tipi di reati: *frode informatica* e *sostituzione della persona*. Nel contesto del *Revenge Porn*, giuridicamente esistono leggi che puniscono coloro che hanno divulgato immagini e video di natura pornografica senza il consenso della persona protagonista.

È sempre più pressante quindi la necessità di creare soluzioni informatiche sempre più sofisticate al fine di contrastare l'uso illecito di questa potente tecnologia.

La maggior parte degli algoritmi di rilevamento di *forgery* classici, non sono in grado di distinguere i Deepfake dai contenuti reali. Come riportato in letteratura, i dati multimediali contengono un'impronta digitale univoca in grado di identificare il modello del dispositivo utilizzato durante la fase di acquisizione (come smartphone e scanner), il social network su cui sono stati caricati i dati e ricostruire la storia dei dati stessi. Recenti studi hanno dimostrato che anche i contenuti Deepfake hanno un'impronta digitale unica lasciata dai modelli generativi. I ricercatori hanno dimostrato di essere in grado di estrarre questa traccia ottenendo

[76] https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9512226#:~:text=I%20deepfake%20sono%20foto%2C%20video,imitare%20fedelmente%20una%20determinata%20voce.

buoni risultati di classificazione nel task *Real Vs Deepfake* e individuando anche l'algoritmo di Deep Learning utilizzato per generarli.

I vari motori GAN disponibili in letteratura utilizzano diverse architetture di generatori e discriminatori (implementati tramite Convolution Neural Network (CNN)) per risolvere diversi compiti (cambiare il colore dei capelli, creare volti di persone che non esistono). Le CNN sono composte da strati convoluzionali e i loro parametri (come la dimensione del kernel e molti altri) caratterizzano i contenuti multimediali sintetici. Risultano essere proprio i layer convoluzionali che lasciano una traccia, la quale può essere rilevata analizzando la correlazione dei pixel dell'immagine[68,69].

Nella continua lotta tra *Forensics* e *anti-Forensics*, con l'avvento delle tecnologie di Intelligenza Artificiale, assistiamo al momento ad uno scontro in cui a modelli generativi sempre più performanti e disruptive, in grado di ingannare l'occhio umano e non solo, si contrappongono tecniche e modelli che ne individuano le tracce nascoste. È chiaro che le competenze e l'esperienza di dominio necessaria per affrontare la sfida deve essere di prim'ordine ma a nostro parere il bravo investigatore digitale può a buon ragione ritenersi ancora ottimista e utilizzare strumenti anch'essi complessi di estrazione, acquisizione e analisi anche nel contesto dei Deepfake.