CAPITOLO PRIMO

ANALISI FORENSE DI IMMAGINI DEEPFAKE DI VOLTI DI PERSONE

Sebastiano Battiato, Oliver Giudice, Luca Guarnera, Cristina Nastasi

Sommario: 1. Introduzione – 2. Tecnologie per la creazione dei Deepfake – 3. Analisi forense dei Deepfake – 4. Trend di ricerca per l'individuazione dei Deepfake – 5. Conclusioni

1. INTRODUZIONE

Manipolare un'immagine o un video consiste nell'alterare in qualche modo la realtà percepita: comunicare un messaggio falso, alterato o anche totalmente fantasioso. Fin da quando esiste la possibilità di rappresentare la realtà con un'immagine grazie alla fotografia, il fenomeno delle alterazioni è sempre stato un rischio [1]: l'immagine viene percepita sempre come reale dal visualizzatore e pertanto il rischio di trasmissione di notizie false o alterate diventa altissimo. D'altro canto, manipolare in maniera realistica un'immagine o peggio ancora un video, richiede delle abilità specialistiche, tempo, strumenti dedicati e una cura del dettaglio altissima: basta un piccolo errore e la manipolazione viene facilmente scoperta. Con l'avvento della fotografia digitale, sono stati realizzati Software sempre più specializzati per l'effettuazione di elaborazioni, addirittura è nato il termine "Photoshoppare" dall'omonimo programma della suite Adobe. Ma sebbene il ritocco sia alla portata di tutti fino ad oggi, il realismo percettivo del contenuto manipolato finale restava sempre legata strettamente all'abilità del manipolatore e alla sua capacità di utilizzare al meglio il relativo Software.
L'evoluzione degli strumenti di Intelligenza Artificiale (IA) però sta cambiando lo scenario appena descritto: complessi modelli matematici, costruiti su miliardi di immagini di volti umani, sono in grado di alterare un volto cambiandone le caratteristiche come ad esempio il colore dei capelli, l'età o addirittura il sesso. Sebbene la costruzione di tali modelli è sicuramente estremamente complessa, l'utilizzo finale è a portata di click. Sui più comuni mobile-store sono apparse applicazioni come FaceApp [2] o Reface [3] che consentono in pochissimi passaggi di creare immagini manipolate di volti umani a un livello di realismo estremamente elevato.
Nasce il fenomeno dei cosiddetti "Deepfake". Politici, attori e personaggi famosi sono i principali bersagli di questo nuovo fenomeno: un famoso esempio è stato il video dell'ex premier italiano Matteo Renzi, mostrato da "Striscia la Notizia" nel settembre 2019, in cui appellava dei suoi colleghi in una maniera "non molto rispettosa". Un altro famoso esempio è stato il video di Mark Zuckerberg in cui esponeva una serie di dichiarazioni sulla capacità di Facebook di rubare dati ai propri utenti. Tanti sono gli esempi di video Deepfake aventi come protagonisti personaggi famosi: uno tra gli esempi più famosi e nel contempo divertenti è sicuramente l'inserimento del volto di Nicholas Cage in alcuni film in cui lo stesso attore non ha recitato, come "The Matrix" o "Fight Club", oppure l'impressionante video in cui Jim Carrey interpreta Shining al posto di Jack Nicholson.
Le tecniche di creazione di Deepfake diventano nuove armi: mentre il fenomeno per quanto riguarda i personaggi famosi è contenibile dai rispettivi uffici stampa che monitorano i canali mediatici e possono immediatamente smentire un Deepfake; il problema diventa legalmente e socialmente più pericoloso se essi vengono impiegati per creare contenuti pornografici. Secondo il rapporto "The State of Deepfakes" [4] del progetto Deeptrace, un lavoro di monitoraggio dei "Deepfake" attivi dal 2018, i video falsi costruiti con tecnologie d'intelligenza artificiale sono poco meno di 15.000. Tra questi il 96% dei video è di natura pornografica (Deep Porn). Purtroppo, questo fenomeno si sta evolvendo rapidamente, diventando sempre più pericoloso: come riportato dal Wall Street Journal, un video Deepfake riguardante l'amministratore delegato di un'azienda tedesca è stato utilizzato dai cyber-criminali per rubare 250.000 dollari da un'azienda inglese.
Il tema del Deepfake assume un'importanza cruciale se analizzato nel campo giudiziario dove una traccia informatica falsa potrebbe addirittura ribaltare il verdetto finale. I problemi, però, non sono solo etici o sociali, poiché i Deepfake rappresentano una tra le emergenti minacce del nostro secolo proprio per la capacità intrinseca di creare "una nuova realtà". La definizione di nuove "Best Practice" di analisi forense e la realizzazione di nuovi strumenti per smascherare i Deepfake, al fine di limitare potenziali danni e abusi, rappresentano al giorno d'oggi una tra le sfide più avvincenti e interessanti per i ricercatori del settore.

Necessariamente, quindi, questa minaccia deve essere contrastata e mitigata. A tal proposito, nella restante parte dell'articolo verranno in prima analisi analizzate, ad alto livello, le tecnologie note ad oggi per la creazione di un Deepfake di volti. La comprensione di tali tecnologie servirà a capire come costruire strumenti adatti ad individuare eventuali caratteristiche invisibili per l'identificazione di un falso.

(a) (b) (c)

Fig 1: Alcuni esempi di Deepfake: (a) Obama, creato da Buzzfeed in collaborazione con Monkeypaw Studios; (b) Mark Zuckerberg, creato dagli artisti Bill Posters e Daniel Howe in collaborazione con l'azienda Canny; (c) Matteo Renzi nel video trasmesso dal Tg Satirico Striscia la Notizia.

2. TECNOLOGIE PER LA CREAZIONE DEI DEEPFAKE

Il termine Intelligenza Artificiale (IA), rappresenta un insieme di tecniche e tecnologie che grazie all'utilizzo congiunto di matematica, statistica e informatica consente di creare sistemi capaci di prendere una decisione dato un problema. Siamo già abituati a vedere l'IA in azione in semplici giochi per computer come gli scacchi. Questi sistemi, seppure efficienti, sono basati su una serie di "regole" codificate dal creatore dell'IA stessa: la decisione dell'IA dipende dalla risoluzione matematica dell'insieme di regole pre-codificate a seconda della situazione. Questo tipo di ragionamento è limitato alla capacità del creatore dell'IA di codificare un insieme completo e corretto di regole. Dall'altra parte è possibile costruire tali regole dall'osservazione della realtà: per esempio guardando una certa configurazione della scacchiera e delle pedine e osservando tutte le partite giocate (o un insieme molto grande) posso costruire "automaticamente" le regole che possono portare alle decisioni migliori per vincere la partita. Un tipo di ragionamento basato su esperienze, su uno storico (grande) di dati. L'insieme di tecniche in questione fanno capo al Machine Learning ed in particolare nelle sue specializzazioni più moderne e attuali al Deep Learning. Con particolare attenzione al mondo delle immagini, è possibile analizzare milioni di immagini di un volto e imparare le regole di distribuzione statistica dei colori dei singoli pixel per permettere di discernere l'immagine di un volto da quella di un paesaggio o di un cane. Questo tipo di ragionamento basato sul Deep Learning è in qualche modo solo discriminatorio: apprendo le regole che mi consentono di decidere se un'immagine rappresenta un volto o un paesaggio. Ma una volta conosciuta la distribuzione statistica rappresentativa di un volto, ci si può chiedere se è possibile generare un qualsiasi volto che rispetti tale distribuzione. Una volta che conosco le regole per discriminare bene un volto da qualcos'altro, posso utilizzare tali regole per costruire un generatore di volti perfetto?
Proprio a partire da questa idea, sono state inventate le Generative Adversarial Network (GAN): una particolare tecnica di Deep Learning introdotta da Ian Goodfellow [5] nel 2014. Goodfellow ha ideato un nuovo framework composto da un generatore G in grado di creare artificialmente immagini a partire della distribuzione statistica di grandi quantità di dati (DT) di un dominio d'interesse (ad esempio volti) e un modello discriminatorio D che stima la probabilità che il campione in esame provenga da DT (reali) piuttosto che da G (generate). La procedura innovativa di addestramento utilizzata per G ha lo scopo di massimizzare la probabilità che D commetta un errore. In altre parole, il discriminatore non sarà più in grado di distinguere tra i dati reali e quelli generati da G. Per capire meglio questa nuova tecnologia consideriamo il seguente esempio: immaginiamo che il modello generativo sia rappresentato da una squadra di falsari che cerca di produrre e usare, senza essere scoperti, una valuta falsa. In questo caso possiamo immaginare che il modello discriminativo sia rappresentato dalle forze dell'ordine che cercano di individuare la valuta contraffatta in un "gioco" in cui G è completo una volta che D non è più in grado di distinguere le valute reali dalle contraffatte.

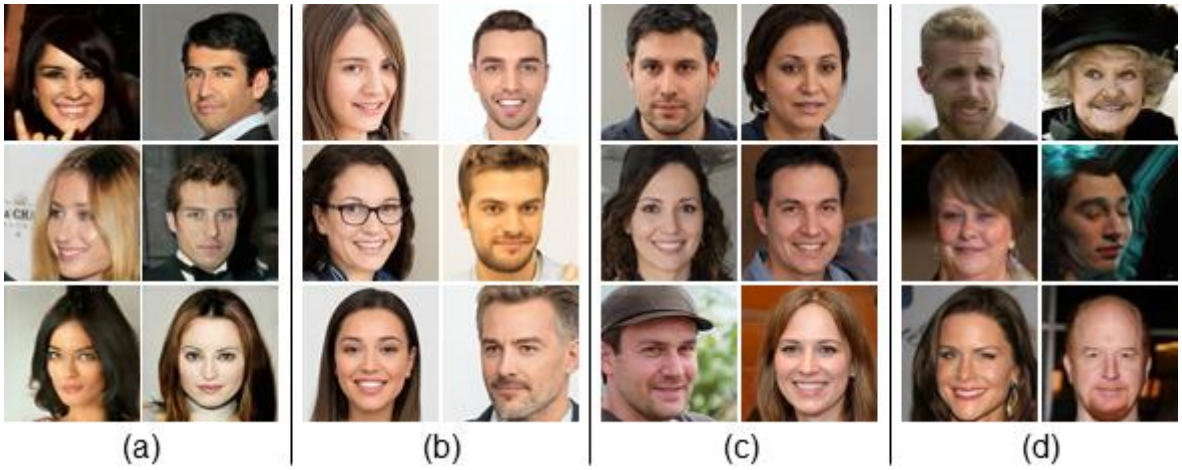

Fig 2: Alcuni esempi di immagini Deepfake generate dalle architetture (a) StarGAN; (b) StyleGAN; (c) StyleGAN2; (d) AttGAN.

Di seguito sono riportate le principali tecniche GAN per la creazione di volti Deepfake:

- **StarGAN [6]:** è un'architettura GAN in grado di eseguire operazioni di "*traslazione immagine da immagine*" (*image-to-image translation*) di volti di persone su domini multipli (cambio di colore di capelli, espressione facciale, ecc.). Data in input un'etichetta come il colore dei capelli, StarGAN esegue l'operazione di *image-to-image translation* dando all'immagine target la caratteristica richiesta.
- **Style Generative Adversarial Network (StyleGAN) [7]:** è un'architettura GAN in grado di generare immagini realistiche di alta qualità e di offrire il controllo di parametri in termini di "stile" per generare l'immagine a diversi livelli di dettaglio.
- **StyleGAN2 [8]:** è un'architettura GAN in cui è stato modificato il modello generativo della struttura originale di StyleGAN in modo da rimuovere alcune imperfezioni presenti nei dati artificiali generati ottenendo immagini ancora più realistiche.
- **AttGAN [9]:** è un'architettura GAN in cui il modello generativo effettua controlli sugli attributi facciali da alterare in modo da garantire le modifiche degli attributi desiderati.

3. ANALISI FORENSE DEI DEEPFAKE

Essendo un Deepfake un'immagine digitale è possibile tentare di affrontare il problema dell'autenticazione di tale segnale tramite l'uso delle tecniche note di analisi forense delle immagini. In generale, nel settore della Multimedia Forensics, vengono analizzati differenti elementi del dato digitale per stabilirne l'autenticità e l'integrità [1]. Un elemento caratterizzante è il cosiddetto Photo Response Non-Uniformity (PRNU), ovvero un'impronta digitale della sorgente di acquisizione, come ad esempio, smartphone, scanner, ecc.. Altri elementi possono riguardare l'analisi dei coefficienti della compressione JPEG e in ultimo l'analisi dell'immagine nel dominio di Fourier. Se una specifica informazione non è presente o risulta essere "anomala", sorge il sospetto che i dati in esame non siano reali.
Attualmente, sembra che le GAN che creano Deepfake non siano in grado di ricreare il processo di acquisizione e digitalizzazione del dato digitale che viene invece effettuato da qualsiasi strumento di acquisizione, questo perché tali tecniche sono addestrate utilizzando immagini acquisite da tantissimi dispositivi diversi.
Per effettuare un'analisi preliminare con tecniche disponibili sul mercato è stata effettuata un'analisi dettagliata tramite l'uso del software "Amped Authenticate" [10], su un campione di immagini Deepfake generate da due particolari tecnologie GAN: StarGAN e StyleGAN. In particolare sono stati analizzati differenti elementi, quali, la struttura JPEG dell'immagine, il PNRU; i diversi spazi colore (RGB, YCbCr, YUV, HSV, HLS, XYZ, LAB, LUV, CMYK); e sono state applicate varie operazioni atte a risaltare anomalie causate da una presunta manipolazione (ELA, ELA, DCT Map, JPEG Dimples Map, Blocking Artifacts, JPEG Ghosts Map, Fusion Map, Correlation Map, PRNU Map, PRNU Tampering, LGA). Infine, è stata effettuata l'analisi a livello pixel dell'immagine, con algoritmi automatici (ClonesBlocks, Clones Keypoints, Orb and Brisk).

La Figura 3 mostra visivamente il risultato dell'applicazione di alcune delle operazioni appena elencate. Dai report finali, ottenuti tramite Amped Authenticate, è possibile notare tutte le anomalie evidenziate.

Fig3: Esempio di analisi eseguita con il software Amped Authenticate. (a) Immagine generata da STYLEGAN. (b) Immagine generata da STARGAN. In entrambi gli esempi mostriamo solo alcuni degli elementi analizzati con Amped Authenticate, come i diversi spazi colore analizzati per un eventuale riscontro di anomalie; ELA (Error Level Analysis) per l'identificazione delle aree di giunzione dell'immagine che sono state compresse in modo diverso; "Correlation map" per analizzare e identificare la correlazione tra i pixel; "Clones Keypoints" chiave per trovare parti che sembrano essere clonate.

Ulteriori interessanti caratteristiche sono individuabili tramite l'analisi di un'immagine Deepfake nel dominio di Fourier con modalità simili a quelle enunciate nell'articolo di Guarnera et al. "Preliminary Forensics Analysis of DeepFake Images" [11]. In generale, data un'immagine, rappresentarla nel dominio delle frequenze significa effettuare un'analisi statistica d'insieme di un'immagine sulla "ripetitività" dei valori dei colori nei pixel di un'immagine. La Figura 4 mostra alcuni esempi del risultato della trasformazione di un'immagine nel dominio di Fourier; ovvero gli *spettri di Fourier*. Si dimostra che le immagini digitali naturali, ovvero non manipolate, hanno uno *spettro* molto simile a quelli riportati in Figura 4 (b) (colonna 1). Invece, la maggior parte delle immagini Deepfake contiene delle componenti frequenziali anomale che risultano essere facilmente evidenti come punti e linee in Figura 4 (a) (colonna 1). È possibile evidenziare ulteriormente queste frequenze, applicando differenti operazioni di filtraggi sulle immagini come ad esempio il "filtro mediano", capace di estrarre le basse frequenze e il "filtro laplaciano", capace di estrarre invece le alte frequenze. L'applicazione e la combinazione di tali filtri permettono di evidenziare ulteriormente le frequenze anomale discusse precedentemente (Figura 4).
Sfortunatamente, quanto fin qui descritto non basta a definire con certezza se le immagini in esame sono o meno Deepfake poiché si può solo dedurre che i dati analizzati sono solo non autentici o che l'integrità è corrotta. Ma quanto ottenuto con strumenti disponibili sul mercato e allo stato dell'arte è un buon punto di partenza. Nel prossimo paragrafo verranno descritti i principali trend di ricerca per contrastare il fenomeno dei Deepfake.

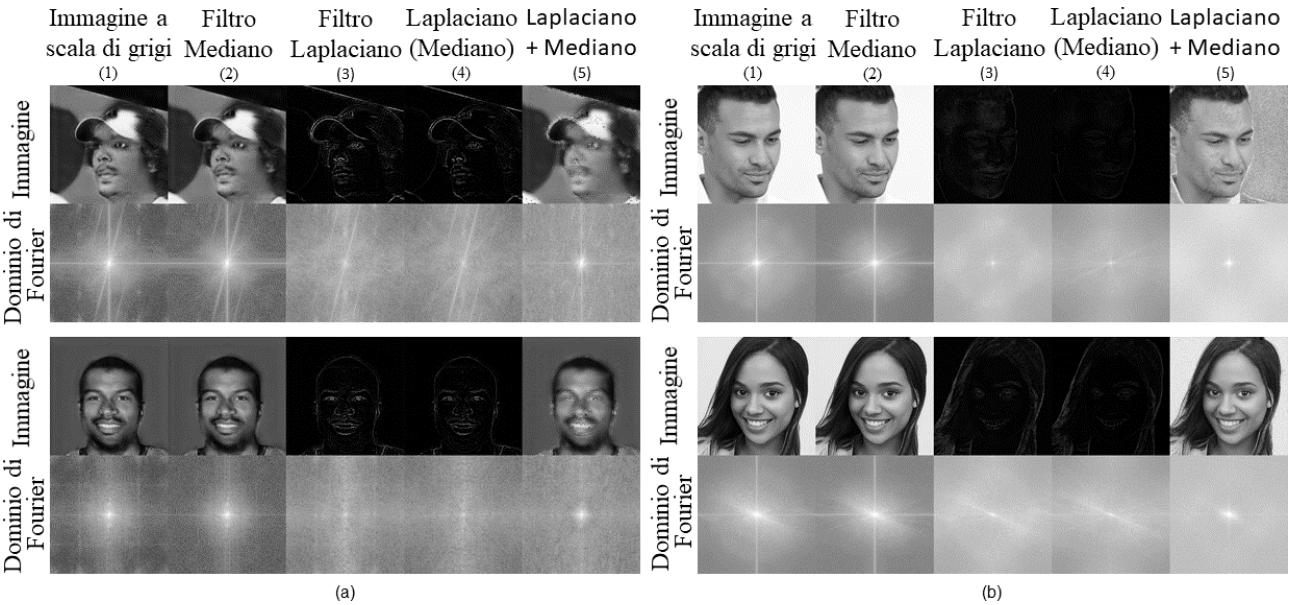

Fig4: Esempi d'immagini analizzate generate da (a) STARGAN e (b) STYLEGAN. Ciascun'immagine di entrambi i set di dati è stata convertita in scala di grigi (1) e applicati progressivamente i filtri: mediano (2), laplaciano (3), laplaciano (4) applicato al risultato di (2), la somma dei risultati tra la mediana e filtri laplaciani (5). Per ogni operazione eseguita, mostriamo la trasformata di Fourier.

4. TREND DI RICERCA PER L'INDIVIDUAZIONE DEI DEEPFAKE

Il Deepfake è un'immagine estremamente realistica, test condotti dalla Polizia Olandese [12] hanno dimostrato che esperti di settore sono capaci solo di riconoscere poco più del 60% dei Deepfake. Ma se i Deepfake sono ottenuti con tecniche di IA, allora si può usare la stessa IA per costruire strumenti per individuarli e contrastare il fenomeno. Differenti metodologie presentate nel consesso accademico, sfruttano le potenzialità delle reti neurali come ad esempio nel lavoro di Wang et al. [13], in cui gli autori hanno realizzato uno strumento in grado di distinguere immagini reali da immagini generate da differenti architetture GAN. Purtroppo il problema non è stato risolto totalmente, infatti una delle problematiche principali legate all'uso delle reti neurali per il rilevamento dei Deepfake è la non generalizzabilità, poiché l'architettura impara a distinguere solo le immagini provenienti da quelle particolari architetture GAN già note non sarà in grado di individuare Deepfake generati da GAN sconosciute.
Attualmente i ricercatori del settore stanno focalizzando la loro attenzione sulla realizzazione di nuove tecniche di rilevamento che quindi siano più generalizzabili e non facciano uso di reti neurali. È di settembre 2020 il risultato ottenuto da Guarnera et al. [14 - 15], dove le risultanze preliminare riportate nel paragrafo precedente trovano riprova con un metodo in grado di individuarle. Infatti in tali studi è stato dimostrato che le attuali architetture GAN lasciano un'impronta digitale sulle immagini (simile al PRNU ma diversa rispetto agli attuali dispositivi di acquisizione.) che caratterizza quella specifica GAN. Per individuare questa traccia però viene utilizzata un'altra tecnica di IA ovvero la classe di algoritmi di "Expectation/Maximization" [16], una tecnica di apprendimento automatico non supervisionata (ovvero non occorre associare ad ogni immagine l'appartenenza al un dato insieme sia esso Deepfake o reali) che ha l'obiettivo di individuare e descrivere quella che viene chiamata "Convolutioanl Trace" ovvero l'impronta digitale delle GAN.

5. CONCLUSIONI

L'analisi preliminare effettuata sul rilevamento di contenuti creati tramite tecniche di intelligenza artificiale mostra che l'utilizzo di software commerciali (come, ad esempio, Amped Authenticate) aiutano ad identificare la presenza di possibili anomalie presenti nel dato digitale in esame. Si evince inoltre che l'applicazione di filtri noti (come il mediano e il laplaciano) e le relative combinazioni sulle immagini digitali risaltano ulteriormente la presenza di anomalie nel dominio di Fourier. Molto c'è ancora da fare per combattere il pericoloso fenomeno dei Deepfake, ma i recenti trend di ricerca fanno ben sperare e presto le soluzioni

attualmente solo teorizzate a livello accademico, saranno disponibili a breve per la comunità investigativa e forense.

REFERENZE:

[1] S. Battiato, F. Galvan, M. Jerian, M. Salcuni, “ Linee Guida per l’autenticazione forense di immagini”. Capitolo VIII, IISFA Memberbook 2013.
[2] FaceApp, https://www.faceapp.com/
[3] Reface, https://get.reface.app/
[4] H. Ajder, G. Patrini, F. Cavalli, and L. Cullen, “The State of Deepfakes: Landscape, Threats, and Impact”, September 2019.
[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[6] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8789–8797, 2018.
[7] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4401–4410, 2019.
[8] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. "Analyzing and improving the image quality of stylegan." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[9] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. Attgan: Facial attribute editing by only changing what you want. IEEE Transactions on Image Processing, 28(11):5464–5478, 2019.
[10] Amped Authenticate, https://ampedsoftware.com/authenticate
[11] L. Guarnera, O. Giudice, C. Nastasi, and S. Battiato, “Preliminary forensics analysis of deepfake images,” arXiv preprint arXiv:2004.12626, 2020.
[12] N. Hulzebosch, I. Sarah, and W. Marcel. "Detecting CNN-Generated Facial Images in Real-World Scenarios." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.
[13] S. Wang, O. Wang, R. Zhang, A. Owens, and A. Efros, “Cnn-generated images are surprisingly easy to spot...for now,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,, 2020.
[14] L. Guarnera, O. Giudice, and S. Battiato. "DeepFake Detection by Analyzing Convolutional Traces." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020
[15] L. Guarnera, O. Giudice, and S. Battiato. "Fighting Deepfake by Exposing the Convolutional Traces on Images." IEEE Access 8 (2020): 165085-165098.
[16] T. K Moon. The expectation-maximization algorithm. IEEE Signal Processing Magazine, 13(6):47–60, 1996.