Metoder för detektion av djupfejkmaterial

Joar Sabel 1800130

Handledare: Mats Aspnäs

Fakulteten för naturvetenskap och teknik

Åbo Akademi

11.11.1444

Innehåll

Introduktion

Under de senaste åren har man kunnat notera en stor ökning i antalet videor och bilder där t.ex. ansikten på kändisar eller andra personer blivit utbytta mot ett annat ansikte. Dessa bilder eller videor är onekligen imponerande men oftast, om producerade av en enda person och är gjorda för komisk effekt, är det oftast lätt att med bara ögat känna igen att videon eller bilden blivit manipulerad.

Förfalskade bilder eller videor som ett fenomen i sig är inget nytt. Detta diskuteras i introduktionen i Güera et al[1]. Man ser förfalskat material vardagligen i film eller tv-program, i formen av datorgenererad grafik eller photoshoppadebilder. Djupfejkning är dock något helt annat. Man kan se en markant skillnad mellan en person i film som blivit skapad med datorgrafik och en person som skapats med djupfejkning, se figuren nedan.

Figur 1.1: Till vänster, karaktären Tarkin, gjord med datorgrafik och till höger karaktären Luke Skywalker som blivit djupfejkad.

Dessa exempel är tagna från Star Wars[1], man bör bekanta sig med de originella karaktärerna för att förstå vad som gjorts, men man ser direkt att den djupfejkade versionen är

[1] *https://www.starwars.com/*

betydligt mindre plastig"eller målningslik"än den som genererats med datorgrafik.

Denna avhandling behandlar fenomenet, som i Engelskan fått benämningen Deepfake", baserat på att förfalskningsprocessen använder sig av djupinlärning, eller som det heter på engelska: 'deep learning'. Avhandlingen kommer att i huvudsak fokusera på djupfejkning av bilder och videor, djupfejkning av röster får bli en annan skribents uppgift. Mer specifikt kommer denna avhandling att behandla de metoder man kan dra nytta av när man försöker detektera om en bild eller video har manipulerats, samt försöka klargöra vad det är man söker för fel i dessa videor eller bilder för att bevisa om de blivit förfalskade.

Till att börja med kommer lite bakgrundsinformation relaterad till fenomenet att presenteras, följt av vilka problem som uppstått på grund av djupfejkning och även relevant statistik. Efter detta kommer produktionen av materialet att behandlas, till en sådan nivå att man har en grundförståelse av vad som sker. I kapitlet om identifiering kommer olika identifieringsmetoder att inspekteras och vi kommer att se på vilka fel man kan hitta i djupfejkat material. Till slut kommer en inblick i resultatet av avhandlingen samt en sammanfattande avslutning.

1.1 Bakgrund

Förut gällde det att ha tillgång till rätt så kraftfulla maskiner för att kunna producera djupfejkat material. I och med den ökade tillgången till kraftfullare maskiner och mjukvara som tillåter även lekmän att producera trovärdiga förfalskade bilder och videor, har dock antalet sådant material ökat enormt under de senaset åren.

Fenomenets ökande popularitet kan till stor del tillskrivas de framsteg som gjorts inom artificiell intelligens. En av de främsta framstegen gjordes 2014, när Goodfellow et al[2] gav ut sin artikel om generativa adversariella nätverk (GAN), men dessa kommer diskuteras utförligare senare i avhandlingen.

Den ökade tillgången till kraftfulla datorer och bra mjukvara[3, 4, 5] har gjort djupfejk-

ningsprocessen markant mer tillgänglig. Onekligen har denna ökning i tillgänglighet givit de flesta ett gott skratt någon gång om man har sett en video [2] eller bild som blivit djupfejkad. Eller kanske imponerat, när man sett någon som inte längre är vid liv, rakt framför sig på sin skärm [3].
Allt djupfejkat material är inte gjort för att vara humoristiskt. En stor del problem har uppenbarat sig i och med populariseringen av djupfejkande och dessa problem är stora bidragande faktorer till att detektion av djupfejkat material är så viktigt.

1.2 Problem

Ett stort problem med internetbaserat material är att man inte kan lita på allt man läser eller ser. En välplacerad förfalskad artikel, bild eller video kan göra stor skada och ha långtgående konsekvenser förrän den bedöms vara falsk.

Tanken dras då kanske omedelbart till olika politiska kampanjer och därtill relaterade missinformationskampanjer. Missinformation i sig är inget nytt, men i och med uppkomsten av djupfejkning, har den fått nytt liv. en av kvalitéerna hos internet är den kvicka spridningen av information, men detta kan också vara en svaghet. Det krävs inte mycket fantasi för att tänka ihop ett scenario där djupfejkmaterial plus snabb spridning via internet kunde leda till en stor katastrof.

Den största potentiella skadan kan orsakas på social media, där fakta och sanning kanske inte är av största vikt. De största plattformerna har dock redan för några år sedan meddelat att de har procedurer igång för att bekämpa sådant förfalskat och skadligt material[6].

En annan grupp som blivit grovt utsatta för djupfejkning är s.k. "kändisar". Filmstjärnor, artister, influencers och kända företagsledare (ex. Mark Zuckerberg[4], Jeff Bezos[5] etc.) är

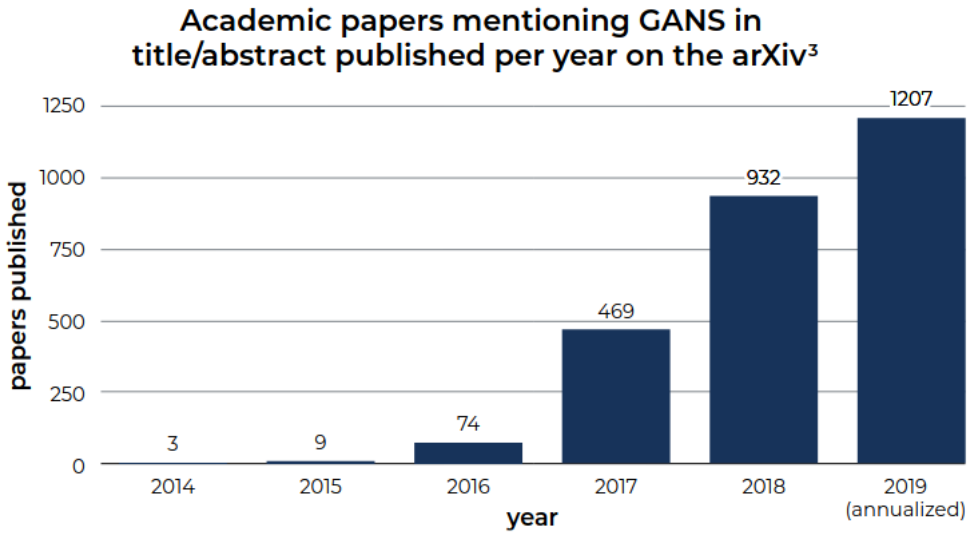
[2] $https://www.youtube.com/watch?v=l_6Tumd8EQI$
[3] https://www.youtube.com/watch?v=mPtcU9VmIIE&t=23s
[4] https://www.youtube.com/watch?v=Ox6L47Da0RY
[5] https://www.youtube.com/watch?v=fqye4Cw0EF0

de grupper som blivit utsatta mest, både för humoristiska syften och mer elakartade syften.

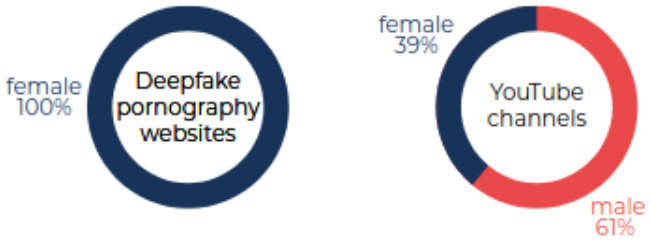
Medans kända män oftast utsätts för komiska djupfejkningar, så utsätts kvinnliga kändisar oftare för pornografiska djupfejkningar, statistik över detta kommer i sektionen om statistik. Detta är problematiskt och är inte lätt att bekämpa, eftersom då något en gång laddas upp på internet så är det svårt att totalt eliminera. Även om större vuxenunderhållningssidor har sagt att de inte tolererar djupfejkmaterial[6], så är det i verkligheten inte så lätt att genomföra.

Ett stort problem som också har uppstått p.g.a djupfejkmaterial är relaterat till polisundersökningar, där bilder och videor ofta används som bevis när man försöker lösa ett brott. I och med att sådant material nuförtiden kan förfalskas kan denna process försvåras markant. I artikeln av Chesney et al[7] diskuteras saken och där presenterar de flera scenarier där djupfejkmaterial kunde vara till skada för samhället. Här kunde man addera scenariot att bild- eller videobevis kan vara förfalskade. Speciellt när djupfejkande är så lätt tillgängligt kan nästan vem som helst förfalska bevis, vilket kan dra ut på, om inte förstöra, en hel polisutredning.

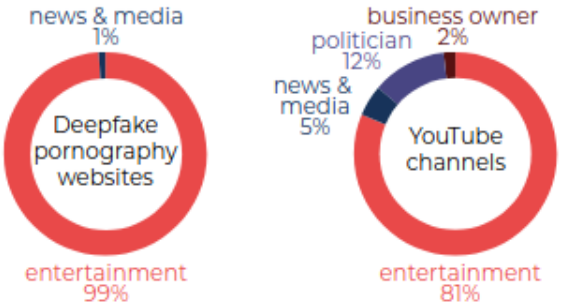
1.2.1 Statistik



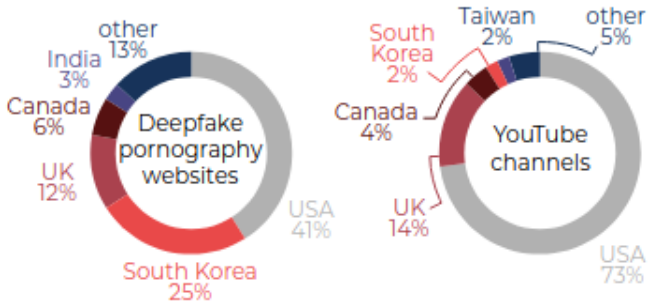
Figur 1.2: Här ser man den ökande populariteten av artiklar relaterade till GAN(2019) [8]



Figur 1.3: Fördelning av kön hos dem som blivit utsatta för djupfejkning (2019) [8]



Figur 1.4: Fördelningen av yrken hos dem som blivit utsatta för djupfejkning (2019) [8]



Figur 1.5: Fördelningen av nationaliteter hos dem som blivit utsatta för djupfejkning (2019) [8]

I denna sektion presenteras statistik relaterad till djupfejkmaterial. Detta kommer kanske mer konkret illustrera några av de problem som diskuterats i föregående sektion.

Låt oss först se på Deeptrace:s undersökning om djupfejkmaterial från 2019. Figur 1.2 beskriver den massiva ökningen i akademiska artiklar relaterade till generativa adversariella nätverk, man kan nog göra antagandet att också ökningen i djupfejkandets popularitet kan ses i denna graf. Det är föga förvånande att intresset för GAN och konsekvent djup-

fejknade ökar, speciellt när man ser vidare på statistik för djupfejkmaterial.

Figur 1.3 illustrerar den oroväckande fördelningen mellan djupfejkmaterial med män i fokus och material med kvinnor i fokus på Youtube och vuxenunderhållningssidor. Om man sedan ser på figur 1.4 som visar vilket yrke personer som blivit utsatta för djupfejkande har, kan man klart se att det är kändisar, eller personer i media, som är i fokus när det kommer till vuxenunderhållning. På Youtube är dock fördelningen aningen mer divers men också där dominerar djupfejkmaterial av personer inom underhållningsindustrin.

Figur 1.5 illustrerar distributionen av folk av olika nationaliteter som blivit utsatta för djupfejkning. Resultaten är kanske aningen annorlunda än man kunde tro. Den största andelen utsatta är från engelsktalande delar av världen men t.ex. i vuxenunderhållningskategorin håller Sydkorea en relativt stor andel, troligen relaterat till den ökade popularitеten av K-pop under de senaste åren. Man kan dock ifrågasätta denna illustrations korrekthet, eftersom information från länder som t.ex. Kina är ytterst svårtillgänglig, om alls tillgänglig.

Generering

Som tidigare nämnt baserar sig djupfejkande på s.k. djup inlärning (cite). Vilket är en form av artificiell intelligens som använder sig av stora mängder data för att lärasig olika saker, t.ex. hur man syntetiserar tal, förutspå kommande händelser baserat på tidigare data eller att rita ett ansikte. Detta ritade ansikte kan sedan ritas på en annan individ i en skild video eller bild.

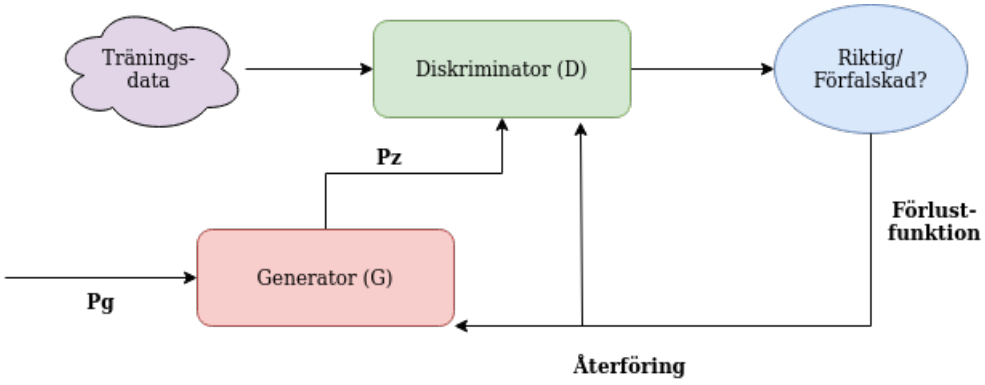
I följande kapitel behandlas genereringen av djupfejkmaterial, samt relaterade ämnen som träningsdata. I primär fokus är generativa adversariella nätverk (GAN) som gav upphov till den växande populariteten av djupfejkning, när de först diskuterades i Goodfellow et al [2] 2014.

2.1 Hur skapas djupfejkmaterial?

Skapandet av djupfejkmaterial är på ytan relativt simpelt, man behöver ett neuralt nätverk och träningsdata, varpå nätverket lär sig av träningsdatan och därefter kan skapa en förfalskad version av en bild eller video.
Nuförtiden finns det tillgång till flera utmärkta program som tillåter en användare att skapa djupfejkmaterial, program som DeepFaceLab(DFL)[4] och DeepFaceLive(DFLive)[5].
DeepFaceLab baserar sig på ett GAN och vi kan således se på genereringsprocessen genom att ta en närmare titt på hur DFL går till väga. Först behöver vi dock förstå vad ett GAN är och vad träningsdata skall bestå av.

2.1.1 Generativa adversariella nätverk (GAN)

Generativa adversariella nätverk[2] är en form av neuralt nät[9], som presenterades första gången 2014 och har sedan dess växt explosionsartat i popularitet Sedan dess har också många variationer av GAN skapats, som STARGAN[10] och STYLEGAN[11] som båda följer samma principer men har lite varierande användingsfall.



Figur 2.1: En förenklad bild av hur ett GAN fungerar. Översatt version av bilden från[12]

GAN består av en generativ modell (G), som fångar upp datafördelningen (Pg) och matar ut en förfalskad bild eller video (Pz) till diskriminatorn (D), som uppskattar sannolikheten att ett prov av datan är från träningsdatan snarare än från G. Således kan G:s uppgift beskrivas som att den försöker maximera chansen att D gör ett misstag. Processen kan således i princip, som sägs i Guarnera et al[12], ses som att G är en grupp med förfalskare som försöker skapa falska pengar medans D kan ses som polisen som försöker ta fast dem. Om D alltså lyckas tyda att bilden är förfalskad, görs ett nytt försök och så går det runt, tills det att D inte lyckas känna igen att bilden är förfalskad mer. I det skedet har man då en genererad djupfejkad bild eller video.

2.1.2 Träningsdata

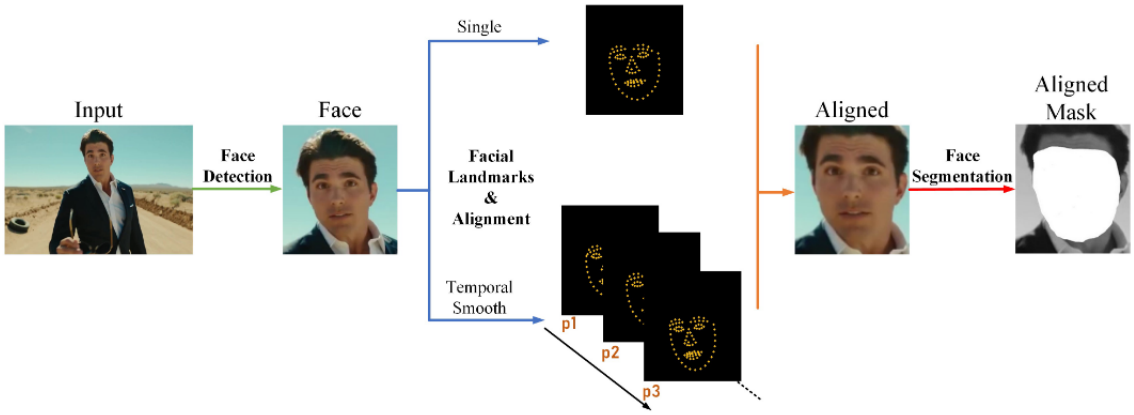
Träningsdatan som behövs för att skapa djupfejkmaterial skall i mån om möjlighet vara så mångsidig men tydlig som möjligt. Detta betyder att man vill ha till exempel en video där ansiktet ska vara i fokus och tydligt men helst också i många olika ljus och att videon fångar ansiktet från olika vinklar och att personen i fråga gör många olika grimaser. Således får nätverket träna på att rita många 'olika' ansikten fast det i verkligheten är frågan om samma ansikte.
Lyckligtvis, finns det många olika paket med träningsdata tillgängliga. Bland annat Meta detection challenge [13] ger använde tillgång till ett stort dataset som innehåller massvis med trängingsdata och målet är förstås att uppnå en så hög grad av detektion som möjligt i denna utmaning.
Exempel på dataset som innehåller bilder på riktiga ansikten är Flickr-Faces-HQ [11] och CELEBA-HQ [14].

2.1.3 Överblick med hjälp av DFL

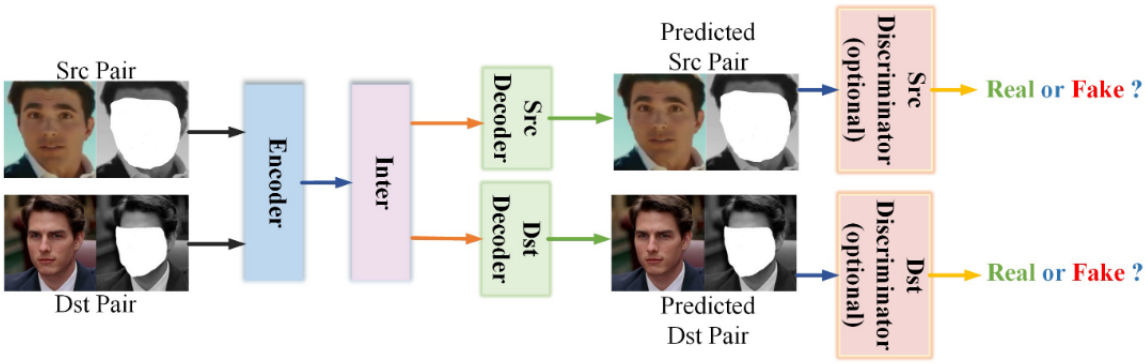
När man då vet vad som behövs, låt oss då se på själva processen. Denna del kommer att baseras mycket på processen som förklaras i Perov et al[4], där de beskriver hur DFL gör för att generera djupfejkmaterial.

För att påbörja processen måste DFL först detektera och extrahera ansikten från både träningsmaterialet(src) såväl som målmaterialet(dst). Detta kan ses i figur 2.4, extraheringen görs på både src och dst, dessa kan också ses som källan och destinationen. Detektionen i DFL görs med S3fd(cite).
Till näst bör ansikten anpassas/justeras så att de passar ihop på varandra. I detta skede är s.k. 'landmärken' i ansiktet viktiga för att kunna anpassa de två ansiktena ordentligt. Detta görs i DFL med hjälp av 2DFAN(cite) och PRNet(cte).
Efter att ansiktsjusteringen är klar är får man en mapp med de justerade ansiktena, varpå TernausNet (cite) algoritmen körs och denna segmenterar ansiktena pixelvis och gör dem till s.k. masker. Efter detta innehas allt som behövs för träningsfasen: utskurna ansikten och deras koordinater i originalmaterialet, ansiktenas landmärken, justerade ansikten och

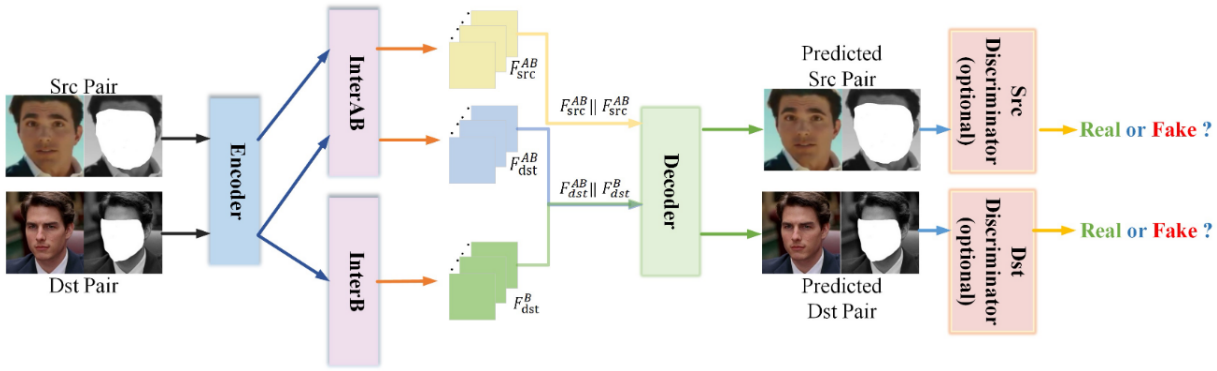
pixelvis segmenterade masker. En sammanfattande illustration kan ses i figur 2.2.



Figur 2.2: En illustration av extraheringsprocessen, från[4]



Figur 2.3: Illustration av träningsprocessen med DF strukturen i DFL, från[4]



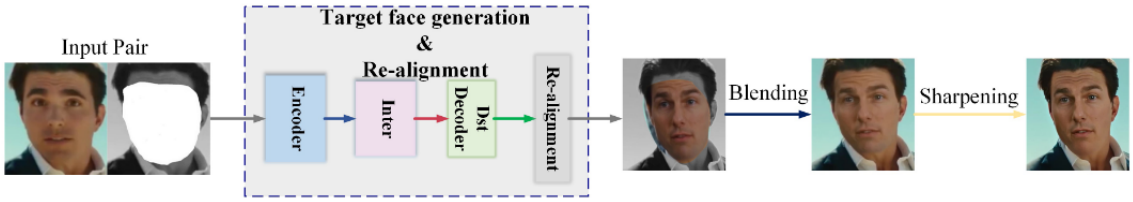
Figur 2.4: Illustration av träningsprocessen med LIAE strukturen i DFL, från[4]

Efter all behövligt material samlats kan träningsskedet påbörjas. DFL använder sig av två strukturer under träningsskedet, DF och LIAE strukturerna. Bägge behövs för att lösa olika problem under träningsprocessen. Dessa två strukturer i konjunktion ämnar lösa dessa problem samt uppnå en fin kvalité på förfalskningarna. För att klargöra dessa bilder en aning kan allting i figur 2.3 och figur 2.4 jämnföras med figur 2.1. Allting före diskriminatorerna kan ses som generatorn

DF strukturen, som kan ses i figur 2.3, är aningen simplare än LIAE och innehåller en enkodare samt en inter-modul med jämnt fördelade vikter på både src och dst materialet, dock har src och dst skilda dekodare.
Eftersom src och dst delar enkodare och inter-modul löser det problem med att de två olika ansiktena inte skulle ha exakt samma ansiktsuttryck. Fastän DF strukturen kan göra ansiktsbytet kan den inte ta tillvara tillräckligt med information från dst, således lider t.ex. belysningen i dessa djupfejkningar.
Här kommer LIAE strukturen in, som ses i figur 2.4. Inom denna struktur har src och dst gemensamma enkodare och dekodare men har skilda inter-moduler. En annan markant skillnad är att medan InterAB producerar latent kod från både src och dst så producerar InterB bara latent kod från dst. Här illustreras denna latenta kod genom t.ex. F_{src}^{AB} för den latenta koden från src. Vi ser också att F_{dst}^{AB} och F_{dst}^{B} sammanfogas och kommer går in i dekodaren tillsammans, härifrån är det sedan möjligt att preservera finare detaljer som finns i dst, till exempel belysningen o.s.v.
Utöver detta finns det ytterligare metoder för att förbättra kvalitén av de genererade ansiktena, t.ex. kan en viktad mask i generell SSIM(cite) användas för att sätta mer vikt på en viss del av ansiktet under träningsprocessen, t.ex. ögonen, vilket resulterar i bättre detalj runt ögonen på det genererade ansiktet.

Till sist är det omvandlingsskedet. I detta skede ser man till att ansiktet är på rätt plats i relation till målmaterialet och att saker som t.ex. hudfärg blandas korrekt mellan ansikten, ansiktsformer och belysningen. Till detta använd färgkorrigerings algoritmer som t.ex. RCT(cite) och Poisson bildmanipulering(cite). Till sist skärps smådetaljer i ansiktet så som födelsemärken etc. En illustration över de sista stegen i processen kan ses i figur

2.5.



Figur 2.5: Illustration av de sista stegen i DFL, från[4]

Så funkar skapandet av djupfejkmaterial i korthet i DFL. Varför ska man dock skapa kraftfulla och användarvänliga verktyg för djupfejkning ifall man vill förhinda spridningen av misinformation? Enligt skaparna av DeepFaceLab så var deras tankegång: ’Det bästa försvaret är ett bra anfall’, det vill säga att de vill visa allmänheten hur lätt det är att skapa djupfejkmaterial och således få folk att tänka efter lite mer förän de litar på allt de ser. Sannerligen en intressant, men aningen icke-konkret tankegång när det kommer till att bekämpa djupfejkmaterial. I nästa kapitel kommer dock mer konkreta preventiva metoder att diskuteras, nämligen kärnan av denna avhandling: metoder för detektion av djupfejkmaterial.

Identifiering

Kriminalteknik inom digital media är inte ett nytt fenomen [15] men djupfejkmaterial tenderar slå vanliga metoder, således behövs nya och mer specialicerade metoder. Detta kapitel kommer att behandla identifieringen av djupfejkmaterial. Först en genomgång av vad det är man söker efter när man försöker detektera djupfejkmaterial samt en kort inblick i Meta detection challange. Efter detta kommer olika detektionsmetoder att presenteras samt deras resultat.

3.1 Vad söker man?

När man försöker identifiera djupfejkmaterial finns det olika ledtrådar man kan söka efter och baserat på vilken metod man använder varierar det vila ledtrådar man söker. För människor är det relativt lätt att se om någonting inte är som det ska, men hur lär man en dator att se sådana saker? I de senare delarna där olika metoder behandlas mer noggrant kommer det framgå vad de metoderna strävar efter att hitta. - **WIP** vilka sorts saker söker man efter för att kunna bestämma om fake

3.2 Meta Detection Challenge

Meta har skapat en 'deepfake detection challenge'[13] för att mäta hur detektion av djupfejkmaterial utvecklas. Detta har gjorts i sammarbete med andra industriledande företag och utmaningen är öppen till alla som vill försöka sig på den. På deras sida hittar man också gott med material för att ta sig an utmaningen; dataset som innehåller videor med hög upplösning, artiklar relaterade till ämnet och ansiktsmodifierande algoritmer. - **WIP** explain things

3.3 Metoder

I denna del kommer olika detektionsmetoder att utforskas, de två som kommer ligga i störst fokus är RNN[1] och CNN(cite). - **WIP** vilka metoder kan man använda

3.3.1 Återkopplade neurala nätverk (RNN)

- WIP tidsdiskret, LSTM

3.3.2 Faltningsnätverk (CNN)

- WIP pixel för pixel dittandattan

Resultat

- vilka metoder verkar bäst? - kan man lita på att en video är äkta eller inte efter användning av metoder (accuracy etc...)

Avslutning

- sammanfattning? - några brillianta insikter!

Källförteckning

[1] D. Guera och E. J. Delp, ”Deepfake video detection using recurrent neural networks,” *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2018.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville och Y. Bengio, ”Generative adversarial nets,” *Advances in neural information processing systems*, årg. 27, 2014.

[3] Y. Goncharov, *FaceApp*, 2016. URL: `https://www.faceapp.com`.

[4] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang m. fl., ”DeepFaceLab: Integrated, flexible and extensible face-swapping framework,” *arXiv preprint arXiv:2005.05535*, 2020.

[5] I. Perov, *DeepFaceLive*, 2021. URL: `https://github.com/iperov/DeepFacelive`.

[6] S. Cole, *Pornhub Is Banning AI-Generated Fake Porn Videos, Says They're Nonconsensual*, 2018. URL: `https://www.vice.com/en/article/zmwvdw/pornhub-bans-deepfakes`.

[7] B. Chesney och D. Citron, ”Deep fakes: A looming challenge for privacy, democracy, and national security,” *Calif. L. Rev.*, årg. 107, s. 1753, 2019.

[8] H. Ajder, G. Patrini, F. Cavalli och L. Cullen, *The State of Deepfakes: Landscape, Threats and Impact*, 2019. URL: `https://enough.org/objects/Deeptrace-the-State-of-Deepfakes-2019.pdf`.

[9] H. Abdi, ”A neural network primer,” *Journal of Biological Systems*, årg. 2, nr 03, s. 247–281, 1994.

[10] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim och J. Choo, ”Stargan: Unified generative adversarial networks for multi-domain image-to-image translation,” i *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, s. 8789–8797.

[11] T. Karras, S. Laine och T. Aila, ”A style-based generator architecture for generative adversarial networks,” i *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, s. 4401–4410.

[12] L. Guarnera, O. Giudice och S. Battiato, ”DeepFake Detection by Analyzing Convolutional Traces,” i *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

[13] Meta, *Meta Deepfake Detection Challenge*, 2019. URL: `https://ai.facebook.com/datasets/dfdc/`.

[14] T. Karras, T. Aila, S. Laine och J. Lehtinen, ”Progressive growing of gans for improved quality, stability, and variation,” *arXiv preprint arXiv:1710.10196*, 2017.

[15] M. C. Stamm, M. Wu och K. R. Liu, ”Information forensics: An overview of the first decade,” *IEEE access*, årg. 1, s. 167–200, 2013.

[16] B. Zi, M. Chang, J. Chen, X. Ma och Y.-G. Jiang, ”WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection,” i New York, NY, USA: Association for Computing Machinery, 2020, ISBN: 9781450379885. URL: `https://doi.org/10.1145/3394171.3413769`.