USTVARJANJE PONAREJENIH VIDEOPOSNETKOV S POMOČJO DIFUZIJSKIH MODELOV

BINE MARKELJ, PETER PEER, BORUT BATAGELJ

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Ljubljana, Slovenija
bm9928@student.uni-lj.si, peter.peer@fri.uni-lj.si, borut.batagelj@fri.uni-lj.si

V članku predstavimo postopke in tehnike generiranja globoko ponarejenih videoposnetkov ali krajše globokih ponaredkov (angl. deepfakes). To so videoposnetki, pri katerih je prišlo do manipulacij s tehnikami globokega učenja. Taki videoposnetki predstavljajo velik problem pri širjenju lažnih novic, politični propagandi, uničevanju podobe posameznikov, izdelavi pornografskih vsebin, izsiljevanju itd. V članku opišemo podatkovno zbirko FaceForensics++ in predstavimo lastno metodo za potencialno izdelavo podzbirke omenjene baze z uporabo najnovejših generativnih difuzijskih modelov. Uporabljene postopke eksperimenta predstavimo in analiziramo njihovo kvaliteto in uspešnost. Komentiramo tudi smiselnost uporabe in nevarnost, ki jo predstavljajo ponarejeni videoposnetki, izdelani z difuzijskimi modeli.

DOI
https://doi.org/10.18690/um.feri.1.2024.8

ISBN
978-961-286-837-6

Ključne besede:
ponarejeni videoposnetki, globoki ponaredki, globoko učenje, nevronska mreža, difuzijski modeli, stabilna difuzija

Prispevek temelji na:
Markelj, B.(2023), *Ustvarjanje ponarejenih videoposnetkov s pomočjo difuzijskih modelov za razširitev zbirke za odkrivanje ponarejenih videoposnetkov*: diplomsko delo, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko.

Univerzitetna založba Univerze v Mariboru

DOI
https://doi.org/10.18690/um.feri.1.2024.8

ISBN
978-961-286-837-6

Keywords:
fake video,
deepfake,
deep learning,
neural network,
diffusion models,
stable diffusion

The proceedings is based on:
Markelj, B.(2023), *Ustvarjanje ponarejenih videoposnetkov s pomočjo difuzijskih modelov za razširitev zbirke za odkrivanje ponarejenih videoposnetkov*: bachelor's thesis, University of Ljubljani, Faculty of Computer Science and Informatics.

University of Maribor Press

CREATING FAKE VIDEOS USING DIFFUSION MODELS

BINE MARKELJ, PETER PEER, BORUT BATAGELJ
University of Ljubljana, Faculty of Computer Science and Informatics, Ljubljana, Slovenia
bm9928@student.uni-lj.si, peter.peer@fri.uni-lj.si, borut.batagelj@fri.uni-lj.si

In the article we present techniques and procedures for generating deepfake videos. These are videos that were subjected to manipulations with deeplearning techniques. Such videos represent a major problem in the spread of fake news, political propaganda, destruction of individual's public image, production of pornographic content, extortion, etc. In the article we describe deepfake database FaceForensics++. We also present our own method for potential creation of a subset of the mentioned database using the latest generative diffusion models. We describe multiple techniques, that we used and tested in our experiment and analyse their quality and success. We also comment on the utility and danger posed by fake videos generated by diffusion models.

1 Uvod

V zadnjih letih vse več slišimo o umetni inteligenci in njenih hitrih in impresivnih napredkih. Orodja generativne umetne inteligence, kot so veliki jezikovni modeli (GPT-3.5 GPT-4, PaLM2), postajajo vse bolj uporabljena, kvalitetna in sofisticirana. Velik napredek generativne umetne inteligence pa je bil v zadnjem času dosežen tudi na področju generiranja slikovnih medijev.

Slike in videoposnetki predstavljajo enega izmed najbolj razširjenih in zaupanja vrednih načinov širjenja informacij. Napredki v umetni inteligenci in razvoj na področju generativnih nasprotniških mrež omogočajo izdelavo zelo prepričljivih ponarejenih videoposnetkov (angl. deepfake), ki postajajo vse boljši in jih danes s prostim očesom praktično ne moremo več ločiti od avtentičnih videoposnetkov.

Danes lahko ponarejene posnetke ustvarimo z zelo naprednimi spletnimi orodji, za katera ne potrebujemo veliko tehničnega znanja. Zato vedno pogosteje prihaja do zlorab. Ocenjujejo, da je velika večina ponarejenih videoposnetkov uporabljena za izdelavo pornografskih vsebin, lažnih novic, zlorabo identitete itd. (Markelj, 2023).

Večina trenutnih ponarejenih videoposnetkov je narejenih z uporabo globokih nevronskih mrež. Največkrat sta uporabljeni 2 tehniki: samokodirniki (angl. autoencoders) in generativne nasprotniške mreže (angl. generative adversarial network, GAN).

Velik napredek na področju generativnih nevronskih mrež predstavljajo difuzijski modeli ter njihova uporaba za generiranje slikovnih medijev. Zato se bo naš članek osredotočil na njihovo uporabnost in tveganje pri ustvarjanju ponarejenih videoposnetkov (Markelj, 2023).

Zaradi vseh nevarnosti, ki jih ti posnetki predstavljajo, se nam zdi pomembno, da se metode in podatkovne zbirke za prepoznavanje in identifikacijo ponarejenih videoposnetkov ves čas razvijajo in izboljšujejo. Zato predlagamo nov postopek uporabe modelov stabilne difuzije, ki bi lahko služil za kasnejšo izdelavo podzbirke že obstoječe podatkovne zbirke ponarejenih videoposnetkov FaceForensics++ (FF++) (Rössler et al., 2022). Med raziskavo področja difuzijskih modelov smo

odkrili več različnih tehnik in načinov za pridobitev ponarejenih videoposnetkov, ki jih v nadaljevanju tudi predstavimo in analiziramo.

2 Pregled področja

2.1 Generiranje ponarejenih videoposnetkov

Večina ponarejenih videoposnetkov je ustvarjenih z uporabo generativnih nasprotniških mrež – GAN in samokodirnikov.
Tehnike generiranja ponarejenih videoposnetkov v grobem delimo v 2 glavni skupini: zamenjava obraza ter uprizoritev obraza (Yu et al., 2020).

Pri zamenjavi obraza algoritem izvede menjavo obraza med prvotno in ciljno osebo, pri čemer ohranja obrazno mimiko in izraz. To je zelo popularna tehnika zaradi različnih dostopnih spletnih orodij, ki omogočajo hiter dostop do dobrih rezultatov.

Uprizoritev obraza pa vključuje prenos obraznih mimik in pozicije obraznih točk prvotne osebe na ciljno osebo. Program zajame mimiko, globino ter osvetlitev obraza, kar omogoča prenos obraznih mimik in značilnih točk (angl. facial landmarks).

2.2 Odkrivanje ponarejenih videoposnetkov

V grobem poznamo 2 načina prepoznavanja ponarejenih videoposnetkov:

1. Klasične metode, ki temeljijo na predprocesiranju vhoda, luščenje značilk (angl. feature extraction) in klasifikaciji. Algoritmi iščejo zabrisane robove, artefakte itd.
2. Metode globokega učenja, za katere so značilni sočasno učenje, luščenje značilk in njihova klasifikacija. V ta namen se uporabljajo konvolucijske nevronske mreže oz. CNN in vizualne transformatorje (Zhang et al., 2022), ki kot vhod lahko sprejmejo vizualne medije. Za zaznavo ponarejenih posnetkov je ključnega pomena upoštevanje časovne komponente (Gu et al., 2021).

2.3 Podatkovna zbirka ponarejenih videoposnetkov FaceForensics++

V našem članku smo se osredotočili na videoposnetke iz zbirke FaceForensics++ ali krajše FF++ (Rössler et al., 2022).

To je ena najbolj citiranih in uporabljenih zbirk na področju odkrivanja ponarejenih videoposnetkov. Pogosto je uporabljena kot merilo uspešnosti algoritmov za odkrivanje obraznih manipulacij.

Sestavljena je iz 1000 originalnih posnetkov, nad katerimi so avtorji uporabili 5 različnih tehnik manipulacije: FaceSwap, Deepfakes, Face2Face, Neural Textures in FaceShifter.

Posnetki te zbirke so nam služili kot osnova za naše delo.

3 Teoretično ozadje in metodologija

Zaradi napredkov v raziskovanju tehnik globokega učenja (angl. deep learning) in razvoja arhitekure konvolucijskih nevronskih mrež – CNN, ki omogočajo obdelavo vizualnih podatkov, je področje ustvarjanja in raziskovanja globokih ponaredkov doživelo velik razcvet in popularnost.

Najbolj uporabljani tehniki generiranja ponarejenih videoposnetkov sta:

1. Variacijski samokodirniki – VAE
   Ti združujejo glavne lastnosti klasičnih samokodirnikov ter verjetnostnih modelov za generacijo novih podatkov z vnaprej naučenimi značilnostmi (npr. obraz).
2. Generativne nasprotniške mreže – GAN
   So posebna vrsta generativnih modelov. Sestavljene so iz dveh nevronskih mrež (generator in diskriminator), ki med učenjem med seboj tekmujeta in se tako hkrati izboljšujeta. Po koncu učenja generatorska mreža generira realistične slike (npr. obrazov).

Najnovejši in zato za nas najzanimivejši in aktualni generativni modeli pa so difuzijski modeli.

3.1 Difuzijski modeli

Difuzijski modeli so sposobni generiranja visoko kvalitetnih in realističnih slik iz šuma in vodenja z besedilom ali sliko (Rombach et al., 2022}. Za razliko od tehnologij, ki temeljijo na GAN, difuzijski modeli generirajo sliko v več korakih, pri čemer ohranjajo njene dimenzije skozi iteracije. Ti modeli se naučijo zadeti porazdelitev podatkov z obračanjem postopka večkoračnega procesa šumenja (angl. noise process). Iz šuma generirajo kompleksne in realistične vzorce in dosegajo primerljive oz. boljše rezultate kot modeli GAN in VAE.

Difuzijski modeli torej vsebujejo dva glavna procesa:

1. Difuzijski proces naprej (angl. forward diffusion process)
   To je proces, ki vhodni sliki v korakih dodaja Gaussov šum. Na koncu vrne sliko, ki je le še šum s standardno normalno razporeditvijo -- šum s povprečno vrednostjo 0 in enotsko varianco. Ker je proces dodajanja šuma vnaprej določen, se ga model ne uči, uporaben je za učenje obratnega difuzijskega procesa.

2. Obratni difuzijski proces (angl. reverse diffusion process)
   Difuzijski proces obrnemo. Začnemo s sliko šuma, ki jo v korakih rekonstruiramo nazaj v realistično sliko – odstraniti moramo šum iz difuzijskega procesa naprej. Postopek poteka v iteracijah in ga nadzoruje model nevronske mreže. Naloga mreže je, da v vsaki iteraciji napove šum, ki je bil sliki dodan pri difuzijskem procesu naprej v prejšnji iteraciji.

Oba difuzijska procesa sta še vizualno prikazana v(Raya, 2023).

Pogosta arhitekturna implementacija generativnih difuzijskih modelov je mreža U-net (Ronneberger et al., 2015), ki vsebuje več zaporednih konvolucijskih in združevalnih slojev ter sloje povečanja vzorčenja (angl. upsampling layer) in preskočne povezave (angl. skip connections).

Difuzijski modeli so zaradi visoke kvalitete generiranih slik v zadnjem času postali zelo priljubljeni. Za naše delo bomo uporabljali stabilno difuzijo, ki jo podrobneje opišemo v poglavju 3.3.

3.2 Modeli LoRA

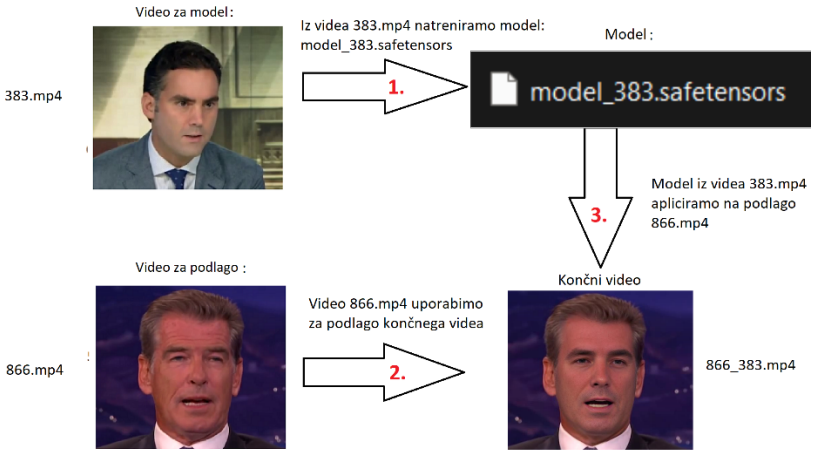
Modeli LoRA oz. prilagoditveni modeli nizkega ranga (angl. low rank adaptation model) se uporabljajo za potrebe manjših prilagoditev, uglaševanja (angl. fine tuning) in domenske prilagoditve velikih modelov. Dosegajo dobre rezultate dodajanja znanja velikim modelom z relativno malim številom učnih primerov ter kratkim časom učenja. Model lahko naučimo koncepta, stila slike, predmeta ali pa osebe.

Model LoRA plastem navzkrižne pozornosti velikega modela z obtežitvijo dodaja matrike nizkega ranga. Ideja izhaja iz matrične faktorizacije na osnovi ranga, kjer matriko predstavimo kot produkt dveh manjših matrik, kar občutno zmanjša število parametrov in s tem čas učenja.

3.2.1 Postopek učenja identitete

V članku želimo odkriti najboljši način generiranja prepričljivih ponarejenih videoposnetkov z uporabo generativnih difuzijskih modelov. Glavni problem takih videoposnetkov, predstavlja povezanost med sličicami, ki tvorijo videoposnetek.
Zato za generiranje sličic potrebujemo nekakšno podlago, ki bo služila vodenju sistema, da bodo zaporedne sličice karseda povezane. Kot podlago smo vzeli obstoječ video iz podatkovne zbirke FF++ in ga razdelili na posamezne sličice. Nato smo iz sličic drugega videa model LoRA naučili generiranja te osebe. Tako smo difuzijski proces omejili, da v celotnem posnetku ohranja identiteto iste osebe. Osnovni potek je prikazan na Sliki 2.

Model LoRA torej doseže, da difuzijski model generira sličice, na katerih je ves čas ista oseba.
Pripravo učnega seta za modele LoRA smo avtomatizirali. Razvili smo poseben program, ki iz videa avtomatsko izlušči zahtevano število sličic, ki so si med seboj čim bolj različne, in jih na koncu še dodatno obdela. Slikam v učni mapi je pred učenjem potrebno dodati istoimenske besedilne datoteke, ki opišejo, kaj je na sliki. Algoritem tako med učenjem ve, na kaj mora biti pozoren in se nauči povezav med videnim in napisanim.



Slika 2: Struktura osnovne ideje: iz sličic videa 383.mp4 s pomočjo modela LoRA naučimo identiteto, video 866.mp4 uporabimo za podlago končnega videa. Model identitete model_383.safetensors apliciramo na podlago. Dobimo video identitete modela 383 na videoposnetku 866.mp4, imenovan 866_383.mp4 (spodaj desno).
Vir: lasten.

Za LoRA učenje je pomembna tudi vnaprej predpisana struktura učne mape, množica regularizacijskih slik, število korakov učenja in optimalni parametri učenja modela, ki smo jih pridobili z daljšim eksperimentiranjem.

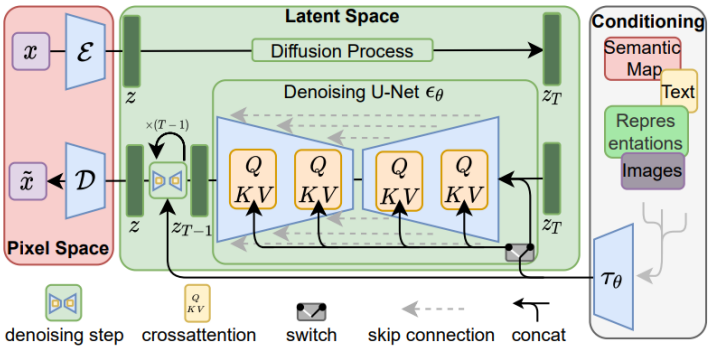
3.3 Stabilna difuzija

Model stabilne difuzije je generativni difuzijski model izdan leta 2022 (Rombach et al., 2021). Naučen je bil na več milijardah slik in na podlagi tekstovnih navodil oz. poziva (angl. prompt) in/ali slike ustvari novo realistično sliko. Stabilna difuzija ima več možnosti generiranja: besedilo v sliko (angl. txt2img), slika v sliko (angl. img2img) in slikovna vrisava (angl. image inpainting).
Arhitektura sistema stabilne difuzije ima 3 glavne komponente (Slika 3):

1. Kodirnik besedila, ki difuzijski proces dodatno usmerja z upoštevanjem besedilnega ukaza (zakodiranega v latentni prostor).
2. Nevronska mreža tipa U-net je glavni del sistema, ki na podlagi besedilnega navodila in latentnega šuma obrne difuzijski proces, tako da v vsakem koraku čim bolj natančno predvidi šum, ki ga mora

odstraniti za generacijo slike. Odstranjevanje šuma poteka v latentnem prostoru, kar zaradi manjših dimenzij znatno pohitri postopek generiranja.

3. Slikovni samokodirnik, ki je zadolžen za kodiranje in dekodiranje slike iz začetnega prostora v latentno predstavitev in nazaj.



Slika 3: Prikaz osnovne arhitekture stabilne difuzije, osrednji del predstavlja difuzijski proces z mrežo U-net, prehajanje med levim in srednjim delom predstavlja delovanje slikovnega kodirnika. Na desni strani so prikazani načini omejevanja nastale slike – besedilno navodilo, slika itd.

Vir: (Rombach, 2022).

3.3.1 Naš postopek

Stabilna difuzija nam pred generiranjem omogoča izbiro velikega števila različnih parametrov, ki vplivajo na končno generirano sliko. Parametri in njihov vpliv so opisani v Tabeli 1.

Videoposnetek generiramo kot serijo zaporednih sličic, ki jih na koncu sestavimo skupaj v zaključen videoposnetek. Največji izziv pri povezavi posebej zgeneriranih sličic v video predstavlja časovna povezanost med njimi. Slaba konsistenca med posameznimi sličicami lahko povzroči nenaravni efekt preskakovanja (angl. flickering). V ta namen smo uporabili orodja izboljšave, kot so: *EbSynth Beta*, ki šum odpravlja s stiliziranjem posnetka po vodilnih sličicah in razširitev *Abysz-LAB-Ext*, ki odstrani šum, utripanje ter normalizira barve.

Tabela 1: Glavni parametri generiranja in njihov vpliv na sliko

PARAMTER	VPLIV NA GENERIRANO SLIKO
poziv	besedilno navodilo, opis željene končne slike
negativni poziv	besedilno navodilo, stvari, ki jih ne želimo na sliki
slika	slika, po kateri želimo, naj se končna slika ravna
metoda vzorčenja	algoritem vzorčenja, uporabljen pri generaciji slike
koraki vzorčenja	število korakov, ki jih izvede izbrana metoda vzorčenja
dimenzije slike	širina in višina končne slike
število serij	koliko serij končnih slik bo zgeneriranih
velikost serije	število zgeneriranih končnih slik v eni seriji
lestvica CFG	stopnja prileganja končne slike pozivu
stopnja odpravljanja hrupa	stopnja prilagajanja končne slike vhodni
naključno seme	naključno število, ki vpliva na začetni šum

Za glavno "podlago" smo vzeli sličice obstoječega pristnega videoposnetka iz podatkovne zbirke FF++ in osebi na posnetku zamenjali obrazno identiteto z naučenim modelom LoRA. Uporaba obstoječega posnetka je nujna za generiranje realističnih videoposnetkov, saj nam le-ta služi kot nekakšen skelet in vodilo, po katerem se nov ponarejeni video ravna.

Eden najpomembnejših mehanizmov nadzora generiranja sličic je razširitev kontrolnih mrež (angl. controlnet) (Zhang et al., 2023). To so posebne nevronske mreže, ki procesu generiranja dodajo dodatne omejitve in pogoje. Uporabne so za maskiranje slik, omejitev robov, sledenje obraznim mimikam, itd. (Slika 4, levo).
Za izdelavo obraznih mask za način vrisovanja smo uporabili razširitev *batch-face-swap*, ki na posamezni sliki zazna obraz in na njegovem mestu naredi poljubno veliko masko (slika 4, desno).

Slika 4: Levo: prikaz izdelanih map kontrolnih mrež različnih predprocesorjev. Desno: prikaz obrazne maske.

Vir: lasten

4 Eksperiment

Najprej predstavimo eksperiment treniranja modelov LoRA. Nato opišemo tudi potek in proces eksperimentiranja generiranja ponarejenih videoposnetkov s stabilno difuzijo, ter analiziramo dobljene rezultate.

4.1 Priprava modelov LoRA

Za učenje uporabnega modela LoRA potrebujemo dovolj raznoliko učno množico slik, ki osebo, ki jo želimo z modelom zajeti, predstavi v več različnih pozicijah, na različnem ozadju in iz različnih zornih kotov. Ker je pri uporabi enega samega videoposnetka to lahko velika omejitev, smo v ta namen razvili program, ki avtomatsko iz videoposnetka pobere ključne sličice in uvede umetno raznolikost.

Ugotovili smo, da je optimalno število sličic za učenje identitete 25. Program najprej izbere vse ključne okvirje s pomočjo orodja *ffmpeg*. Nato naključno izbere preostale sličice, kriterij izbire je, da zajame optimalno razmerje raznolikih sličic. To doseže z obrazno analizo, kjer na vsaki sličici poišče vse glavne obrazne točke s pomočjo knjižnice *mediapipe* in klasificira ali ima oseba zaprte/odprte oči in usta. Ustrezno razmerje tako različnih slik nam kasneje omogoča uprizoritev govorjenja in raznolikih izrazov obraza. Na koncu naključnim fotografijam dodamo še odstranimo ozadje s pomočjo modela *rembg* in jih še dodatno obrežemo. S tem še povečamo raznolikost in dinamičnost učne množice.

Med procesom učenja smo vsak model unikatno poimenovali in dodali večjo regularizacijsko množico slik (okoli 200). Datoteke opisov slik smo zgenerirali z globokim modelom BLIP. Optimalne rezultate smo dobili z 9 ponovitvami učne in 1 ponovitvijo regularizacijske množice, 10 epohami učenja ter serijo velikosti 1. Celotno optimalno učenje modela tako poteka v okoli 2200 korakih.

4.2 Generiranje lažnih videoposnetkov

Trenutno izdelava prepričljivih videoposnetkov z difuzijo zaradi same novosti tehnologije še vedno predstavlja izziv. Hkrati naši ciljni posnetki vedno prikazujejo govorjenje, ki je za replikacijo z difuzijskimi modeli zelo zapleteno. Prav zato smo k delu pristopili zelo eksperimentalno in s pomočjo spletnih virov in različnih

razširitev in programov preizkusili in ustvarili čim več različnih pristopov. Najprej le na kratko opišemo prvotne načine, ki pa so nas končno pripeljali do našega najuspešnejšega načina, ki ga tudi podrobneje analiziramo in komentiramo.

4.2.2 Predhodni eksperimenti

Sprva smo želeli z načinom img2img generirati celotne sličice, ne le obrazni del. Za ohranjanje premikov in poze oseb smo uporabili več kombinacij kontrolnih mrež. Rezultat smo nato dodatno izboljšali z posebno skripto, ki uporabi kot dodatno vodilo generiranja še prejšnje slike. Kasneje smo preizkusili še uporabo orodja za stiliziranja po vodilnih sličical – *EbSynth Beta.* Za izboljšanje enakosti med sličicami smo sličice generirali v obliki mreže 25 sličic, ki so bile tako bolj enotne. Ker omenjeni eksperimenti niso dali zadovoljivih rezultatov, smo se osredotočili na način generiranja z vrisovanjem, kjer menjamo le obrazno identiteto osebe. Ta način smo združili z orodjem *EbSynth Beta.* Generirali smo vsako deseto sličico videoposnetka in jih v orodju združili s še dodatnim zamikanjem sličic in navzkrižnim bledenjem (angl. crossfade). Rezultati omenjenih tehnik, še niso bili dovolj prepričljivi (zabrisani robovi, nenaravno premikanje ust, degradacija barv, itd.), služili pa so nam, kot podlaga za naslednji postopek.

4.2.2 Najuspešnejši postopek generiranja in analiza rezultatov

Najprej za vsako sličico ciljnega videoposnetka z razširitvijo *batch-face-swap* pridobimo obrazno masko ustrezne velikosti. Nato z vrisovanjem poiščemo optimalne nastavitve generiranja sličic in kombinacije kontrolnih mrež. Izbira pravih kontrolnih mrež in njihovih obtežitev predstavlja enega izmed ključnih korakov.

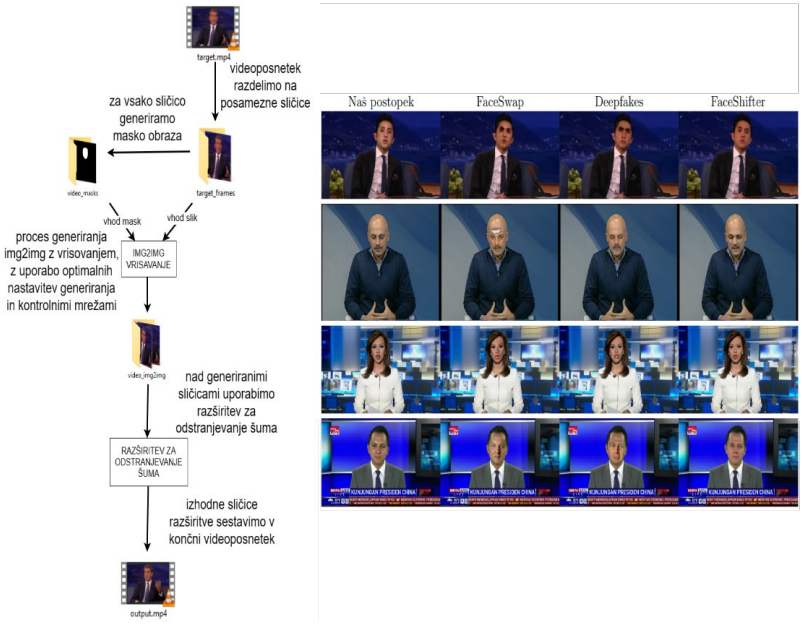
Med najpomembnejše kontrolne mreže, ki dajejo najboljše rezultate, spadajo:

- Kontrolni mreži *OpenPose* in *MediaPipeFace*, ki skrbita za zajem ključnih obraznih točk in mimik med govorjenjem.
- Kontrolna mreža *TemporalNet*, ki je ključnega pomena za doseganje časovne konsistence. To dosega z uporabo konteksta vseh prej zgeneriranih sličic.
- Kontrolna mreža *Tile*, ki pomaga pri popravljanju barv – izenačitev kožnega odtenka novo generiranega obraza in ciljnega obraza.
- Kontrolna mreža *Inpaint*, ki pri vrisovanju poskrbi za glajenje robov maske

Proces iskanja optimalnih parametrov je lahko zaradi raznolikih videoposnetkov lahko zelo dolgotrajen in nekoliko drugačen ob vsakem generiranju.
Enake parametre generiranja in kombinacije kontrolnih mrež nato uporabimo nad vsemi sličicami ciljnega posnetka ter tako na vsako nanesemo novo identiteto z modelom LoRA.

Končno posamezne sličice združimo v ponarejeni videoposnetek. Nato z uporabo razširitve *Abysz-LAB-Ext* še dodatno odstranimo zadnje sledove šuma, utripanja ter izvedemo še končno barvno normalizacijo. Celoten postopek prikažemo na Sliki 5.

Ponarejeni videoposnetek ustvarjen z zgoraj opisanim postopkom ima visoko časovno konsistenco, minimalno prisotnostjo utripanja in visoko kvaliteto. Smo mnenja, da proces predstavlja mejnik na področju generiranja ponarejenih posnetkov z difuzijskimi modeli in daje rezultate, ki so enakovredni oz. celo presegajo obstoječe tehnike ponarejanja videoposnetkov v bazi FF++. Vizualno primerjavo prikažemo na Sliki 5.



Slika 5: Levo: prikaz celotnega poteka generiranja ponarejenega videoposnetka. Desno: vizualna primerjava generiranih istoležnih sličic z našim postopkom in tehnikami v zbirki FF++.

Vir: lasten

5 Zaključek in nadaljnje delo

V članku smo pokazali več načinov generiranja prepričljivih ponarejenih videoposnetkov z difuzijo ter osvetlili teoretično ozadje. Vendar pa problematika generacije ponarejenih videoposnetkov z difuzijskimi modeli še zdaleč ni dokončno raziskana. Pokazali smo, da naša metoda daje rezultate, ki po kakovosti presegajo trenutne tehnike manipulacij. Predvidevamo pa, da bo zaradi hitrega razvoja področja prišlo do vedno boljših načinov manipulacij z difuzijo in bodo taki posnetki v bližnji prihodnosti predstavljali veliko nevarnost.

Prav zaradi hitrega napredovanja v razvoju tehnologije se nam zdi pomembno, da se čim več raziskav posveti področju odkrivanja ponarejenih videoposnetkov nastalih z novimi tehnikami difuzijskih modelov. Izdelava dobre podatkovne zbirke takih manipulacij nam lahko pomaga, da ostanemo v koraku s časom z napadalci, ki bi difuzijsko tehnologijo lahko uporabljali v zle namene. Pomemben naslednji korak bi bila avtomatizacija procesa generiranja posnetkov in izdelava celotne zbirke takih posnetkov, ki bi služili za učenje algoritmov za odkrivanje manipulacij z difuzijo.

Viri in literatura

MARKELJ, BINE, 2023, Ustvarjanje ponarejenih videoposnetkov s pomočjo difuzijskih modelov za razširitev zbirke za odkrivanje ponarejenih videoposnetkov [na spletu]. Diplomsko delo. Pridobljeno s: https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=slv&id=149326

Andreas Rössler in sod. "FaceForensics++: Learning to Detect Manipulated Facial Images". V: International Conference on Computer Vision (ICCV). 2019.

Peipeng Yu in sod. "A Survey on Deepfake Video Detection". V: IET Biometrics 10.6 (2021), str. 607–624. doi: https://doi. org / 10 . 1049 / bme2 . 12031. eprint: https : / / ietresearch . onlinelibrary . wiley . com / doi / pdf / 10 . 1049 / bme2 . 12031. url: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/bme2.12031.

Daichi Zhang in sod. Deepfake Video Detection with Spatiotemporal Dropout Transformer. 2022. arXiv: 2207.06612 [cs.CV].

Zhihao Gu in sod. Spatiotemporal Inconsistency Learning for DeepFake Video Detection. 2021. arXiv: 2109.01860 [cs.CV].

Robin Rombach in sod. High-Resolution Image Synthesis with Latent Diffusion Models. 2021. arXiv: 2112.10752 [cs.CV].

Robin Rombach in sod. High-Resolution Image Synthesis with Latent Diffusion Models. 2022. arXiv: 2112.10752 [cs.CV].

Gabriel Raya in Luca Ambrogioni. Diffusion Models Seminar. url: https://diffusionmodels.nl/ (pridobljeno 25. 8. 2023).

Olaf Ronneberger, Philipp Fischer in Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015. arXiv: 1505.04597 [cs.CV].

Lvmin Zhang. Controlnet - Official implementation of Adding Conditional Control to Text-to-Image Diffusion Models. url: https://github.com/lllyasviel/ControlNet (pridobljeno 30. 8. 2023).