Techniky audiovizuální manipulace

Tomáš Chytil

Bakalářská práce
2023

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky
Ústav počítačových a komunikačních systémů

Akademický rok: 2022/2023

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Tomáš Chytil**
Osobní číslo:	**A19168**
Studijní program:	**B3902 Inženýrská informatika**
Studijní obor:	**Informační technologie v administrativě**
Forma studia:	**Prezenční**
Téma práce:	**Techniky audiovizuální manipulace**
Téma práce anglicky:	**Techniques of Audiovisual Manipulation**

Zásady pro vypracování

1. Vypracujte literární rešerši na dané téma.
2. Charakterizujte nejčastější způsoby audiovizuální manipulace a rámcově popište principy metod, které se k jejich vytváření využívají.
3. Stručně představte několik volně dostupných softwarových nástrojů, použitelných pro dané účely, přičemž jeden vybraný, dále využitý v praktické části práce, popište podrobněji.
4. Připravte vhodné vzorky foto, video či audio materiálu k demonstraci tvorby Deepfake.
5. Ve zvoleném nástroji na jednodušší videosekvenci demonstrujte proces tvorby videa se záměnou obrazové či hlasové identity.
6. Naznačte, na co se zaměřit při posuzování pravosti záznamu a vypracujte 'manuál obrany před audiovizuální manipulací' využitelný např. v pedagogickém procesu.

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam doporučené literatury:

1. ŠMÍD, K. Syntetické médium deepfake: ztrácející se hranice mezi realitou a fikcí. Praha, 2021. Bakalářská práce. Filozofická fakulta, Univerzita Karlova.
2. HAVELCOVÁ, T. Fake News a nástroje jejich odhalování. České Budějovice, 2020. Bakalářská práce. Pedagogická fakulta, Jihočeská univerzita v Českých Budějovicích.
3. MUDROVÁ, N. Fenomén deepfakes jako hrozba žurnalistiky. Praha, 2020. Bakalářská práce. Fakulta sociálních věd, Univerzita Karlova.
4. WESTERLUND, M. The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review [online]. 2019, 9(11), 39-52 [cit. 2022-11-22]. ISSN 19270321. Dostupné z: doi:10.22215/timreview/1282
5. MATOUŠEK, J. Hluboké učení pro každého? Neuronové sítě a spol. In: Data Mind [online]. 2018 [cit. 2022-11-22]. Dostupné z: http://www.datamind.cz/cz/blog/hluboke-uceni-neuronove-site
6. KŘIVAN, M. Úvod do umělých neuronových sítí. Praha: Oeconomica, 2014. ISBN 978-80-245-2024-7.
7. Twelve Best Deepfake Apps and Websites You Can Try for Fun. In: Beebom [online]. 2022 [cit. 2022-11-22]. Dostupné z: https://beebom.com/best-deepfake-apps-websites

Vedoucí bakalářské práce: **doc. Ing. František Gazdoš, Ph.D.**
Ústav řízení procesů

Datum zadání bakalářské práce: **2. prosince 2022**
Termín odevzdání bakalářské práce: **24. května 2023**

doc. Ing. Jiří Vojtěšek, Ph.D. v.r.
děkan

UNIVERZITA TOMÁŠE BATI VE ZLÍNĚ
Fakulta aplikované informatiky
L.Š.

doc. Ing. Petr Šilhavý, Ph.D. v.r.
garant oboru

Ve Zlíně dne 8. prosince 2022

Prohlašuji, že

- beru na vědomí, že odevzdáním bakalářské práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb. o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších právních předpisů, bez ohledu na výsledek obhajoby;
- beru na vědomí, že bakalářská práce bude uložena v elektronické podobě v univerzitním informačním systému dostupná k prezenčnímu nahlédnutí, že jeden výtisk bakalářské práce bude uložen v příruční knihovně Fakulty aplikované informatiky Univerzity Tomáše Bati ve Zlíně;
- byl/a jsem seznámen/a s tím, že na moji bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, zejm. § 35 odst. 3;
- beru na vědomí, že podle § 60 odst. 1 autorského zákona má UTB ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
- beru na vědomí, že podle § 60 odst. 2 a 3 autorského zákona mohu užít své dílo – bakalářskou práci nebo poskytnout licenci k jejímu využití jen připouští-li tak licenční smlouva uzavřená mezi mnou a Univerzitou Tomáše Bati ve Zlíně s tím, že vyrovnání případného přiměřeného příspěvku na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše) bude rovněž předmětem této licenční smlouvy;
- beru na vědomí, že pokud bylo k vypracování bakalářské práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tedy pouze k nekomerčnímu využití), nelze výsledky bakalářské práce využít ke komerčním účelům;
- beru na vědomí, že pokud je výstupem bakalářské práce jakýkoliv softwarový produkt, považují se za součást práce rovněž i zdrojové kódy, popř. soubory, ze kterých se projekt skládá. Neodevzdání této součásti může být důvodem k neobhájení práce.

Prohlašuji,

- že jsem na bakalářské práci pracoval samostatně a použitou literaturu jsem citoval. V případě publikace výsledků budu uveden jako spoluautor.
- že odevzdaná verze bakalářské práce a verze elektronická nahraná do IS/STAG jsou totožné.

Ve Zlíně, dne 24.5. 2023

Tomáš Chytil v r.
podpis studenta

ABSTRAKT

Bakalářská práce se obecně zabývá audiovizuální manipulací. První část je věnována literární rešerši, zejména článkům a pracím na podobné téma a doplňujícím informacím. Dále vysvětluje principy audiovizuální manipulace na konkrétních příkladech a zmiňuje nejčastěji používané softwarové nástroje. Ve druhé části práce je podrobně popsán proces tvorby deepfake videa obohacen o autorovy zkušenosti s doporučením, jak postupovat při odhalování audiovizuální manipulace.

Klíčová slova: audiovizuální manipulace, deepfake, softwarové nástroje, ověřování pravosti

ABSTRACT

The bachelor thesis is generally focused on audiovisual manipulation. The first part is devoted to a literature review, especially articles and theses on a similar topic together with additional information. Then it explains the principles of audiovisual manipulation in selected examples and mentions the most commonly used software tools. In the second part of the thesis, the process of creating deepfake videos is described in more detail, enriched by the author's experiences and recommendations on how to proceed in detecting audiovisual manipulation.

Keywords: audiovisual manipulation, deepfake, software tools, authencity verification

Tímto bych chtěl poděkovat vedoucímu mé bakalářské práce doc. Ing. Františku Gazdošovi, Ph.D. za ochotu, vstřícný přístup a cenné rady během psaní bakalářské práce. Dále bych chtěl vyjádřit poděkování mé rodině za jejich nepřetržitou podporu nejen při psaní práce, ale po celou dobu mého studia.

Prohlašuji, že odevzdaná verze bakalářské práce a verze elektronická nahraná do IS/STAG jsou totožné.

OBSAH

ÚVOD

Nejrůznější formy fake news jsou bezpochyby velkým fenoménem poslední doby a s tím, jak rychle postupují kupředu možnosti v jiných odvětvích informačních technologiích dochází zároveň k rozšiřování a postupnému zdokonalování i v oblasti audiovizuální manipulace. Přestože samotný princip manipulace či falešných zpráv není v historii ničím novým, tak s možnostmi, které informační technologie zejména v posledních letech nabízejí, se šíření a odhalování falešných zpráv stává stále obtížnějším. Platí to zejména u fenoménu deepfake a u fotografií vygenerovaných umělou inteligencí, kde se pomalu ale jistě blížíme k bodu, (či jsme jej již překročili) kdy běžný uživatel internetu již nebude schopný určit, zda vidí opravdového člověka či obličej vytvořený AI.

Nutno ovšem dodat, že rozmachu fake news často napomáhá i samotný lidský faktor, kdy mnoho lidí bezmyšlenkovitě věří první zprávě, kterou na internetu uvidí. A zde vidím obrovský problém, který často pramení z nedostatečné informovanosti a neznalosti této problematiky. Přitom k odhalení fake news často stačí jen několik kliknutí. Existují samozřejmě propracované a sofistikované formy fake news, většina je ale stále relativně snadno odhalitelná. Hlavní problém proto nespatřuji v rozmachu a přesvědčivosti těchto dezinformací, ale spíše v neznalosti možností a nástrojů, jak s nimi bojovat. Dezinformátoři často právě na toto spoléhají. Většina lidí je jednoduše příliš neznalá nebo se jim zkrátka nechce zprávu video či fotografii, jakkoliv ověřovat. Dezinformace je tak sdílena po sociálních sítích, kde si následně žije „vlastním životem“.

Za nejnebezpečnější v současnosti považuji zneužití deepfakes pro politické účely s cílem zmanipulovat co nejvíce lidí a ovlivnit veřejné mínění. Stalo se tak např. v případě ex-prezidentů Trumpa, Obamy a v i případě prezidenta Zelenského v počátcích ruské invaze na Ukrajinu. Právě poslední zmíněné mohlo mít na následující události zásadní vliv, pokud by se stihlo dostat k většímu počtu lidí.

Cílem mé práce je především upozornit na nebezpečí audiovizuální manipulace, popsat a vysvětlit její techniky a principy a zamyslet se nad možnou budoucností těchto typů fake news. První kapitola mé práce je věnována rešerši tématu, kde uvádím několik zajímavých webů a článků zabývajících se tímto tématem. Následně se již věnuji samotné problematice audiovizuální manipulace, a to od fenoménu deepfake přes fotomontáže až po fotografické a hlasové generátory. Taktéž objasňuji, jakým způsobem falešný obsah vzniká a upozorňuji na několik konkrétních případů, kdy byly tyto technologie zneužity k manipulaci. Teoretická

část dále pokračuje výčtem nejpoužívanějších nástrojů a softwaru k tvorbě potenciálně zneužitelného audiovizuálního obsahu.

V praktické části se věnuji tvorbě ukázkového deepfake videa v populárním programu DeepFaceLab. Podrobně popisuji proces tvorby od prvního kroku, čímž je výběr vhodných podkladů až po finální úpravy a export videa. Výsledné video slouží jako praktická ukázka této technologie, na které předvádím jak jednoduché či naopak složité může vytvoření deepfake videa být. Na tomto i dalších příkladech pak upozorňuji na jaké detaily je třeba se zaměřit při posuzování pravosti konkrétního záznamu.

Přílohou práce je taktéž stručný manuál na obranu před audiovizuální manipulací, který může posloužit jako doplňkový výukový materiál např. pro použití na základních a středních školách.

I. TEORETICKÁ ČÁST

1 REŠERŠE TÉMATU

V současné době lze na internetu dohledat poměrně velké množství článků zabývajících se problematikou fake news, deepfake či falešných fotografií. Například na webové stránce tomaskubica.cz [1] zmiňuje autor ve svém článku web [2], kde je na jedno kliknutí možné vygenerovat velmi realistickou fotografii člověka, který ale ve skutečnosti vůbec neexistuje. Autor dále zmiňuje několik způsobů, jak takovouto vygenerovanou fotografii poznat. Můžeme se zaměřit například na detaily brýlí, náušnic, či oblečení – jakákoliv nesrovnalost, podivný tvar nebo chybějící součást může napovědět, jestli se jedná o falešnou fotografii. Podobné je to i s texty, pokud se na takovéto fotografii vůbec objeví, bývají zpravidla nesmyslné. Na webu se dále dočteme o souboji dvou AI, kdy úkolem jedné je vytvořit co nejvěrnější fotografii a úkolem druhé poté naučit se rozpoznávat mezi opravdovými fotografiemi a podvrhy z generátorů. [1]

Podobně zaměřený článek nabízí taktéž např. web medium.com. [3] Autor Kyle McDonald se zde podrobně věnuje analýze obličejů vytvořených generátory a upozorňuje na nejčastější nesrovnalosti, které mohou napovědět, zda se jedná o podvrh. Článek je doplněn množstvím obrázků, zobrazujících detaily, kterých je potřeba si všímat. [3]

Web unite.ai [4] přináší výčet nejlepších AI generátorů tváří pro rok 2023. Na první pozici se umístila placená aplikace Generated photos, využívána i platformami jako BBC, Forbes, nebo Daily Mail. Její síla spočívá především v širokých možnostech přizpůsobení výsledného snímku. Na druhém a třetím místě se umístily aplikace BoredHumans a ThisPersonDoesnNotExist. Obě fungují zdarma, na rozdíl od GeneratedPhotos ale neumožňují pokročilejší přizpůsobení. Na vytvoření náhodné realisticky vypadající lidské tváře však bohatě stačí. [4]

Na to že jsou generované obličeje potenciálním nebezpečím pro uživatele sociálních sítí upozorňuje výroční zpráva společnosti Meta za rok 2022. Na webu společnosti [5] se můžeme dočíst o rapidním nárůstu falešných profilů používajících fotografie z generátorů. Meta uvádí, že za rok 2022 bylo u více než dvou třetin zrušených falešných profilů zaznamenáno použití vygenerovaných obličejů. Podobně je tomu i u sociální sítě LinkedIn a zřejmě u mnoha dalších. [5]

O fenoménu Deepfake píše Jiří Zahrádka pro magazín Finmag na webu finmag.penize.cz. [6] Podle něj jsou Deepfake videa zatím stále poměrně nedokonalé a jejich pravý čas má ještě teprve přijít. Uvádí zde ale několik deepfake aplikací s různými funkcemi např.

Wombo nebo Reface. Obě aplikace jsou zdarma ke stažení na operační systémy Android a iOS. Z aplikací schopných měnit hlas pak uvádí např. Zao nebo Ventriloquism. Aplikacím schopným vytvářet deepfake videa se podrobněji věnuji v pozdější části práce – kapitole 3.1. [6]

Deepfake jako problém z hlediska žurnalistiky ve své bakalářské práci [7] popisuje Nikol Mudrová. Autorka se zde zaměřuje mimo jiné i na psychologickou stránku věci, tedy vysvětlení, proč lidský mozek na první pohled fake news neodhalí a jaké „zkratky" v našem zpracovávání informací dezinformátoři využívají. Praktická část pak obsahuje rozhovor z roku 2020 s novinářem Jakubem Zelenkou. [7]

Na webové stránce the-decoder.com [8] můžeme dále nalézt článek popisující historii deepfakes od prvopočátků v roce 2014 až po současnost. Zaměřuje se zejména na technologii Generativní soupeřících sítí (GAN), která je pro tvorbu současných deepfakes a generování obličejů zásadní. Článek je doplněn o několik tematických obrázků pro lepší pochopení této problematiky. [8]

Společnost Iproov provedla v roce 2019 a 2022 několik výzkumů ohledně povědomí společnosti o problematice deepfake. Respondentů se v otázkách například ptala, zda vědí, co to deepfake je, jestli by dokázali deepfake odhalit či jaké nebezpečí podle nich může představovat. Výzkumy probíhaly v osmi zemích světa a odpovídalo celkem 16 000 respondentů. Výzkumy přinesly několik zajímavých výsledků, především ukazují že deepfake je ve společnosti stále ještě poměrně neznámým pojmem a že zdaleka ne všichni se s ním již setkali. Na druhé straně více než polovina dotazovaných uvedla, že by případný deepfake dokázali odhalit a stejně tak uvedli možné nebezpečí, které podle nich může představovat. Společnost také v článku nabízí srovnání mezi výzkumy z let 2019 a 2022. V pozdější části práce se podrobněji věnuji konkrétním výsledkům těchto výzkumů viz kapitola 2.1.1. [9]

Softwarovým nástrojům sloužícím k odhalování falešných fotografií se pak ve své bakalářské práci [10] podrobněji věnuje Mariana Borisová. Autorka zde zmiňuje několik obrázkových vyhledávačů jako je Bing Image Match, Yandex image nebo TinEye. Tyto vyhledávače jsou schopné často dohledat na internetu původní zdroj fotografie a v mnohých situacích tak mohou pomoct určit kontext či původ fotografie. Lidé na internetu totiž běžně přidávají fotografie, které ve skutečnosti nepředstavují to, za co jsou vydávány, a právě díky podobným možnostem můžeme odhalit manipulaci během několika kliknutí. V případě že nám zpětné vyhledávání nepomůže, představuje autorka několik softwarových nástrojů pro odhalování

digitální manipulace. Uvádí například FotoForensics, Forensically nebo JPEGSnoop. Manipulací s fotografiemi se ve své práci taktéž stručně věnuji – viz kapitola 2.3. [10]

Za další fenomém na poli fake news lze označit taktéž generátory převádějící text na řeč. (Text to speech voice generators). Tyto generátory mají samozřejmě mnoho pozitiv, uveďme např. pomoc s učením pro lidi s dyslexií nebo pomůcku pro studium jazyků. Na druhé straně i zde došlo ke značnému posunu a generovaný „lidský" hlas může působit velmi věrohodně. Proto ve své práci také představuji několik volně dostupných generátorů a uvádím jejich principy. Webové stránky unite.ai [11] a topten.ai [12] představují několik populárních generátorů, kterým se dále podrobněji věnuji.

A samozřejmě na serveru Youtube lze nalézt několik zajímavých kanálů a videí zabývajících se problematikou Deepfake či fake news obecně. Zde bych uvedl například kanál *Jirka vysvětluje věci* [13], na kterém lze nalézt několik videí na toto téma či video od autora Patrika Kořenáře [14] popisující deepfake jako hrozbu budoucnosti.

1.1 Přínos předkládané práce

Ve své práci se zabývám širokou problematikou manipulace pomocí audiovizuálních nástrojů a přináším do ní vlastní zkušenosti a poznatky. Protože se jedná o téma, které se neustále mění a vyvíjí, je důležité jej průběžně aktualizovat a upozorňovat na nové hrozby. Tímto způsobem můžeme zvýšit povědomí a informovanost o této problematice a předcházet možným následkům jako jsou šíření dezinformací, manipulace s veřejným míněním nebo poškozování reputace osob.

V případě fotografií částečně navazuji na již zpracovanou závěrečnou práci na podobné téma [10] kterou doplňuji kapitolou o metadatech a upozorňuji na důležitost kontextu na konkrétním případu. V případě manipulace s fotografiemi je třeba zmínit taktéž generátory obličejů, které se v poslední době stávají zvláště u falešných profilů normou. Ještě větší nebezpečí představují tzv. text-to-speech generátory (TTS), které jsou jednodušší než deepfake videa a v jejich popularita v posledních letech nebo dokonce měsících taktéž výrazně narůstá. Proto v této práci upozorňuji na několik konkrétních případů, kdy byly technologie jako je deepfake či deepfake voice zneužity k manipulaci, objasňuji použité technologie, uvádím možné nástroje a představuji možnosti obrany. Právě u těchto technologií je vidět obrovská rychlost s jakou se umělá inteligence zdokonaluje a nové výzvy kterým je potřeba aktuálně čelit. V některých případech postupuje AI kupředu takovou rychlostí, jako např. v případě fenoménu

posledních několika měsíců – chatbota ChatGPT a podobných, že to znepokojuje i samotné vědce. Konkrétně ChatGPT založený na modelu GPT-4 (březen 2023) dokáže generovat složitější texty a třeba složitou zkoušku pro právníky zvládne lépe, než 90 % lidských řešitelů [15]. Toto všechno následně vedlo k sepsání otevřeného dopisu s žádostí o pozastavení vývoje umělé inteligence na nějaký čas, pod který se podepsala řada známých expertů, akademiků a podnikatelů.

Za relativní novinku, která se v poslední době značně zdokonaluje můžeme označit generátory fotorealistických obrázků, které dokážou z textového zadání vytvořit velmi přesvědčivě vypadající „fotografii“ konkrétní populární osobnosti provádějící libovolnou činnost. V případě, že jsou tyto fotografie zneužity k manipulaci, mohou představovat značné nebezpečí, protože od skutečných jsou na první pohled často téměř k nerozeznání. Nicméně existuje několik způsobů, jak lze takový podvod odhalit, a proto jsem se na ně v této práci taktéž zaměřil a demonstroval je na několika příkladech.

Do přehledu aktuálně nejpopulárnějších a nejznámějších nástrojů a aplikací k tvorbě potenciálně zneužitelného audiovizuálního obsahu jsem zahrnul zejména ty, které jsou dostupné zdarma nebo se jednoduše používají. Stejně postupuji i v případě možné obrany. Obecně zatím platí, že většinu falešného obsahu lze odhalit relativně snadno, pokud zapojíme své kritické myšlení a zjistíme si kontext.

V praktické části dále představuji jeden z aktuálně populárních programů pro tvorbu deepfake videí DeepFaceLab a porovnávám několik výsledků na kterých lze demonstrovat potenciál deepfake a vysvětlit některé principy a úskalí tvorby. Dále prakticky demonstruji, jak „složité“ je vytvořit takové video a diskutuji, zdali by dokázalo někoho oklamat.

2 AUDIOVIZUÁLNÍ MANIPULACE A JEJÍ TECHNIKY

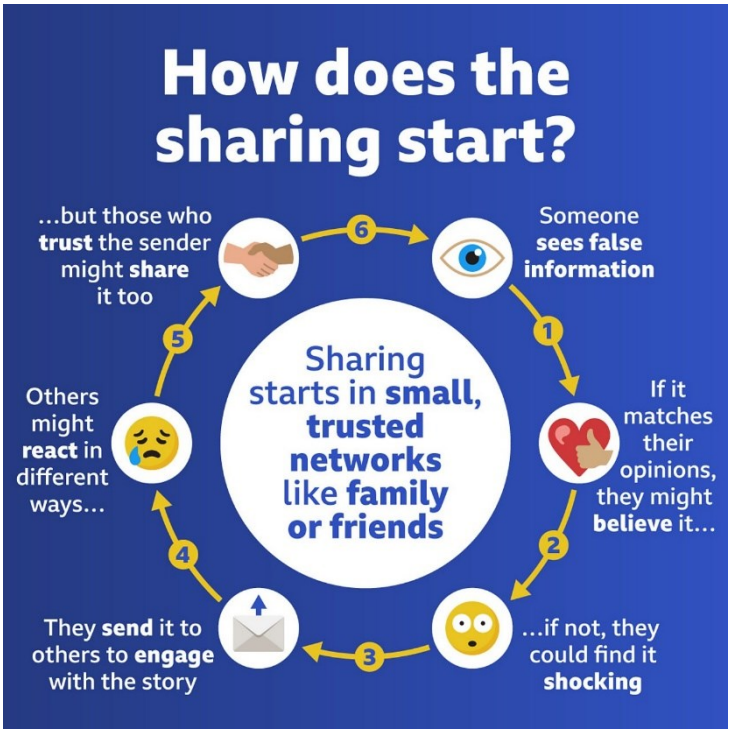
Fejky, hoaxy, dezinformace – pojmy, se kterými se v současnosti setkáváme v prostředí internetu stále častěji. Nejde přitom v historii o nic nového, fake news tu s námi byly odjakživa a jejich účel taktéž zůstal stejný. Mění se pouze způsoby a techniky, které manipulátoři a dezinformátoři v současnosti využívají a je nutno dodat že pokrok v oblasti informačních technologií jim práci často ulehčuje. Mění se taktéž dosah a nebezpečnost falešných zpráv a nezáleží přitom, zda se jedná o falešné video, fotografii či řetězovou emailovou zprávu, dezinformace mají často jeden jediný cíl. Cílem dezinformací totiž není informovat, nýbrž přesvědčovat o „své pravdě“ nebo přimět pochybovat, zda lze pravdu zjistit. Jak se tedy nenechat na internetu „napálit?“ V první řadě je vhodné ujasnit si několik základních pojmů.

Pojmem „hoax“ označujeme nepravdivou zprávu, šířenou za účelem vyvolání paniky. Hoax se snaží vystrašit příjemce a donutit jej k unáhlené či iracionální reakci. Obsahem hoaxu bývají typicky různá varování před neexistujícími hrozbami, nepravdivé informace o aktuálních sociálně-politických problémech či nejrůznější druhy žertíků. [16]

Termínem fake news se označují lživé a nepravdivé zprávy tvářící se jako pravdivé. Občas se tímto pojmem označuje i samotná žurnalistika založená na šíření nepravdivých informací. (zejména na sociálních sítích). Fake news a hoax jsou podobné pojmy a rozdíly mezi nimi se občas stírají [16]

Manipulaci definujeme jako vědomé i nevědomé chování jedince, jímž poškozuje druhého člověka. Slovo pochází z latinského “manus” - ruka, uchopit, a člověk snažící se tímto jednáním získat pro sebe výhody je označován jako manipulátor. Manipulátor se snaží ovlivnit myšlenky, chování a názory ostatních lidí ve svůj prospěch. Tyto vnucené myšlenky, názory a chování jsou přitom pro oběť cizí a můžou představovat nevýhodu či dokonce újmu. Manipulovaná osoba navíc často neví, že je s ní manipulováno, případně neví, jak se manipulaci účinně bránit. [17]

Takto funguje klasická manipulace známá z běžného života, v prostředí internetu tomu ale často bývá jinak. Především tu nemáme úzký vztah mezi manipulátorem a obětí. Produkovat na internetu obsah může dnes v podstatě kdokoliv a kdekoliv, a to i bez větších znalostí a schopností. Stejně tak příjemce není obvykle jedna konkrétní osoba, ale velké množství lidí a skrze nejrůznější způsoby přeposílání a sdílení se může stát manipulátorem nevědomky téměř každý. Životní cyklus dezinformace na internetu představuje následující obrázek.



Obrázek 1 : Životní cyklus dezinformace [18]

V případě audiovizuální manipulace zde potom vidím tři hlavní nástroje používané k těmto účelům. Jedná se především o fenomén deepfake, generátory falešných lidských tváří a taktéž hlasové generátory a převodníky textu na hlas (text-to-speech).

2.1 Deepfake

Prvním významným manipulačním nástrojem, který bych chtěl blíže představit je fenomén deepfake. Videa se známými osobnostmi, které dělají či říkají věci, které by ve skutečnosti dost možná nikdy neřekli se v poslední době šíří na internetu závratnou rychlostí. Rozčilující se americký prezident, světoví lídři zpívající nejrůznější písničky či slavné celebrity, přiznávající se k nejrůznějším věcem. Ačkoliv mnohé z těchto falešných videí mají za cíl pouze pobavit, či pobouřit, deepfakes jsou taktéž silným dezinformačním nástrojem s potenciálem ovlivnit při správném použití velké množství lidí.

Co tedy deepfake videa vlastně jsou a odkud se vzaly? Deepfake je pokročilá forma falzifikace videa, jedné se o videomontáž využívající umělé inteligence a neuronových sítí, které shromažďováním modelů lidského chování vytváří vlastní modely. Podkladem pro tvorbu deepfake bývají obrázky či videa dostupná na internetu, podle kterých je počítač schopný simulovat pohyby a mimiku konkrétních osob v různých situacích. Nově vytvořený obraz a zvuk se poté vkládá do již existujícího videa, které má navodit pocit originality. Samotný

název „deepfake“ pochází ze slov „deep“ tedy hluboký (ve smyslu hluboké učení) a „fake“, tedy falešný. [19], [20]

„Vezmete například videa majitele Facebooku Marka Zuckerberga a natrénujete tu síť na pohyby jeho úst, na jeho hlas. A potom do videa vsunete úplně jiný text, který on sám nikdy neřekl, ale počítač vám ten obraz vygeneruje tak kvalitně, že skutečně vidíte Marka Zuckerberga, jak to říká.“ [19]

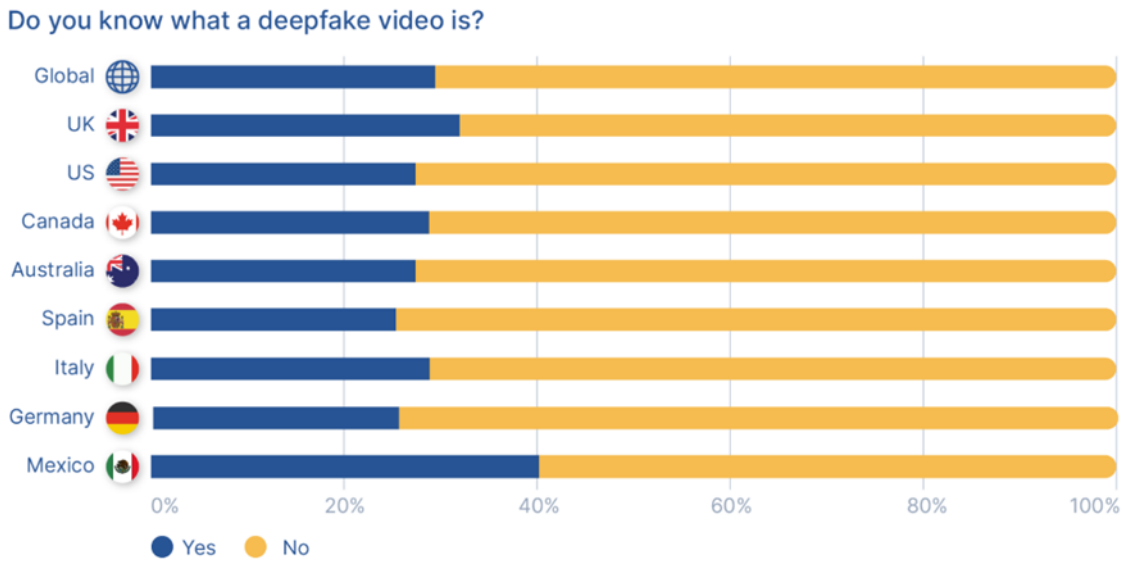
2.1.1 Povědomí lidí o deepfake

Falešná videa generovaná umělou inteligencí poprvé upoutala pozornost veřejnosti koncem roku 2017, kdy účet na platformě Reddit s názvem Deepfakes zveřejnil pornografická videa generovaná pomocí algoritmu pro výměnu obličejů založeného na Hlubokých neuronových sítích. Následně se termín deepfake začal používat šířeji a označuje všechny typy videí generovaných umělou inteligencí, které se vydávají za falešné. [21]

Navzdory rostoucí hrozbě deepfakes pro společnost mnoho lidí stále neví, co to deepfake je. Například společnost iProov provedla v roce 2022 průzkum mezi 16 000 lidmi v osmi zemích (USA, Kanada, Mexiko, Německo, Itálie, Španělsko, Velká Británie a Austrálie) a položila jim řadu otázek týkajících se deepfake. Z výsledků výzkumu mimo jiné vyplynulo, že:

- Celosvětově 71 % respondentů tvrdí, že neví, co to deepfake je. Jen necelá třetina dotazovaných tvrdí, že o deepfakes ví.
- Nejvíce deepfakes znají lidé z Mexika a Spojeného království: 40 % mexických respondentů a 32 % britských respondentů tvrdí, že vědí, co je to deepfake.
- Španělsko a Německo se cítí být o deepfakes nejméně informované: Ve Španělsku i Německu odpovědělo 75 % respondentů záporně. [9]

Část výsledků daného průzkumu můžeme vidět níže na obrázku číslo 2.



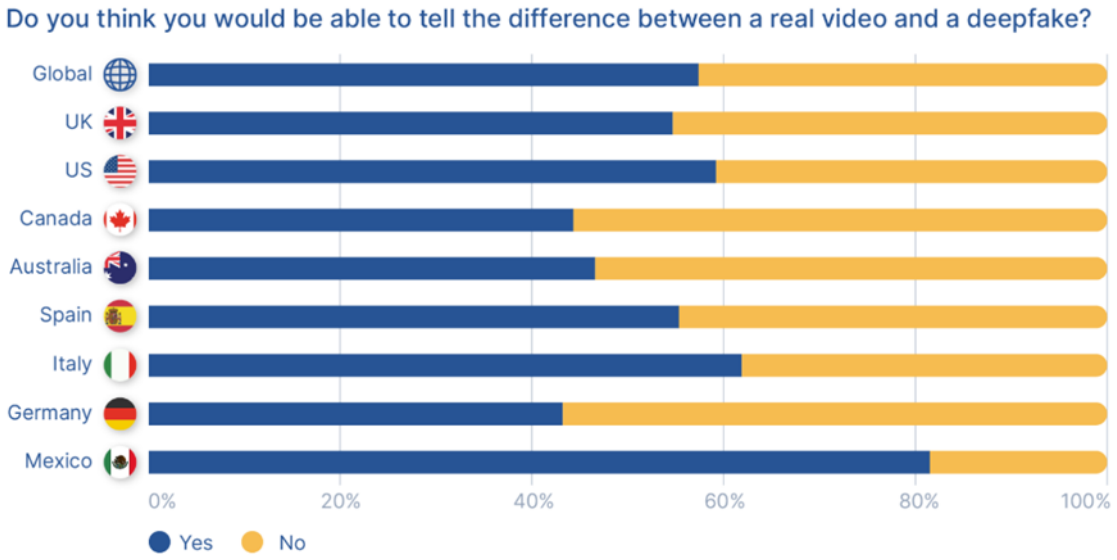
Obrázek 2 : Odpovědi respondentů na otázku „Víte, co to je deepfake?“ [9]

Navzdory na první pohled ne příliš optimistickým číslům se povědomí o deepfakes ve společnosti zvyšuje. Společnost Iproov uvádí že od předchozího výzkumu z roku 2019 se počet lidí, kteří mají povědomí o deepfakes téměř zdvojnásobil. V roce 2019 vědělo, co je to deepfake, jen 13 % lidí, zatímco v roce 2022 to bylo již 29 %. Tyto čísla jsou tak na jedné straně pozitivní, kdy vidíme postupný nárůst informovanosti ale zároveň i znepokojující, jelikož je 29 % v roce 2022 stále velmi malé číslo. Deepfakes mají značný potenciál pro zneužití, manipulaci či podvody a pokud lidé nevědí co jsou zač, je méně pravděpodobné, že budou připraveni rozpoznat, kdy se jedná o podvrh. [9]

Další otázkou, kterou Iproov lidem položila bylo, zdali si myslí, že dokážou rozlišit mezi skutečným videem a deepfake. V tomto případě byly výsledky následující:

- 57 % respondentů z celého světa si myslí, že by dokázali deepfake odhalit.
- 43 % respondentů přiznává, že by nedokázali rozlišit mezi skutečným videem a deepfake.
- Nejvíce si věří respondenti z Mexika: 82 % z nich se domnívá, že by dokázali rozeznat deepfake od skutečného videa.
- Naopak nejméně si jsou jisti respondenti z Německa: 57 % respondentů je přesvědčeno, že by rozdíl nepoznali. [9]

Část výsledků představuje obrázek číslo 3 níže.



Obrázek 3 : Odpovědi na otázku „dokážete rozpoznat deepfake?“ [9]

I v tomto případě Iproov zaznamenala nárůst oproti výsledkům výzkumu z roku 2019, kdy pouze 37 % respondentů uvedlo, že by dokázalo deepfake odhalit (vs 57 % v roce 2022). Opět se ale jedná o znepokojující údaje, jelikož současné sofistikované deepfakes mohou být pro lidské oko nerozeznatelné. Pokud jsme totiž příliš přesvědčeni o své schopnosti deepfake odhalit, můžeme se nakonec nechat snadněji oklamat. K ověření deepfake jsou často zapotřebí technologie tzv. hlubokého učení a počítačového vidění, které analyzují určité vlastnosti, například to, jak se odráží světlo na skutečné kůži v porovnání s obrazem nebo syntetickou kůží. [9]

V souvislosti s deepfakes jakožto potenciální bezpečnostní hrozbou, se lidé nejvíce obávají možnosti zneužití za účelem krádeže identity a získání přístupu k bankovním či jiným účtům. Další rozšířenou obavou je že „uvěřím něčemu, co není pravda“ a možnost krádeže identity za účelem zřízení kreditních karet nebo bankovních účtů na cizí jméno. [9]

2.1.2 Manipulační potenciál deepfakes

Kdy se tedy z na první pohled nevinné zábavy může stát silný manipulační nástroj? Největší nebezpečí představují deepfakes pro politiky a slavné celebrity. U deepfakes jde stejně jako u většiny dalších dezinformačních nástrojů, především o použití ve vhodný čas a na vhodnou skupinu lidí. Vhodným časem máme u politiků na mysli například období před volbami nebo během vyhrocených politických rozepří. V případě celebrit např. v momentě kdy se na povrch dostane nějaký skandál nebo během probíhajících soudů či sporů. Vhodnou skupinu lidí představují v případě politické manipulace nepochybně voliči strany, proti které je

deepfake namířen, ale stejně tak i nerozhodnutí voliči. Vidět a slyšet jednoho z kandidátů říkat věci, které se z mými názory absolutně neztotožňují, může nejednoho voliče šokovat a odradit. Pokud je deepfake namířen proti populární celebritě, může ji vážně poškodit kariéru nebo pověst. V případě probíhajícího sporu, může ovlivnit veřejné mínění v neprospěch této konkrétní osoby. Úspěch deepfakes je dále podmíněn kvalitním zpracováním a dosahem, tzn. čím kvalitnější videomontáž je a k čím většímu počtu lidí se dostane, tím má větší potenciál uškodit.

2.1.3 Konkrétní příklady manipulací

Pojďme se nyní podívat na několik vybraných případů, kdy byly deepfakes použity k manipulaci. Jedním z nejnebezpečnějších případů je ten z března minulého roku. V prvních týdnech ruské invaze na Ukrajinu kolovalo po internetu falešné video s prezidentem Zelenským, který v něm stojí na pódiu a vyzývá ke složení zbraní. Video bylo hackery umístěno na webovou stránku ukrajinských novin a taktéž bylo sdíleno na sociálních sítích. Tento hybridní útok byl ale naštěstí neúspěšný, z webu Ukraine 24 bylo video rychle odstraněno a k žádnému hromadnému složení zbraní nedošlo. Za neúspěchem mohla stát taktéž špatná kvalita této videomontáže. Na obrázku níže si lze všimnout hned několika výrazných nedokonalostí. V první řadě je to hlava, která svými proporcemi neodpovídá zbytku těla, dále nepřirozeně vypadající krk a celkově nízká kvalita rozlišení videa. Na druhou stranu měl tento útok bezpochyby velký potenciál a při lepším provedení, mohl zásadně ovlivnit probíhající konflikt. Lze to proto brát i jako velké varování do budoucna. [22]

Obrázek 4 : Deepfake prezidenta Zelenského [22]

Dalším možným využitím technologie deepfake je zneužití za účelem reklamy. Deník The Wall Street Journal upozornil na několik společností využívajících ve svým reklamách deepfake verze slavných osobností. Ruská telekomunikační společnost MegaFon např. ve svém komerčním klipu použila napodobeninu hollywoodského herce Bruce Willise. Investiční společnost reAlpha Tech Corp pro své účely použila falešného Elona Muska a herci Tom Cruise a Leonardo DiCaprio pro změnu „účinkovali" v promo klipu firmy Paperspace Co. Žádná z těchto osobností nestrávila ani okamžik natáčením takových klipů a Musk, Cruise a DiCaprio navíc ani nikdy nesouhlasili s propagací daných společností. Dostáváme se tak do šedé právní zóny. Podle odborníků by celebrity mohly mít problémy s tím, aby zabránily šíření neautorizovaných digitálních reprodukcí své osoby a manipulaci se svou značkou a pověstí. Autorizované deepfake reklamy by na druhou stranu mohly marketérům ušetřit spoustu času a financí, jestliže není nutné, aby se herci dostavili na natáčení osobně. Do budoucna se tak mohou reklamy využívající deepfake technologie stát běžnou praxí. [23]

Velmi kontroverzní je deepfake pornografie. Za rok 2019 tvořily simulace pornografických snímků ženských celebrit (sdílených pochopitelně bez jejich souhlasu) údajně až 96 % všech deepfakes. Deepfakes se tak mohou jednoduše stát nástrojem k vydírání. [24]

Jedním z nejznámějších deepfakes je video z falešným Morganem Freemanem umístěné na serveru Youtube. Videomontáž je až strašidelně přesvědčivá a dodnes se jedná o jedno z nejlépe zpracovaných deepfake videí vůbec. Video poprvé sdílel nizozemský YouTube deepfake kanál *Diep Nep* [25] v loňském roce, přičemž se jedná o jedno z nejlépe zpracovaných deepfake videí vůbec. Mezi deepfakes tvůrci jsou taktéž populární nejrůznější bizarní filmové/herecké crossovery. Na serveru Youtube jich lze opět najít spoustu. Velmi povedený je např. deepfake herce Jima Carreyho, který ho pasuje do role Jacka Torrence ve filmu Osvícení. Vesměs jde o několik klipů z tohoto filmu a opět se jedná až o děsivě přesvědčivou video montáž – viz obrázek 5 níže.

Obrázek 5 : Deepfake – Jim Carrey [26]

2.1.4 Shallowfake

V roce 2019 vyvolalo skandál video s údajně opilou předsedkyní sněmovny reprezentantů Nancy Pelosi. Zmanipulované video vidělo na Facebooku přes 2,5 milionu lidí, a přestože bylo označeno za deepfake, nejednalo se o deepfake v pravém slova smyslu. Video bylo pouze výrazně zpomaleno, aby projev působil nesouvisle a zmateně. Tímto způsobem zmanipulovaná videa se označují jako shallowfake. Shallowfake jsou videa zmanipulovaná pomocí základních editačních nástrojů jako je právě zrychlení či zpomalení záznamu. Hlavním rozdílem oproti deepfakes je, že shallowfake nevyužívá AI a hlubokého učení. Příkladem zpomaleného záznamu je zmíněné video s N. Pelosi. Ke zmanipulování záznamu pomocí zrychlení, došlo např. v roce 2018 při rozhovoru s Donaldem Trumpem. Novinář Jim Acosta na videu v momentě, kdy si od něj stážistka bere mikrofon působí vlivem zrychlení agresivně. [27], [28]

Někdy se shalllowfakes označují taktéž jako „dumbfake“ a spadají sem taktéž videa, která mají evokovat dojem, že byly natočené jinde a jindy, než tomu bylo doopravdy. Tento způsob manipulace může být opět velmi nebezpečný.

2.2 Jak vzniká deepfake?

K vytvoření deepfake videa lze vesměs použít dvou metod. Obě metody spočívají ve využití hlubokých neuronových sítí, přičemž první využívá neuronovou síť typu enkodér – dekodér, která pracuje s technikou výměny tváří. Druhá metoda využitá neuronovou síť typu GAN (generativní soupeřící sítě), která opakovaně detekuje a vylepšuje nedokonalosti deepfakes do té doby, až jsou od pravého videa (nebo fotografie) prakticky nerozeznatelné. Pojďme se nyní blíže podívat na to, co vlastně neuronové sítě jsou a jak jednotlivé metody fungují.

2.2.1 Neuronové sítě

Neuronové sítě jsou srdcem algoritmů pracujících na bázi hlubokého učení. Jejich název a princip je inspirován lidským mozkem, jelikož pracují na podobném principu, jakým si předávají informace biologické neurony. Neuronová síť obsahuje vrstvy vzájemně propojených uzlů, kdy každý uzel je nazýván perceptron. V neuronové sítí máme vstupní vrstvu, jednu nebo více skrytých vrstev (u složitějších sítí) a výstupní vrstvu. Každý uzel (umělý neuron) se připojuje k jinému a má přiřazenou váhu a práh. Pokud je výstup některého jednotlivého uzlu vyšší než zadaná prahová hodnota, tento uzel se aktivuje a odešle data do další vrstvy sítě. V opačném případě do další vrstvy sítě neodesílá žádná data. [29]

Jakmile je určena vstupní vrstva, dojde k přiřazení váhy. Tyto váhy určují důležitost konkrétní proměnné, přičemž větší váhy přispívají k výstupu významněji než ostatní vstupy. Všechny vstupy se pak vynásobí příslušnými váhami a sečtou. Výstup je následně předán aktivační funkci, která určuje výstup. Pokud tento výstup přesáhne danou mez, dojde k aktivaci uzlu a předání dat další vrstvě sítě. Dochází tedy k tomu, že výstup jednoho uzlu se stává vstupem dalšího uzlu. [29]

Neuronové sítě jsou učeny na tzv. tréninkových datech a s postupem času se jejich přesnost zlepšuje. Jakmile jsou tyto algoritmy vyladěny na přesnost, stávají se mocnými nástroji v oblasti informatiky a umělé inteligence (ale nejen tam), díky nimž můžeme např. velmi rychle klasifikovat a shlukovat data. Například společnost IBM na svých stránkách pomocí binárních hodnot podrobněji rozebírá, jak může vypadat jeden uzel neuronové sítě. Nabízí příklad, kdy se člověk rozhoduje, zda si má jít zasurfovat či ne. Rozhodnutí jít či nejít je predikovaná hodnota *(y-hat)*. Předpokládáme, že naše rozhodnutí ovlivňují tři faktory:

Jsou vhodné vlny? (Ano: 1, Ne: 0)
Je na pláži málo lidí? (Ano: 1, Ne: 0)

Došlo v poslední době k útoku žraloka? (Ano: 0, Ne: 1)
Poté předpokládejme, že máme následující vstupy:

$x_1 = 1$, protože jsou vhodné vlny
$x_2 = 0$, protože na pláži jsou davy lidí
$x_3 = 1$, protože v poslední době nedošlo k útoku žraloka.

Nyní přiřadíme váhy, abychom určily, které proměnné mají pro rozhodnutí nebo výsledek větší význam.
$w_1 = 5$, protože velké vlny se nevyskytují často.
$w_2 = 2$, protože davy lidí nám nevadí.
$w_3 = 4$, protože máme strach ze žraloků.

Nakonec budeme také předpokládat prahovou hodnotu 3, což by znamenalo hodnotu zkreslení -3 (tzv. *bias*). Po zadání všech různých vstupů můžeme začít dosazovat hodnoty do vzorce, abychom získali požadovaný výstup.
Používáme vzorec pro lineární regresi, kdy násobíme váhy vstupními hodnotami.
$\sum w_i * x_i + \text{bias} = w_1 * x_1 + w_2 * x_2 + w_3 * x_3 + bias$
$y\text{-}hat = (1*5) + (0*2) + (1*4) - 3 = 6$
Prahovou hodnotu jsme předpokládali 3 a z výsledku 6 tedy vyplývá že půjdeme surfovat. Pokud by výsledná hodnota byla méně než 3 surfovat nepůjdeme. Binární výsledek tohoto uzlu pak je 1. Případným upravováním váhy nebo prahu můžeme dosáhnout rozdílných výsledků. Příkladem neuronových sítí může být neuronová síť kodér – dekodér či generativní soupeřící síť – GAN. [29]

2.2.2 Autoenkodéry – neuronová síť kodér – dekodér

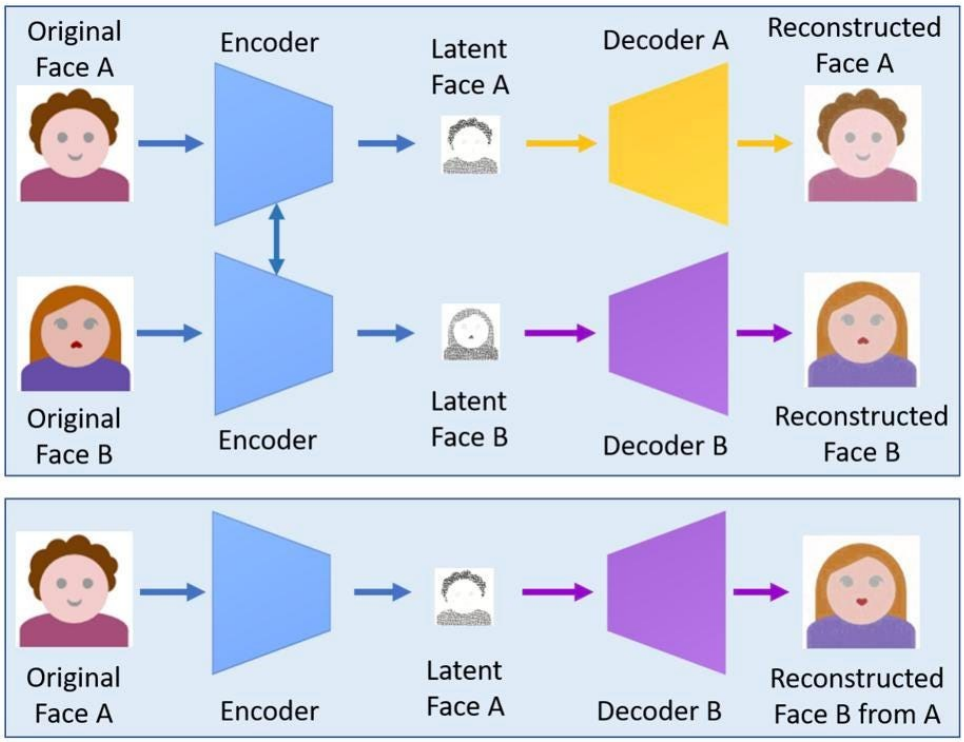
Technika s využitím autoenkodérů je nejběžnějším způsobem tvorby deepfakes. Nejprve je potřeba cílové video, které slouží jako základ deepfake např. slavná filmová scéna nebo projev nějakého politika. Poté je potřeba zdrojové video či klipy, obsahující např. obličej osoby, kterou chceme do cílového videa vložit. Obě videa spolu nemusí mít nic společného, stačí pouze aby obsahovala dostatek informací, které lze z obličejů získat. Pokud bychom chtěli použít např. vlastní obličej, je potřeba nahrát cca 5 ideálně pak až 20 minut dlouhý klip, kde se budeme dívat do kamery a předvádět různé pohyby a mimiku, tak aby byl obličej nasnímán ze všech úhlů.

Autoenkodér je obecně program umělé inteligence s hlubokým učením. Úkolem autoenkodéru je studovat videoklipy a učit se z nich pohyby a rysy obličejů osob z různých úhlů a v různých situacích. Obličej osoby ze zdrojového videa pak přenáší mappingem na osobu z koncového videa a hledá podobné rysy. Čím delší a kvalitnější zdrojové video bude, tím více detailů z naší tváře (nebo kohokoliv jiného) autoenkodér získá. A čím déle hluboké učení poběží, tím kvalitnější výsledný deepfake bude.

Kodér – dekodér

Celý tento proces zajišťuje dvojice kodér a dekodér, která hraje v deepfake generátorech „hlavní roli“. Kódování a dekódování obličejů pomocí kodéru a dekodéru spočívá v získání informací o vlastnostech obličeje a následné rekonstrukci tohoto obličeje z těchto informací. Kódování se zaměřuje na vytvoření skryté (latentní) verze obličeje, která zachycuje emoce a výrazy, zatímco dekódování se zaměřuje na použití této skryté verze k obnovení původního obličeje. Kodér a dekodér jsou obecně rekurentní neuronové sítě (RNN), které se trénováním na tisících zdrojových/cílových obrázcích exponenciálně zlepšují. Rekurentní neuronové sítě obsahují na rozdíl od jednodušších neuronových sítí (s jednou vstupní a jednou výstupní vrstvou) navíc ještě skrytou vrstvu, která neuronové sítí umožňuje „zapamatovat“ si předchozí stavy. [30]

Obrázek 6 níže představuje princip tvorby deepfake pomocí dvou párů kodér-dekodér. Máme zde dvě sítě používající stejný kodér, ale odlišné dekodéry pro tréninkový proces. Obrázek obličeje A je zakódován společným kodérem a dekódován dekodérem B k vytvoření deepfake (rámeček dole). Rekonstruovaný obraz (v pravém dolním rohu) je obličej B s ústy obličeje A. Obličej B má původně ústa ve tvaru srdce obrácená vzhůru nohama, zatímco rekonstruovaná tvář B má pak ve výsledku ústa ve tvaru běžného srdce. [30]



Obrázek 6 : Proces tvorby deepfake [30]

2.2.3 GAN – Generativní soupeřící sítě

Pro tvorbu současných přesvědčivých deepfake je však zásadní technologie tzv. generativních soupeřících sítí (GAN) vytvořená v roce 2014 doktorem a pozdějším zaměstnancem Applu Ianem Goodfellowem. Můžeme říct, že bez GAN by realisticky vypadající deepfake jak je známe dnes nebyly pravděpodobně možné. Technologie generativních soupeřících sítí proti sobě staví dva algoritmy umělé inteligence. Úkolem prvního je vytvořit zfalšovaný obraz a úkolem druhého tento padělek odhalit. Pokud druhý algoritmus odhalí padělek AI padělatele se přizpůsobí a zlepší. Tímto způsobem se oba algoritmy stávají během procesu učení ve svých úkolech stále efektivnější a výsledná videa či obrázky uvěřitelnější. [8]

Historie Vývoje GAN technologie

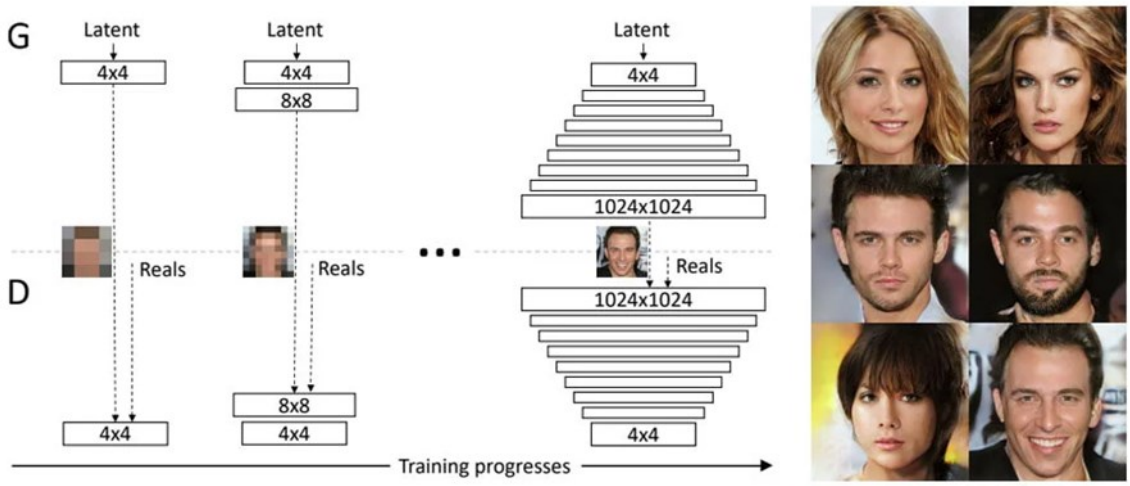
Technologie od svého vzniku v roce 2014 urazila dlouhou cestu a rozdíl mezi vygenerovanými obličeji z let 2014–2015 a těmi současnými je skutečně velmi znatelný. Jak dokazuje obrázek 7 níže na počátku byly obličeje velmi nedokonalé a mnohdy v podstatě až děsivé. GAN se ovšem od té doby dočkaly několika vylepšení. V roce 2015 zkombinovali výzkumníci GAN s konvolučními neuronovými sítěmi (CNN) optimalizovanými pro rozpoznávání obrazu, které mohou paralelně zpracovávat velké množství dat, a zvláště dobře fungují na

grafických kartách. Nahradily jednodušší sítě, které dříve poháněly agenty GAN, což umožnilo dosáhnout důvěryhodnějších výsledků. [8]

Obrázek 7 : Ilustrace vývoje GAN technologie [8]

V roce 2016 přišli vědci s možností kombinace dvou GAN sítí, kdy agenti těchto sítí mezi sebou sdílejí informace a mohou se tak paralelně učit. Každý agent navíc mírně upravuje naučená data a v této době je již možné vygenerovat i osoby se slunečními brýlemi, náušnicemi atd. Falešné portréty se tak dále zlepšují, ovšem stále jsou poměrně jednoduše odhalitelné. [8]

V roce 2017 se výzkumníkům ze společnosti Nvidia se podařilo dosáhnout významného milníku ve vývoji deepfakes. Všechny předchozí GAN měli totiž jeden společný problém, čímž byla nízká kvalita výsledných snímků. Agenti generátorů často vytvářeli snímky s nízkým rozlišením, protože pro zkoušejícího agenta bylo obtížnější odhalit je jako padělky – více pixelů znamená potenciálně více zdrojů chyb. Dávalo tudíž smysl, aby se umělá inteligence padělatelů vyhýbala vysokým rozlišením a prošla tak přes zkoušejícího agenta. Nvidia přišla s řešením, kde trénink sítě probíhá postupně. Padělatelská umělá inteligence se nejprve naučí vytvářet obrázky v nízkém rozlišení, poté se rozlišení postupně zvyšuje – viz obrázek níže. [8]



Obrázek 8 : Tréninkový proces GAN sítě [8]

GAN, která se tímto způsobem dále rozrůstá, vytváří falešné portréty nebývalé kvality: Snímky mají stále nedostatky, ale rozhodně již dokážou oklamat nejednoho uživatele internetu. Mnoho stránek generujících neexistující obličeje využívá právě technologii GAN.

Současná deepfake videa jsou při troše pozornosti ještě stále relativně snadno odhalitelná. Především je potřeba zaměřit se na jakékoliv nesrovnalosti v barvě a mimice obličeje. Nepřirozené mrkání, podivně zdvojené obrysy rtů či uší nebo měnící se barvy v obličeji. Všechny tyto detaily nám mohou při určování pravosti videa velmi napovědět – konkrétní příklady jsou uvedeny v kapitole 5. Do budoucna bude ale odhalování těchto video montáží stále obtížnější. Technologie pochopitelně postupuje kupředu mílovými kroky a ani v případě deepfakes tomu není jinak. Dá se proto předpokládat, že se již v blízké budoucnosti dostaneme do bodu, kdy bude deepfake od pravého videa opravdu pro běžného člověka k nerozeznání.

2.3 Manipulace s fotografiemi

Fotomontáže a úpravy fotek jsou dnes již naprosto běžnou součástí našich životů a zcela jistě nikoho nepřekvapí, že pomocí počítače je možné původní fotografii změnit takřka k nepoznání. Vyretušovat, přidat člověka, který na fotce původně vůbec nebyl, vyměnit dvěma lidem hlavy či někoho nežádoucího zcela odstranit. Všechny tyto úpravy dnes díky aplikacím jako je Adobe Photoshop a dostupnosti nejrůznějších návodů zvládne i méně zkušený uživatel, a proto je téměř každá fotografie, kterou kdekoliv na internetu nalezneme, nějakým způsobem upravena. Pochopitelně tomu tak ale nebylo vždy a v minulosti bylo upravování fotografií podstatně složitější.

2.3.1 Stručná historie upravování fotografií

Ačkoliv rozmach fotografických úprav přišel až s vynálezem digitálních fotografií a osobních počítačů, je historie upravování a manipulace s fotografiemi téměř tak stará jako historie samotné technologie fotografie. Vedle pochopitelných estetických důvodů, k tomu však byly v minulosti často i důvody politické, kdy například docházelo k odstraňování politicky nepohodlných lidí z fotografií. Za jednu z prvních upravených fotografií je považována slavná fotomontáž Abrahama Lincolna, na které je jeho hlava umístěna na tělo politika Johna Clahouna. Fotomontáž vznikla kolem roku 1860 spojením dvou fotografií pravděpodobně aby Lincoln působil na fotografii více ,,heroicky". Dalším podobným příkladem z této doby je vyobrazení generála Granta z období americké občanské války (1864), která je dle historiků složena dokonce ze tří různých snímků. S příchodem 20 století se fotografie začaly upravovat především z politických důvodů. Přispěly k tomu jak obě světové války, tak i množství totalitních režimů, které v této době vznikaly na mnoha místech světa. Na fotkách komunistických či fašistických vůdců běžně mizí lidé nehodící se režimu a s fotkami je obecně manipulováno v rámci propagandy tak, aby konkrétní vůdce působil co nejsilněji. Jako příklad lze uvést fotografii italského diktátora Benita Mussoliniho sedícího na koni z roku 1942. Z tohoto snímku byl odstraněn sluha, který koně přidržoval, aby samotný Mussolini působil více statečně a hrdinsky. Dalším známým příkladem je fotografie Josifa Stalina, ze které postupně zmizely dokonce tři osoby, aby na ní nakonec zůstal stát pouze samotný Stalin. [31]

Před příchodem počítačů a vznikem digitální fotografie se úpravy prováděly nejčastěji pomocí inkoustu či barev. Velmi často používanou technikou byl ale také tzn. airbrushing – stříkání barev pomocí vzduchu, česky fixírování. Původně se předpokládalo, že airbrushing vznikl až v 90. letech 19. století, ovšem podle několika výzkumů tomu bylo již dříve a první airbrush sestrojil v roce 1879 Abner Peeler. Byla to ovšem značně primitivní konstrukce, kdy bylo nutné pro usnadnění procesu rozprašování ručně prohánět vzduch. První moderní airbrush byl představen na veletrhu v Chicagu v roce 1893. Zařízení svým vzhledem připomínalo klasické pero a funkcí se blížilo dnešním moderním airbrushům. Airbrushing se během 20.století používal především na odstraňování osob a objektů z fotek, pokud bylo potřeba do snímku naopak někoho či něco přidat, řešilo se to prostým vystřihováním částí snímků a různým skládáním negativů. Vzhledem k povaze těchto úprav a používaným technikám se v této době mluví o úpravě fotek spíše jako o umění, něž jako o řemeslu. Celé to totiž působilo spíše jako malířství. [31]

2.3.2 Současné fotografické manipulace

V dnešní době se ale stále častěji setkáváme s úpravami za účelem manipulace, ovlivňování veřejného mínění, dehonestování určitých osob či překrucování nebo popírání skutečností. Stejně tak dochází i k situacím, kdy fotografie jako taková není přímo upravena, ale je pozměněn kontext, původ či jednoduše nepředstavuje to, za co je vydávána. Typickým příkladem může být falešný profil na internetové seznamce či sociální síti, kdy se za odcizenou fotografií půvabné modelky skrývá někdo zcela odlišný nebo když se někdo snaží vydávat fotografii z hudební akce jako shromáždění na podporu určitého kandidáta či politické strany. V takovém případě je potřeba zaměřit se na kontext a fotografii např. reverzně vyhledat – příklad takové situace je uveden v kapitole 3.3.

V případě editace fotografií dochází často k odstraňování předmětů, osob nebo celých pozadí. Někdy je naopak do snímku něco přidáno nebo přepsáno. K odstraňování předmětů z fotografií se používají retušovací nástroje např. klonovací razítko, které kopíruje obrazové body z jedné části obrázku do druhé. Postavy a objekty jsou do fotografií vkládány jako nová vrstva, která je vyjmuta z jiné fotografie. Nejznámější a pravděpodobně nejlepší program na úpravu fotografií je placený AdobePhotoshop, který umožňuje množství pokročilých úprav. Alternativou zdarma může být např. GIMP či Krita. Existují taktéž online nástroje schopné provést jednoduché úpravy bez nutnosti program stahovat a instalovat – např. Pixlr. Obrázky 9 a 10 níže například představují fotografii před a po úpravě. Z fotografie byly odstraněny dva automobily a byl přepsán nápis pod značkou upozorňující na začátek obce. Jak je možné (někdy) zjistit, zda bylo s fotografií manipulováno je popsáno dále v kapitole 3.3.

Obrázek 9 : Testovací fotografie před úpravou

Obrázek 10 : Testovací fotografie po úpravě

Metadata

Metadata dokumentů jsou obecně strukturovaná data poskytující informace o datech v digitalizovaných dokumentech. Zjednodušeně můžeme říct, že metadata jsou data o datech. Jsou to například informace o tom, kdy a na jakém zařízení snímek vznikl nebo kde byl pořízen. Většina fotoaparátů zná metadata jako EXIF (Exchangeable Image File Format) a vytváří je automaticky. U fotografií jsou to obvykle tyto informace:

- Informace o poloze ve formě souřadnic GPS
- Datum a čas pořízení snímku
- Model a výrobce zařízení
- Nastavení fotoaparátu jako je clona, rychlost závěrky a ISO
- název a verze použitých nástrojů pro úpravy

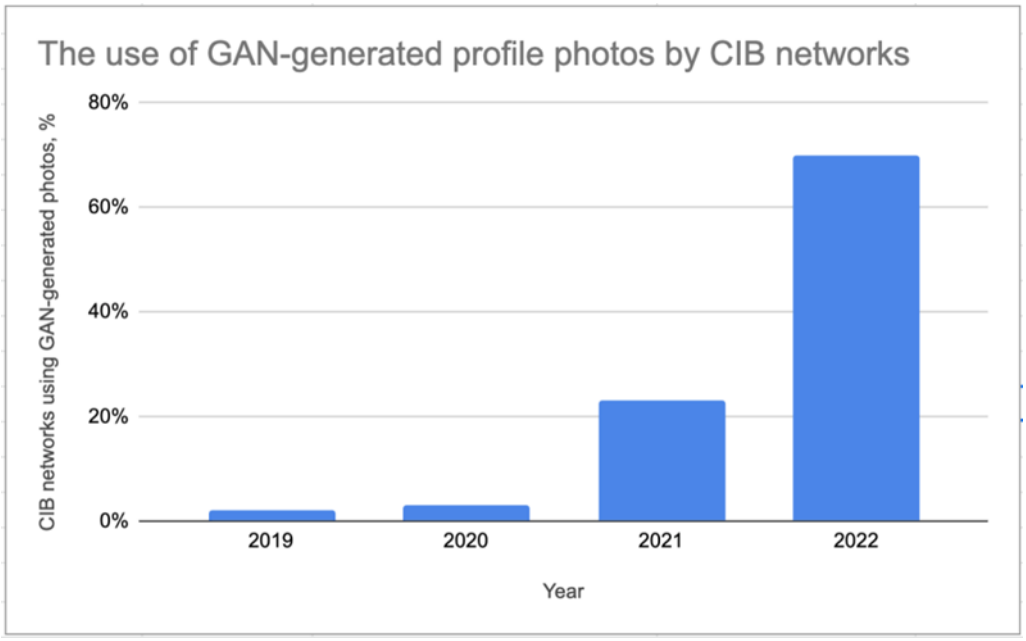
EXIF data si můžeme jednoduše zobrazit u každé fotografie ve vlastnostech v záložce podrobnosti, ale například velké sociální sítě jako Facebook či Twitter je z fotografií automaticky odstraňují a stejně tak je možné metadata přepisovat. Obecně je vhodné z každé fotografie před tím, než ji nahrajeme kdekoliv na internet metadata odstranit. Ve Windows lze tak učinit v záložce podrobnosti nebo použít externí nástroj např. ExifTool, který veškerá data odstraní. S metadaty umí nicméně i pracovat většina fotografických editorů, tj. zobrazit je, editovat či vymazat.

Je tedy možné podle metadat (pokud je máme k dispozici) zjistit, zda bylo z fotografií manipulováno? Některé aplikace jako např. Photoshop zanechávají v metadatech svou historii, podle které můžeme snadno zjistit, zda byla fotografie upravena, pokud fotografie obsahuje informace o poloze, napoví nám více o kontextu. Další možností je podívat se, zda metadata odpovídají skutečnosti. Pokud je například snímek pořízen s nízkou clonou a hloubkou ostrosti, může mít rozmazané pozadí. Stejně tak pomalá rychlost závěrky způsobí rozmazání pohybujících se objektů. V případě že se tyto parametry neshodují je pravděpodobné, že byl snímek nějak upraven.

2.4 Generátory fotografií

Generátory obličejů jsou velkým fenoménem poslední doby. Webové stránky jako je thispersondoesnotexist.com [2], či unrealperson.com [32] dokážou v dnešní době vygenerovat během sekundy fotorealisticky vypadající snímek lidského obličeje a umělá inteligence nás tak opět posouvá blíže k momentu, kdy nebudeme schopni rozeznat pravdu od fikce.

Zajímavostí je výroční zpráva společnosti Meta vlastnící Facebook. [5] Ve zprávě se mimo jiné píše o rapidním nárůstu falešných profilů používajících obličeje vytvořené generátory. Na obrázku 11 níže je vidět graf výskytu falešných profilů s vygenerovanými obličeji mezi lety 2019-2022. Více jak dvě třetiny falešných profilů, které byly společností v roce 2022 zrušeny obsahovaly obličej z generátoru. A důvod je velice prostý. Jednak je to způsobeno širokou dostupností těchto generátorů a jednak také „neprůstřelností“ těchto podvrhů. Pokud použiji neexistující obličej, nikdo si nemůže stěžovat, že jsem mu odcizil ten jeho. Ta osoba přece neexistuje. Vygenerované obličeje navíc nejdou reverzně vyhledat, což dále stěžuje jejich odhalitelnost. [5]



Obrázek 11 : Statistiky společnosti Meta v použití profilových fotografií generovaných pomocí GAN [5]

Podle další studie z roku 2022 [33] jsou obličeje vytvořené generátory již tak přesvědčivé, že lidé mají zhruba 50% šanci na to je odhalit. Samozřejmě je možné lidské oko na tyto podvrhy vytrénovat a člověk, který se touto problematikou zabývá a vygenerovaných tváří viděl již stovky nebo tisíce bude mít procento úspěšnosti nepochybně větší. Běžný uživatel internetu ale tyto schopnosti nemá a v této části práce bych se proto chtěl zaměřit na podvrhy z generátorů a upozornit na několik detailů, které nám mohou pomoct odhalit skutečnost.

2.4.1 Jak se pokusit odhalit vygenerovaný obličej

Konkrétní detaily bych rád předvedl na několika následujících falešných fotografiích. Hned první obrázek (obrázek 12 níže) z webu unrealperson.com [32] okamžitě zaujme na první pohled podivně zdeformovanou náušnicí na levé straně obličeje. Brýle, náhrdelníky, přívěsky nebo jako je tomu v tomto případě náušnice jsou pro generátory stále poměrně obtížným úkolem a pokud se u postavy nacházejí, měly by být první věcí na kterou se zaměřit. Na tomto konkrétním případě tak můžeme vidět, že i když samotná tvář působí realisticky, módní doplňky mohou vygenerovaný obličej jednoduše prozradit.

Obrázek 12 : Vygenerovaný obličej – příklad č.1 [32]

U druhého obrázku (Obrázek 13) se žádné doplňky, které by nám mohly napovědět bohužel nenacházejí. Problémem je zde totiž pozadí, které můžeme označit za další častou nedokonalost vygenerovaných výsledků. U tohoto snímku působí velmi podivně a nerealisticky a v takových případech je potřeba zbystřit a zamyslet se nad původem takové fotografie.

Obrázek 13 : Vygenerovaný obličej – příklad č.2 [32]

Dalšími znaky bývají také oči vycentrované přesně ve středu snímku, nepřirozeně pokroucené pramínky vlasů, které se v některých místech ztrácejí do prázdna nebo brýle či jiné doplňky splývající s obličejem.

Stejným způsobem jsou generátory schopné vytvořit i neexistující fotografie např. koček nebo koní. V případě koček jsou výsledky podobné těm lidským – vygenerovaná fotografie zobrazuje pouze tvář a opět působí velmi realisticky. Problém nastává v případě koní, kdy je zobrazováno celé tělo zvířete, protože generátory si s tímto zatím nedokážou příliš dobře poradit a výsledky vypadají často velmi znetvořeně. Na obrázku 14 níže pro porovnání přesvědčivě a nepřesvědčivě vypadající obrázek koně. Zajímavé je, že oba výsledky jsou ze stejného generátoru, vytvořené jen několik sekund po sobě. Stejný generátor tak skutečně může vytvořit dva kvalitou naprosto odlišné snímky. Generované obrázky zvířat každopádně nejsou pro většinu z nás důvodem proč po generátoru sáhnout a lze tak předpokládat, že ani jejich zdokonalování není prioritou. Na druhou stranu je jen otázkou času, kdy bude i toto pro generátory hračka.

Obrázek 14 : Vygenerovaný obrázek koně – příklad [32]

Generované obrázky aneb výměna textu za pixely

V případě realistických obrázků generovaných umělou inteligencí na základě textového zadání je tu hned několik příkladů, které se na internetu masově rozšířily. V březnu 2023 způsobily na sociálních sítích rozruch fotografie papeže Františka v luxusní bundě a ruského prezidenta Putina klečící před čínským prezidentem Si Ťin-pchingem. Obě byly následně označeny jako falešné – vytvořené generátorem. Nicméně jedná se o velice přesvědčivé padělky, které mohou na první pohled leckoho zmást, a navíc je jejich vytvoření velmi jednoduché. Nástrojům jako je DALLE-2 či Midjourney stačí k vytvoření obrázků pouze textové zadání od uživatele.

Obrázek 15 : Falešná fotografie ruského a čínského prezidenta [34]

Obrázky jsou obecně tvořeny pomocí Generativní AI, která využívá počítačový program zvaný difúzní model. Model je trénován na obrovském množství obrázků s přiřazeným významem – kočka, obloha, televizor apod. Princip difúzního modelu potom spočívá ve „zničení“ a následném obnovení obrázku, kdy jsou jednotlivé pixely postupně promíchávány, dokud se nezmění v čistý šum. Když je tento proces dokončen, je program spuštěn naopak – tedy vrací se postupnou změnou pixelů k původnímu obrazu. Účinnost výsledného modelu je posuzována podle pravděpodobnosti, s jakou bude obnovený obraz významově připomínat ten původní. Zůstává nám tak model, který dokáže efektivně generovat informaci (obraz) z náhodnosti. Celý proces lze přirovnat např. ke skládání Rubikovy kostky, kdy se kostka postupně skládá do finální podoby. [35]

Co ale dělat v případě že se setkáme s těmito mimořádně realistickými syntetickými obrázky a nejsme si jisti jejich pravostí. Stejně jako v případě generovaných obličejů je potřeba si obrázek dostatečně podrobně prohlédnout. Osoby na těchto obrázcích mohou mít např. deformované prsty či jiné části těla a taktéž některé konkrétní předměty jako tužky, sklenice, mobilní telefony apod. mohou působit nepřirozeně nebo rozmazaně. Texty obsahující

nesmyslná písmena nebo znaky jsou dalším detailem poukazujícím na syntetický původ fotografie. Příloha práce obsahuje dva příklady vygenerovaných obrázků a návod, jak je odhalit.

2.5 Audio manipulace

Stejně jako v případě deepfakes a generovaných falešných obličejů se k manipulaci používá taktéž umělou inteligencí generovaný hlas. AI text to speech generátory (TTS) umožňují převést jakýkoliv text na audio soubory s lidsky znějícím hlasem. Fungují na velkém počtu zařízení a díky své jednoduchosti a dostupnosti jsou v současnosti velmi populární. Aplikace schopné generovat syntetický hlas jsou v první řadě nepostradatelnou pomůckou pro osoby s poruchami čtení nebo pro studium jazyků a stejně tak je využívají tvůrci nejrůznějšího obsahu na sociálních sítích. Problémem mohou být TTS generátory schopné věrně napodobit hlas konkrétní osoby. Na to, jakým způsobem může být tato (jinak nepochybně užitečná) technologie zneužita upozorňují následující příklady.

Britská společnost ElevenLabs zabývající se AI hlasovými generátory uvedla v lednu 2023 na trh svůj produkt Prime Voice AI. Hned v prvních týdnech od spuštění se objevily zprávy o zneužití platformy ke tvorbě deepfakes. Platforma umožňuje uživatelům klonovat hlasy z malých zvukových vzorků a objevily se tak například falešné klipy s Emmou Watson předčítající pasáže z knihy Mein Kampf nebo hlas připomínající dabéra Justina Roilanda jak mluví o tom, že se chystá zbít svou manželku. Systém umí podle Gizmodo.com napodobit hlas opravdu přesvědčivě, a i když je u delších klipů jistá umělost znát, u kratších nemusí laik poznat rozdíl. Společnost ElevenLabs samozřejmě nabídla několik řešení, jak tomuto zabránit. Je tak možné, že do budoucna bude nutné ověření účtu či, že celá free verze přestane existovat a každý pokus o klonování hlasu bude manuálně ověřen. [36]

První zaznamenaný příklad podvodu za použití deepvoice se odehrál v roce 2019 ve Velké Británii. Výkonný ředitel britské energetické firmy obdržel telefonát od údajného šéfa mateřské společnosti v Německu se žádostí o převod peněz. Mělo jít o urgentní platbu maďarskému dodavateli, která měla být zaplacena do hodiny. Ředitel tak učinil a společnost přišla o 220,000€. Útočníky se nepodařilo dopadnout. [37]

V roce 2020 se podařilo dalším podvodníkům odcizit díky falešnému hlasu 35 milionů dolarů od společnosti ve Spojených arabských emirátech. Útočníci použili falešný hlas ředitele společnosti a přesvědčili bankovního manažera, který s pravým ředitelem již dříve mluvil,

aby provedl několik transakcí. Útok navíc podpořili falešnými e-maily a manažer tak usoudil, že vše vypadá legitimně a banka transakce provedla. [38]

Podle vyjádření společnosti ESET bude podobných útoků díky jejich jednoduchosti a dostupnosti nástrojů přibývat. Manipulace se zvukem je jednodušší než tvorba deepfake videí. Obranou proti tomuto typu hrozeb může být větší osvěta a informovanost a taktéž lepší ověřovací metody. Patrick Traynor z technické univerzity Herberta Wertheima upozorňuje na možnou jednoduchou obranu před podobnými telefonáty. Pokud nám hovor přijde podezřelý, doporučuje se zavěsit a zkusit zavolat zpátky. Útočníci často používají anonymní telefonní čísla na jedno použití a v takovém případě můžeme snadno zjistit, zda opravdu mluvíme s danou osobou. Další možností může být dále i třeba heslo, kterým se bude dotyčná osoba prokazovat. [37]

2.5.1 Klonování hlasu

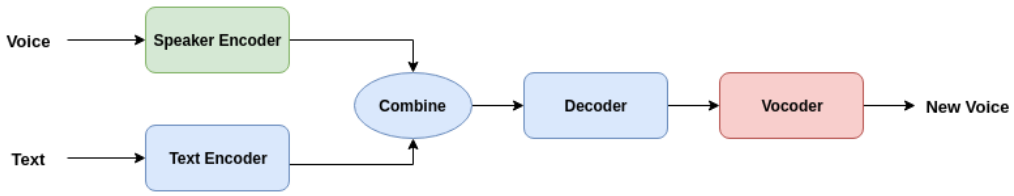
Pojďme se nyní podrobněji podívat na techniku klonování hlasu (taktéž zvanou *deepfake voice* či *deepvoice*), která generátorům umožňuje přečíst text libovolným hlasem. Klonování hlasu je dalším zástupcem metod hlubokého učení. Technologie se poprvé objevila v roce 1998 a od té doby se výrazně zlepšila a pokročila. Výrazný pokrop v technologii klonování hlasu učinila čínská firma Baidu, která v roce 2018 představila službu schopnou ze 3sekundového záznamu rozpoznat lidský hlas a následně tímto hlasem mluvit. V roce 2017 přitom algoritmus potřeboval 30 min audio záznamu. Pro efektivní trénink s přesvědčivým výsledkem je ale potřeba velké množství odpovídajících zvukových nahrávek. Může se jednat i desítky GB či více. Opět platí, že čím více zdrojových dat a delší proces hlubokého učení, tím lepší a přesvědčivější výsledky. Deepvoice nemusí pouze věrně napodobovat hlas konkrétního člověka, může rovněž měnit pohlaví hlasu, styl řeči nebo akcenty. [39]

Metoda vytvoření hlasového klonu spočívá obecně v použití softwaru na rozpoznávání hlasu. Pomocí tohoto softwaru lze vytvořit digitální kopii něčího hlasu tak, že se nahraje mluvící osoba a poté se analyzují její řečové charakteristiky. Digitální kopii určité osoby lze vytvořit taktéž ručně, kdy se „vystřihnou“ útržky řeči a následně poskládají dohromady jako puzzle.

Algoritmus klonování hlasu

Aby počítač dokázal přečíst text, potřebuje znát dvě věci – co čte a jak to vyslovit. Při klonování hlasu se běžně používá jako vstup pro neuronovou síť vzorek řeči a příslušný text,

který chceme, aby tento hlas přečetl. Neuronová síť se pak učí mapovat vztah mezi těmito dvěma vstupy a generovat syntetický hlas na základě textu. Ve schématu níže můžeme vidět že u obou vstupů, tedy u vzorku hlasu a daného textu dochází ke kódování pomocí kodéru řečníka a kodéru textu. Tento proces umožňuje získaní důležitých dat a jejich uložení v jednotné a kompaktní formě. Kombinací těchto dat je dekódován tzv. spektrogram, který je následně převáděn pomocí vokodéru do zvukové podoby, čímž vzniká syntetický hlas. Jednotlivé časti neuronové sítě jsou trénovány nezávisle na sobě. [40]



Obrázek 16 : Algoritmus klonování hlasu [40]

3 VYBRANÉ SOFTWAROVÉ NÁSTROJE

V následující části práce bych rád představil několik softwarových nástrojů potenciálně použitelných za účelem audiovizuální manipulace. Většina z nich funguje v základní verzi zdarma a je uživatelsky přívětivá. V případě aplikací na tvorbu deepfake uvádím taktéž několik placených aplikací, které nabízí značné přizpůsobení. Více podrobněji se v praktické části věnuji programu DeepFaceLab, který jsem se rozhodl použít při tvorbě ukázkového deepfake videa.

3.1 Nástroje na tvorbu deepfakes

Díky možnostem současného internetu a průnikem deepfake fenoménu do mainstreamu, je dnes snadné si nechat vyrobit deepfake video téměř na míru. K tomuto účelu lze na internetu nalézt celou řádku nejrůznějších webů a aplikací. Následující výčet představuje nejznámější a nejpopulárnější aplikace sloužící k těmto účelům.

3.1.1 Deepfakes Web

Deepfakes Web je služba, umožňující vytvářet deepfake videa přímo na webové stránce. Využívá hlubokého učení k absorbování informací a detailů z tváří. Deepfakes Web využívá výkonných grafických karet na cloudu a uživateli tak stačí nahrát pouze zdrojové a cílové video, se kterými aplikace následně pracuje. Učení a trénování ze zdrojového videa a obrázků může webu trvat až 4 hodiny, zatímco výměna obličejů pomocí natrénovaného modelu zabere dalších 30 minut. Služba funguje s i bez něj. Bez předplatného vytvoří deepfake video přibližně za 5 hodin, zatímco prémiová verze „vyplivne“ video za pouhou 1 hodinu. Cena je 3$ na hodinu a jedno deepfake video tak v základní verzi vyjde zhruba na 15$. Obrázek 17 níže je snímek úvodní stránky tohoto nástroje. [41]

Obrázek 17 : Nástroj deepfakes web [41]

3.1.2 Zao a Reeface

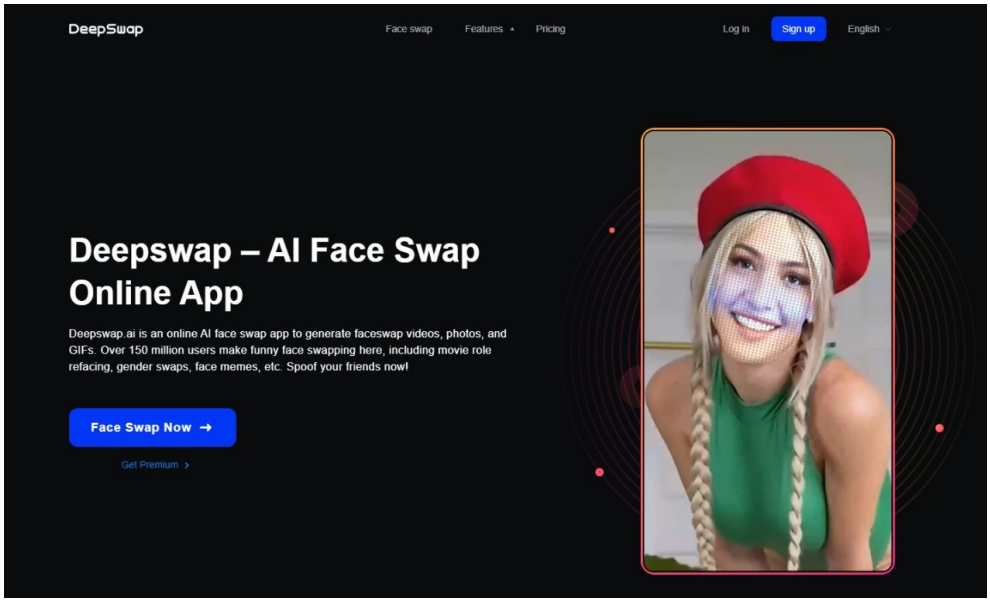
Zao je čínská deepfake aplikace pro záměnu tváří dostupná na operační systémy iOS a Android. Krátce po svém vydání v roce 2019 raketově vystřelila na vrchol žebříčků obchodů s aplikacemi, a i dnes se těší ohromné popularitě. Její obrovskou předností je rychlost, s jakou dokáže deepfake video vytvořit. Zao dokáže téměř okamžitě prohodit vaši vlastní tvář s tváří libovolné televizní hvězdy ze známých filmů či televizních klipů. Aplikace nejprve naskenuje vámi nahraný obrázek a následně jej transponuje na zvolenou postavu. Výsledek je poté plně animovaný v souladu s tím, jak se postava pohybuje. Je ovšem nutné zmínit možná bezpečností rizika související s obavami o zneužití osobních údajů či narušení soukromí. Stejně tak může vyvolat značné znepokojení i samotná rychlost a jednoduchost s jakou Zao deepfake videa vytváří. Aplikace je aktuálně dostupná pouze v Číně. [42]

Reeface nabízí podobné funkce jako Zao. Aplikace má v internetovém obchodě Google play momentálně více jak 100 mil stažení a v roce 2020 nominována na cenu *Google Play Users Choice Award*. Mezi uživateli je populární hlavně díky možnostem prohodit si tvář např. s populárními celebritami ze slavných seriálů nebo klipů. Obrázek níže je oficiálním promo obrázkem produktu. [43]

Obrázek 18 : Aplikace ReeFace [43]

3.1.3 Deepswap

Deepswap je online nástroj pro záměnu tváří ze dvou videoklipů. Nástroj funguje v prohlížeči bez nutnosti instalace. Web contentmavericks.com jej hodnotí jako nejlepší deepfake software pro rok 2023. Vyzdvihuje zejména rychlost a kvalitu, která je údajně srovnatelná s programy jako je např. DeepFaceLab. Nevýhodou je kreditový systém, který uživatele omezuje v počtu a délce nahrávek. Po zakoupení předplatného lze nástroj využívat bez omezení. Deepswap využijí zejména ti, kteří chtějí vytvořit přesvědčivé deepfakes, aniž by museli trávit hodiny u složitějších softwarů. Následující obrázek je z oficiální webové stránky nástroje. [44]



Obrázek 19 : Nástroj Deepswap [45]

3.1.4 DeepFaceLab

DeepFaceLab je bezplatný program pro operační systémy Windows umožňující vytvářet deepfake videa. Na rozdíl od výše zmíněných aplikací jako Wombo, Reeface či Zao schopných vytvářet instantně jednoduchá deepfakes, se jedná spíše o program pro výzkumníky a studenty počítačového vidění. Pro jeho správný chod je nezbytná výkonná grafická karta a taktéž uživatelské rozhraní není nejpřívětivější – používání je nutné se naučit z dokumentace či tutoriálů. Program má nicméně skvělou technickou podporu a flexibilní a na výkon zaměřenou implementaci což mu umožňuje dosáhnou hladkých a fotorealistických výsledků výměny tváří. Kromě nahrazování obličejů na obrázku nebo ve videu umožňuje software DeepFaceLab taktéž měnit hlavu, omládnout či zestárnout obličej nebo dokonce upravovat rty při projevech. Všechno jsou to ale pokročilé funkce a DeepFaceLab tak skutečně není jednoduchým programem pro každého. Použitím tohoto programu se zabývá praktická část práce. [40]

3.1.5 DALL-E 2 a Midjourney

DALL-E 2

DALL-E 2 je systém umělé inteligence schopný vytvářet realistické obrázky a umění na základě textového zadání vytvořený společností OPEN-AI. DALL-E využívá model strojového učení GPT-3 modifikovaný na generování obrázků. Pro použití je nutné si zakoupit kredity. Na obrázku 20 níže můžeme vidět obrázek plyšové hračky na skateboardu vytvořený

nástrojem DALL-E. Pokud se mimochodem podíváme pozorněji na kolečka skateboardu můžeme si všimnout jejich podivné deformace, což může prozradit, že se nejedná o opravdovou fotografii. [46]

Obrázek 20 : Fotografie vytvořená nástrojem DALL-E [46]

Midjourney

Midjourney je online služba na generování obrázků pomocí umělé inteligence dle textového zadání. Vyvíjí ji stejnojmenná společnost amerického podnikatele Davida Holze a funguje na platformě Discord. Nástroj dříve nabízel vyzkoušení zdarma, ale kvůli velkému zájmu a potenciálnímu zneužití byla možnost vyzkoušení zrušena. V současnosti tedy existuje pouze placená funkce a u tarifů se platí čas, kdy grafické jednotky pracují na vašich požadavcích. Jeden obrázek vytvoří bot průměrně za jednu minutu, což se základním tarifem, který nabízí 3,3 hodiny měsíčně odpovídá až 200 obrázkům. Pokud Midjourney porovnáme s DALL-E, tak dříve platilo, že Midjourney byl vhodnější na generování kreativnějších obrázků, zatímco DALL-E na realističtější obrázky. Nejnovější verze Midjourney ale deficit realističnosti úspěšně smazává a výsledky obou generátorů jsou si tak velmi podobné. [47]

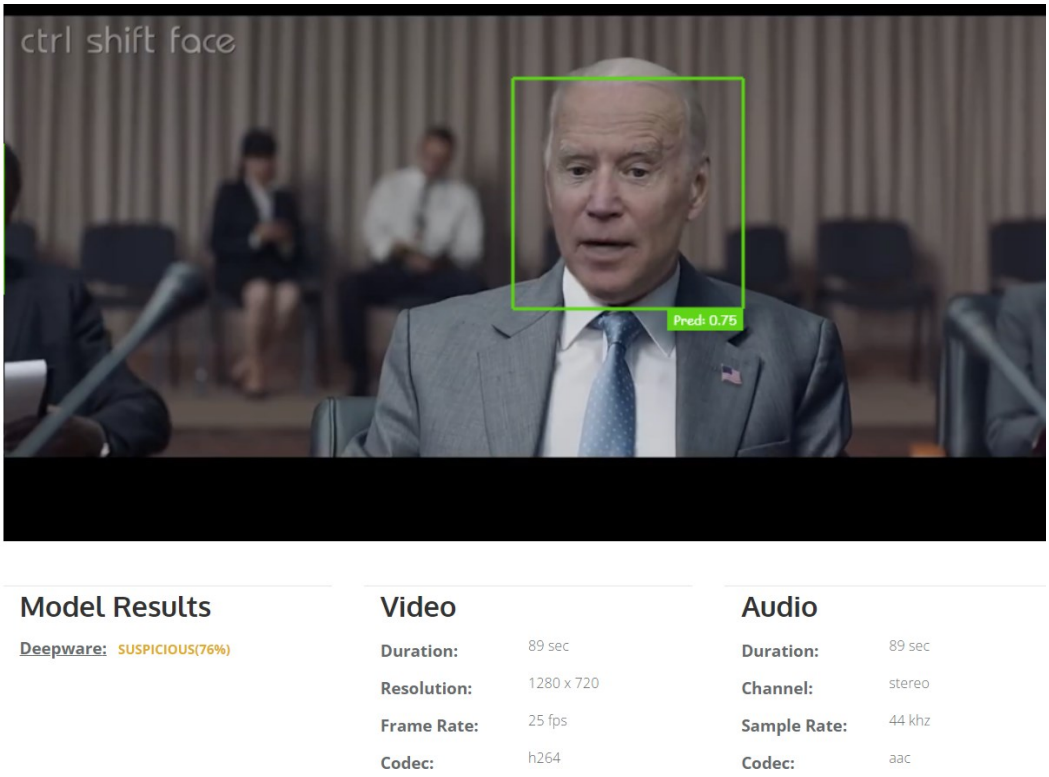
3.2 Nástroje na odhalování deepfakes

V současné době existuje několik společností nabízejících software schopný odhalovat deepfake. Deepdetector od společnosti DuckDuckGoose pracuje v reálném čase s udávanou přesností 93 % a je schopen odhalovat jak videa, tak fotografie. Prvním krokem je extrakce všech

viditelných tváří z videa či fotografie po níž následuje analýza na přítomnost deepfake prvků. Ve výsledku program určí, s jakou pravděpodobností se jedná o deepfake. Dle vývojářů se program „učil“ na stovkách a tisících autentických deepfakes a je průběžně aktualizován dle nejmodernějších detekčních metod. [48]

V listopadu 2022 přišla společnost Intel s tvrzením, že se jejímu oddělení zodpovědnému za vývoj AI aplikací podařilo vyprodukovat FakeCatcher, technologii schopnou odhalit deepfake s 96 % přesností. Technologie využívá hardware a software společnosti Intel (např. systém OpenVino) a běží na serveru, přičemž ovládání probíhá pomocí webové platformy. Platforma má být schopna odhalit deepfake v reálném čase během milisekund. FakeCatcher funguje na lehce odlišném principu než ostatní detektory. Většina detektorů založených na hlubokém učení se zaměřuje na surová data, podle nichž se snaží najít známky neautentičnosti a určit, co je na videu špatně. Naproti tomu FakeCatcher pracuje s autentickými stopami v reálných videích a zaměřuje se na to co nás dělá lidmi – jemný krevní oběh v pixelech videa. Naše žíly mění barvy podle toho, jak srdce pumpuje krev. Systém shromažďuje údaje o průtoku krve z celého obličeje a převádí je do časoprostorových map. Pomocí hlubokého učení pak dokáže téměř okamžitě určit, zda je video pravé či ne. Zůstává každopádně otázkou, jak dlouho může být takový postup účinný a zda se tento detail nedostane naopak do systémů vytvářejících deepfake což by učinilo detektory jako FakeCatcher neúčinné. Další otázkou je účinnost na extrémně komprimovaná videa na sociálních sítích ovlivněné šumem apod. [49]

V současné době bohužel není k dispozici žádný volně dostupný nástroj schopný odhalit kvalitnější deepfakes. Jednou z možností nicméně může být scan od Deepware. Tento online nástroj umožňuje zadat link s videem či video přímo nahrát a provést scan na potenciální deepfakes. Deepware Scanner je každopádně stále ve verzi beta a výsledky nejsou příliš přesvědčivé, proto je jeho účel spíše uživateli napovědět než deepfake přímo odhalit. Z vlastní zkušenosti mohu říct, že nástroj odhalí pouze jednoduché, lehce rozpoznatelné deepfakes a ty propracovanější zatím odhalit nedokáže. Na obrázku 21 níže můžeme vidět nástroj v praxi, kdy po vložení odkazu z YouTube správně vyhodnotil video jako deepfake (podezřelé ze 76 %). [50]



Obrázek 21 : Nástroj Deepware Scanner [50]

Podobným nástrojem je doplňek do prohlížeče DeepfakeProof od společnosti DuckDuckGoose údajně schopný analyzovat v reálném čase obrázky na webových stránkách a upozorňovat na potenciální deepfakes. Nástroj mi však nebyl schopný poskytnout jakékoliv výsledky a každý pokus o detekci končil bohužel hlášením o chybě.

3.3 Nástroje na odhalování pravosti a původu fotografií

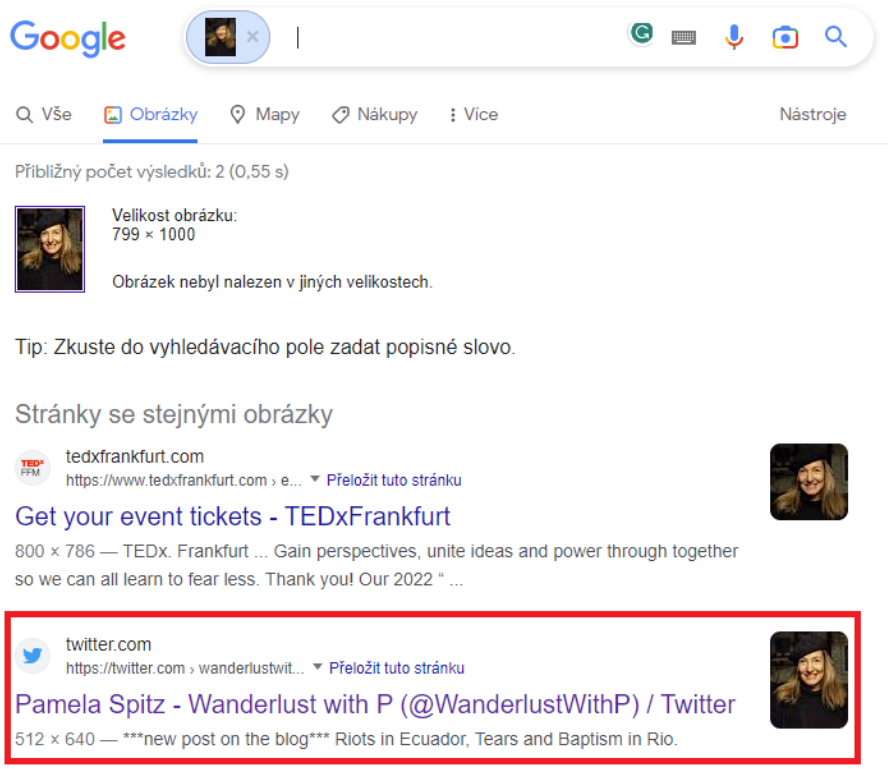
Jedním z nejjednodušších způsobů, jak si ověřit pravost fotografie je použití reverzního vyhledávače TinEye. Tato aplikace obrázek porovná s každým obrázkem v indexu a najde shody. TinEye je taktéž možno přidat do prohlížeče jako rozšíření, kdy se nám možnost vyhledat zobrazí po pravém kliknutí myši na fotografii. Možnou alternativou je pak ruský vyhledávač Yandex.

3.3.1 Google Lens

Podobně funguje taktéž užitečný nástroj od Google – Google Lens, který umí reverzně vyhledávat i konkrétní objekty z fotografií jako jsou např. hrady, zámky nebo sochy. Na základě vizuální podobnosti poté zobrazí výsledky. Praktické využití reversního vyhledávání

obrázků představuje následující příklad. Pokud se na sociální sítí setkáme s profilem, který působí podezřele viz obrázek 22 níže, není nic jednoduššího než použít Google Lens a prozkoumat výsledky. V tomto případě nás druhý odkaz – viz další obrázek 23 zavede na twitter profil, ze kterého lze zjistit, že dotyčná osoba se pravděpodobně jmenuje Pamela Spitz a ne Stadi Gbarle. Na twitterovém profilu lze pak dohledat odkaz přímo na instagramový účet, ze kterého byla fotografie odcizena.

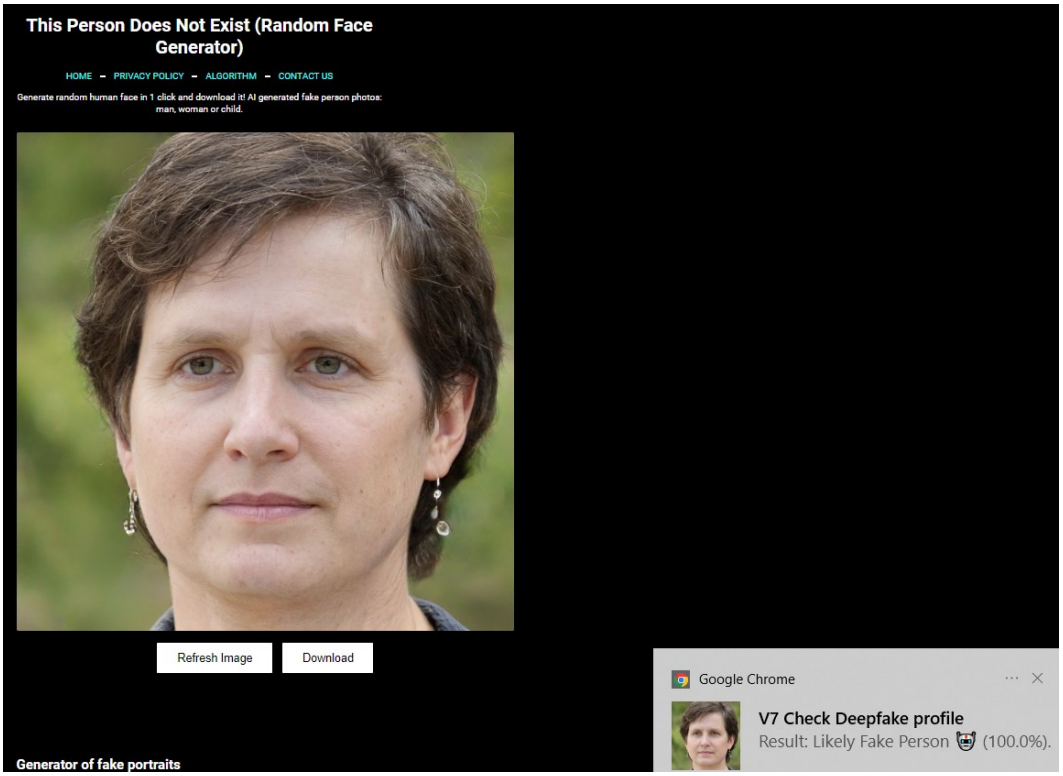
Obrázek 22 : Příklad falešného profilu



Obrázek 23 : Použití reverzního vyhledávání

3.3.2 Fake Profile Detector

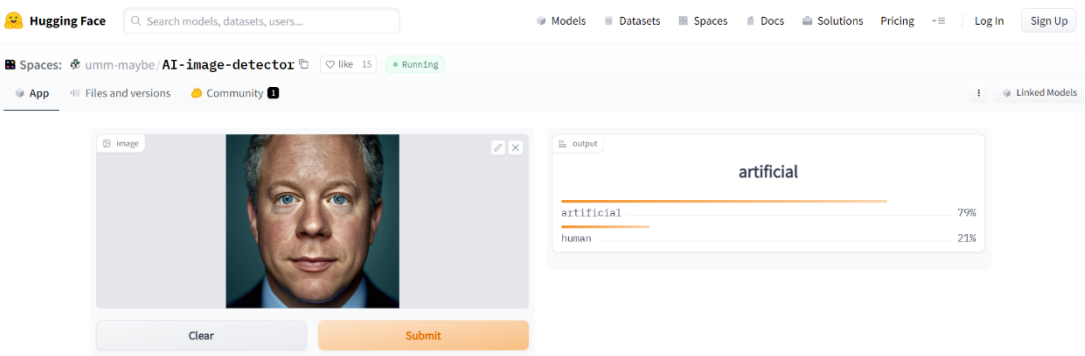
Jedním z možných způsobů, jak určit, zda se jedná o pravý či uměle vygenerovaný obličej je také nástroj od společnosti V7 – Fake profile detector, fungující jako doplněk v prohlížeči Google Chrome. Po nainstalování se při pravém kliknutí na obrázek zobrazí možnost kontroly pravosti a nástroj následně s procentuální přesností určí, zda je fotografie pravá. Na první pohled to vypadá velmi jednoduše, ovšem má to jeden velký háček a tím je samotná funkčnost tohoto doplňku. Z vlastní zkušenosti mohu potvrdit recenze, které si stěžující, že nástroj často vůbec nefunguje a pro jeho znovu zprovoznění je potřeba restartovat celý prohlížeč. Pro jednorázovou kontrolu jedné či dvou fotografií tudíž může posloužit, jinak ale bohužel nemohu doporučit. Nástroj v současné podobě taktéž nedetekuje video deepfake ani záměnu tváří. Funkčnost doplňku lze vidět níže na obrázku 24. [51]



Obrázek 24 : Použití nástroje Fake profile detector [2]

3.3.3 Hugging Face – AI image detector

Francouzsko-americká společnost Hugging Face zabývající se strojovým učením a vývojem softwaru nabízí na svých stránkách volně dostupný nástroj na detekci obrázků vytvořených AI. Snímek stačí jednoduše přetáhnout nebo nahrát ze zařízení a nástroj se následně pokusí určit původ. Na obrázku 25 níže můžeme vidět funkčnost aplikace, kdy správně určila fotografii jako uměle vytvořenou. Ne vždy ale nástroj funguje spolehlivě a je potřeba jej brát s rezervou. [52]



Obrázek 25 : AI Image detector [52]

3.3.4 Bellingcat's Online Investigation Toolkit

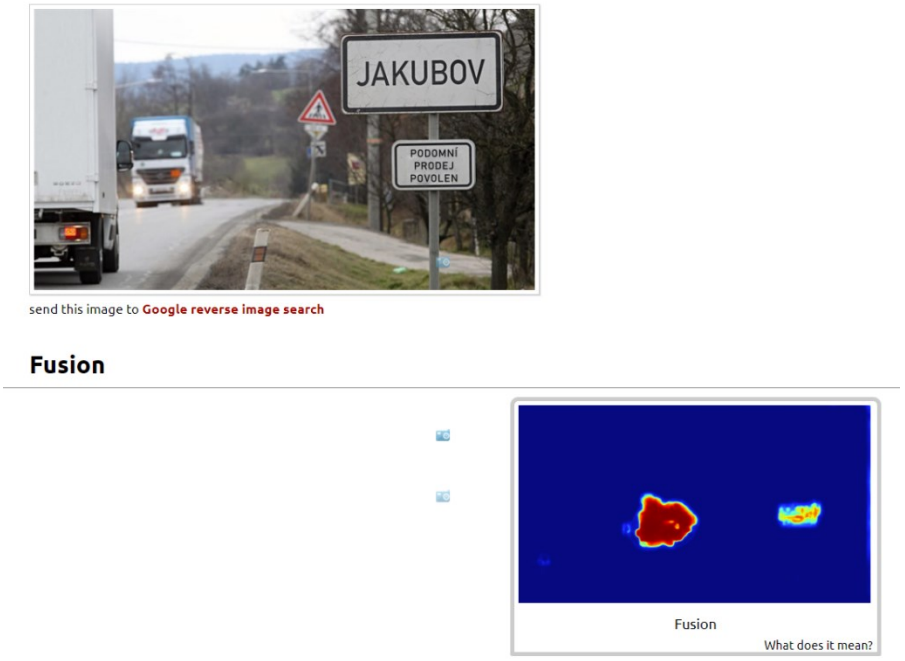
Nizozemský portál investigativní žurnalistiky Bellingcat zveřejňuje svůj seznam nejpoužívanějších volně dostupných online nástrojů na ověřování fotografií, videí, webových stránek a mnoho dalšího. Google dokument je rozdělen dle kategorií a obsahuje odkazy na desítky různých ověřovacích nástrojů téměř na vše. Na ověřování fotografií doporučuje Bellingcat například Metadata2Go či Image Verification Assistant [53].

3.3.5 Metadata2Go

Metadata2Go je jednoduchý nástroj schopný zobrazit veškerá dostupná data o konkrétní fotografii. Metadata mohou být samozřejmě změněna nebo zcela odstraněny. V takových případech nástroj zobrazí pouze minimum informací. Při použití vlastní fotografie je nástroj schopný zobrazit velké množství dat, včetně geolokace či typu fotoaparátu/mobilního telefonu, pokud tam jsou tyto data uložena. [54]

3.3.6 Image Verification Assistant

Image Verification Assistant pomůže odhalit, zda a jak bylo s fotografií manipulováno. Velmi se hodí, pokud chceme například zjistit, zda byl do fotografie přidán nějaký text či objekt. Obrázek 26 níže představuje funkci fúze, což je přístup založený na hlubokém učení, který kombinuje více forenzních filtrů a poskytuje celkovou lokalizaci. Podvržené oblasti (použití klonování, opakovaná komprese) by se měly jevit jako vysoké hodnoty (teplé barvy) na pozadí s nízkou hodnotou (studené barvy). V případě testované fotografie jsou to odstraněná auta a přepsaná značka, jak již bylo zmíněno v kapitole 2.3.2. [55]



Obrázek 26 : Nástroj Image verification assistant

3.4 Nástroje na generování hlasu

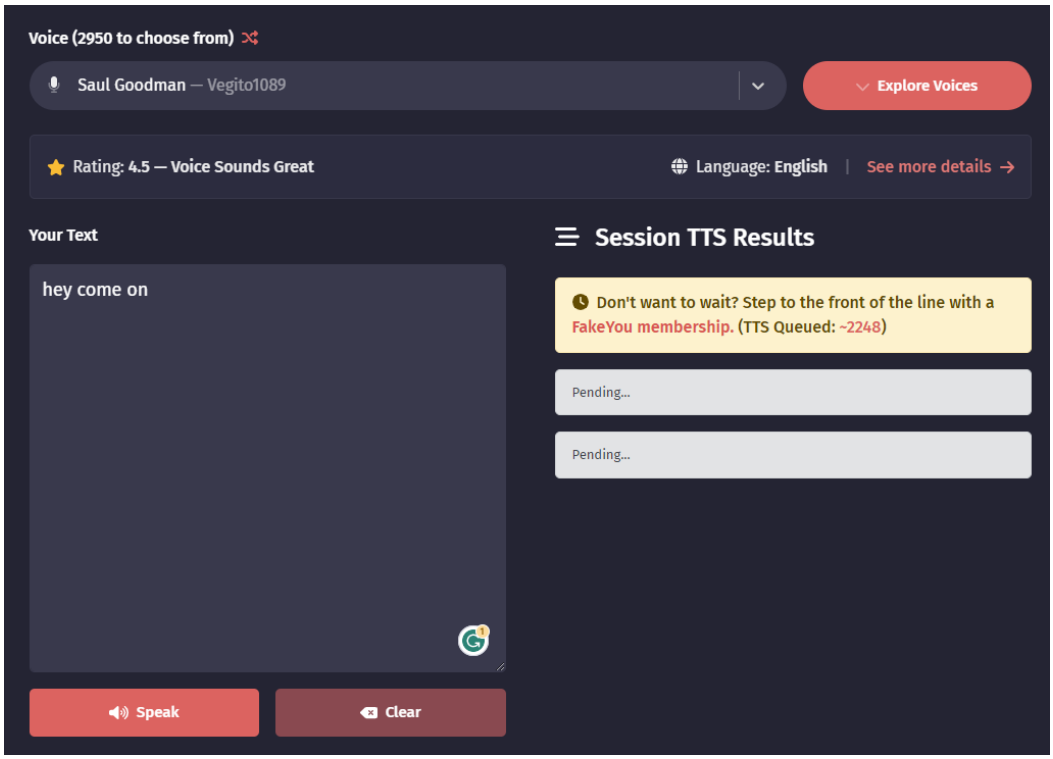
Podobně jako v případě nástrojů na tvorbu deepfakes lze na internetu nalézt množství aplikací sloužících pro záměnu hlasu. Následující výčet stručně představuje ty aktuálně nejpopulárnější.

3.4.1 FakeYou a Resemble

FakeYou je aplikace na převod textu na řeč určená k vytváření realistických zvukových klipů nejrůznějších slavných osobností a postav. Využívá deepvoice technologii ke generování audio-klipů s vybraným hlasem. Aplikace v současnosti nabízí uživatelům knihovnu obsahující téměř 3000 hlasů a taktéž filtry pro výběr podle jazyka nebo kategorie. Po výběru hlasu je zpracován uživatelský vstup a vygeneruje se nahrávka. Poté je možné si zobrazit a stáhnout výsledek. FakeYou nabízí také funkci komunity, kde je možné nahrávat zvukové klipy, generovat žebříček a sledovat kanál nejnovějších dostupných klipů. FakeYou je v základní verzi zdarma, pro vygenerování klipu je ale nutné si stoupnout „do řady“, což může trvat mnoho minut i hodin. Pro smysluplné použití je proto nutné si funkci předplatit. Jak aplikace vypadá lze vidět na následujícím obrázku 27. [56]

Resemble funguje podobně jako FakeYou, ovšem nenabízí tak širokou knihovnu hlasů, ze kterých lze vybírat a vyžaduje registraci. Na druhé straně nabízí dvě minuty audia zdarma a

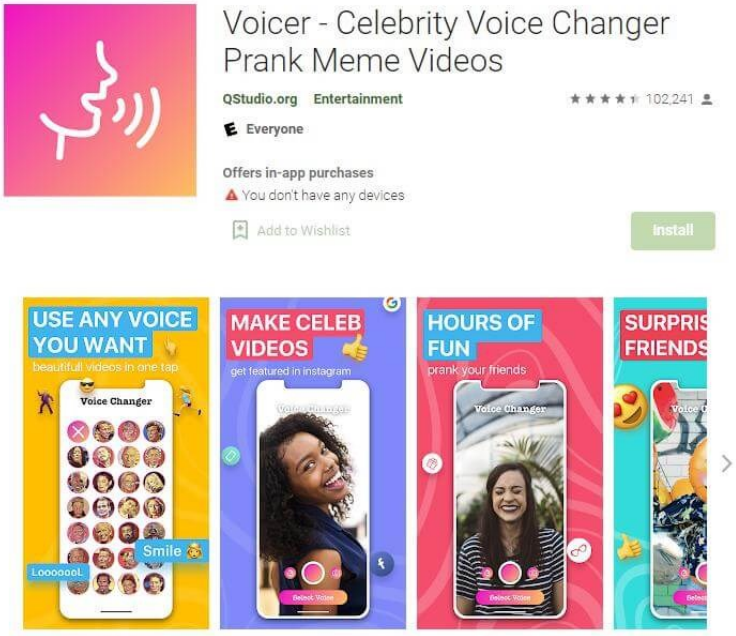
umí taktéž měnit intonaci a dodávat hlasu emoce. Služba samozřejmě umožňuje i nahrávání vlastních klipů. [57]



Obrázek 27 : Nástroj FakeYou [56]

3.4.2 Voicer Celebrity Voice Changer

Voicer Celebrity Voice Changer je aplikace pro Android a iOS umožňující měnit hlas přímo v nahrávce. Na výběr je velké množství známých osobností, případně je možné hlas pouze zrychlit nebo třeba změnit jen výšku. Nástroj pracuje rychle a jeho použití je velmi jednoduché. Na obrázku 28 je tato aplikace nabízena v obchodě s aplikacemi. [58]

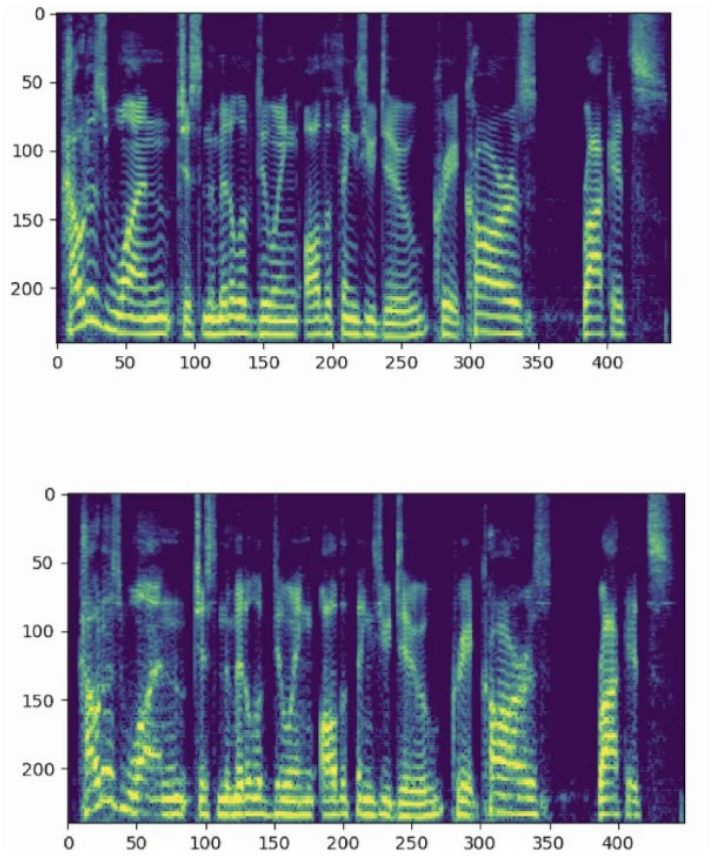


Obrázek 28 : Nástroj Voicer [58]

3.5 Nástroje na detekci falešného audia

V případě podezření na falešné audio je tu např. možnost použití nástroje AI Voice Detector, který umožňuje nahrání až 25sekundového audio klipu a následně zobrazit výsledek ukazující pravděpodobnost umělého hlasu. Pokud bychom chtěli nástroj vyzkoušet, autoři doporučují nahrávat vzorky minimálně 8 sekund dlouhé, které neobsahují hudbu ani žádné jiné zvuky v pozadí nebo šum. Na stránce je možné zobrazit i několik příkladů, kdy nástroj např. správně určil audio klip z YouTube jako falešný. Samozřejmě je opět potřeba zmínit, že volně dostupné nástroje jako je tento, nebývají nejspolehlivější a neměly bychom tedy jejich výsledky brát jako stoprocentní. [59]

Dále se nabízí např. možnost stáhnout si ze stránek github.com [60] již vytvořený model na AI audio detekci. Model pracuje s vizuálním zobrazením spektrogramu, který se u pravého a falešného audia liší – viz obrázek níže, kdy jsou modré a zelené pásy na spodním spektrogramu, který patří falešnému audiu, rozmazanější než u horního. Podle tvůrců model dosahuje více než 90% úspěšnosti detekce. Kód k modelu je navíc open-source a umožňuje tedy s modelem experimentovat nebo vytvářet vlastní. K dispozici jsou všechna předzpracovaná data, tréninkový kód a inferenční kód, aby byly věci co nejpřístupnější. Autoři modelu taktéž připojují manuál pro snadnější seznámení s kódem. [61]



Obrázek 29 : Spektogram pro Deepvoice [61]

II. PRAKTICKÁ ČÁST

4 TVORBA DEEPFAKE VIDEA

Následující praktická část práce popisuje tvorbu deepfake videa v programu DeepFaceLab. DeepFaceLab je populární volně dostupný program na záměnu tváří a tvorbu deepfake videí, dostupný na webu github.com, kde se nachází odkazy pro stažení z cloudových služeb. [62] Mezi jeho pozitiva patří především efektivita, relativně snadné ovládání, široká customizace a dostupnost zdarma. DeepFaceLab podle tvrzení vydavatele stojí až za 95 % všech vytvořených deepfakes. Mimo jiné se využívá i pro tvorbu slavných deepfake videí s Tomem Cruisem populárních na platformě Tiktok. Toma Cruise zde „hraje“ herec Miles Fisher. [62]

Pro vytváření deepfakes skrze program DeepFaceLab se doporučuje výkonný grafický procesor, například Nvidia GTX 1060 nebo lepší. Pokud chceme mít co nejlepší výsledek je ideální použít grafickou kartu RTX 3080 nebo RTX 3090. Software bude fungovat i se slabšími grafickými kartami ale trénink modelu může trvat i několik týdnů. K dispozici je několik verzí – klasická – vhodná pro všechny karty podporující DirectX12 a poté dvě verze optimalizované pro grafické karty Nvidia. Program není třeba instalovat, ovládání po rozbalení archivu probíhá skrze příkazový řádek spouštěním jednotlivých příkazů. Pro účely této práce byla použita grafická karta RTX 2060 s 6 GB VRAM pamětí a trénink sítě probíhal přibližně 24 hodin.

Existuje také možnost použít např. službu Google Colab, která umožňuje si výpočetní prostředky v závislosti na tarifu „pronajmout“.

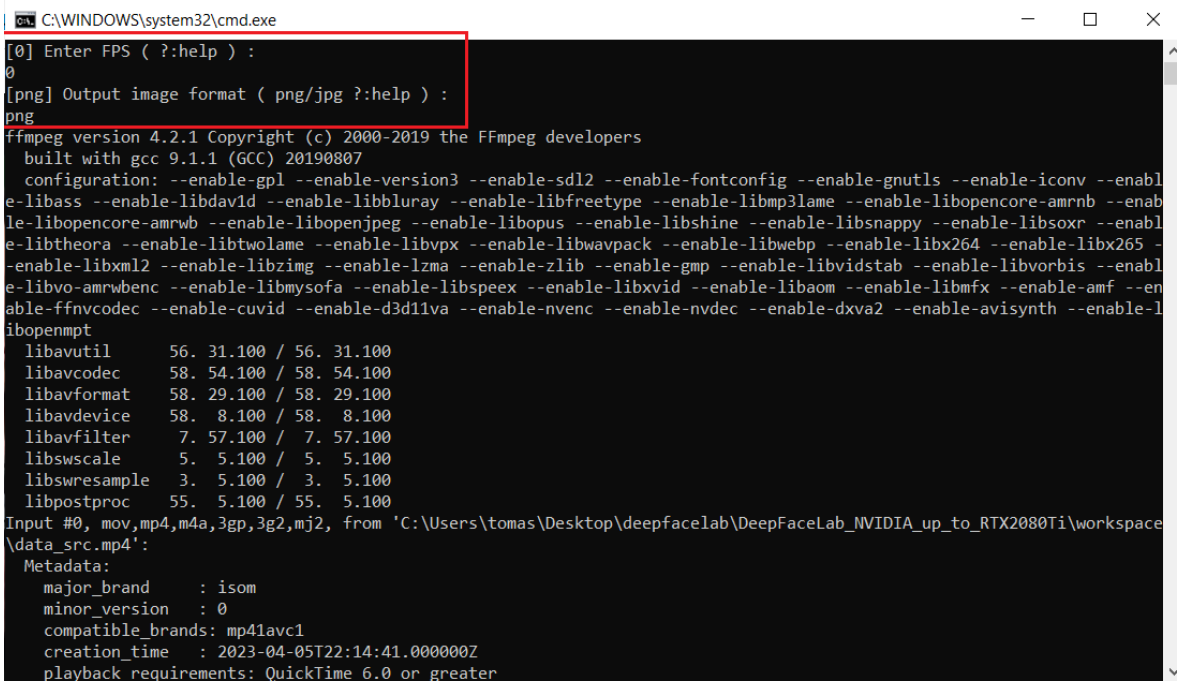
4.1 Podklady pro tvorbu deepfakes

Pro vytvoření deepfake videa v programu DeepFaceLab je nejprve potřeba určit zdrojové a cílové video. Zdrojové video je tím, ze kterého je vyjmuta tvář a cílové tím, které je upravováno tzn. tvář v cílovém videu bude nahrazena tváří ze zdrojového videa. Pokud máme videa, kde postavy výrazně pohybují hlavou je důležité, aby byly obličeje obou postav nasnímány z co největšího počtu úhlů a s potřebnou mimikou. Při hledání vhodných podkladů jsem několikrát narazil na problém, kdy např. u jednoho z videí nebyla tvář dostatečně zobrazena z pravé strany a na výsledném videu byl tento nedostatek znát. Vhodným zdrojem můžou být proto např. projevy, rozhovory nebo reportáže v některých televizních novinách kdy se osoby dívají přímo do kamery. V mém případě jsem pro ukázku a otestování zmíněného nástroje použil projev generálního tajemníka OSN Guterrese [63], jehož tvář jsem se snažil nahradit moderátorem Lawrence O’Donnellem ze stanice MSNBC. [64]

Jakmile máme obě videa nachystána, vložíme je do složky *workspace* v adresáři programu a přejmenujeme jako *data_dst* pro cílové video a *data_src* pro zdrojové video.

4.2 Proces tvorby deepfake videa

Prvním příkazem, který je nutné spustit je *extract images from video data_src*, který nám zdrojový video soubor rozdělí na jednotlivé snímky. Po spuštění se objeví okno příkazového řádku, ve kterém je potřeba zvolit několik možností. K některým možnostem se zobrazí nápověda, která je přístupná po zadání otazníku. Ten k daným možnostem poskytne další informace a užitečné typy. V případě tohoto prvního příkazu – viz obrázek níže je tu možnost výběru počtu snímků za sekundu. Výchozí možnost 0 kterou jsem zvolil znamená použít FPS daného videa. Pokud bychom měli delší video, můžeme zvolit např. 15 snímků za sekundu a z videa o 30 FPS tak extrahovat každý druhý snímek. Dále je potřeba zvolit formát extrahovaných snímků. Na výběr je ztrátový formát JPEG a bezztrátový PNG. Pro lepší kvalitu je vhodnější zvolit výchozí možnost PNG. Snímky je později možné samozřejmě smazat.



Obrázek 30 : Příkaz extract images from video data_src

Stejně se postupuje i v případě extrahování snímků z koncového videa – příkaz *extract images from video data_dst FULL FPS*. Jakmile máme extrahovány snímky z obou videí, doporučuje se ty, které nechceme použít (např. kde není obličej) ze složky workspace vymazat.

Extrahování tváří

Dále jsem pokračoval extrahováním tváří zdrojového videa příkazem *data_src faceset extract*. Zde se nastavuje podstatně více věcí. Důležité je nastavení Face type, které definuje, kterou část obličeje budeme zpracovávat. Přehled nabízí následující tabulka.

Face type	Popis
head	Celá hlava od vlasů až po krk
wf	Celá tvář. Od vrchní části hlavy až po bradu
f	Celá tvář od čela po bradu.
mf	Střední tvář – obočí po bradu.
hf	Poloviční obličej – od očí po ústa

Tabulka 1: Přehled možností pro zpracování deepfake obličeje

Pro vyvážený výsledek jsem zvolil wf - tedy celou tvář od vrchní části hlavy až po bradu.

Parametr *max number of faces* se používá pokud bychom v podkladech měli více tváří. V mém případě je ve zdrojovém i cílovém videu pouze jedna tvář – nechávám tedy výchozí hodnotu. Nálsledující parametry *Image size* nebo *Jpeg quality* taktéž nechávám na výchozí hodnotě – pro ukázkové video jsou dostačující. Větší kvalita znamená větší velikost výstupní složky. Všechny možnosti nastavení jsou potom patrné na obrázku níže.

Obrázek 31 : Příkaz pro extrahováním tváří zdrojového videa

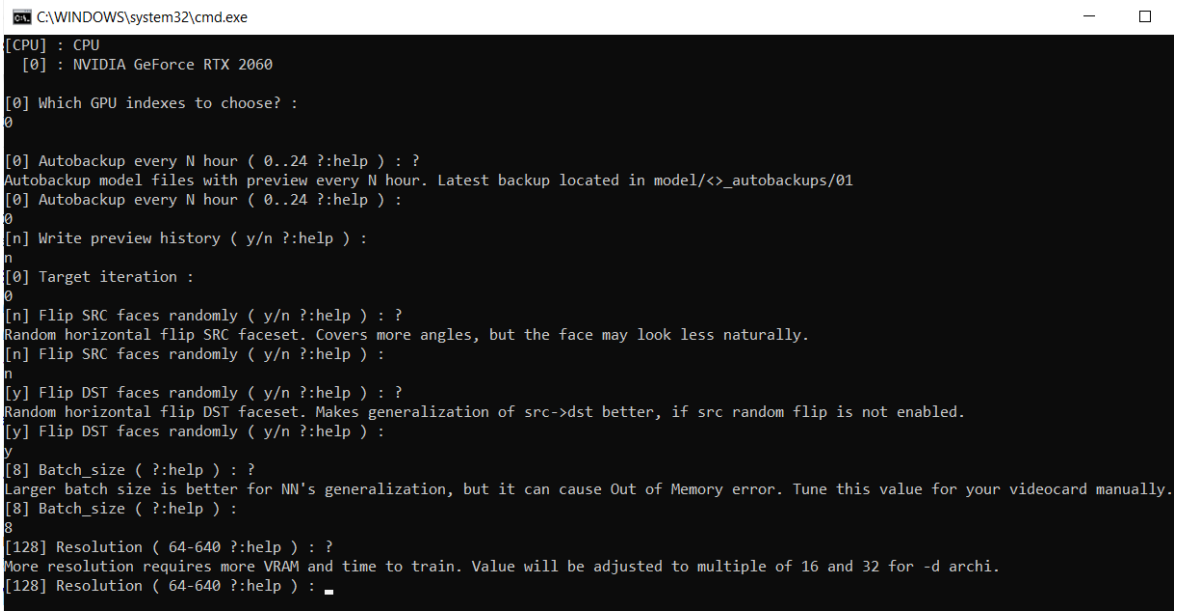
Stejně postupujeme i při extrahování tváří z koncového videa příkazem *data_dst faceset extract*.

Trénování sítě

Pro samotné trénování sítě nabízí DeepFaceLab tyto tři modely:

- **SAEHD** (6 GB+): High Definition Styled Auto Encoder – pro výkonné GPU s alespoň 6 GB VRAM. Nastavitelný. Doporučuje se pro většinu uživatelů.
- **AMP** (6 GB+): Pokročilejší typ modelu, který je zatím stále ve vývoji.
- **Quick96** (2-4 GB): Jednoduchý režim určený pro GPU s nižším výkonem s 2-4 GB VRAM. Má pevně nastavené parametry, používá se primárně pro rychlé testování.

Pro tvorbu ukázkového videa jsem dále použil model SAEHD a opět nastavil několik parametrů – viz obrázek níže.



Obrázek 32 : Nastavení parametrů modelu SAEHD

Pojďme si nyní několik z nich podrobněji vysvětlit.

Autobackup every N hour – umožňuje zapnout automatické zálohování modelu každých N hodin. V mém případě vzhledem k relativně krátkému trénování sítě jsem parametr nechal na výchozí hodnotě – tedy vypnutý. Případnou zálohu bychom našli ve složce programu.

Write preview history – umožňuje zapnout automatické ukládání obrázků s průběžnými výsledky hlubokého učení.

Target iteration – cílový počet iterací není potřeba nastavovat, proces učení lze jednoduše ukončit ve chvíli, kdy to uznáme za vhodné. Pokud bychom chtěli síť natrénovat na přesně 100 000 iterací zadáme hodnotu 100 000.

Flip SRC faces randomly – náhodně horizontálně převrací zdrojové tváře a kopíruje rysy obličeje z jedné strany na druhou, čímž pomáhá pokrýt všechny úhly cílového obličeje. V mnoha případech toto ovšem bude působit nepřirozeně jelikož lidský obličej není perfektně symetrický. Pokud máme zdrojový obličej dostatečně pokrytý, je lepší funkci nechat vypnutou

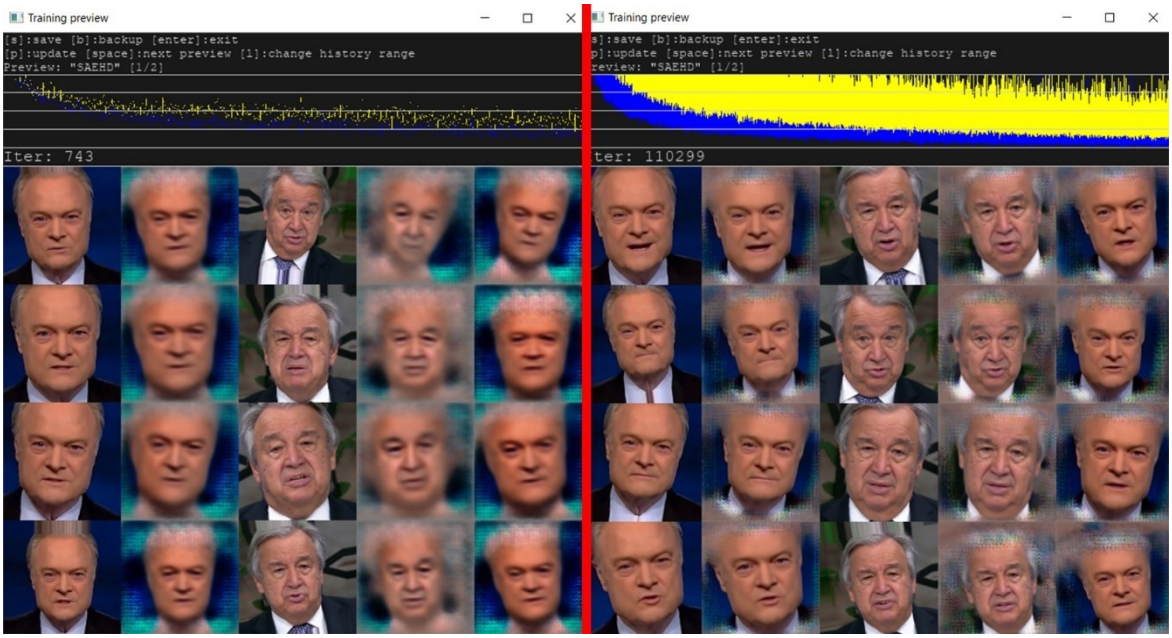
Flip DST faces randomly – náhodně převrací horizontálně koncové tváře. Zlepšuje generalizaci v případě, že je vypnutý předešlý parametr *Flip SRC faces randomly*. Výchozí hodnota je zapnuto.

Batch_size – ovlivňuje kolik obličejů se bude navzájem porovnávat při každé iteraci. Nastavuje se dle výkonu grafické karty. Nejnižší možná hodnota je 2, ale hodnoty menší než 4

není vhodné používat. Optimální hodnota je někde mezi 6-12. Výchozí hodnota 8 byla v mém případě ideální. Samozřejmě čím větší hodnotu zadáme tím bude proces pomalejší a stejně tak můžeme narazit na problémy s nedostatkem paměti VRAM.

Rozlišeni taktéž nechávám na výchozí hodnotě, jelikož zvýšení by vyžadovalo více VRAM paměti a času a použitá grafická karta RTX 2060 není v tomto ohledu z nejvýkonnějších.

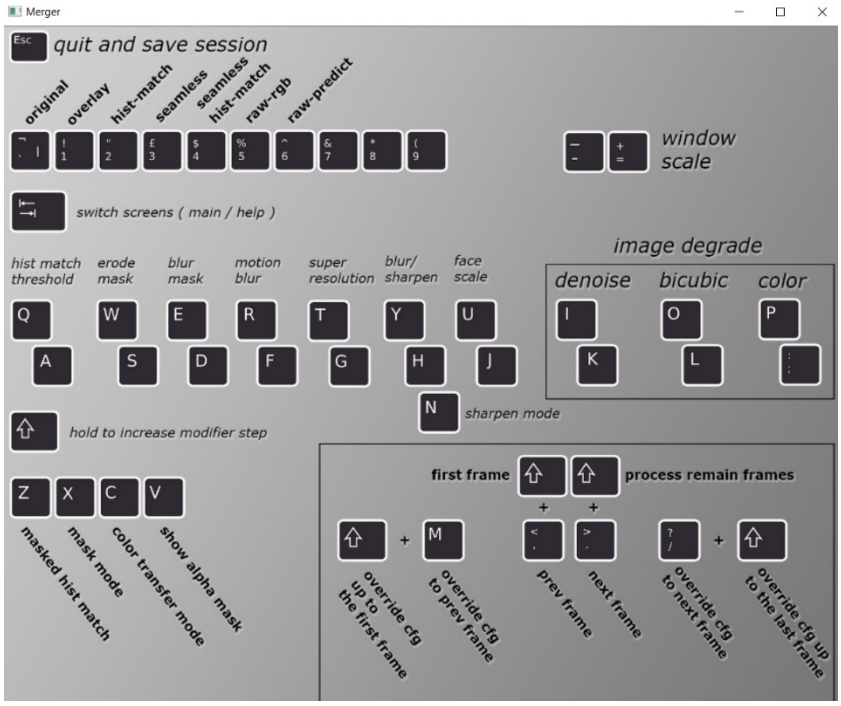
Po zadání všech parametrů se otevře nové okno s průběžnými výsledky hlubokého učení viz obrázek níže. Při trénování sítě sledujeme počet iterací, tedy provádění stejného procesu nebo postupu s cílem dosáhnout požadovaného výsledku. V tomto případě dochází k opakovanému zpracování vstupních dat a úpravě vah sítě, aby se dosáhlo lepší přesnosti výstupu. Pro optimální výsledek by iterací mělo proběhnout asi 100 až 150 tisíc, ale jednoduché pravidlo zní, nechat proces učení běžet tak dlouho dokud nejsme spokojeni s výsledkem. Nicméně video vytvořené pouze pro pobavení může vypadat poměrně dobře už i po 40-50 tisících iterací. V mém případě jsem se snažil dosáhnout minimálního počtu 100 tisíc iterací. Následující obrázek představuje porovnání na začátku procesu hlubokého učení a stav po přibližně 110 tisících iterací. Zajímá nás hlavně poslední sloupec, který zobrazuje průběžné výsledky.



Obrázek 33 : Průběžné výsledky trénování sítě

4.3 Úprava deepfake videa

Jakmile je proces učení dokončen – tzn. dosáhli jsme vytyčeného počtu iterací či jsme spokojeni s aktuálním náhledem, následuje úprava videa skrze interaktivní fúzi (tzv. merger). Spuštěním Příkazu *merge SAEHD* si můžeme zvolit interaktivní merger, který nám umožní deepfake video „doladit“. Tento *merger* umožňuje měnit několik parametrů viz obrázek níže. Pojďme si projít alespoň několik z nich.



Obrázek 34 : Interaktivní merger

V mém případě jsem např. použil efekt *erode mask* – tedy erodování tváře, které se hodí použít, pokud má trénovaný obličej ošklivé okraje nebo digitální artefakty v blízkosti vnější oblasti obličeje. V podstatě se jedná o ořezaní okrajů, ovšem je potřeba si dát pozor abychom trénovaný obličej neořezali příliš, což by odkrylo původní tvář. Následující obrázek představuje názorné použití efektu *erode mask*, který je v mém případě nastaven na hodnotu 5. Můžeme si všimnout že u pravé tváře lze již lehce vidět obočí z původního videa.

Obrázek 35 : Efekt erodování tváře

Dalším užitečným efektem je blur mask (maska rozostření). Zvýšením hodnoty můžeme efektivně skrýt přechod mezi původním videem a deepfake obličejem. Hodnotu jsem zvýšil tak, aby byla hrana co nejméně viditelná viz obrázek níže. Opět je potřeba dát si pozor, aby hodnota nebyla příliš vysoká, čímž by se odkryl původní obličej.

Obrázek 36 : Efekt masky rozostření

Blur/Sharpen (Rozmazání/Vyostření) nabízí možnost jednoduše rozmazat nebo vyostřit obrázek, tuto funkci jsem ovšem nepoužil, jelikož je výsledný rozdíl minimální.

Funkce Motion Blur může udělat deepfake obličej realističtější umělým použitím pohybového rozostření. Zpracování ovšem zabere více času a pokud nemáme zkušenosti např. s programem After-Effects je lepší tuto funkci nepoužívat.

Měřítko obličeje (Face scale) umožňuje deepfake obličej škálovat tak aby lépe „seděl“ na původní video. K funkci je třeba přistupovat opatrně, nevhodným škálováním můžeme video „rozbít“.

Pokud jsme s úpravami spokojeni klávesami Shift +? a následně Shift +> aplikujeme změny na všechny snímky videa. Tím je proces úprav dokončen.

Obrázek níže představuje porovnání původních tváří, které byly k tvorbě použity a výsledného obrazu deepfake videa. Toto výsledné video je dostupné na serveru Youtube pod odkazem. [65]

Obrázek 37 : Porovnání videí před a po

4.4 Zhodnocení vytvořeného videa a porovnání verzí

Celkově můžu říct, že tvorba jednoduššího deepfake videa v programu DeepFaceLab není příliš náročná. Problém může být spíše v ovládání, respektive spouštění jednotlivých příkazů, kdy je nutné si nejprve projít tutoriály. Jak již bylo zmíněno, program neobsahuje snad kromě možnosti zvolit si *interaktivní merger* žádné uživatelské rozhraní. Tvorbu bych dále shrnul v několika bodech níže.

Verze

Během tvorby deepfake videa jsem zkusil vytvořit několik verzí, ovšem např. mezi verzí trénovanou na přibližně 50 000 a verzí trénovanou na 110 000 iterací (finální) jsem na první pohled nezaznamenal výrazné rozdíly v kvalitě. Zde bychom pravděpodobně museli změnit jiné parametry, např. rozlišení. Jediným výraznějším rozdílem bylo mrkání kdy verze o přibližně 50 000 iteracích téměř nemrkala – viz obrázek níže. U finálního videa trénovaného právě na cca 110 000 iterací působí mrkání přesvědčivěji. Objem dat byl u zdrojového videa přibližně 3 minuty záznamu a program z něj extrahoval 5451 snímků tváři, u cílového 1 minuta záznamu, z kterého bylo extrahováno 3565 snímků tváří.

Obrázek 38 : Příklad porovnání dvou verzí deepfake videa

Vhodné podklady

Důležitý je výběr vhodných podkladů – tedy zdrojového a cílového videa. Na obrázku 39 níže je pro porovnání jedno z prvních testovacích deepfake videí, které jsem během seznamování se s programem zkoušel vytvořit. Na levém (falešném) obličeji je na pravé straně

vidět výrazný přechod z původní tváře na falešnou. Toto je způsobeno nedostatečným nasníman̈ím tváře zdrojového videa z pravé strany. U finálního videa se obličej dívá většinu času přímo do kamery, což je pro deepfake video ideální předpoklad. Stejně tak je třeba dávat si pozor např. na brýle, pokud bychom měli obličej s brýlemi je třeba nechat běžet proces učení dostatečně dlouho.

Obrázek 39 : Testovací deepfake video

Mohlo by výsledné video někoho oklamat?

Osobně si nemyslím, že by finální video bylo nějak zvláště přesvědčivé nebo snad dokázalo někoho oklamat, nicméně je vhodnou ukázkou, jak tato technologie může vypadat a jak takové video vzniká. Přesvědčivějších výsledků by bylo možné dosáhnout zvýšením některých parametrů nástroje a delším procesem hlubokého učení, čímž se ovšem pochopitelně zvyšují nároky na čas a výkon grafické karty. Důvěryhodnost videa by taktéž nepochybně zvýšila kombinace deepfake a deepvoice – tedy záměna obrazové a hlasové identity zároveň.

5 JAK ODHALIT DEEPFAKE VIDEO

V závěrečné části práce bych chtěl na několika konkrétních deepfake videích, které lze v současnosti na internetu nalézt, naznačit na co je vhodné se při posuzování záznamu zaměřit. Následující obrázkové ukázky z videí slouží jako vhodné příklady, na kterých lze demonstrovat možné nedokonalosti deepfakes a způsoby jejich odhalení.

Pokožka a pohled očí

Deepfake videa často zobrazují pokožku jako příliš hladkou bez jakýchkoliv vrásek s nepřirozenými stíny v okolí obočí. Právě nedokonalé nasvícení a chybějící stíny mohou být klíčové. Stejně tak je vhodné zaměřit se na pohyby očí a nepřirozené či zcela chybějící mrkání.

V případě obrázku níže je určujícím prvkem u deepfake tváře (levá) příliš hladká pokožka na čele v kombinaci s absencí vrásek a skvrn. U zmíněné deepfake tváře si můžeme všimnout taktéž nedokonalosti pravého oka, které se nedívá přímo na posluchače jako je tomu u skutečné (pravé) tváře.

Obrázek 40 : Deepfake – detaily očí a pokožky [66]

Detaily úst, jazyka a zubů

Některých detailů si na první pohled nemusíme všimnout, ovšem jsou to právě drobnosti, na kterých deepfake videa často stojí a padají. Pohyby úst, detaily zubů, jazyka apod. V případě

obrázku níže si lze všimnout výrazného rozdílu v kvalitě zpracování zubů mezi deepfake tváří vlevo a původní tváří vpravo.

Obrázek 41 : Deepfake – detaily úst, jazyka a zubů [67]

Grafické anomálie

U některých deepfake videí se také mohou vyskytovat výraznější grafické anomálie, které nám okamžitě napoví, že video může být zmanipulované. Jedná se nejčastěji o problikávání, zdvojení obrazu, rozmazání obrazu či např. chybějící část ucha jako je tomu v případě následujícího obrázku. Vhodnou pomůckou může také být si video zpomalit. Při výrazném zpomalení mohou být podobné grafické anomálie a ostatní nedokonalosti více znatelné.

Obrázek 42 : Deepfake – grafické anomálie [68]

Brýle, náušnice, piercing

Jakékoliv doplňky mohou být pro algoritmy hlubokého učení problém. Pokud osoba ve videu, které nám přijde podezřelé nosí např. brýle, měli bychom se na toto zaměřit jako první. Na obrázku níže u levé deepfake tváře v určitý moment brýle zcela zmizí.

Obrázek 43 : Deepfake – brýle a doplňky [69]

U některých deepfake videí je taktéž často vidět hrana určující přechod mezi skutečnou a falešnou tváří. Na obrázku níže je na tváři nad obočím znatelný barevný přechod, který vyznačuje místo, kde splývá původní a falešná tvář.

Obrázek 44 : Deepfake – přechod mezi skutečnou a falešnou tváří [70]

Shrnutí

Na těchto příkladech deepfake videí jsme mohli vidět, jak zásadní mohou být detaily pro věrohodnost výsledného videa, ať už se jedná o oči, ústa nebo např. brýle. Pokud se nesetkáme s opravdu profesionálním videem, téměř vždy platí, že najdeme několik nesrovnalostí. Stejně tak nesmíme zapomínat na kontext a informace si vždy ověřovat z více zdrojů. V dnešní době jsou deepfake videa naštěstí stále spíše nástrojem pro zábavu než vážnou hrozbou, nicméně je důležité mít na paměti, že deepfake technologie se stále rychleji vyvíjí a dokáže vytvářet stále sofistikovanější a méně detekovatelná falešná videa. Dá se proto předpokládat, že v budoucnu bude použití této technologie pro účely manipulace častější a pokud se dostaneme do bodu kdy bude průměrné deepfake video pro člověka nerozlišitelné, nezbude nám než se spoléhat na detekci pomocí umělé inteligence.

ZÁVĚR

Cílem této bakalářské práce bylo provést literární rešerši na téma související s audiovizuální manipulací a představit nejčastěji používané nástroje a metody, které se k manipulaci využívají. Cílem praktické části pak bylo zkusit vytvořit deepfake video, popsat princip jeho tvorby a diskutovat dosažené výsledky. Dále také na vhodných vzorcích demonstrovat způsoby odhalení pravosti záznamu.

Pro literární rešerši v kapitole 1 byly vybrány co nejaktuálnější články a publikace zabývající se tématem včetně zmínění několika závěrečných prací na podobné téma, které jsem se snažil v této práci doplnit novými informacemi a pohledy. Kapitola 2 se zaobírala nejčastějšími typy audiovizuální manipulace a objasněním jejich principů – zejména fenoménem deepfake a hlasovými generátory. V případě manipulace s fotografiemi byla práce zaměřena na generátory tváří a fotografií, které v posledních letech a měsících výrazně pokročily. Zde bych opět upozornil na případy zmíněné v kapitole 2.4.1, které svou realističností představují nebezpečný precedens do budoucna, kdy bude podobných obrázků dozajista přibývat. Na druhou stranu mě hledání konkrétních příkladů při vypracování této práce utvrdilo v přesvědčení, že se v současnosti stále ještě nenacházíme za bodem, kdy by nebylo možné rozeznat deepfake od skutečného videa či vygenerovanou fotografii od pravé. Při důkladném prozkoumání jsem vždy našel něco, co mi napovědělo, že se nejedná o pravé video či fotografii. V průběhu psaní práce jsem zároveň jsem zjistil že rozpoznávání deepfake videí či generovaných fotografií je možné se do jisté míry „naučit“. V vlastní zkušenosti mohu říct, že pokud vygenerovaných fotografií či deepfake videí prozkoumáme desítky či stovky naučíme se více si všímat konkrétních detailů.

Ve kapitole 3 popisující softwarové nástroje byly zmíněny převážně nástroje dostupné zdarma, které jsou v současnosti nejpopulárnější. Určitým zklamáním pro mě byla nedostupnost prakticky použitelných nástrojů pro alespoň základní detekci deepfake videí a falešných fotografií. Dostupné nástroje při testování neposkytovaly požadované výsledky.

Při tvorbě deepfake videa jsem byl do určité míry překvapen, jak jednoduché může být vytvořit základní deepfake video, pokud máme k dispozici vhodný podklad, se kterým můžeme pracovat. Taková video ale nejspíš nikoho neošálí a vytvořit kvalitní deepfake video rozhodně není jednoduché, a naopak vyžaduje spoustu času, znalostí a vhodný software a hardware.

Nakonec bych ještě dodal, že naučit se odhalovat deepfake videa a falešné fotografie nám pomáhá chránit své soukromí a osobní integritu v digitálním prostředí. Budujeme tak kritické myšlení, zdravou skepsi a schopnost rozpoznat manipulaci. Tato práce proto může do jisté míry posloužit jako určitý návod a doporučení, jak se vyvarovat manipulaci, jak s ní pracovat a jak se ji pokusit odhalit. Protože jak jsem zjistil během psaní této práce, stále existuje spousta lidí kteří např. o existenci deepfake nic nevědí, a to i mezi těmi v mladším věku.

SEZNAM POUŽITÉ LITERATURY

[1] Tenhle člověk neexistuje – generativní modely pro zelenáče. Tomaskubica.cz [online]. 11. 7. 2022 [cit. 2023-02-24]. Dostupné z: https://www.tomaskubica.cz/post/2022/tenhle-clovek-neexistuje-generativni-modely-pro.zelenace/

[2] This Person Does Not Exist (Random Face Generator). This-person-does-not-exist.com [online]. [cit. 2023-02-24]. Dostupné z: https://this-person-does-not-exist.com/en

[3] MCDONALD, Kyle. How to recognize fake AI-generated images. Kcimc.medium.com [online]. 5.12.2018 [cit. 2023-02-24]. Dostupné z: https://kcimc.medium.com/how-to-recognize-fake-ai-generated-images-4d1f6f9a2842

[4] ALEX, McFarland. 10 Best AI Random Face Generators (January 2023). Unite.ai [online]. 26.1.2023 [cit. 2023-02-24]. Dostupné z: https://www.unite.ai/random-face-generators/

[5] NIMMO, Ben a David AGRANOVICH. Recapping Our 2022 Coordinated Inauthentic Behavior Enforcements. About.fb.com [online]. 15.12.2022 [cit. 2023-02-24]. Dostupné z: https://about.fb.com/news/2022/12/metas-2022-coordinated-inauthentic-behavior-enforcements/

[6] ZAHRÁDKA, Jiří. Ohromují, manipulují, ale konspirátorům nevoní. Deepfake videa mají problém. Finmag.penize.cz [online]. 27. 9. 2021 [cit. 2023-02-24]. Dostupné z: https://finmag.penize.cz/veda-a-technika/429028-ohromuji-manipuluji-ale-konspiratorum-nevoni-deepfake-videa-maji-problem

[7] MUDROVÁ, Nikol. Fenomen deepfakes jako hrozba žurnalistiky. Praha, 2020. Bakalářská práce. Univerzita Karlova, Fakulta sociálních věd, Katedra žurnalisti-ky. Vedoucí práce Láb, Filip. [online]. [cit. 2023-02-24]. Dostupné z: https://dspace.cuni.cz/handle/20.500.11956/118375

[8] SCHREINER, Maximilian. Deepfakes: How it all began – and where it could lead us. The-decoder.com [online]. 28.4.2022 [cit. 2023-02-24]. Dostupné z: https://the-decoder.com/history-of-deepfakes/

[9] How To Protect Against Deepfakes – Statistics and Solutions. Iproov.com [online]. 26.8.2022 [cit. 2023-02-24]. Dostupné z: https://www.iproov.com/blog/deepfakes-statistics-solutions-biometric-protection

[10] BORISOVÁ, Mariana. Softwarové nástroje pro ověřování pravosti digitálních fotografií. Zlín: Univerzita Tomáše Bati ve Zlíně, 2020, 70 s. (61113). Dostupné také z: http://hdl.handle.net/10563/48021. Univerzita Tomáše Bati ve Zlíně. Fakulta aplikované informatiky, Ústav počítačových a komunikačních systémů. Vedoucí práce Gazdoš, František.

[11] MCFARLAND, Alex. 10 Best "Text to Speech" Generators (February 2023). Unite.ai [online]. 3.2.2023 [cit. 2023-02-24]. Dostupné z: https://www.unite.ai/best-text-to-speech-generators/

[12] SALIA, Rose. Top 10 Realistic Text to Speech Tools Review 2023. Topten.ai [online]. 27.3.2021 [cit. 2023-02-24]. Dostupné z: https://topten.ai/best-realistic-text-to-speech-tools-review/

[13] Jirka vysvětluje věci. Youtube.com [online]. [cit. 2023-02-24]. Dostupné z: https://www.youtube.com/@Jirkavysvetlujeveci

[14] KOŘENÁŘ, Patrik. HROZBA JMÉNEM DEEPFAKE. Youtube.com [online]. 19. 11. 2019 [cit. 2023-02-24]. Dostupné z: https://www.youtube.com/watch?v=RP9bKAHaaAo&t=389s&ab_channel=PatrikKo%C5%99en%C3%A1%C5%99

[15] KASÍK, Pavel. Naléhavá výzva hvězd IT: Zastavte vývoj umělé inteligence, jde moc rychle. Seznamzpravy.cz [online]. 29.3.2023 [cit. 2023-03-31]. Dostupné z: https://www.seznamzpravy.cz/clanek/tech-technologie-zastavte-vyvoj-umele-inteligence-volaji-odbornici-a-musk-lidstvo-to-nestiha-228605

[16] KOPECKÝ, Kamil a René SZOTKOWSKI. Dezinformace a fake news. Pdf.upol.cz [online]. 2019 [cit. 2023-02-24]. Dostupné z: https://www.pdf.upol.cz/fileadmin/userdata/PdF/VaV/2019/odborne_seminare/Kopecky_Dezinformace_a_Fake_News.pdf

[17] Co je to manipulace. Apas.cz [online]. [cit. 2023-02-24]. Dostupné z: https://apas.cz/slovnicek-pojmu/manipulace/

[18] How false information spreads. Bbc.co.uk [online]. [cit. 2023-02-24]. Do-stupné z: https://www.bbc.co.uk/bitesize/articles/zcr8r2p

[19] ŠLERKA, Josef. Josef Šlerka: Deep fake videa vám nedávají žádnou šanci poznat, že vám lžou. Dvojka.rozhlas.cz [online]. 24.7.2020 [cit. 2023-02-24]. Do-

stupné z: https://dvojka.rozhlas.cz/josef-slerka-deep-fake-videa-vam-nedavaji-zadnou-sanci-poznat-ze-vam-lzou-8257113

[20] Tři hrozby Deepfake videí a možné řešení proti nim. Internetembezpecne.cz [online]. 14.11.2019 [cit. 2023-02-24]. Dostupné z: https://www.internetembezpecne.cz/deep-fakes/

[21] LYU, Siwei. Deepfakes and the New AI-Generated Fake Media Creation-Detection Arms Race. Scientificamerican.com [online]. 20.7. 2020 [cit. 2023-03-31]. Dostupné z: https://www.scientificamerican.com/article/detecting-deepfakes1/

[22] EVON, Dan. Bad Deepfake of Zelenskyy Shared on Ukraine News Site in Reported Hack. Snopes.com [online]. 16.3.2022 [cit. 2023-02-24]. Dostupné z: https://www.snopes.com/news/2022/03/16/zelenskyy-deepfake-shared/

[23] COFFEE, Patrick. 'Deepfakes' of Celebrities Have Begun Appearing in Ads, With or Without Their Permission. Wsj.com [online]. 25.10.2022 [cit. 2023-02-24]. Dostupné z: https://www.wsj.com/articles/deepfakes-of-celebrities-have-begun-appearing-in-ads-with-or-without-their-permission-11666692003

[24] DONEGAN, Moira. Demand for deepfake pornography is exploding. We aren't ready for this assault on consent. Theguardian.com [online]. 13.3.2023 [cit. 2023-03-30]. Dostupné z: https://www.theguardian.com/commentisfree/2023/mar/13/deepfake-pornography-explosion

[25] Diep Nep. Youtube.com [online]. [cit. 2023-02-24]. Dostupné z: https://www.youtube.com/@DiepNep

[26] Jim Carrey DeepFake [VFX Comparison]. Youtube.com [online]. 3. 9. 2019 [cit.2023-02-24]. Dostupné z: https://www.youtube.com/watch?v=JbzVhzNaTdI&ab_channel=CtrlShiftFace

[27] Doctored Nancy Pelosi video highlights threat of "deepfake" tech. Cbsnews.com [online]. 26.5.2019 [cit. 2023-05-16]. Dostupné z: https://www.cbsnews.com/news/doctored-nancy-pelosi-video-highlights-threat-of-deepfake-tech-2019-05-25/

[28] ARATANI, Lauren. Altered Video Of CNN Reporter Jim Acosta Heralds A Future Filled With 'Deep Fakes'. Forbes.com [online]. 8.11.2018 [cit. 2023-05-16].

Dostupné z: https://www.forbes.com/sites/laurenaratani/2018/11/08/altered-video-of-cnn-reporter-jim-acosta-heralds-a-future-filled-with-deep-fakes/

[29] What is a neural network? Ibm.com [online]. [cit. 2023-05-15]. Dostupné z: https://www.ibm.com/topics/neural-networks

[30] NGUYEN, Thanh Thi a Quoc Viet Hung NGUYEN. Deep Learning for Deepfakes Creation and Detection: A Survey. In: Arxiv.org [online]. [cit. 2023-05-16]. Dostupné z: https://arxiv.org/pdf/1909.11573.pdf

[31] MÍSAŘ, Jan. Historie fotografických úprav. Fel.cvut.cz [online]. 6.6.2014 [cit. 2023-02-24]. Dostupné z: https://cw.fel.cvut.cz/b212/_media/courses/a7b33dif/sin-slavy/historie_fotografickych_uprav.pdf

[32] Unreal Person, This person does not exist. Unrealperson.com [online]. [cit. 2023-02-24]. Dostupné z: https://www.unrealperson.com/

[33] J. NIGHTINGALE, Sophie. AI-synthesized faces are indistinguishable from real faces and more trustworthy. Pnas.org [online]. 14.2.2022 [cit. 2023-02-24]. Dostupné z: https://www.pnas.org/doi/10.1073/pnas.2120481119

[34] Exposed: AI-Generated Viral Photo of Putin Bowing to Xi Jinping - Fact or Fiction? Timesnownews.com [online]. 22.3.2023 [cit. 2023-05-16]. Dostupné z: https://www.timesnownews.com/technology-science/exposed-ai-generated-viral-photo-of-putin-bowing-to-xi-jinping-fact-or-fiction-article-98912114

[35] HARTSFIELD, Tom. How do DALL-E, Midjourney, Stable Diffusion, and other forms of generative AI work?. Bigthink.com [online]. 28.9.2022 [cit. 2023-05-15]. Dostupné z: https://bigthink.com/the-future/dall-e-midjourney-stable-diffusion-models-generative-ai/

[36] BARR, Kyle. AI Voice Simulator Easily Abused to Deepfake Celebrities Spouting Racism and Homophobia. Gizmodo.com [online]. 30.1.2023 [cit. 2023-05-16]. Dostupné z: https://gizmodo.com/ai-joe-rogan-4chan-deepfake-elevenlabs-1850050482

[37] VINCENT, James. This is what a deepfake voice clone used in a failed fraud attempt sounds like. Theverge.com [online]. 27.7.2020 [cit. 2023-05-15]. Dostupné z: https://www.theverge.com/2020/7/27/21339898/deepfake-audio-voice-clone-scam-attempt-nisos

[38] BREWSTER, Thomas. Fraudsters Cloned Company Director's Voice In $35 Million Heist, Police Find. Forbes.com [online]. 2021 [cit. 2023-05-16]. Dostupné z: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/

[39] CAUGHILL, Patrick. China's Google Equivalent Can Clone Voices After Seconds of Listening. Futurism.com [online]. 27. 2. 2018 [cit. 2023-05-16]. Dostupné z: https://futurism.com/baidu-clone-voices-seconds

[40] SEIF, George. You can now speak using someone else's voice with Deep Learning. Towardsdatascience.com [online]. 2.7. 2019 [cit. 2023-05-21]. Dostupné z: https://towardsdatascience.com/you-can-now-speak-using-someone-elses-voice-with-deep-learning-8be24368fa2b

[41] 12 Best Deepfake Apps and Websites You Can Try for Fun. Beebom.com [online]. 23.12.2022 [cit. 2023-02-24]. Dostupné z: https://beebom.com/best-deepfake-apps-websites/

[42] SMITH, Katie Louise. Is ZAO safe to use? Viral Chinese app creates terrifying deep fakes based on one selfie. Popbuzz.com [online]. 3.9.2019 [cit. 2023-05-16]. Dostupné z: https://www.popbuzz.com/internet/social-media/zao-app-deepfake-privacy-issues/

[43] Reface: Face swap aplikace. Play.google.com [online]. [cit. 2023-05-16]. Dostupné z: https://play.google.com/store/apps/details?id=video.reface.app&hl=cs&gl=US

[44] VON WILPERT, Chris. 7 Best Deepfake Software Apps of 2023 (50 Tools Reviewed). Contentmavericks.com [online]. 10.5.2023 [cit. 2023-05-16]. Dostupné z: https://contentmavericks.com/best-deepfake-software/#DeepSwap

[45] Deepswap – AI Face Swap Online App. Deepswap.ai [online]. [cit. 2023-05-16]. Dostupné z: https://www.deepswap.ai/

[46] DALL·E 2 [online]. [cit. 2023-05-16]. Dostupné z: https://openai.com/product/dall-e-2

[47] Midjourney [online]. [cit. 2023-05-16]. Dostupné z: https://www.midjourney.com/home/?callbackUrl=%2Fapp%2F

[48] Deepfake Detection Made Easy with DeepDetector Software. Duckduckgoose.ai [online]. [cit. 2023-03-31]. Dostupné z: https://www.duckduckgoose.ai/detector

[49] Intel Introduces Real-Time Deepfake Detector. Intel.com [online]. [cit. 2023-03-31]. Dostupné z: https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html#gs.tg4y2x

[50] Scan & Detect Deepfake Videos. Deepware.ai [online]. [cit. 2023-05-16]. Dostupné z: https://scanner.deepware.ai/

[51] V7 Releases Deep Fake Detector for Chrome. V7labs.com [online]. 6.4.2022 [cit. 2023-05-16]. Dostupné z: https://www.v7labs.com/news/v7-releases-deep-fake-detector-for-chrome

[52] Maybe's AI Art Detector. Huggingface.co [online]. [cit. 2023-05-16]. Dostupné z: https://huggingface.co/spaces/umm-maybe/AI-image-detector

[53] Bellingcat's Online Investigation Toolkit. In: Docs.google.com [online]. [cit. 2023-05-16]. Dostupné rz: https://docs.google.com/spreadsheets/d/18rtqh8EG2q1xBo2cLNyhIDuK9jrPGwYr9DI2UncoqJQ/edit#gid=930747607

[54] METADATA2GO.COM [online]. [cit. 2023-05-16]. Dostupné z: https://www.metadata2go.com/

[55] Image Verification Assistant [online]. [cit. 2023-05-16]. Dostupné z: https://mever.iti.gr/forensics/

[56] AI Music, Text to Speech, & Voice Transformation [online]. [cit. 2023-05-16]. Dostupné z: https://fakeyou.com/

[57] Your Complete Generative Voice AI Toolkit [online]. [cit. 2023-05-16]. Dostupné z: https://www.resemble.ai/

[58] Voicer Celebrity Voice Changer [online]. [cit. 2023-05-16]. Dostupné z: https://play.google.com/store/apps/details?id=org.qstudio.voicer&hl=en_US

[59] AI Voice Detector, Your Ears Deserve the Truth [online]. [cit. 2023-05-16]. Dostupné z: https://aivoicedetector.com/

[60] Fake-voice-detection. Github.com [online]. [cit. 2023-05-16]. Dostupné z: https://github.com/dessa-oss/fake-voice-detection

[61] Detecting Audio Deepfakes With AI. Medium.com [online]. 28.9.2019 [cit. 2023-05-16]. Dostupné z: https://medium.com/dessa-news/detecting-audio-deepfakes-f2edfd8e2b35

[62] DeepFaceLab. Github.com [online]. [cit. 2023-05-16]. Dostupné z: https://github.com/iperov/DeepFaceLab

[63] Speech by Mr. António Guterres, United Nations Secretary-General, to WUF11. In: YouTube [online]. 27.6. 2022 [cit. 2023-05-16]. Dostupné z: https://www.youtube.com/watch?v=wU4qZgU9xAY&t=16s&ab_channel=UN-HABITATWORLDWIDE

[64] Watch The Last Word With Lawrence O'Donnell Highlights: April 13. In: YouTube [online]. 14.4. 2023 [cit. 2023-05-16]. Dostupné z: https://www.youtube.com/watch?v=iZzeVULQfdM&ab_channel=MSNBC

[65] Deepfake 3. In: YouTube [online]. 21. 4. 2023 [cit. 2023-05-16]. Dostupné z: https://youtu.be/AU5y8knoiuI

[66] Deepfaking Tarkin & Leia in Rogue One: A Star Wars Story [4K]. In: YouTube [online]. 8. 12. 2020 [cit. 2023-05-16]. Dostupné z: https://www.youtube.com/watch?v=_CXMb_MO3aw&t=53s&ab_channel=Shamook

[67] Deepfake - Nachgefragt Bundeswehr - Nora Tschiner und Mr. Bean 128 LIAE 100k inter. In: YouTube [online]. 20. 5. 2022 [cit. 2023-05-16]. Dostupné z: https://www.youtube.com/watch?v=SeMdWbh3GlI&ab_channel=Deepfacelabfan

[68] Willem Dafoe as Hannibal Lecter [DeepFake]. In: YouTube [online]. 10.3.2020 [cit. 2023-05-15]. Dostupné z: https://www.youtube.com/watch?v=XJB5W2beVZ0&t=181s&ab_channel=CtrlShiftFace

[69] Face with glasses (XSeg+wf in Colab). In: YouTube [online]. 23. 4. 2020 [cit. 2023-05-16]. Dostupné z: https://www.youtube.com/watch?v=h1Rr9X5QuIk&ab_channel=DeepFakeTube

[70] Twins Remake [Deepfake] - Arnold Schwarzenegger. In: YouTube [online]. 10. 1. 2020 [cit. 2023-05-16]. Dostupné z:

https://www.youtube.com/watch?v=GR2sRSB8kUA&t=78s&ab_channel=TheFakeReport

SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK

AI	Artificial Intelligence
GAN	Generative adversarial networks
TTS	Text to speech
RNN	Recurrent neural network
CNN	Convolutional neural network
EXIF	Exchangeable image file format
GPS	Global Positioning System
VRAM	Video Random Access Memory
OSN	Organizace spojených národů
MSNBC	Microsoft and the National Broadcasting Company.
FPS	Frames per second
JPEG	Joint Photographic Experts Group
PNG	Portable Network Graphic
GPU	Graphics Processing Unit
SAEHD	High Definition Styled Auto Encoder

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

SEZNAM PŘÍLOH

PŘÍLOHA I: MANUÁL NA OBRANU PŘED AUDIOVIZUÁLNÍ MANIPULACÍ

Úvodem:

Tento stručný manuál slouží jako úvod do manipulace za použití audiovizuálních děl. Na konkrétních příkladech potvrzuje důležitost této problematiky a uvádí základní postupy při ověřování pravosti záznamu či fotografie.

Proč je důležité se touto problematikou zabývat?

Naučit se odhalovat deepfake videa a falešné fotografie je důležité z několika důvodů. Prvním důvodem je ochrana proti dezinformacím a šíření falešných informací. V dnešní digitální době je snadné vytvořit realisticky vypadající deepfake videa a fotografie, které mohou být použity k manipulaci s veřejným míněním, šíření falešných zpráv nebo poškození pověsti jednotlivců či institucí. Druhým důvodem je ochrana osobního soukromí. Deepfake technologie mohou být zneužity k vytváření falešných obrazů a videí jednotlivců, které by mohly být použity k šikaně, vydírání nebo jiným nekalým účelům. Naučit se odhalovat deepfake videa a falešné fotografie pomáhá lidem chránit své soukromí a osobní integritu v digitálním prostředí.

Co je deepfake video?

Jako deepfake označujeme falzifikovaná videa se záměnou hlasové či zvukové identity. Nejčastěji se jedná o projevy, rozhovory či filmové scény, kdy je původní tvář herce či politika nahrazena jinou (falešnou) tváří. Deepfake video vznikají nejčastěji pro pobavení, ale můžou být taktéž silným desinformačním nástrojem.
První deepfake video vzniklo v roce 2017 a od té doby se velmi rozšířily a zdokonalily.
Je proto vhodné znát několik základních informací, abychom se tzv. nenechali „napálit."

Pokud se setkáme s podezřelým videem, je v první řadě vhodné zaměřit se na kontext a vyhledat si o osobách vystupujících ve videu více informací. Pokud takovou možnost nemáme, sledujeme detaily videa, jako pohyby očí, úst, grafické anomálie, nepřirozené stíny či barvy. Většina deepfake videích stále obsahuje nedokonalosti, které jsou viditelné na první pohled Na prvním obrázku níže si můžeme všimnout rozdílu v kvalitě úst a zubů kdy u je levého deepfake snímku patrné nižší rozlišení. U druhého obrázku si všimněme brýlí, které u levého deepfake snímku v určitý moment zcela zmizí. Pokud bychom si chtěli zkusit vytvořit vlastní video nabízí se např. nástroje Deepfakesweb či Reface.

Generované fotografie

Tvorba fotografií pomocí generátorů je dnes díky množství nástrojů mnohem častější a jednodušší. Nebezpečné jsou zejména generátory dle textového zadání, které dokáží vytvořit fotografii téměř kohokoliv. Stejně tak se můžeme setkat s falešnými fotografiemi neexistujících osob. Můžeme tedy nějak určit, která fotografie je pravá, a která falešná? Stejně jako u Deepfake videí je tu několik detailů na které je třeba se zaměřit:

- **Brýle, náušnice, přívěšky...**
- **Splývající či nepřirozeně pokroucené prameny vlasů**
- **Rozmazané pozadí, které nepřipomíná nic konkrétního**
- **Nesmyslné texty**

Pokud se nám zdá že některý z detailů neodpovídá skutečnosti, je velká šance, že fotografie nebyla nikdy vyfocena. Pojďme si nyní udělat malý test. Podívejte se pozorně na dva následující obličeje a zkuste si tipnout, který z nich je pravý.

Pokud jste řekli ani jeden, gratuluji, je tomu skutečně tak, oba byly vytvořeny nástrojem na generování fotografií. Pokud Vás zmátli, pojďme si ukázat proč jsou falešné. U levé fotografie jsou patmé chybějící obroučky brýlí, u pravé je podezřelá čepice, na které si můžeme všimnout nesmyslného symbolu.

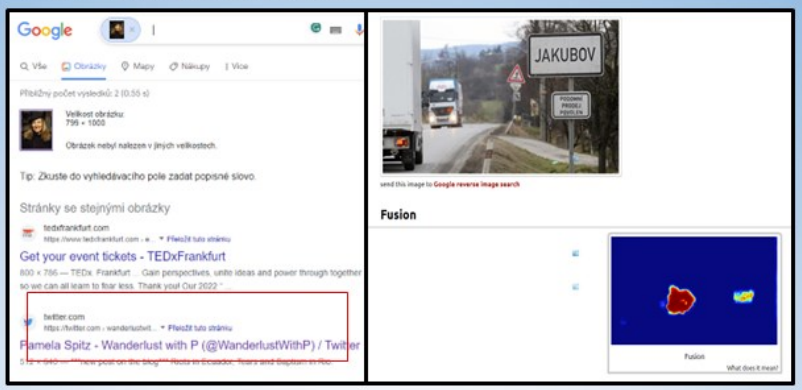
Jak odhalit vygenerovanou fotografii a nenechat se zmást si můžeme názorně ukázat i na dvou fotografiích níže – obě se staly virálními v roce 2023 a obě pochází z generátoru. Jak si můžeme Všimnout problémem jsou zejména ruce, které v obou případech vypadají nerealisticky.

Věděli jste...?
V roce 2022 obsahovalo 66% falešných profilů na Facebooku fotografií z generátoru.

Podezřelé fotografie – použití reversního vyhledávání

Velmi užitečným nástrojem pro ověření původu fotografie může být reversní vyhledávání od Google. Na levém obrázku vidíme názorné použití, kdy jsme reversním vyhledáváním zjistili pravou totožnost ženy z fotografie, která byla použita u falešného profilu.

Na obrázku vpravo vidíme použití nástroje Image Verification Assistant, který zobrazuje podvržené oblasti (aplikování klonů či opakovanou kompresi) jako vysoké hodnoty (teplé barvy) na pozadí s nízkou hodnotou (studené barvy). Z obrázku byly odstraněny automobily a nápis na ceduli byl přepsán.



Manipulace za použití audiozáznamu

Stejné nebo možná ještě větší nebezpečí než deepfake představuje tzn. deepvoice – falešný hlas generovaný pomocí klonování hlasu. Pomocí této technologie je možné napodobit hlas jakékoliv konkrétní osoby a ten následně použít např. k podvodu jako se stalo v případě dvou firem ve Velké Británii a Spojených Arabských Emirátech. Kdy podvodníci použili při telefonním hovoru falešný hlas ředitele společnosti.

V tomto případě je možnou obranou důsledné ověření zda skutečně mluvíme s danou osobou např. pomocí hesla či jiné unikátní informace. Experti taktéž doporučují jednoduchý trik, kterým je v případě podezření na falešný hlas, zavěsit. Podvodníci často používají jednorázová čísla a pokud se nejedná o opravdu sofistikovaný podvod tak již znovu nezavolají.

Závěrem:

V tomto manuálu jsme se seznámili s důležitými postupy a technikami pro odhalování deepfake videí, falešných fotografií a generovaného hlasu. Naučili rozpoznávat znaky deepfake videí a fotografických manipulací, jako jsou nesrovnalosti ve tvářích, nepřirozené pohyby, nekonzistence světla a stínů a další. Pochopili jsme, že deepfake technologie se neustále vyvíjejí, a proto je důležité být ostražití a neustále se vzdělávat o nejnovějších metodách. Při odhalování deepfake je také důležité spoléhat se na další znalosti a zdroje, zejména znalost kontextu. Pamatujte si, že schopnost odhalovat deepfake videa a falešné fotografie je klíčová pro ochranu naší společnosti a digitálního prostoru. Budujeme tak kritické myšlení, zdravou skepsi a schopnost rozpoznat manipulaci.