Metody wykrywania nagrań typu deepfake

Adam Żychowski

Majority say fake news has left Americans confused about basic facts

% of U.S. adults who say completely made-up news has caused ___ about the basic facts of current events

Source: Survey conducted Dec. 1-4, 2016.

PEW RESEARCH CENTER

North Carolina For Donald Trump
October 14, 2016 ·

Pope endorses Trump!
Game changer !!

http://endingthefed.com/pope-francis-shocks-world-endorses-...

Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement

VATICAN CITY – News outlets around the world are reporting on the news that Pope Francis has made the unprecedented decision to endorse a US...

ENDINGTHEFED.COM

[redacted]

Jakie wieści dostałem od znajomych?
-Tajna narada
-Warszawa zamknięta
-W pn. zamkną sklepy
-Zapadła decyzja o zamknięciu stacji benzynowych
-Banki będą zamknięte, wypłacaj z konta

7 2 8

Temat:PD:

"Czesc wszystkim.
Mam Info od kolezanki z Warszawy, ktora pracuje w administracji panstwowej.
Zrobcie zakupy, bo potem ceny moga byc zabojcze w osiedlowych sklepikach ...😩

„Wszystkie centra handlowe beda zamkniete od poniedzialku. Czynne beda tylko sklepy spozywcze. W niedziele bedzie oredzie prezydenta o wprowadzeniu stanu wyjatkowego. Warszawa ma byc podzielona na 3 strefy. Wojsko juz do Warszawy sie zjezdza. Wczoraj juz bylo bardzo duzo w Warszawie."

Deepfake

- technika stosowana jest do łączenia i nakładania obrazów nieruchomych i ruchomych na obrazy lub filmy źródłowe przy użyciu metod uczenia maszynowego (głównie sieci neuronowych)

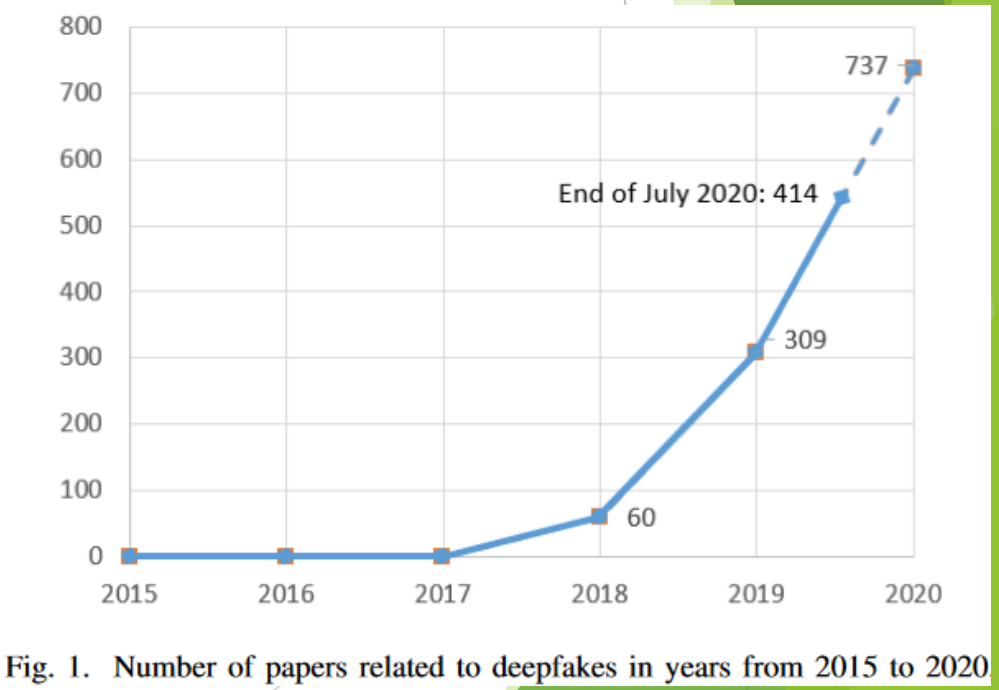


Fig. 1. Number of papers related to deepfakes in years from 2015 to 2020

[1]

Przykłady

„Wkraczamy w erę, w której nasi wrogowie mogą zmusić kogokolwiek do wypowiedzenia dowolnych słów w dowolnym momencie”

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

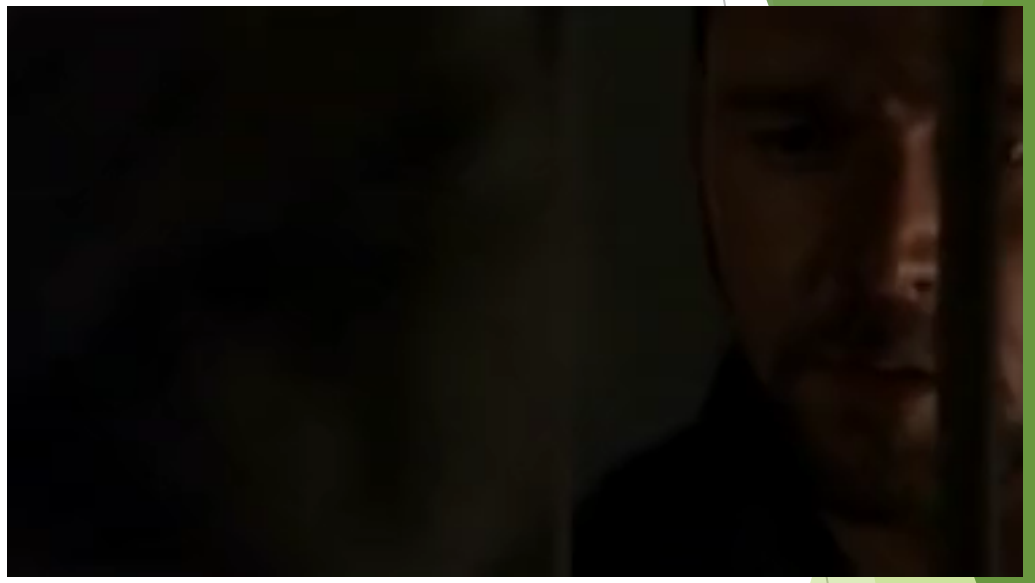
Przykłady

https://www.youtube.com/watch?v=A8TmqvTVQFQ

Przykłady

Home Stallone https://www.youtube.com/watch?v=2svOtXaD3gg

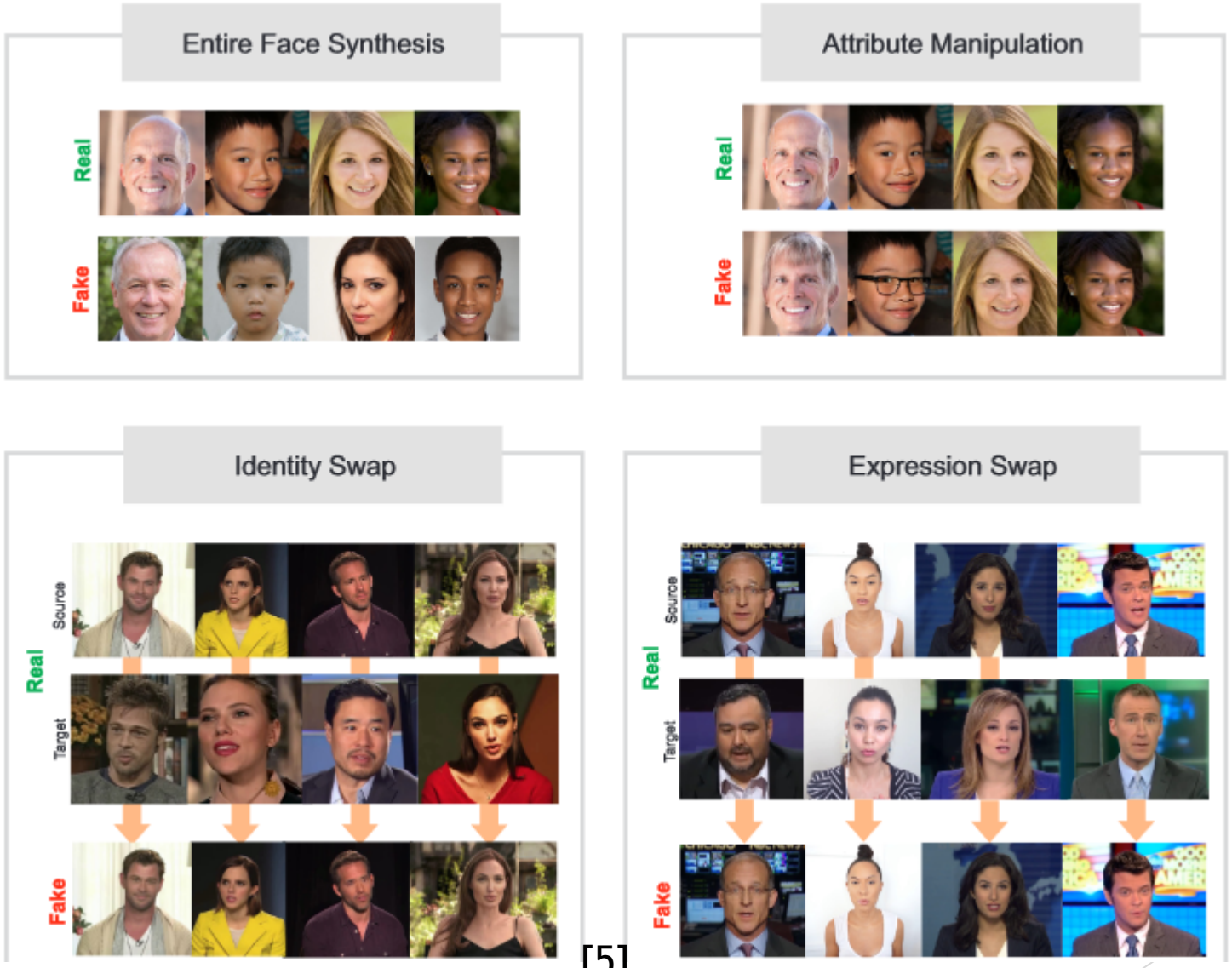
To nic nowego?



Photoshop
CGI (Computer Generated Imagery)

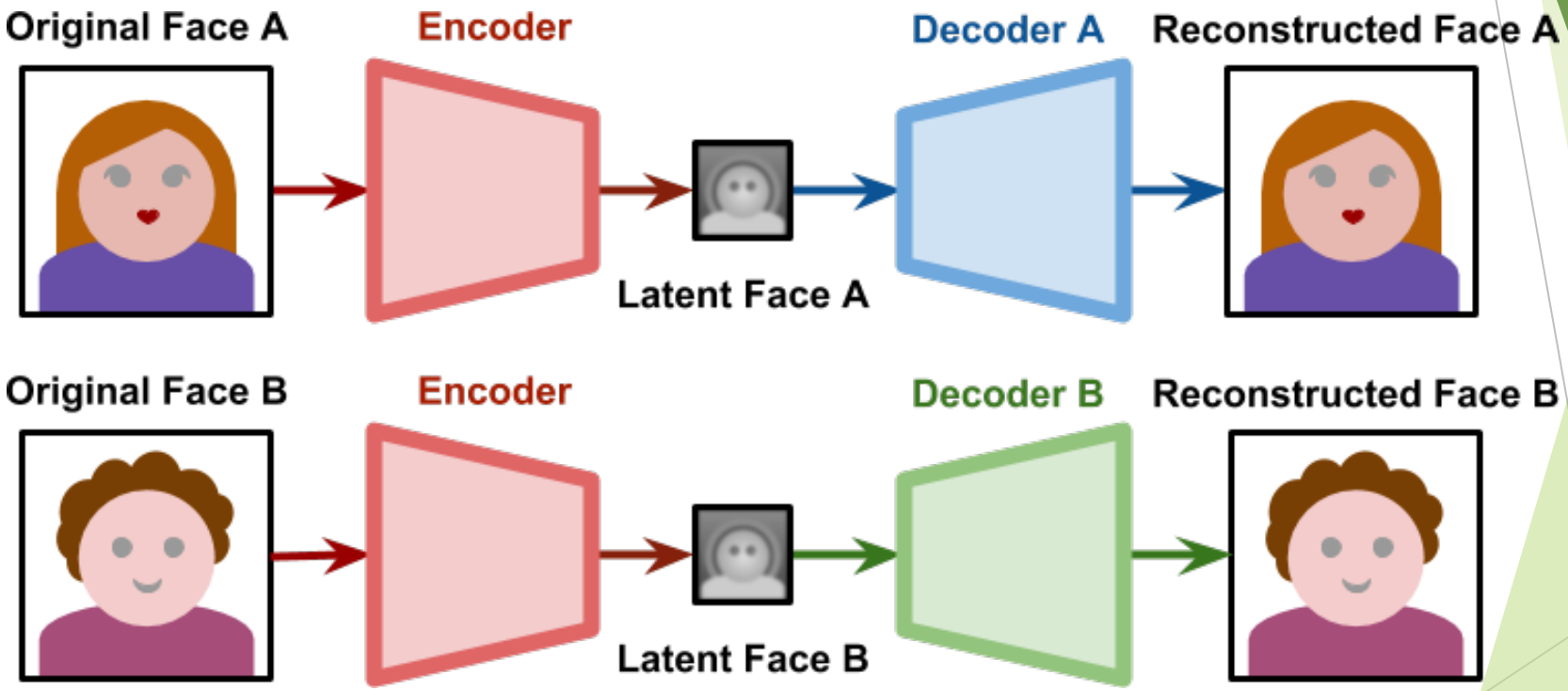
https://colab.research.google.com/github/AliaksandrSiarohin/first-order-model/blob/master/demo.ipynb#scrollTo=SB12II11kF4c

Rodzaje manipulacji



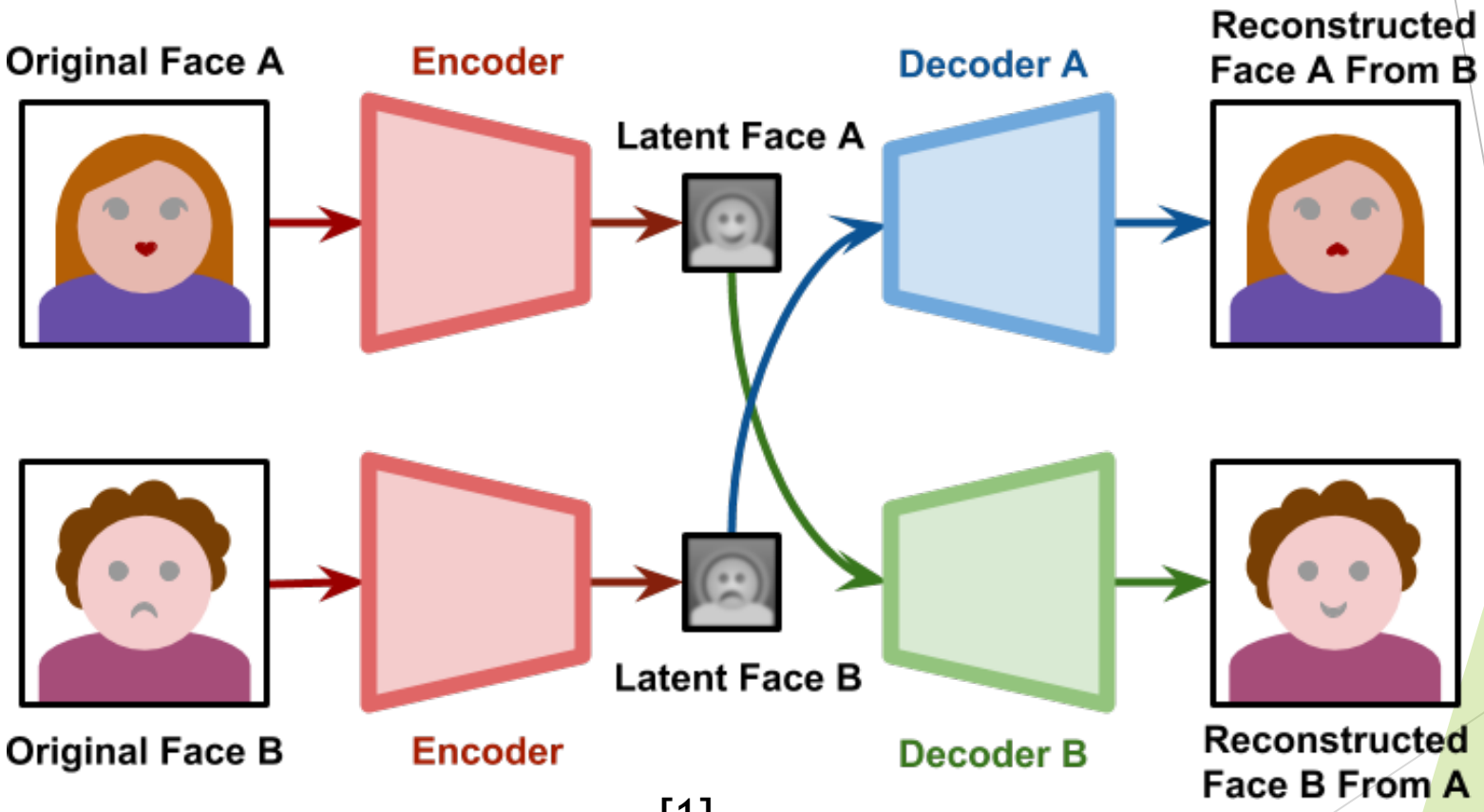
[5]

Autoenkodery



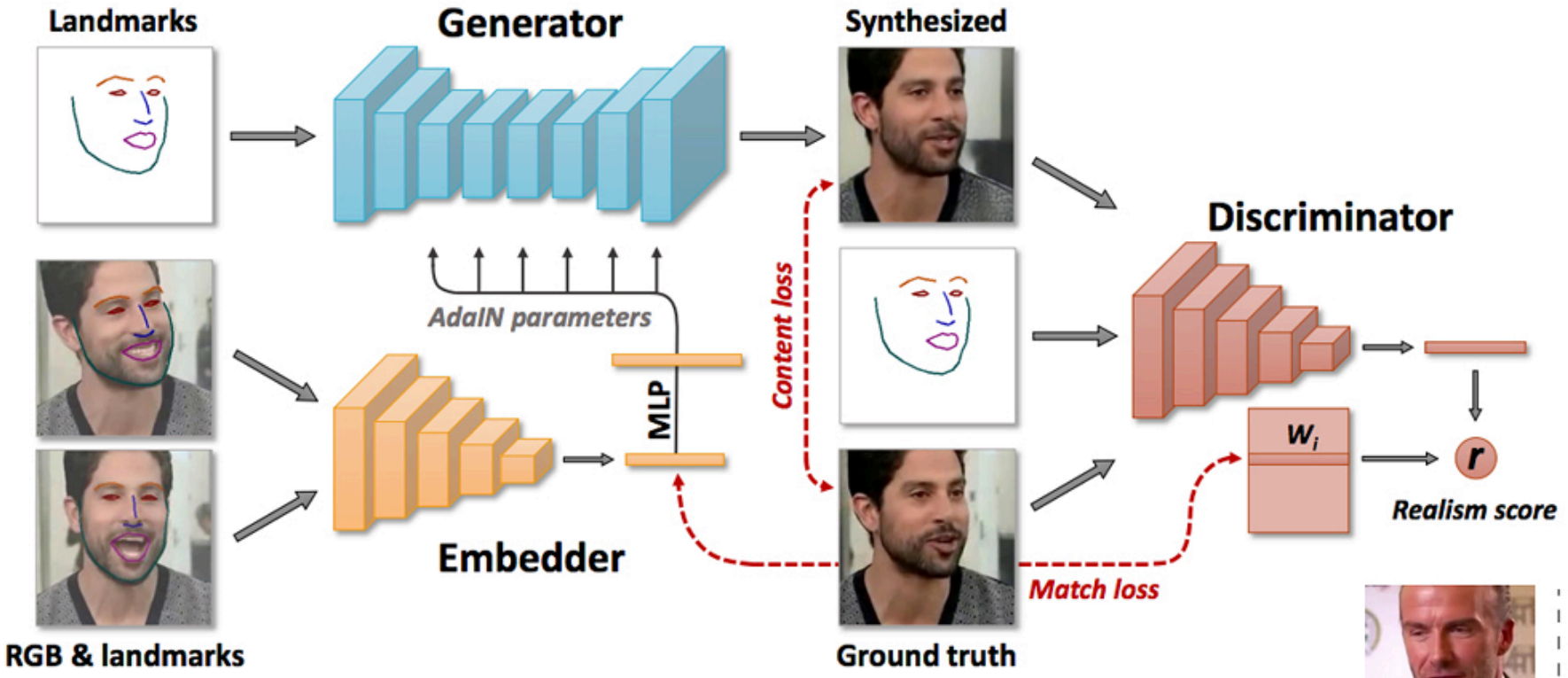
[1]

Autoenkodery

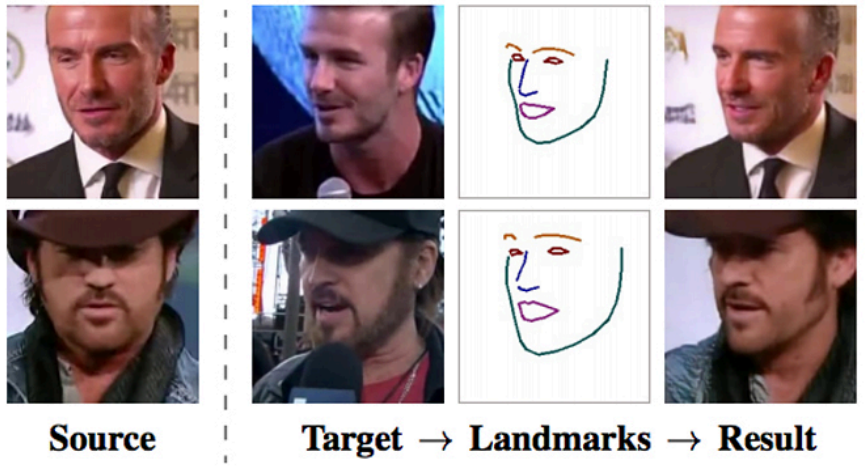


[1]

Generative adversarial networks

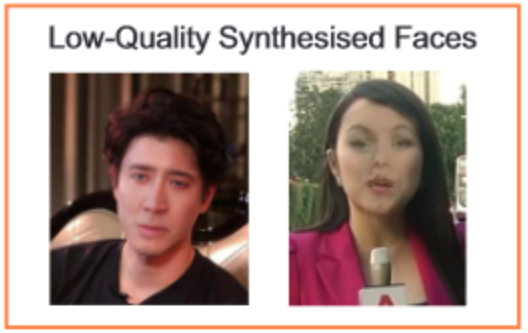


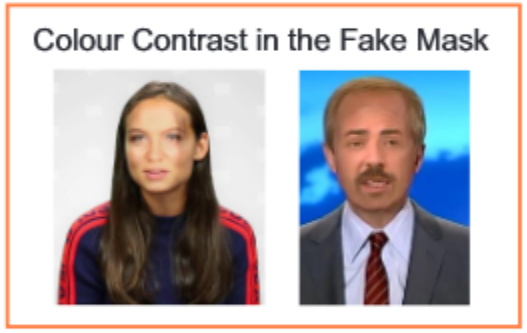
[2]

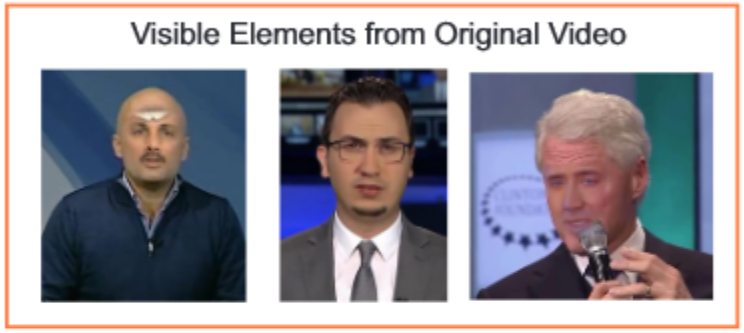


Detekcja manualna

- oczy – poruszanie, mruganie
- nadmierne rozmycie lub wyostrzenie
- nienaturalna pozycja twarzy
- brak emocji
- nienaturalne gesty pozostałej części ciała (lub ich brak)
- włosy i zarost
- oświetlenie, cienie







[5]

„Odcisk palca”

- rozwiązanie zaproponowane przez Microsoft we współpracy z serwisami informacyjnymi (m.in. BBC, New York Times)

- kryptografia - dodawanie odpowiednich sekwencji do plików (metadane)

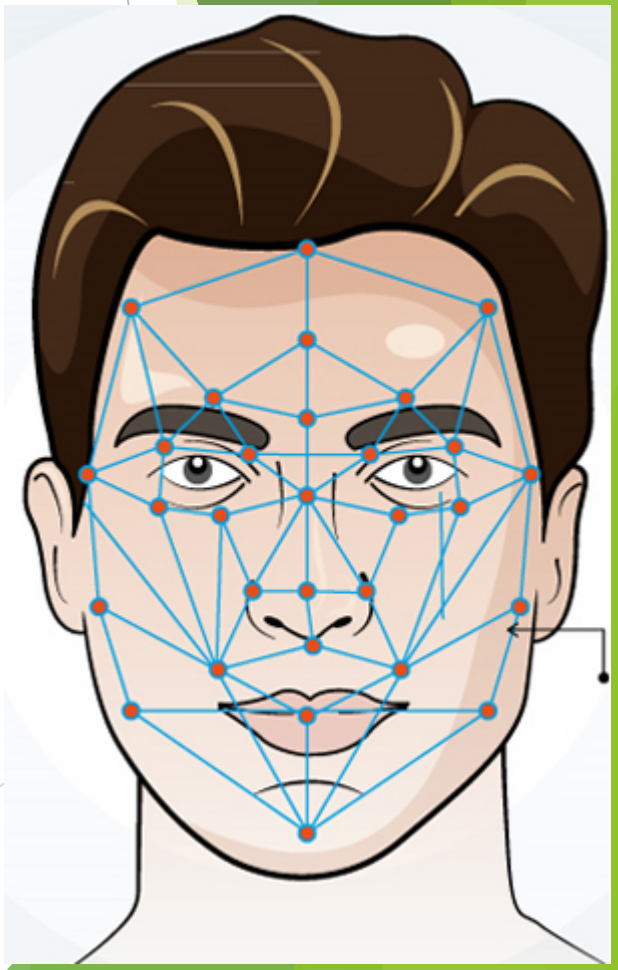
lub

- znak wodny (wygenerowany przez sieć neuronową) [4] niewidzialny dla oka, odporny na typowe zniekształcenia: zmiana wielkości, jasności, kontrastu, kompresja, ale wrażliwy na usuwanie, dodawanie, zmianę obiektów

które potem mogą być weryfikowane przez przeglądarkę lub portale społecznościowe

Punkty charakterystyczne

- wykorzystanie osiągnięć biometrii
- sprawdzanie położenia punktów charakterystycznych w kolejnych klatkach – np. odległość między nimi
- porównywanie twarzy ze zdjęciem referencyjnym
- rozpoznawanie twarzy



Manualna ekstrakcja wektora cech [6]

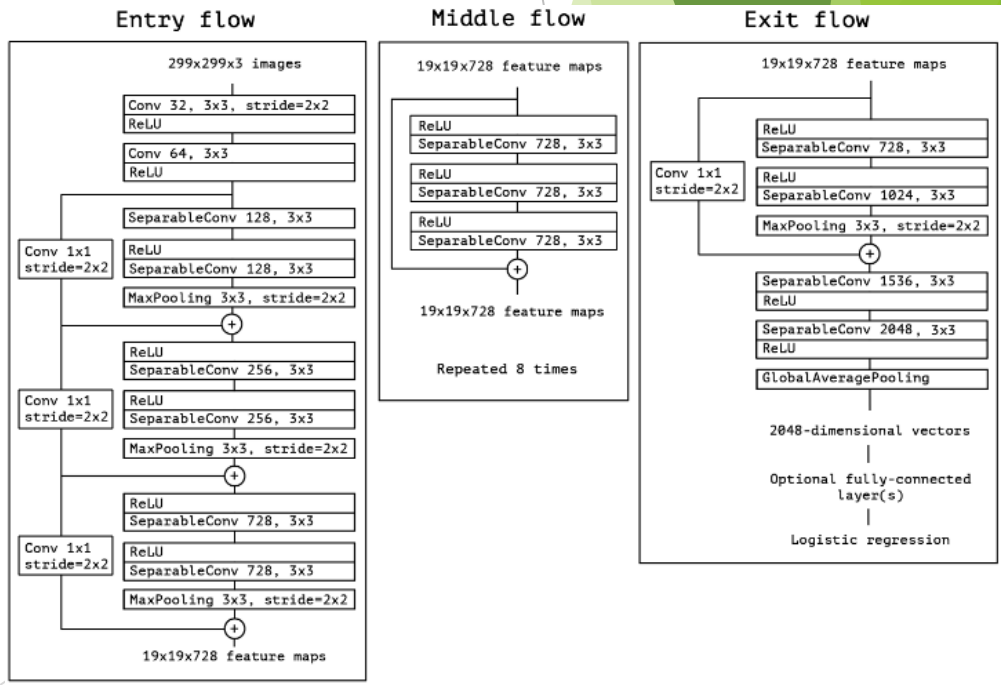
- zdefiniowanie wektora cech twarzy pojedynczego obrazu (np. rozstaw oczu, odległość między kośćmi policzkowymi, szerokość nosa) lub całego filmu (np. częstotliwość mrugania, pozycja twarzy w czasie, mimika, mikroekspresje)

- nauka modelu uczenia maszynowego (np. SVN, MLP) na podstawie wektora cech, a nie bezpośrednio na obrazie

Konwolucyjne sieci neuronowe

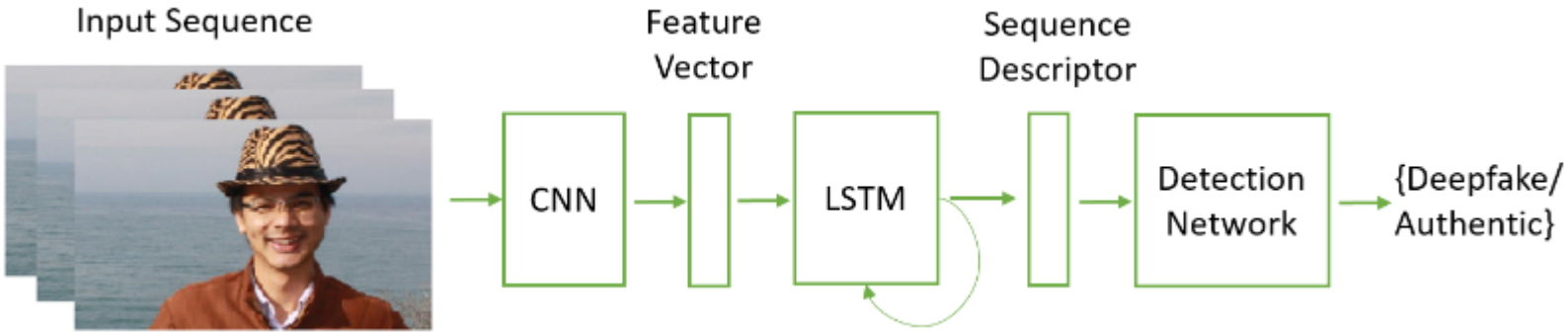
- Dwa podejścia:
  - klasyfikacja poszczególnych klatek nagrania (obrazów)
  - klasyfikacja przejść pomiędzy klatkami (par obrazów)

- różne architektury, często pretrenowane
- wykorzystanie istniejących bardzo głębokich architektur [6]



Rekurencyjne sieci neuronowe

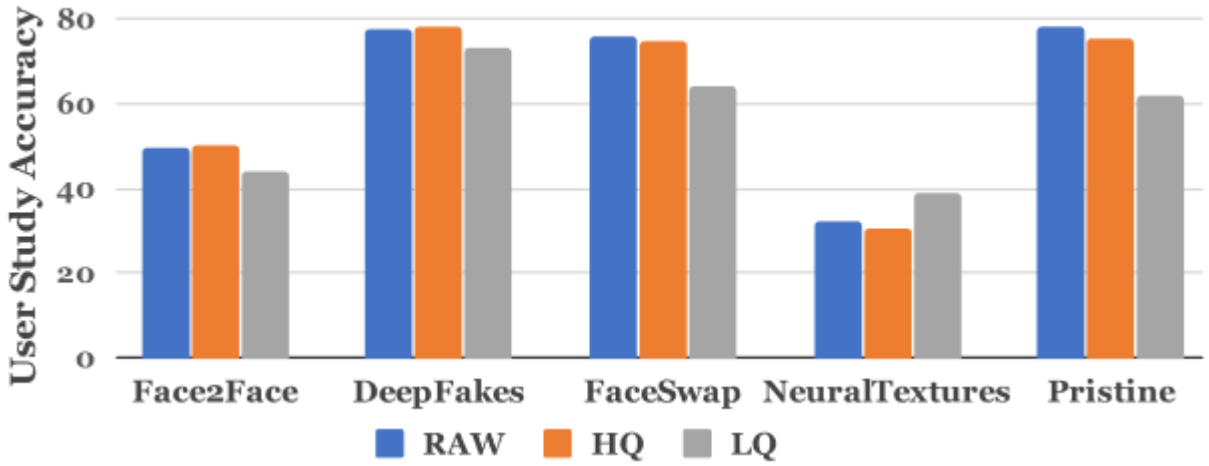
- kolejne klatki filmu -> szereg czasowy
- często łączone z sieciami konwolucyjnymi do ekstrakcji cech



[1]

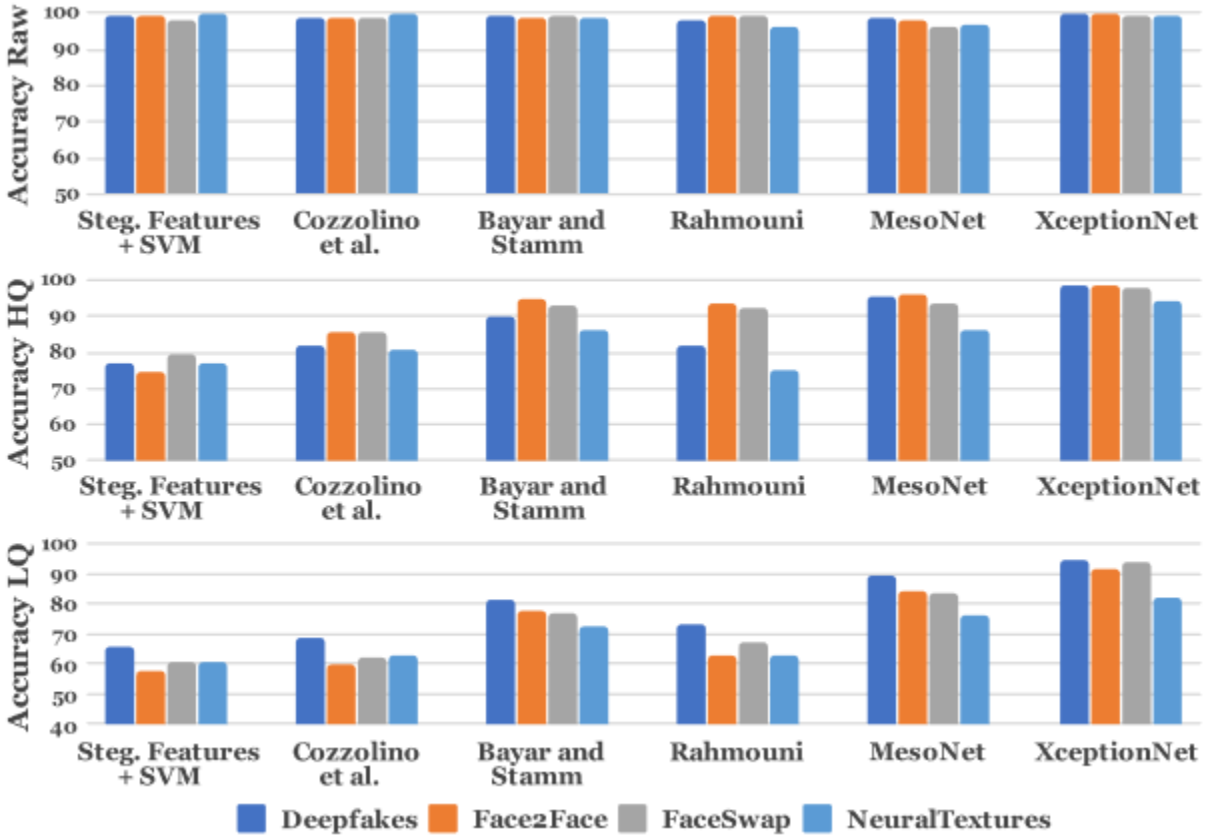
Rezultaty

- Benchmark: 1000 nagrań z youtube i serwisów informacyjnych zmanipulowanych kilkoma najbardziej znanymi metodami tworzenia deepfake w 3 jakościach



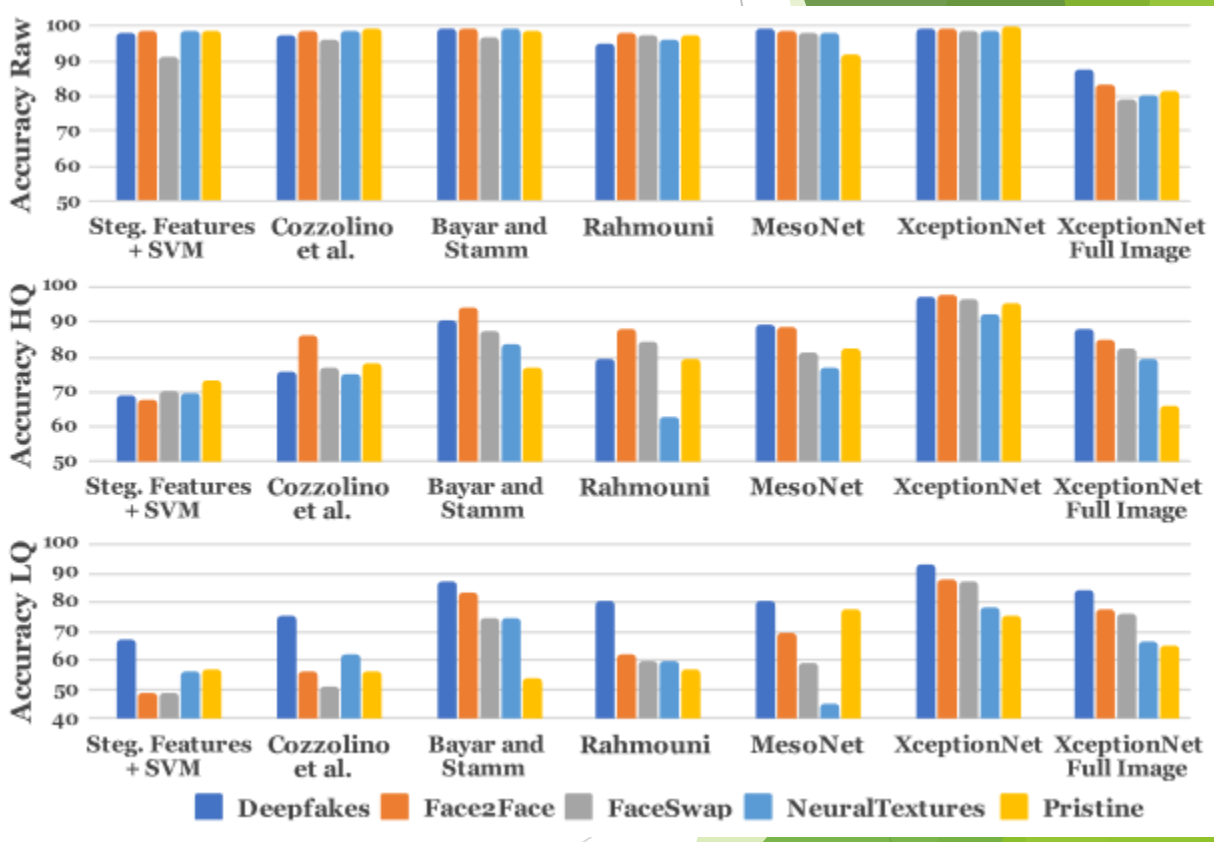
wyniki ludzi [6]

Rezultaty



Nauka tylko na wybranym rodzaju manipulacji

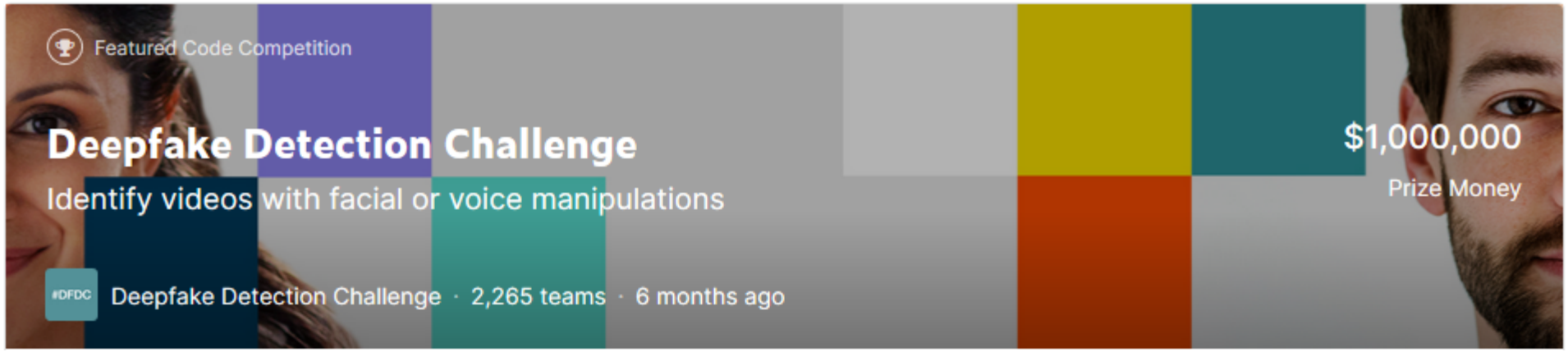
[6]



Nauka na wszystkich rodzajach manipulacji

Deepfake detection challenge

- Amazon, Facebook, Microsoft
- marzec – czerwiec 2020

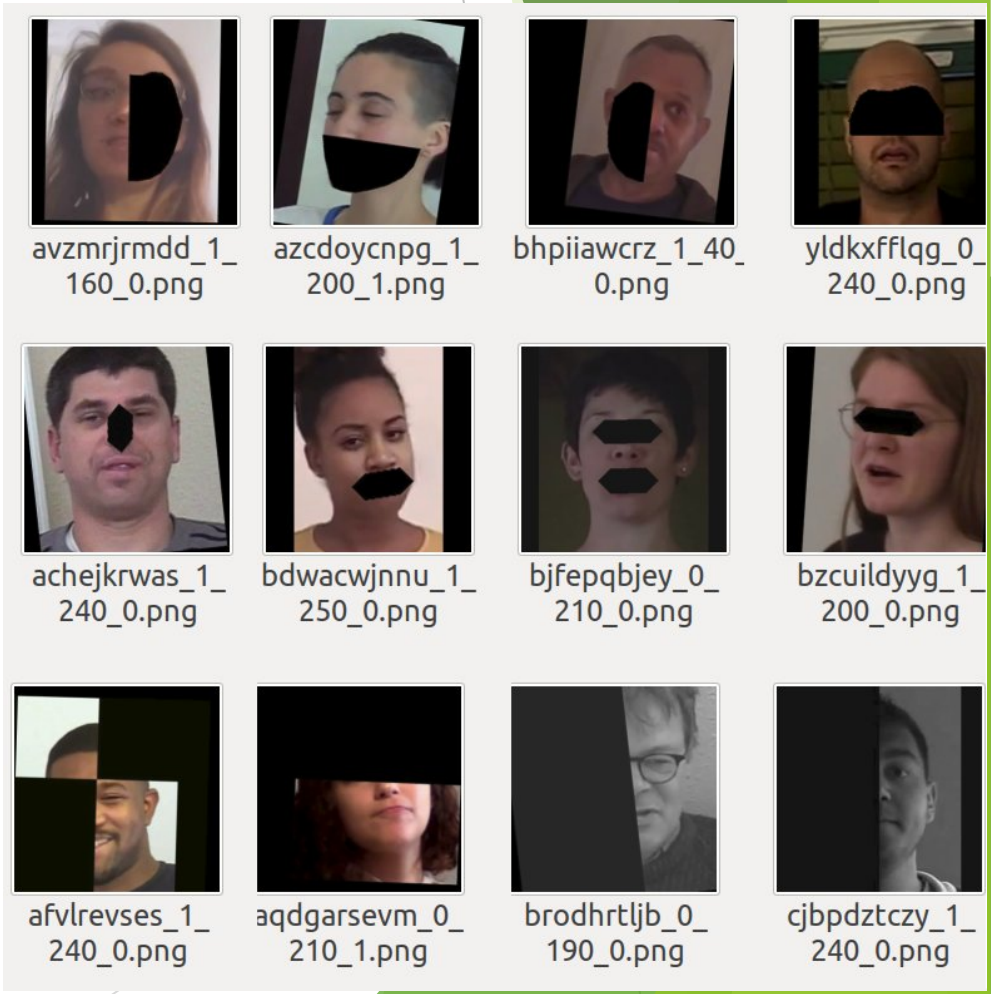


#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	▲ 3	Selim Seferbekov			0.42798	2	7mo
2	▲ 35	\WM/			0.42842	2	7mo
3	▲ 3	NtechLab			0.43452	2	7mo

Deepfake detection challenge

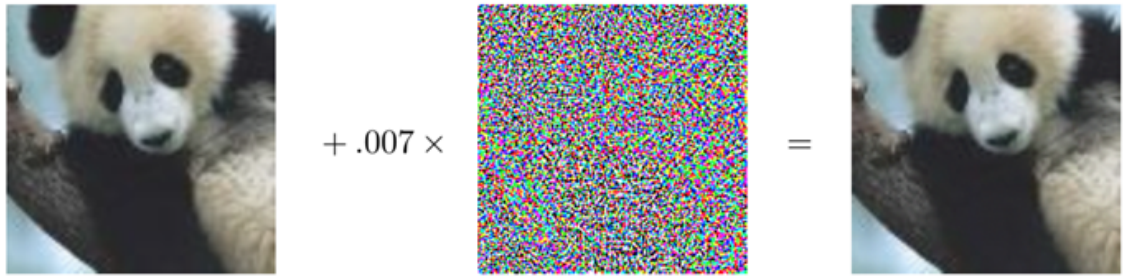
Deepfake detection challenge – zwycięskie rozwiąznanie

- komitet 3 sieci EfficientNet (Google AI 2019):
  - 2 sieci działające na sekwencjach ramek
  - 1 sieć działająca na pojedynczych ramkach (zdjęciach)
- szeroka augmentacja danych treningowych: zmiana kontrastu, jasności, kolorów, kompresja video, usuwanie ramek, łączenie prawdziwych nagrań z deepfake, zakrywanie części twarzy
- użycie pretrenowanej sieci + 5 dni treningu na specjalnym serwerze zoptymalizowanym pod uczenie głębokie

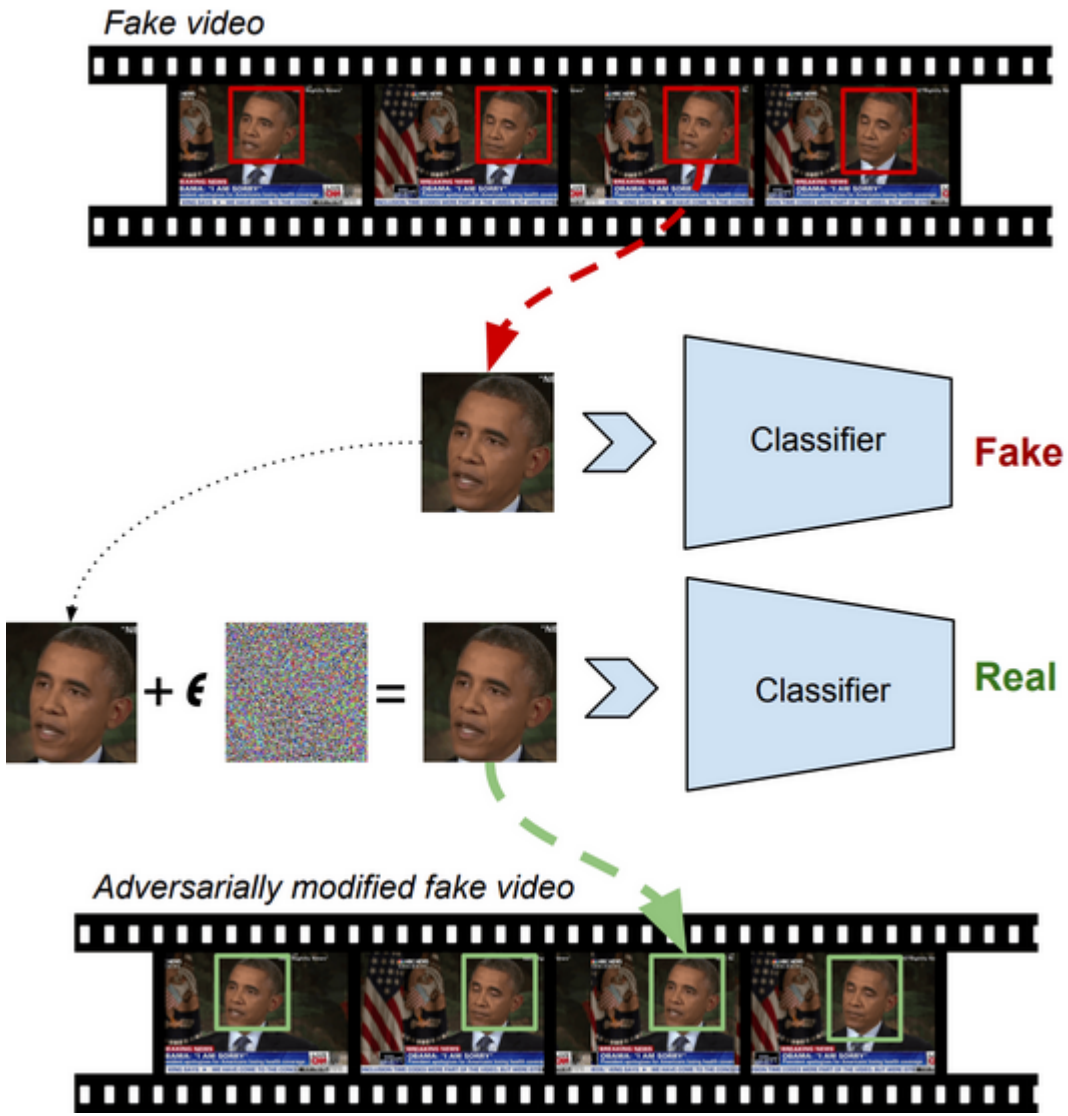


Odpowiedź deepfake

- detektor jako dyskryminator (GAN)
- poprawianie twarzy (np. wyostrzanie rysów) - StyleGAN
- Adversarial Deepfakes



Adversarial deepfakes



[3]

Face2Face

FaceSwap

NeuralTextures

https://adversarialdeepfakes.github.io

Deepfake a prawo

- debata w kongresie USA – propozycja nakazu oznaczania każdego z takich nagrań za pomocą specjalnego widocznego znaku wodnego
- Unia Europejska chce zmusić firmy technologiczne do ustalenia odpowiednich regulacji wewnętrznych i ich egzekwowania
- technologiczni giganci też nie nadążają
- jakiekolwiek działania policji, sądów - zbyt późno

- Tymczasem w Polsce:

„Powiedzieć fake news o czymś takim to zdecydowanie za mało powiedzieć. (…) ***To był deepfake, to był bardzo głęboki fake****, zastosowany przez człowieka, jestem przekonany o tym, jest obcy środowiskom mniejszości seksualnych, bo oni wiedzą, że mogą w Polsce maszerować bez przeszkód, mogą demonstrować swoje racje.”*

Mateusz Morawiecki, 28.09.2020

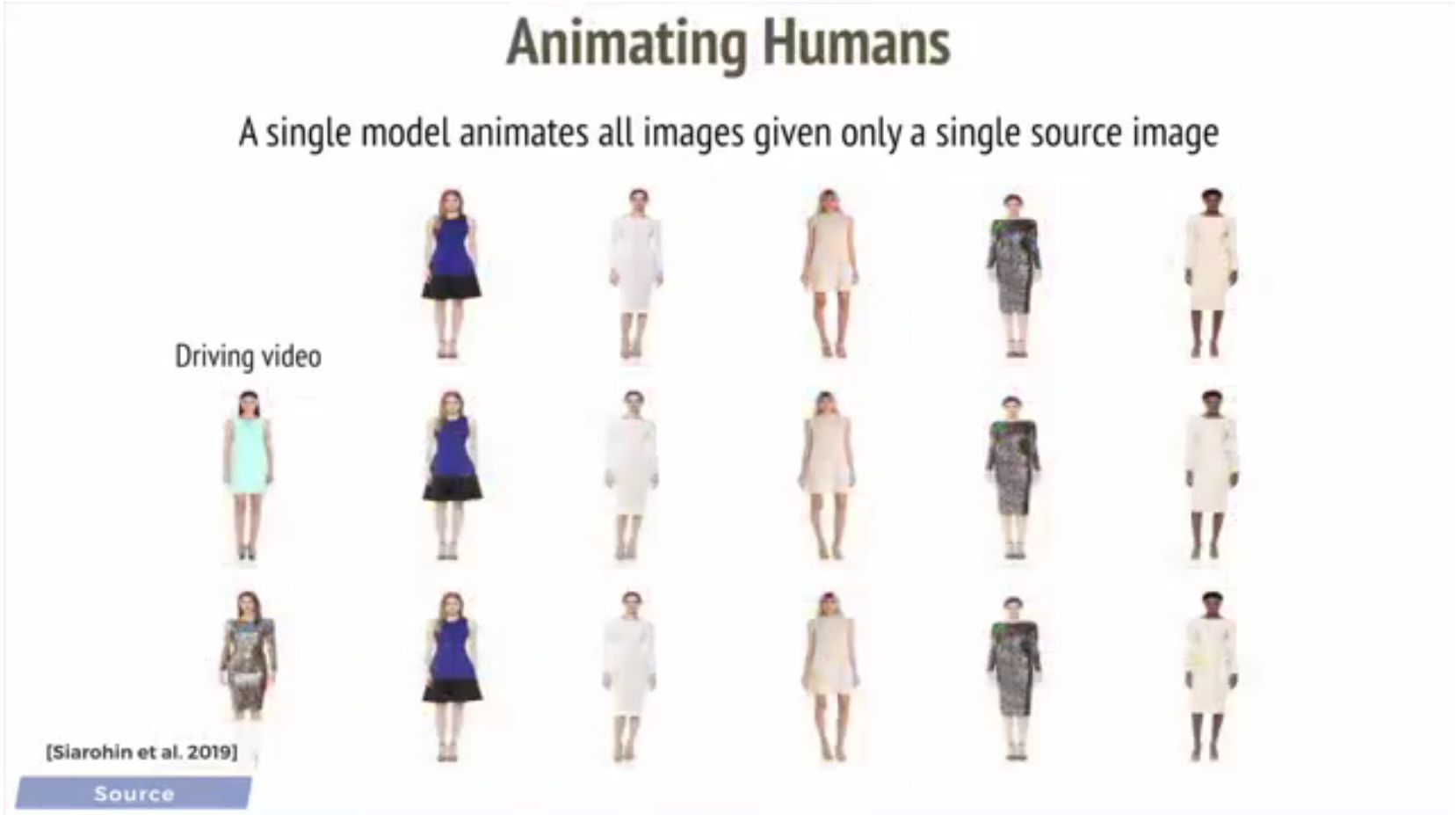
- art. 216 kodeksu karnego:

„Kto znieważa inną osobę za pomocą środków masowego komunikowania, podlega grzywnie, karze ograniczenia wolności albo pozbawienia wolności do roku.”

Powiązane zagadnienia

- generowanie fałszywych nagrań audio (wypowiedzi, rozmów) – imitacja czyjegoś głosu
- animacje
- całe ciało: taniec, ćwiczenia, inne czynności

Nie tylko twarze



Podsumowanie

- detektory: słaba generalizacja, dobre wyniki jeśli wytrenowane na tej samej (znanej) technice manipulacji
- zastosowanie istniejących rozwiązań do nowego problemu
- najlepsze rezultaty osiągają komitety detektorów

- realne zagrożenie – manipulacje, wiarygodność źródeł
- wyścig „zbrojeń”
- mimo obaw na razie brak potwierdzonych faktycznych zastosowań

Bibliografia

[1] Nguyen, Thanh Thi, et al. "Deep learning for deepfakes creation and detection." *arXiv preprint arXiv:1909.11573* 1 (2019).

[2] Zakharov, Egor, et al. "Few-shot adversarial learning of realistic neural talking head models." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.

[3] Neekhara, Paarth, et al. "Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples." *arXiv preprint arXiv:2002.12749* (2020).

[4] Korus, Pawel, and Nasir Memon. "Content authentication for neural imaging pipelines: End-to-end optimization of photo provenance in complex distribution channels." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[5] Tolosana, Ruben, et al. "Deepfakes and beyond: A survey of face manipulation and fake detection." *arXiv preprint arXiv:2001.00179* (2020).

[6] Rossler, Andreas, et al. "Faceforensics++: Learning to detect manipulated facial images." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.