



ROBUST PHYSICAL-WORLD ATTACKS ON FACE RECOGNITION

WOJCIECH SZLOSEK, KONRAD NOWAK

ABSTRACT



Wprowadzenie do zagadnienia



Praca jest odpowiedzią na
zagrożenia związane z wrogimi
pułapkami przy procesie
rozpoznawania twarzy



Prezentacja nowego
frameworku PadvFace



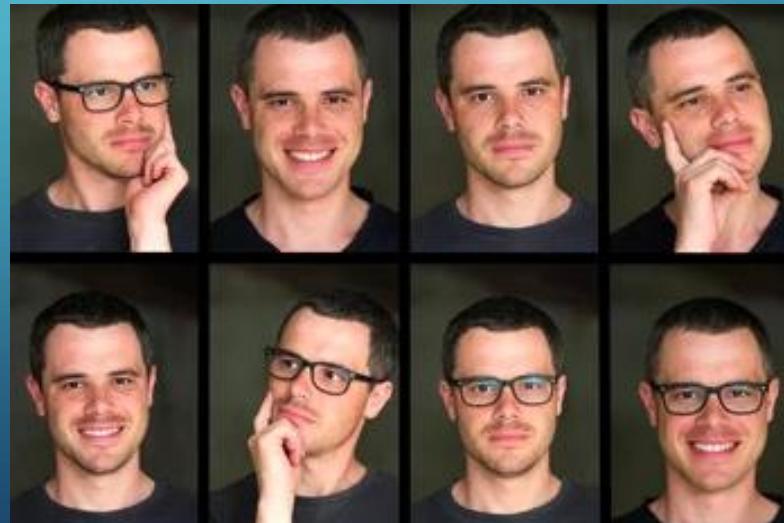
Algorytm CAA – *Curriculum
Adversarial Attack*



Eksperymenty i testowanie

WSTĘP

- Rozwój głębokich sieci neuronowych przyczynił się do rozwoju zagadnienia rozpoznawania twarzy – wiele zastosowań
- Za tym zagadnieniem kryje się wiele pułapek i możliwych sposobów na oszukanie rozpoznawania



Źródło obrazka: shutterstock.com

ATAKI



Ataki zasadniczo możemy podzielić na dwie kategorie: ataki cyfrowe oraz ataki fizyczne



W dalszym ciągu zostaną omówione głównie ataki fizyczne – w skomplikowanych warunkach rzeczywistych

ATAKI FIZYCZNE - ROZWINIĘCIE

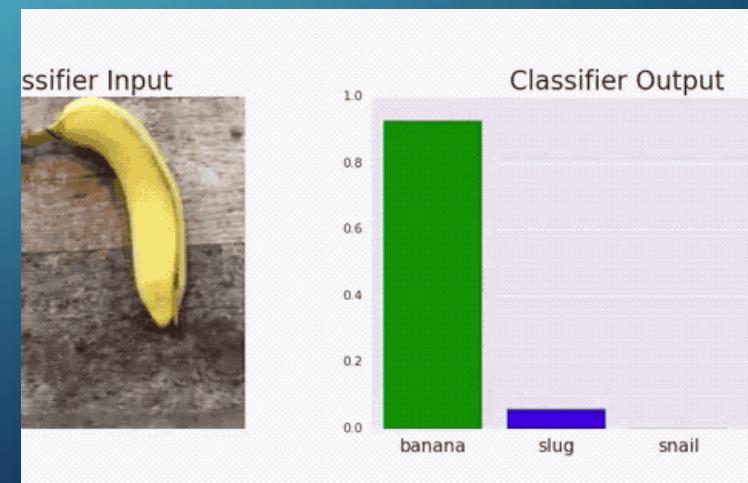
Oszukanie sieci neuronowych najczęściej poprzez zakłócające sens obiekty w świecie rzeczywistym



Niniejsza praca zajmuje się głównie tą właśnie tematyką

PRACE POWIAZANE Z TYM TEMATEM

- Ranga problemu sprawia, że wielu badaczy rozważa temat ataków i innych przeszkód w rozpoznawaniu twarzy
- Obecnie przeszkody typu okulary lub nakrycia głowy nie stanowią ogromnych problemów; wyzwanie mogą stanowić naklejki



PRZYKŁADOWY SCENARIUSZ ATAKU



Dwa zasadnicze etapy *sticker-based attack*:



- wygenerowanie naklejki w przestrzeni cyfrowej



- wydrukowanie naklejki – i jej noszenie

WSTĘPNE RACHUNKI

- Niech $f: \mathbb{R}^m \rightarrow \mathbb{R}^d$ będzie modelem rozpoznawania twarzy

WSTĘPNE RACHUNKI

- Niech $f: \mathbb{R}^m \rightarrow \mathbb{R}^d$ będzie modelem rozpoznawania twarzy
- Mamy przy tej okazji dwa zadania:
 - A) *face identification*
 - B) *face verification*

DALSZE RACHUNKI

x, x^a - dwa obrazy twarzy

Wynik ich dopasowania stanowi podobieństwo cosinus między $f(x)$, a $f(x^a)$.

PODOBIEŃSTWO COSINUSOWE – JAKO MIARA PODOBIEŃSTWA



Matematycznie: podobieństwo cosinus daje podobieństwo dwóch wektorów z n wymiarów przez określenie cosinusa kąta.



Mając dwa wektory, A i B , cosinus ich kąta θ otrzymujemy przez podzielenie ich iloczynu skalarnego przez iloczyn ich norm.



Podobieństwo cosinusowe jest metryką używaną do określania stopnia podobieństwa dwóch podmiotów niezależnie od ich wielkości.

KOLEJNA KATEGORYZACJA ATAKÓW



UNIKANIE -
DODGING ATTACK



PODSZYWANIE -
IMPERSONATION ATTACK

ROZWAŻANIA DLA "UNIKANIA"

$$\min_{\delta} J_D = \mathcal{L}_{\text{sim}}(f(\mathbf{x} + \delta), f(\mathbf{x}^a)), \quad (1)$$

where \mathcal{L}_{sim} denotes the attack loss of measuring the pair-wise similarity, e.g., the cosine loss $\mathcal{L}_{\text{sim}} = \cos(f(\mathbf{x} + \delta), f(\mathbf{x}^a))$. In contrast,

ROZWAŻANIA DLA "PODSZYWANIA"

$$\min_{\delta} J_l = -\mathcal{L}_{\text{sim}}(f(\mathbf{x} + \delta), f(\mathbf{x}^a)).$$

PRZYZIEMNE UTRUDNIENIA I ZAŁOŻENIA DLA NAKLEJEK

- Nie powinna zakrywać całej twarzy
- Należy wziąć pod uwagę ewentualne problemy przy drukowaniu
- Należy uznać problemy fizyczne - oświetlenie, mimikę itd.



PODVFACE

DZIAŁANIE

- Założenia: prostokątna naklejka, przyklejana na czole.

$$\mathbf{x}^{adv} = \mathcal{T}^B((1 - M) \circ \mathbf{x} + M \circ \mathcal{T}^A(f_{D2P}(\delta))) + \mathbf{v}, \quad (3)$$

where $\mathbf{v} \sim \mathcal{N}(\mu, \sigma)$ is a random Gaussian noise with mean μ and standard deviation σ . $M = \mathcal{T}^A(m)$ and \mathbf{x} is a randomly selected facial image. More details of each module are introduced as follows.

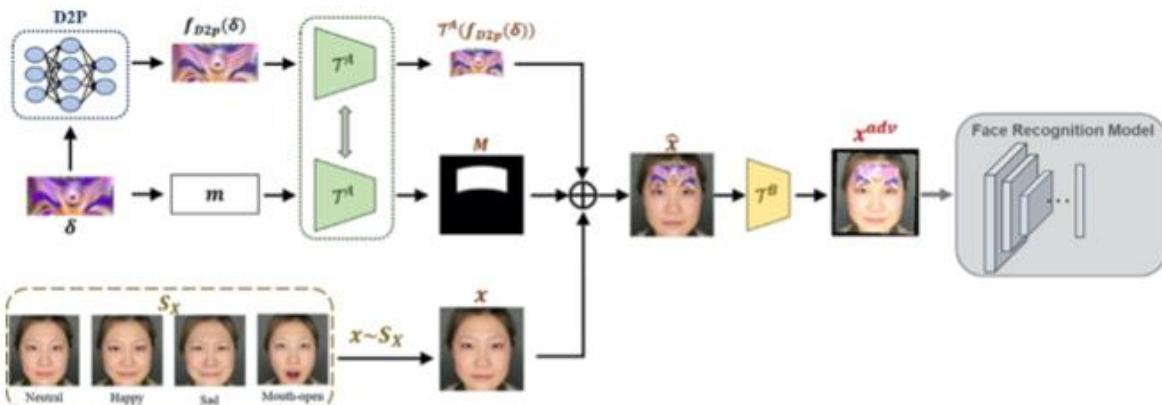


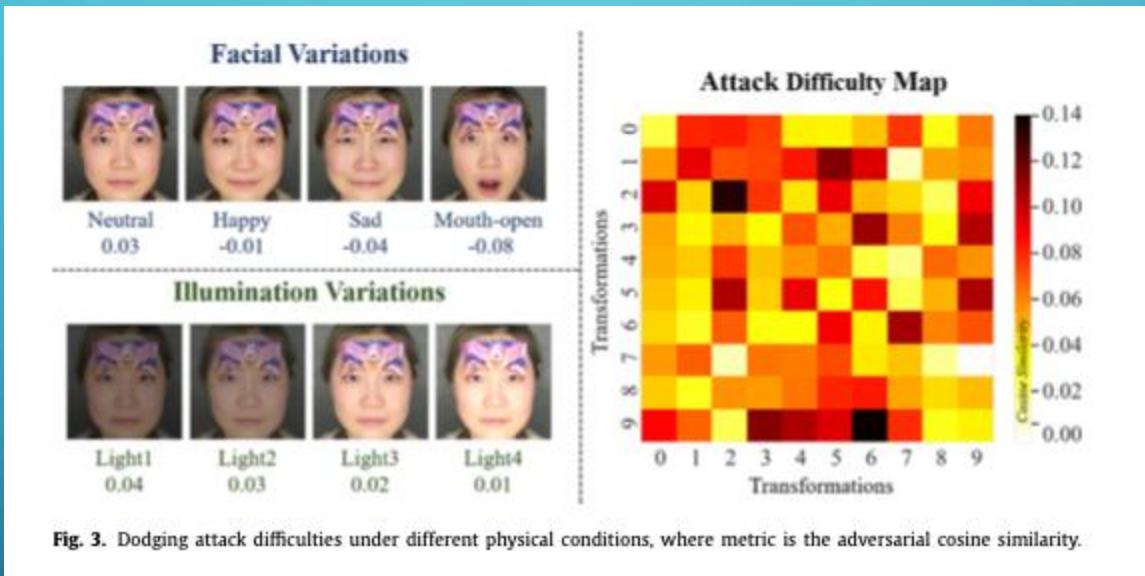
Fig. 1. Overall framework of the proposed robust PadvFace, where 'D2P' denotes Digital-to-Physical Module, ' \mathcal{T}^A ' denotes Sticker Transformation Module, and ' \mathcal{T}^B ' conducts transformations on adversarial faces. Summarized environment variations of each module are provided in Table 1.



DYGRESJA: ABERRACJA CHROMATYCZNA W D2P

- Główne jej rodzaje to:
 - błędy drukowania (np. odchylenia kolorów)
 - błędy aparatów, złe oświetlenie, szумy

NIELINIOWOŚĆ WNIOSKÓW





ALGORYTM DO
OPTYMALIZACJI
NAKLEJEK OD
ŁATWIEJSZYCH
DO BARDZIEJ ZŁOŻO
NYCH WARUNKÓW
ŚWIATA FIZYCZNEGO

ALGORYTM CAA

WZÓR

$$\min_{\delta, p_i \in [0, 1]} \frac{1}{n} \sum_{k_i \in \mathcal{K}} \left\{ p_i \mathcal{L}_{sim, k_i} + \lambda g(p_i) \right\} + \alpha \mathcal{L}_{TV}(\delta), \quad (5)$$

where $g(p_i) = \frac{1}{2} p_i^2 - p_i$ is a regularizer and $\lambda > 0$ is a curriculum parameter.

Dla danego p , możemy zredukować do:

$$\min_{\delta} \frac{1}{n} \sum_{k_i \in \mathcal{K}} p_i \mathcal{L}_{sim, k_i} + \alpha \mathcal{L}_{TV}(\delta),$$

Dla δ :

$$\min_{p_i \in [0,1]} p_i \mathcal{L}_{sim, k_i} + \lambda (\frac{1}{2} p_i^2 - p_i),$$

STÅD...

$$\min_{p_i \in [0,1]} p_i \mathcal{L}_{\text{sim}, k_i} + \lambda (\frac{1}{2} p_i^2 - p_i),$$



$$p_i^* = 1 - \frac{\mathcal{L}_{\text{sim}, k_i}}{\lambda}.$$

ALGORYTM

Algorithm 1 Curriculum adversarial attack.

Require: Attacked face recognition model f , physical transformations $\mathcal{K} = \{k_i\}_{i=1}^n$, anchor image \mathbf{x}^a , initial sticker δ , curriculum parameters $\{\lambda_t\}_{t=1}^T$ with $\lambda_1 < \lambda_2 < \dots < \lambda_T$.

```
1: for  $t = 1, \dots, T$  do
2:   for  $e = 1, \dots, E$  do.
3:     Fixed  $\delta$ , updating  $\mathbf{p}$  as  $p_i = 1 - \mathcal{L}_{sim, k_i}/\lambda_t$ .
4:     Fixed  $\mathbf{p}$ , updating  $\delta$  via gradient descent.
5:   end for
6: end for
```

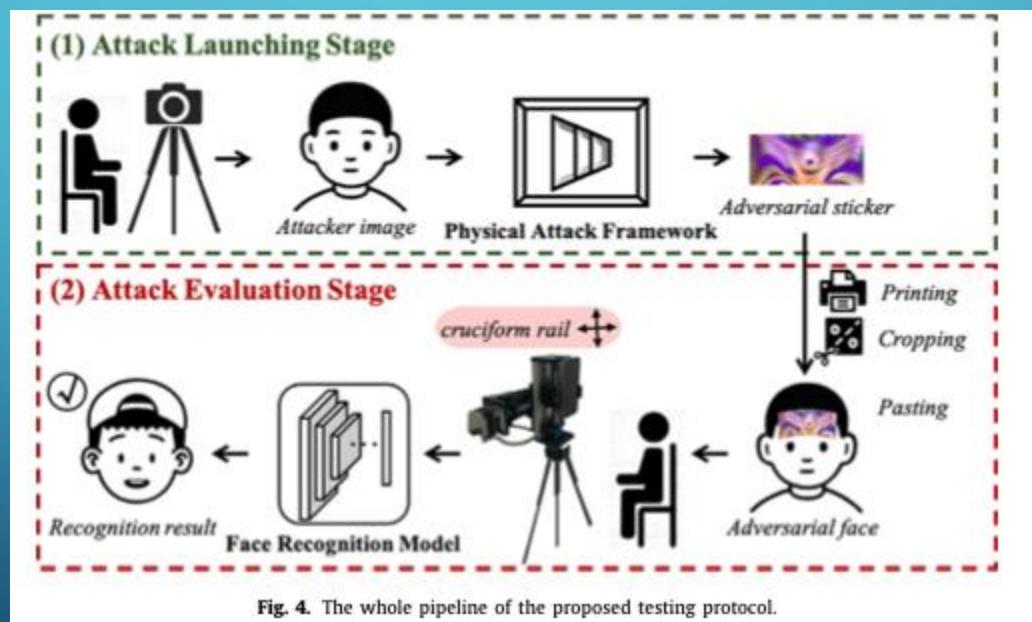
Ensure: Robust adversarial sticker δ .

ETYKA TEMATU

- Zagadnienie jest rozbudowane i wzbudza kontrowersje
- Można zgodzić się co do tego, że niewłaściwe wykorzystanie jest z góry nieetyczne
- Teorie o powszechnym wykorzystaniu mechanizmu naklejek ("oryginalny wygląd")

EKSPERYMENTY - PROTOKÓŁ BADAN

Technicznie rzecz biorąc, autorzy wykorzystali dane modele sprzętowe oraz parametry. Nie jest to jednak wiedza potrzebna w idei tematu.

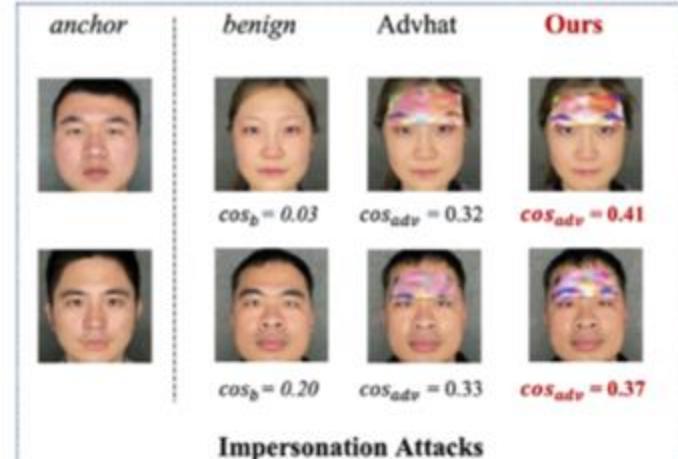
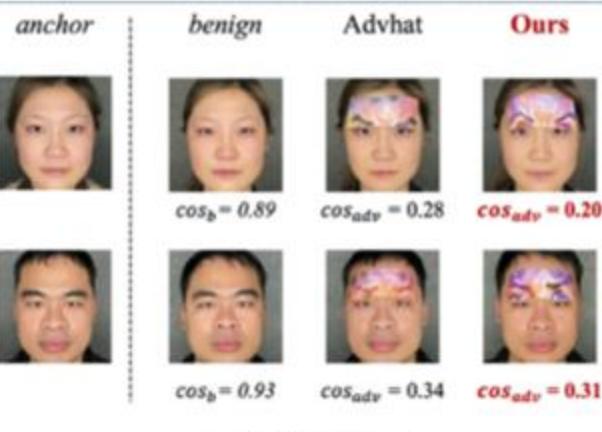


PORÓWNANIE PADVFACE DO ADVHAT

Table 2

Statistic comparison of physical evaluation cases between Advhat [9] and our PadvFace.

Methods	Experimenters	Dodging	Impersonation	Illuminations	Face variations	Poses	# Images
Advhat	10	✓	✗	3	1	8	128
Ours	10	✓	✓	3	4	~ 35	5880



EKSPERYMENTY

Table 4

Results of Advhat and Ours (PadvFace-B) on dodging and impersonation attacks. Best results are in bold. Each ID pair in impersonation attacks indicates ‘attacker → victim’.

Dodging attack (↓)				Impersonation attack (↑)			
ID	benign (\cos_b)	Advhat (\cos_{adv})	Ours (\cos_{adv})	ID	benign (\cos_b)	Advhat (\cos_{adv})	Ours (\cos_{adv})
01	0.91	0.29	0.27	01 → 02	0.16	0.28	0.46
02	0.94	0.43	0.33	02 → 03	0.09	0.26	0.35
03	0.91	0.33	0.33	03 → 04	-0.08	0.17	0.21
04	0.88	0.42	0.25	04 → 09	0.12	0.19	0.21
05	0.94	0.60	0.51	05 → 04	0.04	0.21	0.24
06	0.93	0.38	0.32	06 → 07	0.10	0.26	0.29
07	0.93	0.34	0.31	07 → 08	0.20	0.33	0.37
08	0.95	0.32	0.26	08 → 09	0.12	0.24	0.33
09	0.93	0.37	0.28	09 → 01	0.09	0.19	0.19
10	0.89	0.28	0.20	10 → 03	0.03	0.32	0.41
Average	0.92	0.38	0.31	Average	0.09	0.24	0.30

EKSPERYMENTY

Table 5

Comparison of dodging attacks with the neutral expression under illumination variations. Metric is ‘benign (\cos_b) | PadvFace-B (\cos_{adv}) | PadvFace-F (\cos_{adv})’. Best results are in bold.

ID	Dodging attack (↓)			Average (ID)
	4	7	9	
Dark	0.81 0.48 0.38	0.92 0.38 0.34	0.91 0.56 0.48	0.88 / 0.47 / 0.40
Normal	0.83 0.50 0.38	0.95 0.44 0.38	0.93 0.50 0.46	0.90 / 0.48 / 0.41
Light	0.81 0.49 0.36	0.96 0.39 0.29	0.90 0.49 0.43	0.89 / 0.46 / 0.36
Average (Illus)	0.82 / 0.49 / 0.37	0.94 / 0.40 / 0.34	0.92 / 0.52 / 0.45	-

Table 6

Comparison of impersonation attacks with the neutral expression under illumination variations. Metric is ‘benign (\cos_b) | PadvFace-B (\cos_{adv}) | PadvFace-F (\cos_{adv})’. Best results are in bold.

ID	Impersonation attack (↑)			Average (ID)
	07 → 08	08 → 09	09 → 10	
Dark	0.21 0.36 0.37	0.10 0.33 0.39	0.03 0.32 0.40	0.11 / 0.34 / 0.39
Normal	0.20 0.37 0.40	0.12 0.33 0.37	0.03 0.33 0.41	0.12 / 0.34 / 0.39
Light	0.22 0.36 0.39	0.14 0.37 0.42	0.04 0.37 0.41	0.13 / 0.37 / 0.41
Average (Illus)	0.21 / 0.36 / 0.39	0.12 / 0.35 / 0.40	0.03 / 0.34 / 0.41	-

EKSPERYMENTY

Table 7

Comparison of dodging attacks with internal facial variations under the normal illumination. Metric is 'benign (\cos_b) | PadvFace-B (\cos_{adv}) | PadvFace-F (\cos_{adv})'. Best results are in Bold.

ID	Dodging attack (↓)			Average (ID)
	04	07	09	
Happy	0.80 0.41 0.32	0.94 0.40 0.34	0.90 0.51 0.49	0.88 / 0.44 / 0.38
Sad	0.66 0.29 0.18	0.93 0.43 0.35	0.79 0.47 0.43	0.79 / 0.40 / 0.32
Neutral	0.83 0.50 0.38	0.95 0.44 0.38	0.93 0.50 0.46	0.90 / 0.48 / 0.41
Mouth-open	0.78 0.47 0.32	0.91 0.45 0.35	0.82 0.65 0.54	0.84 / 0.52 / 0.40
Average (Face)	0.77 / 0.42 / 0.30	0.93 / 0.43 / 0.35	0.86 / 0.53 / 0.48	-

Table 8

Comparison of impersonation attacks with internal facial variations under the normal illumination. Metric is 'benign (\cos_b) | PadvFace-B (\cos_{adv}) | PadvFace-F (\cos_{adv})'. Best results are in Bold.

ID	Impersonation attack (↑)			Average (ID)
	07 → 08	08 → 09	09 → 10	
Happy	0.22 0.36 0.39	0.10 0.32 0.39	0.02 0.34 0.37	0.11 / 0.34 / 0.38
Sad	0.18 0.22 0.31	0.12 0.36 0.40	0.07 0.32 0.35	0.12 / 0.30 / 0.35
Neutral	0.20 0.37 0.40	0.12 0.33 0.37	0.03 0.33 0.41	0.12 / 0.34 / 0.39
Mouth-open	0.21 0.34 0.38	0.15 0.34 0.39	0.06 0.26 0.29	0.14 / 0.31 / 0.35
Average (Face)	0.20 / 0.32 / 0.37	0.12 / 0.34 / 0.39	0.04 / 0.31 / 0.35	-

EKSPERYMENTY

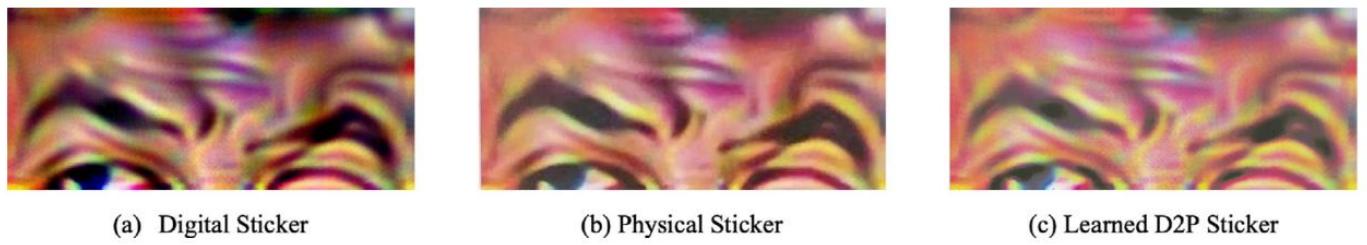


Table 9

Performance of D2P module over adversarial sticker pairs. Higher MSSIM \uparrow and lower PSNR \downarrow and MSE \downarrow denote better results. Best results are shown in bold.

Sticker metrics	PSNR (dB) \downarrow	MSSIM \uparrow	MSE \downarrow
Digital-to-Physical	18.27	0.55	0.014
D2P-to-Physical	21.58	0.62	0.006

Table 10

Ablation study of D2P module.

ID	benign	w/o D2P	w/ D2P
01	0.87	0.29	0.22
02	0.95	0.38	0.32
10	0.89	0.19	0.16

KONKLUZJA

- W tej pracy została zbadana podatność rozpoznawania twarzy w świecie fizycznym, używając do tego specjalnie przygotowanych naklejek. Została podana nowa metoda uwzględniająca różne warunki jak emocje czy natężenie oświetlenia poparte odpowiednimi eksperymentami, które pokazały efektywność nowej metody zarówno dla uniku i podszywania tychże ataków.



KONIEC PIERWSZEGO TEMATU

- Dziękujemy za uwagę.

Możemy przejść teraz do następnej
pracy...

LEARNING A DEEP DUAL-LEVEL NETWORK FOR ROBUST DEEPFAKE DETECTION

ADAM WIELICZKO, KONRAD NOWAK

ABSTRACT

- Wprowadzenie
- Powiązane prace
- Nowa metoda

WPROWADZENIE



Przykład DeepFake

Czym jest DeepFake?

- Technika wykorzystująca sztuczną inteligencję do zamiany twarzy człowieka na video na inną twarz, zachowując przy tym mimikę i oświetlenie.

Powód pracy

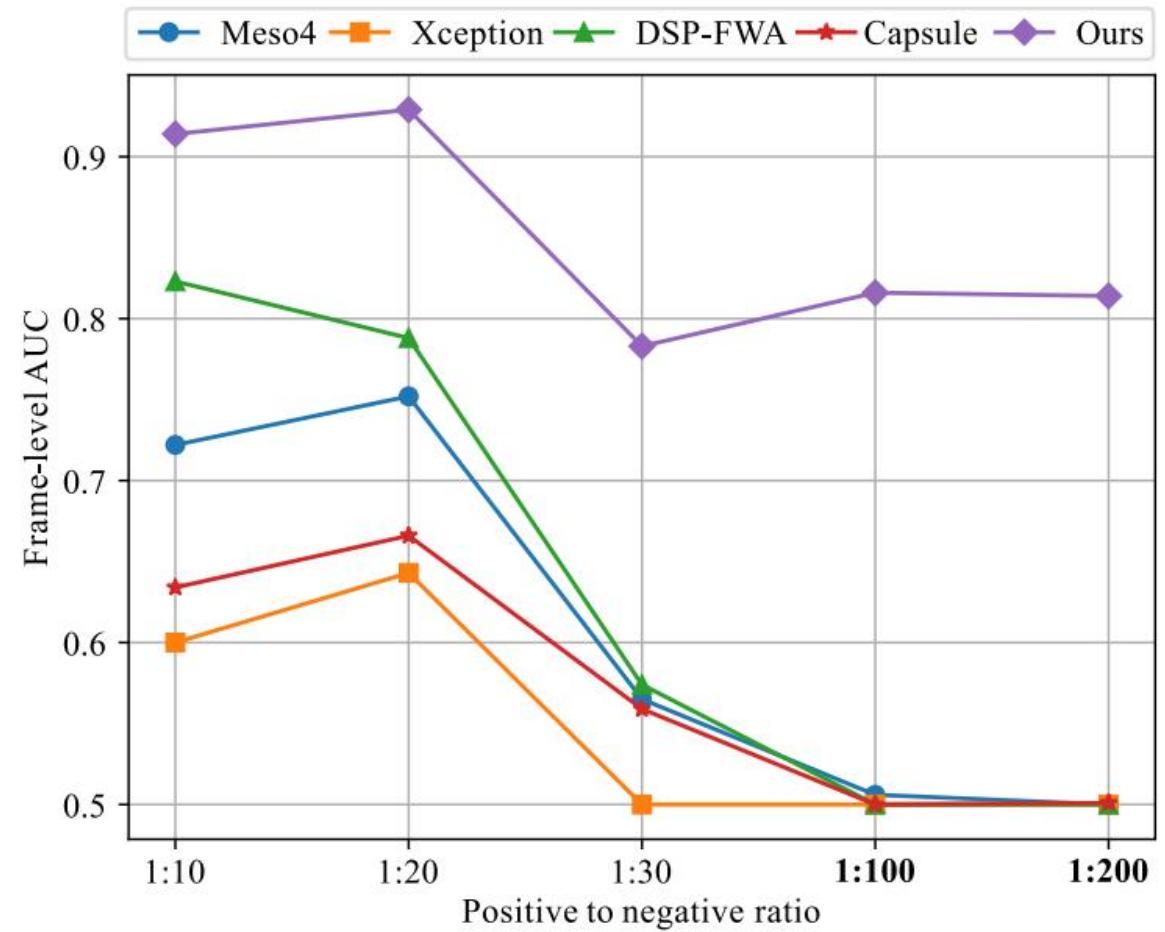
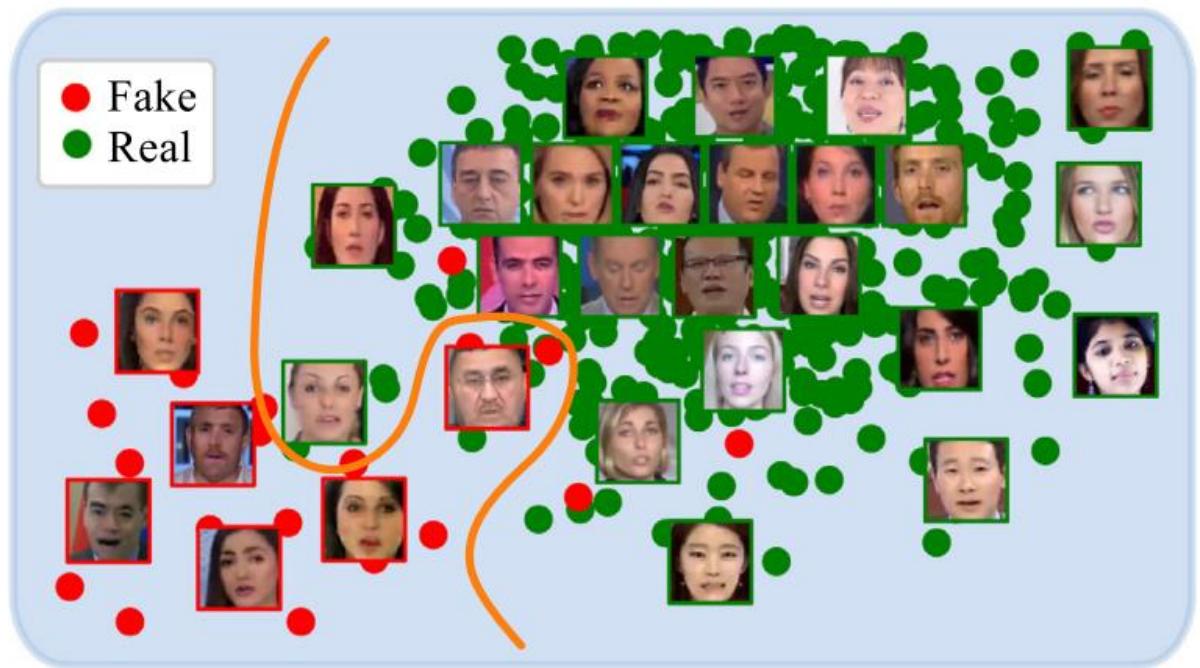
- Wraz z rozwojem technologii, powstają coraz bardziej przekonujące deepfaki.
- Troska o bezpieczeństwo (kradzież tożsamości i związane z tym oszustwa).
- Obecne metody wykrywania deepfaków nie są tak skuteczne jak powinny być.

- W odpowiedzi na coraz bardziej przekonujące deepfaki wykorzystujące sztuczną inteligencję naprzeciw wychodzi sztuczna inteligencja pozwalająca na wykrywanie tych deepfaków.
- Większość prac traktuje problem wykrywania jako klasyfikację binarną.
- Istniejące metody na wykrywanie fałszywych twarzy to np. MesoNet oparty na CNN wyspecjalizowanej specjalnie DeepFake.
- Istnieją też modele takie jak Xception, Capsule, DSP-FWA.
- Fantastyczne wyniki, które upadają po zderzeniu się z rzeczywistością...

DLACZEGO?

Powodów można szukać w:

- Znacznej dysproporcji w danych między prawdziwymi obrazami a fałszywymi.
- Wyrównywanie danych różnymi metodami (np. data augmentation albo oversampling) nie rozwiązuje problemu, gdyż zakrywia obraz rzeczywistości.



KRÓTKO O DEEPFAKE

- Oparte są one na autoenkoderach i GAN, w skład którego wchodzi generator G tworzący fałszywe twarze z wektorów wejściowych i dyskryminator D , który rozróżnia fałszywe twarze od prawdziwych.
- Dwa podejścia: face swapping (podmiana na twarzy i operowania na tym) oraz face re-enactment (rekonstrukcja twarzy na podstawie danych wejściowych).



DWA TYPY METOD

FRAME-LEVEL METHODS

- Te metody biorą z filmiku wszystkie klatki i analizują je osobno, na końcu np. uśredniając wyniki, by dojść do finalnej konkluzji.

Metody poziomie klatek dzielą się na dwie kategorie:

- Używające sieci konwolucyjnych
- Używające autoencodera

- Sieci konwolucyjne są wykorzystywane do wydzielenia cech z obrazu i sprowadzają problem deepfaków do binarnej klasyfikacji (czy obraz jest deepfakiem lub czy też nie jest).

Różne sieci skupiają się na różnych aspektach:

- Luo używa wysokiej częstotliwości szumu do odnalezienia zakłóceń naniesionych twarzy.
- PRRNet potrafi rozpoznać relację na poziomie małych regionów pikseli.
- FakeCatcher skupia się na rozpoznawaniu biologicznych sygnałów, czyli np. emocji do wykrycia zniekształceń obrazów.

- Autoencodery kompresują dane, kodują i próbują odtworzyć obraz jako dane wyjściowe.
- Bappy przykładowo wykorzystuje framework lokalizujący zmanipulowane regiony, bazujący i long short-term memory metodach wykorzystujące cechy obszarów.
- X-ray z drugiej strony używa metody self-supervised learning do wykrywania obszarów.
- Według badań minusem autoencoderów zwłaszcza jest to, że słabo się uczą na niezbalansowanych danych.

VIDEO-LEVEL METHODS

- Te metody biorą pod uwagę cały kontekst filmiku i zależności pomiędzy klatkami, używając do tego sieci konwolucyjnych do wyciągnięcia cech i rekurencyjnych sieci neuronowych, aby wykryć niekonsekwencję między klatkami
- DeepFakesON-Phys wykrywa fałszywe twarze, używając do tego faktów związanych z rytmem serca oraz zdalnej fotopletyzmografii.

WADY, ZALETY

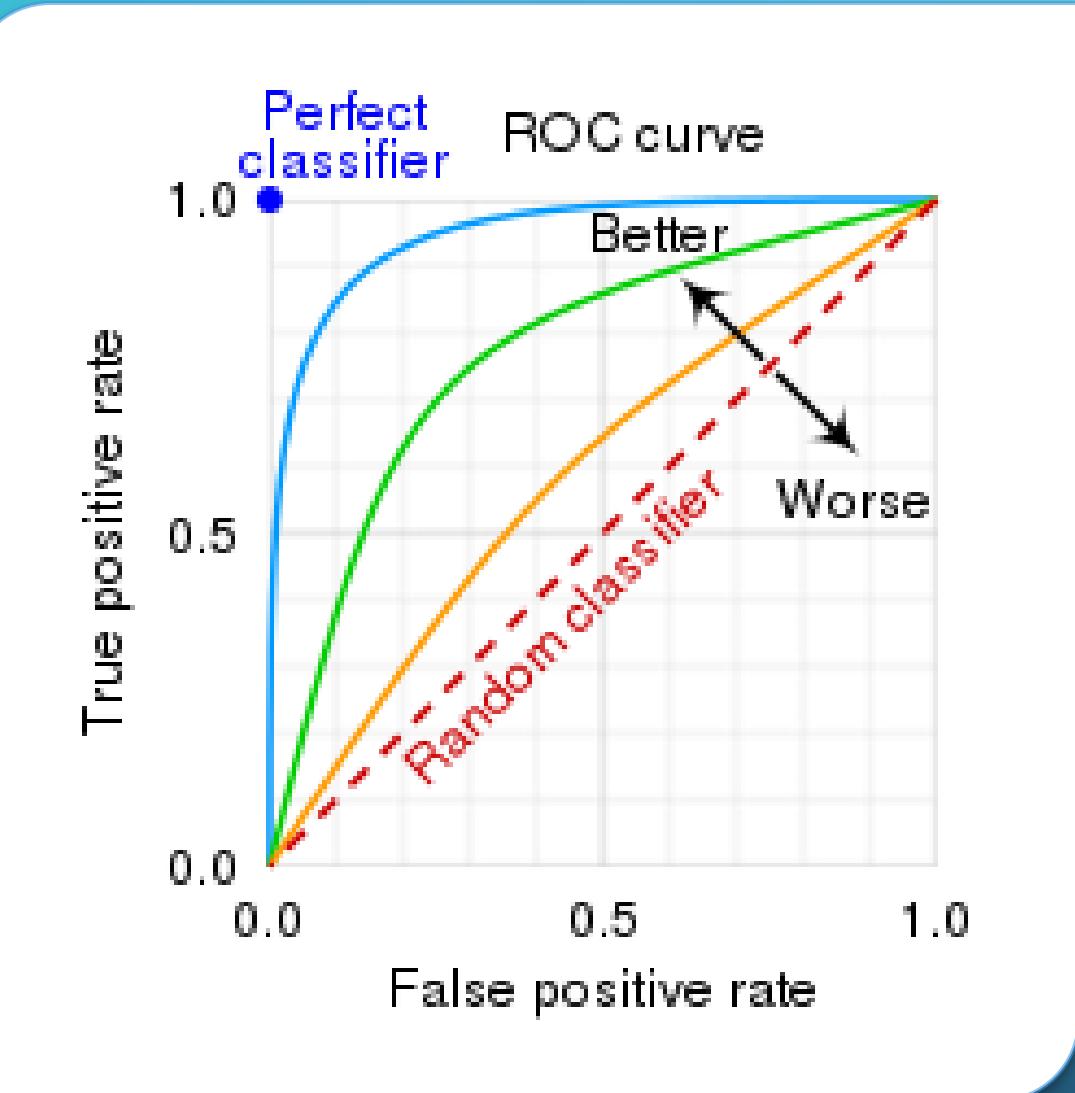
- Metody na poziomie klatek są bardzo precyzyjnie wskazać, czy dana kamera jest fałszywa, jednak nie potrafią przenosić między sobą kontekstu potrzebnego do wykrycia bardzo przekonującego deepfaka, który nienaturalnie się porusza.
- Metody na poziomie video mają zupełnie na odwrót. Nie posiadają umiejętności wykrywania w indywidualnych klatkach i mają problemy, jeśli przykładowo tylko jakąś część klatek zawiera to zniekształcenie.



WNIOSEK?

ROZWIAZANIE TRZECH PROBLEMÓW

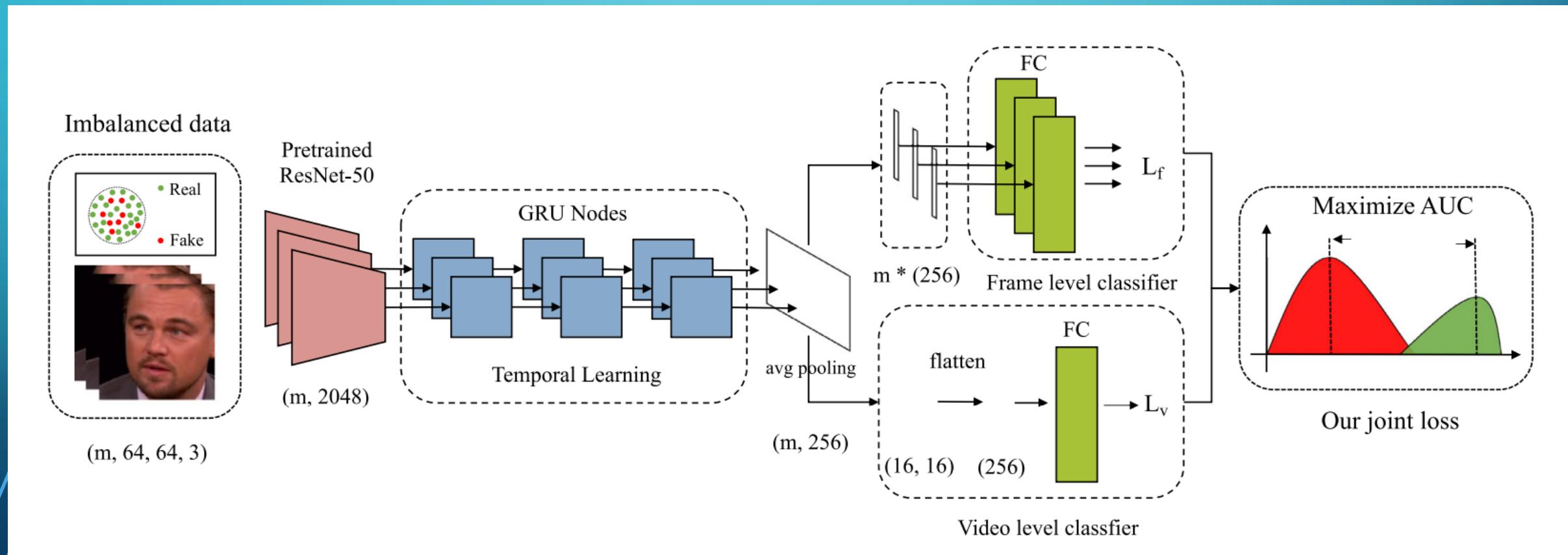
- Niebalansowany zestaw danych
- Wady metod na poziomie klatek
- Wady metod na poziomie filmiku



ANALIZA

- Jedną z najpopularniejszych metryk dla niezbalansowanych danych jest AUC score stworzony na podstawie krzywej ROC. Drugim jest accuracy.
- Potrzebujemy jednak różniczkowalną metrykę - pomoże statystyka WMW (Wilcoxon-Mann-Whitney).

PRZEBIEG METODY



1. Potrzebne są obrazy samej twarzy, więc zostaje użyty Dlib face detector, której dane są użyte jako wejście dla sieci neuronowej.
2. Pierwszym modułem sieci jest wytrenowana sieć konwolucyjna do znalezienia cech każdej twarzy, by wykryć jakieś niekonsekwencje wizualne, można do tego użyć ResNet-50 lub Xception. Wynikiem są dane zawierające domenę każdej twarzy.
3. Potem dane są przekazywane do modułu temporal learningu. Ta warstwa wykrywa migotanie klatek lub niekonsystentne ciągi klatek wybijające się na tle innych.

STATYSTYKA WMW

- Statystyka WMW jest odpowiednikiem AUC i jest wyliczana tak, gdzie:
- P – pozytywne przypadki, N – negatywne przypadki
- X - zbiór danych, Y - zbiór kategorii
- F – funkcja przewidującą wynik i-tego przypadku (i od 1 do M)

$$\begin{aligned} \text{WMW} &= \frac{1}{|\mathcal{P}| |\mathcal{N}|} \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{N}} \mathbb{I}_{[\mathcal{F}(x_i) > \mathcal{F}(x_j)]}, \quad \text{where } \mathbb{I}_{[\mathcal{F}(x_i) > \mathcal{F}(x_j)]} \\ &= \begin{cases} 1, & \mathcal{F}(x_i) > \mathcal{F}(x_j), \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

RÓŻNICZKOWALNE PRZYBLIŻENIE STATYSTYKI WMW

$$\mathcal{L}_{AUC}(\mathcal{F}(x), y) = \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{N}} R(\mathcal{F}(x_i), \mathcal{F}(x_j)).$$

- Różniczkowalna przybliżona wersja statystyki WMW pozwalająca uzyskać jak najbliżej AUC uwzględniając przy tym, że dane są niebalansowane.
- $0 < \gamma \leq 1$ i $p > 1$ to hiperparametry.

$$R(\mathcal{F}(x_i), \mathcal{F}(x_j)) = \begin{cases} (-(\mathcal{F}(x_i) - \mathcal{F}(x_j) - \gamma))^p, & \mathcal{F}(x_i) - \mathcal{F}(x_j) < \gamma, \\ 0, & \text{otherwise,} \end{cases}$$

FUNKCJA STRATY

$$\begin{aligned}\mathcal{L}_v &= \alpha \mathcal{L}_{BCE}(\mathcal{F}_v(\mathcal{V}), Y) + (1 - \alpha) \mathcal{L}_{AUC}(\mathcal{F}_v(\mathcal{V}), Y), \\ \mathcal{L}_f &= \alpha \mathcal{L}_{BCE}(\mathcal{F}_f(\mathcal{I}), \hat{Y}) + (1 - \alpha) \mathcal{L}_{AUC}(\mathcal{F}_f(\mathcal{I}), \hat{Y}),\end{aligned}$$

- \mathcal{V} – video, \mathcal{I} – klatki, Y - etykiety
- α - współczynnik skali dla balansu
- β - hiperparametr dla skalowania \mathcal{L}_v i \mathcal{L}_f

$$\mathcal{L} = \beta \cdot \mathcal{L}_v + (1 - \beta) \cdot \mathcal{L}_f,$$

OCENA EKSPERYMENTU

Model utworzony przez autorów pracy posiada dwie główne fazy eksperymentowe:

- Porównanie z obecnymi metodami na całych datasetach
- Badanie jakości nauczania modelu na w niebalansowanych datasetach

DATASETY

- Pierwsza faza korzysta z datasetu Celeb-DF oraz FaceForensics++
- Druga natomiast używa Celeb-DF i DFDC

W przypadku drugiego podzielono datasety na pięć posiadających nierównomierny rozkład poprawnych nagrań do DeepFakowych

ALGORYTMY DO DEEPFEJKÓW ORAZ UŻYTA JAKOŚĆ NAGRAŃ

Algorytmy:

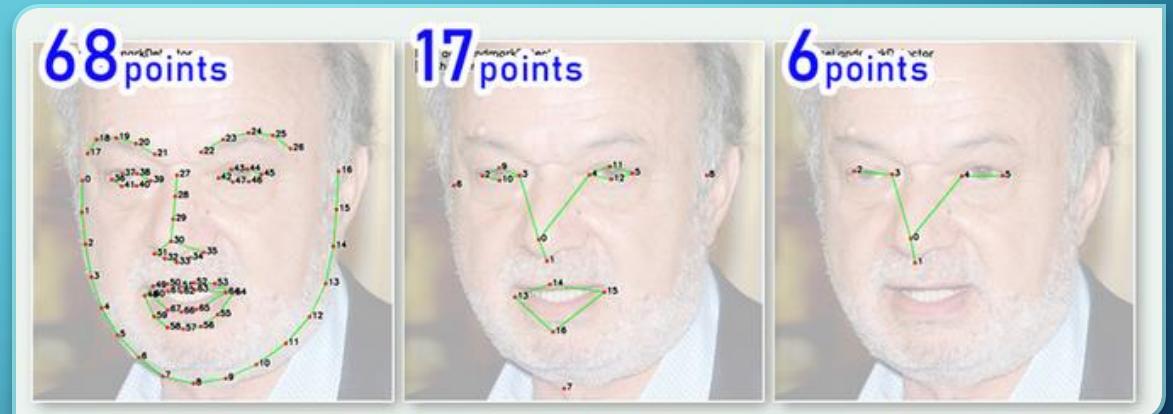
- DeepFakes
- FaceSwap
- Face2Face
- Neural Texture

Jakość nagrań:

- raw
- c23
- c40

PREPROCESSING

- Dlib (widoczny po prawo)
- Przeskalowanie do 64x64 piksele
- Normalizacja
- Grupowanie
- Wejściowa ilość klatek równa 300



INNE METODY WYKRYWANIA DEEPFAKÓW

Na poziomie klatkowym działają:

- MesoNet
- Xception
- Capsule
- DSP-FWA

INNE METODY WYKRYWANIA DEEPFAKÓW

Na poziomie nagrania działają:

- CNN+LSTM
- CNN+GRU

DETALE IMPLEMENTACYJNE

- Backbone: ResNet-50 z modelem przetrenowanym na ImageNecie
- Optymalizator: Adam
- Learning rate: $1*10^{-4}$

Użyte Metryki:

- Accuracy, AUC (ROC Curve), F1 score, recall, precision

DETALE DOTYCZĄCE DATASETÓW

Datasets	Train.		Test		Ratio (Real:Fake)
	Real	Fake	Real	Fake	
Celeb-30	712	23	178	5	30:1
Celeb-20	712	35	178	8	20:1
Celeb-10	712	71	178	17	10:1
DFDC-100	500	5	300	3	100:1
DFDC-200	1000	5	600	3	200:1

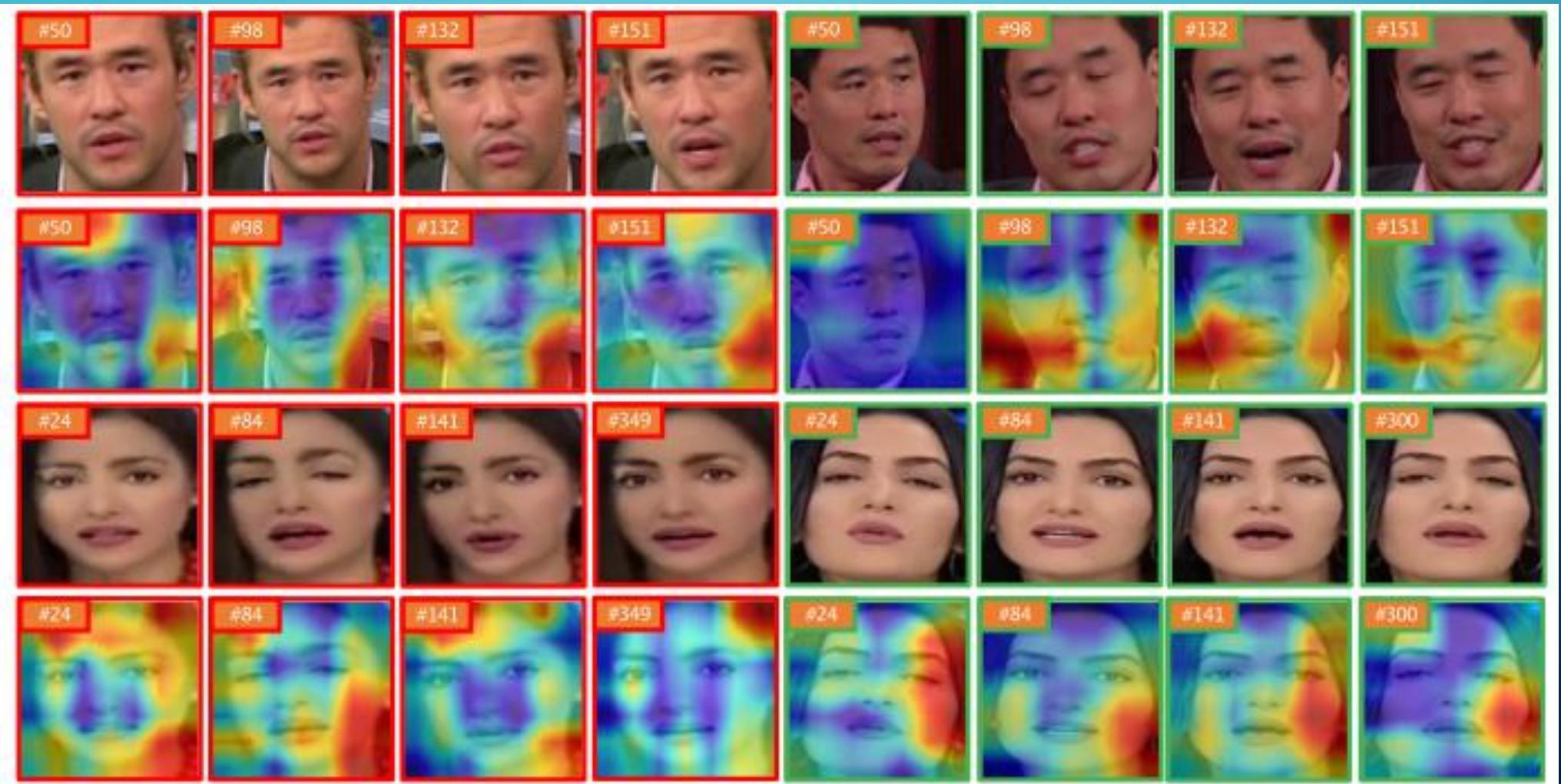
WYNIKI METOD NA FACEFORENSICSIE++

Method	Video Level					Frame Level				
	ACC	AUC	F1	Precision	Recall	ACC	AUC	F1	Precision	Recall
DSP-FWA [5] ^a	-	-	-	-	-	51.2	51.4	67.0	50.9	100
Meso4 [2]	-	-	-	-	-	70.4 ± 1.02	77.6 ± 0.85	72.7 ± 2.69	63.2 ± 1.42	85.6 ± 1.01
MesoInception4 [2]	-	-	-	-	-	82.3 ± 1.32	83.9 ± 0.93	79.4 ± 2.23	77.3 ± 1.57	84.3 ± 1.69
Xception [3]	-	-	-	-	-	83.5 ± 0.75	89.9 ± 0.53	80.7 ± 0.79	81.6 ± 0.45	92.0 ± 0.78
Capsule [6]	-	-	-	-	-	84.6 ± 0.12	84.7 ± 0.13	84.6 ± 0.13	87.9 ± 0.58	81.4 ± 0.62
CNN+LSTM	89.2 ± 0.23	88.7 ± 0.31	86.8 ± 0.59	89.4 ± 0.61	87.3 ± 0.45	-	-	-	-	-
CNN+GRU	91.2 ± 0.65	89.9 ± 0.27	86.6 ± 0.74	88.6 ± 0.53	92.4 ± 0.39	-	-	-	-	-
Ours	95.6 ± 0.29	99.0 ± 0.11	93.2 ± 0.63	94.3 ± 0.33	98.2 ± 0.67	94.8 ± 0.25	98.4 ± 0.23	95.9 ± 0.73	94.7 ± 0.53	99.4 ± 0.64

WYNIKI METOD NA CELEB-DF

Method	Video Level					Frame Level				
	ACC	AUC	F1	Precision	Recall	ACC	AUC	F1	Precision	Recall
DSP-FWA [5] ^a	-	-	-	-	-	65.3	50.0	79.1	64.8	99.3
Meso4 [2]	-	-	-	-	-	72.0 ± 0.99	83.0 ± 1.65	35.9 ± 4.17	93.5 ± 1.23	22.3 ± 3.27
MesolInception4 [2]	-	-	-	-	-	85.3 ± 1.53	89.7 ± 2.11	76.3 ± 3.80	88.1 ± 2.41	66.4 ± 5.41
Xception [3]	-	-	-	-	-	93.6 ± 0.15	91.4 ± 0.26	89.9 ± 0.59	97.9 ± 0.21	83.7 ± 0.62
Capsule [6]	-	-	-	-	-	91.0 ± 0.35	88.5 ± 0.26	86.2 ± 0.51	93.2 ± 0.92	80.2 ± 0.63
CNN+LSTM	87.4 ± 0.23	85.5 ± 0.35	89.2 ± 0.54	90.1 ± 0.82	92.7 ± 0.66	-	-	-	-	-
CNN+GRU	92.3 ± 0.17	89.9 ± 0.37	93.2 ± 0.45	91.7 ± 0.76	94.9 ± 0.68	-	-	-	-	-
Ours	96.5 ± 0.19	98.9 ± 0.45	95.6 ± 0.34	96.4 ± 0.69	98.6 ± 0.72	96.2 ± 0.26	97.4 ± 0.29	94.3 ± 0.56	98.2 ± 0.58	98.8 ± 0.66

PODGLĄD JAK MODEL ROZPOZNAJE DEEPFAKI

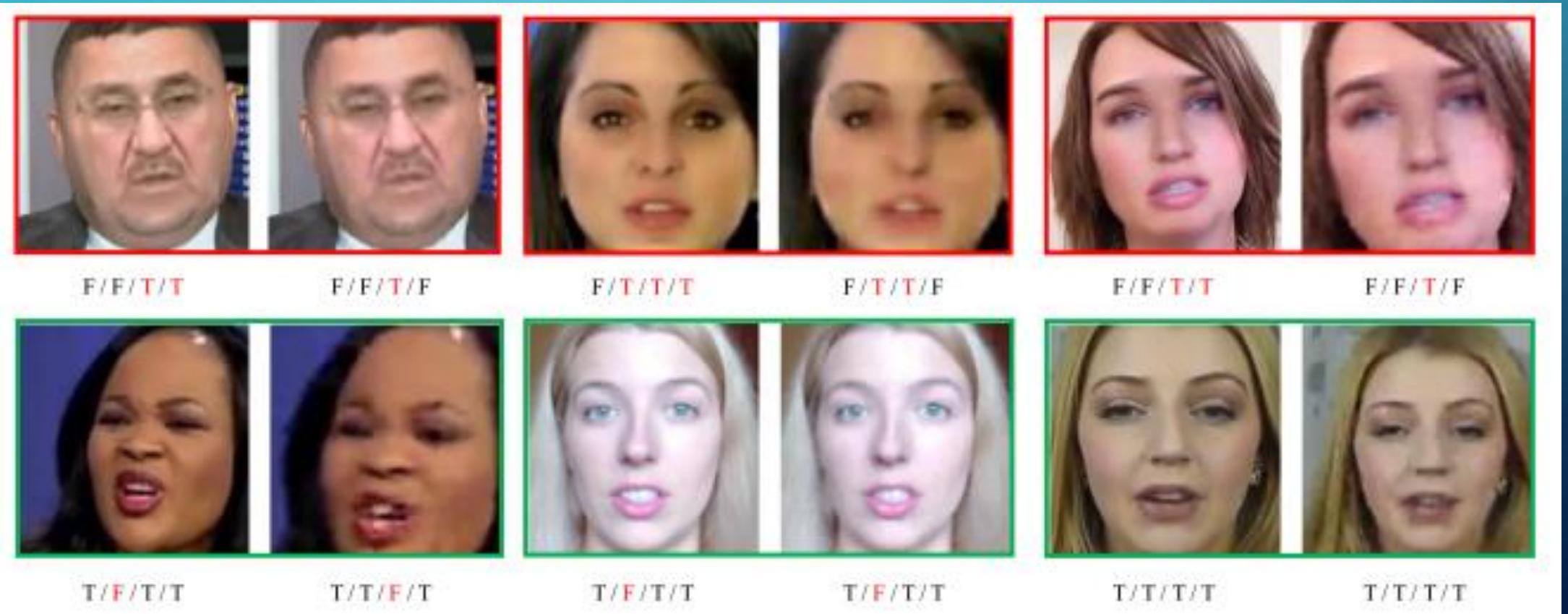


WYNIKI NA CROSS-DATASECIE

Method	Input Size	Frame-level AUC	
		FaceForensics+	Celeb-DF
Meso4 [2]	64	73.9	62.1
Mesolnception4 [2]		86.3	57.9
Xception [3]		91.2	55.4
Capsule [6]		87.3	58.7
Ours		99.2	70.3
Two-branch [10]	224	93.2	73.4
Face X-ray [23]	256	99.2	74.8
High-frequency [21]	256	99.3	79.4

WYNIKI METOD NA FACEFORENSICISE Z RÓŻNA KOMPRESJĄ

Method	Video-level AUC			Frame-level AUC		
	raw	c23	c40	raw	c23	c40
DSP-PWA	-	-	-	51.0	51.4	50.3
Meso4	-	-	-	74.6	73.9	73.1
MesolInception	-	-	-	86.3	86.3	83.0
Xception	-	-	-	91.0	91.2	87.8
Capsule	-	-	-	86.8	87.3	84.2
CNN+LSTM	87.6	88.3	82.2	-	-	-
CNN+GRU	88.6	89.4	84.3	-	-	-
Ours	99.2	99.4	95.4	98.9	99.2	95.2



WYNIKI DLA UŻYTEGO "ABLATION STUDY"

Method	Video-level		Frame-level	
	ACC	AUC	ACC	AUC
CNN +FLC	-	-	74.5	68.2
CNN +FLC,+AUC loss	-	-	81.0	84.9
CNN +VLC	66.1	52.0	-	-
CNN +VLC,+TLM	91.9	89.9	-	-
CNN +VLC,+TLM,+FLC	97.2	96.0	94.9	93.4
CNN +VLC,+TLM,+FLC,+AUC loss	96.6	99.4	95.8	98.3

WYNIKI DLA CELEB-30

Method	Video Level					Frame Level				
	ACC	AUC	F1	P	R	ACC	AUC	F1	P	R
DSP-FWA [5]	-	-	-	-	-	92.2	57.4	96.2	94.8	94.8
Meso4 [2]	-	-	-	-	-	97.3	56.5	98.4	95.3	100
MesoInception4 [2]	-	-	-	-	-	97.3	50.0	98.6	96.2	100
Capsule [6]	-	-	-	-	-	95.1	55.9	96.9	98.4	97.1
Xception [3]	-	-	-	-	-	97.3	50.0	98.1	97.3	100
Xception w AUC loss	-	-	-	-	-	97.3	54.8	99.2	96.6	100
BBN [28]	-	-	-	-	-	95.4	50.0	97.1	97.3	97.8
CNN+LSTM	94.9	58.7	98.8	95.3	98.6	-	-	-	-	-
CNN+GRU	97.1	59.6	97.1	97.0	97.0	-	-	-	-	-
Ours w/o AUC loss	96.2	70.4	98.0	96.2	98.3	95.7	72.1	98.0	95.3	98.3
Ours w Focal Loss [27]	97.1	72.4	92.4	96.6	98.9	91.0	62.1	97.8	96.7	98.8
Ours (default γ)	97.2	76.9	98.6	96.7	100	95.1	78.0	98.6	96.7	100
Ours ($\gamma = 0.1$)	97.2	73.7	98.6	96.7	100	95.1	78.3	98.6	96.7	100

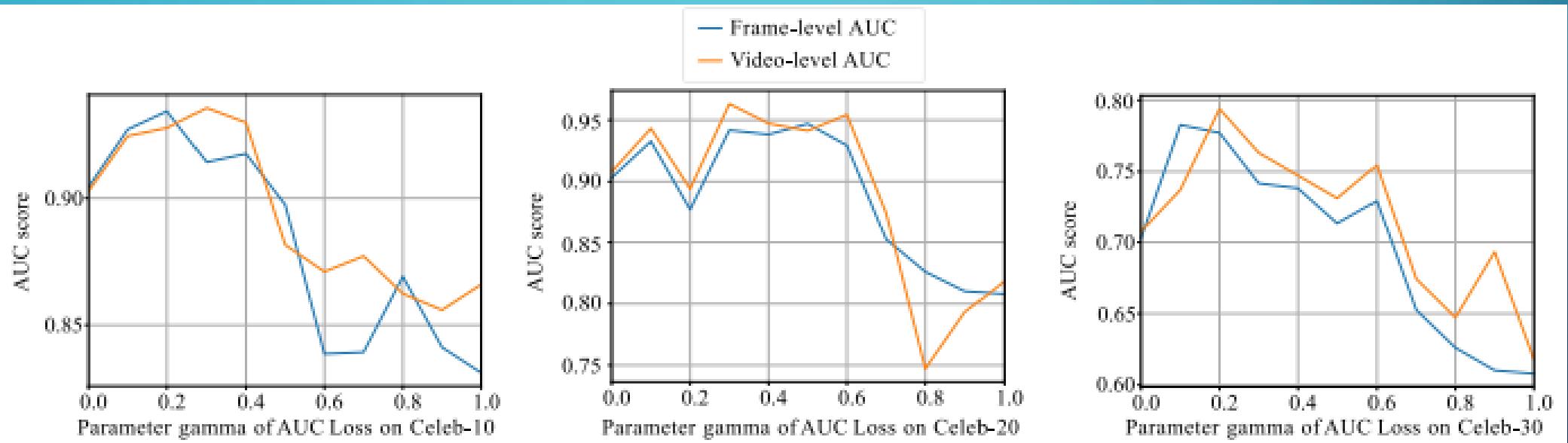
WYNIKI DLA CELEB-20

Method	Video level					Frame Level				
	ACC	AUC	F1	P	R	ACC	AUC	F1	P	R
DSP-FWA [5]	-	-	-	-	-	96.8	78.8	98.3	97.1	98.4
Meso4 [2]	-	-	-	-	-	96.4	75.2	98.2	94.1	100
Mesolnception4 [2]	-	-	-	-	-	93.0	68.8	96.3	95.3	95.3
Capsule [6]	-	-	-	-	-	95.6	66.6	97.7	97.1	98.3
Xception [3]	-	-	-	-	-	93.8	64.3	97.0	96.9	97.1
Xception w AUC loss	-	-	-	-	-	95.6	78.8	97.7	96.3	100
CNN+LSTM	70.3	84.4	81.7	71.0	69.1	-	-	-	-	-
CNN+GRU	87.6	87.2	93.1	86.9	87.6	-	-	-	-	-
Ours w/o AUC loss	96.2	90.1	98.1	95.4	98.8	95.7	87.7	98.1	96.2	98.8
Ours w Focal Loss [27]	95.7	94.9	98.2	97.2	99.0	95.7	92.6	98.3	97.3	99.5
Ours (default γ)	95.7	95.2	97.8	95.7	100	95.7	92.7	97.8	95.7	100
Ours ($\gamma = 0.6$)	95.7	95.4	97.8	95.7	100	95.7	92.9	97.8	95.7	100

WYNIKI DLA CELEB-10

Method	Video Level					Frame Level				
	ACC	AUC	F1	P	R	ACC	AUC	F1	P	R
DSP-FWA [5]	-	-	-	-	-	94.2	82.3	96.8	97.2	96.7
Meso4 [2]	-	-	-	-	-	91.4	72.2	95.5	90.1	100
Mesolnception4 [2]	-	-	-	-	-	88.4	54.2	91.8	89.3	92.1
Capsule [6]	-	-	-	-	-	85.0	63.4	91.2	92.2	89.6
Xception [3]	-	-	-	-	-	85.9	60.0	91.5	93.4	91.2
Xception w AUC loss	-	-	-	-	-	91.0	74.5	94.8	91.4	100
CNN+LSTM	85.1	78.6	91.4	84.3	86.5	-	-	-	-	-
CNN+GRU	92.1	78.6	94.7	93.5	97.1	-	-	-	-	-
Ours w/o AUC loss	88.2	87.3	92.9	87.9	91.6	88.9	88.2	92.9	88.2	91.6
Ours w Focal Loss [27]	90.3	90.5	94.3	93.4	96.4	90.0	90.3	95.4	92.5	95.7
Ours (default γ)	91.3	92.5	95.4	91.2	100	91.3	93.2	95.4	91.2	100
Ours ($\gamma = 0.3$)	91.3	93.5	95.4	91.2	100	91.3	91.4	95.4	91.2	100

WKŁAD HIPERPARAMETRU GAMMA NA WYNIKI ROC CURVE



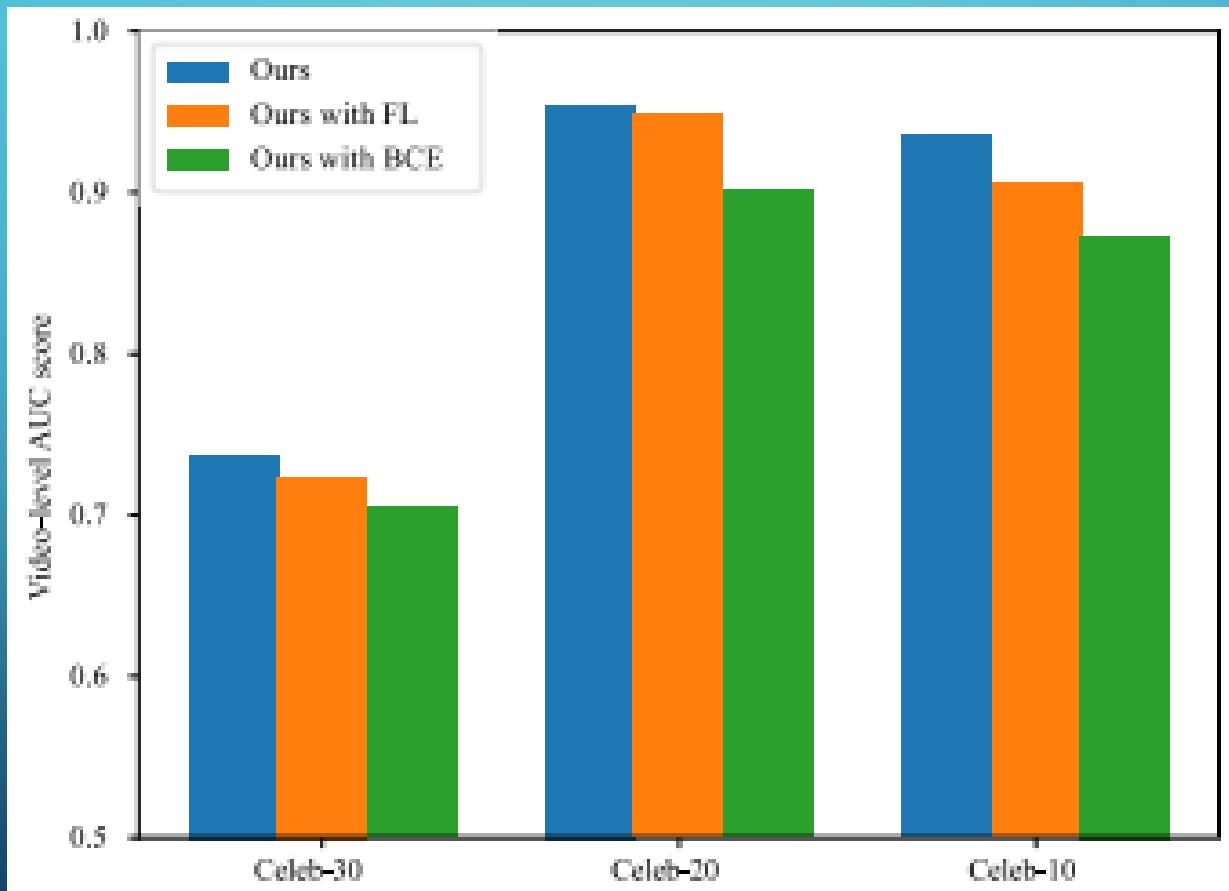
WYNIKI DLA DFDC-100

Method	Video level					Frame Level				
	ACC	AUC	F1	P	R	ACC	AUC	F1	P	R
DSP-FWA [5]	-	-	-	-	-	99.1	50.0	98.7	98.7	98.5
Meso4 [2]	-	-	-	-	-	88.5	50.6	93.8	99.0	90.2
MesoInception4 [2]	-	-	-	-	-	99.0	63.7	99.2	99.3	100
Capsule [6]	-	-	-	-	-	98.5	50.0	98.7	99.4	100
Xception [3]	-	-	-	-	-	98.7	50.0	98.7	99.0	100
CNN+GRU	98.6	50.0	98.6	99.2	100	-	-	-	-	-
Ours w/o AUC loss	98.7	65.6	98.5	99.3	98.9	99.1	65.5	98.9	98.8	100
Ours	98.7	79.9	99.1	99.1	100	99.1	81.6	99.0	99.2	100

WYNIKI DLA DFDC-200

Method	Video level					Frame Level				
	ACC	AUC	F1	P	R	ACC	AUC	F1	P	R
DSP-FWA [5]	-	-	-	-	-	99.3	50.0	99.2	99.0	100
Meso4 [2]	-	-	-	-	-	98.8	50.0	98.8	98.7	100
Mesolnception4 [2]	-	-	-	-	-	98.8	61.5	98.5	98.6	100
Capsule [6]	-	-	-	-	-	98.9	50.1	99.1	99.4	100
Xception [3]	-	-	-	-	-	98.5	50.0	99.0	99.4	100
CNN+GRU	98.8	50.0	98.5	99.1	100	-	-	-	-	-
Ours w/o AUC loss	92.9	62.6	96.6	99.4	94.2	91.3	62.2	96.3	98.7	91.7
Ours	98.9	85.6	98.6	99.0	100	98.5	81.4	99.3	98.6	100

PORÓWNANIE JAKOŚCI DZIAŁANIA NA POZIOMIE NAGRANIA DLA MODELU PROPONOWANEGO PRZEZ AUTORÓW



PODSUMOWANIE

- Bardzo efektywna funkcja kosztu, która podniosła wyniki ROC Curve
- Model tak skuteczny, że wygrywa z każdym innym zarówno na poziomie nagrania jak i klatek...
... Nawet bez wspomnianej funkcji kosztu!