

# Detekcija stvarnih i umjetno generiranih lica ljudi

1<sup>st</sup> Mateja Vuradin  
mateja.vuradin@fer.hr

2<sup>nd</sup> Lucija Prodan  
lucija.prodan@fer.hr

3<sup>rd</sup> Erika Tomakić  
erika.tomakic@fer.hr

4<sup>th</sup> Hana Ujčić  
hana.ujcic@fer.hr

5<sup>th</sup> Marta Vidas  
marta.vidas@fer.hr

**Sažetak**—Ovaj seminar istražuje primjernu modela dubokog učenja u detekciji stvarnih i umjetno generiranih lica ljudi. Isprobane su četiri različite arhitekture te su za svaku od njih istrenirana dva modela, s nasumično inicijaliziranim težinama i s predtreniranim težinama. Najveća točnost (80%) dobivena je predtreniranom EfficientNet arhitekturom.

**Index Terms**—detekcija, lice, umjetno, duboko učenje, slike

## I. UVOD I MOTIVACIJA

Razvojem tehnologija za generiranje sintetičkih slika, kao što su GAN-ovi (generativne suparničke mreže), lažna lica postaju sve realističnija. Ovo predstavlja problem za autentifikaciju jer je često teško razlikovati je li lice stvarno ili prevara, tj. prepoznati napade u sustavima autentifikacije licem [1].

Osim lažnih slika, raste broj i lažnih videa i to za čak 900%. Ovi videi omogućuju stvaranje lažnih profila na društvenih mrežama, ali i širenje raznih ideja i stavova. Često ovakvi lažni profili rade "zajedno", tj. međusobno se prate i dijele objave kako bi se stvorila iluzija o stvarnosti. Premda brojne tvrtke danas izražavaju zabrinutost o umjetno generiranih profilima i prevarama, tek njih 29% je poduzelo korake zaštite. Dodatnih 25% planira u budućnosti pobrinuti se oko ovog problema [2].

Iako se danas detekcija lažno stvorenih lica radi korištenjem GAN-ova, postoje nezaobilazni problemi. Naime, ovi modeli pate od "mode collapse" problema, tj. generiraju ograničen skup uzoraka te se ovime smanjuje varijanca uzorka. Osim toga, generirane slike imaju specifičnu raspodjelu piksela koja ne odgovara stvarnim slikama. Ove informacije mogu se iskoristiti za dobru detekciju umjetno generiranih slika [3].

Umjetno generirane slike, naravno mogu doći iz drugog izvora (stvorene od strane stručnjaka) te tako biti teže raspoznatljive. Ova činjenica i prethodno opisan problem, motivacija su za proučavanje ručno stvorenog lažnog skupa slika i stvaranje vlastitih modela dubokog učenja koji će uspješno detektirati lažne profile.

Pregled postojećih istraživanja dan je u odjeljku II, dok su u odjeljku III prikazani korišteni materijali i metode. Odjeljak IV prikazuje dobivene rezultate koji su onda uspoređeni s postojećim rezultatima u odjeljku V. Dodatna vizualizacija rada modela nalazi se u odjeljku VI.

## II. PREGLED POSTOJEĆIH PRISTUPA

Za detekciju umjetno generiranih lica koriste se tehnike dubokog učenja i računalnogvida. Mogući postupci su analiza tekture, analiza metapodataka i forenzička analiza. Na umjetno generiranim slikama često se nalaze ponavljajući uzorci, nekonistentne boje i nagli prijelazi oštirine, zbog čega je korisna analiza tekture [4]. Analizom metapodataka mogu se uočiti nekonistentnosti informacija o slici, što ukazuje

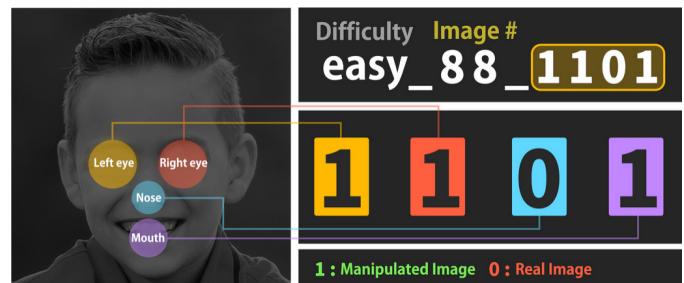
na umjetnu manipulaciju [5]. Za dublju analizu koristi se forenzička analiza [6] koja služi za detekciju kloniranih dijelova slike, odnosno dijelova slike koji su potpuno identični, te analizu artefakata [7] i analizu šuma [8].

## III. MATERIJALI I METODE

### A. Skup podataka

Korišteni podaci preuzeti su s javno dostupnog skupa podataka na Kaggle stranici [9]. Skup se sastoji od ukupno 2041 slike u boji. One su podijeljene na 2 klase: prave i umjetno stvorene slike lica. Slike su stvorene ručno, tj. nisu korišteni GAN-ovi za njihovo generiranje.

Dana je i subjektivna podjela lažnih slika u tri kategorije s obzirom na težinu prepoznavanja: lako, srednje i teško prepoznatljive lažne slike (vidi Sliku 2). Svaka slika ima posebno kodirano ime jedinicama i nulama pri čemu jedinica označava da je određena značajka lica lažna (vidi Sliku 1). Na primjer, slika "easy\_113\_0011" pripada kategoriji lako prepoznatljivih slika, a zamijenjene su joj dvije od četiri značajke lica. Značajke koje su detaljnije promatrane su redom lijevo oko, desno oko, nos i usta. Na danom primjeru se iz njegovog naziva zaključuje da su nos i usta zamijenjeni te slika nije pravo ljudsko lice.



Slika 1: Kodiranje dijelova (značajki) lica koji su izmijenjeni

Podaci su prije razvoja modela biti podijeljeni u tri skupa: za treniranje, validiranje i testiranje u omjeru 80:10:10. Odbirom 80% slučajnih podataka dobiveno je nejednako lažnih i stvarnih lica te je korištena augmentacija kako bi se klase balansirale. Za provođenje augmentacije korišteno je zrcaljenje, rotiranje za 20 stupnjeva i zumiranje. Ovime je povećan broj slika za treniranje s 1632 na 1728.

### B. Korišteni modeli

Kao polazna točka u razvoju dubokih neuronskih mreža za klasifikaciju lažnih i stvarnih lica, korišten je javno dostupni kod s Kaggle stranice [10]. Isprobano je nekoliko baznih modela na koje su dodani slojevi BatchNormalization, Dense,



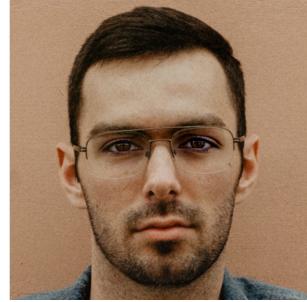
(a) stvara



(b) lažna - easy\_18\_0011



(c) lažna - mid\_127\_0011

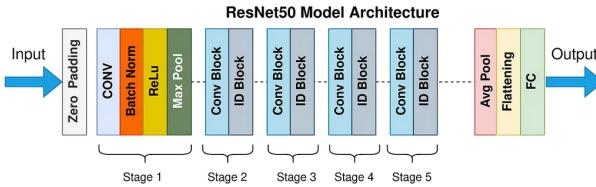


(d) lažna - hard\_124\_1100

Slika 2: Prikaz različitih klasa slika

Dropout i ponovo Dense kako bi se bazni modeli prilagodili za traženi zadatak.

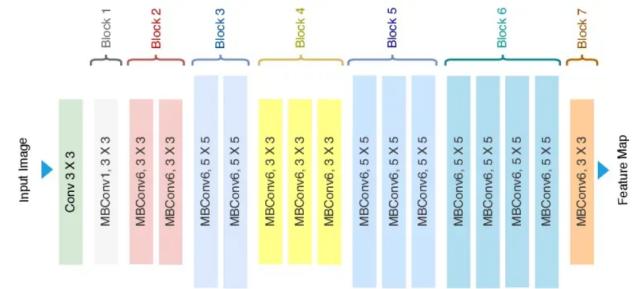
Prvi bazni model bio je ResNet50 (vidi Sliku 3). Ova arhitektura sadrži rezidualne veze koje sprječavaju nestajuće građijente, a pedeset slojeva omogućuje učenje brojnih značajki [11]. U prvom slučaju ResNet50 je treniran na nasumično inicijaliziranim težinama s automatski prilagođljivim koeficijentom učenja čija je početna vrijednost postavljena na 0.001, a u drugom sa težinama dobivenim predtreniranjem na ImageNet skupu podataka i istim koeficijentom učenja.



Slika 3: Arhitektura ResNet50 [11]

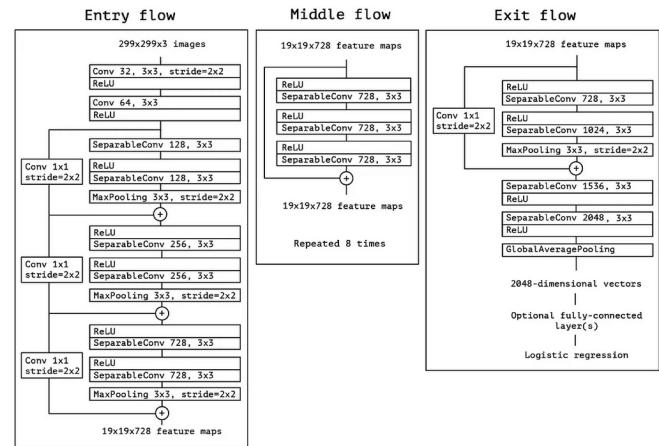
Sljedeća isprobana arhitektura bila je EfficientNetB3 (vidi Sliku 4). Ova arhitektura ističe se po malom broju parametara i skaliranju dimenzija (visine, širine i rezolucije). Cilj joj je naći optimum učinkovitosti i točnosti [12]. Kao i kod ResNet50, isproban je model i s predtreniranim težinama na ImageNet skupu podataka.

Nadalje, Xception arhitektura sadrži rezidualne veze kao i ResNet, ali ističe se po konvolucijama koje se odvajaju po dubini, tj. primjenjuje se filter za svaki ulazni kanal zasebno. U arhitekturi se mogu odvojiti tri sloja: za smanjenje dimenzija,



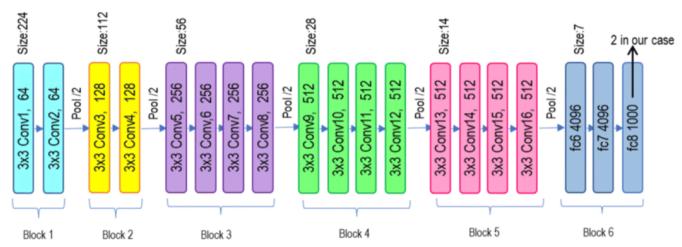
Slika 4: Arhitektura EfficientNet [12]

za izdvajanje značajki te završni za obradu i predikciju [13]. Isproban je model bez i sa predtreniranim težinama.



Slika 5: Arhitektura Xception [13]

Posljednja isprobana arhitektura, ponovno slučajnih težina i predtreniranih, bila je VGG19. Model se sastoji od 19 slojeva, malih konvolucijskih jezgri ( $3 \times 3$ ), ReLU aktivacijske funkcije, max-sažimanja i potpuno povezanih slojeva [14].



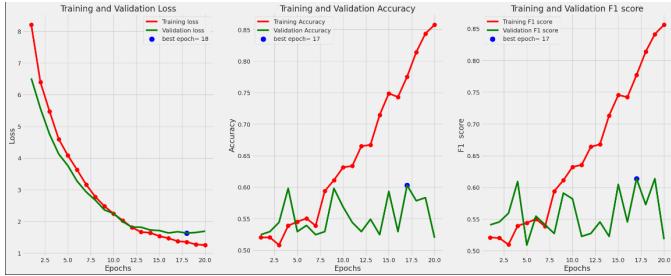
Slika 6: Arhitektura VGG19 [14]

#### IV. REZULTATI

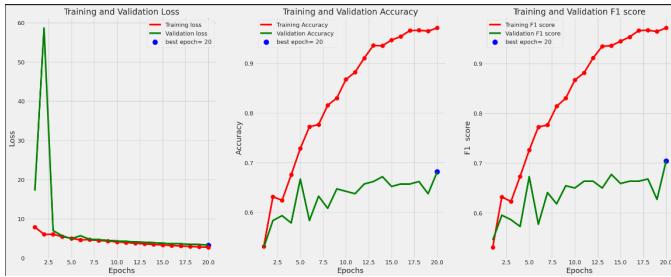
Za sve četiri opisane arhitekture istrenirana su dva modela (s nasumično inicijaliziranim težinama i s predtreniranim težinama na ImageNet skupu podataka). Na slikama 7, 8, 9, 10, 11, 12, 13 i 14 prikazano je kretanje gubitka, točnosti i f1-mjere na podacima za treniranje i validaciju. Vidljivo je da

se modeli općenito bolje ponašaju uz predtrenirane težine, tj. gubitci brže padaju, a točnost i f1-mjera manje osciliraju. Najbolje se ponašaju predtrenirani EfficientNet i Xception. Oboje postižu oko 80% točnosti na skupu za validaciju. Najgore se ponaša VGG19 arhitektura i to i za nepredtrenirane i za predtrenirane težine. Naime, mjere preciznosti jako osciliraju i dostižu tek malo više od 50%.

U tablici I prikazani su rezultati testiranja za sve arhitekture. Može se zaključiti da modeli koji ne koriste predtrenirane težine nisu ništa bolji od slučajnog klasifikatora pošto im se točnost kreće oko 50%. Najveću točnost daje predtrenirani EfficientNetB3 model (80.49%), a slijede ga predtrenirani Xception i ResNet50. VGG19 ne radi dovoljno dobro čak niti s predtreniranim težinama, tj. daje točnost od samo 50.24%. Također, primjećujemo da korištenje predtreniranih težina najviše doprinosi EfficientNetB3-u jer mu točnost najviše poraste u odnosu na ekvivalentan model koji ima slučajno inicijalizirane težine.



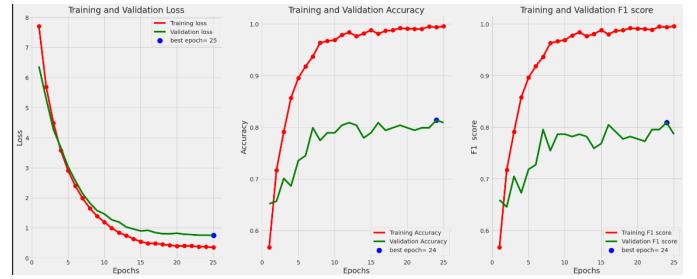
Slika 7: Gubitak, točnost i F1 mjera ResNet50



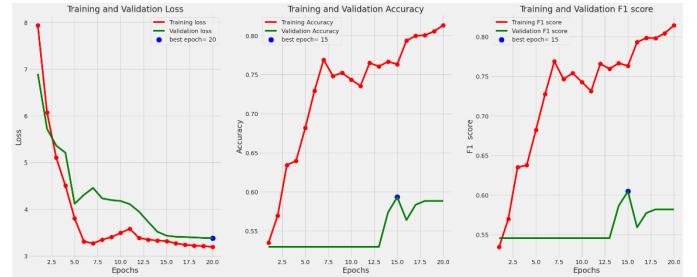
Slika 8: Gubitak, točnost i F1 mjera ResNet50 (predtrenirano)



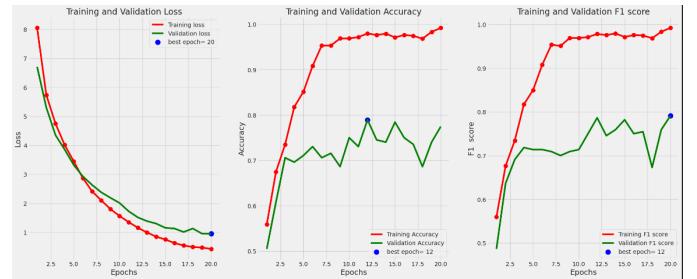
Slika 9: Gubitak, točnost i F1 mjera EfficientNet



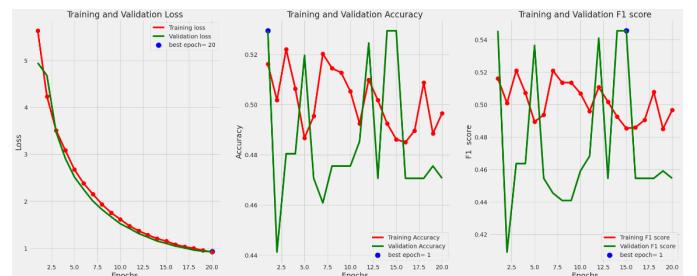
Slika 10: Gubitak, točnost i F1 mjera EfficientNet (predtrenirano)



Slika 11: Gubitak, točnost i F1 mjera Xception



Slika 12: Gubitak, točnost i F1 mjera Xception (predtrenirano)



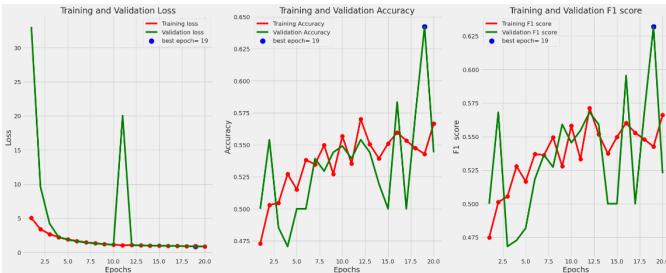
Slika 13: Gubitak, točnost i F1 mjera VGG19

## V. USPOREDBA S POSTOJEĆIM RJEŠENJIMA

Kako bismo procijenili izvedbu našeg modela, potrebno je usporediti naše rezultate s rezultatima objavljenima u prethodnoj literaturi.

U našem radu najveću točnost daje EfficientNet arhitektura (80.49%). U postojećoj literaturi također daje obećavajuće rezultate iznad 80%, s čak 97.90% točnosti u istraživanju koje radi na općenitoj detekciji lažnih videa [15].

Naša druga najbolja arhitektura je XceptionNet s točnošću



Slika 14: Gubitak, točnost i F1 mjera VGG19 (predtrenirano)

Tablica I: Rezultati testiranja različitih modela

Model	točnost na test podacima
ResNet50	54.15
ResNet50 (predt.)	69.27
EfficientNet	54.15
EfficientNet (predt.)	80.49
Xception	55.12
Xception (predt.)	71.22
VGG19	46.83
VGG1 (predt.)	50.24

od 71.22%. XceptionNet se kroz literaturu pokazuje kao općenito dobra arhitektura u borbi protiv lažno generiranih lica, premda se većina istraživanja radi na videima, a ne na slikama, te postiže točnost od 96% [16].

VGG19 se u našem radu pokazao kao najgora arhitektura, s točnošću od 50.24%, dok je u jednoj postojećoj studiji [17] VGG19 rezultirao točnošću od 96% te u drugoj studiji [18] VGG16 u kombinaciji s AlexNet dosegnuo točnost od 94.01%. Takva razlika u rezultatima sugerira da arhitektura VGG nije idealna za analizu lica, ali može biti dobar izbor za općenitu detekciju lažno modificiranih slika.

## VI. VIZUALIZACIJA TEHNIKOM GRADCAM

U svrhu bolje interpretacije i razumijevanja odluka koje donose modeli za detekciju lažno generiranih lica, koristimo alat GradCAM (Gradient-weighted Class Activation Mapping). GradCAM [19] je popularna tehnika vizualizacije koja omogućava da se identificiraju ključne regije u slikama koje najviše doprinose konačnoj odluci modela. Ovo je posebno korisno za modele dubokog učenja, koji su često "crne kutije" i teško ih je interpretirati.

Rezultati vizualizacije za predtrenirani EfficientNetB3 model na lažnim slikama su prikazani na slikama 15, 16 i 17. Iznad originalne slike je napisano koji dio lica je izmjenjen, a iznad izlaza GradCAM predikcija modela.

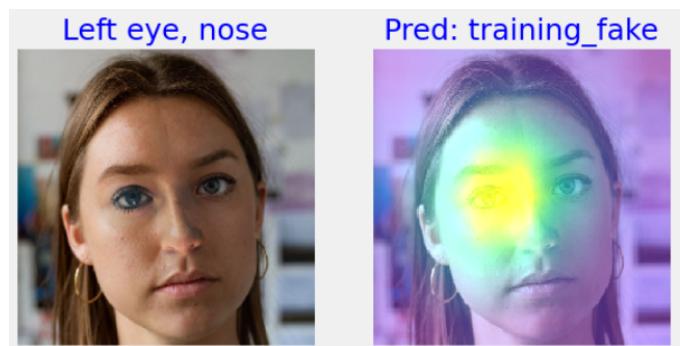
Analiziranjem rezultata možemo vidjeti da model kao regije interesa uzima područja oči, nosa i usta.

## VII. ZAKLJUČAK

U ovom radu istraženo je otkrivanje umjetno generiranih slika lica pomoću modela dubokog učenja. Trenirane su i evaluirane četiri različite arhitekture: ResNet50, EfficientNetB3, Xception i VGG19. Zaključeno je da korištenje predtreniranih težina značajno poboljšava rezultate modela. Najbolji



Slika 15: Model točno detektira lažno lice, te u obzir najviše uzima područje usta i nosa



Slika 16: Model točno detektira lažno lice, a odluci najviše doprinosi lijevo oko



Slika 17: Model krivo označuje sliku kao stvarnom

rezultati ostvareni su s predtreniranim EfficientNet i Xception modelima, koji oba postižu oko 80% točnosti na skupu za validaciju te iznad 70% na skupu za testiranje. Ovakvi rezultati usporedivi su s onima iz postojeće literature te ukazuju na učinkovitost modela dubokog učenja u prepoznavanju umjetno generiranih slika lica.

## LITERATURA

- [1] B. R., Rohini and H. K., Yogish, Focus Challenge Based Presentation Attack Detection in Face Authentication Systems Using Generative Adversarial Network, International Journal of Intelligent Systems and Applications in Engineering, 2023
- [2] Matthew Miller, Deepfakes: Real threat, 2021
- [3] Jonathan Hui, Why it is so hard to train Generative Adversarial Networks!, 2018

- [4] Xu Bozhi, Liu Jiarui, Liang Jifan, Wei Zhuo, Zhang Yue. DeepFake Videos Detection Based on Texture Features. *Computers, Materials & Continua*. 2021
- [5] Venema, Agnes, and Zeno Gerardts. "Digital Forensics, Deepfakes, and the Legal Process." *Scitech Lawyer*, vol. 16, no. 4, 2020
- [6] Samuel Henrique Silva, Mazal Bethany, Alexis Megan Votto, Ian Henry Scarff, Nicole Beebe, Peyman Najafirad, Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models, *Forensic Science International: Synergy*, Volume 4, 2022
- [7] Y. Li, S. Lyu. Exposing DeepFake Videos By Detecting Face Warping Artifacts. *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW*. 2019
- [8] Tianyi Wang, Ming Liu, Wei Cao, Kam Pui Chow, Deepfake noise investigation and detection, *Forensic Science International: Digital Investigation*, Volume 42, Supplement, 2022
- [9] <https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection>
- [10] <https://www.kaggle.com/code/gpiosenka/general-use-image-classifier-f1score-78>
- [11] Suvaditya Mukherjee, The Annotated ResNet-50, *Towards Data Science*, 2022
- [12] Petru Potrimba, What is EfficientNet? The Ultimate Guide, *Roboflow*, 2023
- [13] Sik-Ho Tsang, Review: Xception — With Depthwise Separable Convolution, Better Than Inception-v3 (Image Classification), *Towards Data Science*, 2018
- [14] Khattar Anuradha, Quadri Syed, Generalization of convolutional network to domain adaptation network for classification of disaster images on twitter, *Multimedia Tools and Applications*, 2022
- [15] Deng L, Suo H, Li D. Deepfake Video Detection Based on EfficientNet-V2 Network. *Comput Intell Neurosci*. 2022 Apr 15;2022:3441549.
- [16] A. Mitra, S. P. Mohanty, P. Corcoran, E. Kougianos, A novel machine learning based method for deepfake video detection in social media, in: 2020 IEEE International Symposium on Smart Electronic Systems (iSES)(Formerly iNiS), IEEE, 2020, pp. 91–96.
- [17] Killi, C.B.R., Balakrishnan, N., Rao, C.S. (2023). Deep fake image classification using VGG-19 model. *Ingénierie des Systèmes d'Information*, Vol. 28, No. 2, pp. 509-515. <https://doi.org/10.18280/isi.280228>
- [18] A. Sengur, Z. Akhtar, Y. Akbulut, S. Ekici, U. Budak, Deep Feature Extraction for Face Liveness Detection, in: 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), IEEE, 2018, pp. 1–4.
- [19] Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, in: *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.