Detekcija prezentacijskih napadov s 3D maskami z uporabo globokega učenja

Lovro Sikošek[1], Marko Brodarič[1,2], Peter Peer[1], Vitomir Štruc[2], Borut Batagelj[1]

[1]*Fakulteta za Računalništvo in Informatiko, Univerza v Ljubljani, Večna pot 113, 1000 Ljubljana*
[2]*Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška cesta 25, 1000 Ljubljana*
E-pošta: ls30427@student.uni-lj.si

Detection of Presentation Attacks with 3D Masks Using Deep Learning

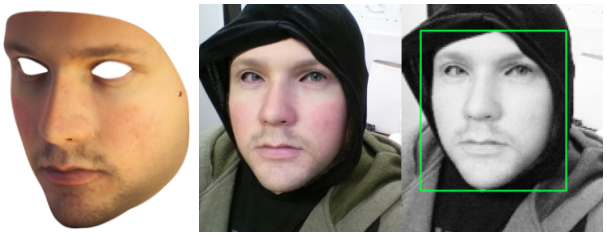
Abstract. *This paper describes a cutting edge approach to Presentation Attack Detection (PAD) of 3D mask attacks using deep learning. We utilize a ResNeXt convolutional neural network, pre-trained on the ImageNet dataset and fine-tuned on the 3D Mask Attack Database (3DMAD). We also evaluate the model on a smaller, more general validation set containing different types of presentation attakcs captured with various types of sensors. Experimental data shows that our model achieves high accuracy in distinguishing between genuine faces and mask attacks within the 3DMAD database. However, evaluation on a more general testing set reveals challenges in generalizing to new types of attacks and datasets, suggesting the need for further research to enhance model robustness*[1].

1 Uvod

V zadnjem desetletju je področje prepoznave obrazov izjemno napredovalo, predvsem zaradi razvoja globokih nevronskih mrež in naraščajoče razpoložljivosti obsežnih podatkovnih baz [6]. Kljub temu pa se še vedno soočamo z izzivom zanesljive detekcije prezentacijskih napadov, kjer se zlonamerni posamezniki poskušajo izdati za druge osebe z uporabo tehnik, kot so fotografije, videoposnetki in tridimenzionalne maske. Prezentacijski napadi s 3D maskami predstavljajo še posebej velik izziv, saj so maske pogosto zelo natančne replike, ki jih je težko razlikovati od pravih obrazov.

V tem članku se osredotočamo na detekcijo prezentacijskih napadov s 3D maskami z uporabo globokega učenja. Predstavili bomo naš pristop, ki temelji na uporabi konvolucijske nevronske mreže za klasifikacijo med pravimi obrazi (bona fide) in napadi z maskami. Uporabili smo model ResNeXt, prednaučen na podatkovni bazi ImageNet, ki smo ga dodatno doučili na podatkovni bazi 3D Mask Attack Database (3DMAD) [16]. Ta podatkovna baza temelji predvsem na visoko kakovostnih 3D maskah, ki jih proizvaja podjetje ThatsMyFace.com, narejenih po podobi človeškega obraza (slika 1). V sklopu tega prispevka smo usposobili model za razpoznavanje

[1]Raziskava je bila sofinancirana iz ARRS raziskovalnega projekta DeepFake DAD (J2-50065) in raziskovalnega programa Računalniški vid (P2-0214).



Slika 1: **Levo**: Primer sofisticirane 3D maske proizvajalca ThatsMyFace.com. **Sredina in desno**: poskus prevare varnostnega sistema z uporabo tovrstne maske.

med resničnimi in neresničnimi primerki te baze in tudi raziskali, kako natančno model deluje na bolj splošnih primerkih prezentacijskih napadov, kot so posnetki zaslona, na katerem je prikazan obraz ali pa obraz, izrezan iz fotografije in uporabljen kot 2D maska [8].

2 Sorodna dela

Prve raziskave na področju odkrivanja prevar z uporabo 3D mask so osnovane na odbojnih lastnostih samega materiala le-teh v primerjavi s človeško kožo [13]. Ta pristop se je izkazal za nepraktičnega, saj je odvisen od tega, da je območje čela jasno vidno in, da je subjekt oddaljen natanko 30 cm stran. V tej raziskavi se avtorji tudi niso spustili v eksperimentiranje na maskah, oblikovanih po človeških obrazih, zgolj testiranje na materialih mask [16]. Naslednja večja raziskava na tem področju je bila [5], ki pa je že koristila podatkovno bazo 3D natisnjenih mask. Odkrivanje je bilo izvedeno na podlagi globinske slike in teksturne analize z uporabo metode Local Binary Pattern (LBP) [14]. Uporabljena podatkovna baza pa ni javno dostopna, kar je otežilo primerjalne raziskave.

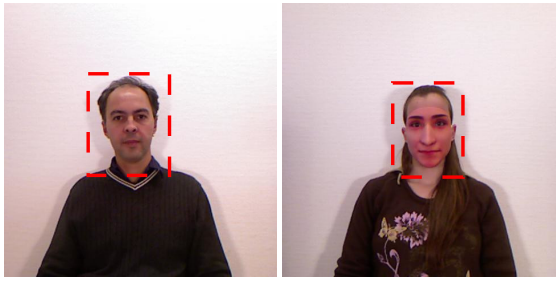
Prva javno dostopna podatkovna baza za odkrivanje 3D mask je bila 3D Mask Attack Database (3DMAD) [16], uporabljena tudi v raziskavi [2]. Avtorji raziskave na 3DMAD bazi so tudi evalvirali, kako učinkovite so njihove 3D maske pri zavajanju sistema za odkrivanje dvodimenzionalnega obraza (torej brez podatkov o globini). V ta namen so uporabili metodo Inter Session Variability modeling (ISV) [10], ki naj bi služila kot izhodišče za primerjavo drugih algoritmov za prepoznavo obraza. To metodo so primerjali z že prej omenjeno metodo LBP [14], ki je bila uporabljena za teksturno analizo in se je

izkazala za bolj zanesljivo pri detekciji napadov s 3D maskami.

K nadaljnemu razvoju pa so pripomogle tudi organizacije, kot je LivDet Face, ki gosti razna tekmovanja v razvoju algoritmov za zaznavanje ponaredkov, med drugimi tudi LivDet Face [8, 12]. Na prvem tekmovanju leta 2021 so zmagovalci tekmovanja razvili tri strategije za zaznavanje ponaredkov, temelj vseh pa je bil model osnovan na konvolucijski nevronski mreži [8]. Prva strategija je temeljila na modelu DeepPixBis [7] in nadzorovanem učenju, druga na modelu ResNeXt, naučenem na javni podatkovni bazi ImageNet in nato dodatno doučen na več privatnih bazah, tretja pa na manjšem modelu, ki za vhod vzame diskretno Fourierjevo transformacijo slike, z namenom, da bi zaznal razlike med frekvenčno domeno pravih slik in prezentacijskih napadov. Za dano sliko so predikcije vseh treh modelov nato združene s pomočjo uteži Fisher Discriminant Ratio (FDR) [4]. Ta ansambel modelov se je izkazal za dovolj natančnega glede na ACER metriko (angl. Average Classification Error Rate) za zmago na tekmovanju v kategoriji slik z ACER vrednostjo 16,47% [8].

3 Metodologija

Sledili smo najnovejšim metodam in razvili algoritem za odkrivanje prezentacijskih napadov, ki uporablja binarni klasifikacijski model za razlikovanje med pravimi obrazi (bona fide) in poskusi vdora (prezentacijski napadi). Model je bil prednaučen na podatkovni bazi ImageNet ter dodatno izpopolnjen na eni izmed največjih in najbolj znanih podatkovnih baz za odkrivanje napadov z maskami, 3DMAD [16].



Slika 2: Primer verodostojne slike in prezentacijskega napada iz podatkovne baze 3DMAD. Na vsaki sliki je s črtkano črto označeno območje obraza, kot ga zazna model MTCNN.

Izbran je bil model ResNeXt zaradi svoje arhitekture, ki omogoča visoko zmogljivost pri obdelavi slik. Njegov pristop grupiranih konvolucij (anlg. grouped convolution) omogoča večjo učinkovitost pri učenju kompleksnih značilnosti [3]. Na tekmovanjih, kot je LivDet Face, se je ResNeXt izkazal za učinkovitega pri nalogah detekcije napadov z maskami, kar je bil eden glavnih razlogov za našo izbiro tega modela. Uporabljena je bila javno dostopna [15] implementacija SE-ResNeXt arhitekture. Vhodni sloj modela vključuje tri zaporedne konvolucijske sloje s 3x3 filtri, ki zmanjšajo dimenzije slike in izluščijo pomembne značilnosti, po vsaki konvoluciji pa se uporabi aktivacijska funkcija ReLU za dodajanje nelinearnosti. Sledijo bloki z grupiranimi bottleneck 3x3 konvolucijami in SE (angl. Squeeze-and-Excitation) mehanizmom, po vsaki konvoluciji pa zopet ReLU funkcija. Uporabljene so tudi Anti-Aliasing metode (Rectangle-2 Anti-Aliasing), katerih cilj je zmanjšanje aliasing artefaktov, ki nastanejo pri procesiranju slik zaradi sprememb ločljivosti ali geometrije vhodnih podatkov. Aliasing artefakti so lahko na primer neželeni robovi ali popačenja, ki jih model napačno prepozna kot pomembne značilnosti. Uporabljene metode vključujejo uporabo nizkopasovnih filtrov, ki zmanjšajo učinek teh popačenj ter izboljšajo zmožnost modela, da se osredotoči na prave značilnosti slike. Sledi še združevalni sloj na podlagi povprečja (angl. Global Average Pooling layer) in polno povezani sloj, ki izvede končno klasifikacijo. Ta različica modela ResNeXt [17] je bila izbrana, ker izmed vseh modelov v svoji družini dosega najboljšo natančnost na bazi ImageNet.

Podatkovno bazo smo pripravili tako, da smo izbrali samo barvne slike, brez globinskih posnetkov, saj naš model ni prilagojen za učenje v tej modalnosti. Vsako sliko je potrebno še dodatno pripraviti, preden jo model prejme v procesiranje. ResNeXt namreč zahteva velikost slike 384×384 slikovnih točk, torej je potrebno sliko vsaj obrezati do kvadrata in prilagoditi njeno velikost. Za večjo natančnost pa smo v predprocesiranje vključili avtomatsko zaznavo obraza subjekta z uporabo modela MTCNN [11]. Ta vzame sliko in vrne mejne točke pravokotnika, ki omejuje obrazno regijo na sliki. Obraz lahko nato izrežemo iz slike, kot je nakazano na sliki 2, in tako zagotovimo standardno obliko podatkov za učenje. (sicer bi lahko bil obraz v različnih slikah na različnih položajih, kar bi vplivalo na učinkovitost učenja modela). Uporabili smo zgolj vsako sedmo sliko iz podatkovne baze 3DMAD, saj so sami posnetki dokaj statični, torej so zaporedne slike v posnetkih med seboj podobne.

4 Rezultati

4.1 Podatkovne baze

Za podatkovno bazo smo izbrali 3DMAD [16]. Dejavniki za izbiro ravno te baze so bili njena obsežnost (to je namreč ena obsežnejših baz za tovrstne poskuse) in uporaba v že obstoječi literaturi, tudi v tekmovanju LivDet Face 2021 [8] . Podatkovna baza 3DMAD zajema posnetke 17 različnih subjektov, pridobljene s senzorjem Microsoft Kinect, ki zajame tako podatke o sliki kot o globini. Za vsakega izmed 17 subjektov so izvedli 3 snemalne seanse: prvi dve sta bili izvedeni 2 tedna narazen, na njih pa so zajeli subjektov pravi obraz. Na tretji je subjekt nosil razne maske v podobi drugih subjektov. Primerka resnične in ponarejene slike si lahko ogledate na sliki 2. Na vsaki seansi in za vsakega subjekta je v bazi 5 posnetkov dolžine 10 sekund. Skupno torej baza obsega 255 posnetkov, sestavljenih iz 300 slik, kar pomeni, da imamo na voljo 76.500 slik. Te so bile brez prekrivanja razdeljene v množico za učenje in testno množico, prva zajema 12 subjektov (54.000 slik) druga pa preostalih 5 subjektov (22.500 slik) kot narekuje protokol baze 3DMAD. Maske so bile pridobljene preko podjetja ThatsMyFace.com, ki

proizvaja 3D maske, narejene na podlagi slik človeškega obraza, zajetih iz različnih kotov. Obenem pa smo za testiranje uporabili tudi manjšo, validacijsko množico, pridobljeno od organizatorjev LivDet Face 2024. Le-ta vsebuje poleg prezentacijskih napadov z maskami tudi druge vrste napadov: 6 primerkov napadov s 3D natisnjeno masko, 4 primerke obrazov, natisnjenih na fotografski papir, 4 primerke natisnjenih obrazov z izrezanimi luknjami za oči (nato uporabljenih kot maska), 5 primerkov natisnjenih obrazov z manj kakovostnim tiskalnikom, 5 primerkov obrazov na zaslonih raznih naprav, 3 primerke 3D mask iz belega filamenta in 4 primerke mask iz umetne mase Resin. Primera slik obrazov, natisnjenih na papir sta razvidna na sliki 3, primera posnetkov obrazov, prikazanih na zaslonih pa na sliki 4. Model smo prijavili tudi na tekmovanje LivDet Face 2024, kjer so ga evalvirali na zasebni zbirki, ki vsebuje iste vrste napadov, kot validacijska množica.

4.2 Metrike

Model za vsako sliko napove, ali gre za poskus napada (idelana vrednost 0.0) ali pa gre za resnični obraz (idealna vrednost 1.0). Končno napoved smo pridobili z uporabo vrednosti praga (angl. threshold) 0.5 za verjetnost, da je oseba na sliki resnična. Če model torej vrne vrednost napovedi večjo od 50% bo to pomenilo, da je obraz resničen, sicer pa, da gre za napad. Za ugotavljanje natančnosti smo uporabili metriki APCER (angl. attack presentation classification error rate), BPCER (angl. bona fide presentation classification error rate), saj se ti že pojavljata v obstoječi literaturi [8, 12], izračunani pa sta kot:

$$APCER = \frac{FP}{TN + FP}\ ,\ BPCER = \frac{FN}{FN + TP} \tag{1}$$

kjer je FP (angl. False Positive) število nezaznanih napadov, TN (angl. True Negative) število zaznanih napadov, FN (angl. False Negative) število napačno klasificiranih resničnih obrazov in TP (angl. True Positive) število pravilno zaznanih resničnih obrazov. APCER meri delež napačno klasificiranih napadov, kar pomeni, da se napad zmotno obravnava kot pravi obraz. BPCER pa meri delež napačno klasificiranih pravih obrazov, kar pomeni, da model zmotno prepozna pravi obraz kot napad. Te metrike so ključne za ocenjevanje uspešnosti modela, saj pomagajo razumeti, kako zanesljivo model razlikuje med pravimi in lažnimi primerki. V fazi evalvacije smo torej izračunali APCER in BPCER našega modela za testno množico baze 3DMAD, kot tudi za validacijsko množico. Za vpogled v natančnost modela ob različnih pragih (angl. threshold) smo izrisali tudi ROC (angl. Receiver Operating Characteristic Curve) krivulji za validacijsko in testno zbirko ter izračunali ploščino pod ROC krivuljo (angl. Area under ROC curve, AUC).

4.3 Učenje

Za eksperimente smo uporabili model ResNeXt s 150 milijoni parametrov, javno dostopen preko knjižnice timm [15]. Čeprav to ni največji model na voljo v sodobnih arhitekturah globokega učenja, predstavlja precej velik model v primerjavi z mnogimi drugimi arhitekturami, ki jih pogosto uporabljamo za podobne naloge (kot so npr. ResNet z manjšim številom parametrov). Ta obseg omogoča boljše zajemanje kompleksnih vzorcev v slikah, kar je še posebej pomembno pri nalogi, kot je detekcija napadov s 3D maskami. Le-ta je na voljo v že naučeni različici na bazi ImageNet. Za predprocesiranje slik je bil uporabljen model za odkrivanje obrazov MTCNN različice 0.1.1. Učenje je nato potekalo na bazi 3DMAD, uporabljen je bil optimizer ADAM za optimizacijsko funkcijo skupaj z Binary Cross Entropy (BCE) za funkcijo za računanje izgube. Hitrost učenja (angl. learning rate) je bila nastavljena na 10^{-5}, in je bila prepolovljena vsakih 5 epohov, da bi preprečili preveliko prilagajanje (angl. overfitting) uporabljeni podatkovni bazi, saj ta vsebuje veliko število slik. S takimi hiperparametri je bil nato model treniran 25 epohov na učni množici, na koncu vsakega epoha pa je bila tudi preverjena natančnost na validacijski množici 37 slik, pridobljenih od organizatorjev tekmovanja LivDet Face 2024 [12]. Model smo testirali tudi na validacijski množici tudi v fazi evalvacije, da bi preverili, kako učinkovito se model posploši tudi na druge tipe prezentacijskih napadov. Za namene tako učenja kot tudi testiranja modela je bila uporabljena Arnesova gruča superračunalniškega omrežja SLING[2]. Specifikacije uporabljenega delovnega vozlišča so: 2 grafični kartici Nvidia V100S, procesor AMD EPYC 7272 in 128GB pomnilnika RAM.

Slika 3: Primera slik natisnjenih obrazov iz validacijske množice.

Slika 4: Primera slik iz posnetkov zaslonov iz validacijske množice.

Med učenjem modela je bila opazovana tudi skupna izguba za vsak epoh, ki jo vrne funkcija BCE. Model je bil shranjen po tistem epohu, v katerem je dosegel

[2] https://www.arnes.si/storitve/superracunalnisko-omrezje/

najnižjo skupno izgubo. Izkazalo se je, da je to po epohu 22.

4.4 Evalvacija

Eksperimentalni rezultati so podani v tabeli 1 in prikazuje, kako uspešen je naučen model ResNeXt v razpoznavanju primerkov resničnih in neresničnih obrazov iz testnega dela podatkovne baze 3DMAD v primerjavi s primerki validacijske zbirke in rezultati s tekmovanja LivDet Face 2024.

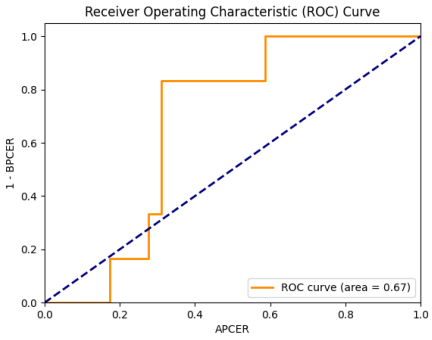
Testna množica	APCER	BPCER
3DMAD	0.00	0.00
LivDet Face validacija	0.34	0.17
LivDet Face 2024 test	0.05	0.98

Tabela 1: Tabela rezultatov testiranja na testni množici baze 3DMAD in validacijski ter testni množici tekmovanja LivDet Face 2024.

Iz rezultatov v tabeli 1 je razvidno dejstvo, da model zelo natančno zaznava poskuse napadov iz baze, ki so podobni tistim, na katerih je bil naučen. Celo vse primerke 3DMAD testne zbirke klasificira pravilno, kar se sklada z rezultati podobnih raziskav na tej zbirki [1]. Na validacijski množici LivDet Face 2024 pa so rezultati pokazali manjšo natančnost ($APCER = 0.34$ in $BPCER = 0.17$), kar pomeni, da model pri zaznavanju novih, nepoznanih vrst napadov ni tako učinkovit kot pri nalogah, na katerih je bil treniran. Ta razlika nakazuje, da model sicer uspešno prepoznava napade, ki so mu znani, vendar težje posploši na nove tipe napadov. Po drugi strani pa rezultati na validacijski zbirki kažejo na to, da model tudi primerke napadov, ki jih še ni srečal, razlikuje od resničnih primerkov bolj natančno kot strategija naključnega ugibanja.

Za bolj podroben vpogled v natančnost modela smo izrisali tudi ROC krivulji na podlagi napovedi, ki jih ta vrne za obe testni množici. Ker model pravilno klasificira vse primerke v testni množici 3DMAD, je ROC krivulja vodoravna črta in je nismo vključili v članek. ROC krivulja za LivDet Face validacijsko zbirko na sliki 5 potrjuje, da je model tudi na bolj splošni validacijski zbirki natančnejši kot naključno ugibanje. Na podlagi ROC krivulje smo tudi izbrali primeren prag za določitev predikcije, ki se je izkazal za verjetnost 0.999 da gre za resnično sliko. To pa smo izvedli šele po tekmovanju LivDet Face 2024, zato tam model dosega nizko natančnost na resničnih primerkih.

Na tekmovanju je model zaradi tega tudi dosegel nižjo uvrstitev v kategoriji za slike, z $APCER = 0.05$ in $BPCER = 0.98$, v primerjavi z zmagovalnim algoritmom, ki je dosegel $APCER = 0.09$ in $BPCER = 0.01$. Vse visoko uvrščene ekipe pa so za treniranje uporabile več raznolikih podatkovnih baz, kar nakazuje na možnost razširitve in izboljšave našega pristopa v prihodnosti. [9]



Slika 5: ROC krivulja za validacijsko množico LivDet Face 2024 tekmovanja.

5 Zaključek

Eksperimentalni rezultati so pokazali, da je naš model zelo učinkovit pri prepoznavanju napadov, še posebej tistih s pomočjo 3D maske. Model je dosegel visoko natančnost pri odkrivanju napadov v testni množici 3DMAD, kar potrjuje njegovo sposobnost za uporabo v praktičnih aplikacijah. Evalvacija na validacijski množici LivDet Face 2024 pa je razkrila izzive pri posploševanju na nove vrste napadov, kar kaže na potrebo po nadaljnjem raziskovanju.

V prihodnosti bi lahko eksperiment razširili z različnimi kombinacijami hiperparametrov za učenje in z uporabo večih modelov, osnovanih na različnih arhitekturah, za primerjalno analizo. Naš pristop bi lahko še izboljšali z vključitvijo večjih in bolj raznolikih podatkovnih baz ter z razvojem metod, ki bi omogočale boljše posploševanje. Nadaljnje raziskave bodo ključne za razvoj robustnih sistemov za prepoznavo obrazov, ki bodo kos vedno bolj prefinjenim biometričnim napadom.

S tem prispevkom smo pokazali, da so modeli, kot je ResNeXt, obetavna rešitev za detekcijo prezentacijskih napadov s 3D maskami in postavili temelje za nadaljnje raziskave.

Literatura

[1] S. Arora, M. Bhatia, V. Mittal: A robust framework for spoofing detection in faces using deep learning, The Visual Computer 38 (2022): 1-12

[2] S. Jia, G. Guo, Z. Xu: A survey on 3D mask presentation attack detection and countermeasures, Pattern Recognition 98 (2020): 107032

[3] S. Xie, R. Girshick, P. Dollar, Z. Tu, K. He: Aggregated Residual Transformations for Deep Neural Networks, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017

[4] N. Damer, A. Opel, A. Nouak: Biometric source weighting in multi-biometric fusion: Towards a generalized and robust solution, European Signal Processing Conference (EUSIPCO), 2014

[5] N. Kose, J.-L. Dugelay: Countermeasure for the protection of face recognition systems against mask attacks, IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013

[6] M. Wang, D. Weihong: Deep face recognition: A survey, Neurocomputing 429 (2021): 215-244

[7] A. Geroge, S. Marcel: Deep Pixel-wise Binary Supervision for Face Presentation Attack Detection, International Conference on Biometrics (ICB), 2019

[8] S. Purnapatra, N. Smalt, K. Bahmani, P. Das, D. Yambay, A. Mohammadi, A. George, T. Bourlai, S. Marcel, S. Schuckers, M. Fang, N. Damer, F. Boutros, A. Kuijper, A. Kantarci, B. Demir, Z. Yildiz, Z. Ghafoory, H. Dertli, H. K. Ekenel, S. Vu, V. Christophides, D. Liang, G. Zhang, Z. Hao, J. Liu, Y. Jin, S. Liu, S. Huang, S. Kuei, J. M. Singh, R. Ramachandra: Face liveness detection competition (LivDet Face) - 2021, 2021 IEEE International Joint Conference on Biometrics (IJCB), 2021

[9] L. Igene et al.: Face Liveness Detection Competition (LivDet-Face) - 2024

[10] R. Wallace, M. McLaren, C. McCool, S. Marcel: Inter-session variability modelling and joint factor analysis for face authentication, International Joint Conference on Biometrics (IJCB), 2011

[11] J. Xiang, G. Zhu: Joint Face Detection and Facial Expression Recognition with MTCNN, International Conference on Information Science and Control Engineering (ICISCE), 2017

[12] LivDet Face 2024: Liveness Detection Competition, 2nd Edition, https://face2024.livdet.org/

[13] Y. Kim, J. Na, S. Yoon, J. Yi: Masked fake face detection using radiance measurements, JOSA A 26.4 (2009): 760-766

[14] T. Ojala, M. Pietikainen, T. Maenpaa: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence 24.7 (2002): 971-987

[15] Hugging Face, *SEResNeXtAA201D 32x8d SW_IN12k_FT_IN1k_384*, 2024. [Online]. Available: `https://huggingface.co/timm/seresnextaa201d_32x8d.sw_in12k_ft_in1k_384`.

[16] N. Erdogmus, S. Marcel: Spoofing in 2D Face Recognition with 3D Masks and Anti-spoofing with Kinect, IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), 2013

[17] J. Hu, S. Li, G. Sun: Squeeze-and-Excitation Networks, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018