Učinkovitost detektorjev globokih ponarejenih slik ob napadih zasnovanih z difuzijskimi modeli

Marija Ivanovska[1], **Vitomir Štruc**[1]

[1]*Fakulteta za elektrotehniko, Univerza v Ljubljani*
E-mail: {marija.ivanovska, vitomir.struc}@fe.uni-lj.si.com

Abstract

The detection of malicious Deepfakes is a constantly evolving problem, that requires continuous monitoring of detectors, to ensure they are able to detect image manipulations generated by the latest emerging models. In this paper, we present a preliminary study that investigates the vulnerability of single-image Deepfake detectors to attacks created by a representative of the newest generation of generative methods, i.e. Denoising Diffusion Models (DDMs). Our experiments are run on FaceForensics++, a commonly used benchmark dataset, consisting of Deepfakes generated with various techniques for face swapping and face reenactment. The analysis shows, that reconstructing existing Deepfakes with only one denoising diffusion step significantly decreases the accuracy of all tested detectors, without introducing visually perceptible image changes.

1 Uvod

S hitrim razvojem digitalnih tehnologij je ustvarjanje lažnih slik in videoposnetkov postalo skoraj samodejen proces. Čeprav so te metode zelo uporabne v zabavni industriji, jih je mogoče uporabiti tudi zlonamerno. Primer so ponarejene slike, pri katerih je obraz osebe spremenjen ali uporabljen kot nadomestek za obraz druge osebe z namenom ustvarjenja želenih prizorov. [13]. Prirejene podatke je nato mogoče zlorabiti za širjenje napačnih informacij, škodovanje žrtvi ali manipuliranje z javnim mnenjem. Razvoj natančnih algoritmov za odkrivanje lažnih informacij je zato ključnega pomena za preprečevanje takšnih kršitev.

V preteklih letih so bili predstvaljeni različni algoritmi strojnega učenja za samodejno odkrivanje manipuliranih podatkov. [12, 10, 7]. Ti algoritmi običajno iščejo neskladnosti v osvetlitvi in sencah, vizualne artefakte ali druge sledove, ki jih generativni modeli pustijo med ustvarjanjem globokih ponaredkov. Kljub temu so detektorji lahko podvrženi napadom, katerih cilj je zavajanje detektorja, da ta napačno razvrsti ponarejene slike kot prave. Ti napadi so običajno ustvarjeni z dodajanjem majhnih motenj obstoječim globoko ponarejenim slikam [6]. Naprednejše napadalne metode se lahko celo naučijo vključiti napad neposredno v postopek ustvarjanja globoke ponarejene slike, da bi jo naredile bolj neprepoznavno [2].

Raziskava je bila sofinancirana iz ARRS projekta J2-2501 (A) in temeljnega programa P2-0250 (B).

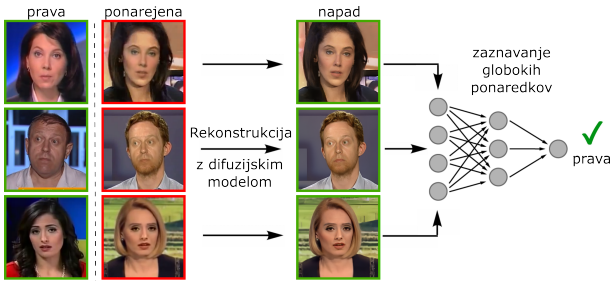

Figure 1: Raziskujemo ranljivost uveljavljenih, vnaprej naučenih enoslikovnih metod za odkrivanje globokih ponaredkov na napade, ki nastanejo z rekonstrukcijo ponarejenih slik z modelom DifFace. [16]. DifFace je sodoben difuzijski model (Denoising Diffusion Model, DDM) za odpravljanje šuma v postopku obnove slike obraza. Naši poskusi so pokazali, da lahko samo en rekonstrukcijski korak z difuzijskim modelom zavede detektorje, da razvrstijo manipulirano sliko kot pravo, ne da bi pri tem škodovali njenemu izgledu.

Nedavni pojav nove generacije modelov, tako imenovanih difuzijskih modelov za zmanjševanje šuma (Denoising Diffusion Models- DDM), je povzročil veliko zaskrbljenost zaradi širjenja lažnih informacij, saj se je izkazalo, da lahko ustvarijo še bolj realistične in prepričljive ponaredke kot njihovi predhodniki - generativna nasprotujoča omrežja (angl. Generative Adversarial Networks- GANs) [4].Zaradi te grožnje raziskujemo zmožnost DDM, da napadejo sisteme za odkrivanje globokih ponarejenih slik, tako da preprosto rekonstruirajo obstoječe lažne slike z vnaprej določenimi koraki dodajanja in manjšanja šuma (Slika 1). V študiji smo se omejili na generativne algoritme za zamenjavo in preoblikovanje obrazov. Kolikor nam je znano, smo prvi, ki smo raziskali možnosti uporabe DDM v tem kontekstu.

V tem članku predstavljamo naslednje prispevke: *i)* preizkušamo zmožnost nemodificiranega modela, ki temelji na difuziji, da ustvari napade na sisteme za odkrivanje globokih ponaredkov; *ii)* vrednotimo modificirane globoko ponarejene slike glede na njihovo vizualno kakovost; *iii)* preizkušamo ranljivost različnih vrst enoslikovnih detektorjev globokih ponarejenih slik na modificirane globoke ponarejene slike.

2 Sorodna dela

Zaznavanje globoko ponarejenih slik. Novejši enoslikovni algoritmi za odkrivanje lažnih slik temeljijo predvsem na različnih metodah globokega učenja.

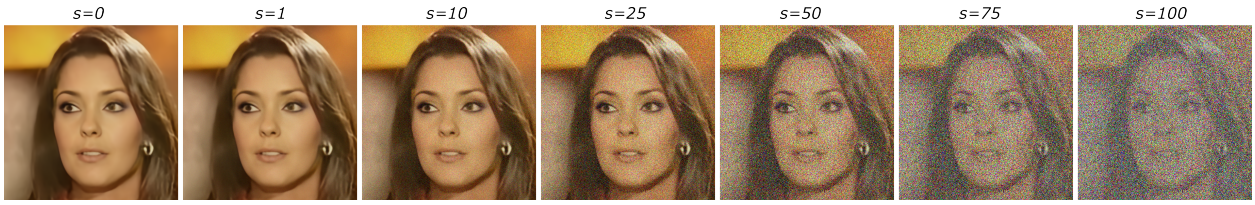

Figure 2: Vizualizacija različnih ravni šuma, dodanih na globoko ponarejenim slikam, pred njihovo rekonstrukcijo z difuzijskim modelom DifFace [16]. Spremenljivka s se nanaša na število časovnih korakov, ki so bili uporabljeni za aplikacijo šuma pred razšumljanjem z izbranim modelom. Začetna, nespremenjena globoka ponarejena slika, je označena z $s = 0$.

Naivni pristopi običajno predstavljajo CNN, ki se nauči zaznavati globoke ponarejene slike z razvrščanjem primerov pravih in lažnih podatkov. Xception [3] in MesoNet [1] sta med najbolj popularnimi v tej kategoriji. Da bi zagotovila, da CNN zajame diskriminatorne lastnosti, sta Nguyen *et al.* in Wang *et al.* [12, 15] uporabila eksplicitno modeliranje tipičnih artefaktov globokih ponaredkov v slikovnem prostoru. Luo *et al.* [10] je na drugi strani predlagal analizo slik v frekvenčnem prostoru, saj imajo pravi in ponarejeni podatki običajno različne frekvenčne spektre. Podoben pristop je predstavil Liu *et al.* [9]. Čeprav so ti algoritmi zelo natančni, ko se uporabljajo za probleme znotraj zaprte množice, pa se slabše obnesejo pri posploševanju vzorcev izven tega območja in pri vzorcih ustvarjenih z neznanimi metodami ustvarjanja globoko ponarejenih slik. Da bi rešil ta problem, sta Li *et al.* [7] in Shiohara *et al.* [14] predlagala samonadzorovane algoritme, ki se ne zanašajo na posebne nabore podatkov globokih ponaredkov, temveč se učijo iz simuliranih manipulacij slik, da bi nadomestili odsotnost podmnožice lažnih podatkov.

Napadi na detektorje globokih ponaredkov. Številna dela so pokazala, da so lahko metode za odkrivanje globokih ponaredkov dovzetne za nekatere vrste skrbno pripravljenih napadov. Ti napadi so na splošno razdeljeni v dve glavni kategoriji, tj. napade bele škatle in napade črne škatle. Prvi so zasnovani s popolnim poznavanjem arhitekture in parametrov ciljnega detektorja globokih ponarejenih slik, drugi pa s poskusnimi napadi probajo oceniti delovanje napadanega modela. Hussain *et al.* [6] in Gandhi *et al.* [5] sta izvedla obe vrsti napadov z optimizacijo motenj, dodanih globoko ponarejenim slikam, tako da so bile te kasneje razvrščene kot pristne. Neekhara *et al.* [11] je uporabil podoben pristop v okvirih "črne škatle" in dodatno preučil prenosljivost ustvarjenih napadov med različnimi detektorji lažnih slik. Poleg motenj v slikovnem prostoru je Carlini *et al.* [2] raziskoval tudi možnost izvedbe napadov v latentnem prostoru generativnega modela globoko ponarejenih slik, tako da je iz njega pridobil nasprotne slike. Novo vrsto napadov črne škatle je predstavil Liu *et al.* in [8], ko je izvedel naknadno obdelavo predhodno ustvarjenih globoko ponarejenih slik z odstranjevanjem zaznavnih sledi, ki jih pusti proces generiranja. Tako dobljeni globoki ponaredki so pristnejši in jih je težje odkriti. V ta namen je bila razvita in optimizirana metoda za prepoznavanje in odstranjevanje vnaprej določenih vrst sledi TR-Net, ki temelji na GAN.

3 Napadi zasnovani z difuzijskimi modeli

Naši napadi globoko ponarejenih slik so ustvarjeni z orodjem DifFace [16], to je sodoben difuzijski model za slepo obnavljanje obrazov. Upoštevajte, da model, ki ga uporabljamo, ni bil zasnovan ali optimiziran za napade na metode odkrivanja globokih ponaredkov. Naš cilj je preizkusiti zmogljivosti izbranega modela kot že pripravljenega generatorja za napade "črne škatle". Generiranje napadov je sestavljeno iz dveh faz. V prvi fazi, se izbrani ponarejeni sliki y_0 postopoma dodaja Gaussov šum $\mathcal{N}(0, \sigma^2\mathbf{I})$ za s korakov. V drugi fazi, sliko z dodanim šumom (x_s), postopoma razšumljamo s parametriziranim generativnim modelom $D_\theta(x, \sigma)$. V naši študiji je D_θ vnaprej usposobljen aproksimator, ki je bil optimiziran za zmanjšanje Kullback-Leiblerjeve (KL) divergence med modelirano $p(x_s|y_0)$ in ciljno distribucijo $q(x_s|x_0)$, pri čemer je x_0 rekonstruirana različica začetne vhodne slike y_0. Splošen pregled generiranja napadov ponarejenih slik je predstavljen na sliki 3. Za več podrobnosti o orodju DifFace predlagamo branje izvornega članka [16].

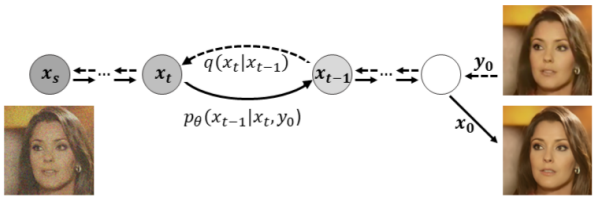

Figure 3: Splošni prikaz ustvarjanja napadov globoko ponarejenih slik. Globok ponaredek (y_0) je vnešen v model, ki temelji na difuziji [16]. Model ustvari ošumljen vzorec x_s, s postopnim dodajanjem Gaussovega šuma za s korakov. Napad (x_0) je nato ustvarjen z razšumljanjem.

4 Eksperimenti

Zbirke podatkov. V naši študiji eksperimentiramo s podatkovno zbirko FaceForensics++ (FF+) [13], ki se pogosto uporablja za ocenjevanje metod za odkrivanje globokih ponaredkov. Zbirka podatkov je sestavljena iz 1000 resničnih videoposnetkov iz YouTube-a, ki so na voljo v treh različnih kakovostih. Mi uporabljamo izključno podatke najvišje kakovosti (z oznako raw). Globoki ponaredki FF+ so ustvarjeni s tremi različnimi tehnikami zamenjave obrazov, to so DeepFakes, FaceSwap in FaceShifter, ter dvema metodama uprizarjanja obrazov, tj. Face2Face in NeuralTextures. Naša študija je omejena na enoslikovne detektorje za globoko ponarejene slike, zato smo iz vsakega posnetka izvozili le vsako deseto sliko. Zaznavanje in obrezovanje področij obraza je bilo izvedeno z algoritmom MTCNN. Vse pridobljene globoke ponaredke smo nato šestkrat rekonstruirali z orodjem DifFace [16], vsakič z drugačno vrednostjo skupnega števila časovnih korakov difuzije: 1, 10, 25, 50, 75 in 100. Nivoji šuma, ki predstavljajo

Fake Data	SSIM ↑						LPIPS ↓						CSIM ↑					
	s=1	s=10	s=25	s=50	s=75	s=100	s=1	s=10	s=25	s=50	s=75	s=100	s=1	s=10	s=25	s=50	s=75	s=100
DF	0.9504	0.9475	0.9330	0.9045	0.8755	0.8457	0.0626	0.0625	0.0714	0.0951	0.1145	0.1286	0.9593	0.9366	0.8577	0.7015	0.5500	0.4008
F2F	0.9528	0.9496	0.9349	0.9066	0.8782	0.8496	0.0551	0.0552	0.0639	0.0850	0.1023	0.1145	0.9699	0.9481	0.8733	0.7210	0.5665	0.4105
FSh	0.9488	0.9451	0.9309	0.9041	0.8772	0.8496	0.0659	0.0660	0.0742	0.0960	0.1138	0.1265	0.9712	0.9482	0.8732	0.7153	0.5490	0.3895
FS	0.9544	0.9507	0.9352	0.9048	0.8744	0.8435	0.0513	0.0519	0.0597	0.0811	0.0994	0.1129	0.9687	0.9479	0.8777	0.7295	0.5709	0.4127
NT	0.9482	0.9455	0.9318	0.9046	0.8772	0.8489	0.0673	0.0670	0.0760	0.0986	0.1167	0.1293	0.9670	0.9462	0.8754	0.7290	0.5771	0.4217
Avg.	**0.9509**	**0.9477**	**0.9332**	**0.9049**	**0.8765**	**0.8475**	**0.0604**	**0.0605**	**0.0690**	**0.0912**	**0.1093**	**0.1252**	**0.9672**	**0.9454**	**0.8715**	**0.7193**	**0.5627**	**0.4070**

Table 1: Kakovost ustvarjenih napadov je ovrednotena s tremi različnimi merili, tj. merilom strukturnega indeksa podobnosti (Structural Similarity Index Measure - SSIM), zaznavno podobnostjo slikovnih delov (Perceptual Image Patch Similarity - LPIPS) in merilom kosinusnega indeksa podobnosti (Cosine Similarity Index Measure - CSIM). Vsak napad se primerja z ustreznim nespremenjenim globokim ponaredkom zbirke FF+, ustvarjenim z eno od petih metod globokih ponaredkov, ki so tukaj označene z DF (DeepFakes), F2F (Face2Face), FSh (FaceShifter), FS (Face Swap) in NT (NeuralTextures).

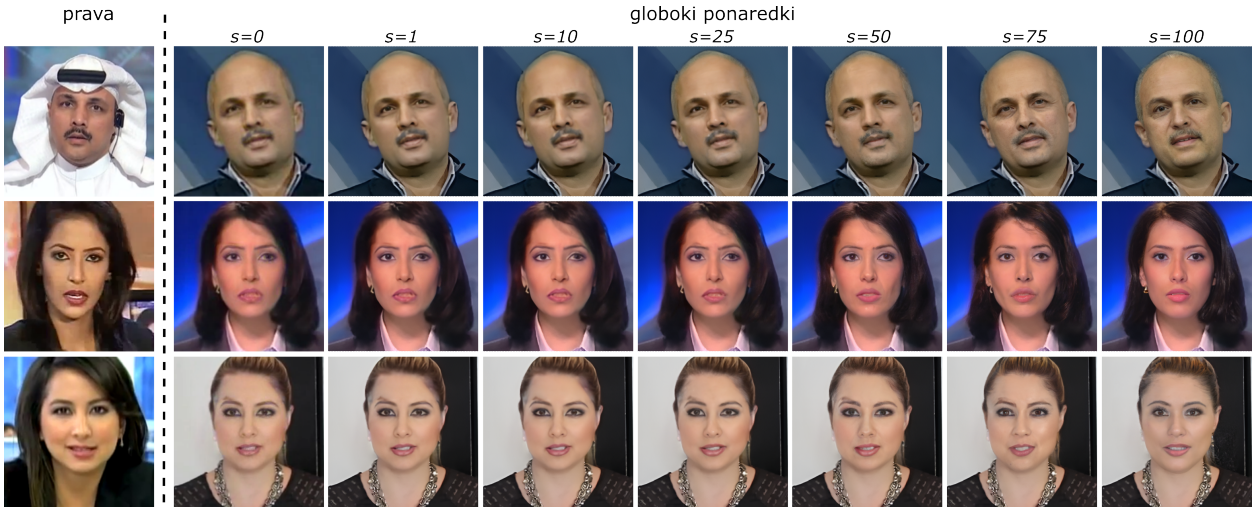

Figure 4: Kvalitativna primerjava ustvarjenih napadov globoko ponarejenih slik. Nizke vrednosti s (število difuzijskih korakov šumljenja) ne povzročijo vizualno zaznavnih sprememb slike. Višje vrednosti s po drugi strani popravijo neskladnosti globokih ponaredkov, vendar lahko v nekaterih primerih spremenijo videz različnih delov obraza.

posamezne vrednosti, so prikazani na sliki 2. Pri izvajanju eksperimentov uporabljamo vnaprej definirane podmnožice učnih, evalvacijskih in testnih FF+ slik.

Podrobnosti eksperimentov. Za generiranje napadov uporabljamo javno dostopno implementacijo orodja DifFace in vnaprej naučene uteži[1]. Ustvarjene napade nato uporabimo za testiranje ranljivosti treh različnih modelov za odkrivanje globokih ponaredkov: Xception [3], Face X-Ray [7] in SRM[2] [10]. Naivni klasifikator Xception in frekvenčni detektor SRM, sta bila naučena z naključno začetno inicializacijo parametrov. Pri tem smo za vsako FF+ metodo generiranja globokih ponaredkov naučili ločene detektorje. Evalvacijo samonadzorovanega detektorja Face X-Ray pa smo izvedli s predhodno naučenim modelom[3]. Vse usposobljene detektorje smo ločeno preizkušeni na običajnih FF+ globokih ponaredkih in napadih, ki so bili ustvarjeni z različnimi stopnjami šuma.

Metrike vrednotenja. Pred testiranjem ranljivosti detektorjev globokih lažnih slik na ustvarjene napade smo ocenili vizualno kakovost rekonstruiranih globoko ponarejenih slik, tako da smo jih primerjali z ustreznimi nespremenjenimi slikami. Njihovo strukturno podobnost smo ocenili z merilom strukturnega indeksa podobnosti (Structural Similarity Index Measure - SSIM). Za primerjavo slikovnih značilnosti višje ravni smo uporabili metodo naučene percepcijske podobnosti slik (Learned Perceptual Image Patch Similarity - LPIPS), ki temelji na predhodno naučenem omrežju SqueezeNet. Ohranitev identitete, predstavljene z nespremenjenimi globokimi ponaredki, smo izračunali s Cosine Similarity Index Measure (CSIM) na podlagi vektorjev identitete, izluščenih z modelom AdaFace.

Natančnost detektorjev globokih ponaredkov smo ocenili z izračunom deleža pravilno zaznanih globokih ponaredkov (True Positive Rate - TPR). Da bi zagotovili pošteno primerjavo posameznih poskusov in simulirali realni scenarij, je vsaka metoda odkrivanja najprej ocenjena na testni podmnožici pravih in nespremenjenih ponarejenih slik. Prag razvrščanja v delovni točki EER (angl. Equal Error Rate) smo nato uporabili tudi za razvrščanje napadov.

5 Rezultati

Ocena kakovosti napadov globoko ponarejenih slik. Ustvarjanje napadov globokih ponaredkov z izbranim pristopom prinaša neizogibne spremembe slike. Kvantitativna ocena njihove vizualne kakovosti je prikazana v tabeli 1. Kot je razvidno iz izračunanih vrednosti SSIM in LPIPS, rekonstrukcija globoko ponarejenih slik z do $s = 25$ korakov šumljenja bistveno ne spremeni njihove slikovne strukture. Večje število korakov šuma po drugi strani izraziteje spremeni začetno sliko, kar posledično zmanjša vrednost CSIM. To pomeni, da se je identiteta obraza do neke mere spremenila. Te ugotovitve potrjuje tudi kvalitativna analiza napadov. Kot je razvidno iz slike 4, ni očitnih, zaznavnih razlik med nespremenjenimi globoko ponarejenimi slikami ($s = 0$) in napadi, označenimi z $s = 1$, $s = 10$ in $s = 25$.

[1] https://github.com/zsyOAOA/DifFace
[2] https://github.com/crywang/face-forgery-detection
[3] https://github.com/wkq-wukaiqi/Face-X-Ray

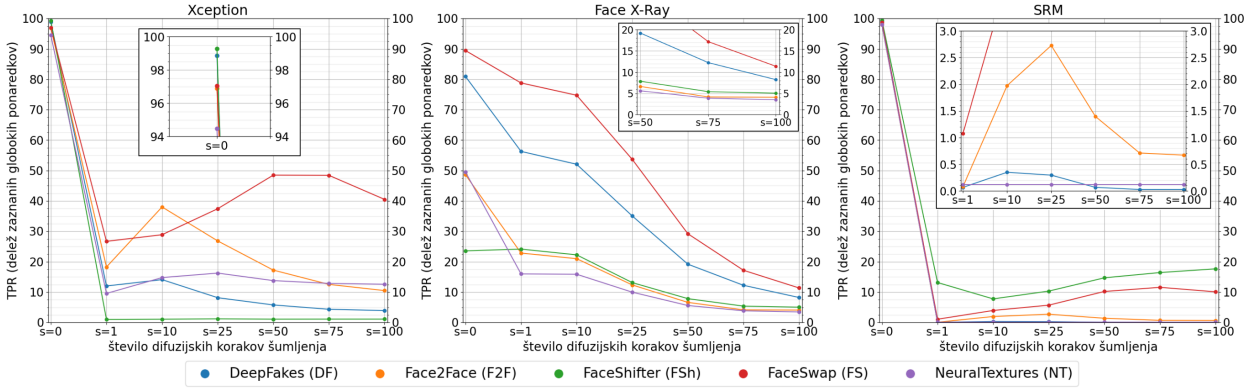

Figure 5: Odstotek odkritih globoko ponarejenih slik FF+ ($s = 0$) in napadov ($s > 0$) s testiranimi detektorji globokih ponaredkov. Prag za izračun TPR je enak pragu EER posameznih detektorjev, ko so ocenjeni na nespremenjenih lažnih slikah. ($s = 0$).

Opazimo, da višje ravni šuma v splošnem ustvarjajo veliko bolj realistične slike. Poleg tega je z $s = 100$ bilo ugotovljeno, da lahko difuzijski model v celoti odstrani tipične artefakte globokih ponaredkov, kot so dvojne obrvi, nenaravne sence, ostre meje šivanja obraza itd., vendar ta raven šuma pogosto spremeni tudi videz (velikost, obliko, barvo) posameznih delov obraza.

Evalvacija detektorjev globokih ponaredkov. Odstotek uspešno odkritih globokih ponaredkov smo najprej izračunali na običajnih pravih in ponarejenih slikah iz podatkovne baze FaceForensics++, da smo dobili vrednost praga EER za binarno razvrščanje vzorcev. S tem pragom smo nato razvrstili še vse generirane napade. Dobljeni rezultati so prikazani na sliki 5. Opazimo, da samo ena iteracija razšumljanja z difuzijskim modelom močno vpliva na natančnost testiranih detektorjev. Na splošno sta deskriminativni metodi (Xception in SRM) v primerjavi s samonadzorovano metodo (Face X-Ray) veliko bolj ranljivi. Opazili smo tudi, da se s povečevanjem števila korakov šumljenja (vrednost s) TPR začne do določene mere izboljševati. Predvidevamo, da pri teh ravneh šuma difuzijski model začne vnašati artefakte, ki jih lahko v nekaterih primerih prepoznajo detektorji globokih ponaredkov. Kljub temu so za razjasnitev tega pojava potrebne nadaljnje raziskave.

6 Zaključek

Nedavno odkriti difuzijski modeli za razšumljanje podatkov (Denoising Diffusion Models - DDM) so pokazali impresivne zmožnosti ustvarjanja zelo realističnih in prepričljivih slik. V tem članku raziskujemo njihovo potencialno uporabo za generiranje napadov globoko ponarejenih slik tipa črne škatle. Naše poskuse izvajamo na globokih ponaredkih iz podatkovne baze FaceForensics++. Napadamo tri različne metode za odkrivanje globokih ponaredkov na osnovi analize ene slike, tj. Xception, Face X-Ray in SRM. Naša študija je pokazala, da so vsi preizkušeni detektorji zelo ranljivi že za manjše spremembe slik, ki jih difuzijski model uporabi.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. MesoNet: A Compact Facial Video Forgery Detection Network. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018.

[2] N. Carlini and H. Farid. Evading Deepfake-Image Detectors with White- and Black-Box Attacks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2804–2813, 2020.

[3] F. Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017.

[4] P. Dhariwal and A. Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems (NIPS)*, volume 34, pages 8780–8794, 2021.

[5] A. Gandhi and S. Jain. Adversarial Perturbations Fool Deepfake Detectors. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.

[6] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley. Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3348–3357, 2021.

[7] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo. Face X-Ray for More General Face Forgery Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5000–5009, 2020.

[8] C. Liu, H. Chen, T. Zhu, J. Zhang, and W. Zhou. Making Deepfakes More Spurious: Evading Deep Face Forgery Detection Via Trace Removal Attack. *arXiv:2203.11433*, 2022.

[9] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu. Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 772–781, 2021.

[10] Y. Luo, Y. Zhang, J. Yan, and W. Liu. Generalizing Face Forgery Detection With High-Frequency Features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16317–16326, 2021.

[11] P. Neekhara, B. Dolhansky, J. Bitton, and C. Canton-Ferrer. Adversarial Threats to DeepFake Detection: A Practical Perspective. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 923–932, 2021.

[12] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen. Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos. In *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8, 2019.

[13] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner. FaceForensics++: Learning to Detect Manipulated Facial Images. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[14] K. Shiohara and T. Yamasaki. Detecting Deepfakes with Self-Blended Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18699–18708, 2022.

[15] C. Wang and W. Deng. Representative Forgery Mining for Fake Face Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14923–14932, 2021.

[16] Z. Yue and C. C. Loy. DifFace: Blind Face Restoration with Diffused Error Contraction. *arXiv:2212.06512*, 2022.