# Detekcija manipulacij za ohranjanje zasebnosti mehkih atributov na slikah obraza

Peter Rot[1,2], Peter Peer[2], Vitomir Štruc[1]
[1]Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana
[2]Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana
peter.rot@fe.uni-lj.si, peter.peer@fri.uni-lj.si, vitomir.struc@fe.uni-lj.si

## Izvleček

V strokovni literaturi se vse pogosteje pojavljajo potrebe po metodah za zagotavljanje zasebnosti v slikovnih podatkih. Na področju analize obrazov so raziskovalci predlagali metode, ki preslikajo sliko obraza tako, da je samodejno luščenje mehkih biometričnih lastnosti oteženo, obenem pa je vizualni videz slike podoben izvirni sliki. V tem članku predlagamo nov detekcijski postopek, ki zaznava, ali je bila slika spremenjena s tovrstnimi metodami. Postopek temelji na dejstvu, da razvrščevalnik mehkih biometričnih lastnosti tipično vrne različen rezultat za zaščiteno sliko in za sliko, ki je obdelana s postopkom za obnavljanje slike. Z eksperimenti pokažemo, da lahko to razliko uporabimo za detekcijo sprememb na sliki, ki so posledica uporabe metod za zaščito zasebnosti. Prednost naše metode je, da ne potrebuje znanja o uporabljeni metodi za zaščito zasebnosti. Detektor ovrednotimo na štirih metodah za ohranjanje zasebnosti mehkih atributov in na treh raznovrstnih zbirkah slik obrazov. Rezultati kažejo, da ima predlagan postopek vrsto prednosti pred konkurenčnimi rešitvami in da z visoko natančnostjo detektira manipulirane slike.
**Ključne besede:** analiza obraza, globoko učenje, mehke biometrične značilnosti, zasebnost

## Abstract

In scientific literature, there is a growing need for methods to ensure the privacy in digital images. In the field of face analytics, researchers have proposed privacy-preserving techniques which transform face images in such a way that the automatic extraction of soft-biometrics (e.g., gender) is prevented, while the visual appearance gets minimally degraded. We present a novel technique to detect whether or not an image was manipulated with such privacy-preserving techniques. Our detector exploits the fact that the soft–biometric classifier gives different results for privacy-enhanced images and their reconstructed versions. In our experiments, we have demonstrated that this difference can be exploited to detect whether an image was privacy-enhanced. The advantage of our method is that the used privacy-enhancing technique does not need to be known in advance (black-box scenario). Our approach is evaluated considering four privacy-enhancing techniques for soft-biometrics on three versatile face datasets. The results show that our approach has a number of advantages over competing techniques and that it detects privacy-enhancement with high accuracy.
**Keywords:** Face analytics, deep learning, soft-biometrics, privacy

## 1 UVOD

Nedaven napredek na področju računalniškega vida je izboljšal zmogljivosti sistemov za avtomatsko analizo obraza. S sodobnimi sistemi je možno povezati sliko obraza z identiteto posameznika in izluščiti obrazne lastnosti, kot so spol, starost, etnična pripadnost ali sorodstvena razmerja [Dantcheva et al., 2015]. Tehnolo- gija se uporablja v mnogih aplikacijah, npr. za zagotavljanje varnosti, kontrolo mej, preiskovanje kaznivih dejanj in za analize na socialnih omrežjih. Kljub številnim pozitivnim vidikom je možno tovrstne tehnologije za analizo biometričnih podatkov zlorabiti, kar močno ogroža zasebnost posameznikov [Meden et al., 2021]. Zato raziskovalci razvijajo mehanizme za zaščito zasebnosti [Mirjalili et al., 2019a, Mirjalili et al., 2020], ki bi hkrati zagotovili

*i*) uporabnost biometričnih podatkov in *ii*) ustrezen nivo zasebnosti uporabnikov. Zanimiva podskupina tovrstnih mehanizmov, s katero se ukvarjamo v tem delu, so tehnike za ohranjanje zasebnosti mehkih biometričnih značilnosti (tj. starost, spol, etnična pripadost ipd.), ki hkrati skušajo *i*) otežiti avtomatsko analizo obraznih značilnosti in *ii*) ohraniti vizualno podobo slike. Take tehnike so v praksi uporabne predvsem pri aplikacijah za deljenje slik ali na socialnih omrežjih, saj onemogočajo nezaželene avtomatske analize brez privolitve uporabnikov za namene ciljnega oglaševanja, demografskih analiz, diskriminacije in podobno.

Kljub številnim tehnikam za ohranjanje zasebnosti [Mirjalili et al., 2018, Mirjalili et al., 2019b] se mnoge zanašajo na to, da napadalec ne bo poskušal rekonstruirati zaščitene informacije [Rot et al., 2021]. V tem delu s pomočjo rekonstrukcije slike, ki izniči učinek metode za zaščito zasebnosti, želimo zaznavati, ali je bila slika obraza manipulirana z metodo za zaščito zasebnosti ali ne. Razumevanje, do katere mere je možno zaznati takšne manipulacije, je v praksi pomembno za uporabnike metod za zaščito zasebnosti, saj s tem spoznajo omejitve uporabljenih metod. Hkrati je to razumevanje pomembno za ponudnike storitev, ki upravičeno želijo zanesljivo izluščiti mehke biometrične značilnosti z avtomatskim postopkom. V tovrstnih sistemih bi bile lahko slike, na katerih je zaznana manipulacija, zavrnjene ali označene za bolj podroben nadaljnji pregled.

## 2 SORODNA DELA

### 2.1 Ohranjanje zasebnosti mehkih obraznih atributov

Raziskovalci so predlagali številne metode [Chhabra et al., 2018, Mirjalili et al., 2020], ki s spreminjanjem slikovnih elementov na sliki obraza otežijo avtomatsko analizo mehkih biometričnih značilnost. Dodaten cilj je ohraniti zmožnost avtomatske verifikacije osebe. Avtorji tehnike SAN [Mirjalili et al., 2018] so predlagali samo– kodirnik, ki temelji na konvolucijski nevronski mreži. Izboljšana verzija SAN je FlowSAN [Mirjalili et al., 2019b], s katero avtorji zvišajo raven splošne uporabnosti za druge razpoznavalnike mehkih atributov (tj. kako dobro metoda generalizira), ki niso bili vključeni v fazo učenja. Avtorji $k$-AAP [Chhabra et al., 2018] so predlagali tehniko, pri kateri se s pomočjo nasprotniških primerov da zaščititi množico več obraznih atributov. Ideja, da se z nasprotniškimi primeri da pretentati razvrščevalnik mehkih biometričnih značilnosti, je predstavljena tudi v delu [Rozsa et al., 2019].

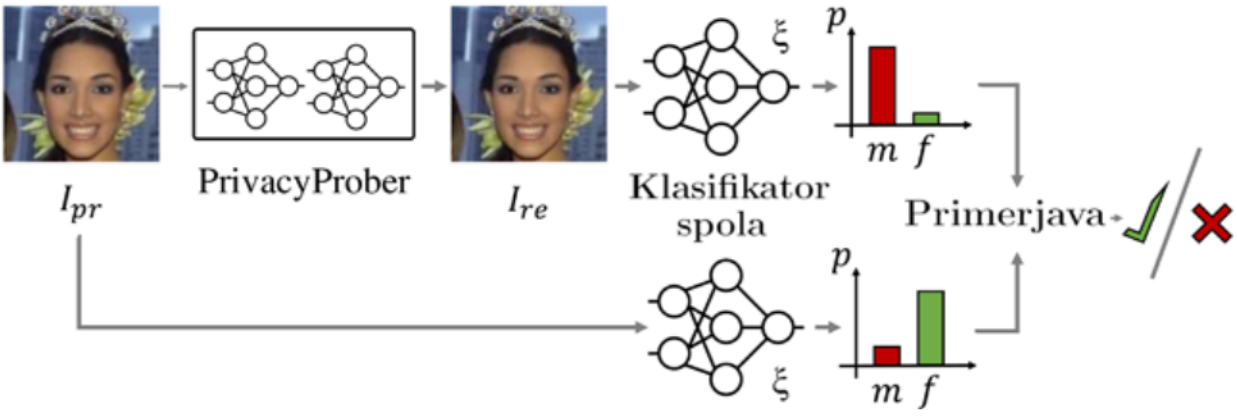
Raziskovalci [Rot et al., 2021] so demonstrirali, da z rekonstrukcijo slike obraza lahko izničimo učinek tehnik za ohranjanje zasebnosti. V tem delu rekonstrukcijo slike uporabimo za detekcijo, ali je bila na sliki uporabljena tehnika za ohranjanje zasebnosti ali ne.

### 2.2 Detekcija manipulacij za ohranjanje zasebnosti

Detekcija manipulacij za ohranjanje zasebnosti mehkih biometričnih značilnosti je nov problem, ki v odprti literaturi še ni bil naslovljen. Čeprav je zaznava tovrstnih manipulacij delno povezana z nasprotniškimi primeri [Cennamo et al., 2019, Meena and Tyagi, 2019], je pomembna razlika v tem, da metode za ohranjanje zasebnosti (med drugim) vsebujejo tudi sintezo slik [Mirjalili et al., 2019b], kar naredi ta problem znatno bolj obširen. Naše delo se razlikuje tudi od obstoječih postopkov za detekcijo globokih ponared- kov [Mirsky and Lee, 2020], ki odgovarjajo na vprašanje »Ali je bila slika manipulirana?«. Naš algoritem namreč odgovarja na bolj specifično vprašanje »Ali je bil manipuliran izbran obrazni atribut (npr. spol)?« Razlika je tudi v tem, da je primarni cilj globokih ponaredkov prelisičiti človeka, medtem ko se obravnavane metode osredotočajo na zaščito zasebnosti pri avtomatski analizi. Sodobne metode za detekcijo manipulacij, kot je npr. [Agarwal et al., 2020], so tipično ovrednotene v transparentnem scenariju (bela škatla), kjer je znano, s katero metodo za ohranjanje zasebnosti je bila slika manipulirana. V tem delu pa predstavimo detekcijsko metodo, ki tega znanja ne potrebuje (črna škatla). Naš algoritem primerjamo s pristopom T-SVM iz [Agarwal et al., 2020].

## 3 ZAZNAVANJE MANIPULACIJ

Detekcijska shema je prikazana na Sliki 1., pri čemer $\xi$ označuje razvrščevalnik, ki napoveduje izbrano obrazno značilnost (v našem primeru gre za binarno napoved spola). Detektor predpostavlja, da bo razvrščevalnik $\xi$ v primeru, ko je bila slika manipulirana z metodami za ohranjanje zasebnosti, generiral drugačno napoved pred in po rekonstrukciji slike. Natančneje, pri podani manipulirani sliki $I_{\mathrm{pr}}$ pričakujemo, da $\xi$ generira drugačne posteriorne verjetnosti $p(C_{\mathrm{k}}|I_{\mathrm{pr}})$ kot pri slikah $p(C_{\mathrm{k}}|I_{\mathrm{re}})$, ki so bile rekonstruirane. V našem delu za rekonstrukcijo slik uporabimo




Slika 1: **Detekcija manipulacij, ki jih povzroči metoda za ohranjanje zasebnosti, s pomočjo postopka za rekonstrukcijo slik PrivacyProber. Predlagana detekcijska shema izrablja razlikovanje med i) napovedjo originalne slike in ii) napovedjo rekonstruirane slike. Za razliko od drugih detektorjev manipulacij, predlagan postopek ne potrebuje faze učenja (t. i. scenarij črne škatle) in deluje le na podlagi primerjave napovedi.**

metodo PrivacyProber [Rot et al., 2021], ki je podrobneje opisana v nadaljevanju. Pri členu $C_k \in \{C_m, C_f\}$ označba $m$ označuje moški spol in $f$ ženski spol.

S primerjavo posteriornih verjetnosti je mogoče določiti, ali je bila slika modificirana ali ne. V naših eksperi- mentih smo za primerjavo posteriornih verjetnosti uporabili simetrično verzijo Kullback-Leibler divergence:

$$D_{SKL}(p,q) = D_{KL}(p||q) + D_{KL}(q||p),$$

kjer

$$D_{KL}(p||q) = \sum_{x \in \mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

in kjer $p = p(C_k | I_{pr})$, $q = p(C_k | I_{re})$, = $C_k$ in $D_{SKL}$ uporabljamo za merjenje manipulacij (oz. ohranjanje zasebnosti).

Pomembno je izpostaviti, da za razliko od obstoječih shem za detektiranje manipulacij predlagani postopek ne potrebuje učenja, zaradi česar ne potrebuje primerkov manipuliranih slik. Zanaša se izključno na ugotovitev, da se po uspešni odstranitvi učinka, ki ga je povzročila metoda za ohranjanje zasebnosti, napoved razvrščevalnika spremeni. Za detektor torej lahko pričakujemo, da deluje, ko PrivacyProber povzroči dovoljšno rekonstrukcijo skrite informacije o obraznem atributu.

## 4 EVALVACIJA IN REZULTATI

Za preizkus v kako splošnem primeru lahko predlagano metodo uporabljamo (tj. kakšna je njena sposobnost generalizacije), smo v poskusih uporabili tri raznovrstne zbirke obrazov, in sicer LFW [Huang et al., 2007], MUCT [Milborrow et al., 2010] in Adience [Eidinger et al., 2014]. Za ovrednotenje postopka smo za vsako izmed preizkušanih metod za zaščito zasebnosti mehkih biometričnih značilnosti ($k$-AAP, FGSM, FlowSAN-3 in FlowSAN-5) izbrali 698 zaščitenih in 698 originalnih slik, pri čemer je bil delež moških in žensk uravnotežen. Slike so bile nadalje naključno razdeljene v štiri disjunktne množice, kar nam omogoča poročanje standardnega odklona. Metoda za rekonstrukcijo slike je v našem delu najbolje delujoča implementacija PrivacyProber-ja iz dela [Rot et al., 2021], natančneje PP-DI (PrivacyProber with Denoising and Inpainting). PP-DI je rekonstrukcijska metoda, ki na sliki obnovi informacijo o mehkih biometričnih atributih s sekvenčno vezavo *i*) mreže za odstranjevanje šuma (angl. denoising) in *ii*) mreže za slikovno vrisovanje (angl. inpainting). Rezultati kažejo na to, da naš detekcijski algoritem dosega skoraj idealne rezultate na podatkovnih zbirkah LFW in MUCT. Povprečne AUC ocene na podatkovni zbirki Adience so slabše, tj. 0.828 za $k$-AAP, 0.828 za *FGSM* , 0.696 za FlowSAN–3 in 0.751 za FlowSAN–5. To pripisujemo dejstvu, da je podatkovna zbirka Adience bolj zahtevna, kar se je izkazalo tudi v delu [Rot et al., 2021]. V primerjavi z drugima dvema zbirkama (LFW in MUCT) ima namreč večjo variabilnost pri svetlobnih pogojih, pozah obrazov in vsebuje slike slabše kvalitete.

Ponovno poudarjamo, da problem detekcije manipulacij, ki jih naredijo algoritmi za ohranjanje zasebnosti mehkih obraznih značilnosti, še ni bil obravnavan v odprti literaturi. Kljub temu smo delovanje našega postopka primerjali s sodobno detekcijsko

metodo T–SVM [Agarwal et al., 2020], ki je sicer razvita za reševanje ožjega problema detekcije nasprotniških primerov. Detekcija nasprotniških primerov je namreč po nekaterih karakteristikah podobna našemu izzivu. T–SVM deluje tako, da slike obrazov najprej preslika z diskretno valjčno transformacijo (discrete wavelet transform, DWT) in z diskretno sinusno transformacijo (discrete sine transform, DST). Iz preslikav nato izlušči GIST značilke, na podlagi katerih se uči model z metodo podpornih vektorjev (support vector machine, SVM). T–SVM za razliko od našega postopka torej potrebuje primerke slik, na katerih so prisotne manipulacije, ki jih povzročijo metode za ohranjanje zasebnosti. V naših eksperimentih zato za primerjavo obravnavamo dve različni strategiji, in sicer:

- **Scenarij bele škatle:** Pri tej strategiji predpostavljamo dostop do vseh modelov za ohranjanje zasebnosti mehkih atributov. V naših eksperimentih je bil T–SVM ločeno naučen za vsak model za ohranjanje zasebnosti. Za učenje detektorjev smo uporabili učni del podatkovne zbirke LFW, testirali pa smo na testnih delih vseh obravnavanih zbirk (LFW, MUCT in Adience).
- **Scenarij črne** škatle: Za to strategijo predpostavljamo dostop le do enega modela za zaščito zasebnosti. T–SVM detektor mora zato generalizirati za vse ostale modele, do katerih ni imel dostopa. V naših eksperimentih obravnavamo dva ločena modela za detekcijo. Prvi model T–SVM (A) je naučen samo na slikah iz podatkovne zbirke LFW, ki so bile manipulirane s FlowSAN–5. Drugi model T–SVM (B) je naučen izključno na slikah, ki so bile manipulirane z metodo $k$–AAP, ponovno samo na slikah iz LFW. Oba obravnavana modela sta testirana na vseh podatkovnih zbirkah in modelih za ohranjanje zasebnosti.

Tabela 1. prikazuje, da je T–SVM pri scenariju bele škatle na podatkovni zbirki LFW primerljiv z našim postopkom, na podatkovnih zbirkah MUCT in Adience pa deluje občutno slabše. Iz tega rezultata lahko sklepamo, da karakteristike podatkovnih zbirk močno vplivajo na T–SVM, kljub temu, da modelu damo občutno prednost in mu omogočimo dostop do primerkov zaščitenih slik. Za razliko od T–SVM naš detektor namreč ne potrebuje tovrstnih učnih primerov. Opazimo, da T–SVM v primeru bele škatle in T–SVM(A) v primeru črne škatle ne naredi nobene napake. Postopka FlowSAN na sliki namreč naredita značilne vzorce (razvidne tudi na Sliki 2.), ki se jih model nauči dobro razpoznavati.

Tabela 1: **ocene AUC ($\mu \pm \sigma$) za detekcijo manipulacij, ki jih povzročijo metode za ohranjanje zasebnosti. Naš postopek, ki ne potrebuje učne faze (tj. scenarij črne škatle), primerjamo s sodobno tehniko T–SVM za detekcijo nasprotniških primerov, pri čemer ločeno obravnavamo oba scenarija (bele in črne škatle).**

| Model za ohranjanje zasebnosti | Zbirka | Naš model | Bela škatla | Črna škatla | |
|---|---|---|---|---|---|
| | | | T–SVM | T–SVM (A) † | T–SVM (B) ‡ |
| *k*–AAP | LFW | 0.980 ± 0.003 | 0.984 ± 0.007 | 0.743 ± 0.013 | 0.984 ± 0.007 |
| | MUCT | 0.984 ± 0.001 | 0.534 ± 0.021 | 0.727 ± 0.030 | 0.534 ± 0.021 |
| | Adience | 0.828 ± 0.002 | 0.541 ± 0.007 | 0.604 ± 0.008 | 0.541 ± 0.007 |
| FGSM | LFW | 0.992 ± 0.001 | 0.955 ± 0.011 | 0.894 ± 0.013 | 0.921 ± 0.016 |
| | MUCT | 0.995 ± 0.000 | 0.852 ± 0.012 | 0.858 ± 0.015 | 0.552 ± 0.030 |
| | Adience | 0.877 ± 0.004 | 0.624 ± 0.009 | 0.595 ± 0.016 | 0.471 ± 0.014 |
| FlowSAN–3 | LFW | 0.980 ± 0.002 | 1.000 ± 0.000 | 1.000 ± 0.000 | 0.997 ± 0.002 |
| | MUCT | 0.991 ± 0.003 | 1.000 ± 0.000 | 1.000 ± 0.000 | 0.524 ± 0.020 |
| | Adience | 0.696 ± 0.003 | 1.000 ± 0.000 | 1.000 ± 0.000 | 0.609 ± 0.005 |
| FlowSAN–5 | LFW | 0.976 ± 0.002 | 1.000 ± 0.000 | 1.000 ± 0.000 | 0.998 ± 0.001 |
| | MUCT | 0.991 ± 0.004 | 1.000 ± 0.000 | 1.000 ± 0.000 | 0.531 ± 0.018 |
| | Adience | 0.751 ± 0.003 | 1.000 ± 0.000 | 1.000 ± 0.000 | 0.563 ± 0.005 |

† Detekcijski model je učen na FlowSAN–5 in LFW.
‡ Detekcijski model je učen na *k*–AAP in LFW.

Slika 2: **Primeri napačnih napovedi predlaganega postopka. Prva vrstica prikazuje slike, na katerih postopek ni zaznal, da so bile slike manipulirane z metodo za zaščito zasebnosti. V drugi vrstici so primeri, kjer je postopek detektiral manipulacije na slikah, kljub temu, da le–te niso bile prisotne.**

Ko gledmo AUC ocene za scenarij črne škatle, vidimo, da so detekcijski rezultati T–SVM na splošno slabši. T–SVM (A), ki je učen na slikah, manipuliranih z FlowSAN–5, še vedno lahko zagotovi idealno detekcijo za oba FlowSAN modela na vseh podatkovnih zbirkah, a je hkrati manj kompetitiven za detekcijo $k$–AAP in FGSM.

T–SVM (B) model, učen na manipuliranih slikah s $k$–AAP na podatkovni zbirki LFW, je med primerjanimi modeli najmanj kompetitiven. Naš postopek premaga le v redkih primerih na LFW podatkovni zbirki.

Predstavljeni rezultati demonstrirajo dodano vrednost našega postopka, ki ne potrebuje učenja in primerkov manipuliranih slik. Učinkovitost našega detektorja je tudi manj občutljiva na spremembo testnih podatkov, za razliko od T–SVM, ki ima manjše sposobnosti za generalizacijo na druge podatkovne zbirke.

Primeri napak predlaganega postopka so prikazani na Sliki 2. Za vsako izmed obravnavanih metod za zaščito zasebnosti sta prikazana primera za oba tipa napake – prva vrstica prikazuje napačno negativne, druga pa napačno pozitivne primere.

## 5 SKLEP

V tem delu smo proučevali problem detekcije manipulacij na slikah, ki jih naredijo tehnike za zaščito zasebnosti mehkih biometričnih značilnosti. Predlagali smo nov algoritem za detekcijo in ga ovrednotili na štirih sodobnih metodah za zaščito zasebnosti in na treh eksperimentalnih podatkovnih zbirkah. Rezultati naše študije kažejo na to, da s predlaganim detektorjem lahko zaznamo manipulacije vseh obravnavanih modelov za ohranjanje zasebnostih. Iz teh rezultatov sledi, da:

- V primeru, da sistem zaščito zasebnosti lahko zazna, zaščitene slike lahko avtomatsko označimo za ročni pregled ali za procesiranje z bolj prefinjenimi tehnikami. Tovrstno procesiranje bi lahko onemogočilo učinek metod za ohranjanje zasebnosti.
- Metode za zaščito zasebnosti mehkih atributov bi morale biti podane z vnaprej definiranimi garancijami, ki bi zagotovile, da zaščitene slike ne bi mogle biti zlorabljene. To vzbuja potrebo po formalnih modelih za zaščito zasebnosti mehkih biometričnih značilnosti, ki bi omogočili matematično izračunljivo stopnjo zagotovljene zasebnosti. Formalne metode za ohranjanje zasebnosti so bile v odprti literaturi za zdaj razvite le za namene deidentifikacije, za zaščito mehkih atributov pa še ne.

V nadaljnem delu zato načrtujemo raziskati možnosti formalnih metod za ohranjanje zasebnosti mehkih atributov, ki bi naslavljale prej omenjene izzive in bi zasebnost ohranjale tudi v primeru bolj

naprednega procesiranja slik. Metodo načrtujemo evalvirati tudi na drugih obraznih atributih.

## LITERATURA


[1] [Agarwal et al., 2020] Agarwal, A., Singh, R., Vatsa, M., and Ratha, N. K. (2020). Image Transformation based Defense Against Adversarial Perturbation on Deep Learning Models. *Transactions on Dependable and Secure Computing*.

[2] [Cennamo et al., 2019] Cennamo, A., Freeman, I., and Kummert, A. (2019). A Statistical Defense Approach for Detecting Adversarial Examples. *International Conference on Pattern Recognition and Intelligent Systems*.

[3] [Chhabra et al., 2018] Chhabra, S., Singh, R., Vatsa, M., and Gupta, G. (2018). Anonymizing k-Facial Attributes via Adversarial Perturbations. *Proceedings of International Joint Conferences on Artificial Intelligence*.

[4] [Dantcheva et al., 2015] Dantcheva, A., Elia, P., and Ross, A. (2015). What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics. *IEEE Transactions on Information Forensics and Security (TIFS)*, 11(3):441–467.

[5] [Eidinger et al., 2014] Eidinger, E., Enbar, R., and Hassner, T. (2014). Age and Gender Estimation of Unfiltered Faces. *IEEE Transactions on Information Forensics and Security (TIFS),* 9(12):2170–2179.

[6] [Huang et al., 2007] Huang, G. B., Berg, M. R., Berg, T., and E., L.-M. (2007). Labeled Faces in the Wild.: A Database for Studying Face Recognition in Unconstrained Environments. *University of Massachusetts, Amherst, Technical Report 07-49*.

[7] [Meden et al., 2021] Meden, B., Rot, P., Terhörst, P., Damer, N., Kuijper, A., Scheirer, W. J., Ross, A., Peer, P., and Štruc, V. (2021). Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, 16:4147–4183.

[8] [Meena and Tyagi, 2019] Meena, K. and Tyagi, V. (2019). Image Forgery Detection: Survey and Future Directions. *Data, Engineering and Applications, Springer*, pages 163–194.

[9] [Milborrow et al., 2010] Milborrow, S., Morkel, J., and Nicolls, F. (2010). The MUCT Landmarked Face Database. *Pattern Recognition Association of South Africa*.

[10] [Mirjalili et al., 2018] Mirjalili, V., Raschka, S., Namboodiri, A., and Ross, A. (2018). Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images. In *International Conference on Biometrics (ICB)*, pages 82–89.

[11] [Mirjalili et al., 2019a] Mirjalili, V., Raschka, S., and Ross, A. (2019a). FlowSAN: Privacy-enhancing Semi- adversarial Networks to Confound Arbitrary Face-based Gender Classifiers. *IEEE Access*, 7:99735–99745.

[12] [Mirjalili et al., 2019b] Mirjalili, V., Raschka, S., and Ross, A. (2019b). FlowSAN: Privacy-enhancing Semi- Adversarial Networks to Confound Arbitrary Face-based Gender Classifiers. *IEEE Access 7*, pages 99735–99745.

[13] [Mirjalili et al., 2020] Mirjalili, V., Raschka, S., and Ross, A. (2020). PrivacyNet: Semi-adversarial Networks for Multi-attribute Face Privacy. *IEEE Transactions on Image Processing (TIP)*, 29:9400–9412.

[14] [Mirsky and Lee, 2020] Mirsky, Y. and Lee, W. (2020). The creation and detection of deepfakes: A survey. *CoRR*, abs/2004.11138.

[15] [Rot et al., 2021] Rot, P., Peer, P., and Štruc, V. (2021). PrivacyProber: Assessment and Detection of Soft–Biometric Privacy–Enhancing Techniques. *under review*, pages 1–18.

[16] [Rozsa et al., 2019] Rozsa, A., Günther, M., Rudd, E. M., and Boult, T. (2019). Facial Attributes: Accuracy and Adversarial Robustness. *Pattern Recognition Letters*, 124:100–108.