

Daljinska avtentikacija obrazov z interpretabilno zaščito zasebnosti spola in izvorne identitete

Peter Rot^{1,2}

¹Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana, Slovenija

²Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana, Slovenija

E-pošta: peter.rot@fri.uni-lj.si

Povzetek. Avtentikacija z uporabo metod za razpoznavanje obrazov je razširjen in prikladen način za preverjanje ujemanja trenutnega biometričnega vzorca z biometričnim vzorcem, registriranim v sistemu. Na tem področju veliko tveganje za zasebnost izhaja iz dejstva, da je mogoče iz biometričnih podatkov obraza poleg identitete razbrati tudi informacijo o mehkih biometričnih značilnostih, kot je spol. V članku predlagamo metodo za zaščito zasebnosti, ki omogoča daljinsko avtentikacijo, pri čemer iz poslanih biometričnih podatkov ni mogoče ugotoviti spola osebe ali jih neposredno povezati z izvorno sliko obraza, iz katere so bili pridobljeni. Zaščita zasebnosti se izvaja izključno na uporabnikovi napravi in ne zahteva nobenih sprememb strežniškega sistema za avtentikacijo. Metoda deluje na podlagi zlivanja obrazov z uporabo generativne nevronske mreže, pri čemer se na strežnik pošlje le biometrične podatke z izboljšano zasebnostjo. Prednost naše metode v primerjavi z obstoječimi rešitvami je višja stopnja razločljivosti, saj vsak zaščiten obraz nastane z determinističnim zlivanjem po vnaprej določenih kriterijih. Na podatkovni zbirki CelebA demonstriramo, da predlagani pristop zagotovi učinkovito deidentifikacijo glede na izvorno sliko osebe in zaščiti informacijo o spolu. Metoda dobro generalizira tudi v primerih, ko v fazi testiranja zamenjamo razpoznavalnik obraza ali razpoznavalnik spola.

Ključne besede: biometrija, razpoznavanje obrazov, avtentikacija, deidentifikacija, zasebnost

Remote Face Authentication with Interpretable Privacy-Enhancement of Gender and Identity

Authentication using face recognition is a widespread and convenient way to verify whether a presented biometric sample matches a previously enrolled template. A major privacy risk, however, is that facial biometric data reveals not only identity but also soft biometric attributes such as gender. We propose a remote face authentication method that ensures the data sent to the server does not disclose the user's gender and cannot be linked to the original face image. Our approach relies on deterministic face fusion with a generative neural network where only the transformed biometrics are transmitted to the server. Compared with existing solutions, our method is interpretable because each protected face is produced through deterministic fusion under predefined criteria. On the CelebA dataset, we show that the method achieves effective deidentification with respect to source face images and protects gender privacy. It also generalizes well even when the face recognition model or gender classifier is swapped at test time.

Keywords: biometrics, facial recognition, authentication, deidentification, privacy

1 UVOD

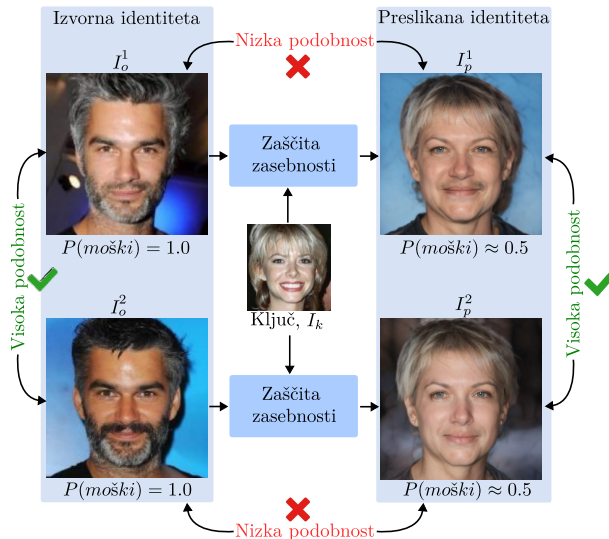
Tehnologija za razpoznavanje obrazov se vse pogostejše uporablja v sistemih za daljinsko avtentikacijo [1], [2].

Prejet 16. januar, 2026

Odobren 24. junij, 2026



Avtorske pravice: © 2026
Creative Commons Attribution 4.0
International License



Slika 1 Predlagana metoda na uporabnikovi napravi zaščiti zasebnost izvorne identitete in spola z zlivanjem slik obrazov iste osebe (npr. I_o^1 in I_o^2) z vnaprej izbranim referenčnim obrazom nasprotnega spola (ključ I_k). Preslikane slike (I_p^1 in I_p^2) ohranijo zadostno podobnost za daljinsko avtentikacijo uporabnika, brez razkritja izvorne identitete ali spola.

V takih sistemih se iz uporabnikove naprave na strežnik običajno pošlje slika obraza ali pa iz nje izluščene značilke. Strežnik nato preveri, ali so prejeti biometrični podatki dovolj podobni tistim, ki so bili prejeti ob registraciji uporabnika. Iz slik obrazov in iz nji-

hovich značilnik pa je poleg identitete mogoče razbrati tudi druge občutljive podatke (npr. spol, raso, starost), ki jih imenujemo *mehke biometrične značilnosti* [3]. Ker njihova nepooblaščen obdelava prinaša tveganje za zasebnost uporabnikov, so raziskovalci predlagali t. i. metode za zaščito mehkih biometričnih značilnosti [4], [5], ki jih krajše označujemo kot metode SB-PET (angl. *soft-biometric privacy-enhancing techniques*). Metode SB-PET sledijo dvema ciljema: (1) ohraniti dovolj diskriminativne informacije za avtentikacijo ter hkrati (2) onemogočiti zanesljivo določanje izbranih mehkih biometričnih značilnosti (npr. spola). Pomanjkljivost podobnih metod SB-PET [6] je, da sprememb, ki jih naredijo na izvornih slikah, ni mogoče na preprost način interpretirati.

Drugačnemu cilju sledijo metode za deidentifikacijo [7], [8], katerih primarni cilj je odstraniti informacije o identiteti oseb, ki se pojavljajo na sliki. Če dodatno zahtevamo, da se pri deidentifikaciji ohranijo mehke biometrične značilnosti oseb na sliki [9], lahko deidentificirane slike uporabimo za nadaljnjo analizo (npr. poročanje demografske statistike).

Oba pristopa (metoda SB-PET in deidentifikacija) se hkrati redko uporabljata v sistemih za daljinsko avtentikacijo, saj sledita *nasprotujočim si ciljem*: metode SB-PET ohranjajo identiteto in skrivajo mehke značilnosti, deidentifikacija pa skriva identiteto in pogosto ohranja mehke značilnosti. V tem delu pokažemo, da je oba pristopa možno združiti v enovit sistem za avtentikacijo uporabnikov.

Da bi jasno opredelili aplikativni okvir v nadaljevanju predlaganega pristopa, je na tem mestu potrebno razločiti med dvema različnima pojmom [10], [11]. (i) *Dokazovanje identitete* (angl. *identity proofing*) označuje postopke, katerih cilj je vzpostavitev in preverjanje vezi med digitalno identiteto in realno fizično osebo, za katero se posameznik predstavlja. Takšni postopki praviloma vključujejo dodatne mehanizme, kot je preverjanje identifikacijskih dokumentov, zaznavanje prisotnosti ali drugi zaupanja vredni postopki zajema. (ii) *Preverjanje avtentikacijskih podatkov* pa označuje postopek preverjanja, ali predstavljeni podatki ustrezajo že vzpostavljeni digitalni identiteti, pri čemer je cilj potrditi kontinuiteto istovetnosti znotraj sistema. V predlaganem pristopu se ta avtentikacija izvaja na osnovi biometričnih podatkov, vendar njen namen ni dokazovanje identitete osebe v realnem svetu, temveč preverjanje podobnosti s predhodno registriranim zapisom. Tak način avtentikacije je po funkcionalnosti primerljiv z avtentikacijo z geslom, vendar je za uporabnika praviloma bolj prikladen, saj ne zahteva ročnega vnosa in je deterministično vezan na uporabnika prek njegovih biometričnih značilnosti.

Predlagamo metodo za daljinsko avtentikacijo, ki na uporabnikovi napravi deterministično transformira sliko obraza, preden je ta poslana na strežnik. Informacijo o spolu zaščitimo tako, da izvorni obraz zlijemo z obrazom osebe nasprotnega spola. Cilj metode je she-

matsko prikazan na sliki 1. Pri tem poskrbimo, da je transformiran obraz dovolj drugačen od izvornega, s čimer dosežemo deidentifikacijo glede na izvorno sliko (tj. podobnost med identiteto osebe na izvorni sliki in identiteto osebe na zaščiteni sliki je nizka). Prednost predlaganega pristopa je boljša možnost interpretacije sprememb na izvorni sliki, saj pri posameznem obrazu natančno vemo, s katerim obrazom je bil zlit. Ključno je, da sta si sliki iste osebe v zaščiteni domeni zelo podobni, kar omogoča uspešno avtentikacijo na strežniku.

2 PREGLED PODROČJA IN SORODNA DELA

2.1 Metode SB-PET

Formulacija problema. Naj bo $\mathcal{I}$ prostor obraznih slik in $\mathcal{G} = \{\text{moški, ženska}\}$ množica binarnih spolnih oznak. Za poljubno sliko $I \in \mathcal{I}$ vnaprej naučen klasifikator spola ξ vrača

$$P_{\xi}(\text{moški} \mid I) \in [0, 1], \text{ pri čemer} \\ P_{\xi}(\text{ženska} \mid I) = 1 - P_{\xi}(\text{moški} \mid I) .$$

Izvorno sliko obraza označimo z $I_o \in \mathcal{I}$. Cilj metod za zaščito zasebnosti mehkih biometričnih lastnosti (SB-PET), ki jih označimo kot $\psi : \mathcal{I} \rightarrow \mathcal{I}$, je ustvariti transformirano sliko $I_p = \psi(I_o)$, pri kateri je informacija o spolu zabrisana, tj.

$$P_{\xi}(\text{moški} \mid I_p) \approx 0.5 .$$

Dodatni cilj metod SB-PET je omogočiti preverjanje identitete v prostoru zasebnostno izboljšanih slik.

Ohranjanje identitete v zaščitenem prostoru. Številne metode SB-PET [12], [13], [14] zahtevajo, da zaščiteni sliki I_p ohrani čim več identitetnih značilnosti svoje izvorne slike I_o . Naj bo $\varphi : \mathcal{I} \rightarrow \mathbb{R}^d$ model za razpoznavanje obrazov, pri čemer podobnost med slikama I in J merimo s kosinusno podobnostjo:

$$F(I, J) = \frac{\langle \varphi(I), \varphi(J) \rangle}{\|\varphi(I)\| \|\varphi(J)\|} .$$

Naj bosta I_o^i in I_o^j dve izvorni sliki iste osebe ter I_p^i in I_p^j njuni ustrezni zasebnostno izboljšani različici. Naj bo τ_{id} prag kosinusne podobnosti, tako da za dve obrazni sliki I in J velja, da gre za ujemanje identitete, kadar $F(I, J) \geq \tau_{id}$. Če veljajo pogoji:

$$F(I_o^i, I_o^j) \geq \tau_{id}, \\ F(I_o^i, I_p^i) \geq \tau_{id}, \\ F(I_o^j, I_p^j) \geq \tau_{id},$$

lahko pričakujemo, da sta si podobni tudi zaščiteni sliki:

$$F(I_p^i, I_p^j) \geq \tau_{id} .$$

Nekatere metode, kot je PrivacyNet [6], omenjeni pogoj sprostijo in ne zahtevajo visoke stopnje podobnosti med izvorno sliko in njeno zaščiteno različico, tj.

$$F(I_o^i, I_p^i) \ll \tau_{id} .$$

V tem delu pokažemo, da tak pristop omogoča učinkovito deidentifikacijo glede na par izvorna slika in zaščiten slika. V tem delu ta pogoj eksplicitno vključimo v našo metodo.

2.2 Obstoječe metode SB-PET za zaščito zasebnosti spola

Zgodnja raziskovalna dela, katerih cilj je prikriti informacijo o spolu [15], so se osredotočala na neposredno manipulacijo slik obrazov. Nekateri pristopi spreminjajo barvne kanale slike, dokler klasifikator spola ne začne delati napak [15], drugi na sliko dodajajo nasprotniški šum [16]. Poznejša metoda Semi-Adversarial Networks (SAN) [17] temelji na zlivanju obrazov, kjer se najprej izračunata prototipa povprečnega moškega in povprečnega ženskega obraza. Za zaščito zasebnosti se izvorna slika zlije s prototipom nasprotnega spola, kar zmanjša zanesljivost klasifikacije spola. Nadgradnja tega pristopa je PrivacyNet [6], pri katerem avtorji naučijo generativno nevronske mreže, da na izvorni sliki dodaja vzorce za prekrivanje informacije o spolu. Glavna omejitev tega pristopa je nerazločljivost, saj ni jasno, kateri latentni dejavniki povzročijo posamezne spremembe na obrazu in zakaj.

V nasprotju s tem je naš pristop bolj razločljiv, saj namesto netransparentnih nasprotniških perturbacij ali latentnega šuma deterministično zlijemo izvorno sliko z izbrano sliko nasprotnega spola (t.i. *ključ*). Slike zlivamo v latentnem prostoru mreže StyleGAN [18]. Ker je postopek strukturiran in sledljiv, je nadzor nad transformacijo boljši. Konceptualno sorodno delo je predstavljeno v [19], kjer avtorji obravnavajo fotorealistično zlivanje prostorsko razčlenjenih segmentov z namenom zaščite zasebnosti spola. V nasprotju s tem delom je bil tamkajšnji pristop ovrednoten predvsem na sintetičnih obrazih, brez analize uspešnosti deidentifikacije glede na izvorno identiteto. V našem delu metodo ovrednotimo v celotnem scenariju daljinske avtentikacije z realnimi podatki ter eksplicitno analiziramo kompromis med zaščito spola, ohranjanjem preverljivosti identitete in deidentifikacijo.

3 SISTEMSKE PREDPOSTAVKE IN MODEL APLIKACIJE

Predlagani mehanizem za zaščito zasebnosti je zasnovan za naslednji razred aplikacij in ob upoštevanju naslednjih predpostavk:

- Sistem temelji na arhitekturi odjemalec-strežnik, kjer se uporabnik avtentificira na daljavo, pri čemer obrazna avtentikacija služi kot uporabniku prijazen nadomestek ročnega vnosa gesla ali PIN-a.
- Namen sistema je avtentikacija v smislu preverjanja kontinuitete digitalne istovetnosti znotraj sistema, ne pa dokazovanje identitete osebe v realnem svetu ali regulirani postopki preverjanja identitete (npr. KYC).

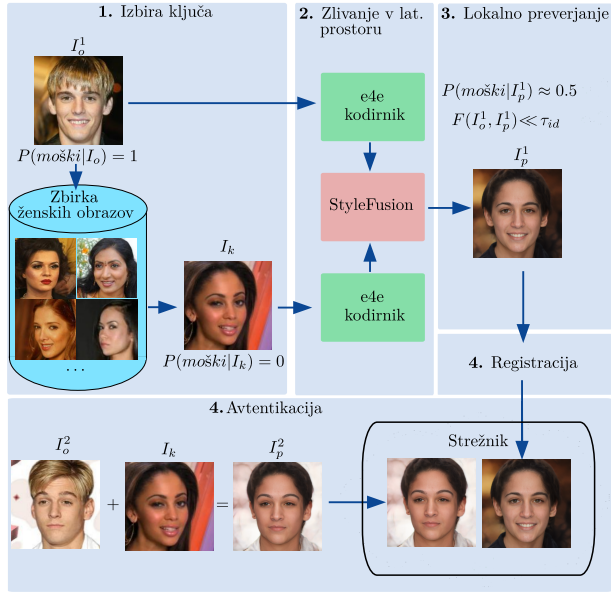
- Strežnik je zaupanja vreden, vendar radoveden (angl. *honest-but-curious*). To pomeni, da ne odstopa od deklariranega funkcionalnega delovanja sistema, vendar lahko shranjuje in analizira prejete podatke. Cilj predlaganega pristopa je izboljšanje zasebnosti uporabnikov na strežniški strani v skladu z GDPR (načelo minimizacije posredovanih podatkov). Strežnik prejme le informacije, ki so nujno potrebne za izvedbo avtentikacije.
- Digitalna identiteta uporabnika (npr. uporabniški račun) je ločena od uporabljenega avtentikacijskega mehanizma; avtentikacija temelji na primerjavi 1:1 z že registriranim zapisom.
- Avtentikator v obravnavanem sistemu je izpeljan iz biometričnih podatkov, vendar ni neposredna predstavitev uporabnikove biometrije. Sprememba referenčnega obraza (ključa) zato omogoča vzpostavitev novega avtentikatorja brez spremembe uporabniškega računa.
- Uporaba istega referenčnega obraza pri različnih uporabnikih je dovoljena in ne pomeni varnostne pomanjkljivosti v obravnavanem modelu, saj se avtentikacija vedno izvaja v kontekstu konkretnega uporabniškega računa (analogno avtentikaciji z geslom).
- Referenčni obraz, uporabljen pri transformaciji, se obravnava kot parameter avtentikacijskega mehanizma in ni kriptografski ključ; po potrebi ga je mogoče zamenjati ali ponovno izbrati.

4 METODA

Visokonivojski pregled. Postopek zaščite zasebnosti poteka v štirih korakih, kot je prikazano na sliki 2. Najprej izberemo obraz nasprotnega spola, ki ga uporabimo kot *ključ* za zlivanje z obrazom uporabnika. V drugem koraku se oba obraza zlijeta v skupno latentno predstavitev, iz katere nastane zaščiten slika. Tretji korak je lokalno preverjanje uspešnosti zaščite: če je zaščita ustrezna, se dobljena slika (ali iz nje izluščene značilke) pošlje na strežnik; sicer se postopek ponovi z drugim ključem.

Ob registraciji strežnik iz prve prejete slike izlušči in shrani biometrične značilke. Pri vsaki nadaljnji avtentikaciji se zajeta slika uporabnikovega obraza ponovno zlije z istim ključem, kot je bil uporabljen ob registraciji. Na strežniku se nato značilke iz novo poslane slike primerjajo z referenčnimi značilkami, shranjenimi ob registraciji. Posamezne faze postopka so podrobneje opisane v nadaljevanju.

Izbor ključa. Predpostavljamo, da ima uporabnik, ki želi zaščititi svojo zasebnost, lokalni dostop do zbirke obraznih slik, bodisi resničnih bodisi sintetičnih. Zbirka lahko vsebuje generične slike iz odprtih virov ali umetno generirane obraze nasprotnega spola, ki jih zagotovi uporabnik ali aplikacija. Če je uporabnik moškega spola, potrebuje zbirko ženskih obrazov, in obratno.



Slika 2 Za izvorno sliko, I_o^1 , najprej izberemo sliko identitete nasprotnega spola (ključ I_k). Slike I_o^1 in I_k pretvorimo v latentni prostor z modelom e4e in s postopkom StyleFusion zlijemo v I_p^1 . Če je zaščita zasebnosti uspešna, osebo na strežniku registriramo z I_p^1 . Pri avtentikaciji se izvorna slika I_o^2 (in vse nadaljnje) zlijejo z I_k , strežnik pa preveri ujemanje med I_p^1 in I_p^2 .

Uporabnik najprej izračuna podobnosti med svojo izvorno obrazno sliko I_o in vsemi slikami iz zbirke, pri čemer uporabi model za razpoznavanje obrazov (npr. ArcFace). Slike v zbirki se nato razvrstijo po naraščajoči podobnosti.

Da bi transformirana slika vsebovala čim manj informacij o izvorni identiteti, kandidate za ključ I_k izbiramo med obrazi z nizko stopnjo podobnosti z izvorno sliko. Uporabnik izračuna podobnosti med izvorno sliko I_o in vsemi slikami v zbirki nasprotnega spola ter jih razvrsti po naraščajoči podobnosti. Kandidati za ključ se nato naključno izberejo iz spodnjega 20-odstotnega kvantila. Ta prag je bil določen empirično kot kompromis med nizko stopnjo podobnosti in stabilnostjo postopka zlivanja, saj lahko pri ekstremno nepodobnih ključih pride do degeneriranih ali nizkokakovostnih rezultatov. V našem primeru je ta kvantil ustrezal približno 800 kandidatom v galeriji.

Zlivanje v latentnem prostoru. Izvorna slika I_o in izbrani ključ I_k se s pomočjo kodirnika e4e [20] pretvorita v latentni prostor mreže StyleGAN. Nato s postopkom StyleFusion [18] združimo njuni latentni predstavitvi v novo latentno kodo, ki jo generator StyleGAN preslika v zaščiten sliki $I_p = \psi(I_o, I_k)$. Postopek StyleFusion omogoča prostorsko lokaliziran nadzor nad sintezo obraza z uporabo vnaprej določenih latentnih komponent, ki ustrezajo posameznim obraznim regijam. Pomembno je poudariti, da postopek ne zahteva ročnega vnosa, segmentacijskih mask ali semantičnih anotacij. V

našem pristopu dosledno zlivamo identitetno relevantne regije (oči, kožo in lase) iz ključa v izvorno sliko, pri čemer so izbira regij in parametri zlivanja fiksni za vse eksperimente in določeni na podlagi empiričnega preverjanja. Takšna eksplicitna zasnova omogoča nadzorovano zatiranje identitetnih in spolnih značilnosti. Ker je postopek determinističen in sledljiv, je mogoče za vsako zaščiten sliki analizirati izvorno sliko, izbrani ključ in uporabljene parametre zlivanja.

Lokalno preverjanje. Uporabnik nato lokalno (tj. na svoji napravi) preveri, ali je zlit obraz ustrezno zaščiten:

- z modelom za razpoznavanje obrazov preveri, ali je podobnost med I_p in I_o dovolj pod pragom τ_{id} ; pri modelih ArcFace in SwinFace je ta prag običajno okoli 0,3,
- z modelom za razpoznavanje spola (npr. FairFace) preveri, ali je napoved spola negotova, tj. $P(\text{moški} | I_p) \approx 0,5$.

Če kateri od pogojev ni izpolnjen, se izbere drug ključ in postopek se ponovi. Lokalno preverjanje služi kot praktičen mehanizem za oceno ustreznosti zaščite pred pošiljanjem podatkov na strežnik in ne pomeni formalne varnostne razsodbe ali nadomestila za strežniško avtentikacijo. V eksperimentih dodatno obravnavamo tudi scenarij, v katerem strežnik uporablja drug model za razpoznavanje obrazov.

Registracija. Ko ustrezna slika I_p uspešno izpolni oba pogoja, se zlit obraz (ali njegov biometrični vektor) pošlje na strežnik kot registracijski zapis.

Avtentikacija (preverjanje obrazov 1:1). Ob vsaki poznejši prijavi se zajame nova uporabnikova slika, ki se ponovno zlije z istim ključem I_k . Uporaba istega ključa za posameznega uporabnika ohranja učinkovitost avtentikacije v zaščitenem prostoru. Zaradi deterministične narave postopka zlivanja bo nastala slika podobna tisti iz postopka registracije, kar strežniku omogoča avtentikacijo brez razkritja spola, pri čemer biometrična identiteta ni podobna identiteti na izvorni sliki.

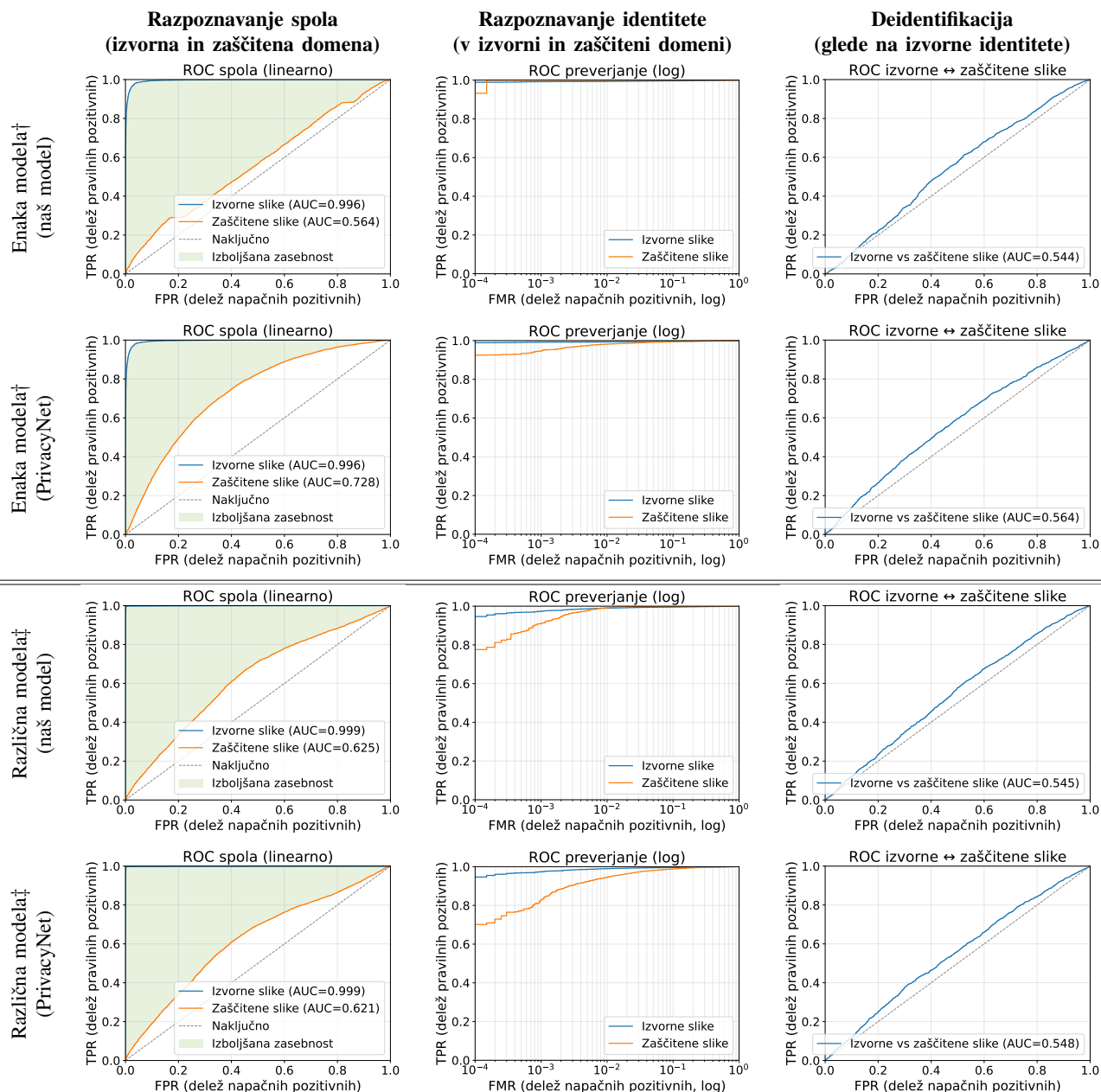
4.1 Opredelitev interpretabilnosti

V tem delu interpretabilnost ne označuje razločljivosti notranjega delovanja globokega modela, temveč razumljivost in sledljivost same transformacije obraza. Interpretabilen pristop v tem kontekstu omogoča, da je za nastalo zaščiten sliki jasno, kateri vhodni podatki in kateri deli obraza so prispevali k posameznim spremembam. Predlagani pristop takšno interpretabilnost dosega z uporabo determinističnega zlivanja z vnaprej izbranim referenčnim obrazom ter z omejitvijo transformacije na vnaprej določene obrazne regije, kar omogoča neposredno povezavo med izvorom in učinkom sprememb.

5 EKSPERIMENTI

5.1 Podatkovna zbirka

Za eksperimentalno evalvacijo smo uporabili široko uveljavljeno podatkovno zbirko obrazov CelebA [21],



† Uporabnik na svoji napravi preveri uspešnost zaščite z istima modeloma (ArcFace in FairFace), kot ju uporablja strežnik.

‡ Uporabnik zaščito preveri z modeloma ArcFace in FairFace, strežnik pa za avtentikacijo in določanje spola uporablja SwinFace.

Tabela 1 Rezultati preverjanja učinkovitosti zaščite zasebnosti. Prvi dve vrstici prikazujeta scenarij, kjer strežnik uporablja enaka modela kot uporabnik (ArcFace za identiteto in FairFace za spol), spodnji dve vrstici pa scenarij z različnima modeloma, kjer strežnik identiteto in spol napoveduje s SwinFace. Prvi stolpec poroča o količini ohranjene informacije o spolu (AUC), drugi o uspešnosti avtentikacije pred zaščito in po njej (FNMR@FMR v logaritmični skali), tretji pa o stopnji deidentifikacije glede na izvorno identiteto.

ki vključuje označene identitete in spol za vsakega posameznika. Zbirko smo najprej predprocesirali: slike smo poravnali glede na obrazne značilnice z metodo MTCNN [22] ter jih prilagodili na standardno velikost 224×224 slikovnih elementov. Nato smo podatke razdelili po spolu in uravnotežili število moških in ženskih identitet z odstranitvijo presežnih ženskih identitet, saj je teh v zbirki več. Končna podatkovna zbirka je tako vsebovala 4 520 moških in 4 520 ženskih identitet.

5.2 Razpoznavanje obrazov

Za ovrednotenje ohranjanja identitete smo pri preverjanju generalizacije v dveh eksperimentih uporabili dva sodobna modela za razpoznavanje obrazov: ArcFace [23] in SwinFace [24]. Predpostavljamo, da ima uporabnik na svoji napravi na voljo model ArcFace za izvedbo zaščite zasebnosti, preden biometrične podatke pošlje na strežnik. Model SwinFace uporabimo kot dodatni model za simulacijo scenarija, v katerem strežnik

uporablja drugačen model, ki ga uporabnik ni predvidel.

Oba modela (vsakega označimo z φ) iz izvornih in transformiranih slik izluščita 512-dimenzionalni vektor značilk $\varphi(I)$, ki ga uporabimo za izračun kosinusne podobnosti med pari obrazov. Izvedemo dva ločena eksperimenta, kot je opisano spodaj.

Eksperiment 1: Avtentikacija. V prvem eksperimentu analiziramo porazdelitev podobnosti med pari slik v zaščiteni oziroma transformirani domeni. Obravnavamo dve vrsti parov: (i) pare slik iste identitete in (ii) pare slik različnih identitet. Za vsak par slik I_p^i in I_p^j izračunamo podobnost $F(I_p^i, I_p^j)$. Namen eksperimenta je preveriti, ali transformirana domena ohranja visoko stopnjo znotrajrazredne podobnosti in hkrati zagotavlja nizko stopnjo medrazredne podobnosti. Na podlagi izračunanih podobnosti naredimo analizo ROC-krivulje.

Eksperiment 2: Deidentifikacija. V drugem eksperimentu ovrednotimo, kako učinkovito predlagana metoda SB-PET ψ odstranjuje informacije o identiteti, pri čemer učinkovitost merimo s stopnjo podobnosti med izvorno in zaščiteni sliko v prostoru identitetnih značilk. Za vsak obraz I_o generiramo njegovo transformirano različico I_p in nato izračunamo podobnost $F(I_o, I_p)$. Nizke vrednosti te podobnosti (tj. bistveno pod pragom za ujemajoče se pare) kažejo na uspešno deidentifikacijo. Za kvantitativno ovrednotenje uporabimo standardni biometrični evalvacijski protokol ter poročamo o stopnjah napačne zavrnitve (FNMR) pri različnih vrednostih stopnje napačnega ujemanja (FMR). Rezultate prikažemo s krivuljami FNMR@FMR. Naj bo $I_{o,u}^i$ i -ta izvorna slika identitete u , pri čemer indeks u označuje identiteto.

- **Ujemajoči se pari (genuine):** $(I_{o,u}^i, \psi(I_{o,u}^i))$
- **Neujemajoči se pari (impostor):** $(I_{o,u}^i, \psi(I_{o,v}^j))$, pri $u \neq v$

Za ovrednotenje avtentikacije smo uporabili 10 213 ujemajočih in 20 048 neujemajočih se parov iz zbirke CelebA. Za ujemajoče se pare smo uporabili slike iste identitete, ki so že v prvem poskusu prestale lokalno preverjanje zaščite.

5.3 Razpoznavanje spola

Za merjenje količine preostale informacije o spolu smo uporabili dva sodobna klasifikatorja spola: FairFace [25] in vejo modela SwinFace za klasifikacijo spola [24]. Podobno kot pri eksperimentih razpoznavanja identitete tudi tu izvajamo napoved spola na izvornih slikah I_o in na transformiranih slikah $I_p = \psi(I_o)$. Predpostavljamo, da ima uporabnik na svoji napravi na voljo model FairFace, ki se uporablja pri generiranju zasebnostno izboljšanih slik. Na strežniški strani ovrednotimo dva scenarija:

- 1) **Enak model:** strežnik uporablja isti model kot uporabnik (FairFace),
- 2) **Različna modela:** strežnik uporablja drug model (SwinFace), da preverimo, ali se učinek zamegli-

Tabela 2 Primerjava lastnosti, povezanih z interpretabilnostjo transformacije, med PrivacyNet in predlaganim pristopom.

Lastnost	PrivacyNet	Predlagani pristop
Drugi obraz kot eksplicitni referenčni vhod	Ne	Da (I_k)
Eksplicitno določene obrazne regije za zlivanje	Ne	Da (StyleFusion regije)
Neposredna sledljivost sprememb na vhodne vire	Ne	Da (znan I_k + regije)

tve spola ohrani tudi pri nepoznanih klasifikatorjih spola.

S tem eksperimentom preverimo robustnost zaščite zasebnosti spola v različnih pogojih razpoznavanja.

5.4 Primerjava s sodobno metodo

Predlagani postopek primerjamo z metodo PrivacyNet, pri čemer obe metodi ovrednotimo na istem naboru vhodnih slik iz zbirke CelebA. Obe metodi ocenjujemo z istim evalvacijskim protokolom. Ker izvirna implementacija metode PrivacyNet ni javno dostopna, smo metodo ponovno implementirali na podlagi opisa v izvirnem članku.

Poleg kvantitativne primerjave rezultatov je smiselno primerjati tudi strukturne lastnosti obeh pristopov, ki vplivajo na interpretabilnost transformacije. V tabeli 2 primerjamo interpretabilnost pristopa PrivacyNet in predlaganega pristopa, pri čemer ocenjujemo, v kolikšni meri je mogoče izvor in naravo sprememb na zaščiteneh slikah neposredno povezati z vhodnimi podatki in uporabljenim postopkom transformacije.

6 REZULTATI

6.1 Zaščita zasebnosti in ohranjanje identitete

V tabeli 1 so predstavljeni rezultati naših eksperimentov. V **prvih dveh vrsticah** je prikazan scenarij, v katerem strežnik uporablja *enake modele* kot odjemalec, in sicer ArcFace za preverjanje identitete ter FairFace za napovedovanje spola. V **tretji in četrti vrstici** je prikazan scenarij z *različnimi modeli*, kjer uporabnik lokalno izvaja preverjanje zaščite z modeloma ArcFace in FairFace, medtem ko strežnik za razpoznavanje identitete in spola uporablja model SwinFace.

V **prvem stolpcu** so prikazani grafi, ki ponazarjajo količino informacije o spolu, ki jo predlagani pristop uspešno zabriše. Model FairFace na izvornih slikah zelo natančno napoveduje spol (AUC = 0,996), model SwinFace pa doseže še nekoliko boljši rezultat (AUC = 0,999). Ko strežnik uporablja enake modele kot odjemalec, FairFace na zasebnostno zaščiteneh slikah doseže zgolj AUC = 0,564. V scenariju z različnimi modeli, kjer strežnik uporablja SwinFace, dobimo AUC = 0,625. To kaže, da zaščita zasebnosti, izvedena z modelom FairFace, dobro generalizira tudi na drug sodoben klasifikator spola, kar je spodbudno, saj pomeni močnejšo zaščito

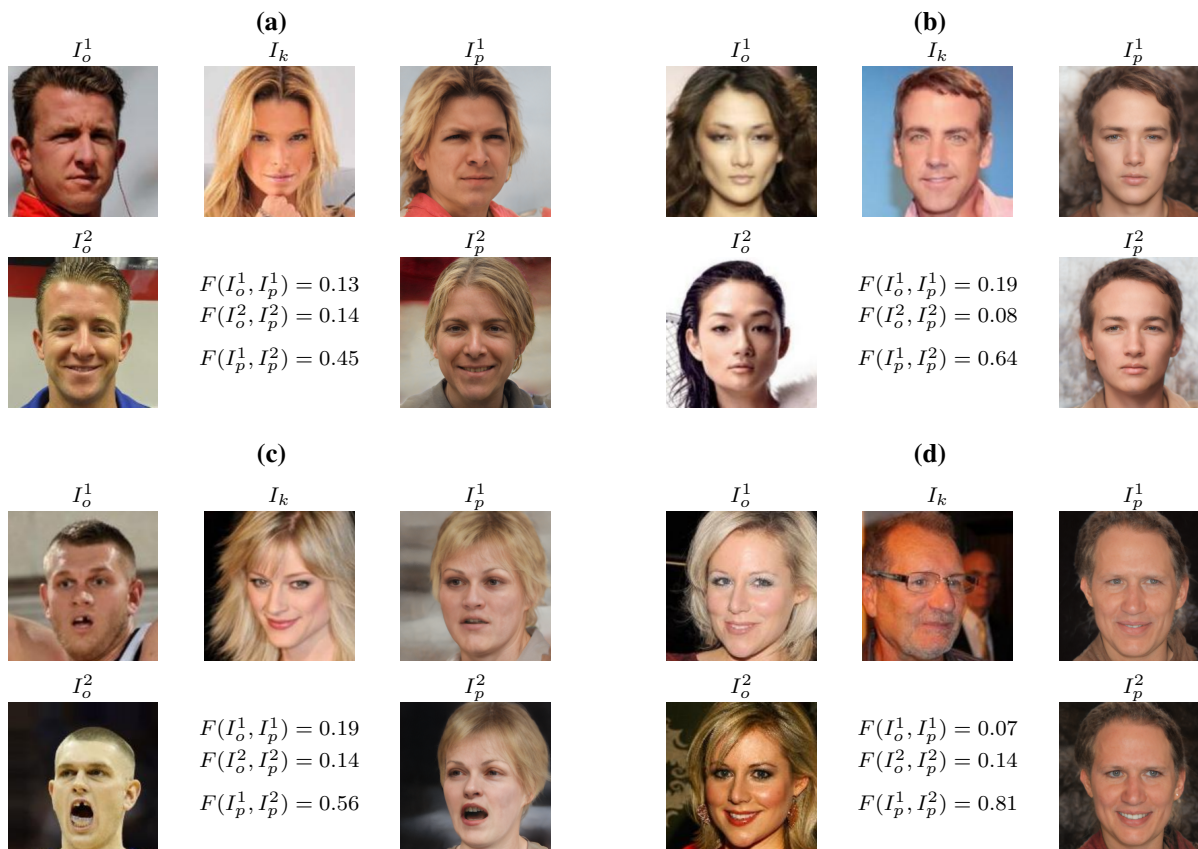


Tabela 3 Primeri uspešnega zlivanja izvornih identitet I_o s ključem I_k , pri katerem dobimo zaščitene slike I_p . Primera (a) in (c) predstavljata zlivanje moške osebe z izbrano žensko identiteto, primera (b) in (d) pa zlivanje ženske osebe z moško identiteto. Prikazani so tudi pripadajoči podobnostni rezultati, kjer nizke vrednosti $F(I_o, I_p)$ kažejo na neuspešno preverjanje identitete med izvirnimi in zaščitnimi slikami (tj. uspešno deidentifikacijo glede na izvirno identiteto), medtem ko visoke vrednosti $F(I_p^1, I_p^2)$ v prostoru zaščiteneh slik omogočajo uspešno avtentikacijo.

zasebnosti. Podoben trend opazimo tudi pri rezultatih metode PrivacyNet (vrstici 2 in 4).

Drugi stolpec prikazuje rezultate preverjanja identitete pred zaščito in po zaščiti zasebnosti. Pri obeh primerjanih metodah opazimo, da je preverjanje identitete v zaščiteni domeni skoraj enako uspešno kot v izvorni domeni, z le minimalnimi odstopanji. Nekoliko slabši rezultati se pojavijo v scenariju z drugačnim razpoznavalnim modelom (SwinFace), kar nakazuje, da lahko zaščita zasebnosti pri modelih z drugačno latentno strukturo nekoliko vpliva na uspešnost preverjanja identitete.

Eksperiment, ki meri *deidentifikacijo glede na izvirne slike*, je prikazan v **tretjem stolpcu**. V obeh scenarijih dosežemo skoraj naključno uspešnost, kar pomeni, da je zaščitene slike težko povezati z njihovimi izvornimi slikami. To potrjuje učinkovitost obeh primerjanih metod pri deidentifikaciji glede na izvirno domeno.

6.2 Kvalitativni rezultati

V tabeli 2 so prikazani konkretni primeri uspešne zaščite zasebnosti spola, kjer zasebnostno zaščitene slike

dosegajo nizke vrednosti podobnosti z ustreznimi izvornimi slikami (tj. $F(I_o^1, I_p^1)$ in $F(I_o^2, I_p^2)$). Hkrati poročamo tudi o vrednostih podobnosti v zaščiteni domeni ($F(I_p^1, I_p^2)$), ki ostanejo nad tipičnim pragom za preverjanje identitete $\tau_{id} = 0,3$ za oba modela, ArcFace in SwinFace.

7 ZAKLJUČEK

V tem delu smo predstavili metodo za zaščito zasebnosti spola pri daljinski avtentikaciji na osnovi obrazov, kjer je zaščitene slike težko povezati z izvornimi slikami obrazov. Predlagani pristop temelji na determinističnem zlivanju obrazov, pri čemer se izvorni biometrični podatki nikoli ne prenašajo na strežnik, temveč se posredujejo zgolj njihove predstavitve z izboljšano zasebnostjo. Na podatkovni zbirki CelebA smo demonstrirali uporabnost predlagane metode in pokazali, da je mogoče metode SB-PET uspešno združiti z deidentifikacijo, kljub temu da ju obstoječa literatura pogosto obravnava kot metodološko nasprotujoča si pristopa. Predlagana metoda dosega rezultate, ki so primerljive z metodo PrivacyNet, vendar je zaradi sledljivega postopka zlivanja slik bolj

transparentna.

Metoda ima tudi določene omejitve. StyleGAN je zasnovan za delo s poravnanimi in visokokakovostnimi obraznimi slikami, zato je učinkovitost postopka neposredno odvisna od kakovosti vhodnih podatkov. Poleg tega mora biti izbrani ključ dovolj različen od izvirne identitete; v nasprotnem primeru se lahko raven zaščite zasebnosti zmanjša, hkrati pa obstaja možnost delnega uhajanja atributov iz izvirne slike.

Pomembno je poudariti, da predlagani pristop ni ekvivalent naključnemu pošiljanju vedno iste slike ali sintetičnega obraza. V takšnem primeru bi bila avtentikacija povsem ločena od uporabnika in ne bi bila vezana na njegove biometrične značilnosti. Predlagana metoda ohranja to vez, saj je zaščiten predstavitev deterministično izpeljana iz uporabnikovega obraza, hkrati pa omogoča stabilno avtentikacijo ob naravnih variacijah zajema.

Uporaba istega referenčnega obraza pri različnih uporabnikih sama po sebi ne pomeni varnostne pomanjkljivosti v obravnavanem avtentikacijskem modelu. Analogno imata lahko v klasičnih sistemih dva uporabnika enako geslo, pri čemer se avtentikacija vedno izvaja v kontekstu konkretnega uporabniškega računa. Pri predlaganem pristopu je referenčni obraz del avtentikacijskega mehanizma, in ne identitete same, zato njegova ponovna uporaba pri različnih uporabnikih ne vodi do zamenjave identitet ali kršitve pravilnega delovanja sistema.

Čeprav se v tem delu osredotočamo na spol kot mehko biometrično značilnost, ki je zlasti primerna za raziskovanje v kontekstu zasebnosti, ostaja odprto vprašanje, v kolikšni meri je mogoče predlagani pristop smiselno razširiti tudi na druge attribute, kot sta starost in etnična pripadnost. Takšna razširitev bi zahtevala dodatno analizo vpliva posameznih obraznih regij na obravnavane attribute in presega obseg tega dela.

V nadaljnjem raziskovalnem delu bi bilo smiselno razviti naprednejši mehanizem za izbiro ključa, ki bi sistematično zagotavljal optimalno razliko med ključem in izvirno identiteto ob hkratnem ohranjanju atributov ciljne spolne domene. Pomembno odprto raziskovalno vprašanje je tudi razširitev predlaganega pristopa v celovit sistem preklicljive biometrije (angl. *cancellable biometrics*), kar vključuje dodatne izzive, kot so zagotavljanje stabilnosti transformacij, upravljanje več ključev ter oblikovanje formalnih varnostnih jamstev. Ti vidiki sicer presegajo obseg tega dela, a so obetavna smer za nadaljnji razvoj metod za zaščito zasebnosti v biometričnih sistemih.

ZAHVALA

Raziskavo sta financirala Javna agencija za znanstveno-raziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS) v okviru raziskovalnega programa P2-0250 "Metrologija in biometrični sistemi" ter raziskovalnega projekta ARIS J2-50069 "Interpretacija mehanizmov za razločljivo biometrično umetno inteligenco (MIXBAI)".

LITERATURA

- [1] K. Fatima, M. Schuckers, G. Cruz-Ortiz, D. Hou, S. Purnapatra, T. Andrews, A. Neupane, B. Marshall, and S. Schuckers, "A large-scale study of performance and equity of commercial remote identity verification technologies across demographics," in *IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2024, pp. 1–8.
- [2] M. A. Hamzah, "Advancing personal identity verification by integrating facial recognition through deep learning algorithms," *International Journal of Information Technology*, vol. 16, no. 7, pp. 4381–4386, 2024.
- [3] P. Terhöst, D. Fährmann, N. Damer, F. Kirchbuchner, and A. Kuijper, "On soft-biometric information stored in biometric face embeddings," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 4, pp. 519–534, 2021.
- [4] P. Melzi, H. O. Shahreza, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, J. Fierrez, S. Marcel, and C. Busch, "Multi-ive: Privacy enhancement of multiple soft-biometrics in face embeddings," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 323–331.
- [5] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, and C. Busch, "An overview of privacy-enhancing technologies in biometric recognition," *ACM Computing Surveys*, vol. 56, no. 12, pp. 1–28, 2024.
- [6] V. Mirjalili, S. Raschka, and A. Ross, "Privacynet: Semi-adversarial networks for multi-attribute face privacy," *IEEE Transactions on Image Processing*, vol. 29, pp. 9400–9412, 2020.
- [7] E. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying facial images," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 232–243, 2005.
- [8] S. Ribarić and N. Pavešić, "An overview of face de-identification in still images and videos," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2015, pp. 1–6.
- [9] A. Jourabloo, X. Yin, and X. Liu, "Attribute preserved face de-identification," in *IEEE International Conference on Biometrics (ICB)*, 2015, pp. 278–285.
- [10] A. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.
- [11] National Institute of Standards and Technology, "Digital identity guidelines," National Institute of Standards and Technology, Tech. Rep. NIST Special Publication 800-63, 2017, including updates through 2023.
- [12] V. Mirjalili, S. Raschka, and A. Ross, "Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers," *IEEE Access*, vol. 7, pp. 99 735–99 745, 2019.
- [13] P. Terhöst, N. Damer, F. Kirchbuchner, and A. Kuijper, "Suppressing gender and age in face templates using incremental variable elimination," in *IEEE International Conference on Biometrics (ICB)*, 2019, pp. 1–8.
- [14] V. Mirjalili, S. Raschka, and A. Ross, "Gender privacy: An ensemble of semi adversarial networks for confounding arbitrary gender classifiers," *CoRR*, vol. abs/1807.11936, 2018. [Online]. Available: <http://arxiv.org/abs/1807.11936>
- [15] V. Mirjalili and A. Ross, "Soft biometric privacy: Retaining biometric utility of face images while perturbing gender," in *IEEE International Joint Conference on Biometrics*, 10 2017, pp. 564–573.
- [16] A. Rozsa, M. Günther, E. M. Rudd, and T. E. Boulton, "Facial attributes: accuracy and adversarial robustness," *Pattern Recognition Letters*, vol. 124, pp. 100–108, 2019.
- [17] V. Mirjalili, S. Raschka, A. Namboodiri, and A. Ross, "Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images," in *2018 International Conference on Biometrics (ICB)*, 2018, pp. 82–89.
- [18] O. Kafri, O. Patashnik, Y. Alaluf, and D. Cohen-Or, "Stylefusion: disentangling spatial segments in stylegan-generated images," *ACM Transactions on Graphics*, vol. 41, no. 5, pp. 1–15, 2022.
- [19] P. Rot, J. Krizaj, P. Peer, and V. Štruc, "Enhancing gender privacy with photo-realistic fusion of disentangled spatial segments," in *ICASSP 2024 - 2024 IEEE International Conference on*

- Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 4615–4619.
- [20] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, “Designing an encoder for stylegan image manipulation,” *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–14, 2021.
 - [21] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.
 - [22] J. Xiang and G. Zhu, “Joint face detection and facial expression recognition with mtcnn,” in *IEEE International Conference on Information Science and Control Engineering*, 2017, pp. 424–427.
 - [23] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.
 - [24] L. Qin, M. Wang, C. Deng, K. Wang, X. Chen, J. Hu, and W. Deng, “Swinface: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2223–2234, 2023.
 - [25] K. Karkkainen and J. Joo, “Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1548–1558.

Peter Rot je leta 2024 doktoriral na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Zaposlen je kot raziskovalec v Laboratoriju za računalniški vid na Fakulteti za računalništvo in informatiko ter v Laboratoriju za strojno inteligenco na Fakulteti za elektrotehniko (Univerza v Ljubljani). Njegovi raziskovalni interesi zajemajo računalniški vid, biometrijo, zasebnost in razločljivo umetno inteligenco. Sodeluje kot recenzent pri mednarodnih znanstvenih revijah in konferencah, med drugim TIFS, ESWA, FG, WACV in IEEE Access. Je član IEEE in SDRV.