Vpliv strojnega pozabljanja identitete na pozabljanje demografskih atributov

Anastasija Manojlovska[1], **Peter Rot**[1,2], **Peter Peer**[2], **Vitomir Štruc**[1]

[1]*Fakulteta za elektrotehniko, Univerza v Ljubljani*
[2]*Fakulteta za računalništvo in informatiko, Univerza v Ljubljani*
E-pošta: [1]*{anastasija.manojlovska, peter.rot, vitomir.struc}@fe.uni-lj.si*
[2]*peter.peer@fri.uni-lj.si*

Abstract

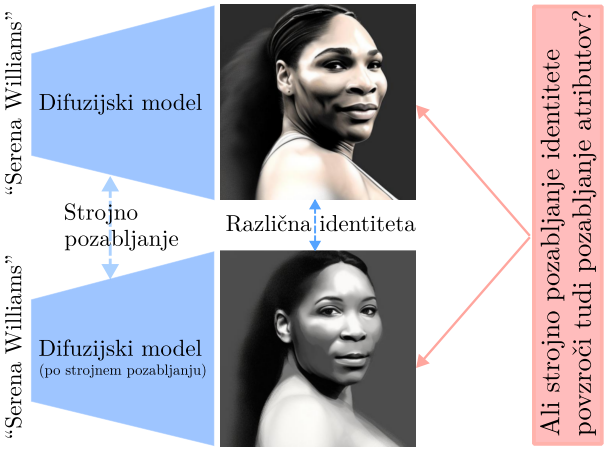
Difuzijski modeli omogočajo ustvarjanje visokokakovostnih slik na podlagi besedilnega opisa. Naučeni so na obsežnih podatkovnih zbirkah, ki vključujejo tudi slike ljudi. Zato lahko z njimi ustvarjamo slike resničnih oseb, kar odpira možnosti za zlorabo. Da bi naslovili ta problem, so raziskovalci predlagali metode strojnega pozabljanja (angl. machine unlearning*), katerih cilj je iz modela odstraniti že naučen koncept, na primer identiteto določene osebe, brez ponovnega učenja modela od začetka. V tem delu raziščemo, kako odstranitev posamezne identitete vpliva na demografske atribute, ki jih generira model, spremenjen s pomočjo strojnega pozabljanja. Zanima nas, kakšne obraze model po uporabi strojnega pozabljanja generira ob zahtevi, naj ustvari pozabljeno identiteto, ter v kolikšni meri se lastnosti obraza (npr. spol, starost, rasa), kljub pozabljanju še vedno ujemajo z izvorno identiteto. V eksperimentih uporabimo dva sodobna modela za strojno pozabljanje, Forget-Me-Not in EraseDiff, pri čemer pokažemo da model po pozabljanju identitet pogosto ne pozabi spola in rase osebe, večkrat pa pozabi starost. Primerjani metodi po pozabljanju v različni meri ohranita globalen izgled osebe v smislu kompozicije slike in barv.*

1 Uvod

Sodobni difuzijski modeli (angl. *diffusion models*) omogočajo ustvarjanje fotorealističnih slik na podlagi besednega opisa [1]. Naučeni so na obsežnih zbirkah označenih slik, pri čemer ponudniku modela in končnemu uporabniku pogosto ni znano, kateri konkretni učni podatki so bili uporabljeni. Ker takšne zbirke vključujejo tudi pare oblike $(ime\ osebe,\ slika\ osebe)$, lahko difuzijski modeli ob vnosu osebnega imena pogosto generirajo sintetične slike identitet, ki so bile prisotne v učni množici [2]. To odpira številna etična vprašanja, saj je model, namenjen vnaprej določeni rabi, zlorabiti za poustvarjanje biometričnih podatkov, ki veljajo za občutljive osebne podatke.

Raziskovalci so predlagali številne metode strojnega pozabljanja (angl. *machine unlearning*), katerih cilj je odstraniti znanje o določenem konceptu ali razredu, ne da bi bilo treba model učiti od začetka [3].

Na področju generiranja slik iz besedila so se dosedanje raziskave osredotočale predvsem na odstranjevanje konceptov, kot so objekti, umetniški slogi ali neprimerne vsebine [4]. Nekatere metode kot odstranljiv koncept



Slika 1: Difuzijski modeli lahko na podlagi opisa, ki vsebuje ime osebe, ustvarijo njeno podobo. Metode strojnega pozabljanja difuzijski model spremenijo tako, da pozabi identiteto te osebe in namesto nje ustvari drugo osebo. V našem delu raziskujemo, koliko informacij o obraznih atributih se kljub pozabljanju ohrani.

obravnavajo tudi identiteto [5]. Uspešna metoda pozabljanja identitete ne bi smela le zmanjšati generiranja ciljne identitete, temveč tudi ohraniti zmožnost modela za ustvarjanje realističnih in semantično doslednih obrazov. Potrebno se je zavedati, da je identiteta osebe bolj kompleksna kakor preprosti objekt ali slog, saj jo sestavljajo številne lastnosti, ki so razvidne iz obraza (npr. starost, spol, barva oči, oblika nosu, itd.). Zato se lahko zdi, da metoda določeno identiteto uspešno pozabi, hkrati pa (ne)namerno ohrani specifične obrazne atribute.

Obstoječe evalvacije metod pozabljanja pri difuzijskih modelih se običajno osredotočajo na vprašanje, ali ciljni koncept izgine, ter na to, ali je splošna kakovost generiranja ohranjena [6]. Ta merila so potrebna, vendar ne zajamejo v celoti stranskih učinkov pozabljanja identitete pri generiranju obrazov. Pogosto ne merijo, ali demografski atributi po pozabljanju ostanejo enaki. Naš glavni prispevek je študija stranskih učinkov pozabljanja identitete pri difuzijskem generiranju obrazov.

2 Sorodna dela

Difuzijski modeli. Difuzijski modeli generirajo slike tako, da se naučijo obrniti postopen proces dodajanja šuma [2, 7]. V sodobnih sistemih za generiranje slik je ta proces pogojen z besedilnim opisom (*angl. text prompt*). Ker so veliki modeli za generiranje slik iz be-

sedila učeni na obsežnih slikovnih zbirkah, si zapomnijo vizualne vzorce iz učnih podatkov, vključno z obrazi oseb. Ko so v besednem opisu vključena imena oseb, takšni modeli znajo zgenerirati sliko njihovih obrazov.

Strojno pozabljanje. Strojno pozabljanje (*angl. machine unlearning*) zajema metode, katerih cilj je odstraniti vpliv izbranih učnih podatkov ali naučenih konceptov na delovanje modela, ne da bi bilo treba model ponovno učiti od začetka [3, 8]. V kontekstu generativnih modelov se sorodne naloge pogosto obravnavajo kot odstranjevanje konceptov (*angl. concept erasure/removal*), slogov (*angl. style erasure*) ali objektov (*angl. object erasure*). Namen teh metod je zmanjšati zmožnost modela, da bi ustvaril izbrane vizualne vsebine (npr. umetniški slog, objekt, identiteto ali neprimerno vsebino), pri tem pa mora čim bolj ohraniti ostale generativne zmožnosti [9].

Avtorji metode Forget-Me-Not [5] so pokazali, da je s strojnim pozabljanjem mogoče odstraniti identiteto osebe iz difuzijskih modelov. Kljub temu, da so kvalitativni rezultati te metode prepričljivi, pa v literaturi še ni znano, v kolikšni meri je identiteta zares odstranjeva v smislu biometrične deidentifikacije in koliko demografskih atributo se pri tem ohrani. Za primerjavo smo v naši raziskavi vključili še metodo EraseDiff [10], ki je bila sicer razvita za splošno odstranjevanje konceptov iz difuzijskih modelov in ni posebej namenjena odstranjevanju identitet. Zanima nas torej, v kolikšni meri so splošne metode za strojno pozabljanje aplikativne na področju biometrije.

Identiteta in demografske skupine. Pri razpoznavanju obrazov se identiteta običajno opiše z biometričnimi predlogami (angl. *templates*), ki jih izračunamo z modeli za razpoznavanje obrazov [11]. Nizka/visoka kosinusna razlika med predlogama dveh obrazov pomeni, da sta si dva obraza podobna/različna [11]. Spremembe v podobnosti teh predlog zato uporabljamo za oceno, ali so generirani obrazi še vedno biometrično podobni ciljni identiteti. Obrazi poleg identitete vsebujejo tudi dodatne lastnosti, ki same po sebi niso nujno edinstveni identifikatorji, vendar vseeno vsebujejo mehke biometrične značilnsti (angl. *soft biometrics* ali demografske atribute.

3 Metoda

Difuzijski model in generiranje podatkovne zbirke. V naših eksperimentih smo uporabili prednaučen model Stable Diffusion, ki je sodoben in prepoznaven difuzijski model [1]. Difuzijski model označimo kot $\mathcal{M}_{\theta_0}$. Ker vemo, da je ta model v fazi učenja videl mnogo slik znanih ljudi, najprej sestavimo seznam imen stotih ($M = 100$) ciljnih identitet $\mathcal{Y}$. Imena smo uravnotežili demografskih po skupinah, natančneje po spolu in rasi (belci, temnopolti, Azijci in Indijci). Vsaki demografski skupini (npr. temnopolte ženske) smo dodelili 25 imen znanih oseb. Za vsako identiteto $y \in \mathcal{Y}$ nato z difuzijskim modelom ustvarimo $N = 10$ slik z opisom $p_y =$ "a photorealistic portrait face photo of $\mathbf{y}$", s čimer ustvarimo množico slik:

$$X_y^0 = \{\mathcal{M}_{\theta_0}(p_y, s_k)\}_{k=1}^N. \quad (1)$$

Pri tem s_k označuje seme pri vzorčenju, N pa število ustvarjenih slik za posamezno identiteto. Te slike služijo kot referenčna množica pred pozabljanjem in se uporabijo kot primeri ciljne identitete v postopku pozabljanja.

Strojno pozabljanje identitete. Strojno pozabljanje izvedemo ločeno za vsako od stotih identitet. Za obe metodi (Forget-Me-Not in EraseDiff) tako pridobimo po 100 različic modela Stable Diffusion, pri čemer je v vsaki izbrisana ena identiteta, drugih pa model eksplicitno ne pozablja.

Vrednotenje ohranjanja identitete. Za pridobitev značilk, ki opisujejo identiteto, sledimo standardnemu protokolu, ki se uporablja na področju razpoznavanja obrazov. Iz vsake generirane slike z detektorjem RetinaFace [12] zaznamo obraz in pridobimo pet ključnih točk, ki služijo za poravnavo. Za izluščanje biometričnih predlog uporabimo dva uveljavljna modela: ArcFace [13] ter MagFace [11]. Oba modela za vsako sliko vrneta predlogo za razpoznavanje obrazov dolžine 512, ki jo pred primerjavo normaliziramo in uporabimo za oceno podobnosti obrazov pred pozabljanjem in po njem. Za vrednotenje izvedemo dva ločena verifikacijska eksperimenta. *Pred* pozabljanjem pozitivne pare (angl. *genuine pairs*) tvorijo veljavni pari izvirnih slik iste osebe, negativne pare (angl. *impostor pairs*) pa do 50.000 naključno vzorčenih parov slik različnih oseb. *Po* pozabljanju pozitivne pare sestavljajo pari izvirnih in pozabljenih slik iste osebe, ustvarjenih z enakim semenom, negativne pare pa pari izvirnih in pozabljenih slik različnih oseb. Za oba eksperimenta izrišemo krivuljo ROC ter poročamo ploščino pod krivuljo (AUC) in stopnjo enakih napak (EER) – višji EER po pozabljanju kaže na oteženo razpoznavanje ciljne identitete.

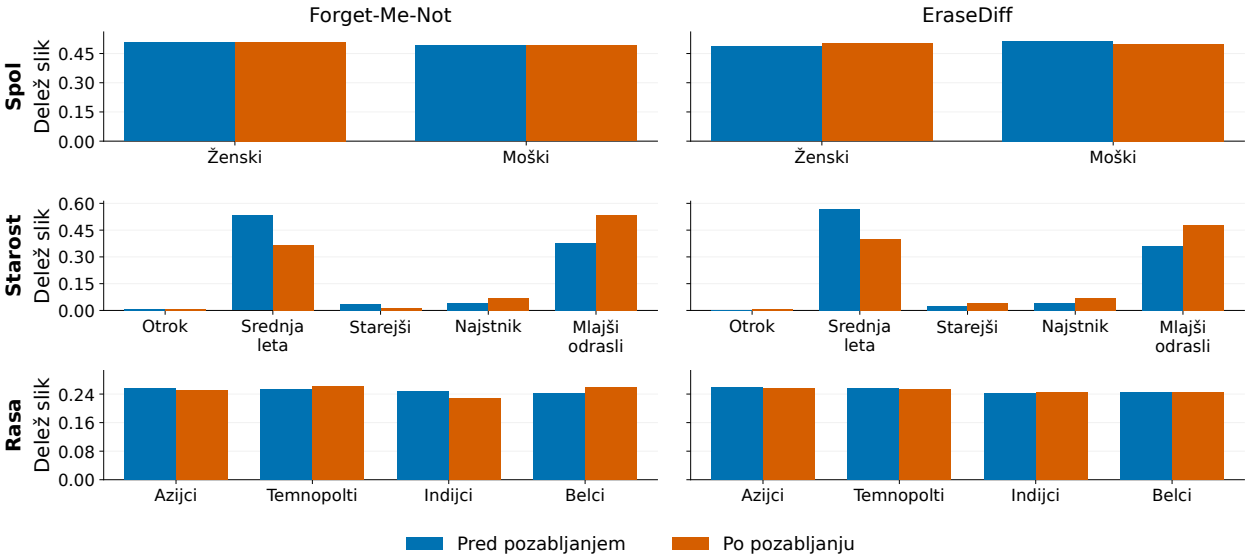
Napovedovanje demografskih atributov. Za napovedovanje spola, starostne skupine in rasne skupine na obraznih slikah, ustvarjenih z modelom Stable Diffusion pred strojnim pozabljanjem in po njem, smo uporabili model CLIP [14] brez dodatnega učenja (angl. *zero-shot*). Za vsako kategorijo značilke smo sestavili besedilni opis, na primer *"a close-up face photo of a {category} person"*. Nato smo izračunali kosinusno podobnost med vektorjem značilk slike in vektorjem značilk vsakega besedilnega opisa ter sliko uvrstili v kategorijo z najvišjo podobnostjo. Spol smo obravnavali kot binarno spremenljivko z dvema kategorijama (*moški in ženska*), starost kot kategorično spremenljivko s petimi starostnimi skupinami (*otroci, najstniki, mlajši odrasli, osebe srednjih let in starejši*), rasno skupino pa kot kategorično spremenljivko s štirimi kategorijami (*Azijci, temnopolti, Indijci in belci*). Rezultate prikazujemo s stolpičnimi grafikoni, na katerih primerjamo napovedi pred pozabljanjem in po njem.

4 Rezultati

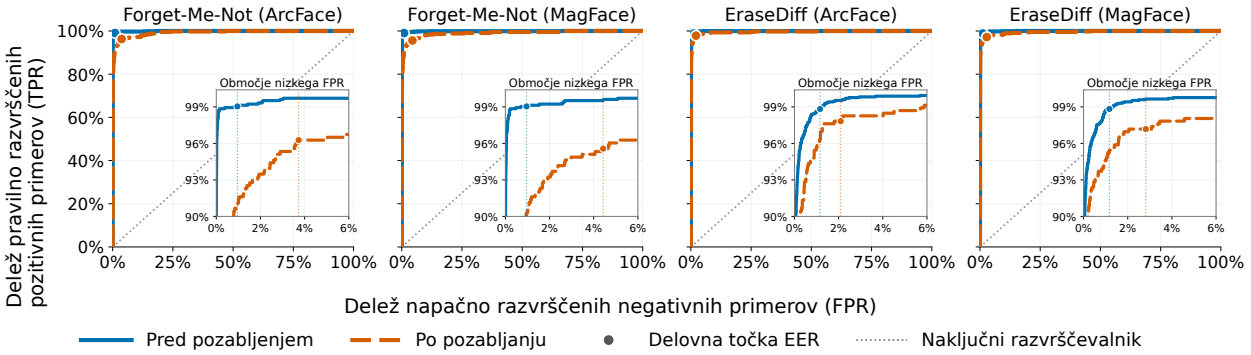
Kvalitativni rezultati. Na sliki 2 so prikazani kvalitativni primeri za vse obravnavane kombinacije demografskih skupin, ki jih model Stable Diffusion ustvari pred strojnim pozabljanjem in po njem. Rezultati kažejo, da metoda Forget-Me-Not spremeni identiteto bolj izrazito kot EraseDiff. Po pozabljanju s pomočjo obeh metodah pogosto ostanejo prepoznavne značilnosti prvotne identitete. Pri večini primerov se ohranijo tudi pričeska, poza in splošna kompozicija portreta. Vizualne spremembe se med identitetami razlikujejo. Pri nekaterih parih so bolj opazne, pri drugih pa je slika po pozabljanju zelo podobna izvirni.



Slika 2: Kvalitativna primerjava generiranih slik pred pozabljanjem identitete in po njem za metodi Forget-Me-Not in EraseDiff. V vsakem stolpcu sta prikazani sliki iste identitete, generirani z enakim pozivom in enakim semenom difuzijskega procesa, pred postopkom pozabljanja in po njem.



Slika 3: Primerjava porazdelitev napovedanih demografskih skupin pred strojnim pozabljanjem identitete in po njem. Rezultati so prikazani za metodo Forget-Me-Not (levo) in za metodo EraseDiff (desno) glede na spol, starostno skupino in raso.



Slika 4: Krivulje ROC za verifikacijo identitete pred pozabljanjem in po njem za metodi Forget-Me-Not in EraseDiff (modela ArcFace in MagFace). Pred pozabljanjem primerjamo pare izvirnih slik iste osebe s pari izvirnih slik različnih oseb, po pozabljanju pa vsako izvirno sliko z njeno pozabljeno različico, ustvarjeno z enakim semenom. Izsek prikazuje območje nizkega FPR.

Tabela 1: Verifikacija identitete pred pozabljanjem in po njem za metodi Forget-Me-Not in EraseDiff (modela ArcFace in MagFace). Poročamo AUC, EER in spremembo EER.

		Pred pozabljanjem		Po pozabljanju		
Metoda	Model	AUC	EER (%)	AUC	EER (%)	ΔEER (pp)
Forget-Me-Not	ArcFace	0.9994	0.96	0.9925	3.73	2.77
	MagFace	0.9989	0.96	0.9898	4.43	3.47
EraseDiff	ArcFace	0.9994	1.16	0.9964	2.13	0.97
	MagFace	0.9993	1.17	0.9951	2.82	1.65

Kvantitativni rezultati. Slika 3 prikazuje porazdelitve napovedanih demografskih skupin pred pozabljanjem in po njem. Porazdelitvi napovedanega spola in etnične pripadnosti se pri obeh metodah v večini primerov ohranita. Večja sprememba je opazna pri starostnih skupinah, kjer se po pozabljanju zmanjša delež obrazov, uvrščenih med odrasle srednjih let, poveča pa se delež mlajših odraslih.

Rezultati kažejo, da Forget-Me-Not biometrično identiteto spremeni bolj izrazito, vendar je pri večini slik še vedno mogoče zaznati podobnost z izvorno osebo. EraseDiff ohrani več informacije o identiteti. Slika 4 in tabela 1 prikazujeta rezultate štiri verifikacijskih eksperimentov. Prvi dve kruvilji ROC na sliki 4 prikazujeta rezultate metode Forget-Me-Not, druga dva grafa pa rezultate EraseDiff, pri čemer za vsako metodo poročamo uspešnost obeh obravnavanih modelov za razpoznavanje obrazov (ArcFace in MagFace). Modra črta prikazuje rezultat pred pozabljanjem identitete, oranžna pa po pozabljanju identitete. Kot pričakovano, je rezultat preverjanja identitete po strojnem pozabljanju vedno slabši, vendar v splošnem ne gre za veliko razliko v uspešnosti preverjanja identitete. Metode za strojno pozabljanje v našem primeru niso zagotovile tolikšnega nivoja izbrisa identitete, kot ga na primer namenske metode za deidentifikacijo [15].To je mogoče deloma pripisati dejstvu, da se večina demografskih atributov po strojnem pozabljanju ohrani, deloma pa zasnovi verifikacijskega eksperimenta, v katerem je posamezna demografska podskupina, na primer belopolti moški, zastopana z zgolj 25 različnimi identitetami.

5 Zaključek

V tem delu smo ovrednotili učinkovitost dveh metod strojnega pozabljanja, Forget-Me-Not in EraseDiff. Osredotočili smo se na odstranjevanje identitete iz generativnih difuzijskih modelov, pri čemer smo uspešnost metod ocenjevali z vidika biometrične verifikacije in ohranjanja demografskih atributov po pozabljanju. Kolikor nam je znano, takšna analiza v obstoječi literaturi še ni bila izvedena. Rezultati kažejo, da se pri "pozabljenih" identitetah v veliki meri ohranijo informacije o spolu in rasi, medtem ko so informacije o starosti ohranjene v manjšem obsegu. Ugotovili smo tudi, da obravnavani metodi v okviru verifikacijskega eksperimenta ne zagotavljata tako učinkovitega odstranjevanja identitete kot specializirane metode za deidentifikacijo obrazov, pri katerih je vrednost AUC praviloma bliže vrednosti 0,5, kar kaže na manjšo zmožnost razlikovanja med identitetami. V prihodnjem delu bomo eksperiment razširili na večje število identitet in s tem nasloviti omejitve preliminarnih rezultatov. Poleg tega bomo predlagali ustreznejše funkcije izgube za učinkovitejše pozabljanje informacij o identiteti.

Literatura

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[3] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, 2015.

[4] N. Kumari, B. Zhang, S.-Y. Wang, E. Shechtman, R. Zhang, and J.-Y. Zhu, "Ablating concepts in text-to-image diffusion models," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, p. 22634–22645, IEEE, 2023.

[5] G. Zhang, K. Wang, X. Xu, Z. Wang, and H. Shi, "Forget-me-not: Learning to forget in text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1755–1764, 2024.

[6] R. Gandikota, J. Materzyńska, J. Fiotto-Kaufman, and D. Bau, "Erasing concepts from diffusion models," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, p. 2426–2436, IEEE, 2023.

[7] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021.

[8] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine unlearning," in *2021 IEEE Symposium on Security and Privacy*, pp. 141–159, 2021.

[9] S. Lu, Z. Wang, L. Li, Y. Liu, and A. W.-K. Kong, "Mace: Mass concept erasure in diffusion models," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 6430–6440, IEEE, 2024.

[10] J. Wu, T. Le, M. Hayat, and M. Harandi, "Erasing undesirable influence in diffusion models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28263–28273, 2025.

[11] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14225–14234, 2021.

[12] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5202–5211, 2020.

[13] 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*, pp. 4690–4699, 2019.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR, 2021.

[15] B. Meden, M. Gonzalez-Hernandez, P. Peer, and V. Štruc, "Face deidentification with controllable privacy protection," *Image and Vision Computing*, vol. 134, p. 104678, 2023.