Razložljivost biometričnih predlog obraza z redkimi samokodirniki in odvedljivo parametrizacijo slik

Peter Rot[1,2], **Peter Peer**[2], **Klemen Grm**[1]

[1]*Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška cesta 25*
[2]*Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Večna pot 113*
E-mail: peter.rot@fe.uni-lj.si

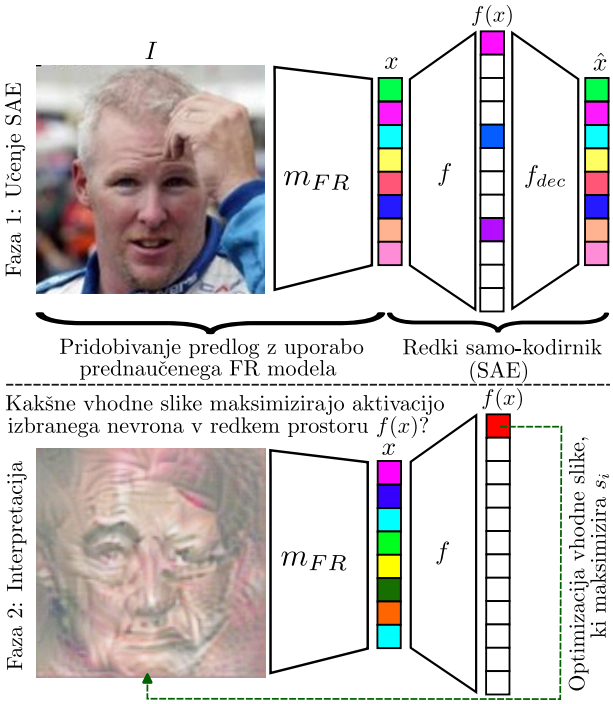
Abstract

State-of-the-art face recognition models rely on deep and complex neural net architectures that produce relatively compact template vectors, making their mechanisms of operation difficult to interpret and understand. Recently, mechanistic interpretability has emerged as a promising approach to explain large language models. In this paper, we aim to apply such approaches to explain face recognition models. Our method involves transforming face image templates into sparse representations and analyzing their components by identifying images that maximize activation. Our results demonstrate that existing mechanistic interpretability techniques generalize well to previously unconsidered tasks and architectures, and that differentiable image parametrizations can serve as a useful additional means of confirming the interpretation of sparse representations.

1 Uvod

Modeli za razpoznavanje obrazov (angl. face recognition, FR), katerih namen je iz slike obraza pridobiti vektor oz. *predlogo* za ugotavljanje podobnosti med obrazi, med drugim postajajo vse bolj pomembni za preverjanje identitete v postopkih kazenskega pregona. Z namenom, da bi povečali natančnost modelov za razpoznavanje obrazov, raziskovalci razvijajo vedno bolj kompleksna globoka nevronska omrežja, zaradi česar je vedno težje razumeti njihov proces odločanja [1]. Modeli so zato prepogosto obravnavani kot *črne škale*, kar vzbuja skrbi glede transparentnosti odločitev, še posebej v primerih, kjer je razložljivost bistvenega pomena. Lastnosti obraza so v predlogi tipično zapisana na zelo kompakten in človeku nerazumljiv način, pri čemer posamezne komponente opisujejo večje število obraznih lastnosti. Cilj razložljivega razpoznavanja obrazov (angl. eXplainable Face Recognition, XFR) je razumeti delovanje modelov in informacije, ki je zapisana v predlogah obrazov [2].

Razlage modelov za razpoznavanje obrazov se lahko lotimo iz različnih zornih kotov, v tem delu pa se zgledujemo po obetavni metodologiji interpretacije mehanizmov (angl. mechanistic interpretability), ki se je izkazala kot uspešna pri razlaganju velikih jezikovnih modelov [3, 4]. Pregled našega pristopa je ilustriran na sliki 1. Cilj je ugotoviti, kakšne vhodne slike maksimizirajo aktivacijo posameznega nevrona v redki predstavitvi obraza, ki



Slika 1: Prikaz naše metodologije. Z izbranim modelom za razpoznavanje obrazov m_{FR} iz slike obraza I najprej pridobimo predlogo $x = m_{FR}(I)$. V prvi fazi predloge uporabimo za učenje redkega samo-kodirnika s kodirnikom f in dekodirnikom f_{dec}, s katerim pridobimo redko predstavitev $f(x)$. V drugi fazi preiskujemo razložljivost komponent $f(x)_i$ z optimizacijo vhodne slike, ki maksimalno aktivira posamezno komponento v redki predstavitvi.

jo pridobimo iz originalnih obraznih predlog. Namesto, da bi se osredotočali na originalni prostor predlog x, za katerega vemo, da informacijo o obrazu kodira v močno stisnjenem in zapletenem (angl. entangled) prostoru, predloge najprej razčlenimo s pomočjo redkega samo-kodirnika (angl. sparse auto-encoder, SAE). SAE se je namreč izkazal za učinkovit način pri interpretaciji jezikovnih modelov. V prvi fazi moramo torej najprej naučiti SAE, ki originalno predlogo x preko redkega latentnega prostora $f(x)$ preslika nazaj v rekonstrukcijo originalne predloge $\hat{x}$, pri čemer $x \approx \hat{x}$. V drugi fazi vizualiziramo slike, ki povzročijo maksimalno aktivacijo posameznega nevrona v redki predstavitvi $f(x)$.

2 Sorodna dela

2.1 Interpretacija mehanizmov

Metode za *interpretacijo mehanizmov* so nastajajoča podskupina metod za razložljivo umetno inteligenco (angl. explainable artificial intelligence, XAI), katerih cilj je pojasniti odločitve modelov s na nivoju analize osnovnih algoritmov, ki jih ti implementirajo [4, 5]. Večina literature se osredotoča na interpretacijo mehanizmov velikih jezikovnih modelov [3, 6], modeli za računalniški vid pa s tovrstnimi pristopi še niso bili veliko raziskovani [7].

2.2 Redki samo-kodirniki in razložljivost

Redki samo-kodirniki so se izkazali za uporabne pri interpretaciji značilk v velikih jezikovnih modelih [8]. V delu [3] avtorji naredijo dekompozicijo značilk iz velikega jezikovnega modela tako, da jih z uporabo redkega samo-kodirnika transformirajo v redko predstavitev višje dimenzije. S tem ugotavljajo, kateri odlomki iz obširnega tekstovnega korpusa aktivirajo posamezne komponente v redki predstavitvi. V tem delu tovrsten način priredimo za reševanje naloge v računalniškem vidu, natančneje za razpoznavanje obrazov.

3 Metodologija

Visoko-nivojski pregled. Pregled naše metode je predstavljen na sliki 1. Z uporabo v naprej naučenega modela SwinFace [9], ki je eden izmed najbolj sodobnih model za razpoznavanje obrazov in ga v tem delu označujemo kot m_{FR}, iz slik obraza I najprej pridobimo predloge $x = m_{FR}(I)$. Nato naučimo redek samo-kodirnik, SAE, ki naredi dekompozicijo originalnih predlog v razširjenem prostoru $f(x)$. Nato analiziramo, katere slike maksimalno aktivirajo posamezno komponento v redki predstavitvi, tj. $f(x)_i = f(m_{FR}(I))_i$, kar prevedemo na naslednji optimizacijski problem:

$$\begin{aligned} I_{opt} &= \arg\max_{I} f\left(m_{FR}\left(I\right)\right)_i, \\ s.t.\, I &= DIP(z), \end{aligned} \tag{1}$$

kjer z uporabo gradientnega spusta optimiziramo vhodno sliko I. Parametri kodirnika f in modela za razpoznavanje obraza m_{FR} so zamrznjeni, $DIP(z)$ pa je odvedljiva parametrizacija slike Deep Image Prior [10], ki definira omejitev optimizacije.

Odvedljiva parametrizacija slike [11]. Pri optimizaciji vhodne slike s ciljem maksimalne aktivacije določene komponente v nevronskem omrežju se izkaže, da parametrizacija slike (npr. optimizacija kompleksnega spektra namesto optimizacije RGB vrednosti) lahko vodi do bolj razložljivih vhodnih slik. V tem delu za parametrizacijo predlagamo uporabo metode Deep Image Prior [10]. Pri tem je vhodna slika generirana kot $I = m_{UNET}(z)$, kjer je m_{UNET} konvolucijsko nevronsko omrežje U-Net [12] in z tenzor naključnega šuma. Optimizacija enačbe (1) je torej izvedena preko parametrov omrežja U-Net.

Samo-kodirnik z mehanizmom zapornic (Gated sparse auto-encoder, G-SAE). Cilj redkih samo-kodirnikov je narediti dekompozicijo predlog za katere je značilno, da informacijo kodirajo v zelo stisnjeni obliki, pri čemer si v idealnem primeru želimo dobiti redke značilke, ki kodirajo informacijo o obrazu v razpleteni (angl. disentangled) obliki, ki je bolj razložljiva. V našem delu za pridobitev redkih predstavitev implementiramo samo-kodirnik z mehanizmom zapornic (angl. gated sparse auto-encoder, G-SAE), za katerega je bilo pokazano, da izboljša učenje s slovarji [6]. Zapišemo ga z enačbo:

$$\begin{aligned} \tilde{\mathbf{f}}(\mathbf{x}) := &\underbrace{\mathbb{1}\left[(\mathbf{W}_{\text{gate}}(\mathbf{x} - \mathbf{b}_{\text{dec}}) + \mathbf{b}_{\text{gate}}) > 0\right]}_{\mathbf{f}_{\text{gate}}(\mathbf{x})} \\ &\odot \underbrace{\text{ReLU}(\mathbf{W}_{\text{mag}}(\mathbf{x} - \mathbf{b}_{\text{dec}}) + \mathbf{b}_{\text{mag}})}_{\mathbf{f}_{\text{mag}}(\mathbf{x})} \end{aligned} \tag{2}$$

kjer $\mathbb{1}$ predstavlja Heavisidovo skočno funkcijo. Parametri $\mathbf{W}_{\text{gate}}, \mathbf{W}_{\text{mag}}, \mathbf{b}_{\text{dec}}$ in $\mathbf{b}_{\text{mag}}$ pa so učni parametri. Dekoder je enako kot prej, enojen linearni sloj.

G-SAE smo nastavili tako, da originalno predlogo $x \in \mathbb{R}^{512}$ razširi za faktor 128, s čimer pridobimo redko predstavitev $f(x) \in \mathbb{R}^{65,536}$. Ker skočna funkcija ni odvedljiva in ker sta uteži zapornice in magnituda uteži enaki, po treningu G-SAE v praksi optimiziramo I tako, da maksimizira $\mathbf{W}_{\text{gate}}(\mathbf{x} - \mathbf{b}_{\text{dec}})$.

4 Postavitev eksperimenta

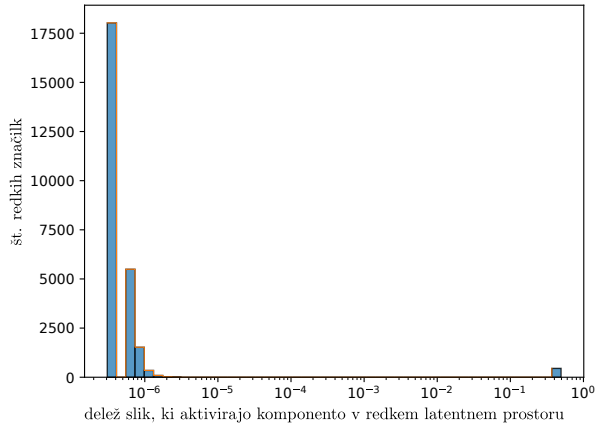
Model za razpoznavanje obrazov in predloge. Sodobni modeli za razpoznavanje obrazov so tipično učeni na v naprej poravnanih slikah obrazov in zato ne delujejo dobro, če slika ni ustrezno poravnana. Za naše eksperimente smo vzeli slike iz podatkovne zbirke VGGFace2 [13], jih poravnali z uporabo metode MTCNN [14] in iz njih z uporabo SwinFace [9] pridobili biometrične predloge.

Optimizacija G-SAE. Za učenje samo-kodirnika uporabimo funkcijo izgube, ki je bila predlagana v delu [6] s faktorjem redkosti (angl. sparsity factor) $\lambda = 10^{-7}$. Model smo učili z optimizatorjem AdamW [15] z začetno stopnjo učenja 5×10^{-4} in z zmanjševanjem uteži 10^{-2}. Velikost paketov smo nastavili na 1024 in izvedli 100 epoh, med katerimi se stopnja učenja zmanjšuje proti 10^{-6} z uporabo kosinusnega razporeda (angl. cosine schedule).

Optimizacija vhodne slike. Za optimizacijo vhodne slike, ki maksimizira aktivacijo izbranega nevrona v redkem latentnem prostoru smo uporabili AdamW optimizator s stopnjo učenja 10^{-4} in optimizacijo izvedli v 4096 korakih. Eksperimentalno smo ugotovili, da dobimo bolj različne rezultate, če namesto hkratne optimizacije večjega števila vhodnih slik, optimiziramo eno po eno, pri čemer pred vsako optimizacijo naključno inicializaciramo parametrizacijo DIP.

5 Eksperimenti in rezultati

Analiza redkega latentnega prostora. Najprej smo prešteli, koliko slik obrazov iz podatkovne zbirke aktivira posamezno značilko v redkem latentnem prostoru $f(x)_i$. Rezultati so predstavljeni na sliki 2. Opazimo, da so (za razliko od rezultatov pri jezikovnih modelov [3]) značilke



Slika 2: Histogram gostote značilk. Os x je v logaritemskem merilu.

porazdeljene v dve gruči: (i) majhna gruča, pri kateri se 451 značilk (od vsega skupaj 65536) aktivira pri 40% slik iz podatkovne zbirke in (ii) velika gruča, pri kateri se značilke aktivirajo na zelo majnhem številu slik (tj. od 1 do 20 slik od vsega skupaj 3.2×10^6 slik). V nadaljnjih sekcijah analiziramo te dve gruči.

Redundanca in rekonstrukcija. Nadalje ovrednotimo redundanco v redkem latentnem prostoru. Opazimo, da se pri vsaj eni sliki iz podatkovne zbirke v redkem prostoru aktivira le 39.64% komponent, kar se zgodi kljub temu, da uporabljamo postopek ponovnega podvzorčenja nevronov iz [3]. Srednja vrednost L_0 norme od $f(x)$ je 105.2. Pri ovrednotenju rekonstrukcijske napake je MSE pri upoštevanju celotne podatkovne zbirke $\mathbb{E}_x||x-\hat{x}||_2^2 = 7.13 \times 10^{-6}$, pri čemer smo z enotnim faktorjem skaliranja normirali predloge x tako da $\mathbb{E}_x||x||_2^2 = 1$.

Razlaga kriterijske funkcije. Zanima nas, kako dobro kodirane in dekodirane predloge delujejo za izvorni namen obrazega modela, t.j. razpoznavanja obrazov. Poleg rekonstrukcijske napake zato preverimo še uspešnost verifikacijskih eksperimentov na podatkovni zbirki Labelled Faces in the Wild (LFW, [16]). Z verifikacijskimi eksperimenti smo ovrednotili kvaliteto celotnih rekonstruiranih predlog, kot tudi kvaliteto rekonstrukcij na podlagi posameznih gruč iz slike 2.

Tabela 1: Rezultati verifikacije z dekodiranimi predlogami.

Eksperiment	**Natančnost verifikacije** $(\mu \pm \sigma)$
Originalne predloge $x = m_{FR}(I)$	0.9948 ± 0.0017
Dekodirane predloge iz goste gruče $f(x)_i$	0.9128 ± 0.0110
Dekodirane predloge iz redke gruče $f(x)_i$	0.6120 ± 0.0511
Dekodirane predloge iz celotne $f(x)$	0.9655 ± 0.0062

Interpretacija posameznih redkih značilk. V tem eksperimentu analiziramo posamezne komponente v redkem latentnem prostoru $f(x)_i$ z optimizacijo vhodne slike, tako da so le-te maksimalno aktivirane, in rezultate primerjamo s konkretnimi primeri slik iz podatkovne zbirke, ki prav tako dosežejo največjo aktivacijo. V naših eksperimentih se je izkazalo, da so mnoge komponente razložljive, saj slike (iz podatkovne zbirke in optimizirane vhodne slike), ki jih aktivirajo, vsebujejo enako semantično vsebino. Izkaže se, da so komponente iz gruče komponent, ki se aktivirane pri velikemu številu slik, pogosto ustrezajo splošnim obraznim atributom. Večina komponent, ki kodirajo splošne obrazne atribute, je invariantna na rotacijo obraza (tj. pri različnih pozicijah obraza na vhodnih slikah povzročijo podobno aktivacijo). Opazili smo tudi, da so aktivacije nekaterih komponente sorazmerne s kvaliteto slike v smislu, da obstaja korelacija med kvaliteto slike in jakostjo aktivacije (tj. bolj ostre slike višje kvalitete povzročijo večjo aktivacijo), pri nekaterih drugih komponentah pa se izkaže, da so invariantne na kvaliteto, saj take korelacije nismo opazili. Preverimo tudi, če interpretacija, ki jo izpeljemo iz slik iz podatkovne zbirke, sovpada z interpretacijo, ki jo dobimo pri optimizaciji vhodne slike. V tabeli 2 so povzete lastnosti o izbranih nevronih iz prostora $f(x)$, za katere smo našli verjetno interpretacijo, vizualizacije konkretnih primerov slik za te nevrone pa so prikazani na sliki 3. Opazili smo, da v skupini značilk z visoko gostoto skoraj nobena značilnost ne ustreza posebni identiteti. Nasprotno pa vse preučevane značilnosti grozda z nizko gostoto ustrezajo določeni osebi in jih je težje razlagati.

6 Sklepi

Pokazali smo, da se metodo za izboljšanje razložljivosti, ki spada v skupino metod za interpretacijo mehanizmov in je bila razvita za razlaganje velikih jezikovnih modelov, da uporabiti za interpretacijo modelov za razpoznavanje obrazov. S pretvorbo originalnih predlog za razpoznavanje obrazov v razširjen prostor z redkimi aktivacijami, smo našli interpretabilne vhodne slike, ki maksimizirajo posamezno komponentno.

Demonstrirali smo, da redke vložitve pripadajo dvema gručama - majhni gruči gostih značilk, ki kodirajo splošne obrazne atribute, in veliki gruči redkih značilk, ki verjetno kodirajo informacijo, ki je specifična za posamezno identiteto iz učne množice.

V nekaterih primerih smo lahko potrdili podobnost med vsebino na slikah iz podatkovne zbirke, ki maksimizirajo posamezno komponento v redki vložitvi, in optimiziranih slikah, ki maksimizirajo to komponento preko optimizacije, kar omogoča neposredno interpretacijo posameznih komponent.

Zahvala

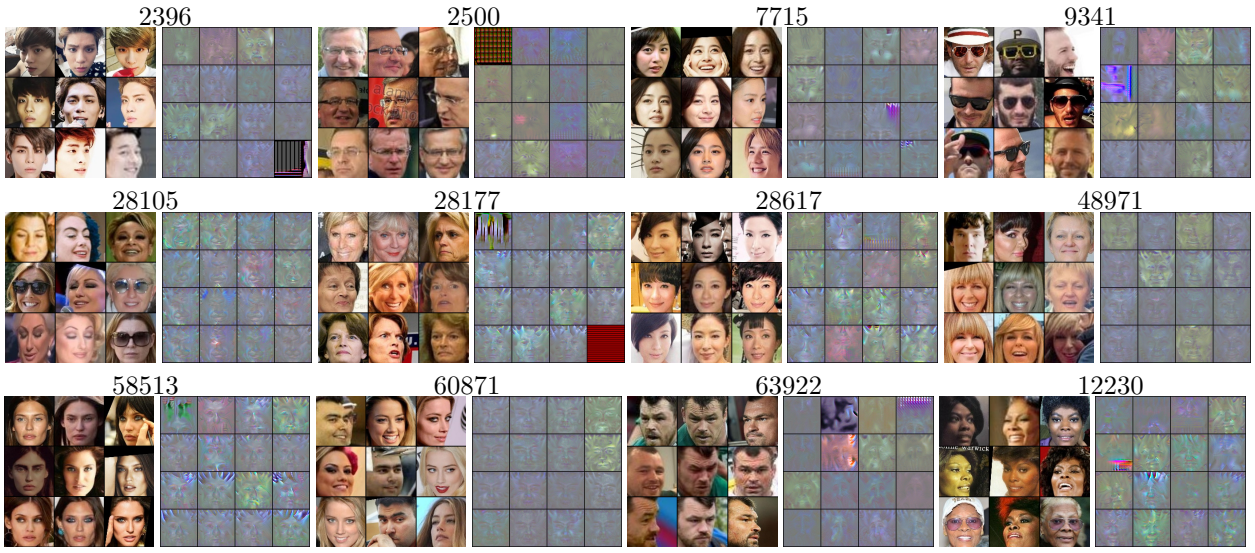
Raziskave v okviru pričujočega prispevka so bile financirane preko ARIS programa P2-0250 “Metrologija in biometrični sistemi” ter ARIS razikovalnega projekta J2-50069 “Interpretacija mehanizmov za razložljivo biometrično umetno inteligenco (MIXBAI)”.

Literatura

[1] A. Rajpal, K. Sehra, R. Bagri, and P. Sikka, “Xai-fr: explainable ai-based face recognition using deep neural networks,” *Wireless Personal Communications*, vol. 129, no. 1, pp. 663–680, 2023.

Tabela 2: Kvalitativna interpretacija značilk.

$i,\ f(x)_i$	**Komentar**	**Aktivacija sorazmerna kvaliteti**	**DIP razložljivost**
2396	Mlajši vzhodnoazijci	✓	✓
2500	Moški srednjih let s pravokotnimi očali	✗	✓
7715	Vzhodnoazijske ženske z daljšimi lasmi	✓	✗
9341	Moški s prikritimi oz. slabo vidnimi očmi	✗	✗
28105	Ženske s prikritimi oz. zaprtimi očmi	✗	✓
28177	Starejše Evropejske ženske	✗	✗
28617	Vzhodnoazijske ženske s kratkimi lasmi	✓	✗
48971	"Frufru" pričeska ("Bangs")	✗	✗
58513	Bianca Balti	✓	✗
60871	Ženske ali moški z širokimi, temnimi obrvmi	✗	✓
63922	Moški z lasno mejo v obliki črke V	✗	✗
12230	Dionne Warwick, invariantno na starost	✓	✗



Slika 3: Prikaz primerov slik, ki maksimizirajo izbrano komponento v redkem latentnem prostoru, pri čemer so na levi slike iz podatkovne zbirke, na desni pa optimizirane vhodne slike.

[2] J. R. Williford, B. B. May, and J. Byrne, "Explainable face recognition," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 248–263.

[3] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell *et al.*, "Towards monosemanticity: Decomposing language models with dictionary learning," *Transformer Circuits Thread*, p. 2, 2023.

[4] L. Bereska and E. Gavves, "Mechanistic interpretability for ai safety–a review," *arXiv preprint arXiv:2404.14082*, 2024.

[5] A. Conmy, A. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso, "Towards automated circuit discovery for mechanistic interpretability," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 16 318–16 352.

[6] S. Rajamanoharan, A. Conmy, L. Smith, T. Lieberum, V. Varma, J. Kramár, R. Shah, and N. Nanda, "Improving dictionary learning with gated sparse autoencoders," *arXiv preprint arXiv:2404.16014*, 2024.

[7] V. Palit, R. Pandey, A. Arora, and P. P. Liang, "Towards vision-language mechanistic interpretability: A causal tracing tool for blip," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2023, pp. 2856–2861.

[8] H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey, "Sparse autoencoders find highly interpretable features in language models," *arXiv preprint arXiv:2309.08600*, 2023.

[9] 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*, 2023.

[10] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9446–9454.

[11] A. Mordvintsev, N. Pezzotti, L. Schubert, and C. Olah, "Differentiable image parameterizations," *Distill*, 2018, https://distill.pub/2018/differentiable-parameterizations.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[13] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[14] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[15] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[16] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.