SEGMENTACIJA TELESA Z UPORABO VEČCILJNEGA UČENJA

JULIJAN JUG,[1] AJDA LAMPE,[1] PETER PEER[1] IN VITOMIR ŠTRUC[2]

[1] Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Ljubljana, Slovenija.
E-pošta: julijan.jug@gmail.com, ajda.lampe@fri.uni-lj.si, peter.peer@fri.uni-lj.si
[2] Univerza v Ljubljani, Fakulteta za elktrotehniko, Ljubljana, Slovenija.
E-pošta: vitomir.struc@fe.uni-lj.si

Povzetek Segmentacija je pomemben del številnih problemov računalniškega vida, ki vključujejo človeške podobe, in je ena ključnih komponent, ki vpliva na uspešnost vseh nadaljnjih nalog. Več predhodnih del je ta problem obravnavalo z uporabo večciljnega modela, ki izkorišča korelacije med različnimi nalogami za izboljšanje uspešnosti segmentacije. Na podlagi uspešnosti takšnih rešitev v tem prispevku predstavljamo nov večciljni model za segmentacijo/razčlenjevanje ljudi, ki vključuje tri naloge, tj. (i) napoved skeletnih točk, (ii) napoved globinske predstavitve poze in (iii) segmentacijo človeškega telesa. Glavna ideja predlaganega modela Segmentacija-Skelet-Globinska predstavitev (ali na kratko iz angleščine SPD) je naučiti se boljšega modela segmentacije z izmenjavo znanja med različnimi, a med seboj povezanimi nalogami. SPD temelji na skupni hrbtenici globoke nevronske mreže, ki se razcepi na tri glave modela, specifične za nalogo, in se uči z uporabo cilja optimizacije za več nalog. Učinkovitost modela je analizirana s strogimi eksperimenti na nizih podatkov LIP in ATR ter v primerjavi z nedavnim (najsodobnejšim) večciljnim modelom segmentacije telesa. Predstavljene so tudi študije ablacije. Naši eksperimentalni rezultati kažejo, da je predlagani večciljni (segmentacijski) model zelo konkurenčen in da uvedba dodatnih nalog prispeva k večji skupni uspešnosti segmentacije.

Ključne besede: računalniški vid, segmentacija, razločanje delov telesa, večciljno učenje, mreža ResNet-101

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.feri.7.2022.4
ISBN 978-961-286-575-7

BODY SEGMENTATION USING MULTI-TASK LEARNING

JULIJAN JUG,[1] AJDA LAMPE,[1] PETER PEER[1] & VITOMIR ŠTRUC[2]

[1] University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia.
E-mail: julijan.jug@gmail.com, ajda.lampe@fri.uni-lj.si, peter.peer@fri.uni-lj.si
[2] University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia.
E-mail: vitomir.struc@fe.uni-lj.si

Abstract Segmentation is an important step in many computer vision problems involving human images and one of the key components that affects the performance of all downstream tasks. Several prior works have approached this problem using a multi-task model that exploits correlations between different tasks to improve segmentation performance. Based on the success of such solutions, we present in this paper a novel multi-task model for human segmentation/parsing that involves three tasks, i.e., (i) keypoint-based skeleton estimation, (ii) dense pose prediction, and (iii) human-body segmentation. The main idea behind the proposed Segmentation-Pose-DensePose model (or SPD for short) is to learn a better segmentation model by sharing knowledge across different, yet related tasks. SPD is based on a shared deep neural network backbone that branches off into three task-specific model heads and is learned using a multi-task optimization objective. The performance of the model is analysed through rigorous experiments on the LIP and ATR datasets and in comparison to a recent (state-of-the-art) multi-task body-segmentation model. Ablation studies are also presented. Our experimental results show that the proposed multi-task (segmentation) model is highly competitive and that the introduction of additional tasks contributes towards a higher overall segmentation performance.

Keywords: computer vision, segmentation, human body parsing, multi-task learning, ResNet-101 net

https://doi.org/10.18690/um.feri.7.2022.4 DOI
978-961-286-575-7 ISBN

University of Maribor Press

1 Uvod

V zadnjih letih je bil na področju računalniškega vida dosežen velik napredek. Sodobni generativni modeli, kot so GAN (angl. generative adversarial network), so omogočili ustvarjanje fotorealističnih slik s prepričljivo vizualno kakovostjo. Veliko raziskav je osredotočeno tudi na uporabo takšnih modelov. Eden takšnih izzivov je generiranje fotorealističnih podob ljudi v želenih oblačilih ali problem virtualnega pomerjanja (Han, 2018; Fele, 2022). Takšne aplikacije imajo velik potencial za uporabo v spletnih trgovinah z oblačili in izboljšanje uporabniške izkušnje spletnega nakupovanja.

Z razvojem globokih nevronskih mrež je bil velik preskok tudi na področju semantične segmentacije (Chen, 2018; Wang, 2020). Vendar pa je na nekaterih področjih, kot je segmentacija človeškega telesa, še veliko prostora za izboljšave. Trenutno najboljši modeli segmentacije še vedno ne delujejo tako dobro, kot bi morali za uporabo v aplikacijah, kot je navidezno pomerjanje oblačil. Večino težav trenutnim modelom povzročajo slike, posnete v slabših svetlobnih pogojih, in delno prikriti pogledi na subjekt.

Slika 1: Na tem primeru vidimo, da dodatni nalogi skeleta in globinske predstavitve doprineseta koristne kontekstualne in strukturne informacije o človeškem telesu. Druga slika prikazuje segmentacijsko masko, ki jo izdela naš večciljni model, ki vsebuje naloge za segmentacijo in napoved skeleta. Tretja slika prikazuje segmentacijsko masko, ki jo je ustvaril naš večciljni model z dodatno nalogo napovedi globinske predstavitve poze. Vidimo lahko, da dodatna naloga bistveno izboljša uspešnost segmentacije.

Vir: lasten.

Pred kratkim je bilo opravljenih veliko raziskav na temo izboljšanja takšnih modelov z uporabo dodatnih informacij za spodbujanje in podporo modelov segmentacije. Z zagotavljanjem dodatnih kontekstualnih informacij se domneva, da lahko model bolje razume vsebino slike in človeško anatomijo.

Obstoječe delo je torej usmerjeno v združevanje modelov segmentacije z drugimi sorodnimi nalogami v tako imenovani večciljni arhitekturi.
Najpogosteje obstoječi modeli vključujejo napoved skeletnih točk kot podporno nalogo, na primer (Gong, 2017).

Prejšnje raziskave so tudi pokazale, da uporaba večciljnega učenja prispeva h kakovosti segmentacije ljudi. Na podlagi tega vpogleda v tem prispevku raziskujemo možnosti za razširitev tovrstnih modelov z dodatnimi nalogami, ki bi lahko dodatno pripomogle k procesu segmentacije.

Medtem ko večina obstoječega dela vključuje napoved skeletnih točk kot podporno nalogo, se naše delo osredotoča na izboljšanje kakovosti rezultatov segmentacije z uporabo dodatne naloge. V ta namen predlagamo novo arhitekturo večciljnega modela, ki poleg naloge segmentacije telesa vključuje še nalogo napovedi skeletnega položaja oziroma drže in globinske predstavitve poze.

Predlagan model smo poimenovali SPD. Črke v imenu predstavljajo naloge: S segmentacija (angl. segmentation), P skelet (angl. pose) in D globinska poza (angl. dense pose). Predlagamo večciljno arhitekturo, ki temelji na skupni nevronski mreži z uporabo treh specializiranih vej na vrhu, po eno za vsako od izbranih nalog. Namen takšnega pristopa je izboljšati nalogo segmentacije.
Predlagani model ocenjujemo na naboru podatkov LIP in ATR in poročamo o zelo spodbudnih rezultatih.

Izvajamo tudi obsežne študije ablacije, da podpremo našo hipotezo, da dodajanje nalog izboljša splošno učinkovitost modela.

Glavni prispevki našega dela so:

- Predstavljamo SPD, nov večciljni model za segmentacijo človeškega telesa, ki vključuje naloge za napoved skeleta in napovedovanje globinske predstavitve poze.

- Pokazali smo, da dodajanje dodatnih nalog izboljša zmogljivost za primarno nalogo.

2 Sorodna dela

Ena izmed bolj specializiranih aplikacijskih področij semantične segmentacije je segmentacija človeškega telesa in oblačil. Potreba po takih algoritmih za segmentacijo izhaja iz zahtev različnih sistemov računalniškega vida, povezanih z analizo človeške podobe, kot je navidezno pomerjanje oblačil (Han, 2018; Fele, 2022) ali ponovna identifikacija (Zhao, 2013).

V zadnjem času je bilo opravljenih veliko raziskav na temo segmentacije ljudi (Liang, 2015; Liang, 2016) z uporabo globokih konvolucijskih nevronskih mrež. Pomanjkljivost teh modelov je v tem, da ne upoštevajo strukture oziroma anatomije človeškega telesa, zato segmentacije pogosto vsebujejo napake, ki so s človeškega vidika nerazumne.

Veliko raziskav se je zato osredotočilo na reševanje tega problema z vključitvijo dodatnih informacij v postopek segmentacije, povezane s telesno držo in anatomijo.

Eden od načinov za uvedbo podpornih informacij v model je pristop učenja z več nalogami, kjer se model hkrati uči reševanje več nalog. Zaradi dobrih rezultatov v zadnjih letih se je večciljno učenje pogosto uporabljalo v različnih aplikacijah za obdelavo naravnega jezika in računalniškega vida (Kokkinos, 2017; Eigen, 2015; Bischke, 2019). Gong *et al.* (Gong, 2017) je na primer predlagal model, ki generira semantične segmentacijske maske in položaje skeletnih točk na podlagi generirane segmentacije. Model je optimiziran na podlagi kakovosti segmentacij in lokacij sklepov in s tem zagotovi, da se model uči semantično dosledne predstavitve človeškega telesa. Liang *et al.* (Liang, 2019) gradi na tem pristopu z uporabo skupne osnovne mreže, ki mu sledita dva manjša modula, specializirana za napoved skeletnih točk in semantične segmentacije. Moduli so zgrajeni na dvostopenjski način tako, da najprej generirajo približne in nato natančne rezultate in si pri tem delijo vmesne približne rezultate.

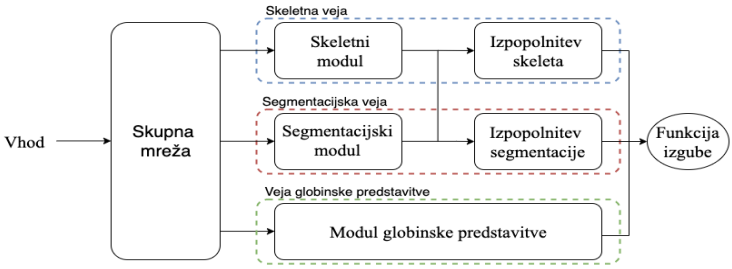
Njihov predlagan model, imenovan JPPNet, dosega impresivne rezultate in prepričljivo presega prejšnje delo. Vendar pa je še vedno prostor za izboljšave v modelovi predstavitvi telesa, saj nekatere strukturne pomanjkljivosti še vedno ostajajo.

3 Metodologija

Predlagamo večciljni model, imenovan SPD, za segmentacijo človeškega telesa, ki se uči na podlagi treh nalog: generiranje segmentacijske maske, napoved položaja skeletnih točk in napoved globinske predstavitve telesa (Guler, 2018). Model je navdihnjen z uspehi obstoječih večciljnih modelov, kot je JPPNet (Liang, 2019), za katere se je izkazalo, da zagotavljajo konkurenčne rezultate, obenem pa imajo tudi zaželene arhitekturne značilnosti.

3.1 Pregled arhitekture modela

Slika 2 prikazuje osnovno arhitekturo našega modela, ki je sestavljena iz skupne mreže za ekstrakcijo značilnosti in treh ločenih vej: *(i)* za segmentacijo človeškega telesa, *(ii)* za napoved skeletnih točk in *(iii)* za napovedovanje globinske predstavitve. Glavni cilj modela je zagotoviti učinkovito segmentacijo telesa, zato je segmentna veja obravnavana kot glavna komponenta modela, preostali dve veji pa opravljata pomožne naloge. Glavni model hrbtenice, ki je skupen vsem nalogam, je ResNet-101 (He, 2016) globoka rezidualna mreža, ki je sestavljeno iz 101 konvolucijskih plasti, razporejenih v 5 rezidualnih blokov. V modelu SPD si del te hrbtenice delijo tri veje, kar deluje kot povezava med tremi obravnavanimi nalogami.



Slika 2: Visokonivojski arhitekturni diagram predlaganega modela SPD. Skupno mrežo ResNet si delijo tri specializirane veje modela, zasnovane za segmentacijo človeškega telesa, napovedovanje skeletnih točk in napoved globinske predstavitve poze.

Vir: lasten.

Tri veje omogočajo definicijo treh ločenih učnih ciljev, ki se nato skupaj uporabijo za učenje modela. Natančneje, skupna funkcija izgube se izračuna kot utežena vsota treh izgub, specifičnih za naloge, t.j.:

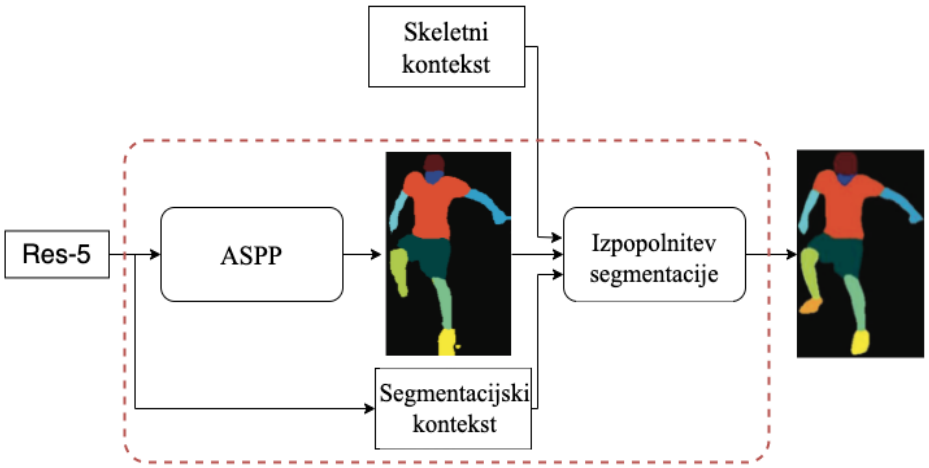
$$\mathcal{L} = \lambda_s \mathcal{L}_s + \lambda_p \mathcal{L}_p + \lambda_d \mathcal{L}_d,$$

kjer so λ_s, λ_p in λ_d uteži, ki ustrezajo funkcijam izgube segmentacije L_s, funkciji izgube skeletnih točk L_p in izgubi globinske predstavitve L_d. Empirično smo izbrali višjo utež za funkcijo izgube segmentacije in nižje vrednosti za drugi dve nalogi, da bi zagotovili prednost segmentacijske naloge. Na podlagi predhodnih poskusov smo izbrali vrednosti $\lambda_s = 1$, $\lambda_p = 0{,}8$ in $\lambda_d = 0{,}6$, da zagotovimo dober kompromis med tremi nalogami.

3.2 Segmentacija

Običajno se za učenje segmentacije ljudi uporabljajo samo informacije iz anotacijskih mask. V našem pristopu v model vključimo tudi kontekstualne informacije skeletnih točk neposredno v segmentacijsko mrežo. Slika 3 prikazuje visokonivojski pregled komponent v segmentacijski veji modela.

Kot je razvidno, se izhod petega rezidualnega bloka uporablja kot prvotna predstavitev za segmentacijski modul. Za generiranje prvotne segmentacijske maske se na vrhu ekstrahiranih značilk ResNet uporablja dodatna plast združevanja poroznih prostorskih piramid (ASPP). ASPP nad vhodnimi podatki izvede zajem več konvolucij pri različnih stopnjah vzorčenja in velikostih mask, pri čemer zajema predmete in kontekstualne informacije na različnih skalah. Vzporedno s komponento ASPP ustvarimo segmentacijski kontekst, tako da obdelamo izhod pete rezidualne plasti skozi dve dodatni konvolucijski plasti. Ta kontekst se kasneje uporablja v drugi fazi segmentacijske veje skupaj z drugimi viri informacij za nadaljnjo izpopolnjevanje rezultatov.



Slika 3: Pregled segmentacijske veje modela SPD. Veja je sestavljena iz dveh delov. Prvi generira začetno segmentacijo na podlagi značilnosti, ki jih ustvari skupni model, medtem ko drugi začetno segmentacijo izpopolni z uporabo različnih vrst vhodnih informacij – tudi iz drugih vej.
Vir: lasten.

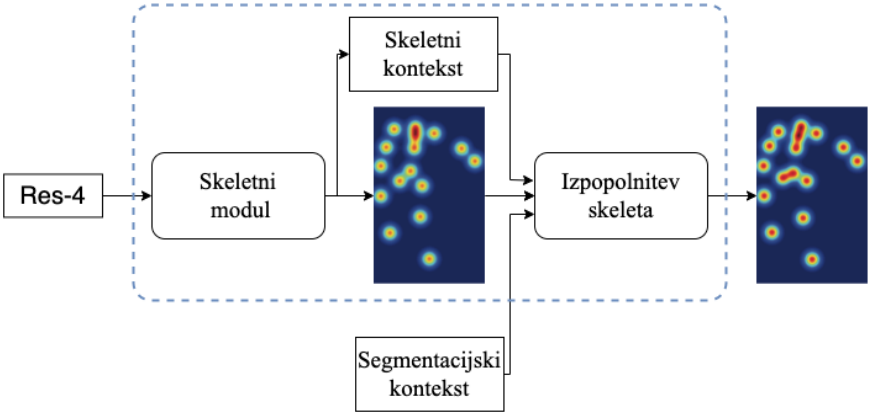
Izpopolnitvena mreža v drugem delu segmentacijske veje vzame kot vhod segmentacijski kontekst in prvotne (grobe) segmentacijske maske in te vhode združi s tako imenovanim skeletnim kontekstom, ki jo ustvari skeletna veja modela. Temu sledijo štirje konvolucijski nivoji, ki služijo zajemu lokalnega konteksta in učenja ključnih povezav med skeletom in segmentacijo. Rezultat teh konvolucijskih plasti je preoblikovan in usmerjen še skozi drugo komponento ASPP. Zadnja ASPP komponenta ustvari končne segmentacijske maske. Izguba segmentacije, se izračuna na koncu te veje in je definirana na nivoju slikovnih elementov kot navzkrižna soft-max entropija, tj:

$$\mathcal{L}_s = \frac{1}{M} \sum_{k=1}^{K} \sum_{m=1}^{M} y_m^k \times log(h_\theta(x_m, k)),$$

kjer je M število vzorcev, K število razredov segmentacije, y_m^k je ciljna klasifikacija za vzorec m in razred k. Vhodni vzorec je označen z x, model napovedi pa z h.

3.3 Veja skeletnih točk

Slika 4 prikazuje visokonivojski pregled nad komponentami, ki sodelujejo pri ustvarjanju predstavitev skeleta. Za razliko od segmentacijske veje je vhod v skeletno vejo izhod četrtega rezidualnega bloka skupne hrbtenice.



Slika 4: Pregled skeletne veje modela SPD. Veja je sestavljena iz dveh delov, kjer prvi generira začetno napoved skeletnih točk na podlagi značilnosti, ki jih ustvari skupen model, medtem ko drugi del prvotno napoved izpopolni z uporabo različnih vrst vhodnih informacij – tudi iz drugih vej.

Vir: lasten.

Začetni skeletni modul v tej veji je sestavljen iz 8 konvolucijskih plasti, prvih šest pridobi skeletne značilnosti, zadnji dve pa generirata prvo različico skeletnih točk v obliki tenzorja s šestnajstimi koordinatami sklepov. Podobno kot v veji segmentacije se v drugi fazi te veje uporablja izpopolnitvena mreža, ki vzame prvotne napovedi skeleta, skeletni kontekst in segmentacijski kontekst, in nato aplicira 4 konvolucije za zajem lokalnega konteksta. Na koncu se za generiranje bolj natančnih koordinat skeletnih točk uporabita dve dodatni konvolucijski plasti.

Funkcija izgube L2 je definirana na koncu skeletne veje, t.j.:

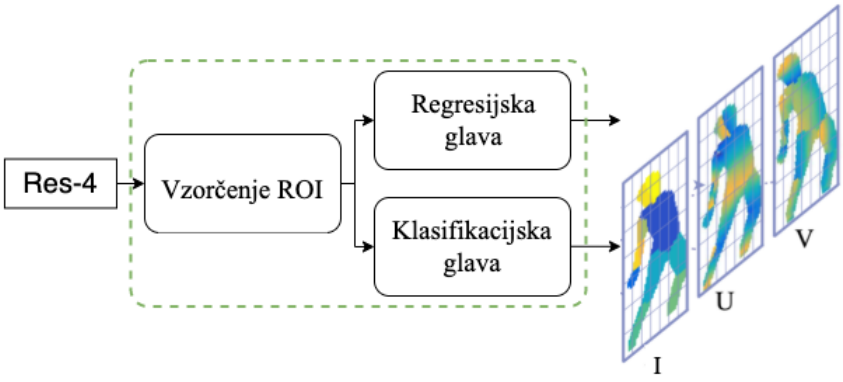
$$\mathcal{L}_p = \frac{1}{2N} \sum_{i=1}^{N} \|p_i - {p_i}'\|^2 ,$$

kjer N predstavlja število definiranih sklepov, p_i' napovedane koordinate sklepov in p_i anotirane koordinate sklepov.

3.4 Veja globinske predstavitve telesa

Slika 5 predstavlja arhitekturo veje globinske predstavitve telesa. Podobno kot v skeletni veji za vhod uporabimo izhod četrtega rezidualnega bloka skupne mreže. Za ResNet mrežo sledi modul za vzorčenje regij interesna (ROI), ki se uporablja za (kaskadni) zajem lokalnih kontekstov na različnih skalah. Nad modulom za vzorčenje ROI je globinska glava, sestavljena iz dveh namenskih CNN glav, klasifikacijske in regresijske.

Prva glava se uporablja za dodelitev slikovnih elementov ustreznemu segmentu telesa, to je klasifikacija komponente *I*. Druga glava določa položaj slikovnih elementov znotraj ustreznih segmentov, torej se uporablja za določanje komponent U in V.



Slika 5: Slika prikazuje visokonivojski diagram arhitekture veje globinske predstavitve telesa.
Vir: lasten, del slike povzet po (Güler,2018)

Funkcija izgube za vejo globinske predstavitve je sestavljena iz dveh delov. Prvi del se nanaša na komponento *I* in je izračunan na enak način kot pri glavni nalogi segmentacije, to je z uporabo navzkrižne entropije. Drugi del, ki se nanaša na koordinate U in V, pa se izračuna s pomočjo Huberjeve funkcije izgube. Celotna funkcija izgube te veje je tore izražena kot

$$\mathcal{L}_d = \sum_{m=1}^{M} CSE(x_I) \cdot L_1(x_U, x_V),$$

kjer so x_I, x_U, x_V komponente predstavitve globine, CSE je transverzna entropijska funkcija za segmentacijski del in L_1 je Huberjeva funkcija izgube za koordinatni del.

4 Eksperimenti in rezultati

V tem delu predstavimo izbrane zbirke podatkov. Opišemo protokol, uporabljen za vrednotenje predlaganega modela SPD, in predstavimo mere uspešnosti, ki smo jih uporabili za vrednotenje nalog modela. Nato komentiramo in analiziramo rezultate modela. Izvedemo tudi ablacijsko analizo in pokažemo prispevek posameznih nalog h končni natančnosti segmentacijskega modela. Na koncu predstavimo tudi kvalitativne primere segmentacijskih mask in jih analiziramo.

4.1 Podatkovne zbirke

Izbira nabora podatkov igra pomembno vlogo pri učenju predlaganega modela SPD. Za naše namene smo uporabili več podatkovnih zbirk, ki vsebujejo slike ljudi v različnih oblačilih, situacijah, kontekstih in položajih telesa. Poseben izziv našega pristopa k modeliranju z več nalogami je potreba po zbirki podatkov, ki vsebuje več različnih vrst anotacij.

Za učenje večciljnega modela, ki vključuje generiranje segmentacij, položajev skeleta in globinskih predstavitev poze, potrebujemo nabor podatkov, ki vsebuje vse tri vrste anotacij. V ta namen smo izbrali zbirko LIP (Gong, 2017), ki vsebuje anotacije segmentacije in skeleta za več kot 50.000 slik.

Za anotacije globinskih predstavitev poze smo uporabili bazo podatkov COCO (Lin, 2014), katere podmnožica je zbirka LIP. Združili smo anotacije obeh zbirk podatkov, da smo pridobili referenčne podatke, potrebne za učenje modela SPD. V končni zbirki imamo anotacije z globinsko predstavitvijo za vse vhodne slike, 19-razredno anotacijo segmentacijskih mask in 16-točkovno oznako za skeletne točke.

4.2 Metrike uspešnosti

Po standardni metodologiji ocenjevanja uporabljamo štiri merila uspešnosti za poročanje o uspešnosti za segmentacijske naloge, to so Jaccardov indeks *IoU*, natančnost, priklic in mera *F1* (Rot, 2020; Emeršič, 2021).

Prvo merilo je Jaccardov indeks ali uteženo povprečje razmerij prekrivanja in unije površin. Mera IoU je definirana kot:

$$IoU = \sum_{i=1}^{K} \frac{S_i' \cap S_i}{S_i' \cup S_i},$$

kjer S' predstavlja označeno območje in S anotirano območje razreda i-ega primerka in K število označenih referenčnih razredov.

Ko gledamo na semantično segmentacijo kot na problem klasifikacije na ravni slikovnih pik, je točnost opredeljena kot razmerje pravilno razvrščenih slikovnih pik med vsemi slikovnimi pikami, razvrščenimi v razred, medtem ko je priklic delež pravilno razvrščenih slikovnih pik med vsemi slikovnimi pikami, ki pripadajo razredu, tj:

$$Pr = \frac{TP}{TP + FP}, \; Rec = \frac{TP}{TP + FN},$$

kjer TP, FP, TN in FN označujejo resnične pozitivne, lažno pozitivne, resnično negativne in lažno negativne.

Rezultat *F1* mere je harmonično povprečje med točnostjo in priklicem:

Za nalogo napovedi skeleta poročamo o povprečni evklidski razdalji (*mED*) med predvideno p_i' in referenčno pozicijo sklepov p_i. Mera je definirana kot:

$$mED = \sum_{i=1}^{N} d_{L_2}(p_i, p_i'),$$

kjer je $d_{L_2}(\cdot)$ funkcija evklidske razdalje, *N=16* pa je skupno število označenih skeletnih točk.

Za nalogo napovedovanja globinske predstavitve poze uporabljamo merilo podobnosti geodetskih točk med ustvarjenimi in referenčnimi točkami globinske predstavitve, kot je definirano v (Guler, 2018). Mera je definirana kot:

$$GPS = \frac{1}{|P|} \sum_{p_i \in P} exp\left(\frac{-d(\hat{p}_i, p_i)^2}{2k(p_i)^2}\right).$$

V zgornji definiciji *P* predstavlja niz označenih površinskih točk, $|\cdot|$ je kardinalnost množice, $\hat{p}_i$ označuje *i*-to napovedano točko na površini in p_i ustrezno anotirano točko na površini osebe. Funkcija *d* predstavlja geodetsko razdaljo med točkami in *k* normalizacijski faktor, specifičen za vsak del telesa.

4.3 Segmentacijski rezultati in ablacijska analiza

S predlaganim modelom SPD želimo izboljšati rezultate obstoječih modelov segmentacije telesa. Natančneje, gradimo na nedavnem pristopu JPPNet iz (Liang, 2019) in zato uporabimo ta model za primerjavo. Tabela 1 prikazuje rezultate segmentacije na zbirki LIP.

Kot je razvidno, na naboru podatkov LIP model SPD doseže *IoU* rezultat 0,547 v primerjavi z modelom JPPNet, ki ima rezultat 0,538. Glede na mero *F1* je predlagani model boljši od JPPNet za približno 5%. Podobne izboljšave zmogljivosti so opažene tudi pri natančnosti in priklicu.

Tabela 1: Rezultati segmentacije in ablacije na naboru podatkov LIP. Puščice poleg metrik nakazujejo kakšna vrednost predstavlja boljši rezultat.

Eksperiment	**Model**	**Metrike uspešnosti**			
		$IoU \uparrow$	$Pr \uparrow$	$Rec \uparrow$	$F1 \uparrow$
Primerjava	JPPNet	0,538	0,68	0,66	0,66
	SPD (naš)	**0,547**	**0,76**	**0,68**	**0,71**
Ablacijska analiza	SP	0,535	0,74	0,52	0,63
	SD	0,478	0,67	0,50	0,57
	S	0,483	0,62	0,49	0,54

Za nadaljnje preverjanje delovanja SPD na neodvisnem naboru podatkov z značilnostmi, ki se razlikujejo od učnih podatkov, smo naš model evalvirali tudi na naboru podatkov ATR. Rezultati segmentacije v tabeli 2 ponovno kažejo, da je SPD boljši od JPPNet glede na vse obravnavane mere uspešnosti. Opažene izboljšave učinkovitosti pripisujemo interakciji treh različnih nalog, kar našemu modelu omogoča, da se bolj učinkovito nauči segmentirati slike v različnih situacijah in svetlobnih pogojih.

Tabela 2: Rezultati segmentacije in ablacije na naboru podatkov ATR. Puščice poleg metrik nakazujejo kakšna vrednost predstavlja boljši rezultat.

Eksperiment	**Model**	**Metrike uspešnosti**			
		$IoU \uparrow$	$Pr \uparrow$	$Rec \uparrow$	$F1 \uparrow$
Primerjava	JPPNet	$0,464$	$0,66$	$0,67$	$0,66$
	SPD (naš)	$\mathbf{0,472}$	$0,67$	$\mathbf{0,70}$	$\mathbf{0,68}$
Ablacijska analiza	SP	$0,423$	$\mathbf{0,69}$	$0,53$	$0,60$
	SD	$0,340$	$0,59$	$0,44$	$0,50$
	S	$0,291$	$0,50$	$0,56$	$0,52$

Da bi prikazali pomen vseh nalog v arhitekturi večciljnega modela SPD, smo izvedli ablacijsko študijo, kjer so iz modela odstranjene različne naloge. Za ta eksperiment so implementirani in naučeni trije dodatni modeli, to so: *(i)* model SPD brez naloge napovedovanja globinske predstavitve (SP v nadaljevanju), *(ii)* model SPD brez naloge za napovedovanje položaja skeletnih točk (SD v nadaljevanju) in *(iii)* model SPD brez obeh nalog, povezanih s pozo (S v nadaljevanju). Rezultati tega poskusa so predstavljeni v tabelah 1 in 2 za nabora podatkov LIP oziroma ATR. Vidi se, da vsaka dodana naloga modelu nudi nove uporabne informacije za izboljšanje rezultatov segmentacije. Odstranitev naloge za napoved globinske predstavitve povzroči padec uspešnosti segmentacije pri vseh merah uspešnosti. Odstranitev naloge za napoved skeletnih točk ima še večji škodljiv učinek na uspešnost. Če sta obe nalogi odstranjeni, opazimo najbolj pomembno poslabšanje zmogljivosti, kar kaže, da obe nalogi, povezani s pozo, zagotavljata pomembne informacije za nadaljnje izboljšanje rezultatov segmentacije. Zanimivo je, da opazimo večje padce zmogljivosti na naboru podatkov ATR kot na LIP. To je verjetno posledica dejstva, da je bil model usposobljen na delu podatkov v LIP, zato so pomožne naloge bolj kritične, ko se spremenijo značilnosti podatkov.

4.4 Rezultati pomožnih nalog

Ker je SPD učen na večciljni način, generira tudi napovedi skeletnih točk in predstavitve vhodnih slik z globinsko predstavitvijo poze. Za boljše razumevanje obnašanja modela tukaj poročamo o rezultatih za naloge napovedovanja skeletnih točk in globinske predstavitve na testnem delu nabora podatkov LIP.

Napoved skeleta. Za prvi poskus ovrednotimo tri modele, predlagani SPD, referenčni JPPNet in SPD model brez naloge napovedovanja globinske predstavitve, to je SP. Na testnih podatkih LIP ima model JPPNet najnižjo vrednost *mED* 51,2 slikovnih pik, sledi model SPD z vrednostjo 55,01 slikovnih pik. Najšibkejši model v tem poskusu je model SP z vrednostjo *mED* 56,82 slikovnih pik. Ti rezultati kažejo, da dodatek naloge z globinsko predstavitvijo jasno izboljša tudi učinkovitost naloge napovedovanja skeletnih točk. Vendar so končni rezultati slabši od JPPNet zaradi dejstva, da so naloge segmentacije dobile višjo prioriteto pri uravnovešanju funkcije izgube.

Napoved globinske predstavitve. Tretja naloga, ki se izvaja znotraj modela SPD, je napoved globinske predstavitve telesa. Ker JPPNet ne ustvarja napovedi globinske predstavitve, poročamo samo o rezultatih za celoten model SPD in model brez naloge za napovedovanje položaja na podlagi skeletnih točk, to je SD.
Na testnih podatkih LIP model SPD dosega oceno *GPS* 48,2%, model SD pa 50,1%. Rezultati kažejo, da dodajanje naloge za napovedovanje položaja skeletnih točk ne pomaga izboljšati učinkovitosti napovedi globinske predstavitve.

4.5 Kvalitativni rezultati

V tem delu predstavimo in analiziramo kvalitativne rezultate, ki jih generira segmentacijska veja modela SPD.

Slika 6 prikazuje primerjavo rezultatov segmentacije, ki jih generirata SPD in JPPNet, skupaj z izvornimi vhodnimi slikami in anotacijskimi maskami segmentacije za tri izbrane slike iz nabora podatkov LIP.

Slika 6: Primerjava rezultatov segmentacije, ki jih generira predlagani model SPD in konkurenčni JPPNet na izbranih slikah iz nabora podatkov LIP. V prvi vrstici so prikazane izbrane vhodne slike, v drugi vrstici segmentacijske anotacije, in v tretji ter četrti vrstici rezultati modelov JPPNet oziroma SPD.

Vir: lasten.

Prva slika prikazuje teniškega igralca in osebo v ozadju, ki je neizostrena in delno zakrita. Vidimo, da je model SPD edini, ki je pravilno zaznal le igralca v ospredju. Referenčni model ima težave z osebo v ozadju, saj je zelo blizu igralca v ospredju.

Razlika v kakovosti segmentacije je vidna tudi pri definiciji prstov na desni roki, kjer je model SPD veliko bolje prepoznal posamezne prste kot model JPPNet. Drugi primer prikazuje žensko, ki je delno skrita za stolom. V tem primeru model JPPNet izpusti celoten predel nog, čeprav je še vedno delno viden skozi stol. Kljub prekrivanju model SPD prepozna položaj noge in jo pravilno označi. Druga edinstvena značilnost te slike je razvrstitev zgornjega dela oblačila. Zgornji del ženskega telesa je označen kot zgornji oblačilni razred, model JPPNet ga napačno klasificira kot plašč, medtem ko model SPD pravilno razvršča območje kot razred zgornjih oblačil, kar je posledica kontekstualnih informacij, ki jih zagotavljata drugi dve nalogi. Na tretji sliki vidimo moškega, ki deska na vodi. V tem primeru model JPPNet daje najboljšo segmentacijo glede na anotacije, saj ustrezno označi zgornji del oblačila in ga loči od hlač. Naš model celotno območje uvršča med enodelne kombinezone, kar je glede na videz slike iz človeške perspektive smiselna klasifikacija.

5 Zaključek

V tem delu smo predstavili večciljni segmentacijski model, imenovan SPD. Poleg primarne naloge segmentacije telesa model vključuje tudi nalogo napovedi skeletnih točk in napovedovanja globinske predstavitve telesa. Segmentacijski del modela je bil ovrednoten na podatkovnih zbirkah LIP in ATR, pri obeh zbirkah podatkov pa je SPD dosegel boljše rezultate kot referenčni model JPPNet. Poleg tega je bilo s strogimi študijami ablacije dokazano, da so modeli, ki so upoštevali manjše število nalog, povzročili slabšo učinkovitost.

V analizi ablacije smo predstavili prispevek vsake od nalog in ugotovili, da skupna uporaba skeleta in globinske naloge dodaja večjo vrednost kot uporaba katere koli od njiju samostojno. Za nadaljnje izboljšanje rezultatov načrtujemo raziskovanje dodatnih nalog v učnem postopku, ki bi lahko dale dodatne napotke za postopek segmentacije.

Opomba

To raziskavo so delno podprli projekt ARRS J2-2501 "Globoki generativni modeli za lepoto in modo (DeepBeauty)", raziskovalni program ARRS P2-0250(B) "Meroslovje in biometrični sistemi" in ARRS Research Program P2-0214 "Računalniški vid".

Literatura

X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "Viton: An image-based virtual try-on network," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7543–7552.

B. Fele, A. Lampe, P. Peer, and V. Štruc, "C-vton: Context-driven image-based virtual try-on network," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2022.

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," in Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4. IEEE, Apr. 2018, pp. 834–848. [Online]. Available: https://doi.org/10.1109/tpami. 2017.2699184

J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high- resolution representation learning for visual recognition," vol. 43, no. 10. IEEE, Oct. 2021, pp. 3349–3364. [Online]. Available: https://doi.org/10.1109/tpami.2020.2983686

K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Jul. 2017, pp. 556–567. [Online]. Available: https://doi.org/10.1109/cvpr.2017.715

R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in Conference on Computer Vision and Pattern Recognition. IEEE, Jun. 2013, pp. 500–518. [Online]. Available: https://doi.org/10.1109/cvpr.2013.460

X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," in Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 12. IEEE, Dec. 2015, pp. 2402–2414. [Online]. Available: https://doi.org/10.1109/tpami.2015.2408360

X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with local-global long short-term memory," in Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Jun. 2016, pp. 710–724. [Online]. Available: https://doi.org/10.1109/cvpr.2016.347

I. Kokkinos, "UberNet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory," in Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Jul. 2017, pp. 1380–1410. [Online]. Available: https://doi.org/10.1109/cvpr.2017.579

D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in International Conference on Computer Vision (ICCV). IEEE, Dec. 2015, pp. 800–820. [Online]. Available: https://doi.org/10.1109/iccv. 2015.304

B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," in International Conference on Image Processing (ICIP). IEEE, Sep. 2019, pp. 630–647. [Online]. Available: https://doi.org/10.1109/icip.2019.8803050

X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," in Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 4. IEEE, Apr. 2019, pp. 871–885. [Online]. Available: https://doi.org/10.1109/tpami.2018.2820063

R. A. Guler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in CVF Conference on Computer Vision and Pattern Recognition. IEEE, Jun. 2018, pp. 1120–1135. [Online]. Available: https://doi.org/10.1109/cvpr.2018.00762

K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in Computer Vision – ECCV. Springer International Publishing, 2014, pp. 740–755. [Online]. Available: https://doi.org/10.1007/978-3-319-10602-1 48

P. Rot, M. Vitek ,K. Grm, Z. Emeršič, P. Peer, and V. Štruc," Deep sclera segmentation and recognition," in Handbook of vascular biometrics. Springer, Cham, 2020, pp. 395–432.

Z. Emeršič, D. Sušanj, B. Meden, P. Peer, and V. Štruc, "ContexedNet: Context–aware ear detection in unconstrained settings," IEEE Access, vol. 9, pp. 145 175–145 190, 2021.