ACADEMIA ROMÂNĂ

INSTITUTUL DE MATEMATICĂ SIMION STOILOW

Modele Integrate pentru Detecția și Reconstrucția Vizuală a Persoanelor

Autor,
Alin-Ionuț POPA

Conducător,
C.S. I Dr. Cristian SMINCHIȘESCU

SUMAR AL TEZEI DE DOCTORAT

București
2018

Capitolul 1

Introducere

În această teză, ne concentrăm pe studiul asupra înțelegerii vizuale a oamenilor din imagini monoculare, în particular detecția și segmentarea suportului spațial al oamenilor din imagini, identificarea părților anatomice și estimarea unei reconstrucții 3D a configurației scheletului. Problema este dificilă deoarece oamenii au multe grade de libertate de-a lungul lanțului cinematic cauzate de articulații si deformări, și deoarece proporțiile corporale și aspectul fizic, incluzând tipuri de îmbrăcăminte, pot varia considerabil. Costul inerent în proiecția perspectivei vizuale, ocluziile cauzate de oameni sau obiecte, sau nivelul de complexitate al fundalului, complică și mai mult înțelegerea vizuală a oamenilor.

Analiza oamenilor din date vizuale este un domeniu științific important ce poate avea numeroase aplicații în alte domenii cum ar fi recunoașterea de activități sau reconstrucția 3d detaliată a scenelor. Deasemenea, există o gamă largă de aplicații industriale ce pot beneficia de problema abordată, acestea incluzând efecte speciale, terapie medicală asistată, sisteme de supraveghere sau industria auto. Considerând toți factorii enumerați mai sus, complexitatea problemei face ca soluția să fie non-trivială.

Obiectivul studiului nostru este de a introduce un model complex util pentru analiza, înțelegerea și reconstrucția oamenilor din date vizuale. Astfel, ne propunem să construim modele capabile să separe silueta unui om de fundal, să determinăm pozițiile 2d/3d ale punctelor cheie de pe lanțul cinematic și să reconstruim omul sub diverse configurații de formă și de îmbrăcăminte. Alt obiectiv principal al tezei este de a ilustra faptul că aceste subprobleme pot fi aplicate în contextul învățarii semi-supervizate. Toate modelele

propuse sunt discutate și evaluate în contextul standardelor din literatura *state-of-the-art*, și al seturilor de date pe scară largă.

Capitolul 2

Segmentare Parametrizată a Oamenilor prin Cunoaștere Anterioară de Formă

În acest capitol studiem problema segmentării oamenilor, bazându-ne pe cunoștințe anterioare de formă folosite în cadrul unor modele de optimizare a energiei. Propunem o metodologie ce are la bază elemente de fuziune particulare unei clase de obiecte și date din cadrul clasei respective, care aliniază segmente candidat cu exemple dintr-o bază de date cu siluete de oameni, ce permite construirea cunoașterii anterioare de formă pentru puncte de observație arbitrare și parțial vizibile. Această cunoaștere anterioară poate fi cu ușurință integrată în modele de optimizare a energiei bazate pe partiționări în grafuri. Segmentarea oamenilor în imagini capturate în medii naturale este o problemă deschisă dificilă cauzată de prezența fundalurilor complexe, a diferitelor proporții ale corpului uman și de diferitele puncte de observație ale imaginilor.

Ne bazăm pe metode de generare de segmentări ale obiectelor din prim-plan *bottom-up* și pe clasificatoare de segmentări ale oamenilor pentru a putea identifica segmente candidat potrivite pentru rafinare. Într-o a doua trecere prin date, aplicăm constrângeri de segmentare de clasă de oameni prin cunoașteri anterioare de formă de oameni și informații cinematice deduse din scheletul uman. Cunoașterea anterioara de formă umană este obținută aliniind segmentele candidat cu o bază de date de scară largă recent construită (Human3.6M [18]) ce conține informații referitoare la configurațiile 2d și 3d ale scheletelor umane și de asemenea a segmentării din fundal. Exploatând modele globale de minimizare de energie de tipul

Metodă	H3D Set de Test [5]			MPII Set de Test [1]		
	Primul	Optim	Dim. Set	Primul	Optim	Dim. set
CPMC [7]	0.54	0.72	783	0.29	0.73	686
CPDC - MAF	0.60	0.72	77	0.55	0.71	102
CPDC - MAF - POSELETS	0.53	0.6	98	0.43	0.58	116

Tabela 2.1: Statistici asupra acurateții și dimensiunii setului de segmente propuse pentru diverse metode, peste seturile de date H3D și MPII. Raportăm media metricii *Intersection over Union* (IoU) peste setul de date de test dintre primul segment din setul de segmente propus și segmentul țintă al imaginii respective (*Primul*), a segmentului cu cel mai mare IoU cu segmentul țintă al imaginii respective (*Optim*) și media dimensiunii setului de segmente propuse (*Dim. Set*).

max-flow, pentru acest caz particular și folosind informații de prim plan specifice clasei segmentate (față de cazul generic și regulat) [11, 20, 7], arătăm că putem îmbunătății considerabil calitatea segmentelor generate. Din cunoștințele noastre asupra literaturii de specialitate, aceasta este una dintre primele formulări ale unei metode de segmentare specifică unei anumite clase de obiecte ce poate rezolva cazuri când obiectul vizat este văzut parțial sau din diferite puncte dificile de observație. De asemenea, este una dintre primele metode care se bazează pe exemple de siluete dintr-o bază mare de date cu oameni, împreună cu alte informații structurale utile disponibile. Arătăm că astfel de constrângeri sunt critice și esențiale pentru acuratețe, robustețe și eficiență computațională.

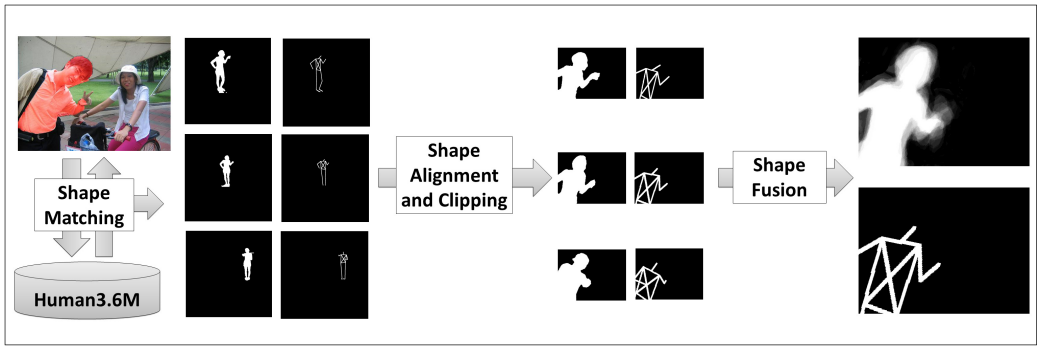

Figura 2-1: Algoritmul nostru de Potrivire și Aliniament de Forme (MAF) bazat pe potrivire semantica, aliniere structurală și decupare, urmată de fuziune. Poate funcționa cu succes și pentru vederi parțiale. A se observa cum algoritmul de construcție a cunoașterii anterioare ne permite să potrivim vederi parțiale pentru un segment candidat la siluete complet vizibile din datasetul Human3.6M. În această manieră putem crea o hartă de activare ce conține cea mai probabilă siluetă umană, pornind de la un segment candidat.

Metodologia prezentată a fost evaluată pe două seturi de date dificile H3D [5] cu 107

imagini și MPII [1] cu 3799 imagini. Avem segmentări țintă disponibile pentru ambele seturi de date. Pentru MPII, am generat segmentări țintă noi înșine. Ambele seturi de date, H3D și MPII, conțin și vederi parțiale și vederi complete asupra persoanelor ocludate, ceea ce face ca problema să devină și mai dificilă.

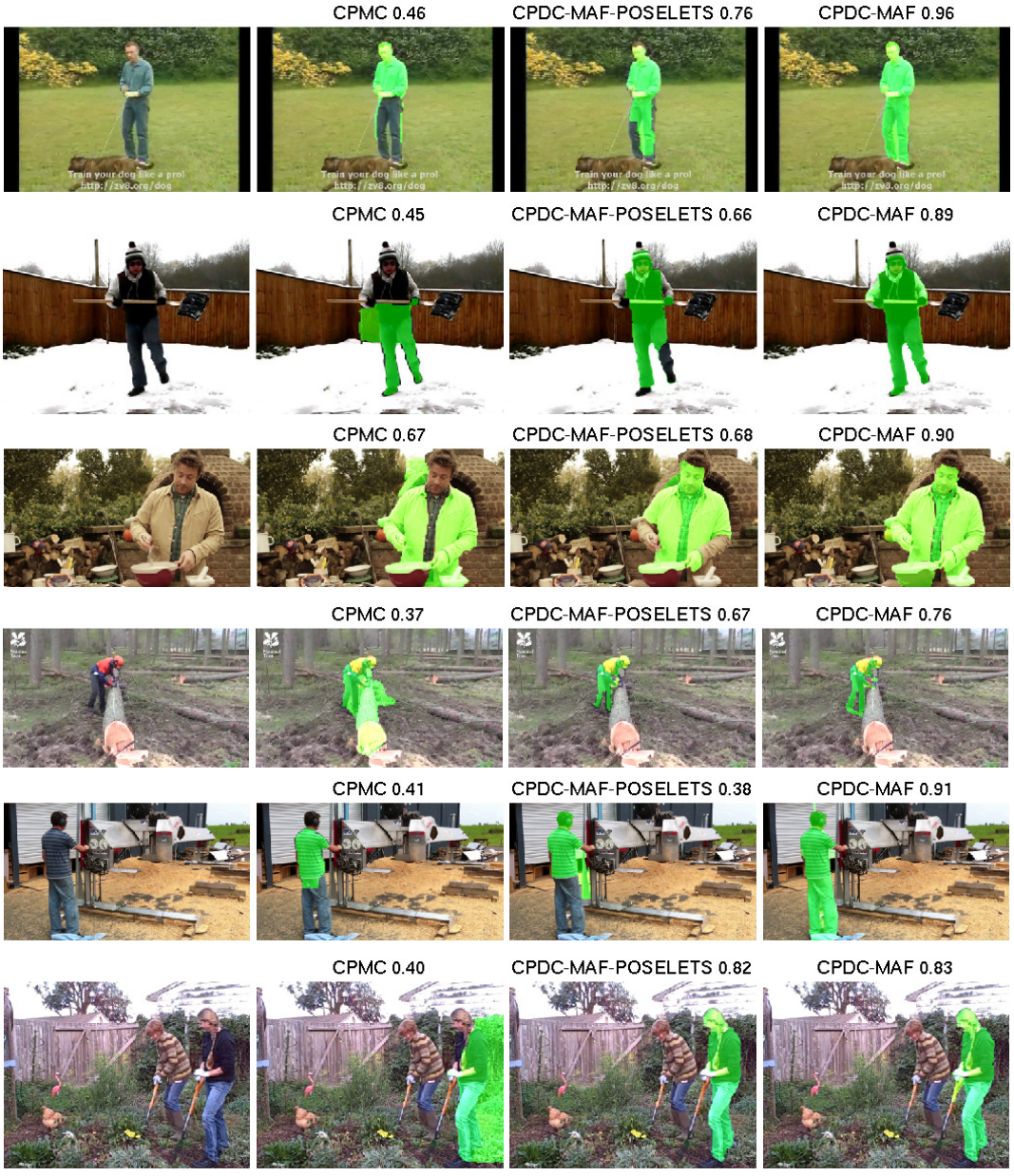

Figura 2-2: Exemple de rezultate ale metodelor de segmentare. De la stânga la dreapta avem, imaginea originală, CPMC cu parametrii normali rulată pe dreptunghiul de delimitare al persoanei, metodele noastre, CPDC-MAF-POSELETS și CPDC-MAF. Verificați de asemenea și tabelul 2.1 pentru rezultate numerice.

Capitolul 3

Aproximare de Kernel Dependent de Structura Datelor pe Scară Largă

În capitolul *Aproximare de Kernel Dependent de Structura Datelor pe Scară Largă*, ne concentrăm pe scalabilitatea modelelor de învățare pentru situațiile când avem de-a face cu predicție de ieșire continuă. Dezvoltăm o formulare bazata pe aproximări de kernel folosind descriptori Fourier aleatori în contextul învățării semi-supervizate. Modelul este bine definit, și este sprijinit și de rezultate teoretice și experimentale. În plus, arătăm că este eficient pentru problema de determinare a configurațiilor de schelete umane 3d cu supervizie slabă. Învățarea eficientă a unui kernel din punct de vedere computațional, din date, este o problema fundamentală pentru comunitatea de învățare automată. Majoritatea kernelurilor din literatura nu iau un considerare structura geometrică a datelor, și cele care fac asta sunt nefezabile computațional pentru seturile de date contemporane (de scară largă). Progresul recent în tehnicile de aproximare a kernelurilor au făcut ca metodologia dedicată lor sa scaleze liniar în raport cu dimensiunea datelor pe care sunt aplicate. Kernelurile dependente de structura datelor [33] nu au beneficiat încă de acest avantaj. În acest studiu, derivăm o *aproximare pentru o procedură de învățare pe scară largă pentru kernelurile dependente de structura datelor* ce este eficientă și a dat rezultate bune în practică.

Metoda noastră nu necesită construirea matricilor de kernel (sau Gram), așa cum se procedează în cazul metodelor pe bază de kernel tradiționale. Propunem o aproximare pentru kernelul dependent de structura datelor [33], printr-o formulare ce implică multiplicarea

Eroare de postura (mm)		Divizare Adnotat vs. Neadnotat	
RFKRR	LapRFKRR	# de Adnotate	# de Neadnotate
57.83	57.72	105,543	949,881
61.6	60.83	10,555	1,044,869
77.99	71.95	1,056	1,054,368
87.41	79.81	528	1,054,896
89.48	84.85	352	1,055,072
94.88	91.68	264	1,055,160
97.37	93.2	212	1,055,212

Tabela 3.1: Evaluarea performanței pentru Human3.6M, pentru diverse divizări ale datelor în adnotat vs. neadnotat. Coloana RFKRR se referă la performanța modelelor antrenate folosind numai datele adnotate, în timp ce LapRFKRR folosește ambele divizări, și adnotate și neadnotate în contextul aproximării kernelului dependent de structura datelor. Se observă că pe măsură ce dimensiunea diviziei de date adnotate crește, performanța modelului RFKRR crește de asemenea. În schimb, observăm că obținem îmbunătățiri în configurația de învățare semi-supervizată, astfel demonstrând și scalabilitatea și avantajul de a folosi metoda de aproximare a kernelului dependent de structura datelor.

descriptorilor Fourier aleatori, obținuți cu [24], împreună cu o matrice de covarianță ponderată construită folosind date complet și parțial adnotate. Acest lucru practic deformează distanțele dintre descriptorii Fourier. Drept urmare, uneori ne referim la ea ca la matricea de deformare. Aproximarea kernelului dependent de structura datelor astfel rezultată are aceleași proprietăți ca metoda din [33], dar nu are aceleași limitări computaționale asociate cu construirea matricii Gram atașate funcției kernel. Construcția aproximării este posibilă printr-o aplicare abilă a identității Woodbury care mută problema de învățare dintr-un spațiu RKHS într-un spațiu Fourier al kernelului, reducînd totodată costul computațional de calcul de la $O(N^3)$ la $O(N)$. Totodată, formulăm și o Lemă care poate fi folosită pentru a deriva o convergență asimptotă a aproximării în limita unui spațiu infinit de descriptori Fourier aleatori, și, în anumite condiții, un estimat al vitezei de convergență. Demonstrăm empiric că putem construi o reprezentare validă, dar în același timp eficientă a aproximării kernelului dependent de structura datelor.

Pentru studiul empiric pe scară largă, considerăm problema de estimare a configurației 3D a scheletului uman, pornind de la informația 2D. Rulăm experimentele pe baza de date Human3.6M [17], de unde eșalonăm un subset de 1,055,424 de configurații de schelețti pentru antrenare și 56,860 de configurații de schelețti pentru testare. Examinăm următorul scenariu de învățare: dându-se o configurație de schelet 2D, să învățăm un model ce este

capabil să estimeze corespondetul acelui schelet în spațiul 3D. Astel, configurația 2D a scheletului uman constitue datele noastre de intrare și configurația 3D a scheletului uman constitue datele noastre țintă. Normalizăm datele 2D astfel încât originea sistemului de coordonate să fie centrată în pelvis. De asemenea, rotim fiecare configurație 2D în plan, astfel încât axa determinată de gât și pelvis să se alinieze cu axa OY și în plus întregul schelet este scalat astfel încât latura gât - pelvis să aibe dimensiunea 1. Ca model de învățare am decis să folosim regularizarea Tikhonov pe bază de kernel. Am decis să folosim un kernel Gaussian în experimentele noastre datorită structurii datelor. Aproximarea pe bază de descriptori aleatori are la bază $d = 4,000$ de dimensiuni. Standard, întreg setul de date este complet adnotat, având și configurațiile 2D și pe cele 3D disponibile. Pentru problema noastră de învățare semi-supervizată, am considerat ca informația 3D să lipsească pentru o parte din date, conform unor divizări prestabilite. Peformanța modelului pentru acest caz poate fi vizualizată în tabelul 3.1. Luați aminte deasemenea, că în timpul acestui experiment am variat raportul dintre numărul de date complet adnotate și cel de date parțial adnotate, ținând numărul totat de date folosite fix. Intuiția din spatele acestei decizii stă în faptul că am vrut să vedem impactul setului de date complet adnotate, dat fiind că numărul de date parțial adnotat este dens ($\simeq 1,000,000$). Scopul acestui experiment este de a demonstra empiric că aproximarea kernelului dependent de structura datelor îmbunătățește problema de învățare de configurații de scheleți 3D într-un scenariu semi-supervizat de învățare, non-trivial, unde am avut de-a face cu seturi de date de peste 1 milion de elemente.

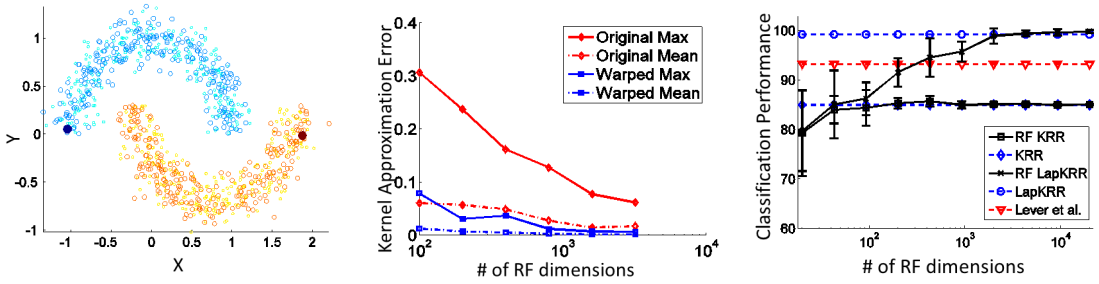

Figura 3-1: (Stânga) Setul de date *Two Moons*. (Mijloc) Erori de aproximare ale kernelurilor (maximul respectiv media erorii absolute) original, K, respectiv kernelul dependent de structura datelor, $\widetilde{K}$. (Dreapta) Performanța problemei de clasificare raportată la setul de date *Two Moons*, folosind un singur exemplu complet adnotat per clasă. De-a lungul axei OX se află numărul de descriptori aleatori Fourier folosiți pentru aproximare. A se observa faptul ca setupul semi-supervizat (folosind aproximarea kernelul dependent de structura datelor), atinge o performanță notabilă chiar și în cazul când avem o aproximare slabă a kernelului (500 de dimensiuni). Folosind 2,000 de descriptori aleatori Fourier obținem aceeași performanță ca atunci când folosim kernelul original dependent de structura datelor. În plus, am făcut o comparație cu o metodă similară de aproximare de kernel [23]. Pentru aproximarea lor am folosit 500 de puncte cheie.

Capitolul 4

Transfer de Aparență Umană

În ultimul capitol tehnic al tezei, *Transfer de Aparență Umană*, introducem și explorăm o nouă problemă, cea a transferului de aparență de la o persoană la alta. Acesta este definit formal ca transferul aparenței dintre două persoane diferite, în ipostaze diferite, cu îmbrăcăminte variată, știind că fiecare se află în câte o imagine. Pentru această sarcină, propunem o soluție care combină modelare geometrică 3d și rețele neuronale profunde, care în medie sintetizează aparența umană la o calitate fotorealistă.

Dându-se o pereche de imagini RGB – sursă și țintă, notate cu I_s și I_t, fiecare conținând câte o persoană –, principalul nostru obiectiv este de a transfera aparența persoanei din imaginea sursă I_s în postura persoanei din imaginea țintă I_t, rezultând astfel o nouă imagine $I_{s \Rightarrow t}$.[1] Platforma noastră computațională poate fi vizualizată in figura 4-1.

Soluția noastră la această nouă problemă este formulată în termeni de o platformă computațională ce combină (1) configurația scheletului 3d și a formei corpului 3d potrivită optimal peste imagine, (2) identificarea triangulărilor formei corpului 3d în termeni de culori RGB vizibile din ambele imagini, care pot fi transferate direct folosind proceduri baricentrice de la sursă la țintă, și (3) prezicerea culorilor suprafețelor ce sunt vizibile în imaginea țintă dar care nu sunt vizibile în imaginea sursă folosind tehnici de sinteză de imagini pe bază de rețele neuronale profunde. Modelul propus de noi se bazează pe predictoare de configurații de schelet 2d împreună cu identificarea părților anatomice ale corpului[6, 31, 16], predictoare de configurații de schelet 3d [31, 3, 35], potrivire optimală

[1] Această procedura este simetrică, lucru care permite ca transferul să se realizeze în ambele direcții.

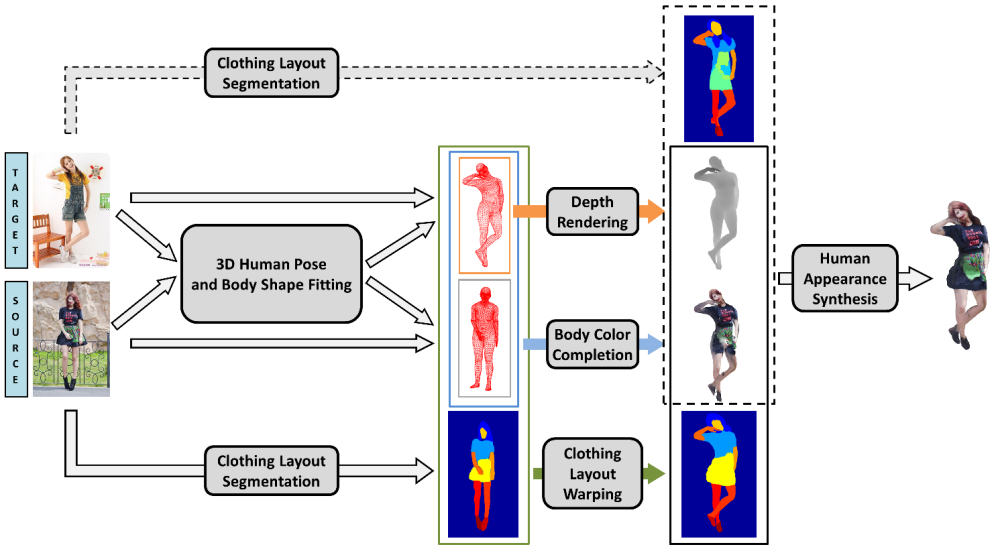

Figura 4-1: Platforma computațională de transfer a aparenței umane. Dându-se o singură imagine sursă și una țintă, cu aparențe, posturi și îmbrăcăminți variate, scopul nostru este de a transfera într-o manieră fotorealistă aparența de la sursă la țintă păstrând forma și aranjamentul de îmbrăcăminte ale omului din țintă. Problema este definită în termeni de componente ale unei platfome computaționale formată din *(i)* configurația scheletului 3d împreună cu forma corpului 3d potrivită optimal peste imagine, împreună cu *(ii)* transferul culorilor de pe suprafața formei corpului 3d sursă în țintă pentru triangulările formelor 3d vizibile din ambele punctele de observație folosind proceduri baricentrice, *(iii)* prezicerea aparenței suprafeței lipsă vizibilă doar în imaginea țintă folosind tehnici pe bază de rețele neuronale profunde. Ultimul pas, *(iv)*, folosește rezultatul anterior împreună cu aranjamentul de îmbrăcăminte al sursei deformat peste țintă și sintetizează rezultatul final. Dacă aranjamentul îmbrăcămințiii sursă este similar cu aranjamentul îmbrăcămințiii țintă, atunci evităm procedura de deformare și, în schimb, folosim îmbrăcămintea țintă.

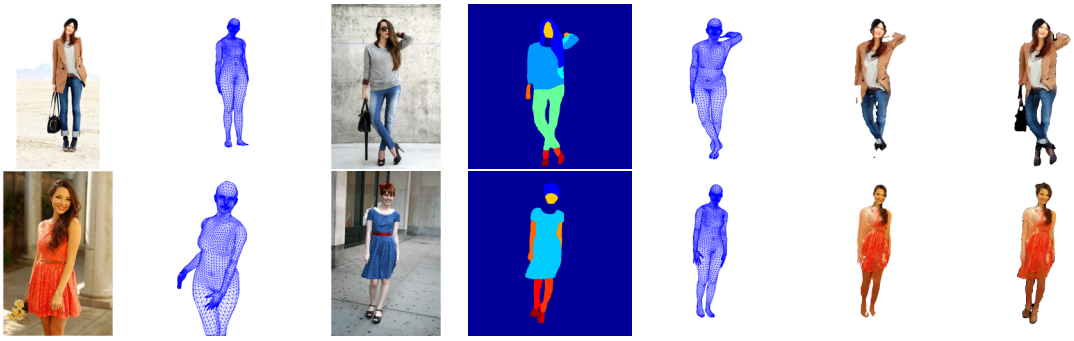

Figura 4-2: Exemple de rezultate ale platformei noastre de transfer de aparență umană. De la stânga la dreapta: imaginea sursă cu forma corpului 3d potrivită optimal peste imagine, imaginea țintă împreună cu aranjamentul de îmbrăcăminte și forma corpului 3d corespunzătoare, colorarea siluetei cu valori RGB (*i.e.* $I_{s \to t}$), sinteza finala a persoanei din imaginea sursă în imaginea țintă (*i.e.* $I_{s \Rightarrow t}$).

parametrică de forme ale corpului umane 3d [13, 39, 29, 28], algoritmi de segmentare semantică a aranjamentelor de îmbrăcăminte [32, 14, 27, 15], și nu în ultimul rând de metode de sinteză de imagini și traducătoare/translatoare de imagini [19, 8, 30, 22, 43, 8, 40].

Modelul nostru produce rezultate vizuale promițătoare ce sunt susținute de un studiu perceptual unde participanți umani au estimat că 65% din rezultatele noastre sunt bune, foarte bune sau perfecte. De asemenea, am efectuate și teste automate (scoruri de Incepție și un detector uman, Faster-RCNN) ce au arătat o rata de răspuns peste imaginile noastre similară cu imagini reale. Am arătat de asemenea ca arhitectura propusă de noi poate fi folosită cu succes pentru a îmbrăca o persoană cu diferite aranjamente de îmbrăcăminte, astfel deschizând orizonturi către aplicații din industriile de diversitment, editare fotografică (*e.g.* pozând ca prieteni sau celebrități), industria de modă, sau magazine online de îmbrăcăminte.

Pentru toate experimentele noastre, am folosit baza de date Chictopia10k [26]. Imaginile din această bază de date descriu diverse topologii de oameni, din puncte de observare parțiale sau complete, capturate frontal. Variabilitatea extinsă în termeni de culoare, îmbrăcăminte, iluminare și postură fac ca acest set de date să fie potrivit pentru obiectivul nostru. Sunt un număr de $17,706$ de imagini disponibile împreună cu segmentările de îmbrăcăminte țintă. Nu folosim segmentările de îmbrăcăminte țintă puse la dispoziție, ci doar segmentarea siluetei umane față de fundal, astfel încât să putem genera imagini de antrenare decupate în jurul siluetei umane. Exemple de rezultate ale modelului nostru pot fi vizualizate în figura 4-2.

Bibliografie

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[2] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016.

[4] Lubomir Bourdev, Subhransu Maji, Thomas Brox, and Jitendra Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010.

[5] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, sep 2009.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, July 2017.

[7] J. Carreira and C. Sminchisescu. CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts. *PAMI*, 2012.

[8] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, October 2017.

[9] Andrew Cotter, Joseph Keshet, and Nathan Srebro. Explicit approximations of the gaussian kernel. *CoRR*, abs/1109.4603, 2011.

[10] V. Ferrari, M. Marin, and A. Zisserman. Pose Seach: retrieving people using their pose. In *CVPR*, 2009.

[11] G. Gallo, M. D. Grigoriadis, and R. E. Tarjan. A fast parametric maximum flow algorithm and applications. *SIAM J. Comput.*, 18(1):30–55, 1989.

[12] Golnaz Ghiasi, Yi Yang, Deva Ramanan, and Charless C. Fowlkes. Parsing occluded people. In *CVPR*, 2014.

[13] Rony Goldenthal, David Harmon, Raanan Fattal, Michel Bercovier, and Eitan Grinspun. Efficient simulation of inextensible cloth. *ACM Transactions on Graphics (TOG)*, 26(3):49, 2007.

[14] Ke Gong, Xiaodan Liang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, July 2017.

[15] M. Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C. Berg, and Tamara L. Berg. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, December 2015.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[17] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *PAMI*, 2014.

[18] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[20] V. Kolmogorov, Y. Boykov, and C. Rother. Applications of parametric maxflow in computer vision. *ICCV*, 2007.

[21] Lubor Ladicky, Philip H. S. Torr, and Andrew Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *CVPR*, 2013.

[22] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

[23] G. Lever, T. Diethe, and J. Shawe-Taylor. Data dependent kernels in nearly-linear time. *AISTATS*, 2012.

[24] F. Li, C. Ionescu, and C. Sminchisescu. Random Fourier approximations for skewed multiplicative histogram kernels. In *LNCS (DAGM)*, September 2010.

[25] F. Li, G. Lebanon, and C. Sminchisescu. Chebyshev approximations to the histogram χ^2 kernel. In *CVPR*, 2012.

[26] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, 2015.

[27] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, pages 1096–1104, 2016.

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *SIGGRAPH*, 34(6):248, 2015.

[29] Rahul Narain, Armin Samii, and James F O'Brien. Adaptive anisotropic remeshing for cloth simulation. *ACM transactions on graphics (TOG)*, 31(6):152, 2012.

[30] Anh Nguyen, Jason Yosinski, Yoshua Bengio, Alexey Dosovitskiy, and Jeff Clune. Plug & play generative networks: Conditional iterative generation of images in latent space. In *CVPR*, 2017.

[31] A. Popa, M. Zanfir, and C. Sminchisescu. Deep Multitask Architecture for Integrated 2D and 3D Human Sensing. In *CVPR*, July 2017.

[32] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. A high performance crf model for clothes parsing. In *ACCV*, 2014.

[33] Vikhas Sindhwani, P Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *ICML*, 2005.

[34] V. Sreekanth, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Generalized RBF feature maps for efficient detection. In *BMVC*, 2010.

[35] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014.

[36] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *PAMI*, 2012.

[37] Huayan Wang and Daphne Koller. Multi-level inference by relaxed dual decomposition for human pose segmentation. In *CVPR*, 2011.

[38] Wei Xia, Zheng Song, Jiashi Feng, Loong Fah Cheong, and Shuicheng Yan. Segmentation over detection by coupled global and local sparse representations. In *ECCV*, 2012.

[39] Feng Xu, Yebin Liu, Carsten Stoll, James Tompkin, Gaurav Bharaj, Qionghai Dai, Hans-Peter Seidel, Jan Kautz, and Christian Theobalt. Video-based characters: Creating new human performances from a multi-view video database. In *ACM SIGGRAPH 2011 Papers*, SIGGRAPH, pages 32:1–32:10, New York, NY, USA, 2011. ACM.

[40] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *CVPR*, 2017.

[41] Jiyan Yang, Vikas Sindhwani, Quanfu Fan, Haim Avron, and Michael Mahoney. Random laplace feature maps for semigroup kernels on histograms. In *CVPR*, 2014.

[42] Yi Yang and Deva Ramanan. Articulated Human Detection with Flexible Mixtures of Parts. *PAMI*, 2013.

[43] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[44] Silvia Zuffi, Javier Romero, Cordelia Schmid, and Michael J Black. Estimating human pose with flowing puppets. In *ICCV*, 2013.