THE ROMANIAN ACADEMY

SIMION STOILOW INSTITUTE OF MATHEMATICS

Integrated methods for the detection and visual reconstruction of people

Author,
Alin-Ionuț POPA

Supervisor,
C.S. I Dr. Cristian SMINCHIȘESCU

PH. D. THESIS SUMMARY

Bucharest
2018

Chapter 1

Introduction

In this thesis, we focus on the study of visual human sensing from monocular images, in particular detecting and segmenting the spatial support of people in images, recognizing their body parts and estimating a three-dimensional reconstruction of the body pose. The problem is challenging because people have many degrees of freedom due to deformation and articulation, and because their body proportions and appearance, including clothing, vary considerably. The inherent loss in the visual perspective projection, occlusions due to other people or objects, or the background complexity, further complicates the visual analysis.

Understanding humans from visual data is an important scientific domain with numerous use cases including activity recognition or complete 3d scene understanding. There is a wide range of industrial applications to our studied research problem, including special effects, assisted medical therapy, surveillance systems or the automotive industry. The complexity of the task makes a robust solution non-trivial.

The scope of our work is to introduce a complex model useful for the visual analysis, understanding and reconstruction of people from visual data. Thus, we aim to build models able to separate the silhouette of a human from the background, determine its 2d/3d position in terms of kinematic joint representation and reconstruct it under various shape and clothing representations. Another objective of the thesis is to illustrate that these tasks can be applied in the context of large-scale semi-supervised learning models. All of our proposed models are discussed and evaluated in the context of current state-of-the-art literature standards, and

on large scale datasets.

Chapter 2

Parametric Image Segmentation of Humans with Structural Shape Priors

We approach the problem of figure-ground human segmentation, by relying on shape-priors within energy optimization models. We propose a data-driven class-specific fusion methodology, based on matching against a large training set of exemplar human shapes, that allows the shape prior to be constructed on-the-fly, for arbitrary viewpoints and partial views. The prior can be seamlessly integrated into efficient optimization methods based on graph-cuts. The figure-ground segmentation of humans in images captured in natural environments is an outstanding open problem due to the presence of complex backgrounds, articulation, varying body proportions, partial views and viewpoint changes.

We rely on bottom-up figure-ground generation methods and region-level person classifiers in order to identify promising hypotheses for further processing. In a second pass, we set up informed constraints towards (human) class-specific figure-ground segmentation by leveraging skeletal information and data-driven shape priors computed on-the-fly by matching region candidates against exemplars of a large, recently introduced human motion capture dataset containing 3D and 2D semantic skeleton information of people, as well images and figure-ground masks from background subtraction (Human3.6M [18]). By exploiting globally optimal parametric max-flow energy minimization solvers, this time, based on a class dependent (as opposed to generic and regular) foreground seeding process [11, 20, 7], we show that we can considerably improve the quality of competitive object

proposal generators. To our knowledge, this is one of the first formulations for class-specific segmentation that in principle can handle multiple viewpoints and any partial view of the person. It is also one of the first to leverage a large dataset of human shapes, together with semantic structural information, which until recently, have not been available. We show that such constraints are critical for accuracy, robustness, and computational efficiency.

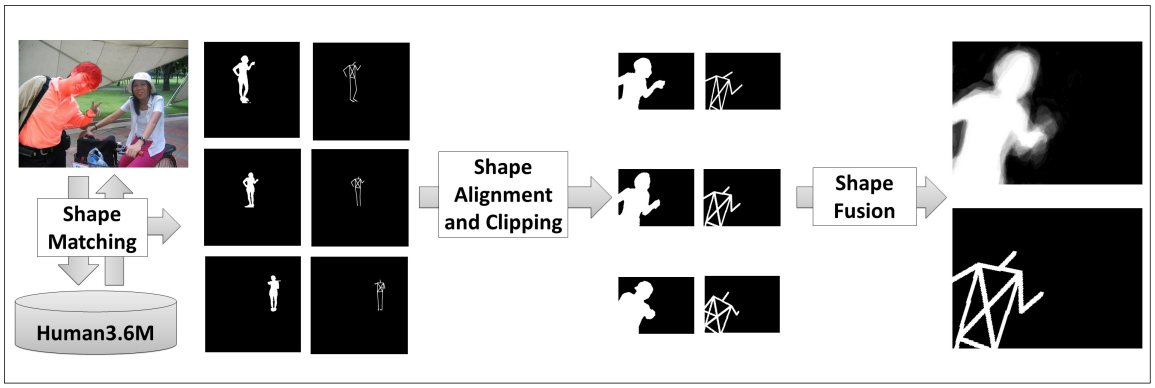

Figure 2-1: Our Shape Matching Alignment Fusion (MAF) construction based on semantic matching, structural alignment and clipping, followed by fusion, to reflect the partial view. Notice that the prior construction allows us to match partial views of a putative human detected segment to fully visible exemplars in Human3.6M. This allows us to handle arbitrary patterns of occlusion. We can thus create a well adapted prior, on-the-fly, given a candidate segment.

We test our methodology on two challenging datasets: H3D [5] which contains 107 images and MPII [1] with 3799 images. We have figure-ground segmentation annotations available for all datasets. For the MPII dataset, we generate figure-ground human segment annotations ourselves. Both the H3D and the MPII datasets contain both full and partial views of persons with self-occlusion which makes them extremely challenging.

Method	**H3D Test Set** [5]			**MPII Test Set** [1]		
	First	Best	Pool size	First	Best	Pool size
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

Table 2.1: Accuracy and pool size statistics for different methods, on data from H3D and MPII. We report average Intersection over Union (IoU) over test set for the first segment of the ranked pool and the ground-truth figure-ground segmentation (*First*), the average IoU over test set of the segment with the highest IoU with the ground-truth figure-ground segmentation (*Best*) and average pool size (*Pool Size*).

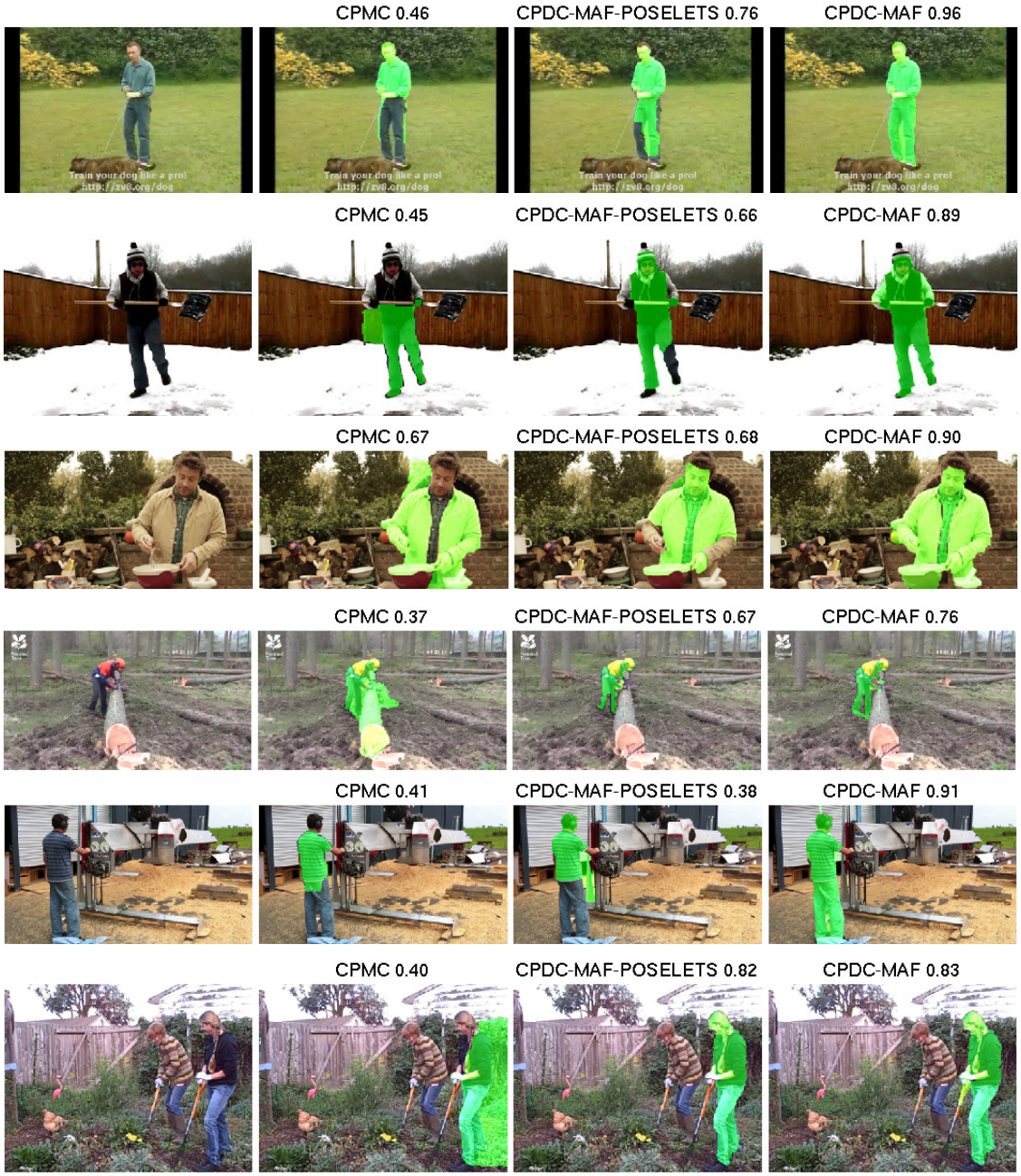

Figure 2-2: Segmentation examples for various methods. From left to right, original image, CPMC with default settings on person's bounding box, and our methods, CPDC-MAF-POSELETS and CPDC-MAF. See also tables 2.1 for quantitative results.

Chapter 3

Large-Scale Data-Dependent Kernel Approximation

In chapter *Large-Scale Data-Dependent Kernel Approximation*, we focus on scalability in learning models for continuous output prediction. We develop a formulation based on random Fourier feature approximations for kernel methods in the context of semi-supervised learning. The model is principled, supported by both theoretical and empirical studies, and we show it is effective for 3d human pose estimation under weak supervision. Learning a computationally efficient kernel from data is an important machine learning problem. The majority of kernels in the literature do not leverage the geometry of the data, and those that do are computationally infeasible for contemporary datasets. Recent advances in approximation techniques have expanded the applicability of the kernel methodology to scale linearly with the data size. Data-dependent kernels [33], which could leverage this computational advantage, have however not yet seen the benefit. Here we derive an *approximate large-scale learning procedure for data-dependent kernels* that is efficient and performs well in practice.

Our method avoids dealing with Gram (or kernel) matrices directly, in the way kernel methods do. We propose an approximation for the data-dependent kernel [33], in a formulation that multiplies the random Fourier features of the data-independent kernel, obtained with [24], by a weighted covariance matrix built using both labeled and unlabeled data. This effectively warps the distances between the Fourier features and therefore, we sometimes

Pose error (mm)		Labeled vs. unlabeled split	
RFKRR	LapRFKRR	# of Labeled	# of Unlabeled
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

Table 3.1: Performance evaluation for Human3.6M, with different splits of labeled and unlabeled data. The RFKRR column gives the performance of models trained using only labeled data while LapRFKRR uses both labeled and unlabeled data within the data-dependent kernel approximation setup. As the size of labeled data decreases, the performance of RFKRR decreases as well. However, we obtain improvements in the semi-supervised learning setting, thus demonstrating both the scalability and the advantage of using data-dependent kernel approximations.

refer to it as the 'warping' matrix. Our resulting data-dependent kernel approximation has the same properties as the one in [33], but no longer suffers from the memory and time constraints associated with building the Gram matrix linked with the kernel function. The construction is made possible by an astute application of a Woodbury identity which moves the learning problem from a reproducing kernel Hilbert space (RKHS) to the Fourier space of the kernel, reducing the computational load from $O(N^3)$ to $O(N)$. We provide a Lemma that can be used to derive the asymptotic convergence of the approximation in the limit of infinite random features, and, under certain conditions, an estimate of the convergence speed. We empirically prove that our construction represents a valid, yet efficient approximation of the data-dependent kernel.

For the large-scale empirical study, we consider a 3D human pose estimation problem based on 2D image information. We run our experiments on the very large Human 3.6M dataset [17], where we sample a subset of 1,055,424 poses for training and 56,860 poses for testing. We examine the following learning task: given the 2D pose, learn a model which is able to estimate the corresponding 3D pose. Thus, our input data consists of 2D human body joint positions and the target data of the corresponding 3D joint positions. We normalize the 2D pose data by setting the origin of the coordinate system in the pelvis joint. Also, we rotate each 2D pose such that the neck pelvis axis would align with the OY axis and scale it such that the average limb size would be 1. Our learning model is kernel ridge regression as

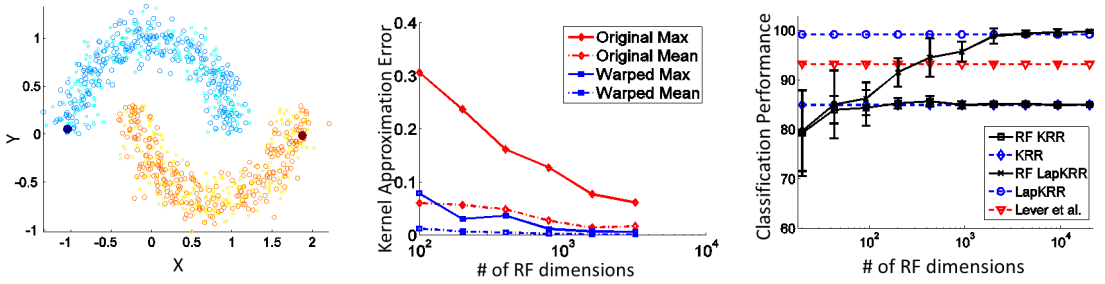

Figure 3-1: (Left) 'Two moons" dataset. (Middle) kernel approximation errors (max and average absolute error) for the original kernel, K, and warped (data-dependent) kernel, $\widetilde{K}$. (Right) classification performance on the two moons dataset, with 1 example per class, plotted against the dimensionality of the Fourier features. Note that the semi-supervised extension delivers some performance boost even with a very poor approximation (500 dimensions). With 2,000 RF features the performance is the same as the exact LapKRR model. We also include a comparison with the method of [23]. For their approximation method we use 500 landmark points.

it is simple and demonstrates the use of kernel methods. We choose the radial basis function kernel approximation for our problem due to the nature of the data. The random features approximation is based on $d = 4,000$ dimensions. By default, the entire dataset is fully labeled with both 2D and 3D information. For the semi-supervised problem, we consider 3D pose to be missing for some of the data, according to different splits. The performance of the model is illustrated in table 3.1. Please note, that during this experiment we varied the ratio between labeled and unlabeled data, keeping the total number of data points used. The reason behind this is that we want to see the impact of the labeled data, given that the quantity of unlabeled data is dense ($\simeq 1,000,000$). The purpose of this experiment is to empirically illustrate that the proposed data-dependent kernel approximation improves 3D pose estimation in a non-trivial, semi supervised learning scenario, where we work with large-scale datasets of over 1 million elements.

Chapter 4

Human Appearance Synthesis

In the final technical chapter of the thesis, *Human Appearance Transfer*, we introduce and explore the new problem of person-to-person appearance transfer, defined formally as transferring the appearance between two different people, in different poses, and differently dressed, given a single image of each. For this task, we propose a solution that combines 3d geometric modelling and deep convolutional neural networks, which on average produce human appearance synthesis results of photorealistic quality.

Given a pair of RGB images – source and target, denoted by I_s and I_t, each containing a person –, the main objective of our work is to transfer the appearance of the person from I_s into the body configuration of the person from I_t, resulting in a new image $I_{s \Rightarrow t}$.[1] Our proposed pipeline is shown in fig. 4-1.

Our solution to this new problem is formulated in terms of a computational pipeline that combines (1) 3d human pose and body shape estimation from monocular images, (2) identifying 3d surface colors elements (mesh triangles) visible in both images, that can be transferred directly using barycentric procedures, and (3) predicting surface appearance missing in the first image but visible in the second one using deep learning-based image synthesis techniques. Our work relies on 2d human detection and body part labeling [6, 31, 16], 3d pose estimation [31, 3, 35], parametric 3d human shape modeling [13, 39, 29, 28], procedures devoted to the semantic segmentation of clothing [32, 14, 27, 15], as well as image translation and synthesis methods [19, 8, 30, 22, 43, 8, 40].

[1] The procedure is symmetric, as we can transfer in both directions.

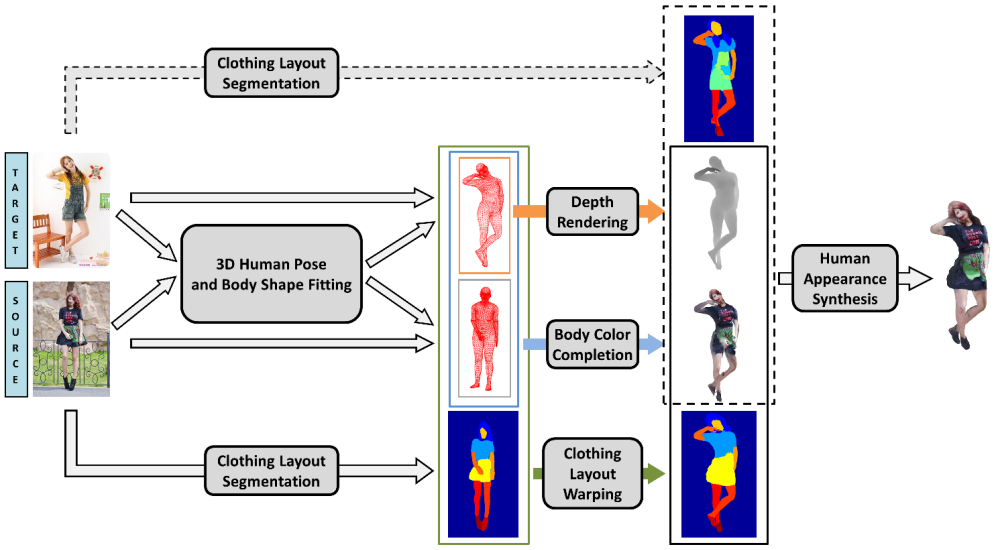

Figure 4-1: Human appearance transfer pipeline. Given only a single source and a single target image, each containing a person, with different appearance and clothing, and in different poses, our goal is to photo-realistically transfer the appearance from the source image onto the target image while preserving the target shape and clothing segmentation layout. The problem is formulated in terms of a computational pipeline that combines *(i)* 3d human pose and body shape fitting from monocular images, together with *(ii)* identifying 3d surface colors corresponding to mesh triangles visible in both images, that can be transferred directly using barycentric procedures, *(iii)* predicting surface appearance missing in the target image but visible in the source one using deep learning image synthesis techniques – these will be combined using the Body Color Completion Module. The last step, *(iv)*, takes the previous output together with the clothing layout of the source image warped on the target image (Clothing Layout Warping) and synthesizes the final output. If the clothing source layout is similar to the target, we bypass the warping step and use the target clothing layout instead.

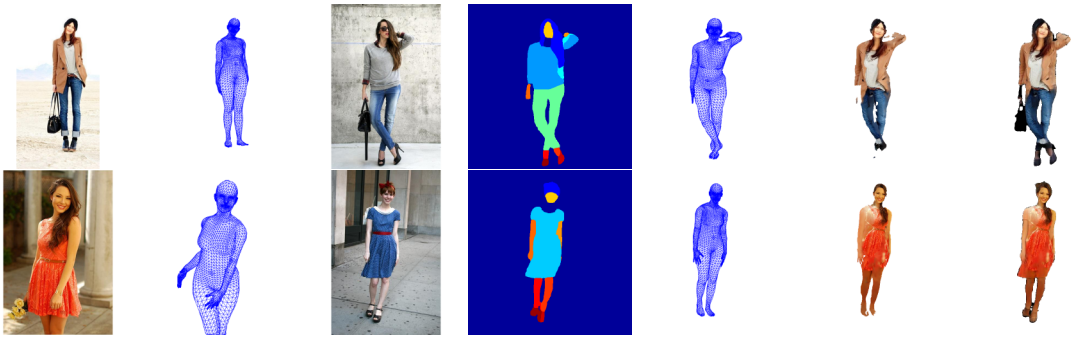

Figure 4-2: Sample results for the human appearance transfer pipeline. From left to right: source image with corresponding fitted 3d body model, target image with its clothing layout and fitted 3d body model, RGB data generated using the Body Color Completion module (*i.e.* $I_{s\to t}$), RGB data generated using the Human Appearance Synthesis module (*i.e.* $I_{s\Rightarrow t}$).

Our model achieves promising results as supported by a perceptual user study where the participants rated around 65% of our results as good, very good or perfect, as well in automated tests (Inception scores and a Faster-RCNN human detector responding very similarly to real and model generated images). We further show how the proposed architecture can be profiled to automatically generate images of a person dressed with different clothing transferred from a person in another image, opening paths for applications in entertainment and photo-editing (*e.g.* embodying and posing as friends or famous actors), the fashion industry, or affordable online shopping of clothing.

For all of our experiments we use the Chictopia10k dataset [26]. The images in this dataset depict different people, under both full and partial viewing, captured frontally. The high variability in color, clothing, illumination and pose makes this dataset suitable for our task. There are $17,706$ images available together with additional ground truth clothing segmentations. We do not use the clothing labels provided, but only the figure-ground segmentation such that we can generate training images cropped on the human silhouette. Sample results of our pipeline can be seen in figure 4-2.

Bibliography

[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.