Fully Convolutional Network with Superpixel Parsing for Fashion Web Image Segmentation

Lixuan Yang, Helena Rodriguez, Michel Crucianu, Marin Ferecatu

Fully convolutional network with superpixel parsing for fashion Web image segmentation

Lixuan Yang[1,2], Helena Rodriguez[2], Michel Crucianu[1], and Marin Ferecatu[1]

[1] Conservatoire National des Arts et Metiers,
292 Rue Saint-Martin, 75003 Paris, France
firstname.lastname@cnam.fr
[2] Shopedia SAS,
16 rue des blancs manteaux, 75004 Paris, France
firstname.lastname@shopedia.fr

Abstract. In this paper we introduce a new method for extracting deformable clothing items from still images by extending the output of a Fully Convolutional Neural Network (FCN) to infer context from local units (superpixels). To achieve this we optimize an energy function, that combines the large scale structure of the image with the local low-level visual descriptions of superpixels, over the space of all possible pixel labelings. To assess our method we compare it to the unmodified FCN network used as a baseline, as well as to the well-known Paper Doll and Co-parsing methods for fashion images.

Keywords: Clothing extraction, Semantic segmentation, FCN, Superpixel parsing

1 Introduction and related work

Although the interest in developing dedicated search engines for fashion databases is several decades old [18], the field only started to develop with the recent massive proliferation of fashion web-stores and online retail shops. Indeed, more and more users expect online advertising to propose items that truly correspond to their expectations in terms of design, manufacturing and suitability. Localizing, extracting, describing and tracking fashion items during web browsing allows professionals to better understand the users' preferences and design web interfaces that make for them a better web shopping experience.

In this work we propose a method to extract, without user supervision, clothes and other fashion items from web images. This is an important building block in designing software systems that solve the problems inspired by the needs of professionals of online advertising and fashion media: present to the users relevant items from a database of clothes, based on the content of the web application they are consulting and its context of use. This goes far beyond the requirements of a search engine: the user is not asked to interact with any search interface or formulate a query, but instead she is accompanied by automatic suggestions presenting in a non-intrusive way a selection of products that are likely to interest her.

1.1 Related Work

Many recent research efforts focusing on fashion images deal with a quite different use case, that of meta search engines federating and comparing several online shops. These

efforts focus on improving existing search engines to help users find products that match their preferences while preserving the "browsing" aspect [5]. Online shops sometimes provide image tags for common visual attributes, such as color or pattern, but they usually form a proprietary, heterogeneous and non-standardized vocabulary, typically too small to characterize the visual diversity of desired clothing [30,7]. Moreover, in many cases users look for characteristics expressed through very subjective concepts to describe a style, a brand or a specific design. For this reason, recent research focused on the development of detection, recognition and search of fashion items based on visual characteristics [5,21,24].

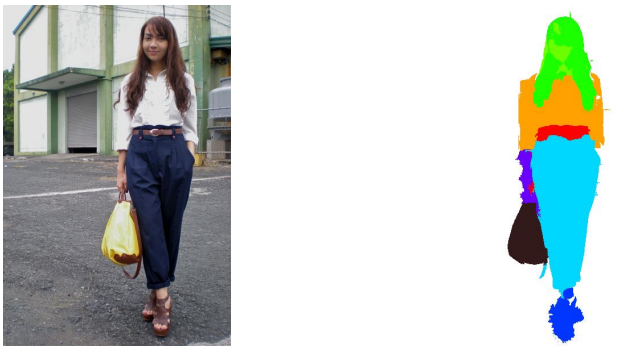
One of the approaches models the target item based on attribute selection and high-level classification [6]. For example, in [7] the authors train attribute classifiers on fine-grained clothing styles, formulating image retrieval as a classification problem; they rank items that contain the same visual attributes as the input, which can be a list of words or an image. A similar idea is explored in [13], where a set of features such as color, texture, SIFT features and object outlines are used to determine similarity scores between pairs of images. In [3], the authors propose to extract low-level features in a pose-adaptive manner and combine complementary features for learning attribute classifiers by employing conditional random fields (CRF) to explore mutual dependencies between the attributes. To narrow the semantic gap between the low-level features of clothing items and the high-level categories, in [24] it is proposed to adopt mid-level clothing attributes (e.g., clothing category, color, pattern) as a bridge. More specifically, the clothing attributes are treated as latent variables in a latent Support Vector Machine (SVM) recommendation model. To address larger fine-grained clothing attributes, [4] introduced a novel double-path deep domain adaptation network for attribute prediction by modeling the data jointly from unconstrained photos and the images issued from large-scale online shopping stores. To be able to propose a set of items to specific users, [14] put forward a functional tensor factorization method to model the interactions between users and fashion items.

A second approach consists in using part-based models to compensate for the lack of pose estimation and to model complex interactions in deformable objects [11]. To predict human occupations, in [33] part-based models are employed to characterize complex details and variable appearances of human clothing on the automatically aligned patches of human body parts, using sparse coding and noise-tolerant capacities. A similar part-based model is proposed in [29], where image patches are described by a mixture of color and texture features. Parts are also employed in [26] to reduce the "feature gap" caused by human pose discrepancy, by employing a graph parsing technique to align human parts for cloth retrieval.

Another approach relies on segmentation and aggregation to select different cloth categories. In [16], articulated pose estimation is followed by an over-segmentation of the relevant parts. Then, clustering by appearance creates a reference frame that facilitates rapid classification without requiring an actual training step. Body joints are incorporated in [15] by estimating their prior distributions and then learning the cloth–joint co-occurrences of different cloth types as part of a conditional random field framework to segment the image into different clothes. A similar idea is proposed in [36], where a CRF is formulated by inter-object or inter-attribute compatibility. Simo-Serra et al. [31]

have exploited another way to formulate a CRF by taking into account the garment's priors, 2D body pose condition and limb segments. The framework in [35] is based on bottom-up clothing parsing from semantic labels attached to each pixel. Local models of clothing items are learned on-the-fly from retrieved examples and parse mask predictions are transferred from these examples to the query image. Face detection is used in [37] to locate and track human faces in surveillance videos, then clothing is extracted by Voronoi partitioning to select seeds for region growing. For the video applications, [25] use SIFT Flow and superpixel matching to build correspondences across frames and exploit the cross-frame contexts to enhance human pose estimation. Also for human parsing and pose estimation, [8] proposed an unified framework to formulate the problem jointly via a tailored And-Or graph.

Recently, convolutional networks have started to achieve good performance in recognition and have also been deployed for semantic segmentation. For exemple, the fully convolutional network introduced in [27] improved the performance on the Pascal VOC database by using convolutional layers for pixel prediction and deconvolution layers for upsampling the result. DeepLab [22] improved on this by integrating a Markov Random Field into the upsampled network. In [39], conditional random fields were formulated as Recurrent Neural Networks and achieved good results on most classes of Pascal VOC [10]. Meanwhile, convolutional networks have also been used in fashion retrieval. Hadi et al. [17] employ three methods for street-to-shop retrieval, including two deep learning methods, and a deep similarity learning for the street and shop domains. In [34], the style retrieval uses Siamese CNN to transform features into a latent space. To achieve fast retrieval in a large scale dataset, [23] devised hashes-like representations learned by a latent layer added to the network during fine-tuning on the clothing dataset. Another emerging topic is *fashionability*: predict how fashionable a person looks. A conditional random field model that reasons about several fashionability factors by using deep network features was put forward in [32].



(a) Original image (b) Desired output (our result)

Fig. 1. Our goal is to produce a precise segmentation (extraction) of the fashion items as in (b).

1.2 Outline of our approach

Our proposal aims to segment *precisely* the object of interest from the background (foreground separation, see Fig. 1(b)), a difficult problem without user interaction and without using an extensive training database. Indeed, to propose meaningful results in terms of high-level expectations (such as product style or design) we need to achieve a good description of the visual appearance. For this, it is clearly much better if the object is segmented to eliminate the effect of mixing with the background and to let the description take into account the shape outline.

Most recent segmentation methods optimize a pixel-based objective function eventually followed by corrections relying on local edge behavior or segmentation templates. However, clothing items are deformable objects that can be encountered in a large variety of poses in real images. Extracting such objects requires an awareness of the object's global properties and context.

To take this into account, we extend the output of a Fully Convolutional Network (denoted by FCN in the following) by optimizing an objective function that iterates over all possible pixel-level labelings. The objective function considers the local adequacy with the class (label) under test, the global mid-scale structure of higher level units (superpixels), as well as the global smoothness of the labeling.

We test our method on the fashion image database CFPD [25] and we compare to the unmodified but fine-tuned FCN introduced in [27] that was shown to achieve state-of-the-art results on the Pascal VOC benchmark. We also compare to Co-parsing [25] and to the Paper-Doll framework [35] (see Sec. 1.1). We provide examples of successful segmentation, analyze difficult cases and evaluate each component of our framework. The rest of the paper is organized as follows. In Sec. 2 we give an outline of our proposal, followed by a detailed description of each component. An experimental validation including both quantitative and qualitative results is then presented in Sec. 3. We conclude the paper in Sec. 4 and provide perspectives for future work.

2 Our proposal

Detecting clothes in images is a difficult problem because the objects are deformable, have large intraclass diversity and may appear against complex backgrounds. To extract objects under these difficult conditions and without user intervention, methods solely relying on optimizing a local criterion (or pixel classification based on local features) are unlikely to perform well. Some knowledge about the global shape of the class of objects to be extracted is necessary to help a local analysis converge to a correct object boundary. In this paper we use this intuition to develop a framework that takes into account the local/global duality to select the most likely object segmentation.

As baseline we employ the output of the fully convolutional network [27] that was shown to perform very well on the Pascal VOC benchmark. For every pixel in a test image, the FCN provides scores for all the labels (object classes) in the training set.

However, convolutional networks tends to produce a smooth output to preserve the network's generality. Thus, the resulting scores can predict the presence and rough position of objects but are less well suited to detect the exact outline of the objects. To

address the localization challenge, some approaches use information from several convolution layers to better estimate the object boundaries [27,9]. Other approaches employ local representations, transforming the localization into a local optimization task [28].

In this paper we pursue an alternative path, by inserting before the softmax layer a fully connected conditional random field model. This idea has emerged recently [22,1,38]. The novelty of our proposal is that our model takes advantage of the middle-scale structure of the image by using superpixel co-localization. As we shall see, this helps the network recover object boundaries at a higher level of detail.

Let X be the test image, of size $W \times H$ pixels. For simplicity, we consider the pixels of the image are enumerated in linear fashion, from 1 to $N = W \times H$. The goal is to associate to each pixel one of the L classes (denoted by the labels $a_1, \ldots, a_L$) learnt by the FCN. Just before the output softmax layer, the FCN produces a vector of L scores for each pixel of the image. The FCN scores for the entire image are given by $F(X)$, the element of $F(X)$ corresponding to pixel i being the L-dimensional vector f_i. A labeling L of the test image consists in attaching a label l_i to each pixel $i \in 1 \ldots N$. The goal is then to maximize a likelihood function $P(L|X, F(X))$ over the space of all labelings L, or equivalently to minimize the corresponding energy:

$$\arg\min_L E(X) = \arg\min_L(-\log(P(L|X, F(X))))$$

The function $E(X)$ contains several terms aimed at preserving the predictive power of the FCN while improving the localization of the contours:

$$E(X) = \sum_{i=1}^{N} \lambda_{l_i}^{(f)} \theta_i^{(f)}(x_i) + \sum_{i=1}^{N} \lambda_{l_i}^{(r)} \theta_i^{(r)}(x_i) + \sum_{i,j=1,i>j}^{N} \theta_{ij}(x_i, x_j) \qquad (1)$$

Sums are over all the pixels of the image X. The first term, containing $\theta_i^{(f)}(x_i)$, encodes the degree of agreement between the produced output and the FCN scores, which is denoted by the superscript (f) in the equation. If only this term were present, the output would be the same as the one provided by the unmodified FCN. To allow for more flexibility, the terms in the sum are weighted by the parameters $\lambda_{l_i}^{(f)}$, where l_i is the label assigned by L to the pixel i. The best values for these parameters, including $\lambda_{l_i}^{(r)}$ from the second term, are found by using a Nelder-Mead simplex method as described in [20]. The second term in Eq. 1 encodes the agreement between the proposed labeling and the visual descriptions of the middle-scale image content units (superpixels). Finally, the third term is a smoothness measure based on the fact that nearby pixels with similar low-level features are likely to belong to the same class. We now provide a detailed description of each of these terms. See also Fig. 2 for an illustration of the results of different stages of our work chain.

Convolutional Term. The FCN [27] takes as input an image of arbitrary size and produces an output of the same size. The classifier transforms fully connected layers into convolutional layers to output a spatial classification map. To make a dense prediction, a deconvolutional layer upsamples the coarse outputs to pixelwise outputs. It employs a skip architecture by combining the final prediction layer with lower layers

with finer strides. The network is initialized by using a pre-trained model learnt on Pascal VOC [10] and then fine-tuned to the dataset employed here following a procedure that is similar to the one described in [2]. The output contains score predictions for each class and each pixel. The first term in Eq. 1 is given by the FCN pixel prediction: $\theta_i^{(f)}(x_i|l_i) = -\log f(i, l_i)$, where $f(i, l_i)$ is the FCN output for pixel i and label l_i.

Region/superpixel prediction. The second term in Eq. 1 encodes the level of agreement between the neighboring labels and the label of the current pixel. The image is first over-segmented in superpixels, following the idea that all pixels in a superpixel should be attributed the same label. This is reasonable because objects are in general delimited by physical contours, and are thus obtained as disjoint unions of superpixels. This procedure should thus improve the contour localization, which was one of the weaknesses of the original FCN. To extract the superpixels we use the well-known method from [12], also employed by other work on fashion retrieval [25]. For each superpixel we compute a single label, obtained by the Softmax procedure applied on the average of FCN class scores of the pixels in the region. Given a label l_i, let $N(l_i)$ be the set of all superpixels having this label. For a given superpixel S and for each possible label l_i we compute the agreement between the superpixels in $N(l_i)$ and S according to:

$$\phi(l_i|S) = \frac{1}{|N(l_i)|} \sum_{s \in N(l_i)} \frac{1}{1 + |h_S - h_s|^2} \cdot \frac{1}{1 + |p_S - p_s|^2} \tag{2}$$

where h_s denotes the low-level feature (see Sec. 3) of superpixel s, p_s is the barycenter of this superpixel and $\| \cdot \|$ the L_2 norm. The first component in Eq. 2 is larger for superpixels having similar low-level descriptions, implementing the idea that visually similar superpixels must share the same label. However, this behavior is weighted by the second term that is smaller for superpixels that are far away in the image. This filters the effect of similar superpixels that are far from S. The 1 is included in the denominator to protect against numerical instabilities.

If S_i is the superpixel to whom pixel i belongs and l_i is its candidate label, then the second term in Eq. 1 is computed using $\theta_i^{(r)}(x_i) = -\log \phi(l_i|S_i)$. To limit even further the influence of superpixels that are too far away from the candidate, we use the superpixels situated inside a circle of a given radius around the candidate. The best radius is likely to depend on the scale of the objects; in our case, by cross-validation we found a radius of 0.2 of the image size.

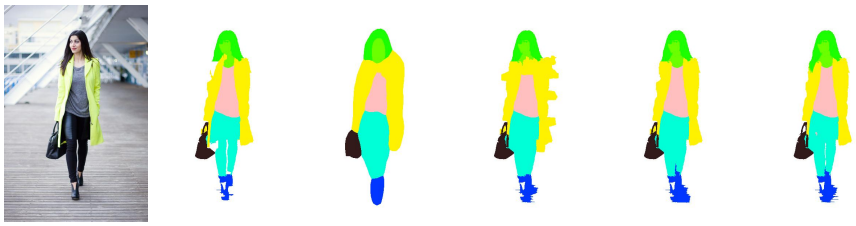
Smoothness term. The third term in Eq. 1 implements a smoothing condition: two pixels are more likely to share the same label if they have similar low-level visual features and are not very distant in the image, corresponding to the idea that objects are localized units in an image. We found that the following formulation, proposed in [19], works well for our purpose:

$$\theta_{ij}(x_i, x_j) = -\log(g_{ij}(x_i, x_j))$$

where

$$g_{ij}(x_i, x_j) = (1 - \delta_{ij})(w_1 \exp(-\frac{\|p_i - p_j\|^2}{2\delta_\alpha^2} - \frac{\|h_i - h_j\|^2}{2\delta_\beta^2}) + w_2 \exp(-\frac{\|p_i - p_j\|}{2\delta_\gamma^2}))$$

where p_i is the position in the image of pixel i, h_i its visual description (see Sec. 3) and δ_{ij} the Kronecker delta. The first part is an appearance kernel inspired by the observation that nearby pixels that are visually similar are likely to be in the same class. The second part is a smoothness kernel that helps removing small isolated regions. The values of $w_{1,2}$ and $\delta_{\alpha,\beta,\gamma}$ are obtained by cross-validation.



(a) (b) (c) (d) (e) (f)

Fig. 2. Original image (a), ground truth (b) and the stages of our approach: (c) softmax FCN, (d) softmax with superpixels labeled by FCN, (e) superpixel parsing, (f) final result.

Training. The FCN model is pre-trained on the Pascal VOC dataset [10] and then fine-tuned on the specific clothing dataset considered here. Fine-tuning is performed with a lower learning rate of 10^{-14} and a high momentum of 0.99. For FCN-32s we employ 200K iterations, then 100K iterations for FCN-16s and FCN-8s.

The CRF parameters λ_f and λ_r are obtained by optimization of the F_1 score on the validation set. Given the size of the search space ($2 \times L$), the Nelder-Mead simplex method [20] is employed.

3 Experimental results

The proposed method is evaluated on the Colorful Fashion Parsing Data (CFPD), put forward in [25]. This clothing dataset consists of 2,682 images with 23 class labels. We employed the same training and test partitions suggested in [25]. We defined a validation set consisting of 100 randomly selected images from the initial training set. As visual features for describing the superpixels we use the concatenation of normalized RGB and HSV joint histograms, each having 10 bins.

To assess the performance of the proposed method, we perform two sets of experiments. First, a class-by-class comparison is performed on CFPD with the recent deep network FCN [27] method for semantic segmentation. Then, the global performance of our proposal is compared to one of the Co-parsing [25] fashion item annotation method. We eventually provide a qualitative evaluation of our porposal.

Class-by-class comparison. The FCN [27] is used as a baseline method in the recent work on object segmentation. The FCN has significantly improved state-of-the-art results in semantic segmentation and has an open implementation. This makes it a good candidate as a baseline and supports a class-by-class comparison. For the purpose of this evaluation, we also fine-tuned the FCN on our training images and employ the `argmax` output of the FCN-8s net.

Table 1. Class-by-class comparison with the FCN [27] using the average F_1 score.

Class	background	T-shirt	Bag	Belt	Blazer	Blouse	Coat	Dress
FCN	95.38	24.30	29.02	10.09	15.44	23.76	13.32	35.08
Our Method	**96.67**	**28.73**	**34.83**	**12.87**	**18.80**	**26.94**	**16.48**	**39.44**
Class	Face	Hair	Hat	Jeans	Legging	Pants	Scarf	Shoe
FCN	46.03	45.11	17.02	26.99	20.77	24.94	11.34	35.40
Our Method	**49.71**	**54.82**	**21.59**	**31.38**	**24.61**	**28.22**	**13.95**	**43.35**
Class	Shorts	Skin	Skirt	Socks	Stocking	Sunglass	Sweater	–
FCN	38.07	43.48	41.90	9.48	28.17	2.43	11.65	–
Our Method	**46.35**	**51.35**	**48.23**	**10.91**	**34.79**	**2.75**	**13.39**	–

Table 2. Comparison of Paper Doll [35], Co-parsing [25], FCN [27] and our method.

Mesure	Accuracy	FG. Accuracy	Avg. Precision	Avg. Recall	Avg. F_1
Paper Doll [35]	82.79	44.08	49.20	32.00	32.66
Co-Parsing [25]	84.7	52.49	42.31	42.31	41.42
FCN [27]	86.09	50.62	47.06	51.13	40.29
CRF w.o. superpixels	86.77	49.87	50.21	49.64	40.45
Our method	**88.69**	**55.69**	**53.40**	**56.93**	**45.91**

In Table 1 we show a class-by-class comparison between the proposed method and the FCN (best results are in boldface). As performance measure we employ the average F_1 score over pixels, further averaged over all the testing images of each class. It can be seen that the proposed method performs significantly better on all the classes. While both segmentation methods are automatic (do not require any interaction on test images), these results speak in favor of better taking into account local image information into the algorithm. In our case this is achieved by parsing superpixels.

Global comparison. We also have to compare our proposal to existing methods that were specifically developed for labelling or extracting fashion items from images. Two prominent frameworks are Paper Doll [35] and Co-parsing [25]. To validate our proposed new term, the algorithm without the superpixel term (second term in Eq. 1) was also tested (denoted "CRF w.o. superpixels" in Table 2). The authors of Paper Doll had introduced the Fashionista database of 685 images that was used to test annotation algorithms. However, this database only contains 456 training images, which is quite small for fine-tuning the FCN, and the classes are not exactly the same, so we did not evaluate on Fashionista but only on CFPD.

Table 2 presents a synthesis of the results obtained on CFPD by Paper Doll [35], Co-parsing [25], FCN [27], CRF w.o. superpixels and by the proposed method. Note that the results for Paper Doll come from [25]. Several performance measures are shown: the accuracy, the foreground accuracy, the average precision, the average recall and the average F_1 score, traditionally used for fashion segmentation evaluation. The measures are averaged over pixels, over all the testing images of each class and over classes. As seen from Table 1, the proposed method compares favorably to the other methods according to all the performance measures considered.

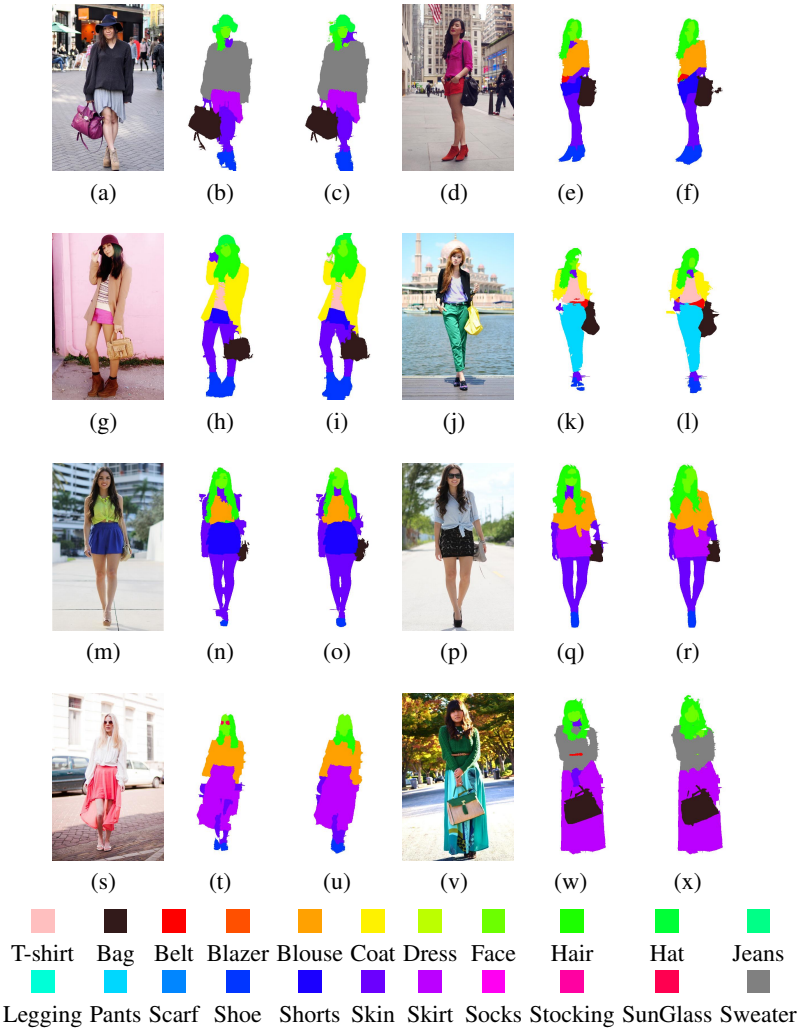


Fig. 3. Qualitative evaluation: original images (a, d, g, j, m, p, s, v), ground truth (b,e, h, k, n, q, t, w) and associated segmentation results (c, f, i, l, o, r, u, x).

Qualitative evaluation. In Fig. 3 we present some difficult but quite successful segmentations: (a, d) for clothes against confusing or cluttered background, (g, j) for deformed clothes (opened jacket) and (m, p) for small object extraction (shoes). Some parts of our results show that the ground truth is not perfect and an automatic segmentation method can do better. Fig. 3 also shows examples where the proposed method is not perfect: when comparing to the ground truth, in (s) it failed to detect the sunglasses and in (v) it failed to detect the belt and the skin (neck). This reflects the lower F_1 score in Table 2 for sunglasses and belt. Small objects are quite difficult to extract and may require a specific setting, including *e.g.* a higher penalty during training.

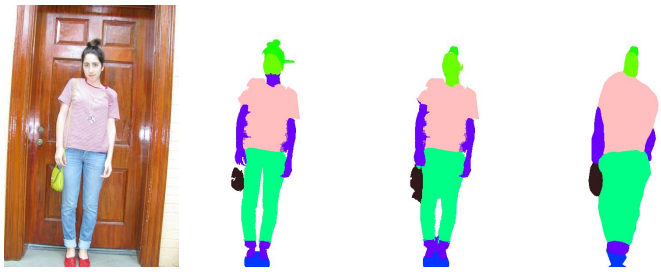


Fig. 4. Comparison with FCN: original image (first), ground truth (second), our method (third) and FCN (last).

A visual comparison with the results of the FCN is also shown in Fig. 4. As hinted by the quantitative results, the FCN leads to an excessive smoothing and the segmented clothes include larger parts of external objects. This occurs on most of the images in the database, explaining the lower performance of FCN in Table 1 and Table 2.

4 Conclusion

To extract clothing objects from web images, we propose to exploit both the output of a Fully Convolutional Neural Network (FCN) used for semantic segmentation and the superpixels obtained from local visual information. By bridging the high-level prediction provided by the deep network and the mid-level image description, the proposed method significantly improves contour localization. The proposed approach is validated by comparisons with the (fine-tuned) FCN alone and to the Co-parsing [25] method, arguably the current state-of-the-art in fashion item extraction.

The results can probably be further improved by the use of more training data and of refined visual features for the superpixels. To better extract small objects we intend to relate the penalty to the relative size of the objects. Also, we believe that some confusion is inevitable between specific classes, *e.g.* legging *vs.* pants or blouse *vs.* sweater, but this may not be a problem for subsequent similarity-based clothing retrieval if object segmentation is correct. The proposed method can easily be extended to other classes at relatively low cost, *i.e.* by manually annotating objects from these new classes to train the FCN and the CRF.

References

1. Bell, S., Upchurch, P., Snavely, N., Bala, K.: Material recognition in the wild with the materials in context database. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3479 – 3487 (2015) 5
2. Bengio, Y.: Deep learning of representations for unsupervised and transfer learning. In: JMLR W&CP: Proc. Unsupervised and Transfer Learning challenge and workshop. vol. 27, pp. 17–36 (2012) 6
3. Chen, H., Gallagher, A., Girod, B.: Describing clothing by semantic attributes. In: Proceedings of the 12th European Conference on Computer Vision - Volume Part III. pp. 609–623. ECCV'12, Berlin, Heidelberg (2012) 2

4. Chen, Q., Huang, J., Feris, R., Brown, L.M., Dong, J., Yan, S.: Deep domain adaptation for describing people based on fine-grained clothing attributes. In: CVPR. pp. 5315–5324. IEEE Computer Society, Boston, MA (2015) 2
5. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40(2), 5:1–5:60 (May 2008) 2
6. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: An experimental comparison. Inf. Retr. 11(2), 77–107 (Apr 2008) 2
7. Di, W., Wah, C., Bhardwaj, A., Piramuthu, R., Sundaresan, N.: Style finder: Fine-grained clothing style detection and retrieval. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 8–13. CVPRW '13, IEEE Computer Society, Washington, DC, USA (2013) 2
8. Dong, J., Chen, Q., Shen, X., Yang, J., Yan, S.: Towards unified human parsing and pose estimation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 843–850. CVPR '14, Washington, DC, USA (2014) 3
9. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. vol. abs/1411.4734 (2014) 5
10. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal visual object classes (VOC) challenge. International Journal of Computer Vision 88(2), 303–338 (Jun 2010) 3, 6, 7
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. 32(9), 1627–1645 (Sep 2010) 2
12. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. Int. J. Comput. Vision 59(2), 167–181 (Sep 2004) 6
13. Hsu, E., Paz, C., Shen, S.: Clothing image retrieval for smarter shopping (Stanford project) (2011) 2
14. Hu, Y., Yi, X., Davis, L.S.: Collaborative fashion recommendation: A functional tensor factorization approach. In: Proceedings of the 23rd ACM International Conference on Multimedia. pp. 129–138. MM '15, ACM, New York, NY, USA (2015) 2
15. Jammalamadaka, N., Minocha, A., Singh, D., Jawahar, C.V.: Parsing clothes in unrestricted images. In: British Machine Vision Conference, BMVC 2013, Bristol, UK, September 9-13, 2013 (2013) 2
16. Kalantidis, Y., Kennedy, L., Li, L.J.: Getting the look: Clothing recognition and segmentation for automatic product suggestions in everyday photos. In: Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval. pp. 105–112. ICMR '13, ACM, New York, NY, USA (2013) 2
17. Kiapour, M.H., Han, X., Lazebnik, S., Berg, A.C., Berg, T.L.: Where to buy it: Matching street clothing photos in online shops. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). pp. 3343–3351. ICCV '15, IEEE Computer Society, Washington, DC, USA (2015) 3
18. King, I., Lau, T.K.: A feature-based image retrieval database for the fashion, textile, and clothing industry in Hong Kong. In: Intl. Symp. on Multi-Technology Inf. Proc. (ISMIP'96), Hsin-Chu, Taiwan. pp. 233–240 (1996) 1
19. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in Neural Information Processing Systems 24. pp. 109–117. Curran Associates, Inc. (2011) 6
20. Lagarias, J.C., Reeds, J.A., Wright, M.H., Wright, P.E.: Convergence properties of the nelder–mead simplex method in low dimensions. SIAM J. on Optimization 9(1), 112–147 (May 1998) 5, 7

21. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Trans. Multimedia Comput. Commun. Appl. 2(1), 1–19 (Feb 2006) 2
22. Liang-Chieh, C., Papandreou, G., Kokkinos, I., murphy, k., Yuille, A.: Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In: International Conference on Learning Representations. San Diego, United States (May 2015) 3, 5
23. Lin, K., Yang, H.F., Liu, K.H., Hsiao, J.H., Chen, C.S.: Rapid clothing retrieval via deep learning of binary codes and hierarchical search. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. pp. 499–502. New York, USA (2015) 3
24. Liu, S., Feng, J., Song, Z., Zhang, T., Lu, H., Xu, C., Yan, S.: Hi, magic closet, tell me what to wear! In: Proceedings of the 20th ACM International Conference on Multimedia. pp. 619–628. MM '12, ACM, New York, NY, USA (2012) 2
25. Liu, S., Liang, X., Liu, L., Lu, K., Lin, L., Yan, S.: Fashion parsing with video context. In: Proceedings of the 22Nd ACM International Conference on Multimedia. pp. 467–476. MM '14, ACM, New York, NY, USA (2014) 3, 4, 6, 7, 8, 10
26. Liu, S., Song, Z., Wang, M., Xu, C., Lu, H., Yan, S.: Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In: Proceedings of the 20th ACM International Conference on Multimedia. pp. 1335–1336. MM '12, ACM, New York, NY, USA (2012) 2
27. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. vol. abs/1411.4038 (2014) 3, 4, 5, 7, 8
28. Mostajabi, M., Yadollahpour, P., Shakhnarovich, G.: Feedforward semantic segmentation with zoom-out features. vol. abs/1412.0774 (2014) 5
29. Nguyen, T.V., Liu, S., Ni, B., Tan, J., Rui, Y., Yan, S.: Sense beauty via face, dressing, and/or voice. In: Proceedings of the 20th ACM International Conference on Multimedia. pp. 239–248. MM '12, ACM, New York, NY, USA (2012) 2
30. Redi, M.: Novel methods for semantic and aesthetic multimedia retrieval. Ph.D. thesis, Univ. Nice, Sophia Antipolis (2013) 2
31. Simo-Serra, E., Fidler, S., Moreno-Noguer, F., Urtasun, R.: A high performance CRF model for clothes parsing. In: Computer Vision - ACCV 2014 - 12th Asian Conference on Computer Vision, Singapore. pp. 64–81 (2014) 3
32. Simo-Serra, E., Fidler, S., Moreno-Noguer, F., Urtasun, R.: Neuroaesthetics in fashion: Modeling the perception of fashionability. In: CVPR (2015) 3
33. Song, Z., Wang, M., sheng Hua, X., Yan, S.: Predicting occupation via human clothing and contexts. In: Proceedings of the 2011 International Conference on Computer Vision. pp. 1084–1091. ICCV '11, IEEE Computer Society, Washington, DC, USA (2011) 2
34. Veit*, A., Kovacs*, B., Bell, S., McAuley, J., Bala, K., Belongie, S.: Learning visual clothing style with heterogeneous dyadic co-occurrences. In: International Conference on Computer Vision (ICCV). Santiago, Chile (2015), *The first two authors contributed equally 3
35. Yamaguchi, K., Hadi, K., Luis, E., Tamara, L.B.: Retrieving similar styles to parse clothing. IEEE TPAMI 37, 1028–1040 (2015) 3, 4, 8
36. Yamaguchi, K., Okatani, T., Sudo, K., Murasaki, K., Taniguchi, Y.: Mix and match: Joint model for clothing and attribute recognition. In: Proceedings of the British Machine Vision Conference (BMVC). pp. 51.1–51.12. BMVA Press (September 2015) 3
37. Yang, M., Yu, K.: Real-time clothing recognition in surveillance videos. In: ICIP. pp. 2937–2940. ICIP '11, IEEE (2011) 3
38. Zhang, N., Donahue, J., Girshick, R.B., Darrell, T.: Part-based r-cnns for fine-grained category detection. vol. abs/1407.3867 (2014) 5
39. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.S.: Conditional random fields as recurrent neural networks. vol. abs/1502.03240 (2015) 3