Human Parsing with Contextualized Convolutional Neural Network

Xiaodan Liang[1,2], Chunyan Xu[2], Xiaohui Shen[3], Jianchao Yang[5], Si Liu[6], Jinhui Tang[4]
Liang Lin[1]*, Shuicheng Yan[2]
[1] Sun Yat-sen University [2] National University of Singapore [3] Adobe Research
[4] Nanjing University of Science and Technology [5] Snapchat Research
[6] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences

Abstract

In this work, we address the human parsing task with a novel Contextualized Convolutional Neural Network (Co-CNN) architecture, which well integrates the cross-layer context, global image-level context, within-super-pixel context and cross-super-pixel neighborhood context into a unified network. Given an input human image, Co-CNN produces the pixel-wise categorization in an end-to-end way. First, the cross-layer context is captured by our basic local-to-global-to-local structure, which hierarchically combines the global semantic information and the local fine details across different convolutional layers. Second, the global image-level label prediction is used as an auxiliary objective in the intermediate layer of the Co-CNN, and its outputs are further used for guiding the feature learning in subsequent convolutional layers to leverage the global image-level context. Finally, to further utilize the local super-pixel contexts, the within-super-pixel smoothing and cross-super-pixel neighbourhood voting are formulated as natural sub-components of the Co-CNN to achieve the local label consistency in both training and testing process. Comprehensive evaluations on two public datasets well demonstrate the significant superiority of our Co-CNN over other state-of-the-arts for human parsing. In particular, the F-1 score on the large dataset [15] reaches 76.95% *by Co-CNN, significantly higher than* 62.81% *and* 64.38% *by the state-of-the-art algorithms, M-CNN [21] and ATR [15], respectively.*

1. Introduction

Human parsing, which refers to decomposing a human image into semantic clothes/body regions, is an important component for general human-centric analysis. It enables many higher level applications, e.g., clothing style recognition and retrieval [5], clothes recognition and retrieval [30], people re-identification [33], human behavior analysis [29] and automatic product recommendation [14].

While there has been previous work devoted to human parsing based on human pose estimation [31] [6] [32] [20] [19], non-parametric label transferring [30][21] and active template regression [15], none of previous methods has achieved excellent dense prediction over raw image pixels in a fully end-to-end way. These previous methods often take complicated preprocessing as the requisite, such as reliable human pose estimation [4], bottom-up hypothesis generation [1] and template dictionary learning [23], which makes the system vulnerable to potential errors of the front-end preprocessing steps.

Convolutional neural network (CNN) facilitates great advances not only in whole-image classification [26], but also in structure prediction such as object detection [10] [16], part prediction [27] and general object/scene semantic segmentation [7][8]. However, they usually need supervised pre-training with a large classification dataset, e.g., ImageNet, and other post-processing steps such as Conditional Random Field (CRF) [8] and extra discriminative classifiers [24][11]. Besides the above mentioned limitations, there are still two technical hurdles in the application of existing CNN architectures to pixel-wise prediction for the human parsing task. First, diverse contextual information and mutual relationships among the key components of human parsing (i.e. semantic labels, spatial layouts and shape priors) should be well addressed during predicting the pixel-wise labels. For example, the presence of a skirt will hinder the probability of labeling any pixel as the dress/pants, and meanwhile facilitate the pixel prediction of left/right legs. Second, the predicted label maps are desired to be detail-preserved and of high-resolution, in order to recognize or highlight very small labels (e.g. sunglass or belt). However, most of the previous works on semantic segmentation with CNN can only predict the very low-resolution labeling, such

*Corresponding author is Liang Lin (E-mail: linliang@ieee.org). This work was done when the first author worked as an intern in National University of Singapore.

as eight times down-sampled prediction in the fully convolutional network (FCN) [22]. Their prediction is very coarse and not optimal for the required fine-grained segmentation.

In this paper, we present a novel Contextualized Convolutional Neural Network (Co-CNN) that successfully addresses the above mentioned issues. Given an input human image, our architecture produces the correspondingly-sized pixel-wise labeling maps in a fully end-to-end way, as illustrated in Figure 1. Our Co-CNN aims to simultaneously capture cross-layer context, global image-level context and local super-pixel contexts by using the local-to-global-to-local hierarchical structure, global image-level label prediction, within-super-pixel smoothing and cross-super-pixel neighborhood voting, respectively.

First, our basic local-to-global-to-local structure hierarchically encodes the local details from the early, fine layers and the global semantic information from the deep, coarse layers. Four different spatial resolutions are used for capturing different levels of semantic information. The feature maps from deep layers often focus on the global structure and are insensitive to local boundaries and spatial displacements. We up-sample the feature maps from deep layers and then combine them with the feature maps from former layers under the same resolution.

Second, to utilize the global image-level context and guarantee the coherence between pixel-wise labeling and image label prediction, we incorporate global image label prediction into our pixel-wise categorization network, illustrated as the *global image-level context* part of Figure 1. An auxiliary objective defined for the global image label prediction (i.e. Squared Loss) is used, which focuses on global semantic information and has no relation with local variants such as pose, illumination or precise location. We then use the predicted image-level label probabilities to guide the feature learning from two aspects. First, the predicted image-level label probabilities are utilized to facilitate the feature maps of each intermediate layer to generate the semantics-aware feature responses, and then the combined feature maps are further convolved by the filters in the subsequent layers, shown as the *image label concatenation* part of Figure 1. Second, the predicted image-level label probabilities are also used in the prediction layer to explicitly re-weight the pixel-wise label confidences, shown as the *element-wise summation* part of Figure 1.

Finally, the within-super-pixel smoothing and cross-super-pixel neighborhood voting are leveraged to retain the local boundaries and label consistencies within the super-pixels. They are formulated as natural sub-components of the Co-CNN in both the training and the testing process.

Comprehensive evaluations and comparisons on the ATR dataset [15] and the Fashionista dataset [30] well demonstrate that our Co-CNN yields results that significantly surpass all previously published methods, boosting the current state-of-the-arts from 64.38% [15] to 76.95%. We also build a much larger dataset "Chictopia10k", which contains 10,000 annotated images. By adding the images of "Chictopia10k" into the training, the F-1 score can be further improved to 80.14%, 15.76% higher than the state-of-the-arts [15] [30].

2. Related Work

Human Parsing: Much research has been devoted to human parsing [31][30][6][32][28][18][25][21]. Most previous works used the low-level over-segmentation, pose estimation and bottom-up hypothesis generation as the building blocks of human parsing. For example, Yamaguchi et al. [31] performed human pose estimation and attribute labeling sequentially and then improved clothes parsing with a retrieval-based approach [30]. Dong et al. [6] proposed to use a group of parselets under the structure learning framework. These traditional hand-crafted pipelines often require many hand-designed processing steps, each of which needs to be carefully designed and tuned. Recently, Liang et al. [15] proposed to use two separate convolutional networks to predict the template coefficients for each label mask and their corresponding locations, respectively. However, their design may lead to sub-optimal results.

Semantic Segmentation with CNN: Our method works directly on the pixel-level representation, similar to some recent research on semantic segmentation with CNN [22] [11] [8]. These pixel-level representations are in contrast to the common two-stage approaches[24] [10] [12] which consist of complex bottom-up hypothesis generation (e.g. bounding box proposals) and CNN-based region classification. For the pixel-wise representation, by directly using CNN, Farabet et al. [7] trained a multi-scale convolutional network from raw pixels and employed the super-pixel tree for smoothing. The dense pixel-level CRF was used as the post-processing step after CNN-based pixel-wise prediction [2] [3] [9]. More recently, Long et al. [22] proposed the fully convolutional network for predicting pixel-wise labeling.

The main difference between our Co-CNN and these previous methods is the integration of cross-layer context, global image-level context, local super-pixel contexts into a unified network. It should be noted that while the fully convolutional network [22] also tries to combine coarse and fine layers, they only aggregate the predictions from different scales in the final output. In contrast, in our local-to-global-to-local hierarchical structure , we hierarchically combine feature maps from cross-layers and further feed them into several subsequent layers for better feature learning, which is very important in boosting the performance as demonstrated in the experiments.

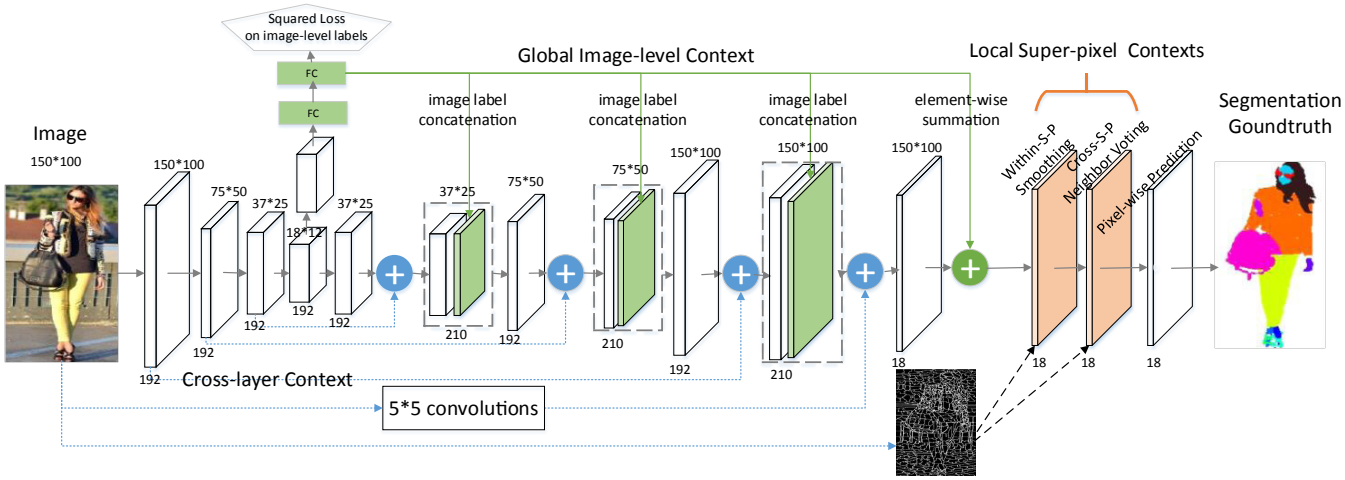

Figure 1. Our Co-CNN integrates the cross-layer context, global image-level context and local super-pixel contexts into a unified network. It consists of cross-layer combination, global image-level label prediction, within-super-pixel smoothing and cross-super-pixel neighborhood voting. First, given an input 150×100 image, we extract the feature maps for four resolutions (i.e., 150×100, 75×50, 37×25 and 18×12). Then we gradually up-sample the feature maps and combine the corresponding early, fine layers (blue dash line) and deep, coarse layers (blue circle with plus) under the same resolutions to capture the cross-layer context. Second, an auxiliary objective (shown as "Squared loss on image-level labels") is appended after the down-sampling stream to predict global image-level labels. These predicted probabilities are then aggregated into the subsequent layers after the up-sampling (green line) and used to re-weight pixel-wise prediction (green circle with plus). Finally, the within-super-pixel smoothing and cross-super-pixel neighborhood voting are performed based on the predicted confidence maps (orange planes) and the generated super-pixel over-segmentation map to produce the final parsing result. Only down-sampling, up-sampling, and prediction layers are shown; intermediate convolution layers are omitted. For better viewing of all figures in this paper, please see original zoomed-in color pdf file.

3. The Proposed Co-CNN Architecture

Our Co-CNN exploits the cross-layer context, global image context and local super-pixel contexts in a unified network, consisting of four components, i.e., the local-to-global-to-local hierarchy, global image label prediction, within-super-pixel smoothing and cross-super-pixel neighborhood voting, respectively.

3.1. Local-to-global-to-local Hierarchy

Our basic local-to-global-to-local structure captures the cross-layer context. It simultaneously considers the local fine details and global structure information. The input to our Co-CNN is a 150×100 color image and then passed through a stack of convolutional layers. The feature maps are down-sampled three times by the max pooling with a stride of 2 pixels to get three extra spatial resolutions (75×50, 37×25, 18×12), shown as the four early convolutional layers in Figure 1. Except for the stride of 2 pixels for down-sampling, the convolution strides are all fixed as 1 pixel. The spatial padding of convolutional layers is set so that the spatial resolution is preserved after convolution, e.g., the padding of 2 pixels for 5×5 convolutional filters.

Note that the early convolutional layers with high spatial resolutions (e.g., 150×100) often capture more local details while the ones with low spatial resolutions (e.g., 18×12) can capture more structure information with high-level semantics. We combine the local fine details and the high-level structure information by cross-layer aggregation of early fine layers and up-sampled deep layers. We transform the coarse outputs (e.g., with resolution 18×12) to dense outputs (e.g., with resolution 37×25) with up-sampling interpolation of factor 2. The feature maps up-sampled from the low resolutions and those from the high resolutions are then aggregated with the element-wise summation, shown as the blue circle with plus in Figure 1. Note that we select the element-wise summation instead of other operations (e.g. multiplication) by experimenting on the validation set. After that, the following convolutional layers can be learned based on the combination of coarse and fine information. To capture more detailed local boundaries, the input image is further filtered with the 5×5 convolutional filters and then aggregated into the later feature maps. We perform the cross-layer combination four times until obtaining the feature maps with the same size as the input image. Finally, the convolutional layers are utilized to generate the C confidence maps to predict scores for C labels (including background) at each pixel location. Our loss function is the sum of cross-entropy terms for all pixels in the output map.

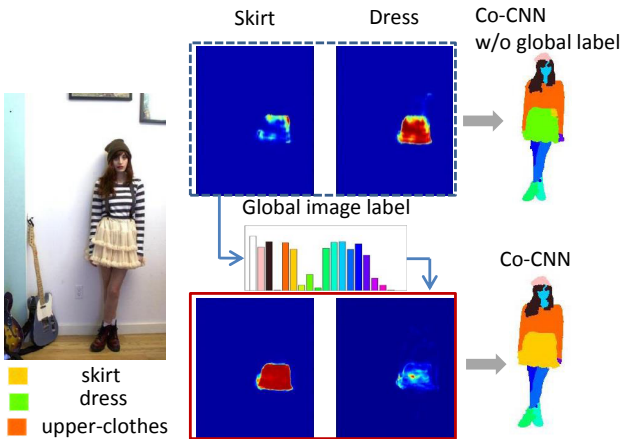

Figure 2. Comparison of label confidence maps between Co-CNN and that without using global labels. By using the global image label probabilities to guide feature learning, the confidence maps for skirt and dress can be corrected.

3.2. Global Image-level Context

An auxiliary objective for multi-label prediction is used after the intermediate layers with spatial resolution of 18×12, as shown in the pentagon in Figure 1. Following the fully-connected layer, the C-way softmax which produces a probability distribution over the C class labels is appended. Squared loss is used during the global image label prediction. Suppose for each image I in the training set, $y = [y_1, y_2, \cdots, y_C]$ is the ground-truth multi-label vector. $y_c = 1, (c = 1, \cdots, C)$ if the image is annotated with class c, and otherwise $y_c = 0$. The ground-truth probability vector is normalized as $p_c = \frac{y_c}{||y||_1}$ and the predictive probability vector is $\hat{p} = [\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_C]$. The squared loss to be minimized is defined as $J = \sum_{c=1}^{C}(p_c - \hat{p}_c)^2$. During training, the loss of image-level labels is added to the total loss of the network weighted by a discount factor 0.3. To utilize the predicted global image label probabilities, we perform two types of combination: concatenating the predicted label probabilities with the intermediate convolutional layers (*image label concatenation* in Figure 1) and element-wise summation with label confidence maps (*element-wise summation* in Figure 1).

First, consider that the feature maps of the m-th convolutional layer are a three-dimensional array of size $h^m \times w^m \times d^m$, where h^m and w^m are spatial dimensions, and d^m is the number of channels. We generate C additional probability maps $\{x_c^p\}_1^C$ with size $h^m \times w^m$ where each $x_{i,j,c}^p$ at location (i, j) is set as the predicted probability p_c of the c-th class. By concatenating the feature maps x^m of the m-th layer and the probability maps $\{x_c^p\}_1^C$, we generate the combined feature maps $\hat{x}^m = [x^m, x_1^p, x_2^p, \cdots, x_C^p]$ of the size $h^m \times w^m \times (d^m + C)$. The outputs $x_{i,j}^{m+1}$ at location (i, j) in the next layer are computed by

$$x_{i,j}^{m+1} = f_k(\{\hat{x}_{i+\delta i, j+\delta j}^m\}_{0 \le \delta i, \delta j \le k}), \quad (1)$$

where k is the kernel size, and f_k is the corresponding convolution filters. We perform this concatenation after each combination of coarse and fine layers in Section 3.1, as shown in Figure 1.

Second, we element-wisely sum the predicted confidence maps with the global image label probabilities. If the class c has a low probability of appearing in the image, the corresponding pixel-wise probability will be suppressed. Given the probability $r_{i,j,c}$ of the c-th confidence map at location (i, j), the resulting probability $\hat{r}_{i,j,c}$ is calculated by $\hat{r}_{i,j,c} = r_{i,j,c} + \hat{p}_c$ for the c-th channel. The incorporation of global image-level context into label confidence maps can help reduce the confusion of competing labels.

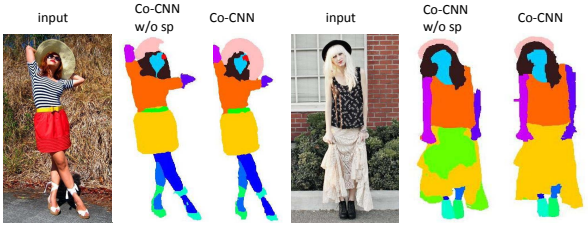

Figure 3. Comparison of example results of using local super-pixel contexts. For each image, we show the results from Co-CNN and "Co-CNN w/o sp", i.e. no local super-pixel information used.

3.3. Local Super-pixel Context

We further integrate the within-super-pixel smoothing and the cross-super-pixel neighborhood voting into the training and testing process to respect the local detailed information. They are only performed on the prediction layer (i.e. C confidence maps) instead of all convolutional layers. It is advantageous that super-pixel guidance is used at the later stage, which avoids making premature decisions and thus learning unsatisfactory convolution filters.

Within-super-pixel Smoothing: For each input image I, we first compute the over-segmentation of I using the entropy rate based segmentation algorithm [17] and obtain 500 super-pixels per image. Given the C confidence maps $\{x_c\}_1^C$ in the prediction layer, the within-super-pixel smoothing is performed on each map x_c. Let us denote the super-pixel covering the pixel at location (i, j) by s_{ij}, the smoothed confidence maps $\tilde{x}_c$ can be computed by

$$\tilde{x}_{i,j,c} = \frac{1}{||s_{ij}||} \sum_{(i',j') \in s_{ij}} x_{i',j',c}, \quad (2)$$

where $||s_{ij}||$ is the number of pixels within the super-pixel s_{ij} and (i', j') represents all pixels within s_{ij}.

Cross-super-pixel Neighborhood Voting: After smoothing confidences within each super-pixel, we can take the neighboring larger regions into account for better inference, and exploit more statistical structures and correlations between different super-pixels. For classes with non-uniform appearance (e.g., the common clothes items), the inference within larger regions may better capture the characteristic distribution for this class. For simplicity, let $\tilde{x}_{\mathbf{s}}$, $\tilde{x}_{\mathbf{s}'}$ denote the smoothed responses of the super-pixel $\mathbf{s}$ and $\mathbf{s}'$ on each confidence map, respectively. For each super-pixel $\mathbf{s}$, we first compute a concatenation of bag-of-words from RGB, Lab and HOG descriptor for each super-pixel, and the feature of each super-pixel can be denoted as $b_{\mathbf{s}}$. The cross neighborhood voted response $\bar{x}_{\mathbf{s}}$ of the super-pixel $\mathbf{s}$ is calculated by

$$\bar{x}_{\mathbf{s}} = (1-\alpha)\tilde{x}_{\mathbf{s}} + \alpha \sum_{\mathbf{s}' \in D_{\mathbf{s}}} \frac{\exp(-||b_{\mathbf{s}} - b_{\mathbf{s}'}||^2)}{\sum_{\hat{\mathbf{s}} \in D_{\mathbf{s}}} \exp(-||b_{\mathbf{s}} - b_{\hat{\mathbf{s}}}||^2)} \tilde{x}_{\mathbf{s}'}. \tag{3}$$

Here, $D_{\mathbf{s}}$ denotes the neighboring super-pixel set of the super-pixel $\mathbf{s}$. We weight the voting of each neighboring super-pixel $\mathbf{s}'$ with the normalized appearance similarities. If the pair of super-pixels $(\mathbf{s}, \mathbf{s}')$ shares higher appearance similarity, the corresponding weight of neighborhood voting will be higher. Our within-super-pixel smoothing and cross-super-pixel neighborhood voting can be seen as two types of pooling methods, which are performed on the local responses within the irregular regions depicted by super-pixels. When back-propagating through the network, the gradients are back-propagated through each super-pixel. Some results with/without incorporating the local super-pixel contexts are shown in Figure 3.

3.4. Parameter details of Co-CNN

Our detailed Co-CNN configuration is listed in Table 1. We use the small 3×3 and 5×5 receptive fields throughout the whole network, and the non-linear rectification layers after every convolutional layer. The network has 21 layers if only the layers with parameters are counted, or 27 layers if we also count max pooling and up-sampling. The dropout (30%) of fully-connected layer in the image-level label prediction is set by the validation set.

4. Experiments

4.1. Experimental Settings

Dataset: We evaluate the human parsing performance of our Co-CNN on the large ATR dataset [15] and the small Fashionista dataset [31]. Human parsing is to predict every pixel with 18 labels: face, sunglass, hat, scarf, hair, upper-clothes, left-arm, right-arm, belt, pants, left-leg, right-leg, skirt, left-shoe, right-shoe, bag, dress and null. Totally, 7,700 images are included in the ATR dataset [15], 6,000 for training, 1,000 for testing and 700 for validation[1]. The Fashionista dataset contains 685 images, in which 229 images are used for testing and the rest for training. We use the Fashionista dataset after transforming the original labels to 18 categories as in [15] for fair comparison. We use the same evaluation criterion as in [30] and [15], including accuracy, average precision, average recall, and average F-1 score over pixels. The images in these two datasets are near frontal-view and have little cluttered background, and are insufficient for real-world applications with arbitrary postures, views and backgrounds. We collect 10,000 real-world human pictures from a social network, *chictopia.com*, to construct a much larger dataset "Chictopia10k", and annotate pixel-level labels following [15]. Our new dataset mainly contains images in the wild (e.g., more challenging poses, occlusion and clothes), which will be released upon publication to promote future research on human parsing.

Implementation Details: We augment the training images with the horizontal reflections, which improves about 4% in terms of F-1 scores. Given a test image, we use the human detection algorithm [10] to detect the human

Table 1. The detailed configuration of our Co-CNN.

component	type	kernel size/stride	output size
local-to-global	convolution	$5 \times 5/1$	$150 \times 100 \times 128$
	convolution	$5 \times 5/1$	$150 \times 100 \times 192$
	max pool	$3 \times 3/2$	$75 \times 50 \times 192$
	convolution	$5 \times 5/1$	$75 \times 50 \times 192$
	convolution	$5 \times 5/1$	$75 \times 50 \times 192$
	max pool	$3 \times 3/2$	$37 \times 25 \times 192$
	convolution	$5 \times 5/1$	$37 \times 25 \times 192$
	convolution	$5 \times 5/1$	$37 \times 25 \times 192$
	max pool	$3 \times 3/2$	$18 \times 12 \times 192$
	convolution	$5 \times 5/1$	$18 \times 12 \times 192$
	convolution	$5 \times 5/1$	$18 \times 12 \times 192$
image-level label prediction	convolution	$1 \times 1/1$	$18 \times 12 \times 96$
	FC (dropout 30%)		$1 \times 1 \times 1024$
	FC		$1 \times 1 \times 18$
	Squared Loss		$1 \times 1 \times 18$
global-to-local	upsampling	$2 \times 2/2$	$37 \times 25 \times 192$
	convolution	$5 \times 5/1$	$37 \times 25 \times 192$
	element sum		$37 \times 25 \times 192$
	concat		$37 \times 25 \times 210$
	convolution	$5 \times 5/1$	$37 \times 25 \times 192$
	upsampling	$2 \times 2/2$	$75 \times 50 \times 192$
	convolution	$3 \times 3/1$	$75 \times 50 \times 192$
	element sum		$75 \times 50 \times 192$
	concat		$75 \times 50 \times 210$
	convolution	$5 \times 5/1$	$75 \times 50 \times 192$
	upsampling	$2 \times 2/2$	$150 \times 100 \times 192$
	convolution	$5 \times 5/1$	$150 \times 100 \times 192$
	element sum		$150 \times 100 \times 192$
	concat		$150 \times 100 \times 210$
	convolution	$5 \times 5/1$	$150 \times 100 \times 192$
	convolution (image)	$5 \times 5/1$	$150 \times 100 \times 192$
	element sum		$150 \times 100 \times 192$
	convolution	$3 \times 3/1$	$150 \times 100 \times 256$
prediction	convolution	$1 \times 1/1$	$150 \times 100 \times 18$
	element sum		$150 \times 100 \times 18$
	convolution	$1 \times 1/1$	$150 \times 100 \times 18$
super-pixel	within-S-P smoothing		$150 \times 100 \times 18$
	cross-S-P voting		$150 \times 100 \times 18$
	Softmax Loss		$150 \times 100 \times 18$

[1] We sincerely thank the authors of [15] for sharing the dataset.

Table 2. Comparison of human parsing performances with several architectural variants of our model and four state-of-the-arts when evaluating on ATR [15]. The ⋆ indicates the method is not a fully end-to-end framework.

Method	Accuracy	F.g. accuracy	Avg. precision	Avg. recall	Avg. F-1 score
⋆ Yamaguchi et al. [31]	84.38	55.59	37.54	51.05	41.80
⋆ PaperDoll [30]	88.96	62.18	52.75	49.43	44.76
⋆M-CNN [21]	89.57	73.98	64.56	65.17	62.81
⋆ ATR [15]	91.11	71.04	71.69	60.25	64.38
baseline (150-75)	92.77	68.66	67.98	62.85	63.88
baseline (150-75-37)	92.91	76.29	78.48	65.42	69.32
baseline (150-75-37-18)	94.41	78.54	76.62	71.24	72.72
baseline (150-75-37-18, w/o fusion)	92.57	70.76	67.17	64.34	65.25
Co-CNN (concatenate with global label)	94.90	80.80	78.35	73.14	74.56
Co-CNN (summation with global label)	94.28	76.43	79.62	71.34	73.98
Co-CNN (concatenate, summation with global label)	94.87	79.86	78.00	73.94	75.27
Co-CNN (w-s-p)	95.09	80.50	79.22	74.38	76.17
Co-CNN (full)	95.23	80.90	81.55	74.42	76.95
Co-CNN (+Chictopia10k)	**96.02**	**83.57**	**84.95**	**77.66**	**80.14**

Table 3. Per-Class Comparison of F-1 scores with several variants of our versions and four state-of-the-art methods on ATR [15].

Method	Hat	Hair	S-gls	U-cloth	Skirt	Pants	Dress	Belt	L-shoe	R-shoe	Face	L-leg	R-leg	L-arm	R-arm	Bag	Scarf
⋆ Yamaguchi et al. [31]	8.44	59.96	12.09	56.07	17.57	55.42	40.94	14.68	38.24	38.33	72.10	58.52	57.03	45.33	46.65	24.53	11.43
⋆ PaperDoll [30]	1.72	63.58	0.23	71.87	40.20	69.35	59.49	16.94	45.79	44.47	61.63	52.19	55.60	45.23	46.75	30.52	2.95
⋆M-CNN [21]	**80.77**	65.31	35.55	72.58	77.86	70.71	81.44	38.45	53.87	48.57	72.78	63.25	68.24	57.40	51.12	57.87	43.38
⋆ ATR [15]	77.97	68.18	29.20	79.39	**80.36**	79.77	**82.02**	22.88	53.51	50.26	74.71	69.07	71.69	53.79	58.57	53.66	**57.07**
baseline (150-75)	28.94	81.96	63.04	74.71	50.91	70.18	53.87	37.32	64.87	60.49	86.02	72.55	72.40	78.54	72.43	63.94	18.86
baseline (150-75-37)	63.12	80.08	36.55	83.12	63.17	81.10	65.38	28.36	65.75	69.94	82.88	82.03	81.55	75.68	76.31	77.36	37.15
baseline (150-75-37-18)	59.41	84.67	69.59	82.75	65.52	80.30	65.29	43.50	75.85	72.71	88.00	85.11	84.35	80.61	80.27	72.25	22.87
baseline (150-75-37-18, w/o fusion)	57.93	79.15	54.01	78.08	65.27	73.25	50.73	20.63	63.00	63.57	82.48	68.20	73.02	73.39	73.37	72.79	27.05
Co-CNN (concatenate with global label)	62.96	85.09	70.42	84.20	70.36	83.02	70.67	45.71	74.26	74.23	88.14	87.09	85.99	81.94	80.73	73.91	24.39
Co-CNN (summation with global label)	69.77	87.91	78.05	79.31	61.81	80.53	57.51	28.16	74.87	73.22	91.34	82.15	83.98	84.37	84.23	79.78	35.35
Co-CNN (concatenate, summation with global label)	65.05	85.11	70.92	84.02	73.20	81.49	69.61	45.44	73.59	73.40	88.73	83.25	83.51	82.74	82.15	77.88	35.75
Co-CNN (w-s-p)	71.25	85.52	71.37	84.70	74.98	82.23	71.18	**46.28**	74.83	75.04	88.76	84.39	83.38	82.84	82.62	78.97	33.66
Co-CNN (full)	72.07	86.33	72.81	85.72	70.82	83.05	69.95	37.66	76.48	76.80	89.02	85.49	85.23	84.16	84.04	81.51	44.94
Co-CNN (+Chictopia10k)	75.88	**89.97**	**81.26**	**87.38**	71.94	**84.89**	71.03	40.14	**81.43**	**81.49**	**92.73**	**88.77**	**88.48**	**89.00**	**88.71**	**83.81**	46.24

body. The resulting human centric image is then rescaled into 150×100 and fed into our Co-CNN for pixel-wise prediction. We choose the resolution of 150×100 for each image, to balance computational efficiency, practicality (e.g., GPU memory) and accuracy. To evaluate the performance, we re-scale the output pixel-wise prediction back to the size of the original ground-truth labeling. All models in our experiment are trained and tested based on Caffe [13] on a single NVIDIA Tesla K40c. We set the weight parameter α in cross-super-pixel voting as 0.3 by using the validation set. The network is trained from scratch using the annotated training images. The weights of all network parameters are initialized with Gaussian distribution with standard deviation as 0.001. We train Co-CNN using stochastic gradient descent with a batch size of 12 images, momentum of 0.9, and weight decay of 0.0005. The learning rate is initialized at 0.001 and divided by 10 after 30 epochs. We train the networks for roughly 90 epochs, which takes 4 to 5 days. Our Co-CNN can rapidly process one 150×100 image within about 0.0015 second. After incorporating the super-pixel extraction [17], we test one image within about 0.15 second. This compares much favorably to other state-of-the-art approaches, as current state-of-the-art approaches have higher complexity: [30] runs in about 10 to 15 seconds, [6] runs in 1 to 2 minutes and [15] runs in 0.5 second.

Figure 4. Exemplar images of our "Chictopia10k" dataset.

4.2. Results and Comparisons

We compare our proposed Co-CNN with five state-of-the-art approaches [31] [30] [21] [15] [25] on two datasets. All results of the competing methods and our methods are obtained by using the same training and testing setting described in the paper [15].

ATR dataset [15]: Table 2 and Table 3 show the performance of our models and comparisons with four state-of-the-arts on overall metrics and F-1 scores of foreground semantic labels, respectively. Our "Co-CNN (full)" can

Table 4. Comparison of parsing performance with three state-of-the-arts on the test images of Fashionista [31].

Method	Acc.	F.g. acc.	Avg. prec.	Avg. recall	Avg. F-1 score
⋆ Yamaguchi et al. [31]	87.87	58.85	51.04	48.05	42.87
⋆ PaperDoll [30]	89.98	65.66	54.87	51.16	46.80
⋆ ATR [15]	92.33	76.54	73.93	66.49	69.30
Co-CNN (full)	96.08	84.71	82.98	77.78	79.37
Co-CNN (+Chictopia10k)	**97.06**	**89.15**	**87.83**	**81.73**	**83.78**

significantly outperform four baselines: 35.15% over Yamaguchi et al. [31], 32.19% over PaperDoll [30], 14.14% over M-CNN [21] and 12.57% over ATR [15] in terms of average F-1 score. Since the code of ATR [15] is not publicly available, we only take our "Chictopia10k" dataset as the supplementary dataset to the training set and report the results as "Co-CNN (+Chictopia10K)". After training with more realistic images in our newly collected dataset "Chictopia10k", our "Co-CNN (+Chictopia10k)" can further improve the average F-1 score by 3.19% and the average precision by 3.4%. This indicates that our "Chictopia10k" dataset can introduce greater data diversity and improve the network generality. We show the F-1 scores for each label in Table 3. Generally, our Co-CNN shows much higher performance than other methods. In terms of predicting small labels such as hat, belt, bag and scarf, our method achieves a very large gain, e.g. 72.81% *vs* 29.20% [15] for sunglass, 81.51% *vs* 53.66% [15] for bag. We also achieve much better performance on human body parts, e.g. 84.16% *vs* 53.79% [15] for left-arm. It demonstrates that Co-CNN performs very well on various poses (e.g. human body parts), fine details (e.g. small labels) and diverse clothing styles.

Fashionista dataset [31]: Table 4 gives the comparison results on the 229 test images of the Fashionista dataset. All results of the state-of-the-art methods were reported in [15]. Note that deep learning based algorithm requires enough training samples. Following [15], we only report the performance by training on the same large ATR dataset[15], and then testing on the 229 images on Fashionista dataset. Our method "Co-CNN (full)" can substantially outperform the baselines by 10.07%, 32.57% and 36.5% over "ATR [15]", "PaperDoll [30]" and "Yamaguchi et al. [31]" in terms of average F-1 score, respectively. We cannot compare all metrics with the CRF model proposed in [25], since it only reported the average pixel-wise accuracy, and only achieved 84.88%, which only slightly improved the results 84.68% of PaperDoll [30] on Fashionista, as reported in [25].

The qualitative comparison of parsing results is visualized in Figure 5. Our Co-CNN outputs more meaningful and precise predictions than PaperDoll [30] and ATR [15] despite the large appearance and position variations.

4.3. Discussion on Our Network

We further evaluate the different network settings for our three components, presented in Table 2 and Table 3.

Local-to-Global-to-Local Hierarchy: We explore different variants of our basic network structure. Note that all the following results are obtained without combining the global image-level label context and the local super-pixel contexts. First, different down-sampled spatial resolutions are tested. The "baseline (150-75)", "baseline (150-75-37)" and "baseline (150-75-37-18)" are the versions with down-sampling up to 75×50, 37×25 and 18×12, respectively. When only convolving the input image with two resolutions ("baseline (150-75)"), the performance is worse than the state-of-the-arts [15]. After further increasing the depth of the network by down-sampling up to 37×25 ("baseline (150-75-37)"), the F-1 score can be significantly increased by 5.44%, compared to "baseline (150-75)". The "baseline (150-75-37-18)" can further improve the F-1 score by 3.4%, compared to "baseline (150-75-37)". We do not report results by further down-sampling the feature maps since only slight improvement is achieved with smaller resolutions.

Second, we also evaluate the effectiveness of the cross-layer context combination. The "baseline (150-75-37-18, w/o fusion)" represents the version without cross-layer combinations. The large decrease 7.47% in F-1 score compared with the "baseline (150-75-37-18)" demonstrates the great advantage of the cross-layer combination. Combining the cross-layer information enables the network to make precise local predictions and respect global semantic information.

Finally, we also test the FCN architecture [22] on semantic segmentation in the human parsing task, i.e., fine-tuning the pre-trained classification network with the human parsing dataset and only performing the combination for the pixel-wise predictions. Its performance is much worse than our network (i.e. 64.63% vs 72.72% of "baseline (150-75-37-18)" in average F-1 score).

Global Image-level Context: We also explore different architectures to demonstrate the effectiveness of utilizing the global image label context. All the following results are obtained without using local super-pixel contexts. After the summation of global image label probabilities ("Co-CNN (summation with global label)"), the performance can be increased by 1.26%, compared to "baseline (150-75-37-18)". After concatenating the global image label probabilities with each subsequent convolutional layer, "Co-CNN (concatenate with global label)", the performance can be improved by 1.84% in F-1 score, compared to the version without using global label ("baseline (150-75-37-18)"). The further summation of global image label probabilities can bring 0.71% increase in F-1 score, shown as "Co-CNN (concatenate, summation with global label)". The most significant improvements over "baseline

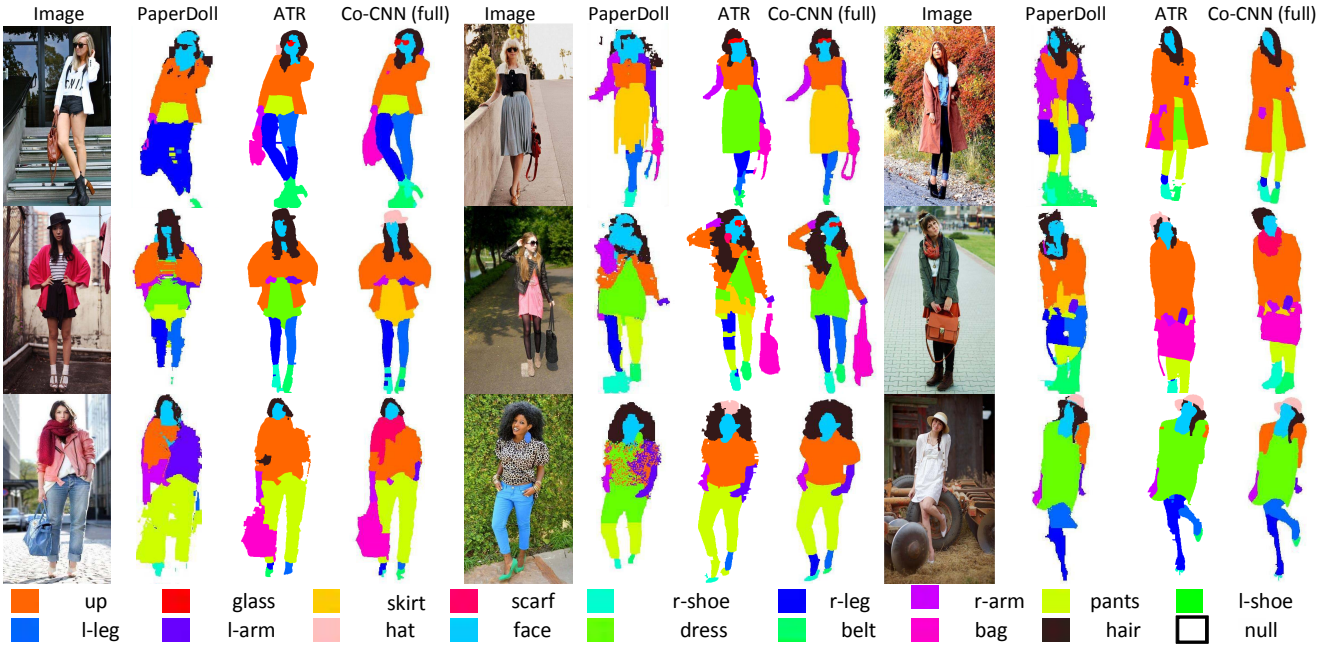

Figure 5. Result comparison of our Co-CNN and two state-of-the-art methods. For each image, we show the parsing results by PaperDoll [30], ATR [15] and our Co-CNN sequentially.

(150-75-37-18)" can be observed from the F-1 scores for clothing items, e.g., 7.68% for skirt and 4.32% for dress. The main reason for these improvements may be that by accounting for the global image-level label probabilities, the label exclusiveness and occurrences can be well captured during dense pixel-wise prediction.

Local Super-pixel Contexts: Extensive evaluations are conducted on the effectiveness of using local super-pixel contexts. The average F-1 score increases by 0.9% by embedding the within-super-pixel smoothing into our network ("Co-CNN (w-s-p)"), compared to the version "Co-CNN (concatenate, summation with global label)". Our full network "Co-CNN (full)" leads to 1.68% increase. For the F-1 score for each semantic label, the significant improvements are obtained for the labels of small regions (e.g. hat, sunglasses and scarf). For instance, the F-1 score for hat is increased by 7.02%, and 9.19% for scarf, compared with "Co-CNN (concatenate, summation with global label)". This demonstrates that the local super-pixel contexts can help preserve the local boundaries and generate more precise classification for small regions. Previous works ofter apply the super-pixel smoothing as the post-processing step.

5. Conclusions and Future Work

In this work, we proposed a novel Co-CNN architecture for human parsing task, which integrates the cross-layer context, global image label context and local super-pixel contexts into a unified network. For each input image, our Co-CNN produces the correspondingly-sized pixel-wise prediction in a full end-to-end way. The local-to-global-to-local hierarchy is used to combine the local detailed information and global semantic information. The global image label prediction, within-super-pixel smoothing and cross-super-pixel neighborhood voting are formulated as the natural components of our Co-CNN. Extensive experimental results clearly demonstrated the effectiveness of the proposed Co-CNN. A new large dataset "Chictopia10k" has been built. In the future, we will further extend our Co-CNN architecture for generic image parsing tasks, e.g., object semantic segmentation and scene parsing. Our online demo website will be released upon publication to demonstrate the efficiency and effectiveness.

Acknowledgement

This work was partially supported by the 973 Program of China (Project No. 2014CB347600), and the National Natural Science Foundation of China (Grant No. 61522203, 61328205). This work was also supported in part by the Guangdong Natural Science Foundation under Grant S2013050014548 and Grant 2014A030313201, in part by the Program of Guangzhou Zhujiang Star of Science and Technology under Grant 2013J2200067, and in part by Fundamental Research Funds for the Central Universities (no. 13lgjc26). This work was also partly supported by gift funds from Adobe Research.

References

[1] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1312–1328, 2012. 1

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *CVPR*, 2015. 2

[3] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. *ICCV*, 2015. 2

[4] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. Human pose estimation using body parts dependent joint regressors. In *Computer Vision and Pattern Recognition*, pages 3041–3048, 2013. 1

[5] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan. Style finder: Fine-grained clothing style detection and retrieval. In *Computer Vision and Pattern Recognition Workshops*, pages 8–13, 2013. 1

[6] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan. A deformable mixture parsing model with parselets. In *International Conference on Computer Vision*, 2013. 1, 2, 6

[7] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Transaction on Pattern Analyisi and Machine Intelligence*, 35(8):1915–1929, 2013. 1, 2

[8] P. George, C. Liang-Chieh, M. Kevin, and Y. A. L. Weakly- and semi-supervised learning of a dcnn for semantic image segmentation. *ICCV*, 2015. 1, 2

[9] K. M. A. L. Y. George Papandreou, Liang-Chieh Chen. Weakly- and semi-supervised learning of a dcnn for semantic image segmentation. *ICCV*, 2015. 2

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, 2014. 1, 2, 5

[11] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. *arXiv preprint arXiv:1411.5752*, 2014. 1, 2

[12] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision*, pages 297–312. 2014. 2

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678, 2014. 6

[14] Y. Kalantidis, L. Kennedy, and L.-J. Li. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *ACM conference on International conference on multimedia retrieval*, pages 105–112, 2013. 1

[15] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan. Deep human parsing with active template regression. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2015. 1, 2, 5, 6, 7, 8

[16] X. Liang, S. Liu, Y. Wei, L. Liu, L. Lin, and S. Yan. Towards computational baby learning: A weakly-supervised approach for object detection. *ICCV*, 2015. 1

[17] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa. Entropy rate superpixel segmentation. In *Computer Vision and Pattern Recognition*, pages 2097–2104, 2011. 4, 6

[18] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan. Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia*, 16(1):253–265, 2014. 2

[19] S. Liu, X. Liang, L. Liu, K. Lu, L. Lin, X. Cao, and S. Yan. Fashion parsing with video context. *Multimedia, IEEE Transactions on*, 17(8):1347–1358, 2015. 1

[20] S. Liu, X. Liang, L. Liu, K. Lu, L. Lin, and S. Yan. Fashion parsing with video context. In *Proceedings of the ACM International Conference on Multimedia*, pages 467–476, 2014. 1

[21] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan. Matching-cnn meets knn: Quasi-parametric human parsing. *CVPR*, 2015. 1, 2, 6, 7

[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, 2015. 2, 7

[23] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010. 1

[24] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. *CVPR*, 2015. 1, 2

[25] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. A High Performance CRF Model for Clothes Parsing. In *Asian Conference on Computer Vision*, 2014. 2, 6, 7

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CVPR*, 2015. 1

[27] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Computer Vision and Pattern Recognition*, 2014. 1

[28] N. Wang and H. Ai. Who blocks who: Simultaneous clothing segmentation for grouping images. In *International Conference on Computer Vision*, pages 1535–1542, 2011. 2

[29] Y. Wang, D. Tran, Z. Liao, and D. Forsyth. Discriminative hierarchical part-based models for human parsing and action recognition. *The Journal of Machine Learning Research*, 13(1):3075–3102, 2012. 1

[30] K. Yamaguchi, M. Kiapour, and T. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *International Conference on Computer Vision*, 2013. 1, 2, 5, 6, 7, 8

[31] K. Yamaguchi, M. Kiapour, L. Ortiz, and T. Berg. Parsing clothing in fashion photographs. In *Computer Vision and Pattern Recognition*, pages 3570–3577, 2012. 1, 2, 5, 6, 7

[32] W. Yang, L. Lin, and P. Luo. Clothing co-parsing by joint image segmentation and labeling. In *Computer Vision and Pattern Recognition*, 2014. 1, 2

[33] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang. Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *IEEE Transactions on Image Processing*, 2015. 1