Human Parsing with Contextualized Convolutional Neural Network

Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang,
Jinhui Tang, Liang Lin, and Shuicheng Yan

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, semantic edge 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. Third, semantic edge context is further incorporated into Co-CNN, where the high-level semantic boundaries are leveraged to guide pixel-wise labeling. 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 [1] reaches $81.72\,\mathrm{percent}$ by Co-CNN, significantly higher than $62.81\,\mathrm{percent}$ and $64.38\,\mathrm{percent}$ by the state-of-the-art algorithms, M-CNN [2] and ATR [1], respectively. By utilizing our newly collected large dataset for training, our Co-CNN can achieve $85.36\,\mathrm{percent}$ in F-1 score.

Index Terms—Human parsing, fully convolutional network, context modeling, semantic labeling

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 [3], clothes recognition and retrieval [4], people re-identification [5], human behavior analysis [6] and automatic product recommendation [7].

While there has been previous work devoted to human parsing based on human pose estimation [8], [9], [10], non-parametric label transferring [2], [4] and active template regression [1], 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 [11], bottom-up hypothesis generation [12] and template dictionary learning [13], 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 [14], but also in structure prediction such as object detection [15], [16], part prediction [17] and general object/scene semantic segmentation [18], [19]. 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) [19] and extra discriminative classifiers [20], [21]. Besides the above mentioned limitations, there are still a few 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 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

- *X. Liang and L. Lin are with the School of Data and Computer Science, Sun Yat-sen University, China. E-mail: xdliang328@gmail.com, linliang@ieee.org.*
- *C. Xu is with the Huazhong University of Science and Technology, School of Computer Science, Wuhan, Hubei, China. E-mail: xuchunyan01@gmail.com.*
- *X. Shen and J. Yang are with Adobe Research, San Jose, CA 95110. E-mail: {xshen, jiayang}@adobe.com.*
- *J. Tang is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. E-mail: jinhuitang@mail.njust.edu.cn.*
- *S. Yan is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576. E-mail: eleyans@nus.edu.sg.*

Manuscript received 21 Sept. 2015; revised 23 Feb. 2016; accepted 24 Feb. 2016. Date of publication 1 Mar. 2016; date of current version 12 Dec. 2016. Recommended for acceptance by P. Gehler.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2016.2537339

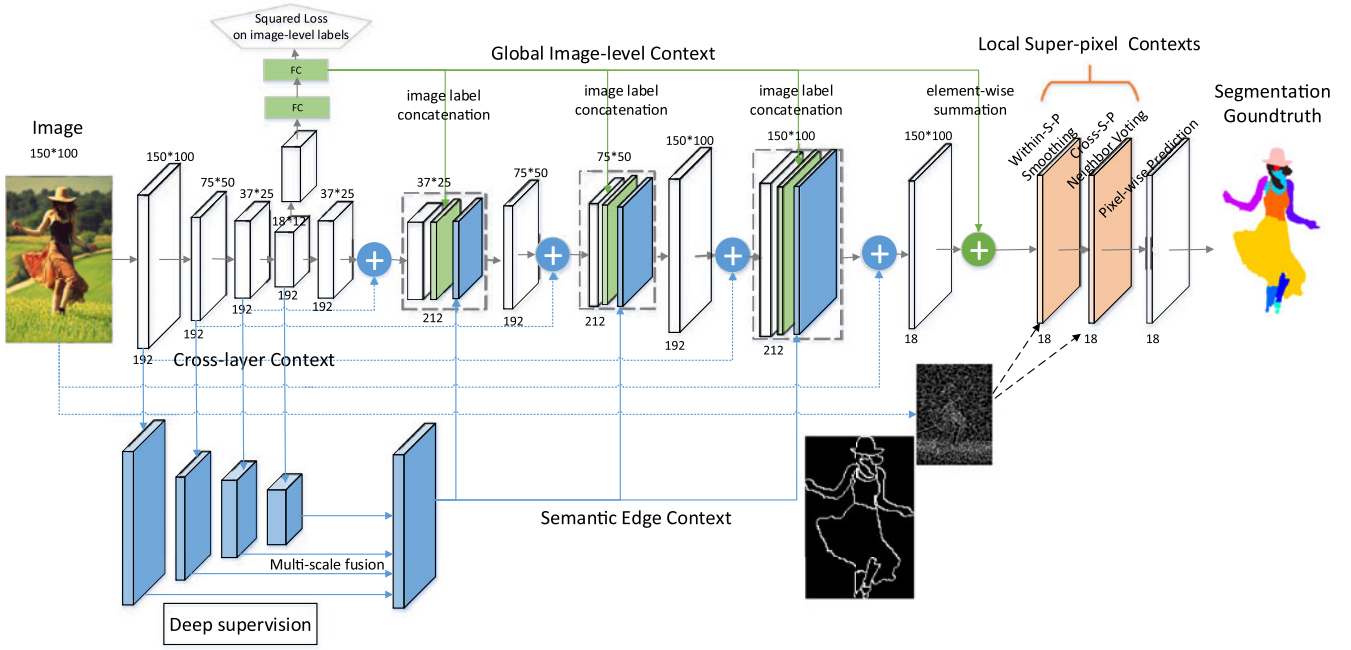


Fig. 1. Co-CNN integrates the cross-layer context, global image-level context, semantic edge context and local super-pixel contexts into a unified network. It consists of cross-layer combination, global image-level label prediction, within-super-pixel smoothing, cross-super-pixel neighborhood voting and semantic edge prediction. 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). Third, the multi-scale prediction streams are appended to predict semantic edges, and then the predicted edge maps are utilized to guide the feature learning in subsequent convolutional layers. 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.

segmentation. Finally, the critical segmentation-specific context constraints, such as the local super-pixel smoothness or the integrity and uniqueness for each semantic region, have not been well considered in previous works on semantic segmentation. For instance, the pixels within the same super-pixel or neighboring super-pixels should have high possibilities to be assigned with the same semantic label. And the label probabilities from neighboring super-pixels should help guide the label inference by leveraging the location priors. Meanwhile, the pixels within the same semantic region (e.g., upper-clothes or dress) should be predicted as the same semantic label to retain the region integrity. Moreover, the semantic edges, i.e., the edges along the boundaries of different semantic regions, could be inferred to better separate these regions. For example, sometimes it is difficult to infer whether the person is wearing a single dress, or upper clothes with a separate skirt below. But these cases can be easily distinguished by explicitly detecting if there is a semantic edge in-between. The challenges for incorporating these segmentation-specific contexts into a unified CNN architecture lie in two aspects: in terms of local super-pixel context, the super-pixel numbers and the neighboring super-pixel graph structures dramatically vary for individual images; in terms of region integrity and uniqueness, long-term contextual information over long-range pixels should be effectively leveraged during CNN training. Some traditional methods [1], [10] resort to bottom region proposal generation or post-processing such as graphical inference to enforce these segmentation-specific contextual constraints. However, their separate stages often suffer from the inconsistent optimization targets and high computation cost.

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 Fig. 1. Our Co-CNN aims to simultaneously capture cross-layer context, global image-level context, semantic edge context and local super-pixel contexts by using the local-to-global-to-local hierarchical structure, global image-level label prediction, semantic edge 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. In this way, the low-level fine details preserved in the early layers can be incorporated back into the deep layers. These enhanced feature maps can be utilized for better feature learning in the subsequent layers. In total, four cross-layer combinations are performed to integrate different levels of context.

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 Fig. 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 Fig. 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 Fig. 1.

Third, the semantic edge prediction is incorporated into our Co-CNN to retain the region integrity and region uniqueness for each semantic label. We define the fine-grained semantic edges as the boundaries between regions belonging to different semantic labels in each image, which is different from the traditional edge prediction task [23] that aims to predict object boundaries. To resolve ambiguity in object boundaries in natural images and semantic edges, multi-scale prediction streams for semantic edges are appended to jointly capture low-level local boundaries from the early convolutional layers and high-level semantic boundaries from deep layers. Then the edge predictions from different streams are combined to provide contextual information for pixel-wise semantic labeling. Due to the favorable characteristics of the learned features for semantic edge prediction and pixel-wise labeling, the proposed Co-CNN can generate more complete and meaningful semantic regions.

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. Unlike the traditional practice of treating the complex super-pixel random field regularization as post-processing [19], we embed the within-super-pixel smoothing and cross-super-pixel neighborhood voting into the training stage and the testing stage. Before the final prediction layer of our network, the within-super-pixel smoothing is performed on the feature maps to constrain the label consistency, and then the weighted cross-super-pixel neighborhood voting is further used to guarantee the consistency in the larger local regions.

Comprehensive evaluations and comparisons on the ATR dataset [1] and the Fashionista dataset [4] well demonstrate that our Co-CNN yields results that significantly surpass all previously published methods, boosting the performance of the current state-of-the-arts from 64.38 [1] to 81.72 percent. We also build a much larger dataset "Chictopia10k", which contains 10,000 annotated images. By adding the images of "Chictopia10k" into the training set, the F-1 score can be further improved to 85.36, 20.98 percent higher than the state-of-the-arts [1], [4]. Notably, by using the semantic edge context, the performance of Co-CNN can be significantly improved by 5.22 percent in F-1 score.

2 Related Work

Human parsing. Much research has been devoted to human parsing [2], [4], [8], [9], [10], [24], [25], [26], [27]. 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. [8] performed human pose estimation and attribute labeling sequentially. 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. [1] 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. Matching CNN [2] was proposed as a quasi-parametric human parsing method, which highly relies on the image gallery set.

Semantic segmentation with CNN. Our method works directly on the pixel-level representation, similar to some recent research on semantic segmentation with CNN [19], [21], [22], [28], [29]. These pixel-level representations are in contrast to the common two-stage approaches [15], [20], [30] 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. [18] trained a multi-scale convolutional network from raw pixels and employed the super-pixel tree for smoothing. Hariharan et al. [21] proposed to concatenate the computed intermediate convolutional features for pixel-wise classification. The dense pixel-level CRF was used as the post-processing step after CNN-based pixel-wise prediction [31]. 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. Meanwhile, we produce the same sized pixel-wise predictions with the input, while [22] can only generate very coarse predictions. Moreover, besides the cross-layer context embedded in the local-to-global-to-local structure, our Co-CNN incorporates global image-level context, semantic edge context and local super-pixel contexts, which have not been utilized in previous CNN-based approaches.

Edge detection. The task of detecting edges and object boundaries is fundamental to many vision tasks such as saliency detection, object detection and segmentation. Recent progress on edge detection has been achieved using convolutional neural networks, including DeepContour [32], DeepEdge [23], CSCNN [23] and holistically-nested edge

detection [33]. For instance, Xie and Tu [33] proposed to use convolutional neural networks and deeply-supervised nets for edge detection. Bertasius et al. [23] exploited object related features as high-level cues for contour detection.

Different from these previous edge detection methods, the semantic edge prediction task addressed in this paper focuses on localizing more fine-grained semantic boundaries, i.e., the boundaries between regions with different semantic labels. In addition, we believe that semantic edge prediction and semantic segmentation are intrinsically two related tasks, and predicting semantic edges can provide a more global perspective for pixel-wise labeling. In our Co-CNN the semantic edge predictions are posed as additional features to guide feature learning in the subsequent convolutional layers. The two targets of semantic edge prediction and semantic segmentation can thus be jointly optimized in an unified architecture.

3 The Proposed Co-CNN Architecture

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

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 Fig. 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 Fig. 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.

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 Fig. 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, $y = [y_1, y_2, \ldots, y_C]$ is the ground-truth multi-label vector. $y_c = 1, (c = 1, \ldots, C)$ 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 $\hat{p} = [\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_C]$. The squared loss to be minimized is defined as $L^{label} = \sum_{c=1}^{C}(p_c - \hat{p}_c)^2$. 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 Fig. 1) and element-wise summation with label confidence maps (*element-wise summation* in Fig. 1).

First, consider that the feature maps of the mth 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 cth class. By concatenating the feature maps x^m of the mth layer and the probability maps $\{x_c^p\}_1^C$, we generate the combined feature maps $\hat{x}^m = [x^m, x_1^p, x_2^p, \ldots, x_C^p]$ of the size $h^m \times w^m \times (d^m + C)$. We perform this concatenation after each combination of coarse and fine layers in Section 3.1, as shown in Fig. 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 cth 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 cth channel. The incorporation of global image-level context into label confidence maps can help reduce the confusion of competing labels.

3.3 Semantic Edge Context

In this paper, we define the semantic edge as the boundaries of regions with different semantic labels, which is intrinsically consistent with the parsing ground-truth. The semantic edge context is utilized to constrain the region integrity and uniqueness of the predicted parsing result, which often cannot be guaranteed by the local pixel-wise or super-pixel-wise prediction. We denote the corresponding ground-truth binary semantic edge map for each image I as G. We

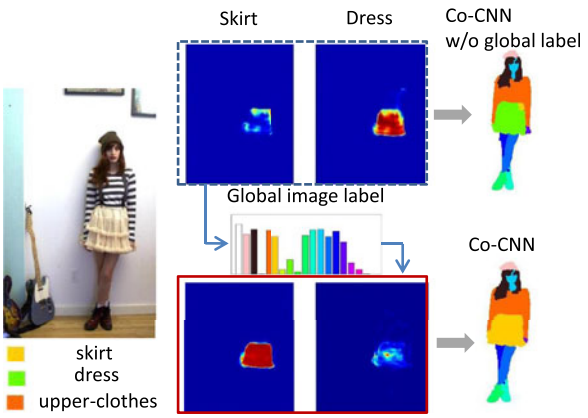


Fig. 2. Comparison of label confidence maps between Co-CNN and that without using global labels.

employ multi-scale prediction streams to increase the semantic edge prediction accuracy. As illustrated in Fig. 1, four multi-scale prediction streams are attached to four convolutional layers with different spatial resolutions in the local-to-global part. For each stream, one prediction layer and one individual loss are utilized. We set the spatial padding for each stream so that the spatial resolution of feature maps is preserved. That is, a corresponding-sized semantic edge map is generated by each stream. The multi-scale predictions from four streams are then accordingly up-sampled to high spatial resolution (i.e., 150×100) and then concatenated to generate the fused feature maps. Then the 1×1 convolutional filters are used to generate the final pixel-wise predictions. Benefiting from multi-scale predictions, the fine local details (e.g., boundaries and local consistency) captured by early layers with higher resolution and the high-level semantic information captured by deep layers with low resolution can jointly contribute to the final prediction.

Suppose we have $Q = 4$ multi-scale prediction streams, and each stream is associated with a loss $\ell_q^o, \{q = 1, 2, \ldots, Q\}$. For each image, the loss for the final predicted edge map after fusing is denoted as ℓ_f^o. The overall loss function for predicting semantic edge maps can be calculated as

$$L^{edge} = \frac{\sum_{q=1}^{Q} \ell_q^o(G, G_q^*)}{Q} + \ell_f^o\left(G, \sum_{q=1}^{Q} \alpha_q G_q^*\right), \quad (1)$$

where G_q^* and G represent the predicted semantic edge confidence maps in each stream and the ground-truth edge map, respectively. α_q is denoted as the fusion weight for each stream. The learning of this fusing weight is equivalent to training 1×1 convolutional filters on the concatenated semantic edge maps from all multi-scale streams. The loss function is computed over all pixels in the image, but over 90 percent of the pixels do not belong to the semantic edge. Following the class-balancing cross-entropy loss function used in [33], ℓ_q^o and ℓ_f^o can be computed by weighting the pixel-wise cross-entropy loss with the ratio of non-edge (negative pixels) and edge (positive) pixels in the image. The loss L^{edge} for semantic edge prediction is jointly optimized with other losses of global image-level label prediction and final

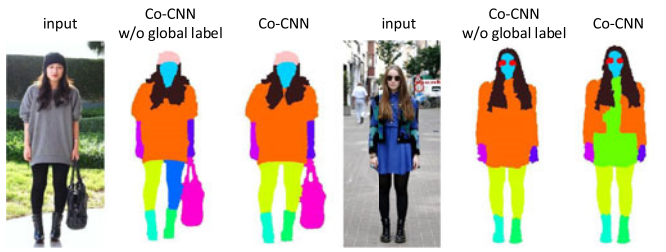


Fig. 3. The comparison of exemplar parsing results between the version of Co-CNN without using global labels and our complete Co-CNN.

semantic segmentation. We set all loss weights for the three losses as 1.

Suppose we are given the fused semantic edge confidence maps denoted as G^* with size $h^q \times w^q \times d^q$, where h^q and w^q are corresponding spatial dimensions for the qth prediction stream, and $d^q = 2$ is the number of channels. We combine the predicted edge confidence maps to learn the features for final parsing prediction. To adapt the predicted semantic edge confidence maps G^* to each convolutional layer, we rescale G^* into G_m^* with the same spatial resolution with that of feature maps in the mth layer. Following the feature map concatenation in the mth layer in Section 3.2, the combined feature map can be further extended to $\hat{x}^m = [x^m, x_1^p, x_2^p, \ldots, x_C^p, G_m^*]$ of the size $h^m \times w^m \times (d^m + C + 2)$. The outputs $x_{i,j}^{m+1}$ at location (i, j) in the next convolutional layer are computed by

$$x_{i,j}^{m+1} = f_k(\{\hat{x}_{i+\delta i, j+\delta j}^m\}_{0 \le \delta i, \delta j \le k}), \quad (2)$$

where k is the kernel size, and f_k is the corresponding convolution filters. We also perform the feature concatenation steps three times with three different spatial resolutions, as shown in Fig. 1. By embedding it into different convolutional layers, the semantic edge context can be conveniently utilized to guide feature learning for final semantic segmentation.

To better show the effectiveness of semantic edge context, some parsing results with/without utilizing semantic edge context are shown in Fig. 4. It can be observed that the

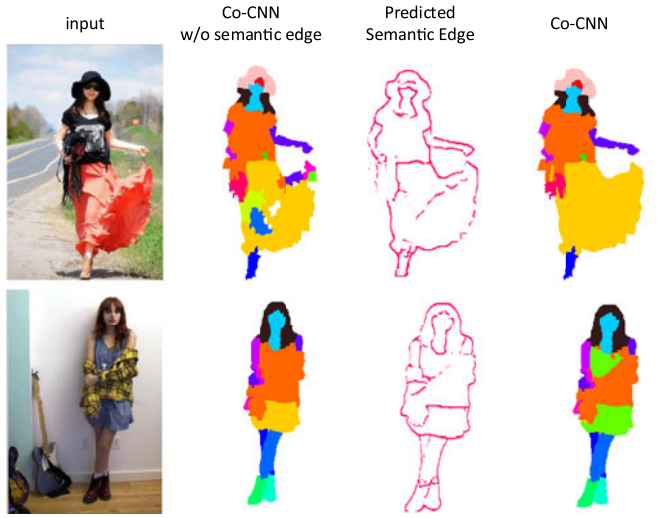


Fig. 4. Comparison of example results of incorporating semantic edge contextual information into Co-CNN. For each image, we show the results from "Co-CNN w/o semantic edge" (i.e., no semantic edge information is used), "Predicted Semantic Edge" and "Co-CNN" sequentially.

wrongly predicted fragmented small regions can be successfully fused into a complete region (in the 1st row) and the semantic edge can help split the upper-clothes into dress and upper-clothes (in the 2nd row). This verifies well that the semantic edge context can help constrain the region integrity and uniqueness, and be complementary to other semantic contexts (e.g., global image-level context and local super-pixel contexts). The predicted semantic edge maps also speak well the effectiveness of the fusion of multi-scale prediction streams.

3.4 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 [34] 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

$$\tilde{x}_{i,j,c} = \frac{1}{||s_{ij}||} \sum_{(i',j') \in s_{ij}} x_{i',j',c}, \tag{3}$$

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

$$\bar{x}_{\mathbf{s}} = (1-\alpha)\tilde{x}_{\mathbf{s}} + \alpha \sum_{\mathbf{s}' \in D_{\mathbf{s}}} \frac{\exp(-||b_{\mathbf{s}} - b_{\mathbf{s}'}||^2)}{\sum_{\hat{\mathbf{s}} \in D_{\mathbf{s}}} \exp(-||b_{\mathbf{s}} - b_{\hat{\mathbf{s}}}||^2)} \tilde{x}_{\mathbf{s}'}. \tag{4}$$

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.

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		$1 \times 1 \times 18$
semantic edge prediction	convolution	$3 \times 3/1$	$150 \times 100 \times 2$
	Softmax Loss		$150 \times 100 \times 2$
	convolution	$3 \times 3/1$	$75 \times 50 \times 2$
	Softmax Loss		$75 \times 50 \times 2$
	convolution	$3 \times 3/1$	$37 \times 25 \times 2$
	Softmax Loss		$37 \times 25 \times 2$
	convolution	$3 \times 3/1$	$18 \times 12 \times 2$
	Softmax Loss		$18 \times 12 \times 2$
	concat		$150 \times 100 \times 8$
	convolution	$1 \times 1/1$	$150 \times 100 \times 2$
	Softmax Loss		$150 \times 100 \times 2$
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		$37 \times 25 \times 212$
	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		$75 \times 50 \times 212$
	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		$150 \times 100 \times 212$
	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		$150 \times 100 \times 18$

3.5 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. Six components are included in the Co-CNN architecture that incorporates four different kinds of contextual information in an end-to-end way. The network has 26 layers if only the layers with parameters are counted, or 32 layers if we also count max pooling and up-sampling. In terms of "global-to-local", except for the last element-wise summation layer, all other element-wise summations are performed on the immediate previous

Fig. 5. Exemplar images of our "Chictopia10k" dataset.

convolutional layers and the last convolutional layers with the same resolution in "local-to-global". The last element-wise summation layer is performed on its two previous convolutional layers. The dropout (30 percent) 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 [1] and the small Fashionista dataset [8]. 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 [1], 6,000 for training, 1,000 for testing and 700 for validation. 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 [1] for fair comparison. We use the same evaluation criterion as in [4] and [1], 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"[1], and annotate pixel-level labels following [1]. Our new dataset mainly contains images in the wild (e.g., more challenging poses, occlusion and clothes), which can help promote future research on human parsing.

Data augmentation. We tried enlarging the training data to increase the diversity by many common techniques, such as the horizontal reflections and translation the human body up to 16 pixels. The horizontal reflections can help improve the testing accuracy by about 4 percent in terms of F-1 scores but the translation yields no noticeable improvement.

Implementation details. We augment the training images with the horizontal reflections, which improves about 4 percent in terms of F-1 scores. Given a test image, we use the human detection algorithm [15] to detect the human 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 [35] 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 four to five days. Our Co-CNN can rapidly process one 150×100 image within about 0.002 second. After incorporating the super-pixel extraction [34], we test one image within about 0.2 second. This compares much favorably to other state-of-the-art approaches, as current state-of-the-art approaches have higher complexity: [4] runs in about 10 to 15 seconds, [9] runs in 1 to 2 minutes and [1] runs in 0.5 second.

1. https://github.com/lemondan/HumanParsing-Dataset/

4.2 Results and Comparisons

We compare our proposed Co-CNN with five state-of-the-art approaches [1], [2], [4], [8], and [24] 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 [1]. The [4], [8] and [24] are three hand-crafted pipelines, which combine hand-crafted feature extraction, pose estimation, Conditional Random Field inference and local classification. ATR [1] and M-CNN [2] are two deep learning pipelines.

ATR dataset [1]. Tables 2 and 4 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 significantly outperform four baselines: 39.92 percent over Yamaguchi et al. [8], 36.96 percent over PaperDoll [4], 18.91 percent over M-CNN [2] and 17.34 percent over ATR [1] in terms of average F-1 score. Our method also largely boosts the performance in terms of F.g. accuracy, which obtains 85.22 percent, while four baselines achieve 55.59 percent of Yamaguchi et al. [8], 62.18 percent of PaperDoll [4], 73.98 percent of M-CNN [2] and 71.04 percent of ATR [1]. The pixel-level accuracy is also increased by at least 5.19 percent over four baselines. For fair comparison, we only take the newly collected "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 3.64 percent. 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 4. Generally, our Co-CNN shows much improvement compared to other methods. In terms of predicting small labels such as hat, belt, bag and scarf, our method achieves a very large gain, e.g., 80.59 *versus* 29.20 percent [1] for sunglass, and 84.53 *versus* 53.66 percent [1] for bag. We also achieve much better performance on human body parts, e.g., 89.58 *versus* 53.79 percent [1] for left-arm, and 90.03 *versus* 68.18 percent [1] for hair. It demonstrates that Co-CNN performs very well on various poses (e.g., human body parts), fine-grained details (e.g., small labels) and diverse clothing styles.

TABLE 2
Comparison of Human Parsing Performances with Several Architectural Variants of Our Model and Four State-of-the-Arts When Evaluating on ATR [1]

Method	Accuracy	F.g. accuracy	Avg. precision	Avg. recall	Avg. F-1 score
⋆ Yamaguchi et al. [8]	84.38	55.59	37.54	51.05	41.80
⋆ PaperDoll [4]	88.96	62.18	52.75	49.43	44.76
⋆M-CNN [2]	89.57	73.98	64.56	65.17	62.81
⋆ ATR [1]	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, post-process)	94.48	78.85	77.22	71.78	73.25
baseline (150-75-37-18, w/o fusion)	92.57	70.76	67.17	64.34	65.25
baseline (150-75-37-18, lessfilters)	94.23	77.79	75.66	70.42	71.82
baseline (150-75-37-18, concat)	93.10	72.17	69.63	66.94	67.82
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 (w-s-p, c-s-p)	95.23	80.90	81.55	74.42	76.95
Co-CNN (full)	96.30	85.22	85.26	80.04	81.72
Co-CNN (w/o edge, +Chictopia10k)	96.02	83.57	84.95	77.66	80.14
Co-CNN (edge w/o multi-scale fusion, +Chictopia10k)	96.87	87.43	86.00	82.85	84.18
Co-CNN (use edge before prediction, +Chictopia10k)	96.65	86.57	85.99	81.88	83.44
Co-CNN (+Chictopia10k)	**97.18**	**88.84**	**87.12**	**84.05**	**85.36**

The ⋆ indicates the method is not a fully end-to-end framework.

Fashionista dataset [8]. Table 5 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 [1]. Note that deep learning based algorithm requires enough training samples. Following [1], we only report the performance by training on the same large ATR dataset [1], and then testing on the 229 images on Fashionista dataset. Our method "Co-CNN (full)" can substantially outperform the baselines by 13.52, 36.02 and $39.95\,\text{percent}$ over "ATR [1]", "PaperDoll [4]" and "Yamaguchi et al. [8]" in terms of average F-1 score, respectively. We cannot compare all metrics with the CRF model proposed in [24], since it only reported the average pixel-wise accuracy, and only achieved $84.88\,\text{percent}$, which only slightly improved the results $84.68\,\text{percent}$ of PaperDoll [4] on Fashionista, as reported in [24].

Chictopia10k dataset. Table 3 shows the parsing results on the 1,000 testing images which are randomly selected from the whole Chictopia10k dataset. "Co-CNN (ATR)" and "Co-CNN (Chictopia10k)" show the parsing results when trained on ATR dataset and the rest of Chictopia10k dataset, respectively. It can be observed that both results are inferior compared to the results evaluated on the test set of ATR dataset. These results demonstrate that our newly collected Chictopia10k dataset is much difficult than the previous human parsing dataset.

4.3 Discussion on Our Network

We further evaluate the different network settings for our four components in Tables 2 and 4.

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, the semantic edge context, 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 [1]. 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 $5.44\,\text{percent}$, compared to "baseline (150-75)". The "baseline (150-75-37-18)" can further improve the F-1 score by $3.4\,\text{percent}$, 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. It well verifies that better features can be learned with a much deeper pyramid that continuously combines the hierarchical feature maps of multiple spatial 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 $7.47\,\text{percent}$ 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. Third, we report the results with different filter

TABLE 3
Human Parsing Performances on the 1,000 Testing Images from Chictopia10k

Method	Acc.	F.g. acc.	Avg. prec.	Avg. recall	Avg. F-1 score
⋆ Co-CNN (ATR)	96.34	85.10	84.00	80.70	82.08
⋆ Co-CNN (Chictopia10k)	96.60	86.29	85.08	81.85	83.21

TABLE 4
Per-Class Comparison of F-1 Scores with Several Variants of Our Versions and Four State-of-the-Art Methods on ATR [1]

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. [8]	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 [4]	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 [2]	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 [1]	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, post-process)	63.78	84.54	69.88	83.08	68.10	80.61	66.56	45.33	72.35	72.36	87.66	84.52	83.48	81.03	79.73	72.78	23.78
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
baseline (150-75-37-18, lessfilters)	57.07	84.40	69.59	82.24	64.65	79.71	63.27	41.57	72.39	72.02	87.97	84.21	83.40	80.21	79.89	71.70	19.46
baseline (150-75-37-18, concat)	59.37	79.50	53.96	79.03	65.63	76.50	58.78	26.64	66.33	66.57	83.18	73.84	76.32	74.83	74.70	73.24	33.81
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 (w-s-p, c-s-p)	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 (full)	78.46	90.03	80.59	87.20	76.82	88.72	71.64	51.25	80.85	80.93	92.78	90.85	91.18	89.58	89.03	84.53	47.15
Co-CNN (w/o edge, +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
Co-CNN (edge w/o multi-scale fusion, +Chictopia10k)	81.13	90.98	81.07	89.02	81.20	91.52	77.30	60.42	83.51	83.78	93.70	92.32	92.45	90.30	90.20	85.79	51.09
Co-CNN (use edge before prediction, +Chictopia10k)	79.48	90.51	81.29	87.89	80.59	90.10	74.95	58.08	82.37	82.77	93.38	91.99	92.22	90.46	90.06	85.50	50.73
Co-CNN (+Chictopia10k)	**81.83**	**91.41**	**82.23**	**89.91**	**84.17**	**92.88**	80.03	**62.07**	**85.07**	**85.43**	**93.86**	**93.89**	**93.90**	**90.76**	**90.83**	**86.02**	52.20

numbers in each layer. Our off-line experiments have shown that the influence of the exact filter number in each layer is relatively minor. After reducing the filter number by half for each layer, i.e., "baseline (150-75-37-18, lessfilters)", the performance on F-1 score is slightly decreased by $0.9\,\text{percent}$ over "baseline (150-75-37-18)". It demonstrates that the depth of our network is much more critical than the filter number. Fourth, we evaluate the other cross-layer combination method ("baseline (150-75-37-18, concat)"), which concatenates the feature maps from deep and fine layers instead of element-wise summations used in our "baseline (150-75-37-18)". The "baseline (150-75-37-18, concat)" is inferior to "baseline (150-75-37-18)" by $4.9\,\text{percent}$ in terms of F-1 score.

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 versus $72.72\,\text{percent}$ 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 semantic edge context and local super-pixel contexts. After the summation of global image label probabilities ("Co-CNN (summation with global label)"), the performance can be increased by $1.26\,\text{percent}$, 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 $1.84\,\text{percent}$ 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 $0.71\,\text{percent}$ increase in F-1 score, shown as "Co-CNN (concatenate, summation with global label)". The gradually improved performance validates that incorporating the predicted global image label probabilities into multiple convolutional layers and the label confidence maps can help achieve better pixel-wise classification. The most significant improvements over "baseline (150-75-37-18)" can be observed from the F-1 scores for clothing items, e.g., $7.68\,\text{percent}$ for skirt and $4.32\,\text{percent}$ 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. For example, the dress is often confused with upper-clothes

TABLE 5
Comparison of Parsing Performance with Three State-of-the-Arts on the Test Images of Fashionista [8]

Method	Acc.	F.g. acc.	Avg. prec.	Avg. recall	Avg. F-1 score
⋆ Yamaguchi et al. [8]	87.87	58.85	51.04	48.05	42.87
⋆ PaperDoll [4]	89.98	65.66	54.87	51.16	46.80
⋆ ATR [1]	92.33	76.54	73.93	66.49	69.30
Co-CNN (w/o edge)	96.08	84.71	82.98	77.78	79.37
Co-CNN (full)	96.59	86.46	86.26	81.14	82.82
Co-CNN (w/o edge, +Chictopia10k)	97.06	89.15	87.83	81.73	83.78
Co-CNN (+Chictopia10k)	**97.64**	**90.85**	**88.55**	**85.93**	**87.08**

and the skirt, and upper-clothes often appear together with skirt and pants.

Local super-pixel contexts. Extensive evaluations are conducted on the effectiveness of using local super-pixel contexts. All the following results are obtained without performing semantic edge prediction. The average F-1 score increases by 0.9 percent 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 version "Co-CNN (w-s-p, c-s-p)" leads to 1.68 percent increase. For the F-1 score for each semantic label, the significant improvements are obtained for the labels of small regions (e.g., hat, sun-glasses and scarf). For instance, the F-1 score for hat is increased by 7.02 percent, and 9.19 percent 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, which is separate with the network optimization and feature learning. To justify the superiority of using the local super-pixel contexts during the training, we test the performance of only using the local super-pixel smoothing and voting as the post-processing steps, i.e., "baseline (150-75-37-18, post-process)" is shown in Table 2. Compared to our "Co-CNN (w-s-p, c-s-p)", the average F-1 score of "baseline (150-75-37-18, post-process)" is decreased by 3.7 percent. When only performing the local super-pixel smoothing and neighborhood voting as the post-processing steps on the "Co-CNN (concatenate, summation with global label)", the F-1 score drops by 1.68 percent compared with "Co-CNN (w-s-p, c-s-p)".

TABLE 6
Performance Analysis on Three Main Influential Factors Affecting Human Parsing, i.e., Diverse Poses, Background Clutters and Viewpoints

Test set	Pose	Background clutters	Viewpoint
ATR (easy)	88.32	87.42	88.29
Chictopia10k (hard)	82.18	83.12	78.10

All results are evaluated by average F-1 score metric.

Semantic edge context. We validate the effectiveness of incorporating semantic edge context into Co-CNN on ATR dataset and Fashionista dataset. The detailed comparison results are reported in Tables 2, 4 and 5. By jointly performing semantic edge prediction, the performance in average F-1 score by "Co-CNN (full)" can be significantly boosted by 4.77 percent compared to "Co-CNN (w-s-p, c-s-p)" on ATR dataset. A similar significant increase in F-1 score can also be observed when evaluating on Fashionista dataset, i.e., 82.82 percent of "Co-CNN (full)" versus 79.37 percent of "Co-CNN (w/o edge)". By training with more data in "Chictopia10k", a 5.22 percent increase in F-1 score can be observed when comparing "Co-CNN (+Chictopia10k)" with "Co-CNN (w/o edge, +Chictopia10k)" on ATR dataset.

In addition, we also conduct experiments on different variants of using semantic edge context. Some interesting observations can be obtained. First, the effectiveness of using multi-scale prediction streams to predict semantic

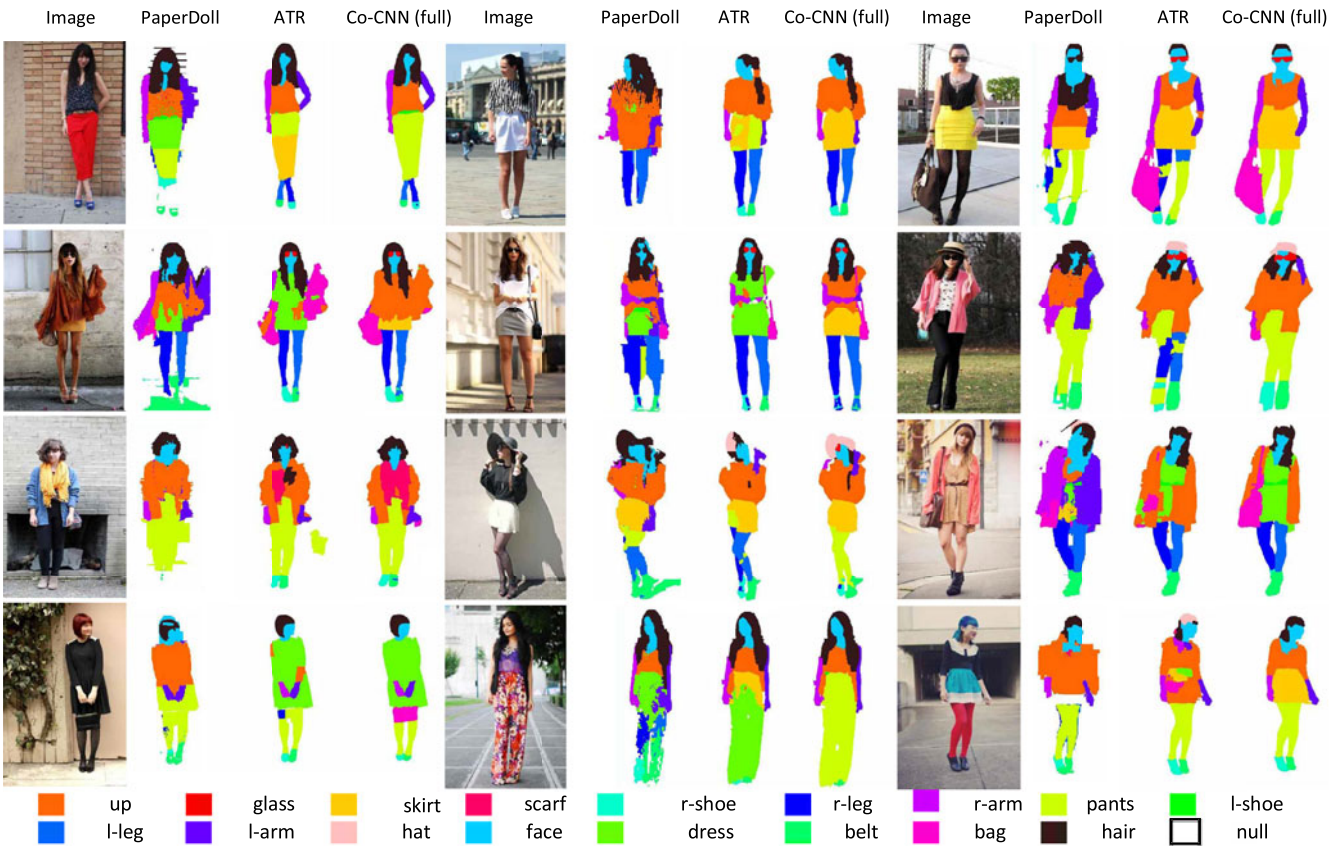


Fig. 6. Result comparison of our Co-CNN and two state-of-the-art methods. For each image, we show the parsing results by PaperDoll [4], ATR [1] and our Co-CNN sequentially.

edges can be observed when comparing "Co-CNN (edge w/o multi-scale fusion, +Chictopia10k)" with "Co-CNN (+Chictopia10k)". The semantic edge is directly predicted from the convolutional layers with resolution of 18×12, and then the predicted confidence maps are combined into the features in later layers. This demonstrates that the multi-scale prediction is critical for fine-grained semantic edge prediction where the local details and semantic information are both captured. All outputs from different convolutional layers with different resolutions and the weighted merging of multi-scale predictions can contribute to the results. Second, we also report the performance of only embedding edge confidence maps in the final prediction layer ("Co-CNN (use edge before prediction, +Chictopia10k)") instead of three convolutional layers in "Co-CNN (+Chictopia10k)". We find combining edge confidence maps into more convolutional layers yields better performance than only using in the prediction layer. This is because the hierarchical feature combinations can lead to better features serving for the final prediction.

4.4 Discussion on Different Test Set

To further facilitate the research in human parsing and help to identify the most promising directions of current methods, we conduct the extensive experiments to evaluate how much different factors such as challenging poses, background clutters and viewpoints influence the final results. All results are reported in Table 6 and evaluated by the average F-1 score metric. For each influential factor, we manually select 100 hard images from Chictopia10k dataset and 100 easy images from ATR dataset as two different test set. In terms of the selection policy, we first evaluate the prediction accuracies for all images in two dataset using our model and the images predicted with low accuracies form the candidate set. Based on this candidate set, the images with twisty arms and legs are regarded as containing hard poses in Chictopia10k dataset. The images with the confusing boundaries between the human body and background are treated as the hard images with large background clutters. The images with lying or sitting people are selected as hard images with diverse viewpoints. The easy subsets in ATR dataset can thus be accordingly identified. The training set includes the rest of ATR dataset and Chictopia10k dataset. As observed from Table 6, all the three factors decrease the parsing performance by a large margin. Among them, the most influential factor affecting the parsing performance is the challenging viewpoints (such as lying person or sitting person). The possible reason for such performance decrease is the lack of enough training images for such challenging images and powerful network capability.

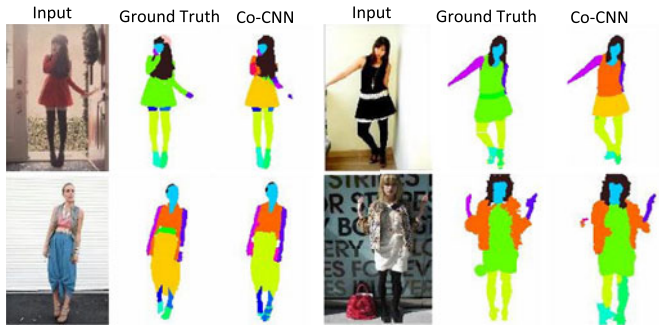


Fig. 7. Some failure cases generated by our Co-CNN.

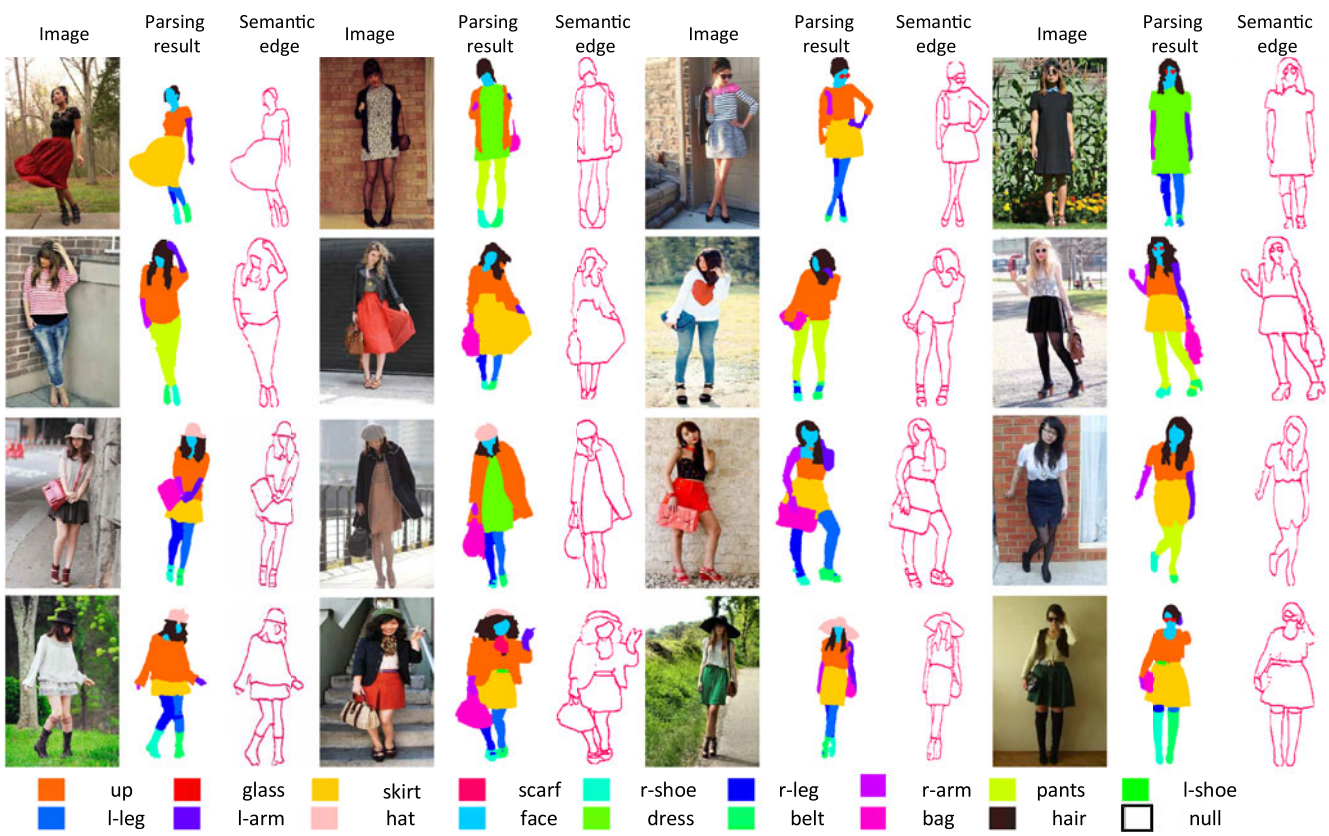


Fig. 8. Some example parsing results and semantic edge prediction results of our Co-CNN.

4.5 Visual Illustration

The qualitative comparison of parsing results is visualized in Fig. 6. Our Co-CNN outputs more meaningful and precise predictions than PaperDoll [4] and ATR [1] despite the large appearance and position variations. Our method can successfully predict the labels of small regions (e.g., hat, scarf, sun-glasses, belt) and the label boundaries are preserved very well. The parsing results of our method are much more complete and cleaner while the results of [4] may be fragmented. The results of [4] may be influenced by the low-level information, e.g., image clarity and color similarity, while the method of [1] shows to suffer from the incorrectly classified label masks and predicted locations of each label.

We also show the parsing results and semantic edge prediction results by the proposed Co-CNN in Fig. 8. It can be observed that the predicted semantic edge can well capture the detailed boundaries and semantic meaning, even for small regions such as scarf, sunglasses and hat.

Finally, we show some representative failure cases by our Co-CNN in Fig. 7, which can help identify some possible directions to improve the human parsing performance in future. As observed from these qualitative results, Co-CNN may fail to parse the challenging images with confusing clothes items, large background clutters and poses.

5 Conclusions and Future Work

In this work, we proposed a novel Co-CNN architecture for the human parsing task, which integrates the cross-layer context, global image label context, semantic edge context and local super-pixel contexts into a unified network. For each input image, our Co-CNN produces the corresponding-sized pixel-wise predictions in a fully end-to-end way. The local-to-global-to-local hierarchy is used to combine the local detailed information and the global semantic information. The global image label prediction, semantic edge 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. In future work, we will further extend our Co-CNN architecture for generic image parsing tasks, e.g., object semantic segmentation.

Acknowledgments

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 2014A030313201, in part by the Program of Guangzhou Zhujiang Star of Science and Technology under Grant 2013J2200067, and in part by Guangdong Science and Technology Program under Grant 2015B010128009. This work was also partly supported by gift funds from Adobe Research. Liang Lin is the corresponding author.

References

[1] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," *IEEE Trans. Pattern Recog. Mach. Intell.*, vol. 37, no. 12, pp. 2402–2414, Dec. 2015.

[2] 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," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1419–1427.

[3] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan, "Style finder: Fine-grained clothing style detection and retrieval," in *Proc. Comput. Vis. Pattern Recog. Workshops*, 2013, pp. 8–13.

[4] K. Yamaguchi, M. Kiapour, and T. Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 3519–3526.

[5] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3586–3593.

[6] Y. Wang, D. Tran, Z. Liao, and D. Forsyth, "Discriminative hierarchical part-based models for human parsing and action recognition," *J. Mach. Learning Res.*, vol. 13, no. 1, pp. 3075–3102, 2012.

[7] Y. Kalantidis, L. Kennedy, and L.-J. Li, "Getting the look: Clothing recognition and segmentation for automatic product suggestions in everyday photos," in *Proc. ACM Conf. Int. Conf. Multimedia Retrieval*, 2013, pp. 105–112.

[8] K. Yamaguchi, M. Kiapour, L. Ortiz, and T. Berg, "Parsing clothing in fashion photographs," in *Proc. Comput. Vis. Pattern Recog.*, 2012, pp. 3570–3577.

[9] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan, "A deformable mixture parsing model with parselets," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 3408–3415.

[10] W. Yang, L. Lin, and P. Luo, "Clothing co-parsing by joint image segmentation and labeling," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 3182–3189.

[11] M. Dantone, J. Gall, C. Leistner, and L. Van Gool, "Human pose estimation using body parts dependent joint regressors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3041–3048.

[12] J. Carreira and C. Sminchisescu, "CPMC: Automatic object segmentation using constrained parametric min-cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1312–1328, Jul. 2012.

[13] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learning Res.*, vol. 11, pp. 19–60, 2010.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1–9.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Comput. Vis. Pattern Recog.*, 2014.

[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," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 999–1007.

[17] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proc. Int. Conf. Comput. Vis. Pattern Recog.*, 2014.

[18] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Analyisi Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[19] P. George, C. Liang-Chieh, M. Kevin, and A. L. Yuille, "Weakly- and semi-supervised learning of a DCNN for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015.

[20] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3376–3385.

[21] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 447–456.

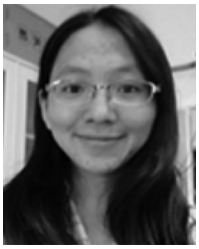
[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015.

[23] G. Bertasius, J. Shi, and L. Torresani, "Deepedge: A multi-scale bifurcated deep network for top-down contour detection," *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015.

[24] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun, "A high performance CRF model for clothes parsing," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 64–81.

[25] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with local-global long short-term memory," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[26] X. Liang, C. Xu, X. Shen, J. Yang, J. Tang, L. Lin, and S. Yan, "Human parsing with contextualized convolutional neural network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015.
[27] S. Liu, X. Liang, L. Liu, K. Lu, L. Lin, X. Cao, and S. Yan, "Fashion parsing with video context," *IEEE Trans. Multimedia.*, vol. 17, no. 8, pp. 1347–1358, Aug. 2015.
[28] X. Liang, Y. Wei, X. Shen, J. Yang, L. Lin, and S. Yan, "Proposal-free network for instance-level object segmentation," *arXiv:1509.02636*, 2015.
[29] X. Liang, Y. Wei, X. Shen, Z. Jie, J. Feng, L. Lin, and S. Yan, "Reversible recursive instance-level object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.
[30] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 297–312.
[31] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," *ICLR*, 2014.
[32] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang, "Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3982–3991.
[33] S. Xie and Z. Tu, "Holistically-nested edge detection," *arXiv:1504.06375*, 2015.
[34] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, "Entropy rate superpixel segmentation," in *Proc. Int. Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 2097–2104.
[35] 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 *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.



Xiaodan Liang is currently working toward the PhD degree at the School of Information Science and Technology, Sun Yat-sen University, China. She is currently working at the National University of Singapore as a research intern. Her research interests include semantic segmentation, object/action recognition, and medical image analysis.

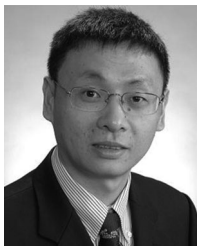
Chunyan Xu received the PhD degree from the School of Computer Science and Technology, Huazhong University of Science and Technology, in 2015. She is a visiting scholar at the National University of Singapore from 2013 to 2015. Her research interests include deep neural network, computer vision, manifold learning, and kernel methods.

Xiaohui Shen received the BS and MS degrees from the Department of Automation, Tsinghua University, China. He received the PhD degree from the Department of Electrical Engineering and Computer Science, Northwestern University, in 2013. He is currently a research scientist at Adobe Research, San Jose, CA. His research interests include image/video processing and computer vision.

Jianchao Yang received the MS and PhD degrees in electrical and computer engineering from the University of Illinois, Urbana-Champaign, Urbana, in 2011. He is currently a research scientist at the Advanced Technology Laboratory, Adobe Systems Inc., San Jose, CA. His research interests include object recognition, deep learning, sparse coding, image/video enhancement, and deblurring.

Jinhui Tang received the BE and PhD degrees from the University of Science and Technology of China, in July 2003 and July 2008 respectively. From July 2008 to Dec. 2010, he worked as a research fellow in School of Computing, National University of Singapore. He is a professor in School of Computer Science and Engineering, Nanjing University of Science and Technology. His current research interests include large-scale multimedia search, computer vision. He has authored more than 100 journal and conference papers in these areas. He received the ACM China Rising Star Award and a co-recipient of the Best Paper Award in ACM Multimedia 2007, PCM 2011 and ICIMCS 2011. He is a member of ACM.

Liang Lin received the PhD degree from the Beijing Institute of Technology, Beijing, China, in 2008. He is a full professor at the School of Data and Computer Science, Sun Yat-Sen University, China. He was a post-doctoral research fellow with the Center for Vision, Cognition, Learning, and Autonamy of University of California, Los Angeles. He is currently an Associate Editor of the *IEEE Transactions on Human-Machine Systems.* His research focuses on new models, algorithms and systems for intelligent processing and understanding of visual data such as images and videos. He has authored more than 100 papers in these areas. He was supported by several promotive programs or funds for his works, such as Program for New Century Excellent Talents of Ministry of Education (China) in 2012, and Guangdong NSFs for Distinguished Young Scholars in 2013. He received the Best Paper Runners-Up Award in ACM NPAR 2010, Google Faculty Award in 2012, and Best Student Paper Award in IEEE ICME 2014.



Shuicheng Yan is currently an associate professor at the Department of Electrical and Computer Engineering, National University of Singapore, and the founding lead of the Learning and Vision Research Group (http://www.lv-nus.org). His research areas include machine learning, computer vision and multimedia, and he has authored/co-authored nearly 400 technical papers over a wide range of research topics, with Google Scholar citation $>$ 12,000 times. He is ISI highly cited researcher 2014, and IAPR Fellow 2014. He has been serving as an associate editor of IEEE TKDE, CVIU, and TCSVT. He received the Best Paper Awards from ACM MM'13 (Best Paper and Best Student Paper), ACM MM'12 (Best Demo), PCM'11, ACM MM'10, ICME'10 and ICIMCS'09, the runner-up prize of ILSVRC'13, the winner prizes of the classification task in PASCAL VOC 2010-2012, the winner prize of the segmentation task in PASCAL VOC 2012, the honorable mention prize of the detection task in PASCAL VOC'10, 2010 TCSVT Best Associate Editor (BAE) Award, 2010 Young Faculty Research Award, 2011 Singapore Young Scientist Award, and 2012 NUS Young Researcher Award.