The AAAI-17 Workshop on
Crowdsourcing, Deep Learning, and Artificial Intelligence Agents
WS-17-07

Real-Time Fashion-Guided Clothing Semantic Parsing: A Lightweight Multi-Scale Inception Neural Network and Benchmark

Yuhang He,[1] **Lu Yang,**[2] **Long Chen**[1]
[1]School of Data and Computer Science
Sun Yat-Sen University, Guangzhou, P.R. China
[2] PRMCT Lab
Beijing University of Posts and Telecommunications, Beijing, P.R. China

Abstract

Currently two barriers exist that sabotage clothing semantic parsing research: existing methods are time-consuming and the lack of large publicly available dataset that enables parsing at multiple scales. To mitigate these two dilemmas, we hereby embrace deep learning method and design a lightweight multi-scale inception neural network which is at both inside and outside multi-scale inception during training. Moreover, atrous convolution block is involved to enlarge the field of view while bringing neither extra computation cost nor parameters. Then the pre-trained model is further pruned and compressed by fine-tuning on a lightweight version of the same network used earlier, in which the inactive feature response and connections below a pre-defined threshold are directly removed. Besides, we construct so far the largest fashion guided clothing semantic parsing dataset (FCP) which contains a total of 5,000 clothing images and each image associates with both pixel-level, object-level and image-level annotations. All clothing in the dataset are recommended by fashion experts or trendsetters and contains as many as 65 common clothing items, accessories. We organize the dataset as Wordnet tree structure so that it enables fashionably parsing hierarchically. Finally, we conduct extensive experiments on three currently available datasets. Both quantitative and qualitative results demonstrate the priority and feasibility of our method, comparing with several other deep learning based methods. Our method achieves 35 FPS in a single Nvidia Titian X GPU with only minimal accuracy loss.

1 Introduction

Semantic parsing, being likened to the holy grail in computer vision community, is a way to understand visual input at a much higher-level way than traditional tasks like object detection or classification. It remains as a challenging task because it tries to assign each individual pixel with a label. The flourish of an artificial intelligence approach dubbed as deep learning in recent years has paved a new way to solve this problem: they broadly depend on deep convolutional neural networks (CNNs) to directly learn the correlation between the color image input and the parsing result output (Noh, Seunghoon, and Bohyung 2015)(Noh, Hong, and Han 2015)(Chen et al. 2015). At the same time, the emergence of various large scene parsing datasets, such as COCO (Lin et al. 2014), VOC (Everingham et al. 2010) and Cityscapes (Cordts et al. 2015), enormously boosted scene semantic parsing research.

As a special branch of image semantic parsing research, clothing semantic parsing has received little attention. The reasons are twofold: on the one hand, clothing semantic parsing has been shown interest mostly by clothing e-commerce companies, like eBay. They require semantic parsing methods to be fast and accurate enough so that they can deploy them on mobile devices to attract more users. However, existing methods (Noh, Seunghoon, and Bohyung 2015)(Noh, Hong, and Han 2015)(Chen et al. 2015) are either time-consuming or too sophisticated. From commercial application scenario perspective, a real-time clothing semantic parsing user experience sometimes is more desirable than parsing accuracy. On the other hand, unlike natural scene parsing task that can turn to several large public datasets, so far there is no large dataset which enables to comprehensively evaluate various clothing parsing algorithms. Existing available dataset including CCP dataset (Yang, Luo, and Lin 2014) and Fashionista dataset (Yamaguchi, Kiapour, and Berg 2013) are either too small in size or poorly annotated so that they is far from satisfying delving deeper into clothing semantic parsing.

To mitigate the two dilemmas above, we hereby propose a lightweight multi-scale inception neural network which achieves real-time clothing semantic parsing without much accuracy loss, and further introduce an up to date the largest and also the most comprehensive fashion guided clothing semantic parsing dataset. We narrow down the clothing semantic parsing to be "fashion-guided" so that we can depend on fashionable outfits recommended by fashion experts or trendsetters. Actually, fashion related research has received much attention in recent years, spanning from fashion feature learning (Serra and Ishikawa 2016), fashion style analysis (Kiapour et al. 2014)(Serra et al. 2015)(Yamaguchi, Kiapour, and Berg 2013)(Vittayakorn et al. 2015) to fashion likeability prediction(Wang et al. 2015)(He, Lin, and McAuley 2016). Nevertheless, fashion-centric clothing parsing is, somewhat, still an uncharted area. Incorporating the two main reasons discussed above, another main reason lies in the fact the fashion is too subjective and abstract to be efficiently modelled by machines.

In this paper, we builds our work on currently popu-

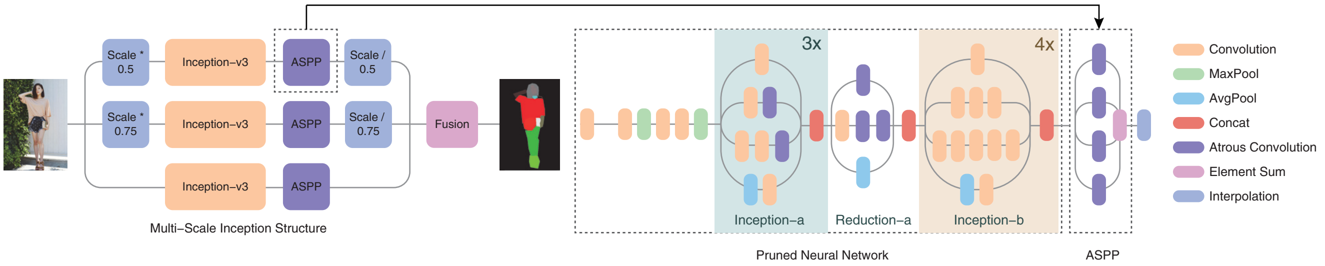

Figure 1: Framework illustration. The learning structure consists of three inception-v3 networks and ASPP module in parallel. Atrous convolution is added all through the structure to enlarge the field of view. The trained model is further compressed and pruned according the feature response and dense connection weight during the fine-tune process. 3x and 4x means repeating the block 3 times or 4 times, respectively. Please note the different operation different color represents and the network structure difference between our pruned and compressed model and the original inception-v3 network.

lar methods in natural scene parsing (Noh, Seunghoon, and Bohyung 2015)(Noh, Hong, and Han 2015)(Chen et al. 2015)(Chen et al. 2016), which overwhelming rely on the new emerging artificial intelligence method dubbed as deep learning to directly learn the internal inference between the original color image and parsing results. Deep learning, especially the Convolutional Neural Networks (CNNs), has dominated various image-centric tasks and shown state-of-the-art performance in classic tasks including classification (Kriszhevsky, Sutskever, and Hinton 2013) and detection (Ren et al. 2015). Unlike conventional tasks that highlight compressed high-level feature representation and transformation invariance, semantic parsing lays emphasis on the same-size mapping, multi-scale object perception and precise localization accuracy. To this end, existing deep learning based methods try to simultaneously learn the compressed feature representation and reconstruct the final parsing image through a stack of downsampling and upsampling neural network, respectively. In upsampling structure, deconvolution (Noh, Seunghoon, and Bohyung 2015) and bilinear interpolation operation have been efficiently employed. As to precise localization, additional global refinement methods like fully connected conditional random field (CRF) (Kraehenbuehl and Koltun 2011) are often adopted as post-processing refinement.

Embracing clothing semantic parsing real-time application requirement, we design a simple yet powerful end-to-end trainable neural network that dynamically learns a clothing image at multiple scales as well as large filed of views in the training stage. Atrous convolution is adopted here to enlarge the field of view without increasing the computation burden nor introducing extra parameters. In order to overcome the scale variability hurdle, not only the input image is tripled to multiple scales (1x, 0.75x, 0.5x in this paper) before feeding to neural network, but a special block called atrous spatial pyramid polling (ASSP) is also inserted to the neural network to percept neural network intermediate layer at multiple scales both inside and outside. Then the pre-trained model is further fine-tuned by the same training dataset on a much lightweight version of the same network it gets trained earlier. The model is further pruned and compressed by removing inactive feature response and connections below a threshold so we achieve real-time processing without explicit parsing accuracy loss.

As to fashion-guided clothing semantic parsing dataset, we introduce 5,000 carefully labelled fashion-centric images. Unlike existing clothing parsing datasets, we take accessories, body torso parts with different spatial locations into account and propose so far the largest items number for fine-grained parsing research. In addition to pixel-level segmentation, object-level and image-level annotations are also provided for motivating more research. Details as well as comparison with other datasets would be thoroughly discussed later. In sum, the main contribution of this paper lies in: a lightweight multi-scale inception neural network that enables real-time clothing semantic parsing with 30 FPS on a Nvidia Titian X GPU. Besides, a large fashion-guided clothing semantic parsing dataset is introduced with the original intention to motivate more research on clothing parsing. This dataset outperforms other relevant datasets at a large scale from various aspects.

2 Multi-Scale Inception Neural Network

We build our lightweight multi-scale inception neural network on two well-established neural networks: inception-v3 (Szegedy et al. 2015) and atrous convolution block (Chen et al. 2016). Our multi-scale inception design derives from two parts: outside multi-scale and inside multi-scale. In the outside multi-scale module, we additionally rescale the original input image with two factors (0.75x and 0.5x) to generate two scaled images and feed them together to the triple parallel inception-v3 neural networks sharing the same parameters. Finally, the two scaled score maps are rescaled to the original size and aggregated together with the non-scaled original score map via linearly max-elementwise operation. Note that forcing neural network to receive the same image but with different scales enables to partially overcome the hurdle that the same clothing item appears in different sizes on different images. The reason why we choose inception-v3 is that its inception module is multi-scale oriented by nature and it trains much faster than deeper neural network like ResNet101 (He, Zhang, and Ren 2016) with subtle or even no accuracy loss.

As to inside multi-scale, we turn to the current powerful atrous convolution operation, which has been successfully employed for enlarging the field of view without adding extra parameters nor computation burden. Atrous convolution evolved from undecimated wavelet transform (Holschneider et al. 1989) for upgrading computation efficiency. In deep learning based image parsing domain, atrous convolution purposedly inserts zero values between a filter's two values to enlarge filter size. The flexibility of insertion enables to percept any intermediate layer at any field of view. Besides, the zero based insertion mechanism introduces neither extra computation cost nor parameters, which is conversely notorious for deconvolution operation (Noh, Seunghoon, and Bohyung 2015).

Considering any given value $x[i]$, atrous convolution convolves it within a pre-defined length N with a filter $w[n]$ to get the output value $y[i]$,

$$y[i] = \sum_{n=1}^{N} x[i + r \cdot n]w[n] \tag{1}$$

where r is the rate parameter controlling the stride step size we use to convolve the initial input. Altering r dynamically simulates various atrous convolution zero insertion schemes, thus resulting in various filed of views. If $r = 1$, atrous convolution collapses into standard convolution. In general, a serial of atrous convolutions are consecutively added the last several layers of inception-v3 to upsample the compressed representation to the original image's resolution.

Inspired by the research in R-CNN (He et al. 2014) that regions of arbitrary scale can be efficiently handled by rescaling intermediate convolutional layer, we adopt atrous spatial pyramid pooling (ASPP) which is originally proposed by L. Chen *et al.* (Chen et al. 2016) to insert multiple atrous convolutions in parallel to extract an intermediate convolution layer at multi-scales. The ASPP module here provides inside multi-scale and we experimentally find it excels at balancing the mutual scale variation of different items within an image.

The graphic illustration is shown in Fig. 1. The initial image together with its two accompanying rescaled versions is fed to the inception-v3 neural network in parallel, part of the standard convolution operations within which are replaced by atrous convolution. ASPP module is then linked to inception-v3 to infer the parsing results in an inside multi-scale and parameter sharing manner. Finally, the output parsing result is bi-linearly interpolated to the same size with the original color image.

With the pre-trained model, we prune and compress the original cumbersome neural network to much lightweight network by directly pruning those inactive responses and part of full connections to reduce the parameters. Specifically, we directly cut off the the reduction_b and incpetion_c module from inception-v3 neural network, then fine-tune the remaining network and further compress the model by removing the connections whose connection weights are below a predefined threshold (see Fig. 2). Model compression has been effectively explored in recent years (Han, Mao, and Dally 2016). Research shows that appropriate compressed model achieves no obviously accuracy loss but maximumly reduces the model size. By following the compression method proposed by S. Han *et al.* (Han, Mao, and Dally 2016), we reduce the size of our model to 7x scale so that it achieves real-time application without much parsing accuracy loss. We will thoroughly discuss it in the experiment section.

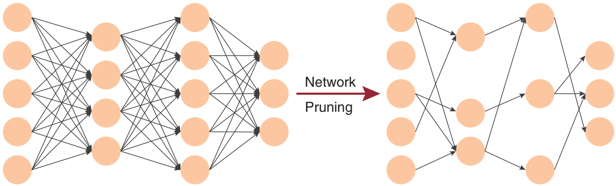

Figure 2: Model compression: by retraining the model, the initially dense connected network can be sparely connected by removing the low-weight connections.

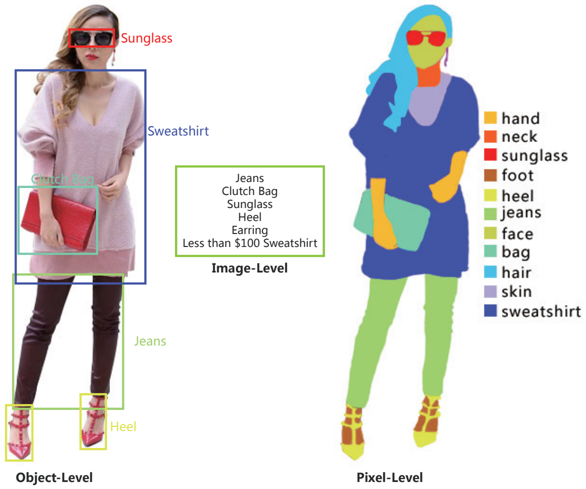

Figure 3: Sample image: Left: the object-level annotation result. Middle: the image-level annotation result which generally shows the detailed item pool. Extra information such as price tag and bag type is available. Right: pixel-level result.

3 FCP: Fashion-Guided Clothing Semantic Parsing Dataset

3.1 Dataset Detailed Introduction

As discussed above, one motivation of this paper is to introduce a large fashion-guided clothing semantic parsing dataset. To this end, we first crawled 5,000 high-quality and fashion experts recommended images from www.chictopia.com. Chictopia is the world's largest fashion style community where bloggers or trendsetters share their style posts and online clothing boutiques sell to the most fashion-forward audience. This guarantees all collected images are fashion-oriented. In addition, excessive annotations are available for most images, including fine-grained clothing item name, item branch name, color information, price tag, etc. This ad-

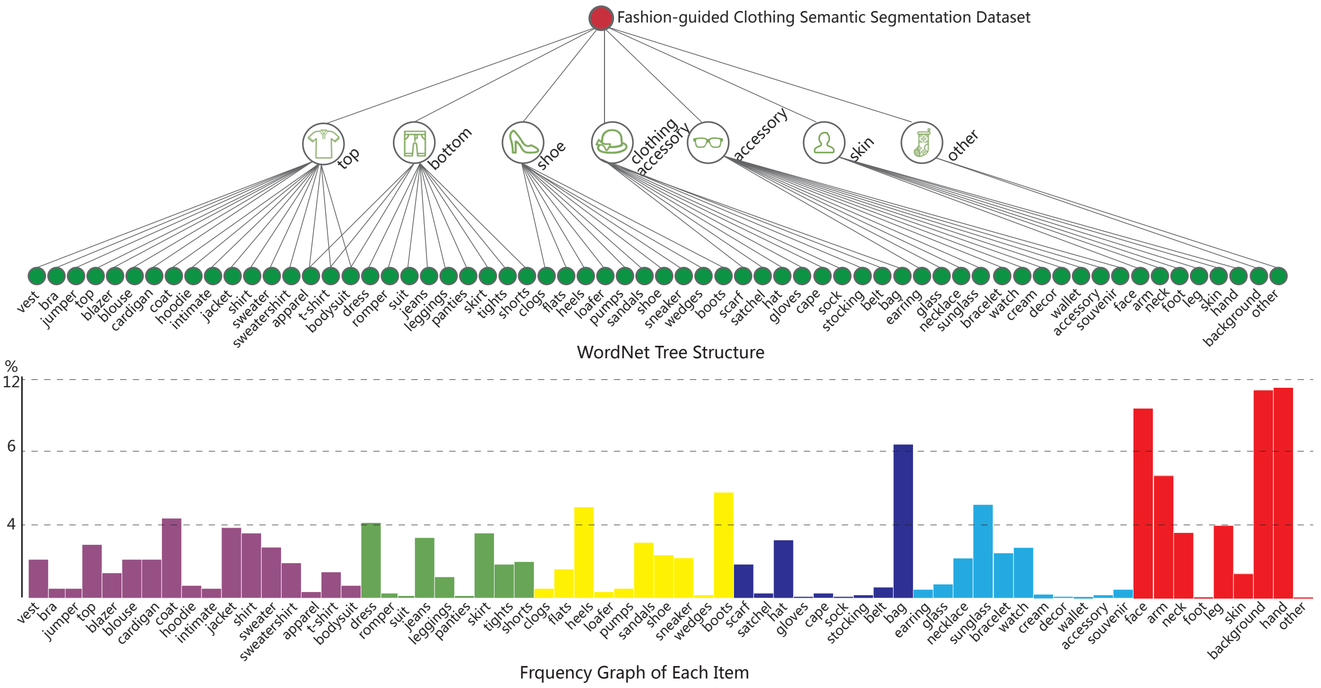

Figure 4: Top: Fashion-guided clothing semantic segmentation dataset WordNet tree structure: each node in the hierarchical tree is depicted by a bunch of relevant items. Bottom: the frequency of all items in the dataset. Note that the ordinate is not linearly coordinated for better visualization.

vantage helps to enrich dataset annotation pool from various perspectives. Therefore, we have created a total of 64 clothing items excluding the background, ranging from all daily clothing items to accessories. This dataset, to our best knowledge, is so far the largest dataset in terms of clothing semantic parsing.

Each image in Chictopia initially associates with multiple attributes, yet these attributes sometimes are incomplete and even erroneous. Thus dataset cleaning and validation in advance are essential. For each image, we provide both pixel-level, object-level and image-level annotations. We first recruit three experienced labellers and train them to familiarize themselves with all clothing item names in advance. For example, since loafer is a special kind of shoes, the labellers have to figure out what makes loafer and the subtle difference between loafer and other types of shoes, such as flats, boots and clogs. The prior training is prerequisite because labellers have to fill the blank or correct the mislabeled annotations under the circumstances of incomplete or erroneous annotations (dataset cleaning and validation work). After training, labellers label these images both pixel-wisely and object-wisely with the tool written by ourselves. For pixel-wise annotation, they label the contour of each single item. In case of item overlap, the overlapping area is labeled to the item category lying atop. The object-level is bounding box like annotation. That is, each clothing item or accessory is labelled by a tight bounding box indicating its location and category name. Image-level annotation corresponds to a set of item name, brand name and color information describing the whole image. Note that we only leverage the pixel-level annotation in this paper to evaluate our proposed framework. Object-level and image-level are crucial to motivate more fascinating research. For instances, bounding box supervised image semantic segmentation (Dai, He, and Sun 2015) and image tagging (Chen, Zheng, and Weinberger 2013), fashion outfits recommendation. A sample image with the three annotations result is shown in Fig. 3, in which the body torso is spatially divided into sub-parts, including face, neck, leg and foot, even though the mutually share visual similarities.

Similar to other natural scene parsing datasets construction process (Zhou et al. 2016)(Martin et al. 2001), we unavoidably encountered three ambiguous problems during dataset construction: boundary ambiguity, item naming ambiguity and saving-deleting ambiguity:

- **boundary ambiguity**: ambiguity around boundary naturally arises as different labellers exhibit different preference towards boundary during labelling. To reduce this ambiguity, we calculate the ratio of inconsistent labelled pixels around the item boundary over the whole image for the three labellers and require to relabel the image if the ratio exceeds 20%.
- **naming ambiguity**: naming ambiguity derives from two ways: Chictopia provided item naming could be either erroneous or non-specific. For instance, a bodysuit can be classified as both top and bottom (trouser). The missing item naming also easily leads to labeller naming ambi-

Instance	Pixel-Level	Object-Level	Image-Level	Item Category No.	Item No.	Data Source
CCP(2014)	1098	N	1000	59	17897	Unknown
Fashionista(2013)	685	N	N	56	9876	Chictopia
FCP (Ours)	5,000	5,000	5,000	65	72,340	Chictopia

Table 1: Comparison between FCP dataset and CCP (Yang, Luo, and Lin 2014), Fashionista (Yamaguchi, Kiapour, and Berg 2013) datasets.

guity because an item can easily be classified as several similar fine-grained categories. To minimize this ambiguity, we require all labellers to reach a consensus before any naming ambiguity happens or give the item a much more coarse-grained name in order to guarantee its naming correctness.

- **saving-deleting ambiguity**: this happens when an item is too small to be efficiently labelled or largely occluded by other items. For example, a necklace may be too thin to be labelled from neck or it blends well with neck skin color. For items in this case, we choose to neglect them to keep the whole image's conformity.

In sum, we have labelled 50,000+ items, averagely an image associates with about 10 items. We follow the ImageNet dataset (Russakovsky et al. 2015) to organise our dataset in WordNet hierarchical tree structure, in which each node corresponds to a bunch of relevant items. The WordNet tree is organized coarse to fine from top to bottom, we first roughly group all items into seven main categories: top, bottom, shoe, clothing accessory, accessory, skin and others according to their spatial locations or functionalities. Each node further corresponds to numerous fine-grained items. One benefit of this hierarchical organization is that it enables to evaluate parsing algorithms at different granularities. The WordNet tree structure and item's frequency graph is given in Fig. 4. We can see some items overlap between top and bottom. Clothing accessory indicates items mainly made by cloth, while accessory indicating jewellery based items, such as earrings, necklace and bracelet.

3.2 Comparison with other Datasets

We compare our FCP dataset with two other existing datasets: CCF dataset (Yang, Luo, and Lin 2014) and Fashionista dataset (Yamaguchi, Kiapour, and Berg 2013). These two datasets are currently available datasets for clothing semantic parsing. We compare from four aspects: annotation richness, item completeness and richness, item total number and data source, the result is shown in Table 1, from which we can obviously see that our FCP dataset far more outnumbers the other two datasets in terms of both pixel-, object- and image-level annotation number. Regarding item category number and item total, our FCP dataset also goes beyond CCP and Fashionista datasets at a large margin. Specifically, rather than generically treating all people's skin as skin, we discriminate it and further divide them into face, neck, leg, foot, hand and skin parts, where skin here indicates body skin not covered by the other five items. In general, the FCP dataset introduced in the paper overwhelmingly outperforms currently existing clothing semantics parsing datasets and we believe it would definitely push forward the clothing parsing research. (Shelhamer, Long, and Darrell 2016)

4 Experiment

We evaluate our proposed framework in both item-level and category-level, in which item-level amounts to the overall 65 items, while category-level only taking the 7 categories into account by hierarchically fusing the 65 items according to the WordNet tree in Fig. 4. Outside our framework, we compare with two Deeplab versions (Chen et al. 2016): inception_v3 (Szegedy et al. 2015), residual network 101 layers (He, Zhang, and Ren 2016). fully connected neural network (FCN-8s) (Shelhamer, Long, and Darrell 2016), dilated convolution (Yu and Koltun 2016) based on VGG16 (Simonyan and Zisserman 2016) (Dilation-VGG16). The four methods involved here overwhelmingly based on deep learning and have shown promising result on various natural scene parsing datasets. We do not compare with traditional semantic parsing methods, like the combination of exemplar-SVM and graph cut proposed by W. Yang *et al.* (Yang, Luo, and Lin 2014), because the main goal of this paper is to design a lightweight deep neural network so that it can reach real-time application without explicit parsing accuracy loss. Thus we choose to compare beneath the deep neural network framework. The evaluation metric we embrace here includes pixel accuracy measuring the ratio of pixels being correctly parsed and the mean intersection over union (mIoU) measuring the the ratio of intersection between parsing result and parsing ground truth over their union combination. Pixel accuracy and mIoU are classic metrics for scene semantic parsing and usually the higher of the two values, the better parsing result it gets.

The 5,000 FCP images are divided into 4,000, 500, 500 for train/validation/test respectively. The model is train on Caffe (Jia et al. 2014) deep learning framework. To test our proposed framework's sensitivity regarding the input image size, we differentiate the input image scale: one is 600×400 and the other is 300×200. The quantitative result is shown in Table 2, from which we can see that all methods' performance on category-level far outweighs the performance on item-level by a margin of about 0.4 for pixel accuracy and 0.3 for mIoU. The reason is twofold: the large fine-grained items' number imbalance and the huge differences between various items' sizes in an image. For example, usually in an image, the clothing items have occupied most of the image area than accessories such as necklace, earrings and bracelet. We argue that an appropriate solution to this problem might have to rely on both top-to-bottom and bottom-to-

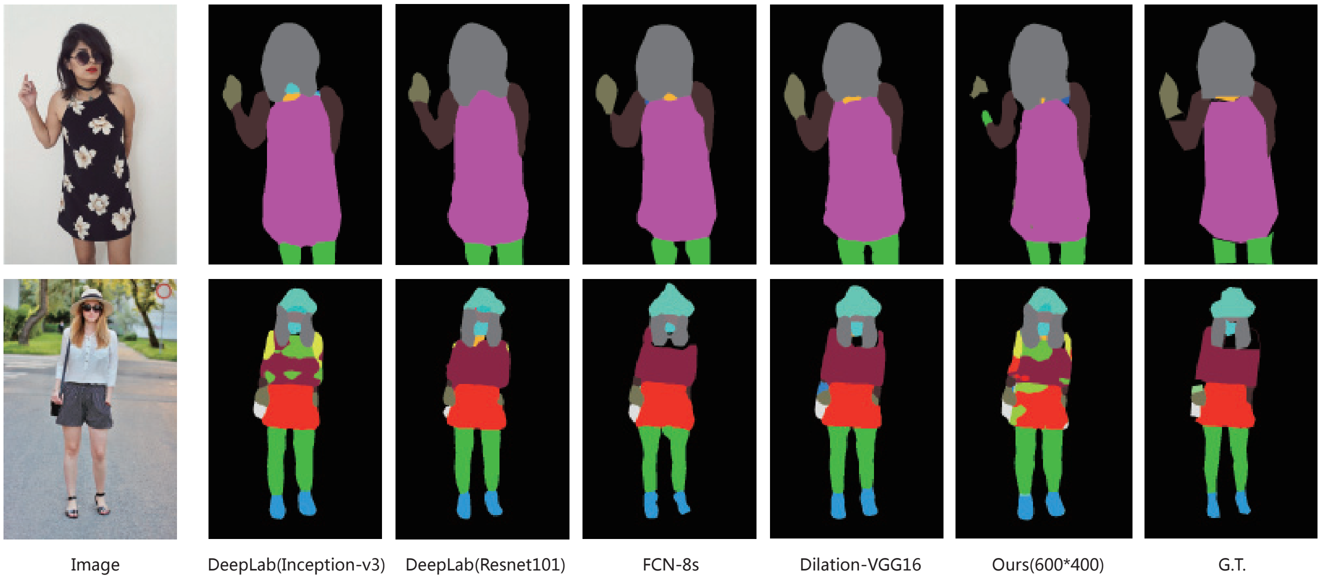

Figure 5: Samples parsing results by various methods.

Level	Item-level		Category-level		Speed (FPS)
Metric	mIoU	Pixel Accu.	mIoU	Pixel Accu.	
DeepLab (inception_v3)	0.2932	0.7701	0.6998	0.9626	2.92
DeepLab (ResNet101)	0.2958	0.7775	0.7033	0.9615	2.04
FCN-8s	0.2611	0.7361	0.6755	0.9511	5.45
Dilation-VGG16	0.2730	0.7389	0.6813	0.9497	3.41
Ours (600 × 400)	0.2736	0.7434	0.6815	0.9624	35.09
Ours (300 × 200)	0.2602	0.7251	0.6449	0.9550	68.98

Table 2: Quantitative result on FCP dataset in terms of pixel accuracy and mIoU. Note that, comparing with the other four methods, our proposed method achieves much faster speed without obvious pixel accuracy nor mIoU loss.

top methods or hierarchical methods treating items of various sizes differently. Besides, our pruned and compressed model achieves the fastest processing time comparing with the other four methods, enabling real-time application in a single Nvidia Titan X GPU processing unit. Moreover, the processing speed is almost doubled if the input image is rescaled from 600 × 400 to 300 × 200, which indicates the processing speed increases exponentially while the input image's scale reduces. This is especially a good news for clothing semantic parsing commercial application.

Note that we trained model separately on CCF dataset (Yang, Luo, and Lin 2014) and Fashionista dataset (Yamaguchi, Kiapour, and Berg 2013), we get slightly lower but also similar quantitative evaluation results with our FCP datasets. The subtle difference, we believe, is caused by the limited amount of dataset. We do not show the detailed results on the two datasets here due to the space limit. Two parsing result sample images are shown in Fig. 5. We can clearly observe that our method achieves comparable parsing result with the other four methods. Moreover, by involving both outside and inside multi-scale inception strategy, our method successfully fused the flower printing with the black dress and parsed the dress as an integral clothing item (sample in the top row). Unfortunately, all methods cannot discriminate small accessories from their neighboring items due to their small sizes (*i.e.* the glass in the bottom sample). We keep this problem open and hope to motivate more research dedicated to resolve it with the help of the FCP dataset.

5 Conclusion

Building on achievements in exploiting deep learning for scene parsing as well as deep model compression and pruning, we have proposed a multi-scale inception neural network to achieve real-time clothing semantic parsing.

6 Acknowledgement

This research is supported by the National Natural Science Foundation of China (NSFC) under grant No. 41401525, the Natural Science Foundation of Guangdong Province under grant No. 2014A030313209, and CCF-Tencent Open Fund under Grant tIAGR20150114.

References

Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2015. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *Proc. ICLR*.

Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2016. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. In *arXiv preprint arXiv:1606.00915*.

Chen, M.; Zheng, A.; and Weinberger, K. 2013. Fast image tagging. In *Proc. ICML*.

Cordts, M.; Omran, M.; Ramos, S.; Scharw´achter, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2015. The cityscapes dataset. In *Proc. CVPR Workshop on The Future of Datasets in Vision*.

Dai, J.; He, K.; and Sun, J. 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proc. ICCV*.

Everingham, M.; Gool, L. V.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*.

Han, S.; Mao, H.; and Dally, W. 2016. Deep compression: Compressi deep compression: Compressing deep neural netoworks with pruning, training quantization and huffmng coding. In *Proc. ICLR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2014. Spatial pyramid pooling in deep deconvolutional networks for visual recognition. In *Proc. ECCV*.

He, R.; Lin, C.; and McAuley, J. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proc. WWW*.

He, K.; Zhang, X.; and Ren, S. 2016. Identity mappings in deep residual networks. In *arXiv preprint arXiv:1603.05027*.

Holschneider, M.; Kronland-Martinet, R.; Morlet, J.; and Tchamitchian, P. 1989. A real-time algorithm for signal analysis with the help of the wavelet transform. In *Wavelets: Time-Frenqucy Methods and Phase Space*, 289–297.

Jia, Y.; Evan, S.; Jeff, D.; Sergey, K.; Jonathan, L.; Ross, G.; Sergio, G.; and Trevor, D. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Kiapour, M.; Yamaguchi, K.; Berg, A.; and Berg, T. 2014. Hispster wars: Discovering elements of fashion styles. In *Proc. ECCV*.

Kraehenbuehl, P., and Koltun, V. 2011. Efficient inference in fully connected crfs with gaussian edge potentials. In *Proc. NIPS*.

Kriszhevsky, A.; Sutskever, I.; and Hinton, G. E. 2013. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; r, P. D.; and Zitnick, C. L. 2014. Microsoft coco: Com- mon objects in context. In *Proc. ECCV*.

Martin, D.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. ICCV*.

Noh, H.; Hong, S.; and Han, B. 2015. Learning deconvolution network for semantic segmentation. In *Proc. ICCV*.

Noh, H.; Seunghoon, H.; and Bohyung, H. 2015. Learning deconvolution network for semantic segmentation. In *Proc. ICCV*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: towards realtime object detection with region propsal network. In *Proc. NIPS*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*.

Serra, E. S., and Ishikawa, H. 2016. Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *Proc. CVPR*.

Serra, E. S.; Fidler, S.; Noguer, F. M.; and Urtasun, R. 2015. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *Proc. CVPR*.

Shelhamer, E.; Long, J.; and Darrell, T. 2016. Fully convolutional networks for semantic segmentation. In *arXiv preprint arXiv:1605.06211*.

Simonyan, K., and Zisserman, A. 2016. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2015. Rethinking the inception architecture for computer vision. In *arXiv preprint arXiv:1512.00567*.

Vittayakorn, S.; Yamaguchi, K.; Berg, A. C.; and Berg, T. L. 2015. Runway to realway: Visual analysis of fashion. In *Proc. WACV*.

Wang, J.; Nabi, A. A.; Wang, G.; Wan, C.; and Ng, T. 2015. Towards predicting the likeability of fashion images. *arXiv preprint arXiv: 1511.05296*.

Yamaguchi, K.; Kiapour, M. H.; and Berg, T. L. 2013. Paper doll parsing: Retrieving similar styles to parse clothing items. In *Proc. ICCV*.

Yang, W.; Luo, P.; and Lin, L. 2014. Clothing co-parsing by joint image segmentation and labelling. In *Proc. CVPR*.

Yu, F., and Koltun, V. 2016. Multi-scale context aggregation by dilated convolutions. In *Proc. ICLR*.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2016. Semantic understanding of scenes through the ade20k dataset. In *arXiv preprint arXiv:1608.05442*.