Semantic Locality-Aware Deformable Network for Clothing Segmentation

Wei Ji[1], **Xi Li**[1,2*], **Yueting Zhuang**[1*], **Omar El Farouk Bourahla**[1], **Yixin Ji**[1], **Shihao Li**[1], **Jiabao Cui**[1]
[1] Zhejiang University, Hangzhou, China
[2] Alibaba-Zhejiang University Joint Institute of Frontier Technologies, Hangzhou, China
{jiwei,xilizju,yzhuang,obourahla,jiyixin,shihaoli,jbcui}@zju.edu.cn

Abstract

Clothing segmentation is a challenging vision problem typically implemented within a fine-grained semantic segmentation framework. Different from conventional segmentation, clothing segmentation has some domain-specific properties such as texture richness, diverse appearance variations, non-rigid geometry deformations, and small sample learning. To deal with these points, we propose a semantic locality-aware segmentation model, which adaptively attaches an original clothing image with a semantically similar (e.g., appearance or pose) auxiliary exemplar by search. Through considering the interactions of the clothing image and its exemplar, more intrinsic knowledge about the locality manifold structures of clothing images is discovered to make the learning process of small sample problem more stable and tractable. Furthermore, we present a CNN model based on the deformable convolutions to extract the non-rigid geometry-aware features for clothing images. Experimental results demonstrate the effectiveness of the proposed model against the state-of-the-art approaches.

1 Introduction

As a challenging problem in computer vision, clothing segmentation has a wide range of applications such as clothing search [Liang *et al.*, 2016a] [Hadi Kiapour *et al.*, 2015] and clothing attribute analysis [Liu *et al.*, 2016] [Veit *et al.*, 2015] [Vittayakorn *et al.*, 2015] [McAuley *et al.*, 2015]. Typically, it is cast as a semantic segmentation problem based on pixel-wise classification into a set of predefined clothing categories, and is implemented by recent studies in an end-to-end deep learning framework [Zhao *et al.*, 2017] [Badrinarayanan *et al.*, 2017]. However, different from generic semantic segmentation [Yang *et al.*, 2017] [Liew *et al.*, 2017], clothing segmentation has the following three characteristics shown in Figure 1: 1) its associated images are human-centric with a great saliency degree; 2) the clothes items (physically attached with fashion models) are non-rigid and geometrically deformable in different directions; and 3) the intra-class appearance variations are extremely large in the forms of different spatial layouts, styles, textures, or materials.

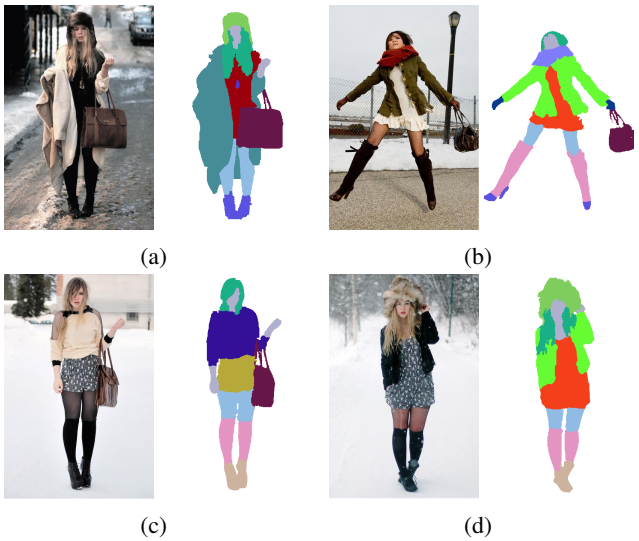


Figure 1: Some example images in the dataset. (a) The structure of the coat is non-rigid. (b) The pose of the model is diverse. The skirt in (c) and the dress in (d) have the same appearance but are classified in different categories.

Therefore, clothing segmentation is a domain-specific problem rather than generic semantic segmentation, such as human parsing which aims at decomposing a human image into semantic body regions such as left-leg, right-leg, etc [Liang *et al.*, 2017] [Gong *et al.*, 2017] [Liang *et al.*, 2016b]. To cope with the aforementioned characteristics for precise clothing segmentation results, the standard solution is to learn the model based on massive training data. However, the pixel-wise label annotations are rather fine-grained and expensive, resulting in the serious lack of sufficient training data. In this case, clothing segmentation has to become a small sample learning problem hindering the power of deep learning.

In the literature, some approaches are proposed to address the small sample learning problem by introducing some extra data or adding some prior rules of label inference. For example, Liang et al. [Liang *et al.*, 2016a] focus on weakly-supervised approach by using SVM to extract semantic con-

*Corresponding authors

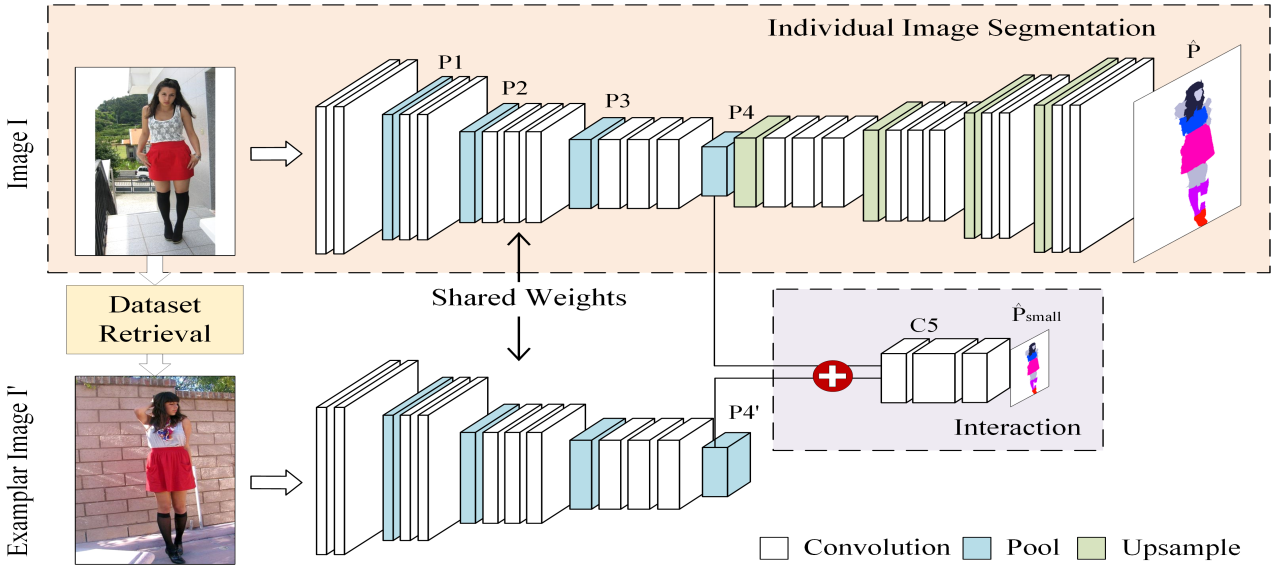


Figure 2: The network structure of the whole framework. Taking one image $\mathcal{I}$ as input, the framework searches the training dataset to obtain a semantic similar image $\mathcal{I}'$ and training the deformable network with image pair ($\mathcal{I}$,$\mathcal{I}'$). When testing, only one image is input into the individual image segmentation pipeline with the learned parameters.

sistent area in image-level images and using Graph Cuts to label the pixels. Tangseng et al. [Tangseng *et al.*, 2017] deal with the problem by adding a new branch which constrains the possibility of items so as to decrease the chances of making errors of semantic matching of cloth items by inference.

In this paper, inspired by the human vision mechanism of visual analogy ability, we propose a semantic locality-aware segmentation model that adaptively attaches an original clothing image with a semantically similar (e.g., appearance or pose) auxiliary exemplar by search. Through the interactions of the clothing image and its exemplar, more intrinsic knowledge about the locality manifold structures of clothing images is discovered to make the small sample learning process more stable and tractable. More specifically, we propose a CNN based on the deformable convolutions in clothing segmentation scheme that takes the clothing image and its exemplar pair as input and outputs the segmentation maps. The learning process of the proposed scheme is only based on the limited training data without using any extra data. In the proposed scheme, the interactions of the clothing image and its exemplar are modeled by a locality-aware reconstruction loss function driven by semantic consistency, which encourages two semantically similar images to have the mutually close deep features. Moreover, the proposed scheme embeds the deformable convolutions into the deep learning process, which makes segmentation flexibly adapt to geometrical clothing deformations. Hence, the main contributions of this work are summarized as follows:

- We propose a locality-aware CNN model driven by semantic consistency with the awareness of the locality manifold structures of clothing images by modeling the interactions of clothing image and its exemplar, which leads to a more stable and tractable small sample learning process.
- We present a deformable version of the proposed CNN model, which takes advantage of the deformable convolutions to extract the geometry-aware features for coping with the non-rigid characteristic of clothing images.

2 Our Approach

2.1 Problem Formulation

Given an input image $\mathcal{I}$, the conventional clothing segmentation solution is based on learning a mapping function $\mathcal{F}$ to output the segmentation map $\mathcal{P}$ with $\mathcal{C}$ predefined categories. In this paper, we aim to deal with the issue of small dataset size by learning the high-level features of both the image $\mathcal{I}$ and a retrieved exemplar image $\mathcal{I}'$. To achieve this, we want the image $\mathcal{I}'$ to be as close in semantics to $\mathcal{I}$ as possible.

When considering a set of clothing images, each image has its context information, which means that similar images in semantic shall have similar segmentation results. The image $\mathcal{I}'$ is chosen based on $\mathcal{I}$. In Section 2.2 we will explain the neighborhood structure that we use in detail.

The design of $\mathcal{F}$ in the framework consists of two parts: individual image segmentation pipeline and image context interaction pipeline. For single image segmentation, we build our segmentation model upon a standard convolutional encoder-decoder network with several convolutional and upsampling layers as shown in the top part of the Figure 2.

The image context interaction part $\mathcal{F}'$ is composed of an encoder sharing the weight parameters with the single image segmentation part, and then followed by an interaction block evaluating the proximity context relationships between the corresponding features generated from both encoders for the image pair $(\mathcal{I},\mathcal{I}')$. Namely, the motivation of this interaction part is to generate mutually close features if the image pair is semantically correlated. As a result, this design aims to make

Figure 3: Semantic similar image pairs by retrieval in CFPD dataset. The appearance of items are similar in (a) and (c), such as the skirts in (a) and pants in (c). The pose of models in (b) and (d) are similar.

feature learning more consistent with semantics. Based on the semantic locality-aware features, the interaction part further carries out a downsampled segmentation task for $\mathcal{I}$ (supervised by the corresponding downsampled segmentation ground truth) to produce a downsampled segmentation map $\hat{\mathcal{P}}_{small}$, which is formulated as:

$$\hat{\mathcal{P}}_{small} = \mathcal{F}'(\mathcal{I}, \mathcal{I}'; \theta') \tag{1}$$

where θ' are the parameters needed to obtain the downsampled output $\hat{\mathcal{P}}_{small}$. Therefore, if the image pair $(\mathcal{I}, \mathcal{I}')$ is very correlated in semantics, the interaction features are mutually reinforced, resulting in the downsampled segmentation performance improvement.

In what follows, we show how to use the context information of the image, how to implement the interaction part in the network design, and how to add deformable information into the mapping function.

2.2 Knowledge-guided Locality-aware Model

Since the clothing items are physically attached to fashion models and vary in styles and textures, the context knowledge of clothing images is manually defined as finding a neighboring image with similar appearances or poses. This is because images with similar poses and appearances also have similar high-level features.

By using the pose and appearance as neighborhood criteria, we obtain the exemplar image $\mathcal{I}'$ that fits our needs as the image has similar segmentation to the original $\mathcal{I}$. Figure 3 illustrates examples of this, the left image in each group is $\mathcal{I}$, and the right is the retrieved exemplar image $\mathcal{I}'$ with similar pose and appearance.

In order to extract the pose information s_{pose} in the images, we use a pre-trained pose parser named OpenPose [Cao *et al.*, 2017]. As for the appearance features $s_{appearance}$, we obtain them from the convolutional layer of OpenPose network.

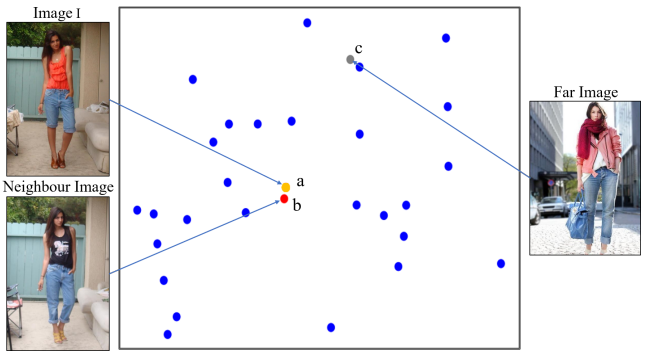


Figure 4: The visualization result of locality in clothing-image-and-exemplar pairs by using the t-SNE. The yellow point a means the query image $\mathcal{I}$ in 2D dimension. The red point b indicates the retrieved exemplar image $\mathcal{I}'$ with similar pose and appearance. The gray point c means image far from $\mathcal{I}$ which is not similar in content.

Then, by concatenating the pose and appearance features as $s_{\mathcal{I}}$ to represent the image $\mathcal{I}$, we use the representative features to retrieve the closest images to it in the dataset. The similarity α of two images can be evaluated by a Euclidean distance between the features of the two images:

$$\alpha = \max(0, 1 - \frac{Euclidean(s_{\mathcal{I}}, s_{\mathcal{I}'})}{Euclidean(s_{\mathcal{I}}, s_{empty})}) \tag{2}$$

The s_{empty} indicates an all-zero feature image with same size of $s_{\mathcal{I}}$. So the score α ranges from 0 to 1 with high scores indicating similar images, and low scores indicating the opposite.

With image pairs training in our network, we have effectively doubled the size of available data compared to the original size of the dataset. Also, it is noted that the matching result is asymmetric. For example, if image B is the closest retrieved exemplar image of image A, it is not necessary that image A is the closest one to image B.

Through using the t-SNE [Maaten and Hinton, 2008] to visualize the features of locality-aware image pairs, Figure 4 shows similar image pairs are truly neighbors in low dimension space. And it is reasonable to reconstruct the segmentation result of image $\mathcal{I}$ with its context information.

With the collected correlated image pairs $\mathcal{I}$ and $\mathcal{I}'$, our network can discover the intrinsic knowledge about manifold structures of clothing images with the semantic consistency and learns from it during the training.

2.3 Exemplar Interaction in the Network

In this section, we will detail the interaction between the image pairs that is input to the framework. As mentioned previously, the interaction of clothing-image-and-exemplar pairs is done by fusing the features of the neighboring images in the network. After several layers of pooling, the feature map is the smallest throughout the network and the semantic information of an image is concentrated in convolutional layer $P4$ as shown in Figure 2.

To use the interaction information, we combine the features of image $\mathcal{I}$ and $\mathcal{I}'$ in a new branch and reconstruct $\hat{\mathcal{P}}_{small}$

to verify that features learned in the semantic-extraction module are appropriate. The combination of features constrains the parameters to take into account the locality described in the neighborhood system. Images with similar semantics are associated with similar features. With the interaction module preserving the locality for the image pairs, we ensure that the training of the parameters is more robust and less likely to overfit, which encourages the network to learn more discriminative and robust features. In order to express the interaction in the network, the feature $s_{interaction}$ in layer $conv5$ is obtained with:

$$s_{interaction} = (1 - \alpha) * s_A + \alpha * s_B, \tag{3}$$

where s_A and s_B are the features obtained at the end of the encoder networks, s_A corresponding to the features obtained from the individual segmentation encode network and s_B from the network responsible for the exemplar image. α is a trade-off factor calculated by Eq. 2, which gives weight for the features. $s_{interaction}$ is then passed into a set of convolutional layers, and results in the downsampled label.

The loss function used in the optimization framework consists of two parts. Let $\hat{\mathcal{P}}$ be the output of individual image segmentation subnetwork and $\hat{\mathcal{P}}_{small}$ shown in Eq. 1 be the output of the image pair interaction subnetwork.

$$\hat{\mathcal{P}} = \mathcal{F}(\mathcal{I}; \theta) \tag{4}$$

In Eq. 5, the first part of the loss is a softmax loss, L_1, of comparing $\hat{\mathcal{P}}$ and the groundtruth $\mathcal{P}$. The second part is a reconstruction loss L_2 calculated by comparing $\hat{\mathcal{P}}_{small}$ and the groundtruth $\mathcal{P}_{small}$.

$$\mathcal{L}(\mathcal{I}, \mathcal{I}') = \mathcal{L}_1(\hat{\mathcal{P}}, \mathcal{P}) + \lambda \mathcal{L}_2(\hat{\mathcal{P}}_{small}, \mathcal{P}_{small}) \tag{5}$$

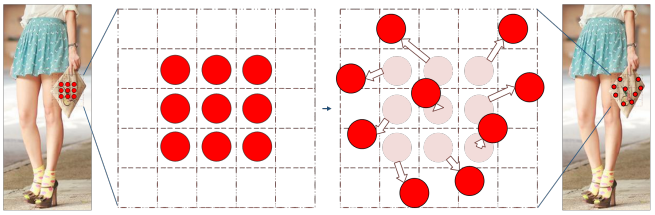
With the first part of the equation depending on θ and the second part on θ', this loss allows the network to learn both the individual segmentation as well as learn from the interaction between the image pair.

After training, the extraction of the segmentation is done by using only $\mathcal{I}$ as the input of the individual image segmentation with its trained parameters. The interaction part and feature extraction part of the exemplar image $\mathcal{I}'$ in the network are removed. All the training and test stage are shown in Algorithm. 1.

2.4 Deformable Convolution

Commonly, clothing items have deformable non-rigid structure due to the artificial flexible materials and different poses of fashion models. In the process of feature extraction in the network, standard convolution extracts square regions in the feature map such as 3*3. As can be seen in Figure 5, the receptive field of tradition convolution is out of the area of the bag. While deformable convolution can match the edge of the skirt effectively, which is significant for the network to learn the edge of the object.

Then, it is more appropriate to use deformable convolution [Dai *et al.*, 2017], which accommodates geometric variations in the images by learning and applying adaptive receptive fields driven by data. In the network, we replace the standard convolution with deformable convolution after layers $pool3$.



(a) (b)

Figure 5: Illustration of the fixed receptive field in standard convolution (a) and the adaptive receptive field in deformable convolution (b), each sample point is moved with an offset according to the shape of the object.

After replacing the 2D convolution filter with deformable convolution, we visualize the response region in the feature map. Figure 5 shows the deformable structure of each cloth items in images. It is very clear to see the structural boundary of each item.

3 Experiments

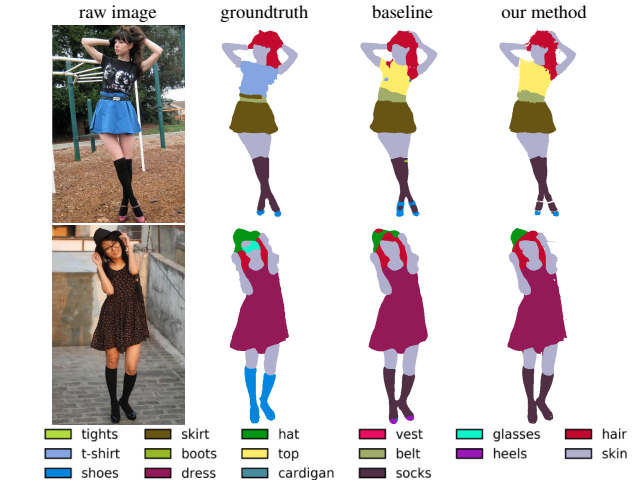
3.1 Datasets

In order to evaluate the performance of the proposed approach, we conduct a set of qualitative and quantitative experiments on two benchmark datasets annotated with pixel-wise groundtruth labeling, including Fashionista [Yamaguchi *et al.*, 2012], refined Fashionista [Tangseng *et al.*, 2017], and CFPD [Liu *et al.*, 2014].

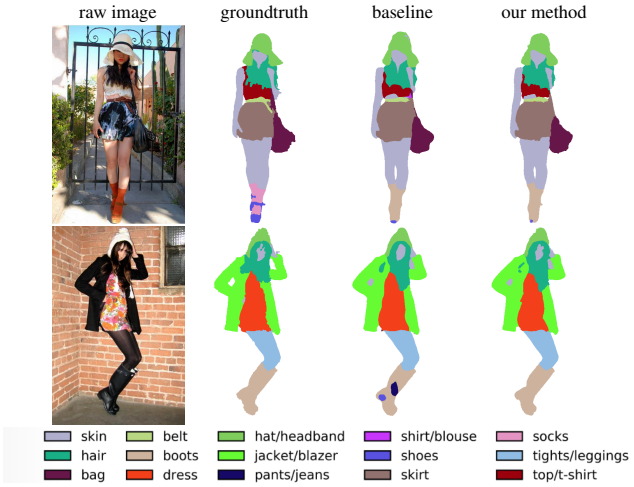
Algorithm 1: Clothing segmentation with semantic locality-aware deformable network

Input: A set of N training samples $\{\mathcal{I}_i\}_{i=1}^N$
Output: The segmentation map $\{\hat{\mathcal{P}}_i\}_{i=1}^N$ and the deep model parameterized by θ: $\mathcal{F}(\mathcal{I}; \theta)$

/* The training stage */
repeat
  /* For the N images, do */
  for $i = 1, \ldots, N$ **do**
    Retrieve the exemplar image $\mathcal{I}'$ in the training dataset with s_{pose} and $s_{appearance}$;
    Calculate the similarity α of image pair $(\mathcal{I}, \mathcal{I}')$ using Eq. 2;
    Take image $\mathcal{I}$ and its context information $\mathcal{I}'$ as input and obtain the segmentation map $\hat{\mathcal{P}}$ using Eq. 4;
    Minimize the objective function Eq. 5;
    /* Update network parameters */
    Update parameters θ by using Adam;
  end
  $iter \leftarrow iter + 1$
until $iter = max\_iter$;
return;
/* The test stage */
Input: Given an image $\mathcal{I}$ and the trained deep model θ: $\mathcal{F}(\mathcal{I}; \theta)$
Output: The predicted segmentation map $\hat{\mathcal{P}} = \mathcal{F}(\mathcal{I}; \theta)$



(a) Fashionista v0.2



(b) Refined Fashionista

(c) CFPD

Figure 6: Successful cases in (a) Fashionista v0.2, (b) Refined Fashionsta, and (c) CFPD. The figure shows the raw image, ground truth, the output of SegNet and our deformable network from left to right respectively. Clearly, with better features trained by our deformable network, our results can achieve the state-of-the-art performance.

Fashionista consists of 685 front-facing full-body images with 56 fine-grained categories. Refined Fashionista is a dataset which reduces the number of categories from 56 to 25 so as to avoid ambiguous labels. CFPD consists of 2682 annotated images based on superpixels for 23 labels. Each image is 400*600 pixels in RGB color.

For the set of training and testing, we randomly divide the Fashionista dataset into train-test splits the same as [Yamaguchi *et al.*, 2012] with 10% of training images leaving for validation, and divide CFPD dataset into 90% train-set and 10% test-set, the same as Fashionista.

3.2 Implementation Details

We implement our architecture by using the Caffe [Jia *et al.*, 2014] toolbox and NVIDIA TITAN X GPU to train the network. The networks are trained in the CFPD and Fashionista training dataset with Adam optimization with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is set to 1e-4, batch size is 1 with the consideration of GPU memory, and the weight decay is 0.0005. The reconstruction loss with the balance parameter $\lambda = 0.125$ since the predicted map $\hat{\mathcal{P}}_{small}$ is 8 times smaller than the raw image. α calculated in Eq. 2 represents the similarity of each image pair and its value varies accordingly with different image pairs. The model trained in the PSPNet is finetuned with a model pretrained in the ADE20K [Zhou *et al.*, 2017] scene parsing dataset, and the SegNet model is trained without any pretrained model. We need about 100K training iterations for convergence.

3.3 Evaluation Methods

In experiments, we utilize two metrics for quantitative performance evaluations, the mean of the pixel-wise accuracy (the ratio of pixels which are correctly predicted) and class-wise intersection-over-union(IoU) which means the intersection of union of pixels averaged over all the semantic categories.

3.4 Ablation Study and Analysis

Table 1 shows the performance of our method compared with various baselines. Results show that we achieve the state-of-the-art performance in accuracy in three datasets. Also, we choose three different baseline networks to verify the effect each module brings in.

Interaction in the Network

The interaction part of the network uses the context information of locality-aware image pairs during the segmentation, which can be normally treated as a kind of regularization. Since clothing segmentation is a fine-grained segmentation task, we constrain the semantic information in a small scale feature map to align the rough structure. Otherwise, the pixel-wise classification will be disturbed in large scale.

SegNet is a simple network with elegant structure and well-adapted to various size of images. After considering the interaction of image pairs, the performance is increased from 90.71% to 91.18% in accuracy. However, the improvement in the PSPNet and DeepLabv2 is not obvious as the SegNet since the performance of the PSPNet and DeepLabv2 baseline is higher, which means the network learns more powerful and discriminated features. Hence, we think the augmentation of

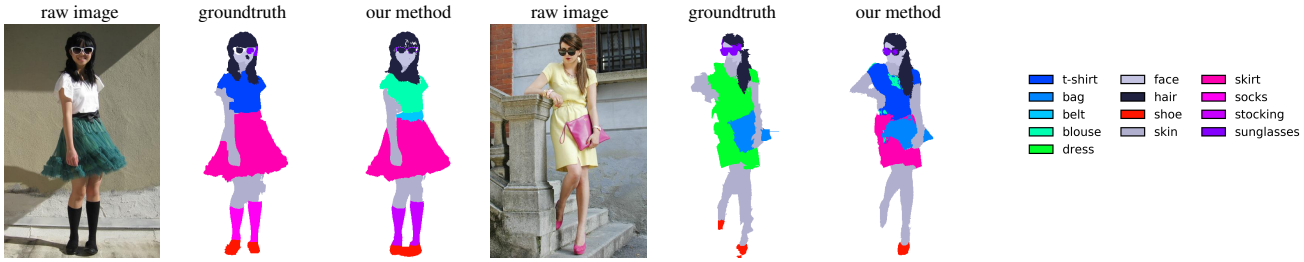


Figure 7: Failure cases in CFPD. The figure shows an input image, a ground truth, and the result of our network from left to right respectively. Failure is caused by either (a) incorrect annotation or (b) the learned model.

Dataset	Method	Acc	IoU
Fashionista v0.2	Paper doll [Yamaguchi *et al.*, 2015]	84.68	-
	Clothelets CRF [Simo-Serra *et al.*, 2014]	84.88	-
	FCN-32s [Long *et al.*, 2015]	85.94	29.61
	FCN-16s [Long *et al.*, 2015]	87.53	34.26
	FCN-8s [Long *et al.*, 2015]	87.51	33.97
	Outfit filter [Tangseng *et al.*, 2017]	87.55	34.26
	Our Network	88.23	36.06
	Our Network +CRF	**89.21**	**37.12**
Refined Fashionista	FCN-32s [Long *et al.*, 2015]	88.56	40.88
	FCN-16s [Long *et al.*, 2015]	89.74	43.96
	FCN-8s [Long *et al.*, 2015]	90.09	44.72
	Outfit filter [Tangseng *et al.*, 2017]	91.50	46.40
	Our Network	92.53	46.68
	Our Network+CRF	**92.93**	**47.85**
CFPD	CFPD [Liu *et al.*, 2014]	-	42.10
	FCN-32s [Long *et al.*, 2015]	90.34	47.65
	FCN-16s [Long *et al.*, 2015]	91.27	50.07
	FCN-8s [Long *et al.*, 2015]	91.58	51.28
	Outfit filter [Tangseng *et al.*, 2017]	91.52	51.42
	Our Network	92.69	52.45
	Our Network+CRF	**93.06**	**53.51**

Table 1: Parsing performance[%] in three datasets.

Network	Acc	IoU
SegNet [Badrinarayanan *et al.*, 2017]	90.71	39.05
+ Interaction	91.18	39.37
+ Deformable conv	91.22	40.22
+ Interaction + Deformable conv	91.43	41.63
+ Interaction + Deformable conv + CRF	**91.86**	**43.61**
DeepLabv2(VGG16) [Chen *et al.*, 2018]	92.07	50.62
+ Interaction	92.44	50.91
+ Deformable conv	92.53	51.14
+ Interaction + Deformable conv	92.58	51.33
+ Interaction + Deformable conv + CRF	**92.85**	**51.89**
PSPNet [Zhao *et al.*, 2017]	92.18	51.37
+ Interaction	92.50	51.54
+ Deformable conv	92.77	52.06
+ Interaction + Deformable conv	92.69	52.45
+ Interaction + Deformable conv + CRF	**93.06**	**53.51**

Table 2: Ablation study in CFPD dataset.

data which context information of image pairs brings does enrich the information of pixel-wise classification. And the improvement is embodied in the learned feature.

Deformable Convolution

Results show that deformable convolution makes an improvement in capturing the edge of clothes items among all datasets. With deformable convolution, the performance using SegNet as baseline improves by 0.51% in accuracy and 1.17% in IoU compared with standard one.

Conditional Random Fields

In general, pixel-wise semantic segmentation ignores the consistency of pixels in the region, which is a significant characteristic in images. We use the softmax output of our deformable network and consider the information by using Gaussian kernels. By using the CRF [Krähenbühl and Koltun, 2011] to process the result, the performance improves the IoU from 41.63% to 43.61% in SegNet, from 51.33% to 51.89% in DeepLabv2(VGG16), and from 52.45% to 53.51% in PSPNet. Truly, the CRF, as a post-processing method, has considerable improvement in segmentation task.

3.5 Qualitative Evaluation

We compare our method with FCN-8s, FCN-16s, FCN-32s [Long *et al.*, 2015], as well as with the reported state-of-art methods for each dataset. Figure 6 shows the visual results of our experiments. Also, there are some failure cases in the results. Some are caused by incorrect annotation in the dataset while some are caused by the learned model. Since the dataset is annotated with extracting superpixel region, the whole semantic region is broken up and the shape is not clear. For example, the sunglasses are not annotated correctly in Figure 7. Also, the learned model can not keep the semantic consistency of dress where the annotated dress is split into shirt and shirt due to the occlusion of the bag.

4 Conclusion

In this paper, we have proposed a deformable network to improve the performance of clothing segmentation in the condition of the small fine-grained annotated dataset. To deal with the problems of deformable structure of semantic object, diversity of cloth items in semantic and small size of the fine-grained dataset, the proposed network uses the method of retrieving semantic similar exemplar image in datasets, mining the interaction parts in locality-aware image pairs correlatively, and using deformable convolution to extract the non-rigid geometry-aware features of clothing images. Experiments show that our method achieves the state-of-the-art performance in the CFPD and Fashionista dataset .

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants (U1509206, 61472353, and 61751209), in part by the National Basic Research Program of China under Grant Grant 2015CB352302.

References

[Badrinarayanan *et al.*, 2017] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 39(12):2481–2495, 2017.

[Cao *et al.*, 2017] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, volume 1, page 7, 2017.

[Chen *et al.*, 2018] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2018.

[Dai *et al.*, 2017] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. *CoRR, abs/1703.06211*, 1(2):3, 2017.

[Gong *et al.*, 2017] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, pages 932–940, 2017.

[Hadi Kiapour *et al.*, 2015] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, pages 3343–3351, 2015.

[Jia *et al.*, 2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, pages 675–678. ACM, 2014.

[Krähenbühl and Koltun, 2011] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, pages 109–117, 2011.

[Liang *et al.*, 2016a] Xiaodan Liang, Liang Lin, Wei Yang, Ping Luo, Junshi Huang, and Shuicheng Yan. Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval. *IEEE TMM*, 18(6):1175–1186, 2016.

[Liang *et al.*, 2016b] Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with graph lstm. In *ECCV*, pages 125–143. Springer, 2016.

[Liang *et al.*, 2017] Xiaodan Liang, Liang Lin, Xiaohui Shen, Jiashi Feng, Shuicheng Yan, and Eric P Xing. Interpretable structure-evolving lstm. In *CVPR*, pages 2175–2184, 2017.

[Liew *et al.*, 2017] JunHao Liew, Yunchao Wei, Wei Xiong, Sim-Heng Ong, and Jiashi Feng. Regional interactive image segmentation networks. In *ICCV*, pages 2746–2754. IEEE, 2017.

[Liu *et al.*, 2014] Si Liu, Jiashi Feng, Csaba Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan. Fashion parsing with weak color-category labels. *IEEE TMM*, 16(1):253–265, 2014.

[Liu *et al.*, 2016] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, pages 1096–1104, 2016.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[McAuley *et al.*, 2015] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *SIGIR*, pages 43–52. ACM, 2015.

[Simo-Serra *et al.*, 2014] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. A high performance crf model for clothes parsing. In *ACCV*, pages 64–81. Springer, 2014.

[Tangseng *et al.*, 2017] Pongsate Tangseng, Zhipeng Wu, and Kota Yamaguchi. Looking at outfit to parse clothing. *CoRR, abs/1703.01386*, 2017.

[Veit *et al.*, 2015] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *ICCV*, pages 4642–4650, 2015.

[Vittayakorn *et al.*, 2015] Sirion Vittayakorn, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg. Runway to realway: Visual analysis of fashion. In *WACV*, pages 951–958. IEEE, 2015.

[Yamaguchi *et al.*, 2012] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *CVPR*, pages 3570–3577. IEEE, 2012.

[Yamaguchi *et al.*, 2015] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Retrieving similar styles to parse clothing. *IEEE TPAMI*, 37(5):1028–1040, 2015.

[Yang *et al.*, 2017] Rui Yang, Bingbing Ni, Chao Ma, Yi Xu, and Xiaokang Yang. Video segmentation via multiple granularity analysis. In *CVPR*, pages 3010–3019, 2017.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017.

[Zhou *et al.*, 2017] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.