EXPLOITING TEXTURE CUES FOR CLOTHING PARSING IN FASHION IMAGES

Tarasha Khurana[⋆,1] *Kushagra Mahajan*[⋆,1] *Chetan Arora*[1] *Atul Rai*[2]

[1]IIIT Delhi [2]Staqu Technologies

ABSTRACT

We focus on the problem of parsing fashion images for detecting various types of clothing and style. The current state-of-the-art techniques for the problem are mostly based on variations of the SegNet model [1]. The techniques formulate the problem as segmentation and typically rely on geometrical shapes and position to segment the image. However, specifically for fashion images, each clothing item is made of specific type of materials with characteristic visual texture patterns. Exploiting the texture for recognizing the clothing type is an important cue which has been ignored so far by the state-of-the-art. In this paper, we propose a two-stream deep neural network architecture for fashion image parsing. While the first stream uses the regular fully convolutional network segmentation architecture to give accurate spatial segments, the second stream provides texture features based upon Gabor filters and helps in determining the clothing type resulting in improved recognition of the various segments. Our experiments show that, the proposed two-stream architecture successfully reduces the confusion between the clothing types, having similar visual shapes in the images but different material. Our approach achieves state-of-the-art results on the standard benchmark datasets, such as Fashionista [2] and CFPD [3].

Index Terms— Fashion parsing, Fully convolutional network, Texture features, SegNet, Gabor

1. INTRODUCTION

In the recent years, deep convolutional neural networks have been successfully applied for semantic segmentation, overcoming challenges such as large visual variations in the objects, reduced feature resolution and segmenting objects at multiple scales. However, often such models, implicitly or explicitly, exploit domain specific features, and are restricted to the focused domain only. Consider the case of highly successful scene parsing models [4, 5] for segmenting higher-level image regions such as roads, buildings, sky etc. The same models fail miserably on human parsing for segmenting human body parts such as arms, legs, or torso, where the successful models [6, 7] have exploited joint labels, pose estimation and customized losses that are sensitive to the human body configuration. On the other hand, these human parsing models fail to extend to clothing parsing where the target is to segment various clothing items worn by a person. While the difference in labels, such as torso in human parsing vs shirt and scarf in clothing parsing is one problem, the models for human parsing are trained inherently towards ignoring the texture. For example, a torso is classified as a torso irrespective of the texture (shirt/top/jacket). Therefore, such models, when applied for clothing parsing, fail to distinguish between clothing items which have similar shapes and position but different material, for example, denims vs trousers, or sweater vs top. An example is shown in Fig. 1.

Fig. 1: We show how the additional supervision from texture descriptors improves garment labelling. First segmentation map denotes the groundtruth annotation, followed by the output of Outfit Encoder [8] and our proposed model. While [8] mispredicts a portion of 'top' as a 'sweater' in the first image and 'stockings' as 'skin' in the second, using characteristic textures of these clothing items helps our model disambiguate between them.

We observe that the texture cues complement the shape and position information exploited by the contemporary segmentation pipelines, by providing fine-grained features associated with the material of a particular clothing object. For instance, in the case of sweaters vs tops, it may be hard to distinguish between the two classes on the basis of spatial or shape cues alone. However, typically the two clothing items not only have different materials, but also contain very differ-

⋆ The first two authors have contributed equally

ent visual patterns on them. Thus, in this work, we propose to augment the standard segmentation pipelines with a second stream based on the texture based features. We have experimented with various texture features such as Gabor [9] and LBP [10] and finally chosen Gabor features for its improved experimental performance. The proposed architecture gives state-of-the-art performance on Fashionista [2] and CFPD [3], outperforming techniques like [2, 8, 11–13].

2. RELATED WORK

Clothing Parsing In the recent years, clothing parsing has seen active research in computer vision [11, 14–18] as a variant of human parsing [6, 7, 13, 19]. Much of the existing works formulate the problem based on pose estimation or non-parametric label transfer [2, 11, 12, 14, 20]. Some recent works focus on joint segmentation and labeling [11] as well as combinatorial preference of clothing items to assist in the prediction [8]. These approaches fail to resolve conflicts between objects which are found at the same semantic locations in the body. Yang *et al.* [11] propose a two-phase inference approach in which the first phase uses *exemplar-SVM* for extraction and refinement of image segments, while the second phase uses multi-image graphical models to classify the segments. Yamaguchi *et al.* [2] show an 89.0% accuracy on the contributed Fashionista dataset using image meta-tags and pose-estimation. However, in a scenario like ours, this external metadata is not available. Tangseng *et al.* [8] claim that higher level judgement regarding clothing combinations is an important prior for boosting the parsing performance.

Texture Characterization Texture analysis is useful in places where shape related features render insufficient for characterization. Texture cues have been explored prominently in texture segmentation [21] and texture recognition [21, 22] tasks. Traditional pooling encoders have been used by some recent works [22–25] who build hybrid representations of deep CNNs. Zhang *et al.* [26] integrate texture in the deep learning models by means of an Encoding layer that learns visual vocabularies directly from the loss function.

3. PROPOSED APPROACH

In this paper, we propose an end-to-end texture assisted segmentation pipeline whose architecture has been shown in Fig. 2 and details are described below[1].

3.1. Segmentation Stream

We have used Fully Convolutional Networks (FCNs) [5] as the base for one of the streams in the proposed model. We use the 8s variant which is trained progressively from its 32s and 16s variants. As FCN-8s improves segmentation detail by incorporating information from layers with different strides, it gives superior results on the two datasets that have been used in this paper. The progressive finetuning of FCN-8s has been done on the datasets under consideration and in the proposed model, we call this branch the 'Segmentation Stream'. We have also experimented with DeepLabV2 [4] for the segmentation stream but found it to be less accurate than FCN-8s.

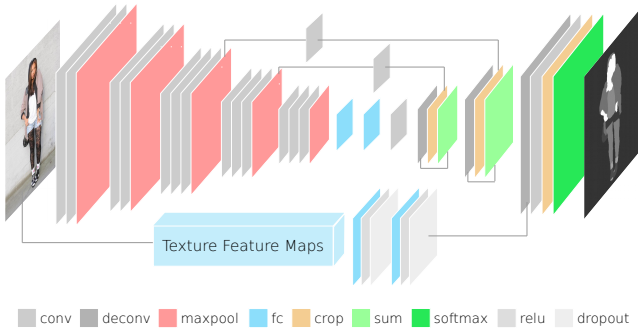

Fig. 2: Proposed two-stream architecture for clothing parsing. We use a separate stream to exploit texture cues for improved clothing type recognition.

3.2. Texture Stream

We have experimented with two commonly used texture descriptors: the Gabor feature descriptor [9] and Local Binary Patterns (LBP) [10]. We note that other texture descriptors could have been used here as well.

Gabor Features Gabor feature [9] responses are extracted corresponding to different wavelengths, orientations and phases. The combined set of these feature maps are either early fused or late fused (refer Section 3.3) into the main network. The following parameters are adopted for feature maps extraction - wavelength: [3-8] pixels, orientation: [0°, 45°, 90°, 135°] and 5 phase values spaced uniformly from 0 to the wavelength(λ). Only 4 orientation values are chosen since most of the textures are essentially aligned along these angles. We use an 11×11 sliding window for gabor feature map extraction at each pixel. We also experimented with windows of size 9, 13, 15 but found 11 to be working best.

Close attention has been paid to represent the wavelengths accurately such that all the essential frequencies in the input images are captured. We choose only small values of λ i.e. from 3 to 8 pixels, since texture on the clothing items is inherently a fine-grained semantic attribute of the cloth and adding responses for larger wavelengths to the set of feature maps does not add value. We further validated the choices for the texture descriptor parameters by conducting experiments on a surrogate task to classify cropped clothing images based upon their texture descriptors alone.

[1]The complete source code and the pretrained models are available at `http://to.be.released.later`

	Fashionista		CFPD	
	Gabor	**LBP**	**Gabor**	**LBP**
Early Fusion				
scorefr	87.7	88.1	91.7	91.8
upscore4	88.9	88.3	92.3	92.5
upscore8	89.4	88.7	92.8	91.9
Late Fusion				
upscore8 + 1 conv	**91.1**	89.8	**93.5**	92.9
upscore8 + 2 conv	90.4	90.2	93.0	92.3

Table 1: Results obtained from various configurations of the proposed model on benchmark datasets.

LBP Features We extract LBP [10] features over a sliding window of size 11×11 as done for Gabor. The number of neighbours is set to 8. Multiresolution analysis is accomplished by varying the neighbourhood radius and the number of points to be considered in the circularly symmetric neighborhood. We experiment with radii starting from 2 pixels and extending to 9 pixels and observe a nearly 2% increase in accuracy for the set of multiresolution feature maps comprising of 2, 3, 4 radius values. However, overall, the two-stream LBP model performs poorly as compared to the two-stream Gabor model by about 0.9% and 0.6% on the Fashionista and CFPD datasets respectively. Therefore, all the results reported in the experiments section have used Gabor filters.

3.3. Stream Fusion

We have explored early and late fusion strategies for merging the texture and segmentation streams. In the early fusion strategy, we merged the texture feature maps with feature maps from the segmentation stream (we experimented with merging at various layers). The merging was followed by a 5x5 convolutional layer to learn local context in the fused information. In late fusion, we let the two streams generate score maps for each of the clothing labels independently. We then concatenate the two score maps and apply a 1x1 convolutional layer to obtain the final category maps for each label. The whole network is trained in an end-to-end fashion.

4. DATASETS

Fashionista Dataset [2] The Fashionista dataset was introduced for evaluating clothing estimation techniques. It comprises of 685 full body images extracted from *chictopia.com* in frontal/near-frontal view, with clean background and complete visibility of all clothing items. Pixel annotations for 56 clothing categories, including a background class are provided. Due to the larger set of labels, it contains instances of similarly shaped classes of upper body clothes as well as lower body clothes. We have used 229 images for testing and the remaining for training.

	Fashionista		CFPD	
	Accuracy	**IoU**	**Accuracy**	**IoU**
OE [8]	88.6	38.0	92.3	54.7
PaperDoll [12]	84.7	-	87.1	-
SSL [6]	84.8	33.2	88.5	49.1
CCP [11]	90.2	-	-	-
DLV2 (ResNet) [4]	86.6	36.8	89.9	48.3
DLV2 (VGG) [4]	86.2	35.4	89.2	47.2
FCN-8s [5]	87.5	33.8	91.6	51.2
Ours	**91.1**	42.1	**93.5**	58.7

Table 2: Comparison with the state-of-the-art in terms of overall accuracy and overall IoU. Here 'IoU' stands for Intersection over Union as a metric of evaluation, 'OE' stands for Outfit Encoder and 'DLV2' stands for DeepLabV2.

CFPD Dataset [3] The Colorful-Fashion dataset is about 3 times larger than the Fashionista dataset, consisting of 2,682 images scraped from *chictopia.com*. This dataset has 23 clothing category labels including a background class. Here, 894 images are used for testing and the rest for training.

5. EXPERIMENTS AND RESULTS

All the experiments have been conducted on a workstation with 1.728 GhZ CPU, 128GB RAM, NVIDIA Quadro P5000 GPU and running Ubuntu 14.04. We augment the training data for both the datasets using flips and crops, making sure that no portion of the object of interest gets cropped out.

5.1. Effect of Hyperparameters

Compared to the segmentation stream (FCN-8s), we observe a gain of 1.9% and 3.6% on the Fashionista dataset using the texture stream in early and late fusion respectively. For CFPD, the corresponding numbers are 1.2% and 1.9%. Augmenting with the texture stream helps in all the cases though, late fusion seems to perform better.

The above mentioned results are achieved when texture information is fused with the segmentation stream at the final pixel classification layer 'upscore8' of the FCN-8s model, and convolutional layers are added post concatenation to obtain a cumulative set of feature maps taking both the complementary features maps into consideration. Experiments are also conducted by fusing this information at other locations of the encoder and decoder (scorefr and upscore4 layers). However, concatenating at 'upscore8' performs the best.

For early fusion strategy, we add a convolutional layer, post fusion, to gather the local context of a pixel before making the final prediction about the class. We observe that the 5x5 kernel outperforms the smaller kernels of size 3 and larger ones such as those of size 9. In the texture stream, we have

Fig. 3: Left column shows results of the proposed model on Fashionista [2] dataset and right column shows results on CFPD [3] dataset. For each image, first segmentation map denotes the groundtruth annotation followed by results of FCN-8s [5], Outfit Encoder [8] and the proposed model respectively. Notice how the characteristic textures of 'top', 'skirt', 'pants', 'jeans' etc. in the images shown help in successfully refining the segments and their corresponding class label predictions.

experimented with adding few convolutional layers and obtain a full score map before late fusing with the segmentation stream for the final prediction. Best results are obtained with convolutional kernels of size 7. After concatenation in the late fusion style, we have experimented by adding a 1x1 kernel convolutional layer for the final pixel-wise prediction and also by applying 2 convolutional layers of kernel sizes 5x5 followed by 1x1. The results from various such configurations are illustrated in Table 1.

5.2. Comparison with State-of-the-Art

We compare our results with the state-of-the-art on the Fashionista and CFPD datasets given by [11] and [8] respectively. Our two-stream architecture, combining the segmentation and texture streams in a late fusion style outperforms [11] for Fashionista dataset by 0.9%. Exploiting texture features helps in disambiguating similarly shaped clothing items and improve results reported by Tangseng *et al.* [8] by 2.5% and 1.2% on the Fashionista and CFPD datasets respectively. We also compare our approach with the FCN-8s architecture. These results are compared in Table 2. Some examples of the improvement in segmentation are given in Fig. 3. To show the advantage of our model over the state-of-the-art semantic segmentation architectures, we also compare with the DeepLabV2 models. The proposed model yields an accuracy improvement of 4.5% and 3.6% over DeepLabV2 (ResNet) on the Fashionista and CFPD datasets respectively. To get a better understanding of the strengths and weaknesses of our approach, we give the confusion matrices and failure cases of our approach in the supplementary material.

6. CONCLUSION

Texture is an important characteristic for understanding the different clothing types in human perception. However, its use in clothing parsing has been largely ignored where the state-of-the-art have used standard segmentation pipelines which have been designed and trained to ignore the texture. In this paper, we have proposed a two-stream architecture, using a standard segmentation pipeline in one stream, but exploiting Gabor based texture features in the second. We show in our experiments that the proposed model helps in disambiguating similarly shaped but different textured clothing items, and achieves state-of-the-art performance on the various benchmark datasets. In future, we would like to extend our model by also including human pose information which can help disambiguate in case of self occlusion.

7. REFERENCES

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *TPAMI*, vol. 39, no. 12, pp. 2481–2495, 2017.

[2] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg, "Parsing clothing in fashion photographs," in *CVPR*, 2012, pp. 3570–3577.

[3] Si Liu, Jiashi Feng, Csaba Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan, "Fashion parsing with weak color-category labels," *IEEE Trans. on Multimedia*, vol. 16, no. 1, pp. 253–265, 2014.

[4] 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," *TPAMI*, 2017.

[5] Evan Shelhamer, Jonathan Long, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," *TPAMI*, vol. 39, no. 4, pp. 640–651, 2017.

[6] 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*, 2017, pp. 932–940.

[7] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan, "Human parsing with contextualized convolutional neural network," in *ICCV*, 2015, pp. 1386–1394.

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

[9] Dennis Gabor, "Theory of communication. part 1: The analysis of information," *JIEE-Part III: Radio and Comm. Engg.*, vol. 93, no. 26, pp. 429–441, 1946.

[10] Timo Ojala, Matti Pietikainen, and Topi Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *TPAMI*, vol. 24, no. 7, pp. 971–987, 2002.

[11] Wei Yang, Ping Luo, and Liang Lin, "Clothing co-parsing by joint image segmentation and labeling," in *CVPR*, 2014, pp. 3182–3189.

[12] Kota Yamaguchi, M Hadi Kiapour, and Tamara L Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *ICCV*, 2013, pp. 3519–3526.

[13] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan, "Deep human parsing with active template regression," *TPAMI*, vol. 37, no. 12, pp. 2402–2414, 2015.

[14] Yannis Kalantidis, Lyndon Kennedy, and Li-Jia Li, "Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos," in *ICMR*. ACM, 2013, pp. 105–112.

[15] Andrew C Gallagher and Tsuhan Chen, "Clothing cosegmentation for recognizing people," in *CVPR*, 2008, pp. 1–8.

[16] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun, "A high performance crf model for clothes parsing," in *ACCV*. Springer, 2014.

[17] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg, "Retrieving similar styles to parse clothing," *TPAMI*, vol. 37, no. 5, pp. 1028–1040, 2015.

[18] Nan Wang and Haizhou Ai, "Who blocks who: Simultaneous clothing segmentation for grouping images," in *ICCV*, 2011, pp. 1535–1542.

[19] Si Liu, Xiaodan Liang, Luoqi Liu, Xiaohui Shen, Jianchao Yang, Changsheng Xu, Liang Lin, Xiaochun Cao, and Shuicheng Yan, "Matching-cnn meets knn: Quasi-parametric human parsing," in *CVPR*, 2015.

[20] Nataraj Jammalamadaka, Ayush Minocha, Digvijay Singh, and CV Jawahar, "Parsing clothes in unrestricted images.," in *BMVC*, 2013.

[21] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi, "Deep filter banks for texture recognition and segmentation," in *CVPR*, 2015, pp. 3828–3836.

[22] Shin Fujieda, Kohei Takayama, and Toshiya Hachisuka, "Wavelet convolutional neural networks for texture classification," *arXiv preprint arXiv:1707.07394*, 2017.

[23] Vincent Andrearczyk and Paul F Whelan, "Using filter banks in convolutional neural networks for texture classification," *Pattern Recognition Letters*, vol. 84, pp. 63–69, 2016.

[24] Peng Tang, Xinggang Wang, Baoguang Shi, Xiang Bai, Wenyu Liu, and Zhuowen Tu, "Deep fishernet for object classification," *arXiv preprint arXiv:1608.00182*, 2016.

[25] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi, "Describing textures in the wild," in *CVPR*, 2014, pp. 3606–3613.

[26] Hang Zhang, Jia Xue, and Kristin Dana, "Deep ten: Texture encoding network," in *CVPR*, 2017, pp. 708–717.

Fig. 4: Examples of failure cases of the proposed model.

A. CONFUSION MATRIX

We show a comparison between the confusion matrices of Outfit Encoder [8] and our proposed model in Fig. 5 to highlight the improvement in the segment labelling of similarly shaped upper body and lower body clothes.

Fashionista Dataset [2] First row in Fig. 5 shows the confusion matrices for Outfit Encoder [8] and the proposed model respectively for Fashionista dataset. They clearly show that the texture based solution reduces errors among the commonly confused classes like jeans-pants-leggings and sweater-blouse-top.

CFPD Dataset [3] The confusion matrices in second row of Fig. 5 for Outfit Encoder and our model respectively demonstrate a reduction in errors between jeans-leggings-pants, shorts-skirt, socks-stockings and t-shirt-blouse-sweater when texture is incorporated in segmentation.

B. FAILURE CASES

Fig. 4 shows the failure cases for our proposed approach for the Fashionista dataset in the first column and the CFPD dataset in the second. For each image, the first segmentation map illustrates the groundtruth, the second map shows the

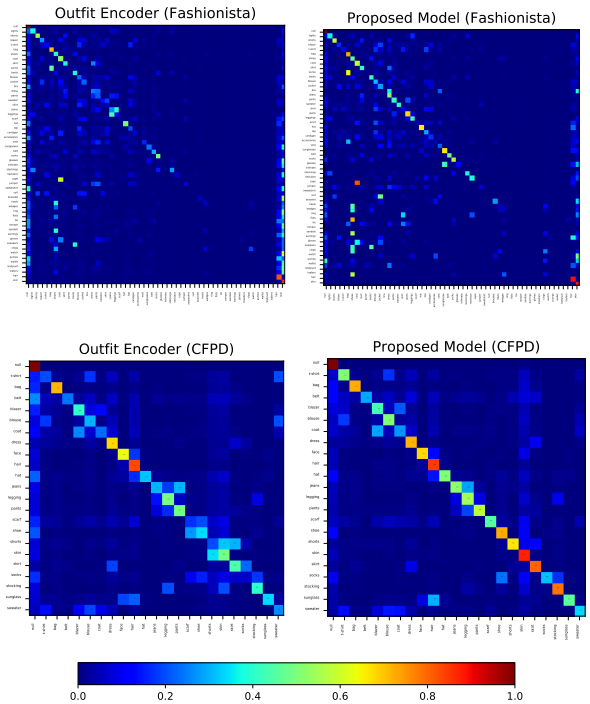

Fig. 5: Normalized confusion matrices for Outfit Encoder [8] and the proposed model on the Fashionista (1st row) and CFPD datasets (2nd row). Please zoom in the pdf to read the fine text in the matrices.

output from our model. Image (a) shows that our model perceives the skin texture for nearly transparent stockings leading to erroneous segment labelling. In (b), the texture of the lower body cloth changes within the segment from creased to plain, leading to the intuitive labelling of a skirt and pant respectively. For (c), a coat is mistaken to be a jacket and the top is largely predicted as a t-shirt. In (d), t-shirt is wrongly predicted as a sweater. Both the textures and the semantic locations of the mispredicted and the true classes are highly indistinguishable in the specific instances reported, leading to the failed predictions.