Semantic Segmentation of Fashion Images Using Feature Pyramid Networks

John Martinsson
RISE Research institutes of Sweden
john.martinsson@ri.se

Olof Mogren
RISE Research institutes of Sweden
olof.mogren@ri.se

Abstract

In this work, we approach the problem of semantically segmenting fashion images into different categories of clothing. This problem poses particular challenges because of the importance of both textural information and cues from shapes and context. To this end, we propose a fully convolutional neural network based on feature pyramid networks (FPN), together with a backbone consisting of the ResNeXt architecture. Our experimental evaluation shows that the proposed model achieves state-of-the-art results on two standard fashion benchmark datasets, and a qualitative study verifies its effectiveness when applied to typical fashion images. The approach has a modest memory footprint and can be used without a conditional random field (CRF) without much degradation of quality which makes our model preferable from a computational perspective. When comparing all methods without a CRF, our approach outperforms all state-of-the-art models on both datasets by a clear margin in all evaluated metrics. In fact, our approach achieves a higher accuracy ***without*** *the CRF than the state-of-the-art models* ***using*** *CRFs.*

1. Introduction

Analysing trends is an important strategic activity in the fashion industry, which is increasingly becoming a task of identifying trend setting individuals, following them on blogs and social media platforms. Fashion is to a large extent communicated with images, and going through large quantities of photos is one of the key tasks. Visual data is useful for successful fashion forecasting. Al-Halah, et.al. [1] used image features produced by a convolutional neural network (CNN) model trained for image classification to perform fashion style forecasts. We assume that the richer information given by semantic segmentation is beneficial of such downstream tasks. Having the right tools at hand to aid this work can allow analysts to work more effectively, and extracting semantically rich representations from images can be such a tool, providing detailed information to sort through the massive stream of data.

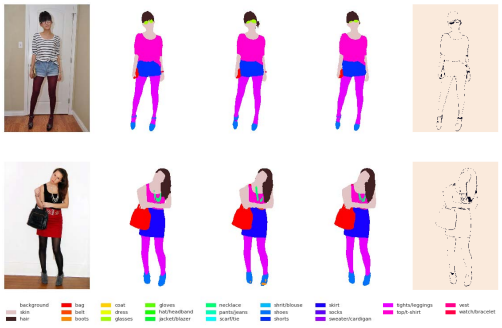

Figure 1: From left to right: the input image, the ground truth segmentation, the predicted segmentation (ResNeXt-FPN), the prediction with CRF, and the incorrectly classified pixels (shown in black) for two top scoring test data image predictions from refined Fashionista.

We consider an important part of such a toolchain: semantic segmentation of fashion images. That is, the division of the image into different regions of clothing.

Figure 1 shows an example image, with the ground truth segmentation, and predicted segmentation maps produced by our approach. Each pixel in the image is classified, and the output is a semantic analysis showing which clothing items are present, where in the image these are, and what shape they have. Shapes of clothing items are an important feature for fashion analysis. For example, a hat could have many different shapes, and the shape of a hat may go in and out of fashion over time.

We use a feature pyramid network (FPN) [8] with a ResNeXt [11] backbone for the semantic segmentation of fashion images. The feature pyramid structure makes the model more robust against images of different scales, and allows both high and low level features to be used in the prediction of the semantic segmentation map. The filters learned by the early layers in a CNN usually resemble Gabor filters or color blobs [13] and such low level features have been shown to improve the accuracy of cloth-

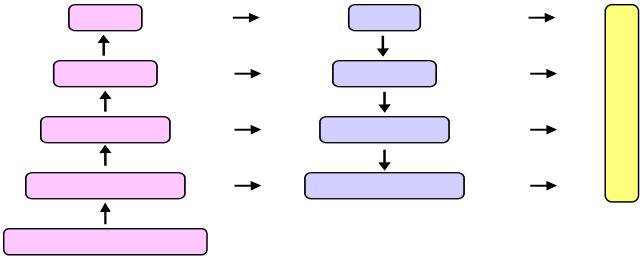

Figure 2: The feature pyramid network architecture used in this work. Images are fed into the bottom layer of the left column. Predictions are produced in the yellow module.

ing parsing methods when fused in the late stages of a fully CNN [5]. We assume that the extraction of Gabor-like features can be learned by the early layers of the ResNeXt backbone, and the FPN enables these to be used directly in the prediction of the segmentation map.

We perform a thorough experimental analysis and show that our approach obtains strong results. In fact, our model achieves higher prediction accuray, without using conditional random fields (CRFs), than the baselines with the CRF enabled. When our model is used together with the CRF, we obtain even stronger results.

2. ResNeXt-FPN

We introduce a model consisting of a ResNeXt [11] backbone, which uses an aggregated residual transformation [11], arranged as a feature pyramid network (FPN) [8]. The model can be used with or without a CRF, and both configurations are explored in the experimental evaluation.

Feature pyramid network. A typical CNN use pooling layers or convolutional layers with a stride of two to reduce the spatial dimension of the input to extract higher-level features. All layers with output features that have the same spatial dimensions are said to belong to the same stage. The last features from each stage are combined into a feature pyramid [8] (see figure 2). This means that we extract a set of features $\{C_2, C_3, C_4, C_5\}$, corresponding to features from the stages with a spatial dimension reduced by 4, 8, 16, and 32 times in the ResNeXt backbone, respectively (C_1 is omitted for computational efficiency).

A set of feature maps F_i are then extracted by a point-wise convolution with each C_i to reduce the channel dimension. The high-level features are upsampled and combined with lower-level features via element-wise addition by

$$P_i = \text{Upsample}(P_{i+1}, 2) + F_i, \quad (1)$$

where $P_5 = F_5$ and $\text{Upsample}(\cdot, k)$ is a nearest neighbor upsampling by a factor of k. We end up with a pyramid of features $\{P_2, P_3, P_4, P_5\}$.

We upsample each P_i to match the spatial dimension of P_2. These upsampled layers and P_2 are concatenated and go through two convolutional layers, a softmax layer and a final upsampling layer to get the final prediction.

Conditional random field. Post-processing the predictions using a CRF has been shown to increase the accuracy in clothing parsing tasks [10, 4]. The densely connected conditional random field [7] is a probabilistic graphical model which is used to improve the same class consistency of predicted labels for similar pixels and for pixels that are close. This is encoded as pair-wise potentials using an *appearence* kernel and a *smoothing* kernel.

3. Previous work

Clothing parsing, semantic segmentation for clothing items, has gained interest recently [12, 9, 10] due to potential applications in the fashion industry such as fashion trend prediction and image-based information retrieval. Pioneering work used meta tags, in addition to the raw fashion images, to parse the content in an image [12, 9]. In this work we consider the scenario where only image data is available.

Recent work has used fully-convolutional neural networks (FCNs) with additional feature branches fused at a late stage in the network to improve performance [10, 5]. Tangseng et.al. [10] use features from an image-level classification network and Khurana et.a. [5] use features extracted by Gabor filters. Ji et.al. [4] propose training a PSPNet [14] using semantically similar image pairs and an auxiliary reconstruction loss to make the training more stable.

4. Experiments

In this section we present the datasets that we evaluate our model on and the implementation and training details for the model.

4.1. Datasets

We evaluate our method on the refined Fashionista [10] and color-fashion (CFPD) [9] datasets. There are relatively few training examples in these datasets, making it a challenge to train a model without overfitting. The model needs to be data efficient and learn from relatively few examples. The images in the datasets are mainly of single person female models in varying poses taken with full body frontal view.

There are 25 different classes in refined Fashionista. The dataset consists of 685 fashion images split into a training set of 456 images and a test set of 229 images. We use 45 of the training images as a validation set. Predicting only the background class gives 77.58% accuracy on the test data.

The color-fashion dataset (CFPD) [9] consists of 2,682 fashion images with pixel level annotations for 23 different

	Refined Fashionista		CFPD	
	Accuracy	IoU	Accuracy	IoU
OE [10]	91.50	46.40	91.52	51.42
PSPNet [4]	92.53	46.68	-	-
FPN	**93.26**	**49.81**	**93.52**	**53.00**

Table 1: The average per pixel accuracy and the mean intersection-over-union for two state-of-the-art models [4, 10] and FPN on the refined fashionista dataset and the CFPD dataset *without* CRF.

	Refined Fashionista		CFPD	
+ CRF	Accuracy	IoU	Accuracy	IoU
OE [10]	91.74	**51.78**	92.35	**54.65**
PSPNet [4]	92.93	47.85	-	-
FPN	**93.62**	50.64	**93.82**	54.39

Table 2: The average per pixel accuracy and the mean intersection-over-union for two state-of-the-art models [4, 10] and FPN on the refined fashionista dataset and the CFPD dataset *with* CRF.

classes. We use the same data split as in [10] where 2,146 images are used as a training set and 536 images are used as a test set. We use 108 of the training images as a validation set. Predicting only the background class gives 79.41% accuracy on the test data.

Pre-processing. The input data is augmented by random horizontal flips and since the weights of the backbone have been pre-trained on ImageNet we have subtracted the channel-wise mean from the RGB channels of the input data and divide by the channel-wise variance. The mean and variance have been pre-computed on ImageNet.

4.2. Implementation details

We implement the model in the Keras [2] framework using the ResNeXt backbone with weights pretrained on ImageNet. The model is trained on a NVIDIA TESLA V100 GPU using the Adam [6] optimizer with a learning rate of 1e-4, β_1 of 0.9, β_2 of 0.999, no decay and a batch size of five. We choose the model with the highest validation accuracy after 100 training epochs as the final model.

The hyper parameters of the CRF [7] are chosen with respect to validation accuracy using random search. We uniformly sample 50 hyper parameter configurations for the appearance kernel with $\theta_\beta \in \{1, 2, \ldots, 100\}$, $\theta_\alpha \in \{1, 2, \ldots, 40\}$ and $w_1 \in \{1, 2, \ldots, 15\}$ using the same notation as in [7]. The number of iterations for the mean field approximation is set to 10.

5. Results

Quantitative results. When comparing all methods *without* a CRF we increase the state-of-the-art accuracy by 0.73 percentage points on refined fashionista and 2.00 percentage points on CFPD, and the state-of-the-art mean intersection-over-union by 3.13 on refined fashionista and 1.58 on CFPD (see table 1). When comparing all methods *with* a CRF we increase the state-of-the-art accuracy by 0.69 percentage points for refined Fashionista and 1.47 percentage points for CFPD while the mean intersection-over-union is marginally worse on both datasets (see table 2).

Notably, the FPN model has a consistently higher accuracy *without* a CRF than other methods *with* a CRF (see table 1 and table 2). Adding a CRF to the FPN does improve results, but only marginally. Using a CRF is not always desirable since it adds a significant computation overhead. In our setting we are able to make predictions for 1.95 images per second without a CRF but only 0.20 images per second with a CRF. An order of magnitude in difference.

We also tried using densely connected CNNs [3], with up to four times fewer parameters, as backbones in the FPN, and observed only a marginal decrease in accuracy. Making the FPN a suitable choice for applications with a low computational budget.

Note that both OE [10] and PSPNet [4] have been fine-tuned on other large-scale semantic segmentation tasks containing tens of thousands of annotated image examples prior to training on refined Fashionista and CFPD, while our model uses weights pre-trained only on ImageNet.

We should note that [4] report a 93.06% accuracy and a mean intersection-over-union of 53.51 on another test split for the CFPD data. Similarity, [5] report a 93.5% accuracy and a mean intersection-over-union of 58.7 for another test split of the CFPD data. We have chosen not to make a direct comparison to these methods in table 1 and table 2 since the test splits are not publicly reproducible. However, the results are interesting and the results in [5] highlights the potential need for models that are able to discriminate between different textural features for clothing parsing.

Qualitative results. In our qualitative analysis we show two of the best scoring predictions form our model, and the confusion matrix for the CFPD dataset.

In figure 1 we show two top scoring predictions on the refined Fashionista dataset. The main source of prediction errors in these examples are predictions along the edges of the semantic regions. There is a bag present in each image which the model detects. The model also detects the bracelet in the upper image and the necklace in the lower image. However, applying the CRF leads to over-smoothing such that skin is predicted instead of the bracelet.

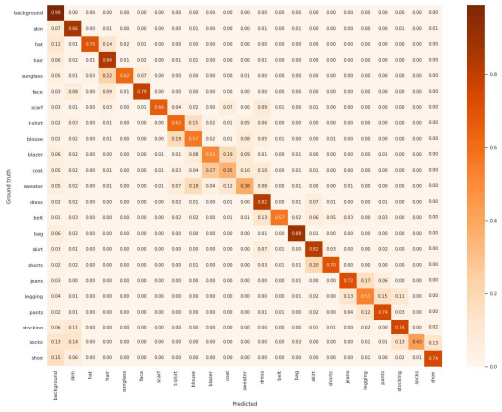

Figure 3: Confusion matrix for predictions on the test set for the CFPD dataset. (Best viewed zoomed in.)

We also plot the confusion matrix for the CFPD test dataset (see figure 3). The class labels have been ordered such that clothing items that are typically worn on neighbouring parts of the human body are close to each other on the diagonal of the confusion matrix. We can see that clothing items that are typically worn on different parts of the human body, such as pants and sweater, are not often confused with each other. However, there are some confusions between classes along the diagonal of the confusion matrix, meaning that clothing items that are typically worn on the same part of the body are sometimes confused by the model. We can also see that small objects, such as belt, necklace, scarf/tie, watch/bracelets are often confused for the clothing items that they are typically occluding.

6. Discussion

In this work, we have demonstrated that a simple feature pyramid network gets state-of-the-art accuracy and mean intersection-over-union on two standard clothing parsing datasets. The method does not need a conditional random field to perform well, in fact it outperforms previous state-of-the-art methods in accuracy even without a CRF, and uses fewer parameters than previous state-of-the-art methods which is an advantage in settings with low computational budgets.

The model learns which body parts of the human body that different clothing items are typically placed on. However, distinguishing between, for example, a t-shirt and a blouse or a pair of leggings and a pair of pants remains a challenge. To discriminate between, e.g, a blouse and a t-shirt low level textural features, for example from Gabor filters, may be necessary [5]. The early layers of a CNN usually learns Gabor-like filters and color blobs [13]. We argue that the importance of the feature pyramid is that both low and high level features are used in concert in the final pixel-wise prediction which potentially enables discrimination between clothing items that have similar shapes and colors but have different textural patterns.

References

[1] Z. Al-Halah, R. Stiefelhagen, and K. Grauman. Fashion forward: Forecasting visual style in fashion. may 2017. 1

[2] F. Chollet et al. Keras. https://keras.io, 2015. 3

[3] G. Huang, Z. Liu, L. v. der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. aug 2016. 3

[4] W. Ji, X. Li, Y. Zhuang, O. E. F. Bourahla, Y. Ji, S. Li, and J. Cui. Semantic locality-aware deformable network for clothing segmentation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, jul 2018. 2, 3

[5] T. Khurana, K. Mahajan, C. Arora, and A. Rai. Exploiting texture cues for clothing parsing in fashion images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, oct 2018. 2, 3, 4

[6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. dec 2014. 3

[7] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 109–117. Curran Associates, Inc., 2011. 2, 3

[8] T.-Y. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. dec 2016. 1, 2

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

[10] P. Tangseng, Z. Wu, and K. Yamaguchi. Looking at outfit to parse clothing. mar 2017. 2, 3

[11] S. Xie, R. Girshick, P. Dollr, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. nov 2016. 1, 2

[12] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2012. 2

[13] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? nov 2014. 1, 4

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2