What Makes a Style: Experimental Analysis of Fashion Prediction

Moeko Takagi Edgar Simo-Serra Satoshi Iizuka Hiroshi Ishikawa
Department of Computer Science and Engineering
Waseda University, Tokyo, Japan
pinkmin@fuji.waseda.jp {esimo, iizuka}@aoni.waseda.jp hfs@waseda.jp

Abstract

In this work, we perform an experimental analysis of the differences of both how humans and machines see and distinguish fashion styles. For this purpose, we propose an expert-curated new dataset for fashion style prediction, which consists of 14 different fashion styles each with roughly 1,000 images of worn outfits. The dataset, with a total of 13,126 images, captures the diversity and complexity of modern fashion styles. We perform an extensive analysis of the dataset by benchmarking a wide variety of modern classification networks, and also perform an in-depth user study with both fashion-savvy and fashion-naïve users. Our results indicate that, although classification networks are able to outperform naïve users, they are still far from the performance of savvy users, for which it is important to not only consider texture and color, but subtle differences in the combination of garments.

1. Introduction

Due to the high level of variability and subjectivity, fashion understanding remains a complicated problem for computer vision. Unlike traditional problems which have consistent and specific definitions, fashion not depends greatly on individual taste, but also has an important temporal component: outfits and garments are constantly falling in and out of style. In order to be able to design computer vision algorithms to solve fashion understanding, it is first important to understand not only how algorithms see style, but also how different individuals see style.

Recent fashion research has been focusing on using weakly labelled and readily available downloaded from the internet [14, 11]. However, even when learning with weak data and annotations, the different approaches have to be evaluated on strongly annotated datasets [15, 7]. Given the subjectivity and diversity of fashion, creating high quality datasets proves to be a challenge.

Figure 1: Overview of the 14 different fashion style classes in our proposed dataset, which captures a large diversity of styles and subjects.

In this work, we propose a new expert-curated dataset[1] for prediction of fashion styles formed by 13,126 images, each one corresponding to one of 14 modern fashion styles. An overview of the different classes can be seen in Fig. 1. We focus images with full outfits visible, and representing a wide diversity of scenes.

We evaluate our dataset by establishing benchmarks with

[1]Dataset available at http://hi.cs.waseda.ac.jp/~esimo/data/fashionstyle14/.

Table 1: Overview of different fashion-oriented computer vision datasets. We consider whether or not the datasets consists of only worn items or if shop gallery images are also included, whether or not the styles are annotated, and the number of styles and images considered. If the style annotations are automatically or semi-automatically computed from user data we denote them as "weak".

Dataset	Worn Items Only	Style Annotations	Number of Styles	Number of Images
DeepFashion [11]	No	Weak	-	**800,000**
Fashionista [22]	**Yes**	No	-	685
Paperdoll [21]	**Yes**	No	-	339,797
Runway [19]	**Yes**	No	-	348,598
Fashion144k [14]	**Yes**	Weak	-	144,169
HipsterWars [10]	**Yes**	**Yes**	5	1,893[2]
FashionStyle14	**Yes**	**Yes**	**14**	13,126

Convolutional Neural Network (CNN) models commonly used for image classification, and also perform an in-depth user study with both fashion-savvy and fashion-naïve users. Our results indicate that CNN models are still far from human level performance on the fashion style prediction task, although recent models do show a significant increase in performance.

2. Related Work

Analysis of fashion has recently seen an increase in interest in the computer vision community. Initially a focus has been on traditional problems applied to the fashion domain such as semantic segmentation of clothing items [22, 21, 13, 23, 25], classification of apparel [2, 3, 19, 24], and image retrieval [8, 9]. More recently, higher level tasks such as predicting fashion styles [10, 18, 15, 7], predicting fashionability [14] and popularity [20], or forecasting fashion styles [1] have been explored. In this work, we propose a new dataset that we hope can serve as a benchmark style prediction approaches.

Initial approaches have focused on creating high quality, albeit small, datasets such as the Fashionista dataset [22] which consists of only 685 images although each with pixel-level garment annotations. In order to increase the size of the data to improve compatibility with deep learning-based approaches, crowd sourcing has recently seen an increase in usage [11], allowing for the creation of large datasets, although with high levels of noise in both images in labels. One approach to reduce this noise is to use Bayesian optimization with pairwise annotations, which allows obtaining cleaner annotations which is critical for subjective and complicated tasks such as fashion style prediction [10]. However, this approach relies on having many users annotate many images, which is limiting when annotating similar fashion styles that non-fashion experts are unable to distinguish. Another approach is to forego annotating entirely and approaches that are able to leverage weak and noisy labels to learn concepts [18, 15]. While weak and noisy labels are easy to obtain and show promising results for training models, they are not suitable for evaluating and benchmarking these models unlike our proposed dataset.

In this work, we propose a new dataset for evaluation which is based on expert annotations of fashion style. We focus on natural images in which the garments are worn and the outfit is visible. A comparison with existing datasets is shown in Table 1. While there exists larger datasets such as DeepFashion [6], Paperdoll [21], Runway [19], and Fashion144k [14], these lack curated style annotations and rely on either crowd-sourced annotations or weak labels. The Fashionista dataset [22] provides per-pixel garment labels, but no information about the fashion style. The closest dataset to the one we propose is that of HipsterWars [10], however, instead of relying on user annotations for a small number of very dissimilar classes, we focus on more complicated classes with large variability, and rely on expert-curated annotations. Furthermore, we provide an extensive analysis of our dataset evaluated both expert and non-expert human performance, as well as the performance of state of the art machine learning techniques for classification.

3. *FashionStyle14* Dataset

We have collected a new expert-curated dataset, which we denote as *FashionStyle14* that consists of 14 fashion style classes: *conservative*, *dressy*, *ethnic*, *fairy*, *feminine*, *gal*, *girlish*, *casual*, *lolita*, *mode*, *natural*, *retro*, *rock*, and *street*. These classes were chosen by an expert as being representative of modern fashion trends, and covering a large diversity of fashion styles.

The general procedure of obtaining the images was to use a search engine in combination with fashion styles as

[2]Up to half of the images are discarded by filtering with labelling confidence.

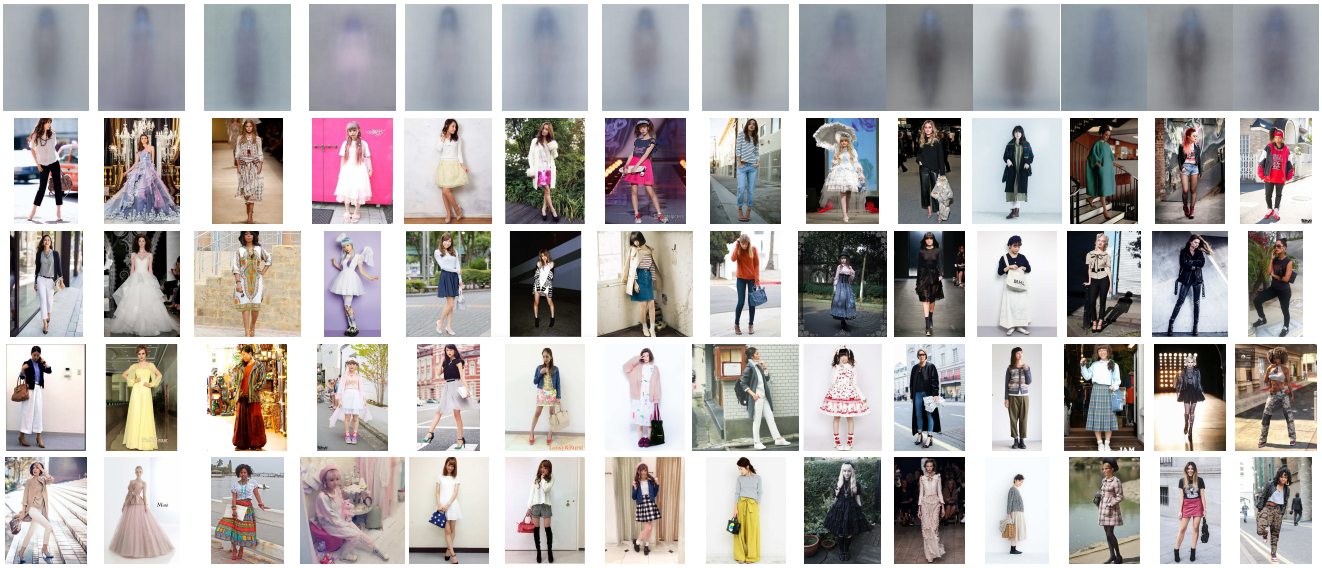


Figure 2: Overview of the proposed dataset. The first row shows the mean image of each class, while the next rows show random examples taken from each class.

query. Afterwards, for each class, roughly 1,000 images were manually selected using the following criterion: 1) being representative of a fashion style, and 2) having the key objects of the fashion coordinate visible. After collecting the images in a first pass, we performed a second pass for quality control in which dubious labeled images were removed. In total, 13,126 images were obtained for all 14 fashion styles.

Examples of images of the different styles and the mean image for each style can be seen in Fig. 2 We can see that some classes do show clear characteristics, such as "fairy", however, the large diversity of the classes, in combination with a variety of poses and backgrounds make the dataset challenging for computer vision techniques.

4. Experiments

We perform analysis of style prediction on our dataset using both state of the art approaches for classification, as well as evaluation with both fashion expert and non-expert user subjects. For evaluation, we use 60% of the dataset for training, 5% of the dataset for validation and 35% of the dataset for testing.

4.1. Classification Networks

We evaluate the state of the art classification networks by fine-tuning them on the training set, using the validation set to choose the best performing model. We compare the VGG16 and VGG19 models [16], which are variants of the model that won the ILSVRC2014 image classification competition, the Inception v3 model [17], the ResNet50 model [6] model, which won the ILSVRC2015 image classification competition, and the Xception model [4]. The VGG16 model consists of 16 layers that can be learnt: 13 convolutional layers and 3 fully connected layers. The VGG19 model builds upon the VGG16 model by adding 3 additional convolutional layers and further training the model. The Inception v3 model uses modules which consist of combining convolutions with different kernel sizes denominated "inception modules". ResNet50 consists of 50 layers, where the basic building block is formed by two convolutional layers with skip-connections, which allows training models with more layers. The Xception model is an improvement over the Inception v3 model that introduces a depth-wise separable convolution operation that is able to more efficiently use the model parameters.

For all approaches we initialize the weights from models trained on ImageNet [12] for 1000-class classification. Afterwards, the models are fine-tuned using the training split with the Stochastic Gradient Descent (SGD) algorithm. We train on the training data with learning rates of 10^{-4}, 10^{-5}, and 10^{-6}, and use the best performing model evaluated on the validation set for each architecture. In particular, we find that the VGG19 model uses a learning rate of 10^{-6} while the rest of the models perform best with a learning rate of 10^{-5}.

Results are shown in Table 2. We can see that for this application, the choice of network heavily influences the result. In particular, ResNet50 shows significant performance, while other networks fail significantly behind, especially on easy to confuse classes such as "conservative" or "retro".

Table 2: Comparison of different networks fine-tuned for classification. Best result is highlighted in bold.

Model	conserv.	dressy	ethnic	fairy	feminine	gal	girlish	casual	lolita	mode	natural	retro	rock	street	**mean**
ResNet50	**0.66**	**0.91**	**0.74**	**0.88**	**0.64**	**0.74**	**0.47**	**0.66**	**0.92**	**0.72**	**0.70**	**0.62**	**0.68**	**0.69**	**0.72**
VGG19	0.54	0.79	0.57	0.81	0.43	0.50	0.26	0.54	0.80	0.62	0.56	0.42	0.53	0.60	0.58
Xception	0.44	0.79	0.63	0.84	0.45	0.50	0.33	0.54	0.80	0.61	0.56	0.44	0.52	0.53	0.58
Inception v3	0.37	0.73	0.54	0.78	0.41	0.39	0.27	0.45	0.78	0.55	0.44	0.35	0.47	0.46	0.51
VGG16	0.31	0.78	0.49	0.78	0.42	0.45	0.22	0.43	0.81	0.58	0.57	0.23	0.43	0.43	0.51

4.2. User Study

We further evaluate the dataset by performing a user study with both users who are self-defined as fashion-savvy, and naïve users that do not profess fashion knowledge. A total of 8 fashion-savvy users and 11 naïve users participated in the user study. All users were shown examples of the different classes taken from the training data and told to annotate images from the test set. In particular, fashion-savvy users classified 1,400 images each (roughly 100 per class), while the naïve users classified 420 images each (roughly 30 per class).

Results are shown in Table 3. We can see that the savvy users should very high performance, outperforming all networks in mean performance. We also see a large gap between naïve and savvy users for all classes. It is worth pointing out that for the "conservative" and "mode" classes, the fine-tuned ResNet50 model outperforms the savvy users, and in the case of the "dressy" class performance is tied.

4.3. Human vs Machine

We perform a more in-depth analysis of the different mistakes of both humans and machines in order to analyze the level of objectivity of the dataset. We first perform a quantitative analysis using the Normalized Mutual Information (NMI) score, which is a value in the $[0, 1]$ range, where 0 corresponds to no mutual information and 1 corresponds to perfect correlation. An overview of the NMI performance of the different networks can be seen in Table 4. We can see a larger gap in NMI score in comparison with mean accuracy. Furthermore, we analyze both the savvy users and naïve users comparing both to the ground truth labels and the labels predicted by the ResNet50 model in Table 5. We can see that even though the accuracy performance is not that different, there is a large gap between the NMI scores of the ResNet50 model and the savvy users, and the margin between the naïve users and the ResNet50 model becomes very small. This indicates that the classification mistakes of the ResNet50 model and the users are significantly different. We next further analyze this phenomena qualitatively.

We show some erroneous classification examples in Fig. 3. The top row shows results in which even though the savvy users don't agree, although the Convolutional Neural Network (CNN) is able to predict the right class, while the bottom row shows cases in which the savvy users generally all agree, but the CNN fails at predicting the right class. In general, we find that the CNN tends to make different mistakes than the users, highly influenced by textures and colours, such as the image on the bottom-left. On the other hand, users tend to make mistakes on conceptually similar classes such as the image in the top-right.

4.4. What makes the style?

We also perform a qualitative analysis of what the best performing network is looking at inside the network in order to evaluate what is critical for the different fashion styles. In particular, we use the approach of Fong et al. [5], which is based on learning a mask that attempts to suppress the accurate classification of an image, in combination with the top-performing fine-tuned ResNet50 model.

Results are shown in Fig. 4. We can see the network focuses on the individuals and the region of interest varies greatly from image to image. It is curious to see how the sunglasses are used to determine that the fourth image from the left on the top row is of the class "rock", how the frills on the skirt are what allows the network to classify the third image from the left on the top row as "lolita", and how the belt of the second image from the left on the bottom row is what makes the outfit "conservative" .

5. Discussion and Conclusions

We have presented a new expert-curated dataset for fashion style prediction of 14 modern fashion styles. We also have provided an in-depth evaluation of the dataset by benchmarking modern classification networks, in which we find that more recent architectures based on residual learning show a significant improvement over other approaches. We also perform a user study with both fashion-savvy and fashion-naïve users. While machine learning approaches are able to outperform the naïve users, we find that performance is still far from the savvy users. The classification networks, similar to the naïve users, tends to make decisions based on textures and salient objects instead of considering subtle nuances like the savvy users, indicating that there is

Table 3: Classification accuracy for different users. We compare savvy users and naïve users with the best performing fine-tuned network for all classes. For the mean value we also display the standard deviation in parenthesis. Best results are shown in bold.

Model	conserv.	dressy	ethnic	fairy	feminine	gal	girlish	casual	lolita	mode	natural	retro	rock	street	**mean**
ResNet50	**0.66**	0.91	0.74	0.88	0.64	0.74	0.47	0.66	0.92	**0.72**	0.70	0.62	0.68	0.69	0.72 (0.117)
Savvy users	0.59	**0.92**	**0.80**	**0.89**	**0.84**	**0.92**	**0.71**	**0.75**	**0.95**	0.69	**0.81**	**0.79**	**0.74**	**0.91**	**0.82 (0.101)**
Naïve users	0.35	0.87	0.64	0.83	0.60	0.62	0.51	0.50	0.83	0.29	0.57	0.50	0.58	0.74	0.62 (0.161)

2 natural, 1 girlish
95% natural, 4% girlish

2 conservative, 1 casual
31% conservative, 23% natural, 21% feminine, 11% casual

1 mode, 2 casual
59% mode, 28% rock

1 retro, 1 girlish, 1 mode
97% retro, 1% girlish, 1% mode

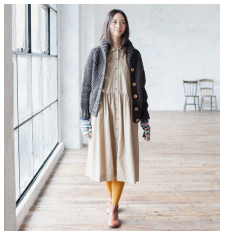
1 rock, 1 street, 1 gal
92% rock, 6% street

3 dressy, 1 fairy
87% fairy, **9% dressy**

3 retro
63% dressy, **15% retro**

3 casual
81% conservative, 17% mode

4 girlish, 1 natural
68% natural, 30% casual, 2% **girlish**



3 natural
99% dressy, 1% feminine

Figure 3: Qualitative analysis of classification mistakes by both users and the fine-tuned ResNet50 model on the test set. Below each image, we display the labels provided by different users, and the prediction results of the network model. The user labels are shown as integer values in which the number of users that assigned the image the label is displayed, while the network model prediction results are shown in percentage values. As each user randomly classified a different subset of the test set, the number of user labels varies per image. We only show images in which at least 3 users have provided labels.

Table 4: Comparison of the NMI score for the different fine-tuned networks. Best result is highlighted in bold.

Model	ResNet50	VGG19	Xception	VGG16	Inception v3
NMI	**0.58**	0.43	0.42	0.37	0.35

Table 5: Comparison of the NMI score between the labels provided by the different users, the ground truth labels, and the labels computed by the fine-tuned ResNet50 network.

	Savvy Users	Naïve Users
Ground Truth	0.75	0.58
ResNet50	0.55	0.49

still significant room for improvement in fashion style classification.

Although in this work we have benchmarked supervised approaches, we believe that the proposed dataset will be useful for the evaluation of unsupervised and semi-supervised approaches for fashion style such as [15, 7]. Using readily available weak data is a very promising direction for learning to understand fashion that can greatly benefit

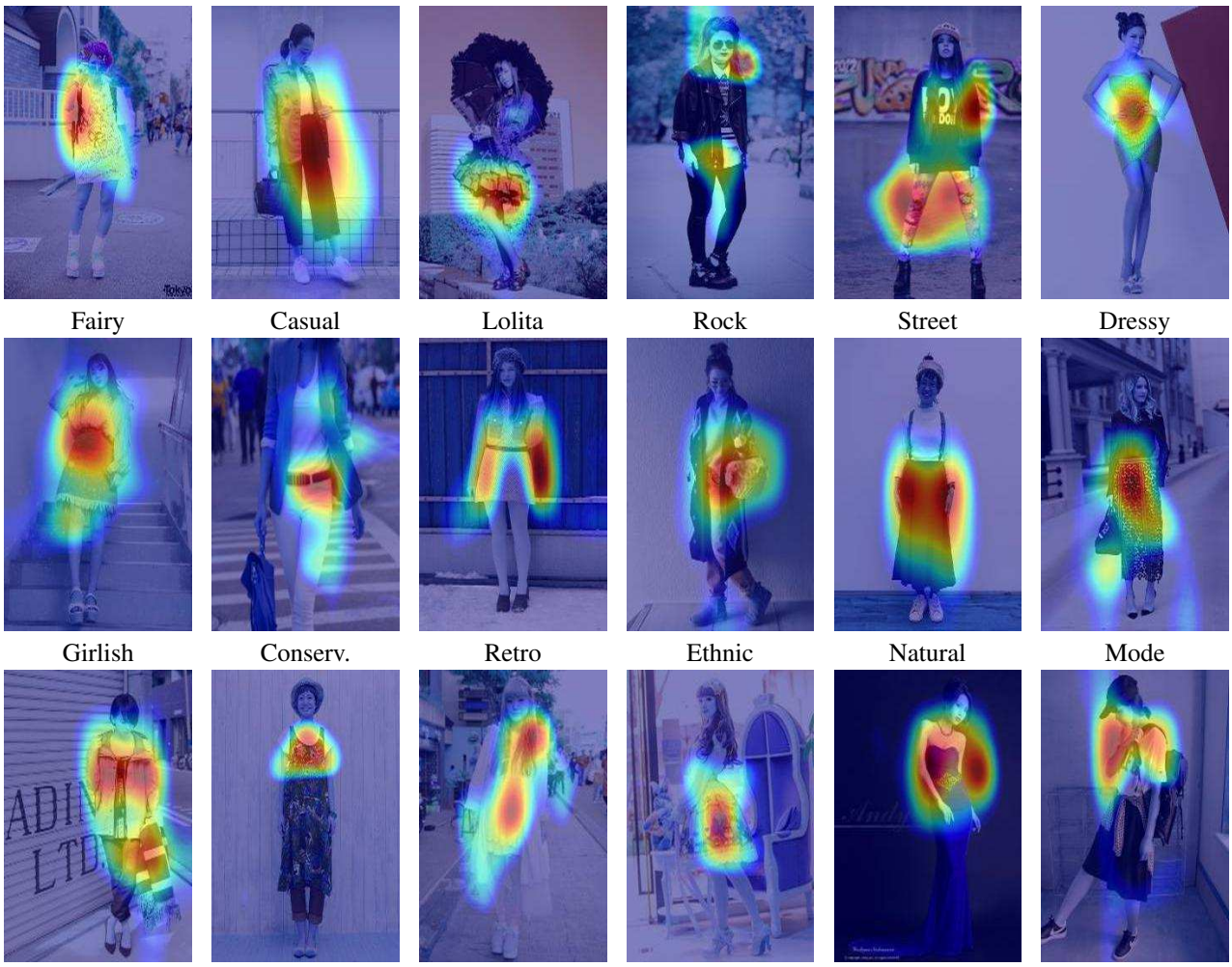


Figure 4: Visualization of the region of interest the fine-tuned ResNet50 model is focusing on to correctly classify the images. The true and predicted class label is shown below each image.

from more rigorous evaluation on datasets such as the one we presented in this work.

Acknowledgements: This work was partially supported by JST-CREST Grant Number JPMJCR14D1.

References

[1] Z. Al-Halah, R. Stiefelhagen, and K. Grauman. Fashion forward: Forecasting visual style in fashion. *arXiv preprint arXiv:1705.06394*, 2017. 2

[2] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. V. Gool. Apparel classifcation with style. In *ACCV*, 2012. 2

[3] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*, 2012. 2

[4] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2016. 3

[5] R. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *arXiv preprint arXiv:1704.03296*, 2017. 4

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 3

[7] W.-L. Hsiao and K. Grauman. Learning the latent" look": Unsupervised discovery of a style-coherent embedding from fashion images. *arXiv preprint arXiv:1707.03376*, 2017. 1, 2, 5

[8] J. Huang, R. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *ICCV*, 2015. 2

[9] M. Kiapour, X. Han, S. Lazebnik, A. Berg, and T. Berg. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, 2015. 2

[10] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*, 2014. 2

[11] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 1, 2

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015. 3

[13] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. A High Performance CRF Model for Clothes Parsing. In *ACCV*, 2014. 2

[14] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. Neuroaesthetics in Fashion: Modeling the Perception of Fashionability. In *CVPR*, 2015. 1, 2

[15] E. Simo-Serra and H. Ishikawa. Fashion Style in 128 Floats: Joint Ranking and Classification using Weak Data for Feature Extraction. In *CVPR*, 2016. 1, 2, 5

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 3

[18] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, and S. Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *ICCV*, 2015. 2

[19] S. Vittayakorn, K. Yamaguchi, A. C. Berg, and T. L. Berg. Runway to realway: Visual analysis of fashion. In *WACV*, 2015. 2

[20] K. Yamaguchi, T. L. Berg, and L. E. Ortiz. Chic or social: Visual popularity analysis in online fashion networks. In *ACMMM*, pages 773–776, 2014. 2

[21] K. Yamaguchi, M. H. Kiapour, and T. L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, 2013. 2

[22] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012. 2

[23] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Retrieving similar styles to parse clothing. *PAMI*, 2014. 2

[24] K. Yamaguchi, T. Okatani, K. Sudo, K. Murasaki, and Y. Taniguchi. Mix and match: Joint model for clothing and attribute recognition. In *BMVC*, 2015. 2

[25] W. Yang, P. Luo, and L. Lin. Clothing co-parsing by joint image segmentation and labeling. In *CVPR*, 2014. 2