Multi-Label Fashion Image Classification with Minimal Human Supervision

Naoto Inoue[1] Edgar Simo-Serra[2] Toshihiko Yamasaki[1] Hiroshi Ishikawa[2]
[1]The University of Tokyo [2]Waseda University
inoue@hal.t.u-tokyo.ac.jp esimo@aoni.waseda.jp yamasaki@ay-lab.org hfs@waseda.jp

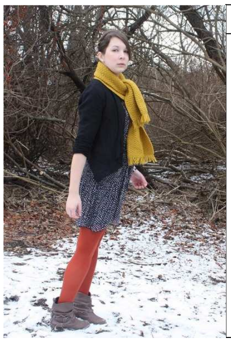
Abstract

We tackle the problem of multi-label classification of fashion images, learning from noisy data with minimal human supervision. We present a new dataset of full body poses, each with a set of 66 binary labels corresponding to the information about the garments worn in the image obtained in an automatic manner. As the automatically-collected labels contain significant noise, we manually correct the labels for a small subset of the data, and use these correct labels for further training and evaluation. We build upon a recent approach that both cleans the noisy labels and learns to classify, and introduce simple changes that can significantly improve the performance.

1. Introduction

Fashion plays an important role in everyday lifestyle, yet understanding fashion is still a very complicated task for computer vision. In particular, due to the subjectiveness of the fashion, obtaining high-quality data for training learning-based models for fashion task remains an open problem. In this work, we analyze the importance of the quantity of the data and evaluate the modern tools for automatically cleaning the data in the context of multi-label prediction.

Unlike more standard computer vision tasks, the fashion domain is variable and subjective, and networks pre-trained on large-scale datasets such as ImageNet and Places are not necessarily best-suited for fashion-related tasks. Simo-Serra and Ishikawa [21] introduced Fashion144k, which consists of fashion images in a variety of scenes with weak labels. Furthermore, they showed that exploiting a large number of fashion images with only weak labels and designing a CNN to learn from them can provide a better feature representation than ImageNet or Places for one of the fashion-related tasks, i.e., fashion style classification. Thus, obtaining a larger-scale fashion image dataset is desirable as a basis for fashion recognition tasks. We examine two aspects that have an impact on the recognition performance on fashion image datasets: (i) the size of the dataset and (ii) the quality of the annotations in the dataset. We first extend the Fashion144k dataset [20] by crawling over 1M images with metadata, which contains weak labels, from Chictopia[1]. To evaluate the recognition performance, we perform an experiment in multi-label classification, as shown in Fig. 1, and predict the colors and the garments worn by the person in each image. To evaluate the quality of annotations, we randomly pick 5,300 images in our new dataset, which we denote as *Fashion550k*, and manually correct the weak labels to obtain "clean" labels. We use these limited annotations to both evaluate and improve the recognition accuracy. Note that the number of the labels in this subset is quite small, compared to the whole dataset. We build upon the approach of Veit *et al.* [25] that uses a neural network to clean the labels by learning a mapping between the noisy labels and clean labels. By learning this mapping, it is possible to afterwards jointly clean the dataset and train a prediction model. We evaluate this approach and show that with some simple modifications we are able to significantly improve performance.



Label	Confidence
Stockings & Tights	0.99
Scarf	0.98
Mustard	0.95
Blazer	0.93
Gray	0.90
Black	0.89
Boots	0.82
Dress	0.76
Red	0.38
..	..

Figure 1: Example of fashion image classification results provided by our model. We achieve this performance by combining several thousand manual annotations with our automatically-collected large-scale very noisy dataset.

To summarize, our contributions are as follows:

[1] Dataset available at http://hi.cs.waseda.ac.jp/~esimo/data/fashion550k/.

- We present a large-scale weakly-labeled fashion image dataset that includes 5,300 human-annotated images for analyzing the label noise.
- We investigate the effect of the size of the training data for multi-label fashion classification.
- We evaluate the methods to exploit weak labels jointly with human-verified data.

2. Related Work

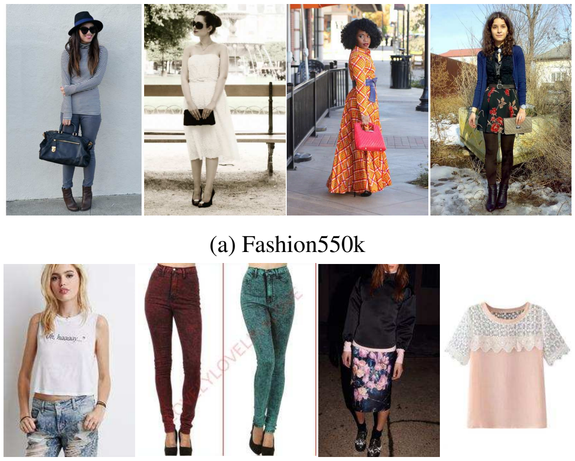
2.1. Fashion

There are many fashion-related problems in computer vision research, such as semantic segmentation of garments [19, 29, 31, 13, 12], image retrieval [8, 5, 33], fine-grained classification of garments and styles [2, 3, 14, 10, 30], fashion landmark detection [15], and image-conditional domain transfer [32]. Predicting more subjective properties such as popularity [26] and fashionability [20] have also been studied. Many of these approaches collected large scale image datasets by crawling the Internet. They also exploit weak labels [20, 28] or manually annotate the images by crowd-sourcing [14]. In this paper, we construct the Fashion550k dataset that consists of fashion images with weak annotations, and evaluate the influence of noisy annotations. We believe this dataset is also useful for other tasks such as fashionability estimation and clothing parsing (as a weak supervision).

2.2. Learning from Noisy Data

There are two approaches to learn from noisy labeled data with a neural network: the first approach aims to directly learn from noisy labels and focus mainly on constructing a model that takes noise into consideration. The simplest approach is to model a label noise that is conditionally independent of the input image [18, 23]. Xiao *et al*. [27] and Misra *et al*. [17] proposed image-conditioned noise model. However, these methods face the challenge of distinguishing difficult samples from mislabeled training samples, and rely on heuristic-based noise models.

Another approach is called semi-supervised approach, which is to learn from noisy labeled data with neural network is to combine a small set of clean labels. Veit *et al*. [25] proposed to learn the mapping between noisy and clean labels and then to exploit the mapping for training deep neural networks by imitating the student-teacher models used in [16, 7]. In this approach, a base CNN is used as a feature extractor and is combined with two additional networks consisting of a label cleaning network and an image classifier. The label cleaning network is input the noisy labels in addition to the visual features, and is trained to predict cleaned labels that are provided as a privileged information. The output of the cleaning network is regarded as the target label to predict for the image classifier. This approach can handle both incorrect and missing labels without assuming the type of noise and greatly boost the classification performance. We evaluate this approach, propose simple changes to improve the performance, and furthermore propose an alternative scheme in which the confidence of the label cleaning network is used.



(a) Fashion550k

(b) DeepFashion

Figure 2: Example images from our Fashion550k and DeepFashion [14]. Images in Fashion550k often have fully-visible centered individuals with natural and diverse background. On the other hand, DeepFashion is constructed to recognize details and landmarks of each clothing. Thus, images in DeepFashion are focused on each item and with more clear backgrounds and surroundings.

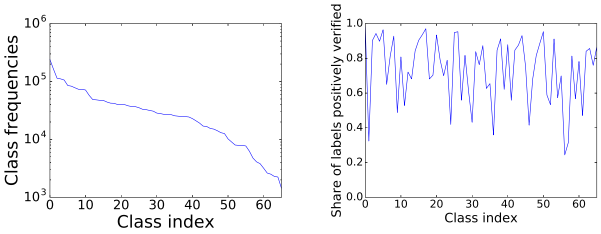
3. Dataset

We extended the Fashion144k dataset [20]. The extended dataset, called Fashion550k, consists of 550,661 user posts from the clothing-oriented website Chictopia. Fashion550k is about four times the size of the original Fashion144k. In Fig. 2, example images of Fashion550k are shown. Each post contains at least a single image with metadata including label tags. Each photograph shows a different angle of the user or a zoom-in on different garments. Users often add a short description of what they are wearing and tags of the types and colors of the garments. Note that this information is very noisy: e.g., not all garments are tagged, and some users make only a part of the information available. We parsed all the information provided with each post. For a representative statistics of the dataset, refer to Table 1.

Since Fashion550k has been collected without any filtering, some of the images are not suitable for learning. For example, some images focus on only one object, or have strong filters. Thus, we cleaned the data following the ap-

Table 1: Statistics of the dataset. The number between the brackets indicates standard deviation of each property.

Property	Total (Unique)	Per user	Per post
posts	550,661	36.19 (71.41)	-
users	15,217	-	-
photos	1,061,468	1.76 (0.90)	1.93 (1.26)
tags	3,627	1.98 (1.55)	2.14 (2.03)



(a) Class frequencies.

(b) Quality of original labels.

Figure 3: Label statistics for Fashion550k. We show and order by the class frequency, and represent the quality of the original labels as the accuracy of the noisy labels compared to manually cleaned labels.

proach of [21]. In [21], to filter out such images the authors fine-tune a VGG16 model [22] pre-trained on ImageNet for the binary classification task of whether or not given image is suitable for training. They achieve about 94% accuracy. By directly using this classifier on our crawled images, we obtain images with a fully-visible centered individual. Additionally, we filtered out all images with no tags, to obtain 407,772 images in total.

As weak annotations, we utilize tags that consist of *colour-garment* pairs, such as *blue-jeans* and *red-sweater*. We split the tags into colors and garments, resulting in a total of 123 unique weakly-annotated tags. However, some tags, such as *orange* and *carrot-orange*, or *watch* and *bracelet*, are hard to visually discriminate under various illumination and lighting conditions. Furthermore, some tags are hard to detect because occurrence is minimal or because they are highly occluded, as in the case of *iPhone-case* and *earrings*. Thus, we discard or merge such tags, resulting in a total of 66 unique weakly-annotated tags. The classes in this dataset are not evenly distributed, as shown in Fig. 3a. The class *black* has more than 200,000 annotations, whereas the class *tie* has only 1,435 annotations.

We additionally verified the tags in a subset of the collected images to provide minimal supervision. The number of the images verified was 3,000 for training, 300 for validation, and 2,000 for testing. The rest of the images (called the noisy dataset) are used to train the baseline classifier. We estimate the quality of the noisy labels using the verified tags for the 5,300 images. Fig. 3b shows the distribution of the quality of the original noisy labels. The noise occurs regardless of the class frequency. We observe that 26.4% of the original labels are false positives. Further, we also observe that originally positive labels accounts for only 54.2% of the labels actually verified positive. These observations show that directly using these weak labels as ground truth will lead to poor performance, which we will later validate in our experiments.

4. Proposed Method

In this paper, our goal is to train a multi-label image classifier on a large dataset with extremely noisy labels, where additionally a small subset of the dataset that has human-verified labels available. This setting can often happen when we collect images from the web or social media and have experts to correct some of the labels.

Formally, we have a very large training dataset T. T consists of tuples of noisy labels $\boldsymbol{y}$ and images $\boldsymbol{x}$, $T = \{(x_i, y_i), \ldots\}$. Additionally, we have a small dataset V with human verified labels v, $V = \{(x_j, y_j, v_j), \ldots\}$. The number of the data in T is significantly larger than that in V. In our experiments, T exceeds V in the number of data by two orders of magnitude. Each $\boldsymbol{y}$ and $\boldsymbol{v}$ is a sparse d-dimensional vector with a binary annotation for each of the d classes indicating whether it is present in the image or not. Our aim is to fully utilize the accurate annotation verified by a human in V and noisy but huge number of labels in T.

4.1. Multi-Task Label Cleaning Network

We base our approach on the model of Veit *et al*. [25], in which a label cleaning network is used in combination with a classification network. We next briefly summarize [25].

4.1.1 Model architecture

The network is designed to jointly learn to generate accurate labels from noisy labels and to learn a more accurate multi-label classifier from the generated labels. An overview of the model is shown in Fig. 4. There are two classifiers g and h on top of a CNN-based feature extractor f.

The first classifier g, shown on the bottom of Fig. 4, is called the label cleaning network. It learns a mapping from noisy labels $\boldsymbol{y}$ to human-verified labels $\boldsymbol{v}$, conditional on the input image. Its output $\boldsymbol{c}$ denotes the cleaned labels. The classifier g has two separate inputs, the noisy labels $\boldsymbol{y}$ and the visual features $f(\boldsymbol{x})$. Each input is projected into an embedding by a linear layer and the two are concatenated, then transformed with a hidden linear layer. Finally, $\boldsymbol{y}$ is added to the output by an identity-skip connection and clipped to $[0, 1]$ to remain in the valid label space. In short, $\boldsymbol{c}$ is com-

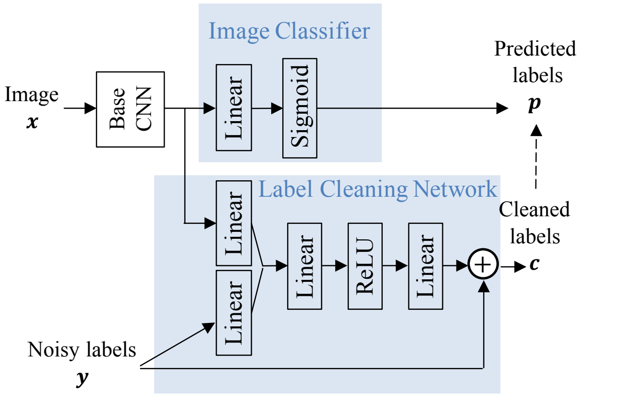


Figure 4: Overview of the multi-task label cleaning network in [25]. Dashed arrow in the figure indicates a data flow without gradient back-propagation.

puted as follows:

$$\boldsymbol{c} = \max(\min(g(f(\boldsymbol{x}), \boldsymbol{y}) + \boldsymbol{y}, 1), 0) \quad (1)$$

The second classifier h is called the image classifier. It learns to predict labels by imitating the first classifier using only the image as input. The image classifier h is shown in the top row of Fig. 4. It is composed of a linear layer followed by a sigmoid as an activation function. We denote the predicted labels by $\boldsymbol{p}$. It is a d-dimensional vector in $[0, 1]^d$ and computed by $\boldsymbol{p} = h(f(\boldsymbol{x}))$. It indicates the likelihood of the visual presence of the d classes and is used to evaluate the quality of the whole network.

4.1.2 Training strategy

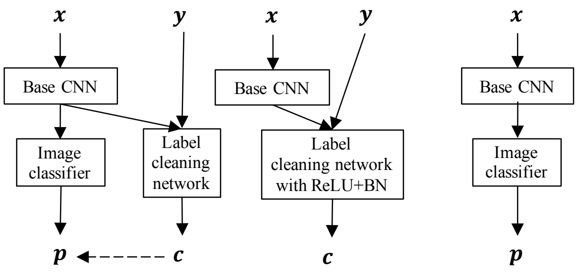
Two losses are used to train the model: the label cleaning loss L_{clean} to enhance the quality of the cleaned labels $\boldsymbol{c}$ and the classification loss L_{classify} to enhance the quality of the predicted labels $\boldsymbol{p}$.

The label cleaning network is supervised by the verified labels of all samples j in the human verified set V. The cleaning loss is based on the L_1-distance between the cleaned labels y_j and the verified labels v_j.

$$L_{\text{clean}} = \sum_{j \in V} |c_j - v_j| \quad (2)$$

The classification network is also supervised by two terms. For all samples i from the noisy subset T, the image classifier is supervised by the cleaned labels c_i that is produced by the cleaning network. For samples j from verified dataset V, we can directly supervise p_i by verified labels v_i. For both terms, a cross-entropy loss is used.

$$L_{\text{classify}} = -\sum_{j \in V} p_j \log(v_j) - \sum_{i \in T} p_i \log(c_i) \quad (3)$$



(a) Veit *et al.*

(b) Phase1

(c) Phase2

Figure 5: Comparison between Veit *et al.* and the proposed method.

The cleaned labels c_i is regarded as constant in order to prevent a trivial solution $c_i = p_i = 0$.

4.2. Proposed Method

In Veit *et al.*, as shown in Fig. 5a, the two networks are trained simultaneously. However, they assume a large number of verified labels actually. They use about 40K images with verified labels in their paper. If we have much smaller number of verified labels, the label cleaning network tends to overfit.

Thus, we propose to improve the label cleaning network. To avoid overfitting, ReLU and batch normalization (BN) [9] are added after each linear layer in the label cleaning network. Since this network tends to overfit even with this modifications, we use the validation subset of Fashion550k to find the best model and do early stopping.

Further, for more stable training, we also propose to separate the whole training process into two phases. We train the label cleaning network on a pre-trained base CNN as shown in Fig. 5b. Further, we freeze the whole network and feed all images in the noisy subset and get cleaned labels $\boldsymbol{c}$. Subsequently, we train the image classifier on another pre-trained base CNN as shown in Fig. 5c. As target labels, $\boldsymbol{v}$ is used for the training subset, and $\boldsymbol{c}$ is used for the noisy subset. The loss used in our method is the same as employed in Section 4.1.2.

5. Experimental Results

We train our model on the noisy and training subset, find the best model and hyper-parameter using the validation subset, and evaluate on the test subset in Fashion550k. We evaluate our approach using multi-label classification. For each of the 66 target classes, we predict a score which indicates the likelihood of the concept described by the class presenting in the image.

As evaluation metrics, the class-agnostic average precision (AP_{all}), and the mean of the each class-average precision (mAP) are used. AP_{all} regards every annotation for all classes equally by handling them as coming from one single class and thus is biased towards more frequent classes.

5.1. Baseline and Compared Methods

As the baseline model for our evaluation, we train a CNN on the noisy labels from Fashion144k and Fashion550k. 117,746 and 402,472 images are used to train on Fashion144k and Fashion550k, respectively. This model is called **Baseline** and is used as the initial weight for all the other compared methods except ours. The **Baseline** model is based on a 50-layer ResNet [6] pre-trained on ImageNet [4]. The last softmax layer of the network is replaced with a 66-way sigmoid layer to predict the probability of each of the labels.

We compare the following approaches:

Fine-tune with clean labels. Most common approach with cleaned labels is to feed the clean labels directly to the network and supervise the last layer. However, this approach is prone to overfitting, since the number of clean labels is limited and cannot utilize huge noisy labels fully.

Fine-tune with mix of clean and noisy labels. This approach handles the insufficient training examples. We fine-tune the last layer of the base CNN with a combination of training samples from the small clean and the large noisy subset (in a 1 to 9 ratio).

Veit *et al.* [25]. This approach utilizes the clean labels using the label cleaning network. There are two variants of this method, "with pre-training" and "trained jointly". In "with pre-training", first, we train just the label cleaning network. Subsequently, the classification network and the image classifier are jointly trained. In this training, a learning rate of the classification network is smaller ($\times 10^{-1}$) than the image classifier. In "trained jointly", the classification network and the image classifier are trained jointly from the beginning. A difference with the original implementation is that in this case we opt to use the ResNet50 [6] as the Base CNN instead of the Inception v3 model [24]. For further details, please refer to the original paper.

Improved Model. This approach only implements our modification of adding ReLU and BN to the label cleaning network, which is proposed in Section 4.2. This model is trained in the same way as original method of Veit *et al.*

Ours. This approach implements both using **Improved Model** and the two-phase training shown in Fig. 5b and Fig. 5c. To reduce the computational overhead, we initialize the base-CNN and the prediction network with the weight of **Fine-tune with clean labels**.

Table 2: Comparison of the models evaluated by the test subset of Fashion550k. We use all 3000 images with cleaned labels for the training. (noisy:clean $\simeq$ 136 : 1)

Model	AP_{all}	mAP
Baseline (Fashion144k)	62.23	49.66
Baseline (Fashion550k)	69.18	58.68
Fine-tuning with mixed labels	72.38	61.50
Fine-tuning with clean labels	79.39	64.04
Veit *et al.* (pretrain)	78.60	62.81
Veit *et al.* (joint)	78.92	63.08
Improved Model (pretrain)	**80.01**	64.34
Improved Model (joint)	79.70	64.03
Ours	79.87	**64.62**

5.2. Training Details

All approaches are implemented using PyTorch [1] and optimized with Adam [11] by a batch size of 64. All the images are resized to 256×256. In the training phase, the images are then randomly cropped into 224×224 with random horizontal flipping. In the test phase, the images are center-cropped into 224×224. We employed early stopping using AP_{all} on the validation subset.

The baseline network is initially trained with the binary cross-entropy loss between the noisy labels and the predictions of the network for 100,000 iterations using an initial learning rate of 10^{-3}. For fine-tuning, we use a learning rate of 10^{-4} for the last linear layer and 10^{-5} for the other layers, for additional 20,000 iterations.

The other variants are trained for additional 20,000 iterations. In this paper, we regard the ResNet except the last two layers, i.e., linear layer and sigmoid layer, as the base-CNN. The last two layers are regarded as the image classifier. For the label cleaning network, the arrangement of the linear layers are the one used in Veit *et al.* as in Fig. 4 and the number of filters for the each linear layer is set to 512. We use an initial learning rate of 10^{-5} for the base-CNN and 10^{-4} for the other layers. To train the cleaning network and image classifier jointly, we sample training batches that contain samples from T as well as V in a ratio of 9:1. To balance the losses, we weight L_{clean} with 0.1 and L_{classify} with 1.0 for the variants of Veit *et al.*, which is the same as the parameters used in the paper.

5.3. Results

We first discuss the overall performance of the proposed method in Table 2. The performance regarding AP_{all} is higher than mAP. This means that the AP for popular classes is higher. Training solely on the noisy labels from our Fashion550k, which is about four times bigger than Fashion144k [20], shows significant performance gain and shows about $+7\%$ improvement in AP_{all} and about $+9\%$ improvement in mAP compared to training solely on noisy labels from Fashion144k. This result suggests that collecting larger dataset is still necessary for the recognition performance improvement.

Table 3: Examples from the test subset in Fashion550k. We show the top 5 most confident predictions along with whether the prediction is correct or incorrect. Our approach consistently removes false predictions.

Image from the test subset	Top5 predictions	Baseline	Fine-tuning with clean labels	Ours
	Dress	✓	✓	✓
	Bag & Purse	✓	✓	✓
	Sunglasses		✓	✓
	White	✓	✓	✓
	Pink	✓		✓
	Sandals		×	
	Brown	×		
	Skirt	✓	✓	✓
	Black		✓	✓
	Stockings & Tights		✓	✓
	Brown	✓	✓	✓
	Red			✓
	Shirt	×	×	
	Ruby-red	×		
	Shoes	×		
	Dress	✓	✓	✓
	Black	✓	✓	✓
	Red	✓	✓	✓
	Orange	×	×	×
	Brown			✓
	Socks		×	
	Shoes	×		

Then we conduct an analysis on combining noisy and clean labels. Simply fine-tuning with cleaned labels improves AP_{all} and mAP significantly because Fashion550k is very noisy. We would like to emphasize that the two variants of the original Veit *et al.* do not work and are even worse than fine-tuning with clean labels. Further, there is still certain improvement in improved model and ours in both AP_{all} and mAP over fine-tuning with clean labels. This shows the importance of carefully designing the label cleaning network, which results in over 1% improvement both in AP_{all} and mAP compared to the original method of Veit *et al.* Notably, ours achieved best mAP. The difference in improved models and ours are not so remarkable. This suggests that the order of the training is less important. In Table 3, the example results of classification are shown.

We perform an analysis on a harder setting, where we only have access to only 1,000 cleaned annotations. The gap between fine-tuning with clean labels and the label cleaning network based methods are clear as shown in Table 4. Notably, the performance improvement by the proposed methods, i.e., $+0.6\% \sim +0.8\%$ in AP_{all} is not subtle compared to the improvement when we use 3 times larger number of cleaned labels , i.e., $+1.4\%$ in AP_{all}, as shown in Table 2.

Table 4: Comparison of the models evaluated by the test subset of Fashion550k when we are restricted to use only **1000** cleaned annotations for the training. (noisy:clean $\simeq$ 407 : 1)

Model	AP_{all}	mAP
Baseline (Fashion550k)	69.18	58.68
Fine-tuning with clean labels	78.03	62.48
Improved Model (pretrain)	**78.83**	62.75
Improved Model (joint)	78.70	**62.85**
Ours	78.58	62.60

Table 5: Comparison of the models in the test subset of Fashion550k when we are restricted to use all 3000 images with cleaned labels and noisy labels only from Fashion144k for the training. (noisy:clean $\simeq$ 39 : 1)

Model	AP_{all}	mAP
Baseline (Fashion144k)	62.23	49.66
Fine-tuning with clean labels	74.90	**57.79**
Improved Model (pretrain)	**74.99**	57.36
Improved Model (joint)	74.45	56.53
Ours	74.79	57.28

We perform another analysis, where we only have access to Fashion144k for noisy labels in Table 5. Here, AP_{all} and mAP get worse compared to fine-tuning with clean labels in almost all proposed methods. This result suggests that the label cleaning network-based approach does not work if we have relatively enough cleaned annotations.

6. Conclusion

We have presented a new large-scale weakly-labelled dataset for multi-label classification of garments of full pose images, providing a small subset of validated annotations for evaluation. We provide experimental evaluation of the effect of the number of images and number of clean annotations, along with various variants of multi-task classification networks that make use of the clean and noisy annotations. Results show that a few simple modifications are able to improve the performance of previous approaches significantly, and larger amounts of noisy data are useful for improving classification results.

Acknowledgements: This research was supported by GCL program of The Univ. of Tokyo by JSPS. N. Inoue and T. Yamasaki are partially supported by the Grants-in-Aid for Scientific Research (no. 26700008) from JSPS, JST-CREST (JPMJCR1686), and Microsoft IJARC core13. E.

Simo-Serra and H. Ishikawa are partially supported by JST-CREST Grant Number JPMJCR14D1, JST ACT-I Grant Number JPMJPR16UD,

References

[1] PyTorch. http://pytorch.org/. 5

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

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

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

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

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[7] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

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

[9] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 4

[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] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[12] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan. Semantic object parsing with local-global long short-term memory. In *CVPR*, 2016. 2

[13] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, 2015. 2

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

[15] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang. Fashion landmark detection in the wild. In *ECCV*, 2016. 2

[16] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015. 2

[17] I. Misra, C. Lawrence Zitnick, M. Mitchell, and R. Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *CVPR*, 2016. 2

[18] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari. Learning with noisy labels. In *NIPS*, 2013. 2

[19] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. A high performance crf model for clothes parsing. In *ACCV*, 2014. 2

[20] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *CVPR*, 2015. 1, 2, 5

[21] 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, 3

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[23] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014. 2

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 5

[25] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie. Learning from noisy large-scale datasets with minimal supervision. *arXiv preprint arXiv:1701.01619*, 2017. 1, 2, 3, 4, 5

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

[27] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015. 2

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

[29] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Retrieving similar styles to parse clothing. *PAMI*, 37(5):1028–1040, 2015. 2

[30] 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

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

[32] D. Yoo, N. Kim, S. Park, A. S. Paek, and I. S. Kweon. Pixel-level domain transfer. In *ECCV*, 2016. 2

[33] B. Zhao, J. Feng, X. Wu, and S. Yan. Memory-augmented attribute manipulation networks for interactive fashion search. CVPR, 2017. 2