DESCRIBING CLOTHING IN HUMAN IMAGES: A PARSING-POSE INTEGRATED APPROACH

Yanghong Zhou[1,2], Runze Li[1,2], Yangping Zhou[1,2] and Pik-Yin Mok[1,2,*]
[1] *The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China*
[2] *Institute of Textiles & Clothing, The Hong Kong Polytechnic University, Hunghom, Hong Kong*

ABSTRACT

With the advent of information technology, digital product information grows exponentially. People are exposed to far too much information, and information overload can slow down, instead of speeding up, a simple decision-making process like searching for suitable clothing online. Traditional semantic-based product retrieval may not be effective due to human subjectivity and cognitive differences. In this paper, we propose a method by integrating the state-of-the-arts deep neural models in pose estimation, human parsing and category classification to recognise from human images all clothing items and their fine-grained product category information. The proposed fine-grained clothing classification model can facilitate a wide range of applications such as the automatic annotation of clothing images. The effectiveness of the proposed method is validated through experiment on a real-world dataset.

KEYWORDS

Clothing Retrieval, Clothing Recognition, Fine-Grained Classification, Pose Estimation, Human Parsing, Deep Learning

1. INTRODUCTION

With the rapid growth of digital information online, people can easily share almost everything of their daily life with friends and peers via social networks, but at the same time people also face a new challenge of mass information retrieval. Clothing product retrieval is an attractive research area to both academia and the fashion industry because it improves the users' experience and promotes sales of online shopping. Typically, users search for products at online stores by inputting keywords such as the attributes and styles of clothing. Semantic-based product retrieval is efficient, and the method is widely used by online shopping sites (Egozi et al., 2011). A same category of goods, however, often can be expressed by different text vocabularies, and each clothing attribute is assigned a semantic description according to the type of clothing attributes (such as type, brand, etc.). Because of human subjectivity and cognitive differences, clothing attributes do not have standard definitions, and the accuracy of textual and semantic-based product recommendations is not high.

Another group of researchers proposed to recognise clothing attributes from images. Clothing items are soft and do not have a fixed form, however; thus, clothing classification is a challenging computer-vision problem due to possible deformation and occlusion, and always requires correct localisation of clothing regions in the input images. There are mainly three classes of method to estimate clothing location: pose estimation (Chen et al., 2012), clothing parsing (Yamaguchi et al., 2012; 2013; Arbelaez et al., 2011; Simo-Serra et al., 2014) and clothing detection (Bossard et al., 2012; Eichner et al., 2012). Chen et al. (2012) used the pose estimation method, extracted image features from different human body parts and trained a classification for each attribute using the extracted images features. Considering the natural relationship between clothing attributes, a conditional random field (CRF) is employed on the classification predictions to produce more reasonable attribute lists. Pose provides only the location of human joints, whereas the detail locations of the clothing region remain missing.

Some researchers proposed to take advantage of the clothing segmentation to improve the clothing prediction and treat the clothing classification problem as an image parsing problem, which predicts the clothing type on a pixel level. To do this, Yamaguchi et al. (2012) built a clothing parsing Fashionista

[*] Corresponding author: P.Y. Mok, tracy.mok@polyu.edu.hk

dataset, consists of 685 photos which label the clothing or body clothing on a pixel level. They segmented the image to super-pixels using an image segmentation algorithm (Arbelaez et al., 2011) and exploited a CRF model to predict the clothing labels for each segmented super-pixel. To improve the parsing performance, they used a clothing retrieval approach to predict the clothing items, namely tag prediction, and then combined the tag prediction into the parsing model (Yamaguchi et al., 2013). Considering the complex dependencies between clothing and human pose, Simo-Serra et al. (2014) proposed a pose-aware CRF model for clothing parsing.

In addition, some researchers focused on clothing detection and applied the detection results to predict clothing classification. For example, Bossard et al. (2012) used the Calvin upper-body detector (Bossard, et al., 2012) to estimate the location of upper-body clothing and then trained the classification model based on the detected clothing. With the development of deep learning technology, deep learning technique then was applied for the fast and accurate fashion item detection. Hara et al. (2016) used RCNN (Girshick et al., 2014) to generate object proposals and extract features from each object proposal. Smirnov et al. based on fast-RCNN (Girshick, 2015) and switch selective search to MultiBox (Szegedy et al., 2014) to further improve the detection accuracy.

In this paper, taking advantages of pose estimation and clothing parsing, we design a framework that detects all the clothing items presented in an input human image and classify the clothing type and fine-grained category of each identified clothing item. To distinguish 'dress' and 'upper-clothing and skirt', we combined an image-type classification model with human parsing. The framework is shown in Figure 1. The main contributions of this paper include:

- A novel dataset consisting of 79,153 images annotated with a structure clothing category.
- A new framework for accurate clothing classification combining image type classification, human parsing and pose estimation.
- A coarse attention-based neural network to predict the fine-grained clothing category.

Our fine-grained clothing classification can facilitate a wide range of applications, such as the automatic annotation of clothing images, similar clothing product retrieval, fashion coordination rule mining from online images and fashion trend analysis.

2. METHOD

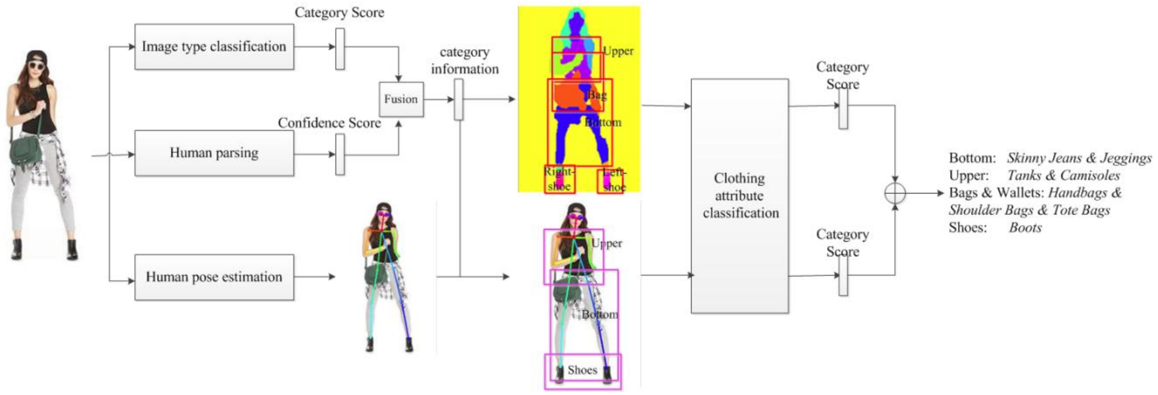

Figure 1. Framework of Fine-Grained Clothing Recognition

Figure 1 shows our framework to recognise the fine-grained category of clothing that the human is wearing. We collected about 10,000 full-body fashion images. As there is no dataset that can recognise all the clothing items and their fine-grained clothing category in the full-body image, we proposed to detect the clothing items in the image first and train fine-grained clothing category classification models to analyse the fine-grained category of these detected clothing items. To do so, we built a dataset labelling with fine-grained clothing categories to train the clothing category classification models for various clothing items. Finally, the detailed category of detected clothing items will be given.

2.1 Clothing Item Detection

In our method, both human parsing and pose estimation were used to detect the clothing region, and image category information was injected to refine the detection result. The clothing type classification then was used to complement the global feature of the image.

We employed Fully Convolutional Networks (FCN) method (Long et al., 2015) and used the ATR dataset to train a human parsing model (as shown in Figure 2). The model can segment the human body into 18 different classes (background, hat, hair, sunglasses, upper-clothing, skirt, pants, dress, belt, left shoe, right shoe, face, left leg, right leg, left arm, right arm, bag, and scarf). If z denotes the confidence score of the last layer of the network, the output of z provides a probability for each label through the softmax function. Given image *I*, we denote $z^{I}_{k,i,j}$ the probability of pixel $(i,$j) belongs to the k-th label and $k^* = argmax\, z^{I}_{k,i,j}$ will be the predicted label of pixel (i,j). We denote the number of pixels having the predicted label k^* as $N(k^*)$ and compute $z^{I}_{k^*}$=$\sum_{k^*} z^{I}_{k^*,i,j}/N(k^*)$, and regard the mean probability $z^{I}_{k^*}$ as the probability that the image have the k^* label.

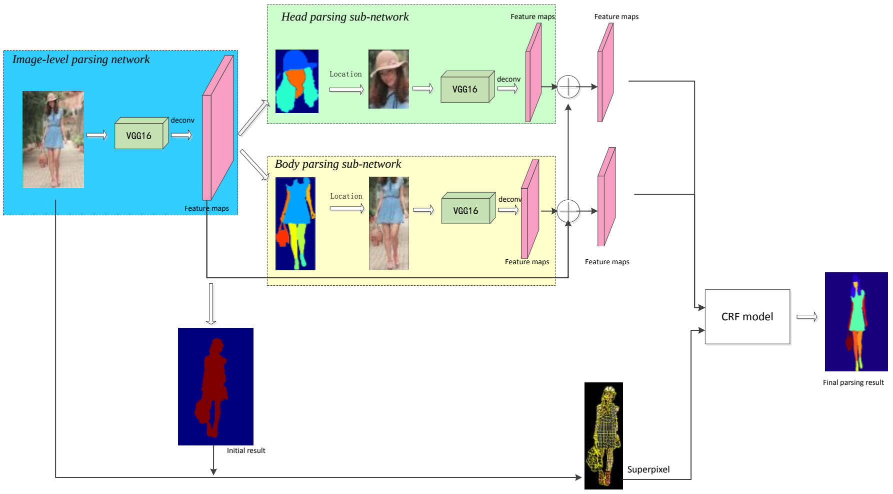

Figure 2. Human Parsing Model

Since human parsing model provide pixel-to-pixel annotation with little global context information, we proposed to train an image-classification model to predict the image category – such as upper clothing, bottom clothing, dress, bag, shoes, sunglasses and overalls – to provide global context information.

Given image *I*, we denote the classifier score as C^{I}_{k} and $k^* =$ argmax C^{I}_{k} as the predicted label of image *I*.

Firstly, we put the image *I* into the image classification model which is a Resnet-34 classification model and determine its category. If the predicted label is ‘dresses’ or ‘overalls’, there will be no skirts in the image and the probability of having upper clothing in the image also is low.

To crop out the corresponding clothing region, we proposed an algorithm which exploits the human parsing model and the image classification model. The image classification model gives the image category (such as dress, upper clothing, bottom clothing (skirt, pants), hat, belt, bag and scarf) and the human parsing model gives the pixels having these labels. The proposed algorithm is shown as follows:

```
Input Image $I$
Extract the probability $z^{I}_{k,i,j}$ of pixel $(i,$j) belonging to the $k$-th label for all the pixels via human parsing
Extract the classifier score $C^{I}_{k}$ for the image
If $k^* =$ argmax $C^{I}_{k}$ is “dresses” or “overalls” do
    $k^* = argmax\, z^{I}_{k,i,j}$
  For $k^* = argmax\, z^{I}_{k,i,j}$ is skirt do
     $k^*$ = “dresses” / “overalls”
  If $k^*$ = “upper clothing” and $z^{I}_{k^*}$ < 0.8:
     Eliminate $k^*$
  Else do
  If $k^*$ = “dress” and $z^{I}_{k^*}$ < 0.8:
     Eliminate $k^*$
Output: $k^*$
```

Moreover, the accuracy of pose estimation model has substantial improvements recently, we used the pose estimator proposed by (Cao et al., 2017) to compute the 2D locations of human joints. The pose estimator has satisfactory performance across the dataset. We proposed an algorithm, which exploits the location of human joints to define the locations of upper clothing, bottom clothing, glasses, dress and shoes as follows:

```
Input the location of joints using pose estimator
Neck (x_neck, y_neck), Nose (x_nose, y_nose), Right-ear (x_rear, y_rear), Left-ear (x_lear, y_lear),
    Right-eye (x_reye, y_reye), Left-ear (x_leye, y_leye), Right-shoulder (x_rshoulder, y_rshoulder),
    Left-shoulder (x_lshoulder, y_lshoulder), Right-hand (x_rhand, y_rhand), Left-hand (x_lhand, y_lhand),
    Right-hip (x_rhip, y_rhip), Left-hip (x_lhip, y_lhip), Right-knee (x_rknee, y_yknee), Left-knee (x_lknee,
    y_lknee), Right-shoe (x_rshoe, y_rshoe), Left-shoe (x_lshoe, y_lshoe)
Glasses = [[max(0, y_reye - (y_neck - y_nose) / 6), y_reye + (y_neck - y_nose) / 6, x_rear, x_lear]]
Upper-clothing = [max(y_rshoulder - (y_rsholder - y_neck) /4, 0), max(y_lhand, y_rhand, y_rhip),
    min(x_rhand, x_rshoulder),max(x_lhand, x_lshoulder + (y_lshoulder - y_neck) / 3)]
Bottom-clothing = [max(y_rhip - (y_rsholder - y_ neck) / 3, 0), max(y_lshoe, y_rshoe, y_rhip), min(x_rhip -
    (y_rsholder - y_ neck) / 2, x_rshoe - (y_rsholder - y_ neck) / 2), max(x_lshoe, x_lhip + (y_lshoulder -
    y_ neck) / 2)]
Dress = [max(y_rshoulder - (y_rsholder - y_ neck) / 3, 0), max(y_lshoe, y_rshoe, y_rhip), min(x_rshoulder -
    int((y_neck - y_nose)) / 2, x_rshoe - (y_rsholder - y_ neck) / 2), max(x_lshoes, x_lshoulder +
    (y_lshoulder - y_ neck) / 2)]
Shoe = [y_rknee, max(y_rshoe + (y_rshoe - y_ rknee)/3, y_lshoe + (y_rshoe - y_ rknee)/3), min(x_rshoe -
    (y_rshoe - y_ rnkee)/ 3, x_lshoe -(y_rshoe - y_ rnkee)/ 3), max(x_lshoe +(y_rshoe - y_ rnkee)/3, x_rshoe
    + (y_rshoe - y_ rnkee)/ 3 )]
Output: glasses, upper-clothing, bottom-clothing, dress, shoe
```

Therefore, we obtain the clothing detection using human parsing and pose estimation, respectively.

2.2 Fine-grained Clothing Category Classification

Our goal is to recognise the correct category information and fine-grained attributes of all the clothing items presented on the input image. After detecting the clothing regions, we can build a big fine-grained clothing dataset and train Convolution Neural Netwroks (CNNs) to predict category and fine-grained attributes for given clothing regions.

2.2.1 Datasets

We built the dataset by crawling images from several online shopping websites. As the text description of clothing on different websites is quite different, we defined the clothing category and attribute and relabelled these images. Each image was annotated with a fine class and the corresponding coarse label. Each fine class belongs to exactly one coarse class. Table 1 shows all the clothing classes defined in this work. There are 15 coarse labels and 111 fine clothing labels. The dataset contains 79,153 images, which include 19,520 full-body clothing images, 23,579 upper clothing images and 8,414 bottom clothing images. The number of images for each fine-grained clothing category is shown in Figure 3.

Table 1. A list of Clothing Categories Considered in this Work

Image type	Clothing Category	Subclasses
Full-body clothing	Dresses	‘Layered Dresses’ (1), ‘A-line Dresses’ (2), ‘Mini Dresses’ (3), ‘Midi Dresses’ (4), ‘Maxi Dresses’ (5), ‘Evening Dresses’ (6), ‘Bodycon Dresses’ (7), ‘T-shirt Dresses’ (8), ‘Wrap Dresses’ (9), ‘Slit Dresses’ (10), ‘Shirt Dresses’ (11), ‘Oversized Dreses’ (12), ‘Mid-length Dresses’ (13)
	Overalls	‘Dungarees’ (14), ‘Jumpsuits’ (15), ‘Playsuits’ (16), ‘Bodysuits’ (17)
Upper clothing	Tops	‘Cropped Tops’ (1), ‘Shirts’ (2), ‘Short-sleeved Blouses’ (3), ‘Long-sleeved Blouses’ (4), ‘Sleeveless/Short-sleeved Shirts’ (5)
	Tanks & Camisoles	‘Tanks’ (6), ‘Singlets’ (7), ‘Camisoles’ (8)
	T-Shirts	‘Short-Sleeved T-shirts’ (9), ‘Long-sleeved T-shirts’ (10), ‘Polo T-shirts’ (11)
	Shirts	‘Formal Shirts’ (12), ‘Short-sleeved Shirts’ (13), ‘Denim Shirts’ (14), ‘Polo Shirts’ (15)

	Outerwear	‘Bombers & Blousons’ (16), ‘Trench Coats & Pea Coats’ (17), ‘Jackets’ (18), ‘Sweaters & Jumpers’ (19), ‘Hoodies & Sweatshirts’ (20), ‘Cardigans’ (21), ‘Denim Jackets’ (22), ‘Coats’ (23), ‘Parkas’ (24), ‘Gilets’ (25), ‘Blazers & Suits’ (26), ‘Waistcoats & Vests’ (27), ‘Down Jackets’ (28), ‘Capes & Ponchos’ (29), ‘Leathered Jackets & Coats’ (30), ‘Overcoats’ (31), ‘Knitted’ (32)
Bottom Clothing	Shorts	‘Denim Shorts’ (1), ‘Casual Shorts’ (2), ‘Running Shorts’ (3), ‘Tailored Shorts’ (4), ‘Cargo Shorts’ (5), ‘Beach Shorts’ (6), ‘Culottes’ (7), ‘Pyjama Shorts’ (8)
	Pants & Trousers	‘Sweat Pants’ (9), ‘Slim-Fit Pants’ (10), ‘Chino Pants’ (11), ‘Cropped Pants’ (12), ‘Skinny Pants’ (13), ‘Wide-Leg Trousers’ (14), ‘Leggings’ (15), ‘Flared Trousers’ (16), ‘Peg Trousers’ (17), ‘Straight-Leg Pants’ (18)
	Skirts	‘Slit skirts’ (19), ‘Pencil skirts’ (20), ‘Mini Skirts’ (21), ‘Maxi Skirts’ (22), ‘Wrap skirts’ (23), ‘Layered skirts’ (24), ‘Pleated skirts’ (25), ‘Midi Skirts’ (26), ‘Panel skirts’ (27), ‘A-line Skirts’ (28)
	Jeans	‘Slim-Fit Jeans’ (29), ‘Boyfriend Jeans & Causal Jeans’ (30), ‘Skinny Jeans & Jeggings’ (31), ‘Cropped Jeans’ (32), ‘Trumpet Jeans’ (33)
Others	Bags & Wallets	Wallets & Card Holders, ‘Messengers & Laptop Bags’, ‘Clutches’, ‘Purses & Small Item Bags’, ‘Backpacks’, ‘Duffle Bags’, ‘Hangbags & Shoulder Bags & Tote Bags’, ‘Waist/Chest Bags’, ‘Bag Accessories’, ‘Satchels & Sling Bags’, ‘Luaggages & Storage Bags’
	Shoes	Slip-ons’, ‘Sandals’, ‘Flip-flops & Slippers’, ‘Boots’, ‘Brogues’, ‘Boat Shoes & Loafers’, ‘Trainers & Sneakers’, ‘Heels & Platforms & Wedges’, ‘Flats’
	Belts	‘Leather Belts’, ‘Non-leather Belts’
	Sunglasses & Glasses	Squared Sunglasses’, ‘Round Sunglasses’, ‘Retro Sunglasses’, ‘Aviator Sunglasses’, ‘Oversized Sunglasses’, ‘Wayfarer’, ‘Cat Eye Sunglasses’

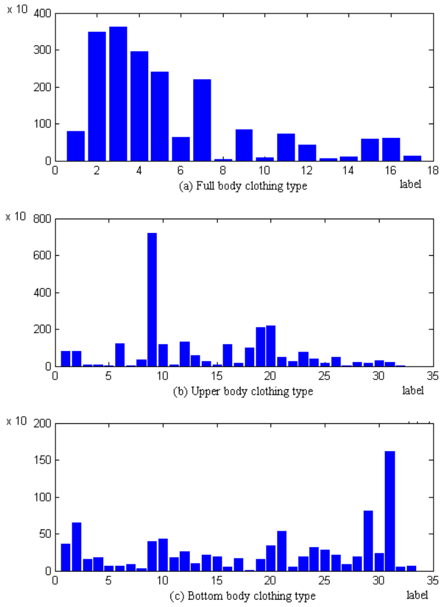

Figure 3. Number of images for each fine-grained category

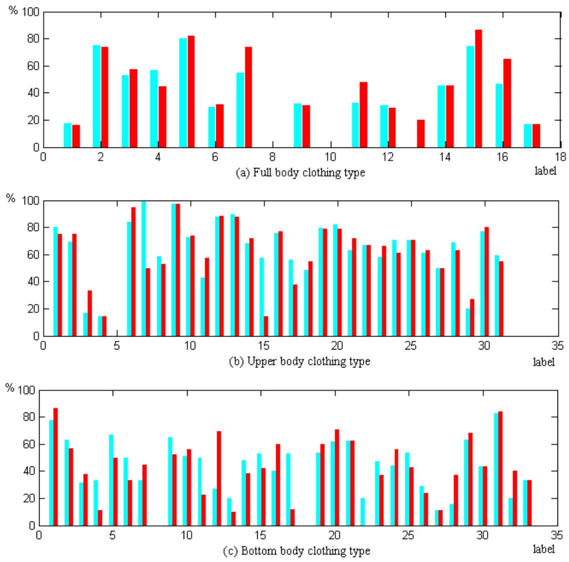

Figure 4. Comparison of per class accuracy between basic Resnet (blue) and our method (red)

2.2.2 Coarse attention-based neural network

Considering the hierarchical relationship between coarse and fine labels, we proposed a new model which jointly learns the coarse and the fine-grained classifiers in a cascaded manner. The proposed coarse attention-based network architecture is depicted in Figure 5 (a). Given image I, we use a feature extractor to obtain the low-level feature

$$x_l = \text{Extractor}(I, W_{\text{extractor}}) \tag{1}$$

where $W_{\text{extractor}}$ represents the feature extractor’s parameters. Secondly, based on the low-level feature x_l, we design the coarse and fine-grained classifiers, respectively. We formulate the coarse classification process as follows:

$$\hat{y}_{\text{coarse}} = \text{Predictor}_{\text{coarse}}(x_l, W_{\text{coarse}}) \tag{2}$$

where W_{coarse} indicates the parameters of the coarse classifier and $\hat{y}_{\text{coarse}}$ indicates the category score of coarse category. To take advantage of the hierarchical relationship between the coarse and fine-grained labels, we also designed a coarse information injection module, which enabled the coarse information providing prior knowledge for the fine-grained category classification. Figure 5 shows the overall architecture of the coarse attention-based neural network.

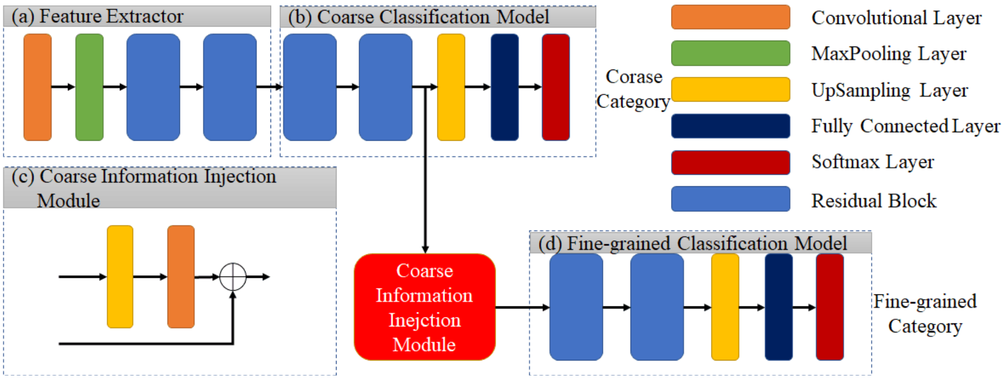

Figure 5. Coarse Attention-based Neural Network

a) Coarse information injection module

We realized the coarse information injection from low-level feature to the fine-grained classifier by designing a coarse information injection module (Figure 5(b)). As the high-level features have semantic information, the coarse information injection module injects the high-level features from the coarse classifier into the low-level features to provide prior knowledge to the fine-grained classifier. Let x_s be the high-level features with semantic meaning and x_l be the low-level features; we pass x_s and x_l through the coarse information injection module (Figure 5(c)) to obtain the features with semantic information as:

$$x_{\text{fusion}} = \text{Injector}(x_l, x_s) \tag{3}$$

The injection module merged the low-level features and the high-level features with semantic meaning using an element-wise addition operation. To match the spatial size of x_l, we performed a bi-linear up-sampling on the high-level features and then added a 1×1 convolutional layer on the up-sampled features. As such, the output of coarse information injection module represents the fused feature maps characterized by high-level semantic knowledge embedding as the learning guidance of the coarse classifier. As shown in Figure 6, in shirt images, semantic prior knowledge of coarse predictor highlights the shirt part features.

Finally, we put the fused feature maps into the fine-grained classifier (Figure 5(d)). and obtain the fine-grained category score as follows:

$$\hat{y}_{\text{fine}} = \text{Predictor}_{\text{fine}}(x_{\text{fusion}}, W_{\text{fine}}) \tag{4}$$

where W_{fine} indicates the parameters of fine label predictor.

b) Loss function

For the model training, we adopted cross-entropy loss for both the coarse and fine-grained classifiers. We deployed a separate loss function at the output of each classifier. The cross-entropy loss is defined as:

$$L(y, \hat{y}, m) = -\sum_{c=1}^{m} y_c \log(\hat{y}_c). \tag{5}$$

where m is number of labels. The loss function for our proposed model is formulated as:

$$L = L(y_{\text{coarse}}, \hat{y}_{\text{coarse}}, m_{\text{coarse}}) + L(y_{\text{fine}}, \hat{y}_{\text{fine}}, m_{\text{fine}}) \tag{6}$$

where m_{coarse} and m_{fine} denote the number of coarse labels and the number of fine-grained labels, respectively.

3. EXPERIMENT

In this section, we firstly evaluate the performance of proposed coarse attention-based neural network on our fine-grained clothing dataset. For comparison purposes, the broadly applied state-of-the-art network Resnet (He et al., 2016) is used as the baseline method. As our goal is to recognise all the fine-grained clothing categories on the full-body images, we used the proposed network to train specific models on full-body clothing images, upper clothing images and bottom clothing images (shown in Table 1). We selected full-body images and detected all the clothing regions using the method in Section 2.2. We then input the detected clothing regions into the corresponding models and obtained the fine-grained clothing category of detected clothing regions.

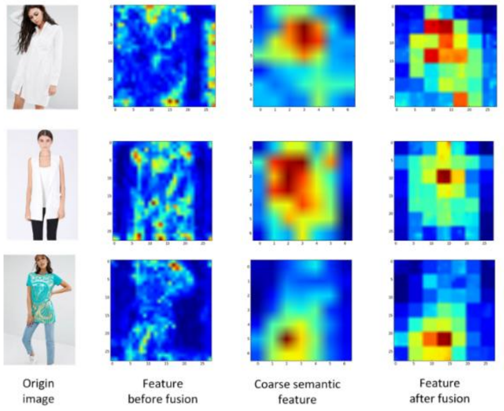

Figure 6. Some Examples of Feature Maps Before Fusion and After Fusion

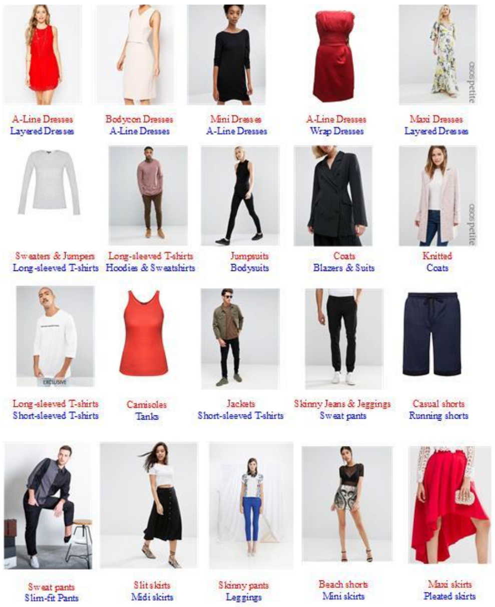

Figure 7. Some Failure Examples, the Classification Result of our Model (Red) And Ground-Truth (Blue)

3.1 Evaluation of Fine-grained Classification

To evaluate the performance of our method, we trained three models on full-body clothing images, upper clothing images and bottom clothing images. All the datasets were split into three parts: 70% for training, 10% for validation, and 20% for testing. The training implementation set was the same. Input images were resized into 256×256 images. The network was trained with a mini-batch stochastic gradient descent with a momentum of 0.9, and weight decay of 0.0005; the learning rate was set to fixed 0.001. The final model was trained on pytorch for 100 epochs on GPU gtx 1080Ti.

Table 2. Accuracy Comparison of Basic Resnet and Our Model on Full-Body Clothing Type Images, Upper-Body Clothing Type, and Bottom-Body Clothing Type Images

Image type	Accuracy of basic Resnet	Accuracy of our model
Full-body clothing type	56.28%	58.88%
Upper-body clothing type	80.22%	81.12%
Bottom-body clothing type	56.95%	58.88%

We measured performance based on the top-1 accuracy. Table 2 shows the top-1 accuracy of our model and the basic Resnet on image type classification. Compared to basic Resnet-34, our model performed better on these three image types. In particularly, the accuracy of our model on full-body clothing type increased by more than 2%.

We also compared the per-class accuracy of our method with that of basic Resnet. Figure 4 illustrates that our method outperformed basic Resnet on most of categories. The class accuracy of basic Resnet exceeded our method because the number of images was very limited most of the time. To illustrate, the basic Resnet accuracy for Polo Shirts (15) was greater by about 43% and for Trench Coats & Pea Coats (17) by about 21%, but the number of images were 70 and 160, respectively, and the number of images for validation were only 7 and 16. As the number of images were so limited, their effect can be ignored. By filtering out classes with fewer than 1,000 images, there are 4 classes in which the accuracy of our method outperformed basic Resnet in a total of 6 classes for full-body clothing type images, and 6 classes in which the accuracy of our method outperforms basic Resnet in a total of 7 classes for upper-body clothing type images. Therefore, when the number of images is large enough, our method performs better than the basic Resnet.

3.2 Fine-grained Clothing Category Recognition on Full-Body Images

To recognise the fine-grained clothing category of all presented clothing regions in an input human image, we trained a human parsing model on the ART dataset (Liang et al., 2015) to parse the human image, and we estimated the location of human joints using OpenPose (Cao et al., 2017). Based on the human parsing result and estimated human pose, we detected all the clothing regions and input these regions into the corresponding classifiers. For example, the detected upper clothing region was input into the upper clothing classifier and bottom clothing region into the bottom clothing classifier. Figure 8 shows some clothing recognition of our methods.

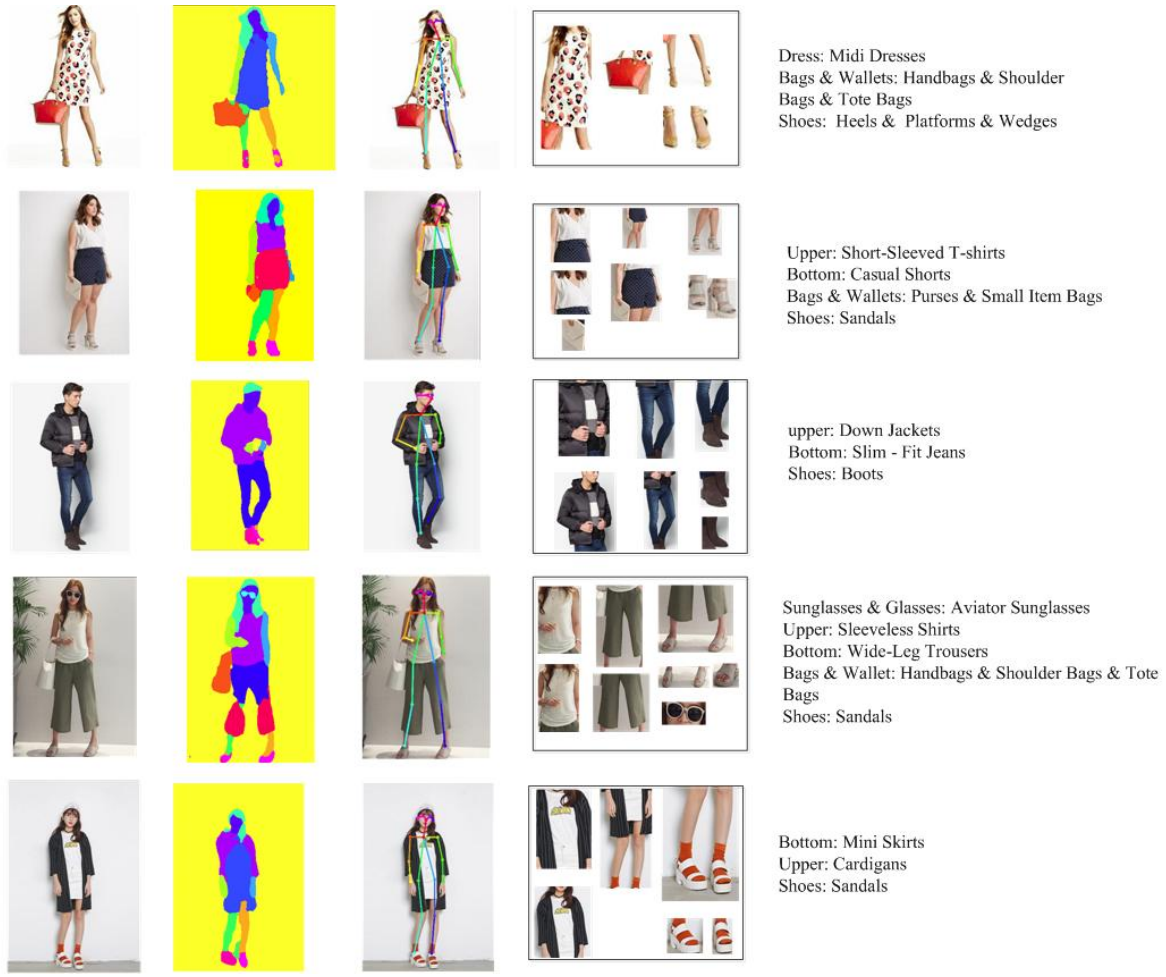
Dress: Midi Dresses
Bags & Wallets: Handbags & Shoulder Bags & Tote Bags
Shoes: Heels & Platforms & Wedges
Upper: Short-Sleeved T-shirts
Bottom: Casual Shorts
Bags & Wallets: Purses & Small Item Bags
Shoes: Sandals
upper: Down Jackets
Bottom: Slim - Fit Jeans
Shoes: Boots
Sunglasses & Glasses: Aviator Sunglasses
Upper: Sleeveless Shirts
Bottom: Wide-Leg Trousers
Bags & Wallet: Handbags & Shoulder Bags & Tote Bags
Shoes: Sandals
Bottom: Mini Skirts
Upper: Cardigans
Shoes: Sandals

Figure 8. Examples of Clothing Recognition on Full-Body Images

Figure 7 also shows some failure examples. Some failures were caused by vague clothing definitions. To illustrate, some Short-sleeved T-shirts were mistaken as Long-sleeved T-shirts. Actually, these T-shirts were not exactly Short-sleeved T-shirts and seem to be the Middle-sleeved T-shirts, which are not defined in our clothing categories. Also, when a clothing image has multiple category features, our model can recognise only one and fail to recognise another correct category. For example, some Long-sleeved T-shirts were recognised as Sweaters & Jumpers or Hoodies & Sweatshirt, and Layered Dresses were recognised as Maxi Dresses. Similarly, when one person wears multiple pieces of clothing, like a T-shirt and a jacket, our model can recognise only one article. In addition, when the clothing features are very similar, our model may have difficultly recognising the correct one, such as Tanks and Camisoles, skinny pants and leggings, and casual shorts and running shorts.

4. CONCLUSION

In this paper, we have studied the problem of reinforced clothing category and fine-grained attributes recognition. We developed a new method that integrates human parsing, pose estimation and category classification to segment all clothing regions on input image and carry out fine-grained category and attributes recognition for each clothing region. Using this approach, we can obtain the fashion coordination information of all clothing items on input images.

ACKNOWLEDGEMENT

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 5218/13E). The supports from Guangdong Provincial Department of Science and Technology (Project No. R2015A030401014) Shenzhen Science and Technology Innovation Commission (Project No. JCYJ20170303160155330), are gratefully acknowledged.

REFERENCES

Arbelaez P., M. Maire, C. Fowlkes, and J. Malik, 2011. Contour detection and hierarchical image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 33, pp. 898-916.

Bossard L., M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool, 2012. Apparel classification with style, in *2012 Asian Conference on Computer Vision*, pp. 321-335.

Cao Z., T. Simon, S.-E. Wei, and Y. Sheikh, 2017. Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR),* p. 7.

Chen H., A. Gallagher, and B. Girod, 2012. Describing clothing by semantic attributes, in *European Conference on Computer Vision (ECCV),* pp. 609-623.

Egozi O., S. Markovitch, and E. Gabrilovich, 2011. Concept-based information retrieval using explicit semantic analysis, *ACM Transactions on Information Systems (TOIS),* vol. 29, p. 8.

Eichner M., M. Marin-Jimenez, A. Zisserman, and V. Ferrari, 2012. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images, *International Journal of Computer Vision,* vol. 99, pp. 190-214.

Girshick R., J. Donahue, T. Darrell, and J. Malik, 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR),* pp. 580-587.

Girshick R., 2015. Fast r-cnn, *arXiv preprint arXiv:1504.08083.*

Hara K., V. Jagadeesh, and R. Piramuthu, 2016. Fashion apparel detection: the role of deep convolutional neural network and pose-dependent priors, in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV),* pp. 1-9.

He K., X. Zhang, S. Ren, and J. Sun, 2016. Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778.

Liang X., S. Liu, X. Shen, J. Yang, L. Liu, J. Dong*, L. Lin and S. Yan*, 2015. Deep human parsing with active template regression, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 37, pp. 2402-2414.

Long J., E. Shelhamer, and T. Darrell, 2015. Fully convolutional networks for semantic segmentation, in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3431-3440.

Simo-Serra E., S. Fidler, F. Moreno-Noguer, and R. Urtasun, 2014. A high performance CRF model for clothes parsing, in *2014 Asian Conference on Computer Vision,* pp. 64-81.

Szegedy C., S. Reed, D. Erhan, D. Anguelov, and S. Ioffe, 2014. Scalable, high-quality object detection, *arXiv preprint arXiv:1412.1441.*

Yamaguchi K., M. H. Kiapour, L. E. Ortiz, and T. L. Berg, 2012. Parsing clothing in fashion photographs, in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3570-3577.

Yamaguchi K., M. H. Kiapour, and T. L. Berg, 2013. Paper doll parsing: Retrieving similar styles to parse clothing items, in *2013 IEEE International Conference on Computer Vision (ICCV),* pp. 3519-3526.