ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Issue 10 , October 2019

Apparel key-points localization by Mask R-CNN and attribute recognition

Sharofiddin Allaberdiev， Richard Odol, Haiyang Jiang

P.G. Students, department of Mathematics and Computer Science, Wuhan Textile University, Hubei, China

ABSTRACT: Recent advances have been driven in fashion clothes recognition by rich clothes datasets and high-quality annotations. Nevertheless, approaches that are being applied treat clothes with spatial relations, symmetry, proportions and key characteristics of clothes as common images, ignoring the prior clothing knowledge. We propose New-key-detection, a model using the prior symmetric constraint to refine the key-points located by any backbone detection networks to combine the semantic information with the advantages of deep learning. We've introduced a new loss to utilize all available data which contain "maybe" labels in order to engage with uncertainty in labeling clothing. About 2.53% Normalized Error in FashionAI dataset and 3.2% AP in human key-points dataset coco2017 has been reduced and improved by New-key-detection when compared to the Mask R-CNN baseline. The most experimental results showed that the proposed approach achieved better results in different recognition datasets (resp., FashionAI, and Deepfashion) with about (resp., 2.57% mAP and 10% recall) improvements.

KEY WORDS: Feature extraction, Key-points Localization, Attribute Recognition, Mask R-CNN.

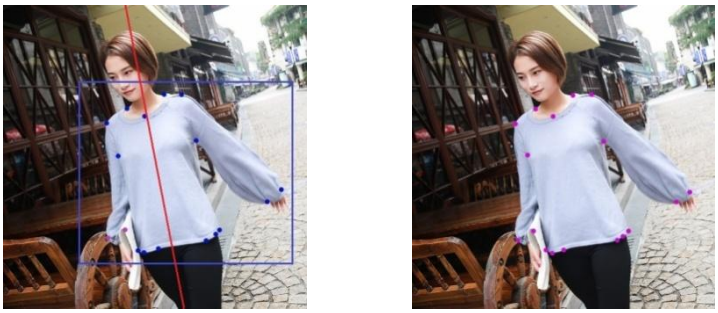
I.INTRODUCTION

Popularity and growth have been recently observed in the E-commerce of apparel in the online market. The critical for e-commerce plat-form is considered as automatic and certain garments recommendation in order to grow up its sales and profits. That recommendation has depended heavily on high-performance apparel recognition and detection.

(a) Origin image.

(b) Mask R-CNN.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 6, Issue 10 , October 2019



(c) With axial symmetric constraint. (d) With New-key-detection.

Fig. 1. The advantages of the New-key-detection algorithm. Mask R-CNN misses two cuffs and one invisible point in Fig.1(b). The new-key-detection algorithm performs best in Fig.1(d).

Nevertheless, while we were developing an effective algorithm for key-points of localization and clothing recommendation, there appeared two fundamental challenges. Firstly, because of the fact that garments possess deformation of dresses, viewpoint variation flexibility as well as style diversity, clothes frequently gets some variations on images which are appearance changes. These variations cause extreme difficulties in terms of attribute recognition and category classification. Secondly, occlusions are often observed owing to the various postures of cloth models, which ruins the represent-ability of neural networks, as shown in Fig.1 and Fig.3 (w/o occlusions). How to localize the key-points concealed in the occlusions are still challenging.

In order to recognize apparel attributes, the most used approach naming Convolutional neural networks have been applied [1, 2, 3, 4, 5, 6, 7]. In spite of being several methods, those approaches cannot reach satisfactory performance to predict the attributes and key-points, because this problem was mainly formulated as a normal image classification without using any prior knowledge of clothes.

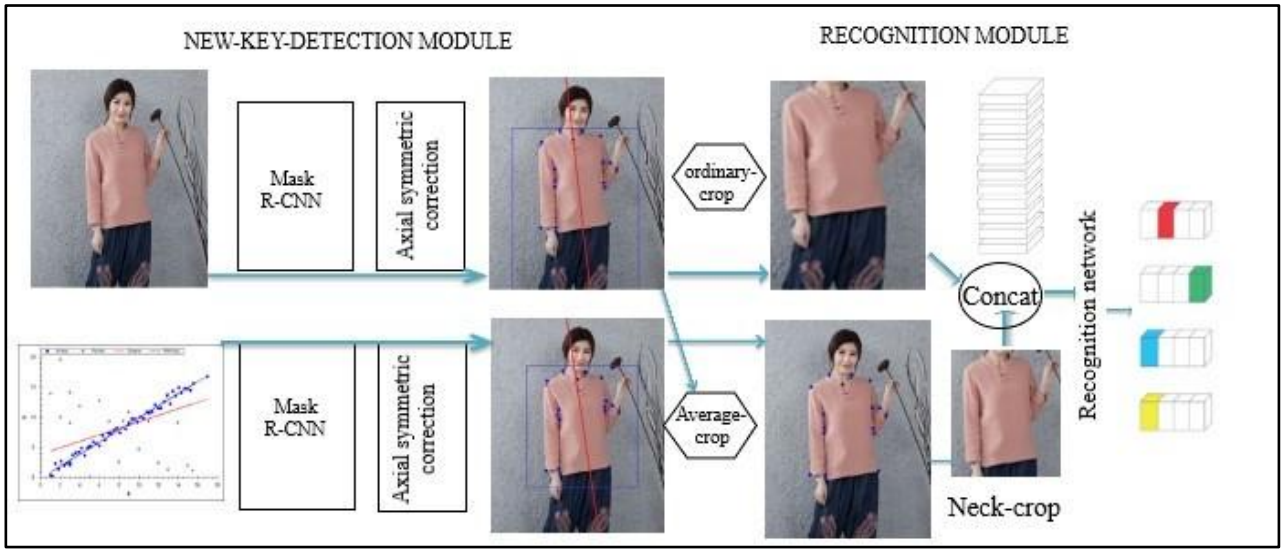


Fig. 2. The full pipeline for our model (best viewed in color). The model for the upper-body design tasks in FashionAI attributes recognition dataset, which contains four different tasks was made up, red as collar design, green as neck design, blue as neckline design and yellow as lapel design (each part is marked in the input image). In this research, we tried to make up the incorporation of garments as a prior knowledge into key-points localization and attribute recognition, according to the aim of this investigation which is to research the challenging problem, locating clothing key-points in difficult circumstances such as occlusions

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 6, Issue 10 , October 2019

and the deformations of appearance. Once an accurate localization of key-points has been derived, it will become much more facile to identify the clothing attributes even in some difficult situations such as occlusions and sleeves folding. The main contributions of this paper are:

- ✓ New-key-detection is purposed, as a new prior knowledge-based key-points localization method. New-key-detection modifies the key-point localizations by using prior knowledge such as symmetric relation and proportion of clothes. It successfully overcomes the complexity of occlusions in key-points localization and the performances on both FashionAI and coco2017.
- ✓ We attempt in a recognition network with a new multi-label loss function defined such as to deal with "maybe" labels. Our model has been shown the best results on DeepFashion, improving about 10% on average for top3 and top5 recall.

II.RELATED WORKS

Clothes Classification. Because of being initially aimed Alexnet, excellent performance has been established in large-scale image recognition tasks on Convolutional neural networks [8]. Residual Networks enriched with some developments to refine recognition performance through a deep cascaded structure and residual propagation cross-layer[9]. Moreover the recognition networks we used in this paper are Inception-tionV4 [10], InceptionResnetV2 [10] and NASNet [11]. The rich prior information required to be further explored to improve the performance of both key-points detection and clothes classification was incorporated into the clothing images.

Object Detection and Key-points Localization. Object detection is estimated as another field in which the deep neural networks have been successfully applied [12]. By adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition, the latest two-stage method Mask R-CNN extends Faster R-CNN [13]. Stacked hourglass [14] and Cascaded Pyramid Network (CPN) [15] were proposed for human key-points localization. But, the performance of these existing methods degenerated substantially once occlusions occur in clothes images. Such a sophisticated problem is the main concern of this investigation. Some related works about using structural features of clothes exist. Considering the target problem, Agns et al. [16] used the structural features of five classes of upper body composition to do image retrieval by minimizing the matching cost. Vittorio et al. [17] detect the spatial layout of different parts of the body and then regards a human as a tree structure with additional two repulsive edges between arms for evaluating the probability of whether the arms are overlapped, finally outputting a coarse pose heatmap. Considering the prior constraints in methodology, [16, and 17] haven't used any symmetry knowledge of the whole body. Our method is different from these works on both target problems and methodology. Furthermore, the key-points localization task in our paper is more challenging.

III. METHODOLOGY

The aimed framework is demonstrated in Fig.2. The full model includes two modules: New-key-detection and recognition network. A preprocessing module outputting key-point localizations, Normal-crop and Neck-crop are New-key-detection. The recognition network obtains cropped images as inputs and predicts final at-tribute values with the multi-label loss and joint learning.

A.New-key-detection: prior knowledge constraint. In this section, we will illustrate our New-key-detection algorithm in detail (showed in Alg.2). The new-key-detection module first detects key-points by using Mask R-CNN. We then modify incorrect or invisible key-points with the axial symmetric constraint on two steps: fitting a symmetric axis (M_{best}) by RANSAC [18] (Alg.1) and improving the points far away from the axis (the points not in the A_{best}). The refinements (6, 7 in Alg.2) of the key-points are based on EMA (exponential moving average).

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Issue 10 , October 2019

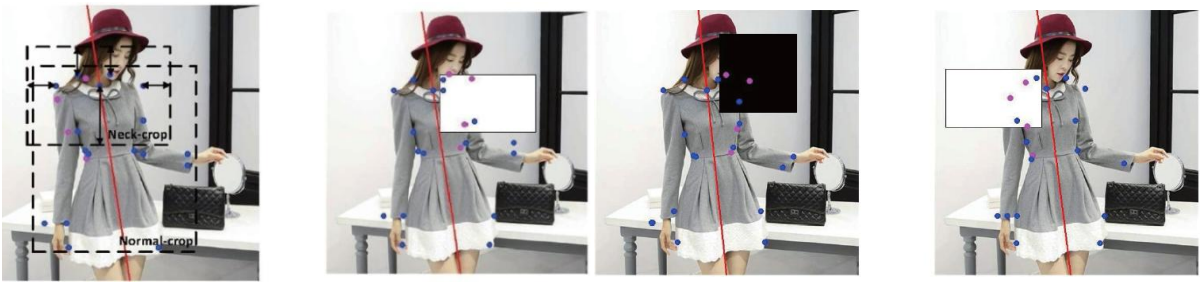


Fig.3. The bigger bounding box in the first figure is the Normal-crop for universal tasks and the smaller one is the Neck-crop, a refined bounding box specialized for upper body design tasks. The remaining figures show the detection results with occlusions.

On the next stage, the minimum and maximum coordinates of key-points and feed a corresponding bounding box to New-key-detection again to obtain the second time refined key-points repeatedly in case of missing parts of clothes like the right cuff in Fig.1(b). Finally, we have derived the average of two scales' predicted key-points as the final output of key-points. We have elucidated the illustration of dealing with different occlusions in Fig.3.

Algorithm 1: RANSAC

Input: Input points*M ids*, number of iterations*k*,threshold distance *d*, inline threshold elements *m*, number of points to fit a symmetric axis *n*

Output:M_{best}, A_{best}

```
• for i = 0, ..., k do
•     A = {n random points from M ids};
3     M_base = model fitted to A;
4     for point in M ids − A do
•         if M_base(point) ≤ d then
•             A+ = {point};
•         end
•     end
•     if |A| ≥ m then
•         M_candidate = model fitted to A;
•         err = the mean distance error of M_candidate;
•         if err ≤ err_best then
•             err_best = err, M_best =
              M_candidate, A_best = A;

14        end
15    end
16 end
```

We defined the hyper parameters in the New-key-detection algorithm as follows: The parameter k means we will do RANSAC k times to find the best symmetric axis which minimizes the square loss of the points not in the A_{best}. Whether a point is in asymmetric axis depends on the threshold distanced. There we denote m is a threshold that a candidate symmetric axis $M_{candidate}$ has to fit at least m mid-points and n is the randomly sampled number to fit a base symmetric axis model M_{base} α is the weight of EMA, for combining original key-point and refined key-point symmetric to the axis. We choose α = 0.4 for best performance. So as to show the advantages of our New-key-detection algorithm, we showed some experimental results by using Mask R-CNN (blue points in Fig.1 (b)), axial symmetric constraint (Fig.1(c)) and New-key-detection (Fig.1(d)). Mask R-CNN misses two right cuff points.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 6, Issue 10 , October 2019

Algorithm 2: New-key-detection

Input: Input an imageimg
Output: key-points
To estimate $keypoints_1$ using Mask-RCNN;
$M\ ids$ = {mid-points of paired points in $keypoints_1$};
M_{best}, A_{best} = $RANSAC(M\ ids)$;
for $point\ in M\ ids\text{-}A_{best}$ do
  find corresponding two endpoints in $keys_1$;
  $point_{left} = EMA_{\alpha}(point_{left}, point^{sym}_{right})$;
  $point_{right} = EM\ A_{\alpha}(point_{right}, point^{sym}_{left})$;
end
Find the minimum rectangle closure of $keys_1$ namely $R_1 = [x_{min}, y_{min}, x_{max}, y_{max}]$ and crop img with the R_1 to get $croppedimg$;
Repeat 1-8 for $croppedimg$ to get refined $keys_2$;
$keypoints = (keys_1 + key_2 + [x_{min}, y_{min}])/2$;

Consequently, the corresponding minimum rectangle closure of key-points (used as a bounding box) does not the right cuff of clothes consist, leading to poor performance in the attribute learning of sleeves. However, as we can see in Fig.1(c), the Normalized Error (key-points localization metric defined in the next section) is still big even though we use our axial symmetric constraint once to obtain two missing right cuffs (purple) points.

When compared to these two results, New-key-detection outputs more reasonable results as shown in Fig.1 (d). The NE for the blue points in Fig.1 (d) is smaller than the ones in Fig.1 (b) and Fig.1(c).

By using New-key-detection, we can efficiently detect clothes and locate key-points even when there are occlusions. Suppose if we drop the axial symmetric constraint, just run Mask R-CNN twice for the input image (the second time we input the bounding box from the first run) and we will get a smaller bounding box, a worse localization, and a wrong sleeves recognition result. This observation indicates our axial symmetric constraint is indispensable.

	FashionAI	coco2017 (%)				
	Normalized Error(%)	AP@0.50:0.95	AP@0.50	AP@0.75	AP@medium	AP@large
Mask R-CNN	13.70	62.7	87.0	68.4	57.4	71.1
CPN	13.55	65.3	90.4	74.2	64.3	74.5
New-key-detection	10.96	65.9	89.5	75.2	63.6	75.9

Table 1. The performance of New-key-detection on FashionAI and coco2017 keypoints localization dataset.There the CPN is implemented by ourselves without ground truth bounding boxes on coco2017 for fair.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Issue 10 , October 2019

	Collar	Neck	Neckline	Lapel	Skirt	Pant	Sleeve	Coat	mAP
NASNet(including) Mask R-CNN)	84.27%	80.82%	84.00%	79.15%	73.49%	78.76%	69.50%	65.46%	92.52%
+joint learning	86.02%	83.42%	85.11%	81.82%	74.49%	79.57%	69.85%	65.73%	92.86%
+multi-label loss	86.41%	82.52%	85.68%	80.88%	75.46%	80.29%	70.01%	65.81%	92.94%
+New-key-detection	86.74%	83.67%	85.51%	81.09%	75.56%	80.37%	70.02%	66.31%	92.98%

Table 2. Three significant improvements from our contributions compared with the same baseline. The digits under eight tasks are the accuracy on the validation set.

The bigger bounding box we defined by the key-points is called Normal-crop in the leftmost subfigure of Fig.3, which removes almost all the background objects. However, after Normal-crop, the remaining parts of the image still contains noise information such as sleeves and skirts when dealing with the upper body design tasks. We then propose Neck-crop (the smaller bounding box), a region near the neck of clothes, specialized for upper-body design tasks. With two stacked crops, our recognition model can focus on the corresponding attributes rather than taking the full image as an input.

B.RECOGNITION MODULE: A MULTI-LABEL LOSS. Labeling clothes precisely for training data is a challenging task. FashionAI1 is a large clothing dataset for competition. It contains labels "maybe", reflecting some uncertainty in labeling the attributes for clothing. It is commonly used to soften the labels, that is, a training sample may belong to multiple categories with different certainty levels. We tolerate the samples to be classified as the uncertain labels by assigning soft weights (i.e., β) on the labels marked as "maybe" in annotations, which is a reasonable and effective way to take full ad-vantage of the label information. Thus, we design a weighted cross-entropy loss with the weight ratio β for the "m" labels and (1 − t · β) for the "y" labels. Here t is the number of attributes labeling "m" in this sample. We define our loss function like this:

$$\text{Loss} = -\sum_{i=1}^{N}\sum_{j=1}^{M} W_{ij}\log(p_{ij})$$

There're $p_{i1}, p_{i2}, \ldots p_{iM}$ is the prediction probability of its sample after softmax. N is the number of samples. M is the number of classes and GT is the ground truth. The weight W ij is defined as

$$W_{ij} = \begin{cases} \beta, & \text{GT}_{ij} = m \\ 1 - t * \beta, & \text{GT}_{ij} = y \\ 0, & \text{GT}_{ij} = n \end{cases}$$

The β is chosen as 0.1 in this paper for the best result.

IV. EXPERIMENTS

Our method which is on three different datasets from easy to challenging (resp., FashionAI, Deepfashion, and coco2017) for different tasks (resp., clothing key-points localization and at-tribute learning, attribute learning and human key-point localizations) was evaluated to test the general performance of the aimed method. We first focused on clothing key-point localizations and then generalize our method to similar human key-points. Another purpose of using Deep fashion is that it is more challenging than FashionAI because it has more hard images with a complex background and different views. We aim to show that our method still works well even if the input photos were taken from the side view of a person.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Issue 10 , October 2019

A. KEY-POINT LOCALIZATIONS ON FASHIONAI AND COCO2017. In FashionAI dataset, images have a totally of 24 (10 paired mid-points and 4 non-symmetric points) key-points including some points that are invisible or do not exists in an image.Thus, we choose m = 8, n = 5, d = 1 and k = 10 for best results in New-key-detection. In coco2017 [22， 23], there are 17 (8 pairs of mid-points and 1 non-symmetric point) human key-points. We implement New-key-detection on coco2017 with m = 6, n = 4, d = 1 and k = 10. The input size is 224 × 224, with the learning rate = 0.001. We also implement CPN with the same hyper parameters on these two datasets. We illustrate our experiments on FashionAI key-points dataset in Table.1. In FashionAI, the evaluation metric is NE (Normalized Error), the average normalized distance be-tween predicted key-point coordinates and annotation coordinates. Mask R-CNN's NE is 13.70% while our New-key-detection module reduces it to 10.96%.

	Category		Texture		Fabric		Shape		Part		Style		All	
	Top3	Top5	Top3	Top5	Top3	Top5	Top3	Top5	Top3	Top5	Top3	Top5	Top3	Top5
WTBI [2]	43.73	66.26	24.21	32.65	25.38	36.06	23.39	31.26	26.31	33.24	49.85	58.68	27.46	35.37
DARN [19]	59.48	79.58	36.15	48.15	36.64	48.52	35.89	46.93	39.17	50.14	66.11	71.36	42.35	51.95
FashionNet [20]	82.58	90.17	37.46	49.52	39.30	49.84	39.47	48.59	44.13	54.02	66.43	73.16	45.52	54.61
Weakly [21]	86.30	92.80	53.60	63.20	39.10	48.80	50.10	59.50	38.80	48.90	30.50	38.30	23.10	30.40
Inception-resnet-v2(ours)	90.60	96.45	60.26	69.57	44.82	54.57	60.00	68.47	45.57	55.47	32.58	41.36	48.86	58.12
+ Joint learning and New-key-detection (ours)	92.75	96.71	64.52	73.80	49.87	59.89	64.46	72.46	51.73	60.95	36.45	45.00	53.64	62.69

Table 3. This table shows top3 and top5 recall (%) on DeepFashion attribute recognition dataset.On coco2017, our New-key-detection method also improves the human key-point localizations AP@0.50:0.95 (evaluation metric based on oks [22]) from 62.7 to 65.9 compared with the baseline Mask R-CNN in Tabel.1. Our method achieves better AP@0.50:0.95 and AP@large than CPN because larger humans in images are more likely to contain prior knowledge information (the symmetric prior is more likely to hold when a person dominates the entire image).

B. ATTRIBUTES RECOGNITION ON FASHIONAI. Dataset and Implementation the training/testing images are about 200,000 / 40,000 for all attribute dimensions on FashionAI. In order to learn the common features, we train two parallel multi-task recognition networks for eight at-tribute dimensions (four-length tasks and four upper-bodies de-sign tasks). With learning rate 10−4 and decay rate 0.1 every four epochs, we train our model for a total of 12 epochs. The optimizer is Adam and the batch size is 24. We choose NASNet as base nets of recognition network with input size 399 × 399. The improvements from our model's three methods in different tasks are shown in Table 2. New-key-detection gives the most significant improvement in almost all tasks due to the correct bounding boxes of clothes. M-label loss helps four-length tasks better than upper-body design tasks since some length attributes are confusing. Joint learning helps all tasks to get useful representation. Altogether, we have shown that the proposed approach is efficient for fashion clothes to attribute recognition. Our whole model achieves the top 10th (95.09% mAP) ranking in FashionAI competition.

C. ATTRIBUTES RECOGNITION ON DEEPFASHION. Dataset and Implementation Deepfashion evaluate the performance of clothing category and attribute prediction. It consists of 289,222 (249,222 for training and 40,000 for testing) clothing images, 50 clothing categories and 1,000 clothing attributes. Attributes are divided into six tasks: category, texture, fabric, shape, part, and style. We use the same hyper-parameters during training and it takes four days to converge. The recognition network here we used is Inception-Resnet-v2 for faster speed. However, this dataset does not have "maybe" labels, which limits our model's performance of length tasks.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Issue 10 , October 2019

The experimental results are shown in Table 3. Our model gets almost all tasks' best top3 and top5 recall [20] accept the style task. The reason for the low recall of style task may be the change of some annotations after the dataset was released. We make significant improvements on the leaderboard for two fundamental tasks: the whole attribute recognition task — "all" and clothes category classification task —"category", about 6% and 8%. The results show that our model has good generalization ability in these more challenging datasets.

V. CONCLUSION

A key-point localizations and attributes recognition model which are based on the prior knowledge of garments and individuals are purposed in this research. Two modules such as New-key-detection and the recognition present on the proposed framework. As far as our attribute recognition network shares the attribute values, the New-key-detection module satisfactorily mingles the symmetric relation with various backbones. In accordance with achievements on our approach, a great quantity for experiments elucidated better results both in key-point localizations and attributes recognition.

VI. REFERENCES

1. Lukas Bossard, Matthias Dantone, Till Quack and Luc Van Gool, "Apparel classification with style," in ACCV, 2012.
2. Huizhong Chen, Andrew Gallagher and Bernd Girod, "Describing clothing by semantic attributes," in ECCV, 2012.
3. YannisKalantidis, Lyndon Kennedy and Lijia Li, "Get-ting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos," ICML, 2013.
4. Yongxi Lu, Abhishek Kumar, ShuangfeiZhai, Yu Cheng, Tara Javidi and Rogerio Schmidt Feris, "Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification," CVPR, 2017.
5. Edgar Simoserra, SanjaFidler, FrancescMorenonoguer and Raquel Urtasun, "Neuroaesthetics in fashion: Modeling the perception of fashionability," CVPR, 2015.
6. Zhengzhong Zhou, Xiu Di, Wei Zhou and Liqing Zhang, "Fashion sensitive clothing recommendation us-ing hierarchical collocation model," in ACMMM, 2018.
7. Laiping Zhou, Zhengzhong Zhou, and Liqing Zhang, "Deep part-based image feature for clothing retrieval," in ICONIP, 2017.
8. Alex Krizhevsky, IlyaSutskever, and Geoffrey E. Hin-ton, "Imagenet classification with deep convolutional neural networks," in NIPS, 2012.
9. Kaiming He, Xiangyu Zhang, ShaoqingRen, and Jian Sun, "Deep residual learning for image recognition," CVPR, 2016.
10. Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke and Alexander A Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learn-ing," AAAI, 2016.
11. BarretZoph, Vijay Vasudevan, Jonathon Shlens and Quoc V Le, "Learning transferable architectures for scalable image recognition," CVPR, 2018.
12. Kaiming He, Georgia Gkioxari, Piotr Dollar and Ross B Girshick, "Mask r-cnn," ICCV, 2017.
13. ShaoqingRen, Kaiming He, Ross B Girshick and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," PAMI, vol. 39, no. 6, pp. 1137–1149, 2017.
14. Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," ECCV, 2016.
15. Yilun Chen, Zhicheng Wang, YuxiangPeng, Zhiqiang Zhang, Gang Yu, and Jian Sun, "Cascaded pyramid net-work for multi-person pose estimation," CVPR, 2018.
16. AgnsBorrs, FrancescTous, JosepLlads and Maria Van-rell, "High-level clothes description based on color-texture and structural features," in PRIA, 2003.
17. Vittorio Ferrari, Manuel Marln-Jimnez and Andrew Zisserman, "Pose search: Retrieving people using their pose," in CVPR, 2009.
18. Martin A Fischler and Robert C Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Readings in Computer Vision, pp. 726–740, 1987.
19. Junshi Huang, Rogerio Schmidt Feris, Qiang Chen, and Shuicheng Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," ICCV, 2015.
20. Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in CVPR, 2016.
21. Charles Corbiere, HediBenyounes, AlexandreRame and Charles Ollion, "Leveraging weakly annotated data for fashion image retrieval and label prediction," ICCV, 2017.
22. Tsungyi Lin, Michael Maire, Serge J Belongie, James Hays, PietroPerona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," ECCV, 2014.
23. Bin Xiao, Haiping Wu, Yichen Wei, "Simple Baselines for Human Pose Estimationand Tracking," The European Conference on Computer Vision (ECCV), 2018.