Spatial and Semantic Relations for Pedestrian Attribute Recognition

1st Lei Li
Beijing University of
Posts and Telecommunications
Beijing, China
lei.li@bupt.edu.cn

2nd Yuan Dong
Beijing University of
Posts and Telecommunications
Beijing, China
yuandong@bupt.edu.cn

3rd Fengye Xiong
Beijing FaceAll Co
Beijing, China
fengye.xiong@faceall.cn

4th Hongliang Bai
Beijing FaceAll Co
Beijing, China
hongliang.bai@faceall.cn

***Abstract*—This paper addresses the problem of pedestrian attribute recognition. Previous works typically treat the different attributes independently with each other, without considering possible dependencies between them, or just take semantic relations or spatial relations into consideration. In our work, we propose an end-to-end learning framework combining with a subnet for multi-task classification to take both spatial and semantic relations into account, which proves to be more accurate and effective. Our work can not only deal with binary attributes, but also multi-class attributes in a single network, overcoming the drawbacks in many methods that can only recognize the binary attributes in joint-training. Experiments have been carried out on several benchmarks and positively demonstrated the superiority of our method.**

***Index Terms*—Pedestrian Attribute Recognition, CNN, Spatial, Semantic, Relations**

I. Introduction

Pedestrian attribute recognition plays a vital role in video surveillance application due to its positive effect on several related applications such as person retrieval [1] and person re-identification [2]. Besides, with high accuracy and recall of pedestrian attribute, the recognition has broad application prospects in security fields. Aiming at making predictions for a set of attributes for an image of pedestrian, the problem is generally treated as a multi-label classification, as some semantic attributes including gender, long hair and carrying backpack are binary while others like color of clothes and the type of pants, have multiple choices. Generally speaking, the challenge of pedestrian attribute recognition task lies in the low resolution, pose variations, and occlusions.

In previous works, Layne et al. [2] use SVM and a series of color and texture features to recognize attributes (e.g. gender, backpack) in order to assist person re-identification. In recent years, Convolutional Neural Networks (CNNs) have gained great popularity since the success of Krizvhesky et al. [3] in the ILSVRC-2012 classification task. Some methods transform the multi-label attributes into multiple single-label classification to fine-tune every single attribute with CNN-based models [4]. Since attributes are intrinsically tied to image regions in small parts, some methods [5], [6] focus on the different visual regions of the image. Zhu et al. in [5] roughly divided pedestrians into multiple overlapping body parts to get relevant regions.

However, since the annotations of spatial regions are not available for training, body-part-based methods can be laborious to execute. Apart from that, these approaches ignore the semantic relationships of different attributes. Therefore, recent proposals focus on exploiting the semantic relations to get better performance by rendering the recognition as a direct multi-label classification task. Li et al. [7] rely on full images and leverage body parts along with scene context information to more accurately determine person attributes in a combined deep model. Though multi-label classification methods generally improve the accuracy than multiple single classification ones, few approaches still capture the spatial relationships of the attributes. Zhu et al. in [8] propose a Spatial Regularization Network that captures both spatial and semantic label relations for multi-label classification, simultaneously applying to person attribute recognition. According to this method, the attributes are strictly limited to binary attributes while the multi-class attributes (e.g. color of clothes) are excluded.

Inspired by recent success of attention mechanism in many vision tasks [9], [10], we propose a multi-task convolutional neural network with a subnet using attention maps to make the best use of spatial attentions and capture the semantic dependencies of the attributes. In addition, we use Class Activation Maps [11] to generate the spatial attentions for the classification of different attributes. Our work can not only fit binary attributes, but also multi-class attributes in a single network, overcoming the drawbacks of many methods that can only recognize the binary attributes in joint-training.

The main contributions of this paper are as follows:

- We propose an end-to-end multi-task network which can easily get the semantic relations by a learnable subnet. Besides, the network can get the relevant spatial regions by Class Activation Maps (CAMs). Therefore, we are allowed to get spatial and semantic relations simultaneously, leading to the state-of-the-art performance on several pedestrian datasets.
- The network can be easily transferred to basic networks

978-1-7281-3723-0/19/$31.00 ©2019 IEEE

like GoogleNet[12], ResNet[13] etc.

- Our work is applicable for both binary attributes and multi-class attributes, overcoming the drawbacks in other methods [8], [14], [15],which are only capable of recognizing the binary attributes in joint-training.

In the remainder of this paper we first introduce the detailed method that we adopt to improve the performance of pedestrian attribute recognition in Section II. Section III illustrates the effectiveness of our method by conducting experiments on several benchmarks. Finally, section IV draws a conclusion.

II. PROPOSED METHOD

We propose a multi-task learning structure with a subnet designed for extracting semantic relations from channels as well as capturing spatial information. Besides, we use different fully connected layers for each attribute. Combined with Class Activation Maps[11], the net can focus on the probable regions of every attribute. With these approaches, it is much easier for us to get spatial and semantic relations of the attributes, contributing to further improvement on the accuracy of the recognition. We pre-process the pedestrian images by resizing them to 280×112 before feeding them to the network. In this way, we can avoid the deformation of pedestrian since some attributes are quite sensitive to the shape.

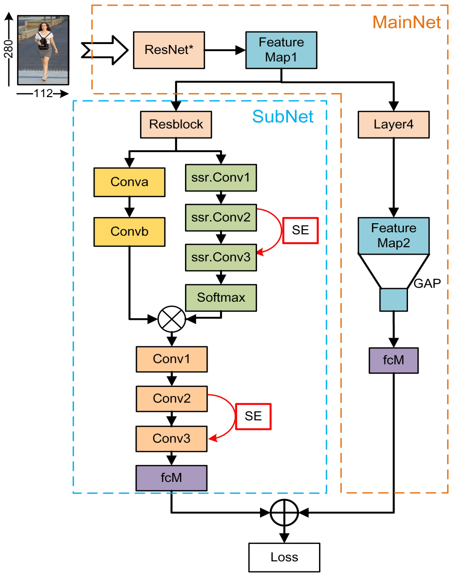

Fig. 1. Overview of our proposed framework. The right part is the structure of ResNet[13]combined with the CAM block, ResNet* is the first three layers of ResNet while the left part is our proposed subnet. The convolutional layers are designed to make the weight and height of the features fixed. The SE is an Squeeze-and-Excitation[16] block.The fcM represents different layers for each attribute.

A. Class Activation Maps

The work by Zhou et al. [11] has shown the global average pooling layer enables CNN to have remarkable localization ability despite being trained on image-level labels. They add the global average pooling layer before the fully-connected layer and propose the Class Activation Maps(CAMs) to highlight the discriminative image regions for an object category.

In order to use the localization ability, we suggest using CAMs[11] for pedestrian attribute recognition. An input image first goes through the ResNet-50, after which, a Global Average Pooling (GAP) layer is applied to the feature maps after the layer4 of ResNet-50. The output of GAP is then taken into fully-connected layers followed by Softmax layers, deciding whether the attribute is present or not. In this network, the parameters of all the layers before fully-connected layers are identical for the different attribute. Although the network is trained for classification, we can also obtain the discriminative regions for each attribute.

B. Subnet Architecture Illustration

As shown in Fig. 1, the main framework containing a subnet is the key-point of our proposal. We will explain the specific structure of the sub-network in this part.

The subnet is added before the last layer of ResNet. Just as Zhu et al. demonstrated in [8], we also use a learnable convolution operation to capture the channel attentions. Let $\mathrm{x}_{i,j}$ denote the visual feature vector at location (i,j) of X, where X is the feature map of Resblock. The three convolution layers with a kernel size of 1×1,3×3 and 1×1 are named ssr.conv, which are designed to get the channel attentions. Note that the spatial information is fixed, the channel relations are enabled to be captured. Before the last convolution that makes the number of channels equal to the number of attributes, we add an SE module that Hu et al. proposed in [16] to get sufficient semantic relations. And then we use a softmax operation to spatially normalize the attention features to $a_{i,j}$.

Therefore, we can get spatially weighted features by element-wise multiply for every attribute:

$$Y^l=\sum_{i,j}a^l_{i,j}(W^lx_{i,j}+b^l) \tag{1}$$

where W_l and b_l are the classifier parameters for label l,$l\in1...M$. This equation can be viewed as an application of label-specific linear classifier to every location of the X. Therefore, the spatial information can be fully used. Since the $a_{i,j}$ sums up to 1 in every channel, we can convert the previews formula into

$$Y^l=W^l\sum_{i,j}a^l_{i,j}x_{i,j}+b^l \tag{2}$$

From the above two equations, $\sum_{i,j}a^l_{i,j}x_{i,j}$ is the attention map for every attribute, which is more related to image regions corresponding to attribute l. We will then learn the channel relations by simple convolution layers and SE module.

Next, we use another two convolution layers with kernel size of 1×1 to optimize the learned channel relations. Afterwards, an SE module is also added before we use a W×H kernel convolution to merge the spatial information. Furthermore, there is one fully-connected layer for each attribute after the three convolution layers.

C. Loss function and prediction

We train the whole net with Softmax-Loss for each attribute so that we can simultaneously train multi-class attributes. Finally, we combine all the M attributes loss.

$$F_{Loss}^{l} = -\sum_{j=1}^{C} y_j log s_j \quad for \quad l = 1...M \tag{3}$$

$$s_j = -e^{\hat{y}_l} / \sum_{d=1}^{C} e^{\hat{y}_d} \quad for \quad j = 1...C \tag{4}$$

The outputs of the main net as well as the subnet are aggregated to make a prediction for every attribute, $\hat{y} = \alpha \hat{y}_{main} + (1-\alpha)\hat{y}_{sub}$, where α is a weighing factor. In our work, we fix the α to 0.5.

III. EXPERIMENTS

We demonstrate the superiority of our method on three representative pedestrian benchmarks(Market-1501[17], RAP dataset[18] and PETA[19]). The Market-1501[17] dataset is collected for person reId and 9 binary attributes are annotated to improve re-identification performance[20]. The RAP dataset [18] consists of 41,585 person images recorded by surveillance cameras. Each image is annotated with 72 attributes, viewpoints, occlusions and body parts. The PETA dataset[19] is a combination of several person surveillance datasets and consists of 19,000 cropped images. Each image is annotated with 61 binary and 5 multi-value attributes. Most of the attributes in the three benchmarks are distributed in different spatial locations to reflect spatial associations, and also include semantic associations. In this section, we will show how our proposed method works in capturing the spatial and semantic relations for attributes recognition.

A. Spatial regions by CAMs

We use CAMs to focus on the exact regions of the attributes that need to be recognized. The different fully-connected layer outputs different CAM for classification. With the assistance of CAM block, we can get more spatial information for attribute recognition. As shown in Fig. 2, the backpack is more likely to appear in the middle of the image while the hat on the top and the handbag at the bottom. The obtained class activation maps in Fig. 2 are basically concentrated on the desired relevant regions when recognizing attributes.

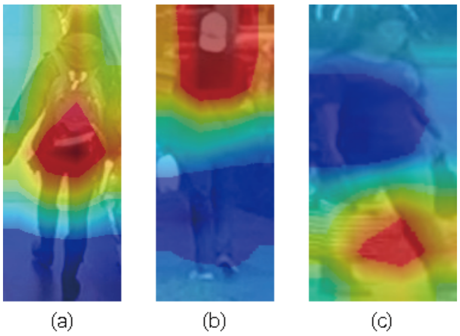

Fig. 2. CAMs from different fully-connected layers for different attributes, (a)backpack attribute (b) hat attribute (c) handbag attribute

B. Spatial and Semantic Relations(SSR)

In this part, We will prove the effectiveness of the subnet that designed to capture both the spatial and semantic relations of the attributes. We compare the result of the single mainnet with the complete network. The experiments are carried out on the Market-1501 dataset, and all the binary attributes are closely tied to the specific regions. Since all the attributes are binary, we use ROC curve to show the results. As Fig. 3 show, some attributes related to spatial locations like backpack and hat have been significantly improved. Apart from that, other attributes considerably relevant in the semantic level like the hair and gender are also obviously promoted. It is obvious that with the SSR subnet we can get better results through the AUC in the legend of Fig. 3.

From the result, we can say that our subnet is effective to assist the main network to take both spatial and semantic relations into consideration. We can confirm that the mean AUC of all the attributes is respectively improved by 0.0278 on Market-1501 with the proposed SSR subnet included.

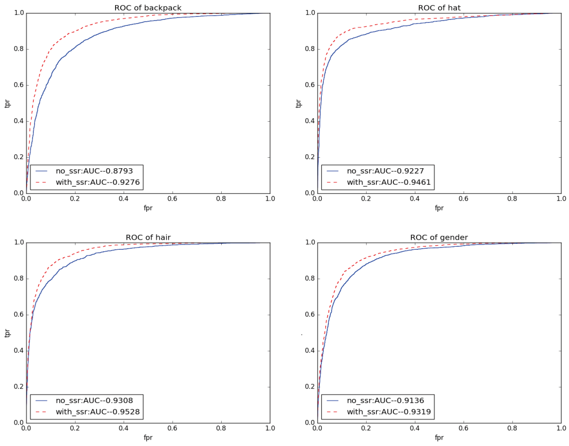

Fig. 3. ROC curves of some attributes in Market-1501. The blue one does not include the SSR subnet, while the red one is with SSR subnet.

C. Comparison with former state-of-the-art approaches

We compare the performance of our method with a number of recent state-of-the-art pedestrian attribute recognition works, including VeSPA[14], DeepMAR [4], $DeepMAR^*$ [18] and WPAL[15]. For a label-based evaluation we compute the mean accuracy (mA) as the mean of the accuracy among positive examples and the accuracy among negative examples of an attribute. This metric is not affected by the imbalance of classes and thus penalizes errors made for the less or more frequent label value equally. In order to account for attribute relationships, we further use example-based metrics like accuracy, precision, recall and F1 score averaged across all examples in the test data. Results of our approach in context of these works are given in Table I for the RAP dataset and Table II for the PETA dataset. We also include results of our mainnet baseline (without SSR subnet) to demonstrate the effectiveness of the proposed spatial-and-semantic-aware architecture.

From the complete evaluation results in Table I and Table II, we can see that our proposed approach achieves state-of-the-art results in all example-based metrics. Besides, compared to the mainnet, we can also improve 1-2% on all the metrics with the SSR subnet contained. It demonstrates that the SSR subnet and CAM block are effective to optimize the recognition of some spatially and semantically sensitive attributes.

Our results of label-based mean accuracy (mA) come second to the unpublished WPAL[15] approach on both datasets. However, the example-based results of WPAL[15] are much lower than ours in comparison. With consideration of all the mertrics, our approach shows competitive performance across the two representative datasets. Compared to the previous published state-of-the-art, our approach has a superior precision-recall trade-off since we can make full use of the spatial and semantic relations.

TABLE I
CAMPARISON OF FORMER METHODS ON RAP

	Evaluation				
Method	**mA**	**Acc**	**Prec**	**Rec**	**F1**
DeepMAR	73.79	62.02	74.92	76.21	75.56
$DeepMAR^*$	74.44	63.67	76.53	77.47	77.00
WPAL-GMP	**81.25**	50.30	57.17	78.39	66.12
WPAL-FSPP	79.48	53.30	60.82	78.80	68.65
VeSPA	77.70	67.35	79.51	79.67	79.59
Our Mainnet-res50	78.21	66.5	79.21	78.82	79.01
Our SSR-res50	79.92	**67.45**	**80.46**	**80.23**	**80.34**

TABLE II
CAMPARISON OF FORMER METHODS ON PETA

	Evaluation				
Method	**mA**	**Acc**	**Prec**	**Rec**	**F1**
DeepMAR	82.89	75.07	83.68	83.14	83.41
WPAL-GMP	**85.50**	76.98	84.07	85.78	84.90
WPAL-FSPP	84.16	74.62	82.66	85.16	83.40
VeSPA	83.45	77.73	86.18	84.81	85.49
Our Mainnet-res50	82.97	76.83	85.43	84.76	85.09
Our SSR-res50	84.28	**78.31**	**86.52**	**85.97**	**86.24**

IV. CONCLUSIONS

In this paper, we focus on the spatial and semantic relations(SSR) for pedestrian attribute recognition. We propose a multi-task network with a subnet capturing the spatial information in the specific regions and semantic information in the channels. In addition, we use CAM block by adding the global pooling layer before fully connected layer to take full advantage of the convolutional layer's positioning capability so that we can get spatial information. With the assistance of SSR subnet and CAM block, we can get the state-of-art performance on several pedestrian benchmarks.

ACKNOWLEDGMENT

This work is supported by Chinese National Natural Science Foundation under Grants 61532018, and Beijing FaceAll Co.

REFERENCES

[1] R. Feris, R. Bobbitt, L. Brown, and S. Pankanti, "Attribute-based people search: Lessons learnt from a practical surveillance system," in *Proceedings of International Conference on Multimedia Retrieval*. ACM, 2014, p. 153.

[2] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes." in *Bmvc*, vol. 2, no. 3, 2012, p. 8.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[4] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*. IEEE, 2015, pp. 111–115.

[5] J. Zhu, S. Liao, Z. Lei, and S. Z. Li, "Multi-label convolutional neural network based pedestrian attribute classification," *Image and Vision Computing*, vol. 58, pp. 224–229, 2017.

[6] G. Gkioxari, R. Girshick, and J. Malik, "Actions and attributes from wholes and parts," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2470–2478.

[7] Y. Li, C. Huang, C. C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *European Conference on Computer Vision*. Springer, 2016, pp. 684–700.

[8] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5513–5522.

[9] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 21–29.

[10] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016, pp. 2921–2929.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich *et al.*, "Going deeper with convolutions." Cvpr, 2015.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] M. S. Sarfraz, A. Schumann, Y. Wang, and R. Stiefelhagen, "Deep view-sensitive pedestrian attribute inference in an end-to-end model," *arXiv preprint arXiv:1707.06089*, 2017.

[15] K. Yu, B. Leng, Z. Zhang, D. Li, and K. Huang, "Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization," *arXiv preprint arXiv:1611.05603*, 2016.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, 2017.

[17] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[18] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, "A richly annotated dataset for pedestrian attribute recognition," *arXiv preprint arXiv:1603.07054*, 2016.

[19] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 789–792.

[20] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang, "Improving person re-identification by attribute and identity learning," *arXiv preprint arXiv:1703.07220*, 2017.