Weakly Supervised Pedestrian Attribute Recognition with Attention in Latent Space

Mingjun Sun[1,2], Hua Yang[1,2(✉)], and Guangtao Zhai[1,2]

[1] Institution of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China
{sun-mj,hyang,zhaiguangtao}@sjtu.edu.cn
[2] Shanghai Key Lab of Digital Media Processing and Transmission, Shanghai, China

Abstract. Pedestrian attribute recognition is a key problem in intelligent surveillance. Relations between attributes and human body structures or relations among attributes are beneficial to attribute recognition, while the annotations are just image-level binary labels. In this work, we propose a novel pedestrian attribute recognition network that takes advantage of latent attribute localizations and local attribute relations to improve the performance of pedestrian attribute recognition. Our method generates latent attribute localization maps by weakly-supervised learning in latent attribute localization (LAL) module. These latent attribute localization maps are fed into the local attribute attention (LAA) module to extract local attributes, and local attributes are interacted with each other with the attention mechanism. Extensive experiments made on the publicly pedestrian attribute datasets of PETA and RAP show that our model outperforms previous methods.

Keywords: Pedestrian attribute recognition · Latent attribute localization · Local attribute attention

1 Introduction

Pedestrian attribute recognition aims to extract semantic descriptions from target person image, including low-level descriptions (e.g., wearing, hairstyle) and high-level ones (e.g., gender, age). Pedestrian attribute recognition is one of the active research areas in computer vision because of its wide applications in intelligent video surveillance systems. Accurate attributes recognition also benefits other applications such as person re-identification and person retrieval.

This work was supported in part by National Natural Science Foundation of China (NSFC, Grant No. 61771303), Science and Technology Commission of Shanghai Municipality (STCSM, Grant Nos. 19DZ1209303, 18DZ1200102, 18DZ2270700, 20DZ1200203), and SJTUYitu/Thinkforce Joint laboratory for visual computing and application.

Mingjun Sun is a student.

Y. Peng et al. (Eds.): PRCV 2020, LNCS 12306, pp. 316–327, 2020.
https://doi.org/10.1007/978-3-030-60639-8_27

In recent years, researchers pay more attention to solve the pedestrian attribute recognition problem. At first, pedestrian attribute recognition mainly relies on hand-crafted features such as color and texture histograms [1,2]. Recently, methods based on deep learning achieve great success, which formulates it as a multi-label classification problem [3]. These methods can roughly be divided into spatial attention methods and label relation methods. Spatial attention methods are proposed to pay more attention to discriminative local features of pedestrian [4–6], which are proved useful due to the existing relations between pedestrian attribute and attribute location on the human body. Label relation methods are proposed to exploit semantic relations to assist attribute recognition [7,8], which improve the performance of pedestrian attribute recognition by considering the dependency and conflicts of labels.

The main problems of existing methods are at least one of the following: (1) Ideally, one attribute corresponds to one specific region according to spatial attention methods. However, the relation between attributes and the human body is quite complicated, and these regions could be disconnected. (2) Some pedestrian attributes are predicted from shared feature vectors, which is beneficial to train the feature extractor. However, classifier needs to handle more redundant features when predicting a specific attribute. For example, The attribute *female* is inferred from many low level features (long hair, dress style), and it doesn't correspond to a specific human body region.

To address these problems, we propose to learn the latent attributes localization and handle local attributes with attention. The target attributes can be seen as combinations of latent attributes, which could be related to more precise regions and easier to represent. In our method, the latent attribute features are first extracted with spatial constraints. And then target attributes are predicted with attention mechanism to model relations among latent attributes.

Different from previous methods, we propose the latent attribute localization module (LAL) to localize the latent attributes, and generated latent attribute localization maps are used to extract local attributes. Then local attribute attention (LAA) methods are adopted to model relations among local attributes. The final predictions are obtained through a voting scheme to output the maximum predictions among different feature levels. The proposed framework is end-to-end trainable and requires only image-level annotations.

The main contributions of this work are as follows:

1. We propose a framework to handle latent attribute localizations and local attribute relations simultaneously in a weakly supervised manner. The latent attribute localization (LAL) module is proposed to localize discriminative latent attributes with image-level labels.
2. The local attribute attention (LAA) method is proposed to process latent attributes simultaneously and model relations among local attributes.
3. We conduct extensive experiments on publicly available pedestrian attribute datasets PETA and RAP, and our method outperforms previous methods.

2 Related Work

Many works have been proposed in the field of pedestrian attribute recognition. At first, pedestrian attribute recognition mainly relies on hand-crafted features such as color and texture histograms [1,2]. Recent years CNN-based approaches make great success in pedestrian attribute recognition and outperform most of traditional methods. The problem is formulated as a multi-label classification problem [3]. These methods can roughly be divided into spatial attention methods and label relation methods.

Spatial attention methods are proposed to pay more attention to discriminative local features of pedestrian [4–6,9], which are proved useful due to the existing relations between pedestrian attribute and attribute location on the human body. Liu et al. [4] propose a multi-directional attention model to learn multi-scale attentive features for pedestrian analysis, which extracts attention maps with convlution methods. Fabbri et al. [5] propose a generative adversial models,which uses features extracted from different human body parts. Li et al. [9] combine pose estimation and spatial transform network to extract local features. Tang et al. [6] extract features from different regions with spatial tranform network. Li et al. [10] model the spatial relations by simply dividing the image into rigid grids. However, these methods try to learn spatial constraints for all attributes, which is unnecessary and hard to learn.

Label relation methods are proposed to exploit semantic relations to assist attribute recognition [7,8], which improve the performance of pedestrian attribute recognition by considering the dependency and conflicts of labels. Wang et al. [7] propose a CNN-RNN network to exploit the relations among attributes. Zhao et al. [8] divide the attributes into several groups and attempt to explore the intra-group and inter-group relationships. However, these methods are mainly reply on pre-defined rules and don't take advantage of relations between attributes and human body regions.

3 Methods

The overview of our proposed framework is illustrated in Fig. 1. The proposed framework consists of a backbone network, Latent Attribute Localization (LAL) modules and Local Attribute Attention (LAA) modules applied to different feature levels. The key idea of this work is to take advantage of latent attribute localizations and local attributes relations to improve the effect of pedestrian attribute recognition.

3.1 Network Architecture

Formally, given an input pedestrian image along with its corresponding labels $y = [y^1, y^2, ..., y^C]^T$ where C is the total number of attributes, and y^c is a binary label that indicates the presence of c-th attribute if $y^c = 1$. We adopt the Inception-V3 [11] as the backbone network in our framework.

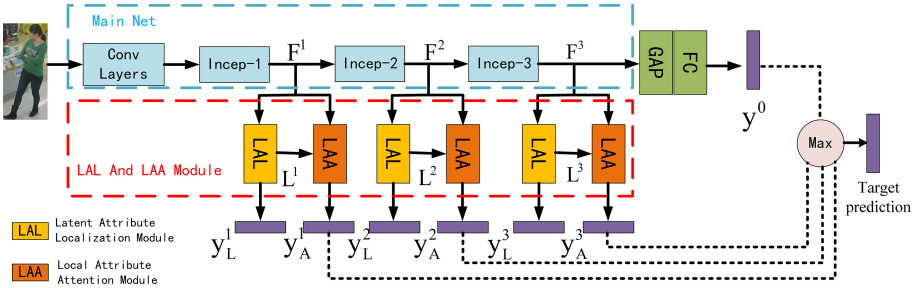

Fig. 1. Overview of the proposed framework. The latent attribute localization (LAL) modules and local attribute attention (LAA) modules are appended after different feature levels to handle attributes of different levels. Outputs from different branches are trained with intermediate supervision ways. Inference process is shown in dashed line, predictions from different levels vote for final prediction in a maximum way. The prediction output from LAL modules are not involved in inference process.

Pedestrian attribute recognition deals with attributes of various levels, so we extract features after different inceptions to take both low-level details feature and high-level semantics feature into account. For each pedestrian image, we can obtain feature representation $F^i, i = 1, 2, 3$ from 3 different levels in backbone network. We then conduct the LAL module and LAA module with different level features F^i. LAL module extracts latent attribute localization maps, and these maps are fed into LAA module. In LAA module, the local attributes are first extracted with features F^i and latent attribute localization maps. The relations among local attributes are modeled in LAA modules by applying corresponding weights on local attributes, and then full connect layers are used to make target attribute predictions. Attributes predictions are generated from different feature levels. During the inference process, final predictions are predicted by voting the maximum prediction among predictions made in different levels. The whole pipeline is shown in Fig. 1.

3.2 Latent Attribute Localization Module

As mentioned in Sect. 1, previous spatial attention methods extract regions related to specific attributes to improve performance. There are two things that should be considered. First, regions related to target attributes could be disconnected and hard to learn. Second, there are correlations among the attributes. Thus it is not suitable to learn each attribute location independently. To address such problems, we propose to learn latent attributes locations.

The details of the latent attribute localization module are shown in Fig. 2. The latent attribute localization method is motivated by weakly supervised detection and localization method [12–14]. Given input $F^i, i \in 1, 2, 3$, stacked convolution layers with kernel size equals to 1 are used to extract latent attribute localization maps z^i. In our experiment, the convolution layer number is set as 3. The kernel number of last convolution layer referred to as N_i equals the number

of the latent attributes. The pixel value of c-th channel at position (h, w) of latent attribute localization maps are referred as $z^i_{c,h,w}$. The extracted latent attribute localization maps are then spatially normalized to put more attention on discriminative regions, and we get normalized latent attribute localization maps $a^i \in \mathbb{R}^{H^i_L \times W^i_L \times N^i}$. The nomalization process of $a^i_{c,h,w}$ is shown as Eq. 1:

$$a^i_{c,h,w} = \frac{exp(z^i_{c,h,w})}{\sum_{h,w} exp(z^i_{c,h,w})} \tag{1}$$

Following [14], the feature maps F^i are concurrently passed to convolution layers followed by the sigmoid function. This branch is used to decrease the influence when the latent attribute is absent. The parameter in this branch is set the same as previous branch. And we can get the latent attribute confidence maps s^i from this branch. The output from two branches are element-wise multiply to get the final latent attribute localization map L^i as Eq. 2:

$$L^i = a^i \odot s^i \tag{2}$$

And we get the latent attribute localization maps L^i that the LAL module learns to represent in the weakly supervised method. Then we convert latent attribute into target attribute predictions for training. The localization map L^i is fed through pooling layers and full connect layers to make target attribute prediction, and predictions from LAL Module on i-th level are referred as y^i_L. These predictions are not involved in reference process.

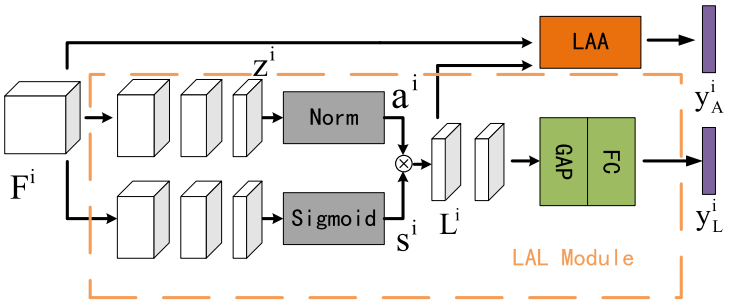

Fig. 2. Details of LAL module

3.3 Local Attribute Attention Module

The LAA module is proposed to handle relations among local attributes and make target attribute recognition. The target attribute can be seen as combination of attributes, and attention mechanisms are adopted in LAA module to handle the relations. The details are shown in Fig. 3.

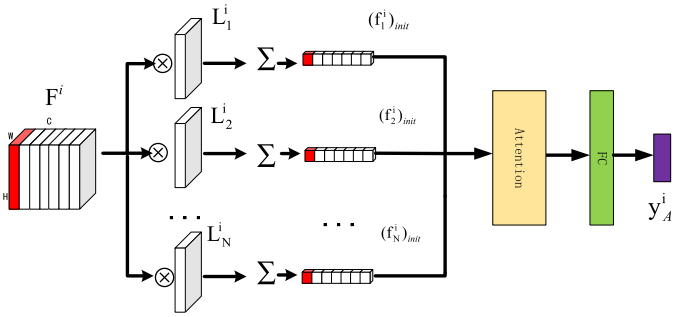

Fig. 3. Details of LAA module

With feature maps F^i from i-th level and extracted latent attribute localization maps L^i, we can extract local attributes vectors, shown as Eq. 3:

$$(f_n^i)_{init} = \sum_{h,w} L_n^i \odot F^i, n \in 1, 2, ..., N \tag{3}$$

To make local attributes distinguishable from each other, we adopt hard position encoding methods [15]. For the n-th local attribute, it computes cosine and sine functions of different wavelengths and adds to the local attribute vector [15]. The position encodered n-th local attribute is referred as f_n^i. With input set of N local attributes $f_1^i, f_2^i, ...f_N^i$, the ralation feature r_n^i of the whole local attribute set with respect to the n-th local attribute f_n^i is computed as follows:

$$r_n^i = \sum_{m=1}^{N} w_{mn}^i \cdot \phi_W^i(f_m^i) \tag{4}$$

where ϕ_W^i is learnable linearly transform, w_{mn}^i is the relation weight that indicates the ralation of m-th local attribute and n-th local attribute. And w_{mn}^i in computed as Eq. 5:

$$\begin{aligned} (w_{mn}^i)_{init} &= \frac{\phi_K^i(f_n^i) \cdot \phi_Q^i(f_m^i)}{\sqrt{d_k}} \\ w_{mn}^i &= \frac{(w_{mn}^i)_{init}}{\sum_k (w_{kn}^i)_{init}} \end{aligned} \tag{5}$$

where ϕ_K^i, ϕ_Q^i are learnable linearly transform. d_k is the dimention of local attribute vector. It calculates the similarity of f_n^i in key space and f_m^i in query space as the relation between n-th local attribute and m-th local attribute. Then the relation features are passed through full connect layers to get the target attribute predictions y_A^i.

3.4 Loss Function

We adopt the weighted binary cross-entropy loss function [3] as the loss function, formulated as follows:

$$Loss(\hat{y}, y) = -\frac{1}{K}\sum_{c=1}^{K} e^{-p^c}(y^c log(\hat{y}^c) + (1-y^c) log(1-\hat{y}^c)) \tag{6}$$

where y^c is the ground truth label which represents whether target person has c-th attribute, $\hat{y}$ denotes the output label probability of the c-th label, p^c denotes the positive ratio of c-th attribute in the training set, K denotes the number of target attributes. As stated in [3], the usage of weights of positive samples in the loss function can alleviate the problem of imbalanced label distribution. All predictions(y^0, y_L^i, y_A^i) are trained with the above weighted entropy loss. During inference process, predictions(y^0, y_A^i) from different levels vote for final prediction in a maximum way. The predictions(y_L^i) from LAL module are just for auxiliary training.

4 Experiments

4.1 Implementation

Datasets. For evaluations, we used two publicly available pedestrian attribute datasets: (1) **PETA Dataset** [16]. The dataset consists of 19000 person images. Following [3], we divide the whole dataset into three nonoverlapping partitions: 9500 for model training, 1900 for verification, and 7600 for model evaluation. And we choose 35 attributes that positive ratio is higher than 5% in our experiment. (2) **RAP Dataset** [17]. The dataset has 41585 images drawn from 26 indoor surveillance cameras. Following the official protocol [17], we split the whole dataset into 33,268 training images and 8,317 test images. We choose 51 attributes that positive ratio higher than 1% in our experiments.

Evaluation Metrics. We adopt two types of metrics for evaluation [17]: (1) Label-based metric: we calculate the **mA** as the mean of positive accuracy and negative accuracy for each attribute. (2) Instance-based metric: we adopt four well-known criteria: **accuracy**, **precision**, **recall** and **F1 score**.

Implementation Details. Our model is implemented with Pytorch. Inception-V3 model pretrained from the ImageNet image classification task is adopted as the backbone network. We train the network with batch size equals to 32. The initial learning rate of training is 1e–4 and changes to 1e–5 after 10 epochs. The optimization algorithm used in training is Adam optimization algorithm [18].

Table 1. Performance comparisons against previous methods on RAP dataset. Best results are in **bold**, and second best results are underlined.

Method	Metric				
	mA	Acc	Prec	Recall	F1
ACN [19]	69.66	62.61	80.12	72.26	75.98
DeepMar [3]	73.79	62.02	74.92	76.21	75.56
JRL [7]	77.81	–	78.11	78.98	78.58
GRL [8]	81.20	–	77.70	**80.90**	79.29
RA [20]	81.16	–	79.45	79.23	79.34
HP-Net [4]	76.12	65.39	77.33	78.79	78.05
PGDM [9]	74.31	64.57	78.86	75.9	77.35
Ours	**83.22**	**69.91**	**80.56**	80.04	**80.29**

Table 2. Performance comparisons against previous methods on PETA dataset. Best results are in **bold**, and second best results are underlined.

Method	Metric				
	mA	Acc	Prec	Recall	F1
ACN [19]	81.15	73.66	84.06	81.26	82.64
DeepMar [3]	82.89	75.07	83.68	83.14	83.41
JRL [7]	85.67	–	86.03	85.34	85.42
GRL [8]	86.70	–	84.34	**88.82**	86.51
RA [20]	86.11	–	84.69	88.51	86.56
HP-Net [4]	81.77	76.13	84.92	83.24	84.07
PGDM [9]	82.97	78.08	86.86	84.68	85.76
Ours	**86.91**	**78.62**	**87.10**	87.04	**87.06**

Table 3. Performance comparisons on RAP dataset when gradually adding proposed component to the baseline model.

Dataset	PETA		RAP	
Component	Metric			
	mA	F1	mA	F1
Baseline	80.63	82.58	75.80	77.78
LAL(single level)	82.48	83.07	77.59	79.24
LAL+LAA(single level)	84.11	83.48	80.37	79.40
LAL+LAA(multi-level)	**86.29**	**85.02**	**83.22**	**80.29**

4.2 Evaluation

In this section, we compare the performance of our proposed method against several previous methods. We choose the representative methods of three catagories mentioned in Sect. 1 and compare proposed method with them: (1) Holistic methods including ACN [19] and DeepMAR [3]. (2) Label relation methods including JRL [7], GRL [8] and RA [20]. (3) Spatial attention methods including HP-Net [4] and PGDM [9].

Tabel 1 and Tabel 2 shows the comparison results of the proposed methods against previous methods on PETA dataset and RAP dataset. The result suggests that our proposed method achieves superior performance compared with existing works. On RAP dataset, our proposed method achieves significant performance. The proposed method achieves the best performance on mA, accuracy, precision, and F1. The mA and F1 can be selected as main metrics to evaluate the methods performance for classification problems. The results suggest that our proposed method achieves superior performances compared with existing methods. The precision and recall metrics are mutually exclusive. Moreover, our method aims to extract more precise local features with complex network structures, which improves the credibility of results. So the precision of our method is quite high. Besides, our method has a faster speed compared with relation-based network [7,8] for its parallel structure. The improvement of performance on PETA dataset is not so obvious. In fact, there are more attributes at high levels depend on human and object interaction in the RAP dataset (e.g., action, attachment), which are better improved by the proposed methods. The comparison details are shown in Sect. 4.4.

4.3 Ablation Study

To validate our contributions, we further perform ablation studies on the PETA dataset and RAP dataset. We choose mA and F1 score as the representation of label-based and instance-based metrics to evaluate the effect. The result with the Inception-V3 net is chosen as the baseline method for comparison. In the baseline method, the input image is passed through convolution layers, pooling layers and full connect layers to make attribute recognition.

As shown in Table 3, based on Inception-v3 as the baseline network, we gradually add modules on it to analyze the effect. (1) **Latent Attribute Localization Module** We first evaluate the contribution of the LAL module by appending LAL Module after backbone convolution layers. The predictions from LAL modules are selected as target attribute predictions. The mA and F1 both increase, which demonstrates the spatial regularization of latent attributes is effective. And the improvement is quite obvious. (2) **Local Attribute Attention Module** We evaluate the impact of the LAA module by appending the LAA module with the LAL module at the same time. The improvement of mA is quite significant, which demonstrates the attention mechanism on local attributes is useful. The effect is further improved by handling the relationship among attributes with LAA module. (3) **Modules on Different Levels** Then

we evaluate the effect of applying the LAL module and the LAA Module on multiple feature levels. We apply the proposed Modules after Incep-1, Incep-2, and Incep-3 and get final predictions with the voting max scheme. The multi-level framework is proved useful because it takes advantage of features extracted from different levels. Considering the model complexity, LAL and LAA modules are applied on 3 levels in our method.

4.4 Improvements on Different Attributes

Table 4. 5 attributes with the greatest improvement in RAP dataset

Attribute	Improvement
Attach paper bag	12.4%
Upbody suitup	12.1%
Shoes cloth	11.5%
Attach handbag	11.3%
Bold head	9.6%

Compared with Baseline methods, the proposed methods increase the accuracy better in the following attributes on RAP dataset: attach paper bag (increase 12.4%), upbody suitup (increase 12.1%), shoes cloth (increase 11.5%), attach handbag (increase 11.3%), bold head (increase 9.6%), as shown in Table 4. We can find that these attributes are more focused on understanding the structure of the human body. These attributes are local attributes or combinations of local attributes. On th contrary, the improvements on abstract attributes like age and body shape are not so obvious. Our proposed method improves the ability to understand human body structure by learning latent attributes. Thus our method performs better on recognition of the attributes related with human body.

5 Conclusion

In this work, we propose a novel end-to-end model for pedestrian attribute recognition by learning latent attributes. Latent Attribute Localization (LAL) module learns the relation between latent attributes and human bodies, and the localization maps are used to extract local attributes. Relations among local attributes are modeled with Local Attribute Attention (LAA) methods. Our proposed model outperforms a wide range of existing pedestrian attribute recognition methods. Extensive experiments demonstrate the effects of proposed modules LAL and LAA.

References

1. Layne, R., Hospedales, T.M., Gong, S., Mary, Q.: Person re-identification by attributes. In: BMVC, vol. 2, p. 8 (2012)
2. Liu, C., Gong, S., Loy, C.C., Lin, X.: Person re-identification: what features are important? In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 391–401. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33863-2_39
3. Li, D., Chen, X., Huang, K.: Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), pp. 111–115. IEEE (2015)
4. Liu, X., et al.: Hydraplus-net: attentive deep features for pedestrian analysis. In: Proceedings of the IEEE international conference on computer vision, pp. 350–359 (2017)
5. Fabbri, M., Calderara, S., Cucchiara, R.: Generative adversarial models for people attribute recognition in surveillance. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE (2017)
6. Tang, C., Sheng, L., Zhang, Z., Hu, X.: Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4997–5006 (2019)
7. Wang, J., Zhu, X., Gong, S., Li, W.: Attribute recognition by joint recurrent learning of context and correlation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 531–540 (2017)
8. Zhao, X., Sang, L., Ding, G., Guo, Y., Jin, X.: Grouping attribute recognition for pedestrian with joint recurrent learning. In: IJCAI, pp. 3177–3183 (2018)
9. Li, D., Chen, X., Zhang, Z., Huang, K.: Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In: 2018 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2018)
10. Li, Q., Zhao, X., He, R., Huang, K.: Visual-semantic graph reasoning for pedestrian attribute recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8634–8641 (2019)
11. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
12. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2846–2854 (2016)
13. Sarafianos, N., Xu, X., Kakadiaris, I.A.: Deep imbalanced attribute classification using visual attention aggregation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 680–697 (2018)
14. Zhu, F., Li, H., Ouyang, W., Yu, N., Wang, X.: Learning spatial regularization with image-level supervisions for multi-label image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5513–5522 (2017)
15. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
16. Deng, Y., Luo, P., Loy, C.C., Tang, X.: Learning to recognize pedestrian attribute. arXiv preprint arXiv:1501.00901 (2015)
17. Li, D., Zhang, Z., Chen, X., Ling, H., Huang, K.: A richly annotated dataset for pedestrian attribute recognition. arXiv preprint arXiv:1603.07054 (2016)

18. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Sudowe, P., Spitzer, H., Leibe, B.: Person attribute recognition with a jointly-trained holistic CNN model. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 87–95 (2015)
20. Zhao, X., Sang, L., Ding, G., Han, J., Di, N., Yan, C.: Recurrent attention model for pedestrian attribute recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9275–9282 (2019)