Multi-attribute Learning for Pedestrian Attribute Recognition in Surveillance Scenarios

Dangwei Li[1], Xiaotang Chen[1], Kaiqi Huang[1,2]
[1]CRIPAC & NLPR, CASIA
[2]CAS Center for Excellence in Brain Science and Intelligence Technology
Email:{dangwei.li, xtchen, kaiqi.huang}@nlpr.ia.ac.cn

Abstract

In real video surveillance scenarios, visual pedestrian attributes, such as gender, backpack, clothes types, are very important for pedestrian retrieval and person re-identification. Existing methods for attributes recognition have two drawbacks: (a) handcrafted features (e.g. color histograms, local binary patterns) cannot cope well with the difficulty of real video surveillance scenarios; (b) the relationship among pedestrian attributes is ignored. To address the two drawbacks, we propose two deep learning based models to recognize pedestrian attributes. On the one hand, each attribute is treated as an independent component and the deep learning based single attribute recognition model (DeepSAR) is proposed to recognize each attribute one by one. On the other hand, to exploit the relationship among attributes, the deep learning framework which recognizes multiple attributes jointly (DeepMAR) is proposed. In the DeepMAR, one attribute can contribute to the representation of other attributes. For example, the gender of woman can contribute to the representation of long hair and wearing skirt. Experiments on recent popular pedestrian attribute datasets illustrate that our proposed models achieve the state-of-the-art results.

1. Introduction

Visual attribute recognition is an important research area in computer vision due to its high-level semantic knowledge, which could bridge low-level features and high level human cognitions. It has achieved much success in areas such as image retrieval [16,19], object recognition [4,5,20], face recognition [11, 17], person re-identification [12–14]. It has also shown great potential in smart video surveillance [8] and video-based business intelligence [15].

Current attribute recognition methods mainly focus on two application scenarios: natural scenario and surveillance scenario. Many researchers pay great attention on natural scenarios attributes recognition and get great success in object recognition, face recognition etc. For example, attributes recognition in natural scenarios is first proposed by Ferrari et al. [6]. In this work, a probabilistic generate model is proposed to learn the low-level visual attributes, such as "striped" and "spotted". Siddiquie et al. [16] explicitly model the correlation between different query attributes and generate the retrieval list. Kumar et al. [11] explore comparative facial attributes and model them through binary classifiers for face verification. Zhang et al. [20] propose a pose aligned neural networks to recognize human attributes (e.g. age, gender, expression) on images under unconstrained scenarios. Totally, these methods [11, 16, 20] focus on high quality images. However, images are blurry and have low resolution, large pose difference and illumination variations in surveillance scenarios. As a result, attributes recognition in surveillance scenarios is much more challenging.

Figure 1. Popular databsets for attributes recognition in surveillance scenarios. Positive and negative example images are indicated by red broken line and blue solid line boxes, respectively. Some attributes are shared with the same person, such as long hair, skirt, and shorts.

There are also some pioneering works on attribute recognition in surveillance scenarios. Layne et al. [12] first using Support Vector Model(SVM) to recognize attributes (e.g. "gender", "backpack") to assist pedestrian re-identification. To solve the attribute recognition problem

978-1-4799-6100-9/15/$31.00 ©2015 IEEE

in mixed scenarios, Zhu et al. [21] introduce a pedestrian database(APiS) and use boosting algorithm to recognize attributes. Deng et al. [1] construct the biggest pedestrian attribute database(PETA) and utilize SVM and Markov Random Field to recognize attributes. However, these methods [1, 12, 21] use handcraft features, which cannot represent the images in surveillance scenarios effectively. In addition, the relationship among attributes is ignored, which is very important to attributes recognition tasks. For example, long hair feature has a higher probability for women than men. So the hair length could help to recognize the gender.

Being inspired by the Convolutional Neural Networks(CNN)'s outstanding performance on different traditional computer vision tasks [7, 10, 18], we proposes two CNN based attributes recognition methods (DeepSAR and DeepMAR) to recognize attributes in surveillance scenarios. In the DeepSAR, each attribute is treated as an independent component and a binary classification network is trained to recognize each attribute. In the DeepMAR, pedestrian attributes recognition is treated as a multi-label classification problem. The proposed methods obtain the state-of-the-art results on existing popular pedestrian attributes datasets.

In this paper, there are two contributions.

- To handle the complicated surveillance scenarios, the automatically learned features are introduced into pedestrian attributes recognition instead of handcrafted features. Treating each attribute as an independent component, the DeepSAR model is proposed to recognize each attribute one by one.
- To exploit the relationship among attributes effectively, the unified multi-attribute jointly learning framework DeepMAR is proposed to recognize multi-attribute simultaneously. In addition, the weighted sigmoid cross entropy loss is proposed to handle the unbalance among attributes and obtain the state-of-the-art results.

2. Methods

In this section, two methods are proposed to solve pedestrian attributes recognition problem. At the first part, the DeepSAR model is proposed to recognize each attribute one by one. At the second part, the DeepMAR model is proposed to recognize multiple attributes jointly.

2.1. Single Attribute Recognition

Before the details of our algorithm are introduced, some basic symbols will be described first. Consider there are N pedestrian images that have been labeled with L attributes. Each image is represented as x_i, $i \in 1, ..., N$. The corresponding attribute label vector of x_i is y_i. Each element of label vector y_i is represented as y_{il}, $l \in 1, ..., L$ and $y_{il} \in \{0, 1\}$. If $y_{il} = 1$, it means that the training example x_i has the l'th attribute and vice versa.

Treating each attribute as an independent component, the DeepSAR method is proposed to predict each attribute. The basic structure of DeepSAR has been shown in Figure 2(a). The ConvNet in Figure 2(c) is a shared network structure between DeepSAR and DeepMAR. It includes five convolutional layers and three full connective layers. ReLu neural units are applied after each layer. The max pooling layer and local normalization layer are added after the first two ReLu layers. There is also a max pooling layer after the fifth ReLu layer. The input of DeepSAR is an image with its attribute label in the training stage. The output of DeepSAR has two nodes, which are the probabilities of the image belonging to the attribute or not.

For each attribute, an independent DeepSAR model is finetuned based on CaffeNet [3], which is the same to AlexNet [10] except switching the order of normalizing and pooling layer. The softmax loss is adopted to compute the final classification loss in proposed DeepSAR model. The loss function used for l'th attribute prediction model is $Loss_l$ in Formula 1. $\hat{p}_{i,y_{il}}$ is the softmax output probability of the convolutional neural networks for l'th attribute.

$$\begin{aligned} Loss_l &= -\tfrac{1}{N}\textstyle\sum_{i=1}^{N} log(\hat{p}_{i,y_{il}}) \\ l &\in \{1, ..., L\}, y_{il} \in \{0, 1\} \end{aligned} \tag{1}$$

$$\hat{p}_{i,y_{il}} = exp(x_{i,y_{il}}) / \sum_{y_{il}=0}^{1} exp(x_{i,y_{il}}) \tag{2}$$

2.2. Multi-attribute Recognition

Generally, attributes are interconnected. As in the Figure 1, it is clear that some attributes are shared by the same person in the dataset. How to utilize these relationships among attributes is still a challenge. To better utilize the relationship among attributes, the unified multi-attribute jointly learning model (DeepMAR) is proposed to learn all the attributes at the same time.

The basic structure of proposed DeepMAR has been shown in Figure 2(b). Different from DeepSAR, the input of the DeepMAR is an image with its attribute label vector and the loss function considers all the attributes jointly. Different from the loss function in DeepSAR, the sigmoid cross entropy loss, which is defined in Formula 3, is introduced in multi-attribute recognition.

$$\begin{aligned} Loss = -\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{l=1}^{L}(y_{il} log(\hat{p_{il}}) \\ +(1 - y_{il}) log(1 - \hat{p_{il}})) \end{aligned} \tag{3}$$

$$\hat{p_{il}} = \frac{1}{1 + exp(-x_{il})} \tag{4}$$

$\hat{p}_{il}$ is the output probability for the l'th attribute of example x_i. y_{il} is the ground truth label which represents whether the example x_i has the $l'th$ attribute or not.

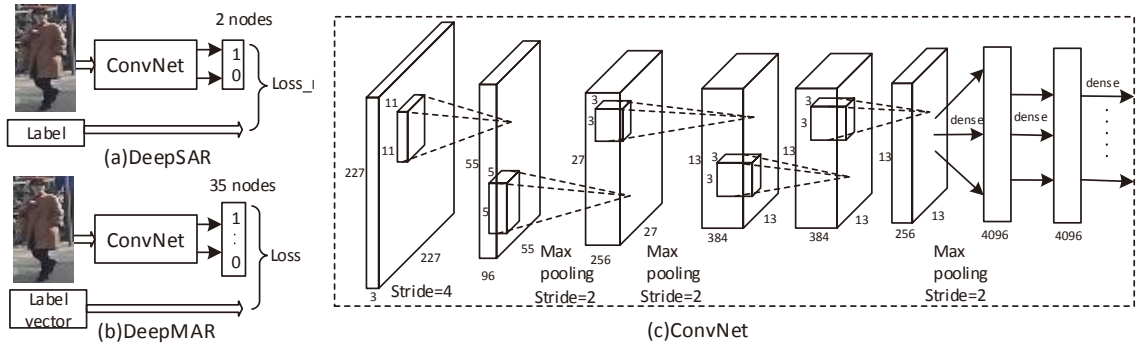

Figure 2. An illustration of the architecture of our network. (a) is the proposed DeepSAR method which consists of a input image, a shared network (c), 2 output nodes. (b) is the proposed DeepMAR method which consists of an input image, a shared network (c), and 35 output nodes. (c) is a shared sub network between DeepSAR and DeepMAR. Given an image, the DeepSAR outputs a label which represents whether it has the attribute or not, and the DeepMAR output a label vector which represents whether it has each attribute or not.

It is very clear that the loss function of Formula 3 considers all the attributes together. However, attributes do not always have uniform distribution. In fact, attributes always have extremely unbalanced distribution especially in surveillance scenarios. For example, the attributes that "has V-Neck clothes" and "has no hair", have very little positive examples than attributes that "has causal upper" and "is male". To handle the unbalance data, the improved loss function defined in Formula 5 is proposed.

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L} w_l(y_{il}log(\hat{p_{il}}) + (1-y_{il})log(1-\hat{p_{il}})) \quad (5)$$

$$w_l = exp(-p_l/\sigma^2) \quad (6)$$

w_l which is defined in Formula 6 is the loss weight for $l'th$ attribute. p_l is the positive ratio of $l'th$ attribute in the training set. σ in Formula 6 is a tuning parameter which is set as 1 in our experiments. In this paper, the improved loss function is used in DeepMAR method in the following experiments.

3. Experiments

In this section, the proposed methods are evaluated on the current popular PETA [1] dataset first. After that, to further verify our method, the proposed DeepMAR has been evaluated on APiS [21] dataset.

3.1. Experiments on PETA

PETA [1] is the current biggest challenging pedestrian attributes dataset that has been used for benchmark evaluation. It contains 19000 pedestrian images which are captured by real surveillance cameras. All the images in PETA are collected in current popular person re-identification databases, such as [9]. They are labeled with 61 binary attributes and 4 multi-class attributes. Because some attributes have extremely unbalance example distribution, previous methods mainly focus on 35 attributes whose positive proportions are bigger than 1/20. Images in PETA have large variation in background, illumination, and viewpoint. Some images in PETA has been shown in Figure 1. The basic evaluation standard on PETA is to calculate each attribute's mean recognition accuracy, which is the average of positive examples' recognition accuracy and negative examples' recognition accuracy. The widely adopted experimental protocol is to randomly divide the dataset into three parts, 9500 for training, 1900 for verifying and 7600 for testing. This paper also follows the same parameter settings.

For each single attribute, a DeepSAR model is finetuned based the CaffeNet. Due to lack of positive training data, only the last full connective layer has been finetuned. To handle the unbalance distribution, the images are randomly copied to make the positive and negative examples to be equel in the training set. In addition, the images are resized to 256×256 first. After that, they are randomly mirrored and cropped to 227×227 to add the training data. For different attributes, different learning rate, weight decay and iterations are adapted to train a better model.

To utilize the relationship among different attributes, the DeepMAR model is trained based on CaffeNet. To make a fair comparison with DeepSAR, the same data partition is adopted in DeepMAR. Generally, the bottom layers of convolutional neural networks could learn some local color and texture information for common object recognition. The top layers could learn some high level semantic information. In this experiment, all the layers are finetuned based on CaffeNet for better learning the low level and high level features to adapted to surveillance scenarios from natural scenarios. The initial learning rate used is 0.001 and weight decay is 0.005 in this experiment. To handle the unbalance data distribution, the loss that defined in Formula 5 is introduced to train the DeepMAR.

The experiment results on PETA have been shown in Table 1. The MRFr2 [2] is current state-of-the-art method

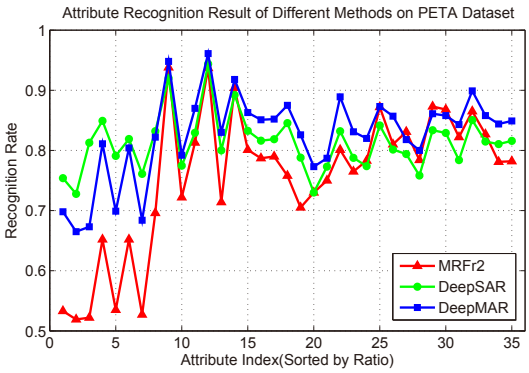

Figure 3. Horizontal axis is attribute index. The attributes are sorted by ascending from left to right according to ratio in Table 1. For example, the 1'th index represent the attribute "V-Neck" and the 35'th index represent the attribute "Casual Lower".

which uses Markov Random Field algorithm to recognize attributes. MRFr2 use handcrafted features instead of automatically learned features. The relationship among different attributes is also not explicitly modeled in MRFr2. The results show that both the DeepSAR and the DeepMAR methods have achieved higher mean accuracy percent(mAP) than current state-of-the-art method MRFr2 on PETA. The proposed DeepSAR and DeepMAR always get higher recognition results on those attributes whose positive examples ratios are small, such as "V-Neck", "Sunglasses", and "Stripes" etc. However these attributes are more important in real surveillance scenarios applications when to search a person based on some attributes descriptions. They have low occurrence ratios and sometimes are saliency attributes compared with other attributes, such as "Casual lower" and "Casual upper."

To have a better view about the results, a graph according to the ratio in Tabel 1 has been drawn in the Figure 3. The proposed DeepSAR method has a larger improvement in low ratio attributes than current state of art. It also has comparable results in high ratio attributes. This could own to the contribution of automatically learned features by CNN. Compared with DeepSAR, the proposed DeepMAR method considers the relationship among different attributes and obtains the best mean recognition accuracy. As in the Tabel 3, the gender, hair length and message box are inter connected and these three attributes have high improvement in recognition accuracy. With the help of relationship among attributes, the proposed DeepMAR utilizes the attributes who have low ratio of positive examples to assist to recognize the attributes that own high ratios of positive examples. This improves the performance on attributes recognition of high positive examples ratios at a large margin. Due to the lack of effective positive examples, DeepMAR achieves low recognition accuracy in attributes whose positive example rates are too low. Given more data, the DeepMAR may exceed the DeepSAR to some extend.

Table 1. Attributes recognition accuracy on PETA.

Attribute	Ratio	MRFr2	DeepSAR	DeepMAR
Age16-30	0.497	**86.8**	82.9	85.8
Age31-45	0.329	**83.1**	79.4	81.8
Age46-60	0.102	80.1	83.3	**86.3**
AgeAbove61	0.062	93.8	92	**94.8**
Backpack	0.197	70.5	78.8	**82.6**
CarryingOther	0.199	73	73	**77.3**
Casual lower	0.861	78.2	81.6	**84.9**
Casual upper	0.853	78.1	81.1	**84.4**
Formal lower	0.138	79	81.9	**85.2**
Formal upper	0.134	78.7	81.6	**85.1**
Hat	0.102	90.4	89.2	**91.8**
Jacket	0.069	72.2	77.5	**79.2**
Jeans	0.306	81	80.2	**85.7**
Leather shoes	0.296	87.2	84.2	**87.3**
Logo	0.04	52.7	**76.1**	68.4
Long hair	0.238	80.1	83.2	**88.9**
Male	0.549	86.5	85.1	**89.9**
MessengerBag	0.296	78.3	77.4	**82**
Muffler	0.084	93.7	94.4	**96.1**
No accessory	0.749	82.7	81.5	**85.8**
No carrying	0.276	76.5	78.8	**83.1**
Plaid	0.027	65.2	**84.9**	81.1
Plastic bag	0.077	81.3	82.9	**87**
Sandals	0.02	52.2	**81.3**	67.3
Shoes	0.363	78.4	75.8	**80**
Shorts	0.035	65.2	**81.9**	80.4
ShortSleeve	0.142	75.8	84.6	**87.5**
Skirt	0.046	69.6	**83.2**	82.2
Sneaker	0.216	75	77.3	**78.7**
Stripes	0.017	51.9	**72.8**	66.5
Sunglasses	0.029	53.5	**79.1**	69.9
Trousers	0.515	82.2	78.4	**84.3**
Tshirt	0.084	71.4	80	**83**
UpperOther	0.456	**87.3**	83.4	86.1
V-Neck	0.012	53.3	**75.4**	69.8
Average	*	75.6	81.3	**82.6**

3.2. Experiments on APiS

The APiS dataset [21] is a popular attributes recognition dataset which include 3661 images in total. The dataset is collected from both surveillance scenarios and natural scenarios. All the images are resized into 128×48 pixels by bilinear interpolation. The dataset is labeled with 11 binary attributes, such as "male", "long hair" and 2 multi value attributes, including upper body color and lower body color. Some images in this dataset are displayed in second row of Figure 1. The common evaluation standard is to part the image into five parts, and compute the average receiver operating characteristic curves(ROC) for different attributes and the nAUC scores. This experiment also follows this rules.

Compared with PETA, the images size in this dataset is too small to train our DeepSAR model. It is easy to be overfitting. However the proposed DeepMAR model could handle the small dataset more flexibly. So in this section, only the DeepMAR model has been verified on this dataset. The experiment standard is the same as the defination [21]. The loss fuction, initial learning rate, and weight decay are the same to the experiment on PETA. After each 20 epoches, the learning rate will decrease by 1/10. The model will convergence with less 100 epoches.

The experiment results has been shown at Figure 4. The Fusion method [21] in Figure 4 is current state-of-the-art method which fuses multiple handcrafted features, such as color, HOG. In the Figure 4, our proposed DeepMAR has

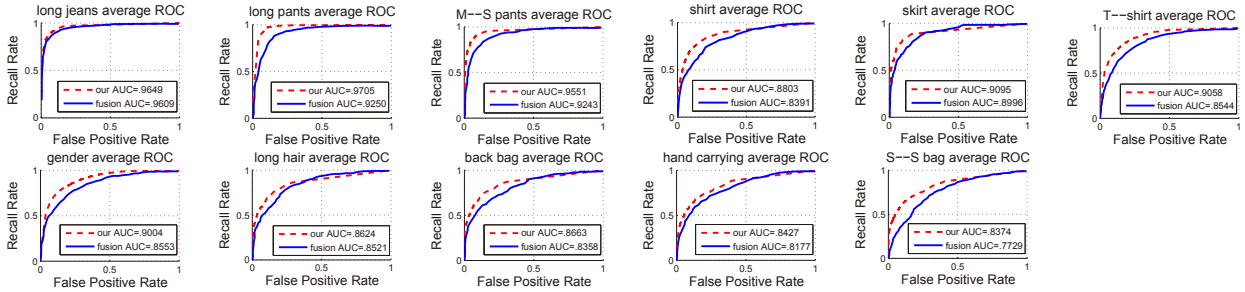

Figure 4. Different attributes' ROC curve on APiS dataset. The read broken line is our method. The blue solid line is current state-of-the-art method which fuses many handcrafted features.

achieve new state-of-the-art mean nAUC score in all the attributes. The average nAUC score of all the attributes has been improved more than 3 points than Fusion method. The attribute "skirt", "long pants", "male", and "Single Shoulder bag" have improved a lot using the proposed multi-attribute jointly learning model DeepMAR. These attributes are inter-connected, for example women often wear single shoulder bag and have short pants or skirt. The proposed DeepMAR has utilized these information to some extent.

4. Conclusion and Future Work

In this paper, two deep learning based methods have been proposed to recognize pedestrian attributes in surveillance scenarios. The proposed DeepSAR has achieved state-of-the-art results in attributes that have low positive examples ratios in PETA dataset. After that, a unified multi-attribute jointly learning model DeepMAR has been proposed, which utilizes the relationship among attributes and has got state-of-the-art results in PETA and APiS. In addition, the proposed DeepMAR model could also be expanded to many multi-label learning problems such as face attributes recognition, multi-object recognition. The experimental results have shown that our proposed methods are effective in pedestrian attributes recognition. In the future, we will explore new loss functions into multi-attribute jointly learning model and apply our multi-attribute learning algorithm to assist attributes based pedestrian re-identification problem.

5. Acknowledgement

This work is funded by the National Basic Research Program of China (Grant No. 2012CB316302), National Natural Science Foundation of China (Grant No. 61403383, Grant No. 61322209 and Grant No. 61175007), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant XDA06040102).

References

[1] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *Proc. ACM Multimedia*, 2014. 2, 3

[2] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Learning to recognize pedestrian attribute. *arXiv preprint arXiv:1501.00901*, 2015. 3

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013. 2

[4] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *Proc. CVPR*, 2012. 1

[5] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proc. CVPR*, 2009. 1

[6] V. Ferrari and A. Zisserman. Learning visual attributes. In *Proc. NIPS*, 2008. 1

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. CVPR*, 2014. 2

[8] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person re-identification*. Springer, 2014. 1

[9] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. ECCV*. 2008. 3

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012. 2

[11] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable visual attributes for face verification and image search. *TPAMI*, 33(10), 2011. 1

[12] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *Proc. BMVC*, 2012. 1, 2

[13] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan. Clothing attributes assisted person re-identification. *TCSVT*, 25, 2015. 1

[14] X. Liu, M. Song, Q. Zhao, D. Tao, C. Chen, and J. Bu. Attribute-restricted latent topic model for person re-identification. *Pattern recognition*, 45(12), 2012. 1

[15] C. Shan, F. Porikli, T. Xiang, and S. Gong. *Video Analytics for Business Intelligence*. Springer, 2012. 1

[16] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *Proc. CVPR*, 2011. 1

[17] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. *arXiv preprint arXiv:1412.1265*, 2014. 1

[18] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proc. CVPR*, 2014. 2

[19] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *Proc. WACV Workshops*, 2009. 1

[20] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Proc. CVPR*, 2014. 1

[21] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Z. Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *Proc. ICCV Workshops*, 2013. 2, 3, 4