Weakly-supervised Learning of Mid-level Features for Pedestrian Attribute Recognition and Localization

Yang Zhou¹

unswork@163.com

Kai Yu²

kyu_115s@buaa.edu.cn

Biao Leng²

lengbiao@buaa.edu.cn

Zhang Zhang³⁴

zzhang@nlpr.ia.ac.cn

Dangwei Li³⁴

dangwei.li@nlpr.ia.ac.cn

Kaiqi Huang³⁴

kqhuang@nlpr.ia.ac.cn

Bailan Feng⁵

fengbailan@huawei.com

Chunfeng Yao⁵

chunfeng.yao@huawei.com

¹ Beijing University of Chemical
Technology

Beijing, China

² Beihang University
Beijing, China

³ CRIPAC & NLPR, CASIA
Beijing, China

⁴ University of Chinese Academy of
Sciences
Beijing, China

⁵ Noah's Ark Laboratory, 2012Labs,
Huawei Technologies Co., Ltd.
Beijing, China

Abstract

Most existing methods for pedestrian attribute recognition in video surveillance can be formulated as a multi-label image classification methodology, while attribute localization is usually disregarded due to the low image qualities and large variations of camera viewpoints and human poses. In this paper, we propose a weakly-supervised learning based approaching to implementing multi-attribute classification and localization simultaneously, without the need of bounding box annotations of attributes. Firstly, a set of mid-level attribute features are discovered by a multi-scale attribute-aware module receiving the outputs of multiple inception layers in a deep Convolution Neural Network (CNN) e.g., GoogLeNet, where a Flexible Spatial Pyramid Pooling (FSPP) operation is performed to acquire the activation maps of attribute features. Subsequently, attribute labels are predicted through a fully-connected layer which performs the regression between the response magnitudes in activation maps and the image-level attribute annotations. Finally, the locations of pedestrian attributes can be inferred by fusing the multiple activation maps, where the fusion weights are estimated as the correlation strengths between attributes and relevant mid-level features. To validate the proposed approach, extensive experiments are performed on the two currently largest pedestrian attribute datasets, i.e.

the PETA dataset [1] and the RAP dataset [2]. In comparison with other state-of-the-art methods, competitive performance on attribute classification can be achieved. The additional capability of attribute localization is also evaluated.

1 Introduction

In video surveillance system, the recognition of pedestrian attributes, such as gender, glasses and wearing style, has great application potentials. For example, in one study on attribute-based people search [3], two suspects in the Boston marathon bombing event can be retrieved rapidly based on their fine-grained attributes, e.g. a light-skin man with black and white hat, wearing sunglasses.

Although much work has been reported on visual attribute recognition with various research backgrounds and application goals, e.g., low-level visual attribute recognition for general object categorization [4], facial attribute recognition for face verification [5] and human attribute recognition from customer photos [6], it is still at a preliminary stage to parse human attributes in surveillance scenarios. With some previous size-limited surveillance datasets, a few studies based on classical hand-crafted features with Support Vector Machine (SVM) classifiers [7, 8, 9] and end-to-end deep neural network models [10, 11, 12] have been proposed. However, the above methods simply solve the problem in multi-label classification framework, i.e. only a set of binary outputs are obtained to indicate the corresponding attributes either exist or not. As far as we know, research on localization of pedestrian attributes in surveillance scenario is still a blank.

In fully-supervised localization methods [13, 14, 15], models are trained with ground-truth bounding boxes of targets. However, it is high-cost to annotate the bounding boxes of multiple attributes across the large datasets. What's more, some attributes have ambiguous boulder definition, such as the "wearing glasses" attribute. Therefore, the approaches based on fully-supervised learning are not feasible on this task.

In this paper, we formulate the problem of pedestrian attribute recognition within a weakly-supervised learning framework, where a Weakly-supervised Pedestrian Attribute Localization Network (WPAL-network) is proposed to perform attribute classification and localization simultaneously. Considering the difficulty to detect all the attributes directly at once, while we want to make use of the internal relationship among attributes, we detect mid-level features instead of the attributes. The existence of attributes is inferred from the response magnitude of correlated mid-level features, and their location is viewed as a fusion of locations of highly-related mid-level features. The multi-scale attribute-aware module we design is trained indirectly with image-level attribute labels, thus no bounding box labeling is required. Both the feature definition and the correlation between mid-level features and attributes are also automatically learned during training of the network, so no priori knowledge is required. One kind of mid-level feature might be shared between multiple attributes, so the relationship among attributes is well-utilized.

To demonstrate the effectiveness of the proposed method, extensive experiments are performed on two largest pedestrian attribute datasets in surveillance scenario, i.e., PETA [1] and RAP [2]. Compared to the state-of-the-art methods, the WPAL-network can achieve competitive performance with higher mean accuracy on attribute classification. The results of attribute localizations are evaluated by both quantitative and qualitative analysis.

The remainder of this paper is structured as follows: In Section 2, we review the related work on attribute recognition and weakly-supervised learning based object localization. In

Section 3, the architecture and mechanism of the WPAL-network is illustrated in detail. In Section 4, experimental results on attribute classification and attribute localization are presented. Conclusions are drawn at the final section.

2 Related Work

In this section, we review the related work on pedestrian attribute recognition and weakly-supervised learning based object localization.

2.1 Pedestrian Attribute Recognition

Early works on pedestrian attribute recognition adopt some classical hand-crafted features and usually trains classifiers for multiple attribute independently. Layne et al. [8] first adopted SVM classifier and some to recognize human attributes (e.g. "gender", "backpack") to assist pedestrian re-identification. Zhu et al. [20] introduced a pedestrian attribute database in surveillance (APiS) and used boosting method to recognize human attributes. Deng et al. [9] constructed the pedestrian attribute database (PETA) and utilized SVM and Markov Random Field to recognize attributes.

Recently, deep learning models enable researchers with powerful feature representations and learning methods to mine the relationships among multiple attributes. Sudowe et al. [14] proposed the ACN model to jointly learn all attributes, and showed that parameter sharing can improve recognition accuracies over independently trained models. This routine was also adopted in the later proposed DeepMAR model [5] and the WPAL-network in this work.

It is another popular idea to make use of spatial part information to improve the performance of attribute recognition. In [3] part models like Deformable Part based Model (DPM) are used for aligning input patches for CNN training. Gaurav et al. [15] proposed an expanded parts model to learn a collection of part templates which can score an image partially with most discriminative regions for attribute classification. The multi-label convolutional neural network (MLCNN) proposed in [22] divided a human body into 15 parts and trained CNN model for each part, then chose the corresponding part model to recognize a given attribute, according to pre-defined spatial priors. The DeepMAR model [10] took three block sub-images in addition with the whole image as the input of the model, where the three blocks correspond to the head-shoulder part, upper body and lower body of a pedestrian respectively. The idea of dividing the image into parts is also adopted in the WPAL-network, which further drives us to adopt a flexible spatial pyramid pooling layer to improve the localization of pedestrian attributes at local scale rather than the whole image.

2.2 Weakly-supervised Object Detection

Indeed, there are some outstanding work on attribute localization existing, including but not limited to pedestrian attribute localization, such as [5, 18]. Most of these methods utilize manually annotated object bounding boxes, or only consider the object regions in the images with clean background in training and testing processes. However, it is high-cost to label bounding boxes of objects manually. To avoid such onerous work, researchers proposed various weakly-supervised learning approaches for object detection and localization. In [14], Pandey et al. demonstrate capability of SVM and deformable part models on weakly-supervised object detection. In [20], Wang et al. proposed unsupervised latent category

Figure 1: The overall architecture of the WPAL-network. The input image goes through the trunk convolution layers derived from GoogLeNet. Then three data flows branch out from different levels. After the FSP layers, the resulting vectors are concatenated for final attribute prediction.

learning, which can discover latent information in backgrounds to help object localization in cluttered backgrounds. Cinbins et al. [8] proposed a multi-fold multiple-instance learning procedure featuring prevention of weakly-supervised training from prematurely locking onto erroneous object locations.

In [8], the proposed network has convolution layers followed by a global max-pooling layer. Each channel of the global max-pooling layer is viewed as a detector for a certain class of object. It is assumed that the positions of max value point in the feature map correspond to the locations where the objects of the target class exist in. However, this method cannot be directly applied to our attribute localization task. Different from objects, some attributes, such as gender and age, are abstract concepts, which do not correspond to certain spatial regions. Meanwhile, some attributes such as "wearing hat" is expected to appear within a certain partition in a pedestrian sample, which can be used to improve the localization of those attributes. Thus, to better fit the task of attribute localization, we embed above-mentioned structure in the middle stage of the network to discover mid-level features relevant to attributes rather than attributes themselves, and propose to use FSP layers instead of a single global max-pooling layer to help constraining location of certain attributes.

3 Weakly-supervised Pedestrian Attribute Localization Network

In this section, we describe the proposed WPAL-network. The network architecture is firstly illustrated. Then, detailed implementations are presented.

3.1 Network Architecture

The framework of WPAL-network is illustrated in Figure 1. The trunk convolution layers, which we view as the feature engine, are derived from the GoogLeNet model [17] pre-trained on the ImageNet dataset. We select the "inception4a/output", "inception4d/output" and "inception5b/output" layers for features at three scale and abstraction levels. This mechanism guarantees that features of some fine-grained attributes can be retained at early stage for recognition, considering that the multiple alternations of convolution and pooling operations can easily eliminate some small-scale and low level features. Each of these three layers is attached a convolution layer (namely CONV1_E, CONV2_E and CONV3_E) to transform the features learnt from general object categories to the mid-level attribute-related features.

Then, we perform weakly-supervised detection of these features by inputting them into the FSPP layers respectively, which perform max-pooling operations at multiple pyramid levels to examine the response magnitudes of the mid-level attribute features, forming multi-scale attribute-aware module. As shown in Figure 2, on one feature map, at the first level, the maximal response of the feature over the whole feature map is output. At the second level, max pooling is performed on 3×3 bins. Each of these bins is 40% as high as corresponding feature map, and are spatially uniformly distributed with overlap. To avoid a high computational cost, we limit the height of pyramid to 2. Each input feature map is processed into a small vector with dimension of the total number of bins at all the pyramid levels. These small vectors of all FSPP layers are concatenated into a high-dimensionality vector (2048 dims), which is further regressed into a 51-dim vector (35-dim for PETA dataset) at the following fully-connected (FC) layers, corresponding to the attribute labels to be predicted.

This weakly-supervised mid-level feature detection mechanism is designed following [20]. Here we assume that during training, the following FC layers can provide image-level existence labels of these features. In [20], the feature maps are connected to the labels with a global max-pooling layer. If a target is labeled existing, its highest response point on the feature map should be stimulated. Otherwise, the point is suppressed. The feasibility and effectiveness of this mechanism has been shown in [20], thus omitted in this paper. Here we use the FSPP layer instead of the max-pooling layer, based on the observation that some features make sense only they appear in some specific positions. For example, for features inferring the type of shoes, apparently those detected on the upper part of the body should not contribute to the decision of shoes types. Therefore, we constraint them spatially in some specific bins at the high levels of the FSPP layers. Results from the other bins as well as the first level bin, the global max-pooling, are suppressed during training.

3.2 Training Mechanism

At the very beginning, the network does not know which mid-level feature contributes to which attribute, and even the features might not be effective at all. As well, it does not know which bin to take to determine an attribute. All of these are learnt gradually during the end-to-end training process of the network. At each training round, gradients are passed backward from the loss function at the final attribute predicting layer to the FSPP layer, through the FC layers. These gradients are viewed as the existence label of the mid-level features, as mentioned in the assumption in the above description of the weakly-supervised detection mechanism. The FC layers, though non-linear, encode positive or negative correlations between the mid-level features and the attributes. For a feature positively related to an attribute, when the attribute is marked existing, positive gradient is passed to the bin

Figure 2: This picture illustrates the structure of a two-level FSPP layer. For a feature map in the bottom layer, on the first level, the bin covers the whole image; on the second level, several overlapping small bins together covers the image. For each bin, max-pooling is performed. The results are concatenated into a large vector of the whole FSPP layer.

outputting the feature, encouraging the feature extraction module to produce higher response magnitude of this feature on images with this attribute. At the same time, the correlation encoded in the FC layer is also enhanced by the gradients, and the feature extractor is adjusted to produce feature more suitable for predicting the attribute. Vice versa, when the attribute is labeled not existing, the existence of the feature shall be suppressed. On the other hand, for a feature negatively related to an attribute, when the attribute is labeled existing, the feature response is suppressed, meaning that this feature should not exist on images with this attribute. The correlation, however, is also suppressed, making this feature still possible to contribute to other attributes.

The selection of bins in the FSPP layer is learnt together with the correlation between mid-level features and the attributes. For example, when a feature is determined positively related to the attribute "wearing hat", it is expected to take effect only on the upper part of the cropped pedestrian image. Therefore, with enough training data, it can be assumed that there are some training samples where the pedestrian does not wear a hat but this feature is found in other parts of the image. In these cases, the bins in the upper part are not effected, but the bins that contain the high response of this feature are suppressed, together with the correlation between the bin and the attribute suppressed. One might argue that this may damage the feature extractor. However, considering that the wrong correlation vanishes in the later stages, we can assume that the feature extractors can be adjusted back with the positive samples.

The above discussion focuses on relationship between feature bins and a single attribute. In the real training scenario, gradients are synthesized from losses of multiple attributes. Therefore, the correlation shifting from one attribute to another of a mid-level feature can be efficient, and it becomes possible that one identical feature existing at different parts of the image contributes to different attributes.

It is worth noting that the shape and size of input image is not fixed, because the feature maps from the feature engine with any size will be turned into vectors with fixed dimensionality in the FSPP layers, which is acceptable for the FC layers. This means the WPAL-network can process images with arbitrary size and resolution without the need of warping or transforming in preprocessing, thus the original shape information of the pedestrian body and accessories can be preserved.

3.3 Loss Function For Unbalanced Training Data

In current pedestrian attribute datasets (e.g. the PETA dataset [14] and the RAP dataset [15]), the distributions of positive and negative samples of most attribute categories are usually

Methods	RAP					PETA				
	mA	Acc	Prec	Rec	F1	mA	Acc	Prec	Rec	F1
ELF-mm	69.94	29.29	32.84	71.18	44.95	75.21	43.68	49.45	74.24	59.36
FC7-mm	72.28	31.72	35.75	71.78	47.73	76.65	45.41	51.33	75.14	61.00
FC6-mm	73.32	33.37	37.57	73.23	49.66	77.96	48.13	54.06	76.49	63.35
ACN	69.66	62.61	80.12	72.26	75.98	81.15	73.66	84.06	81.26	82.64
DeepMAR	73.79	62.02	74.92	76.21	75.56	82.89	75.07	83.68	83.14	83.41
DeepMAR*	74.44	63.67	76.53	77.47	77.00	-	-	-	-	-
WEAK SUP	62.15	47.32	61.83	65.15	63.45	64.43	53.01	71.38	65.00	68.04
ours-GMP	81.25	50.30	57.17	78.39	66.12	85.50	76.98	84.07	85.78	84.90
ours-FSPP	79.64	60.25	69.10	80.16	74.21	81.67	75.72	84.84	83.10	83.96

Table 1: This table shows recognition performance of 7 benchmark algorithms and two models mentioned in this work, evaluated on the RAP and the PETA dataset using mA and example-based evaluation criteria. The DeepMAR* algorithm has no results on the PETA dataset because it depends on ground-truth body part annotations, which is not available on the PETA dataset. The method "ours-GMP" represents the WPAL-network using global max-pooling rather than FSPP.

extremely imbalanced. Some attribute categories, such as "wearing V-neck", are seldom labeled positive in the training data. This imbalance can cause imbalance of gradients passed to the network, which suppress everything in the network. Therefore, a weighted cross entropy loss function is adopted to rebalance the gradients from positive and negative samples:

$$Loss_{wce} = \sum_{i=1}^L \frac{1}{2w_i} \cdot p_i \cdot \log(\hat{p}_i) + \frac{1}{2(1-w_i)} (1-p_i) \cdot \log(1-\hat{p}_i) \quad (1)$$

where L is the number of attributes; p is the ground-truth attribute vector, and $\hat{p}$ is the predicted attribute vector; w is a weight vector indicating the proportion of positive labels over all attribute categories in the training set.

3.4 Attribute Localization & Shape Estimation

To localize an attribute, we first estimate the correlation strength between the attributes and the mid-level feature bins. This is simply calculated as the ratio between the average response magnitude on positive samples to that on negative samples of a bin. Then, a existence possibility map of the attribute can be estimated by superposing all the feature maps before the FSPP layers, weighted by the normalized correlation strengths. The extent of active regions on the map with response magnitude above a threshold indicates the rough shape of the attribute. To determine the fine-position of the attribute, we perform clustering on these regions, and choose the cluster centroids as the location indicators. These centroids can be sorted by the average response magnitude of the cluster region. The number of cluster centroids is defined empirically. For some attributes taking up most of upper or lower body part of a pedestrian, i.e., from attribute "Shirt" to "TightTrousers" listed in Figure 5 in the paper, we find that the scores of the first two clusters are generally much higher than those of the rest clusters. Thus, besides shoes-type, we set the number of cluster centroids to two for these attributes as well. Otherwise, one centroid is set to other small-scale attributes.

Figure 3: These five pie charts are the distribution summaries of independent attribute recognition accuracy of the four benchmark algorithms and the WPAL-network on the RAP dataset. The WPAL-network has more attributes recognized with accuracy above 80%.

4 Experiments

4.1 Datasets and Evaluation Protocols

Extensive experiments have been conducted on the two large-scale pedestrian attribute datasets, i.e., the PETA [9] dataset and the RAP dataset [10]. The PETA dataset includes 19,000 pedestrian samples, each annotated with 65 attributes. The RAP dataset is the largest pedestrian attribute dataset so far, including 41,585 samples with 72 attributes annotated.

For attribute recognition evaluation, we select 35 attributes in the PETA dataset following the protocol in [9], and 51 attributes in the RAP dataset as in [10]. In test phase, images are resized to a fixed size, retaining the original shape. We adopt the mean accuracy (mA) and the example-based criteria proposed in [10] as evaluation metrics.

For localization evaluation, we choose 26 non-abstract attributes including baldhead, glass, skirt, etc. Since no bounding boxes are available in these datasets, we use the part information in the RAP dataset instead for localization evaluation. This information is given in a form of bounding boxes of the head-shoulder, upper-body and lower-body parts respectively. However, these bounding boxes are too rough for the widely-used criterion Intersection over Union (IoU). Therefore, we downgrades it to Intersection over Predicted bounding box (IoP), defined as:

$$IoP_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{\text{area}(B_{ij}^{part} \cap B_{ij}^{pred})}{\text{area}(B_{ij}^{pred})} \quad (2)$$

where N_i is the number of samples with the i 'th attribute, B_{ij}^{pred} is the predicted bounding box on the j 'th sample, and B_{ij}^{part} is the roughly estimated bounding box cropped from the bounding box of the part that includes the attribute.

4.2 Recognition and Localization Performance

For attribute recognition evaluation, we use ACN [10], DeepMAR [9], DeepMAR* [10], WEAK SUP [10] and SVM with CNN features as benchmarks. Here the WEAK SUP method is modified to use GoogLeNet as feature engine and treat each attribute as an object category. Performance comparison on the PETA and the RAP datasets is listed in Table 1. It can be found that our method performs competitively with the state-of-the-art methods. Note that the mA criterion is strongly affected by recall where our method mainly displays strength.

We also compare the individual attribute recognition accuracies of our model with other benchmarks. The accuracy distribution of these models is shown in Figure 3, and Table 2

Model	ACN	DeepMAR	DeepMAR*	WEAK SUP	WPAL
Bald Head	60.7815	62.3884	69.7718	52.3671	85.1139
Long Hair	88.6562	90.2707	92.5585	50.1767	53.1287
Hat	57.3349	62.3532	78.0958	73.5906	85.7659
Sweater	58.5836	66.7341	64.8774	55.4375	71.7950
Suit-Up	71.9403	78.0978	77.8946	66.3667	83.3662
Skirt	50	50	50	58.3059	90.2771
Short Skirt	70.8015	76.8846	78.0803	58.6831	90.2771
Single-Shoulder Bag	64.7848	73.5757	71.8415	57.1089	82.6311
Handbag	63.1261	72.4594	68.198	61.7953	86.4266
Box(Attachment)	64.9486	72.0743	69.5466	63.4732	79.6089
Plastic Bag	58.297	66.9855	61.1198	54.3022	80.2822
Hand Truck	76.8333	81.7525	76.8827	70.9387	88.6439
Other Attachment	65.5210	68.4905	68.6669	59.7538	74.0611
Pulling	65.9776	71.0962	65.5994	58.0522	80.1573
Carrying by Hand	73.1266	78.9767	76.4502	62.3654	85.7927

Table 2: This tables shows attributes in the RAP dataset where independent recognition accuracy differs larger than 5% between our model and best of the benchmarks. We can find that most of these attributes are better recognized by our model, but our model performs extremely bad on the Long Hair attribute.

Figure 4: This figure shows localization performance of the benchmark method proposed in [13] and our method, evaluated on the RAP dataset with the IoP criterion.

shows some selected attributes whose recognition accuracy difference between our model and the best of the benchmarks is larger than 5%. Recognition performance of all attributes can be found in the supplemental materials.

For attribute localization evaluation, we use the WEAK SUP algorithm in [13] for attribute localization benchmark. The comparison results on localizing the 26 non-abstract attributes as well as the overall performance are shown in Figure 4. The mIoP has been improved from 47.14 to 63.40. And in most attributes, our method improves the baseline method clearly, due to the learning of mid-level attribute features rather than the attribute categories themselves. Detailed accuracy of all the 26 attributes is given in the supplemental materials. We also visualize some example localization results in Figure 5.

5 Conclusion and Future Work

In this work, we formulate the pedestrian attribute recognition as a weakly-supervised detection framework for joint pedestrian attribute classification and localization. We proposed the WPAL-network, where, instead of directly predicting and localizing the attributes, a set

Figure 5: The first row contains examples of accurately located fine-scale attributes. The second row contains examples of roughly located abstract or large-scale attributes. The third row contains failed cases, where the attribute labels are correctly predicted, but the locations predicted greatly differ from expectation. The red boxes represent the predicted bounding boxes, and the green boxes represent the expected appearing region of the attributes, roughly estimated and cropped from the part bounding boxes.

of mid-level attribute-relevant features is firstly discovered, and then attributes are predicted based on the response of these features. Furthermore, the location of attributes can be inferred from the response map of these features. Our method achieved competitive recognition accuracy on the two large-scale pedestrian attribute datasets, and its capability of attribute localization is also evaluated.

In the future, we will seek more powerful detectors utilizing additional information, such as background context and location relationship between discovered mid-level features, to improve accuracy and solve recognition failure on attributes like "long hair".

6 Acknowledgement

This work is jointly supported by the National Key Research and Development Program of China (2016YFB1001005), the National Natural Science Foundation of China (Grant No. 61473290, Grant No. 61673375), the National High Technology Research and Development Program of China (863 Program) under Grant 2015AA042307, the Projects of Chinese Academy of Science (Grant No. QYZDB-SSW-JSC006, Grant No. 173211KYSB20160008), and Huawei Technologies Co., Ltd (Contract No.:YBN2017030069).

References

- [1] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: A poselet-based approach to attribute classification. In *2011 International Conference on Computer Vision*, pages 1543–1550. IEEE, 2011.
- [2] Lubomir Dimitrov Bourdev. Pose-aligned networks for deep attribute modeling, February 7 2014. US Patent App. 14/175,314.

- [3] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Multi-fold mil training for weakly supervised object localization. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2409–2416. IEEE, 2014.
- [4] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792. ACM, 2014.
- [5] Kun Duan, Devi Parikh, David Crandall, and Kristen Grauman. Discovering localized attributes for fine-grained recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [6] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester. Cascade object detection with deformable part models. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 2241–2248. IEEE, 2010.
- [7] Rogerio Feris, Russel Bobbitt, Lisa Brown, and Sharath Pankanti. Attribute-based people search: Lessons learnt from a practical surveillance system. In *Proceedings of International Conference on Multimedia Retrieval*, page 153. ACM, 2014.
- [8] Ryan Layne, Timothy M Hospedales, Shaogang Gong, and Q Mary. Person re-identification by attributes. In *BMVC*, volume 2, page 8, 2012.
- [9] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. *Proc. ACPR*, 2015.
- [10] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016.
- [11] Yi Li, Kaiming He, Jian Sun, et al. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*, pages 379–387, 2016.
- [12] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.
- [13] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694, 2015.
- [14] Megha Pandey and Svetlana Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *2011 International Conference on Computer Vision*, pages 1307–1314. IEEE, 2011.
- [15] Gaurav Sharma, Frédéric Jurie, and Cordelia Schmid. Expanded parts model for human attribute and action recognition in still images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–659, 2013.

- [16] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 87–95, 2015.
- [17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.
- [18] R. Tao, A. W. M. Smeulders, and S. F. Chang. Attributes and categories for generic instance search from one example. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 177–186, June 2015. doi: 10.1109/CVPR.2015.7298613.
- [19] Daniel A Vaquero, Rogerio S Feris, Duan Tran, Lisa Brown, Arun Hampapur, and Matthew Turk. Attribute-based people search in surveillance environments. In *Applications of Computer Vision (WACV), 2009 Workshop on*, pages 1–8. IEEE, 2009.
- [20] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. Weakly supervised object localization with latent category learning. In *European Conference on Computer Vision*, pages 431–445. Springer, 2014.
- [21] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Z. Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 331–338, Dec 2013. doi: 10.1109/ICCVW.2013.51.
- [22] Jianqing Zhu, Shengcai Liao, Dong Yi, Zhen Lei, and Stan Z Li. Multi-label cnn based pedestrian attribute learning for soft biometrics. In *2015 International Conference on Biometrics (ICB)*, pages 535–540. IEEE, 2015.