MULTI-TASK LEARNING FOR PEDESTRIAN BODY PARTS DETECTION AND MULTI-ATTRIBUTE CLASSIFICATION

Miaomiao Lou[1,2], *Lin Chen*[1*], *Feng Guo*[2]

[1]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Science
[2]Chengdu University of Information Technology, China
{loumiaomiaocuit, jackeyguofeng}@163.com, chenlin@cigit.ac.cn

ABSTRACT

Pedestrian analysis plays a vital role in intelligent video surveillance and security-centric computer vision systems. Despite that the deep convolutional neural networks (DCNNs) achieved remarkable performance in computer vision, learning fine-grained features of pedestrian attribute tasks is a challenging task in complex surveillance scenarios. In this paper, we proposed a new DCNNs framework for pedestrian analysis, with the main idea of integrating different learning tasks of pedestrian body parts detection and pedestrian attribute classification, which called Hyper-pedestrian convolutional neural network (HP-CNN). Additionally, the proposed HP-CNN bring some advantages: 1) Squeeze-and-excitation block (SE-block) strengthens the representational power of networks by selectively emphasising informative features; 2) Multi-scale feature fusion concatenates more fine-grained information from both the low-level and high-level and enhances the contextual information from different convolutional layers. Experimental results conducted on the largest pedestrian attributes dataset show that the proposed HP-CNN obtained better pedestrian analysis results of both body parts detection and attribute classification, compared to the tested state-of-the-art methods. Especially, the auxiliary task of body parts detection can dramatically boost the performance of attribute classification.

Index Terms— Pedestrian Attribute Classification, Pedestrian Body Parts Detection, Multi-task Deep Learning

1. INTRODUCTION

The attributes of pedestrian (such as gender, age, type of hair, upper-body clothing, lower-body clothing etc.) have attracted a lot of attention in the smart surveillance field in recent years, which offered useful clues for smart video analysis systems, e.g., person retrieval [1] and human identification [2, 3, 4]. Although previous methods based on deep convolutional neural networks (DCNNs) [5, 6, 7, 8] have achieved remarkable results for pedestrian attribute classification, it is still a challenging task due to the following reasons. Firstly, fine-grained attributes (i.e., glasses and accessories) are hard to recognize as the small size of samples cropped out from far distance. Secondly, the large intra-class variations (e.g., appearance diversity and appearance ambiguity) and pose variations exist in pedestrian images in real surveillance scenarios. Recently, it has been shown that using the information of body parts generated the better attribute recognition performance [9, 10], instead of feeding whole pedestrian samples into an end-to-end CNN classifier. However, previous methods based on the fixed image parts or patches ignored the implied relationship between the location of body parts and attributes and also increases the computation and memory consumption.

In this paper, we present a novel DCNNs framework called, Hyper-pedestrian convolutional neural network (HP-CNN), integrating body parts detection and pedestrian multi-attribute classification simultaneously in a unified network based on multi-task learning. The CNN architecture is carefully designed to learn fine-gained features for different learning tasks and exploit the synergy among them. The experiments conducted on the large dataset showed that the proposed HP-CNN achieved better performance than the all tested models.

2. RELATED WORK

In this section, we briefly review the developments on pedestrian attribution classification. Early work on attributes analysis are usually based on the full body images. Layne et al. [11] uses support vector machines (SVM) to identify pedestrian attributes such as backpacks and genders. Zhu et al. [12] introduced the Apis database and proposed a boosting algorithm to solve the attribute classification in mixed scenes. Deng et al. [13] proposed Markov Random Field (MRF) based methods to exploit context of neighboring images to improve attribute inference at far distance. Patrick et al. pro-

* Lin Chen is the corresponding author. This paper is supported by the National Key Research and Development Program of China (No. 238), the Natural Science Foundation of Sichuan Provincial Department of Education (No.14ZA0174), Social Livelihood Foundation of Chongqing (No. cstc2017shmsA120010) and Chongqing research program of technology innovation and application(cstc2017rgzn-zdyfX0020)

978-1-5386-9214-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMEW.2019.00100

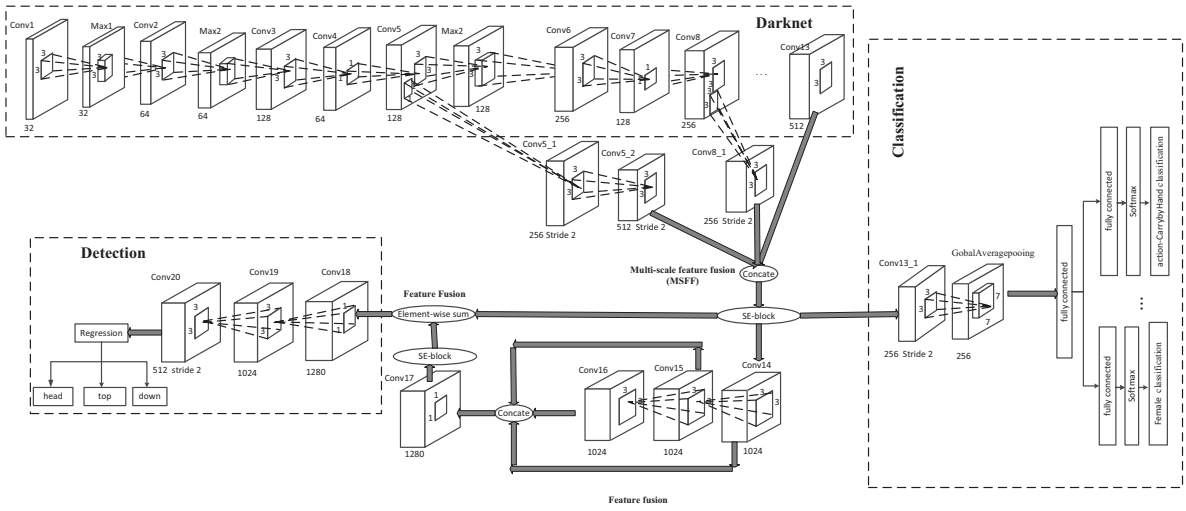

Fig. 1. The architecture of the proposed Hyper-pedestrian CNN.

posed the ACN model [6] to learn all the attributes jointly in a single CNNs model, the experimental results shows that parameter sharing can improve classification accuracy over independently trained CNNs models. This routine is also adopted in the DeepMAR model proposed in [14] to improve attributes recognition with the association among attribute features.

Recently, body parts information is also introduced to assist attribute classification. Zhang et al. [7] employed deformable parts model (DPM) for aligning input patches to improve the attribute classification under large variation of pose and viewpoint. Gaurav et al. [15] propose an expanded parts model to learn a collection of part templates which scored body images partially with most discriminative regions for classification. The MLCNN [16] divides a human body into 15 fixed parts and train CNN models on each part of body, then choose the representative parts of the models according to the spatial constraint prior. The DeepMAR* model [17] takes three block images as input in addition to the whole body image corresponding to the head-shoulder part, upper body and lower body, respectively. The WPAL-network [10] make use of flexible spatial pyramid pooling layers to help locating mid-level features of some attributes in only local patches rather than the whole image. HPlus-net [8] used multi-directionally feeds the multi-level attention maps to different feature layers for pedestrian attribute recognition and human re-identification. Li et al. [18] introduce pedestrian body structure into this task, fusing the part-based Pose Guided DeepModel (PGDM) results with global bodybased results for final attribute prediction.

However, most of methods above are based on the weakly body regions information, which cannot localize human parts accurately. In this paper, we fused the body parts detection and attribute classification in a unified DCNN model to explore the implied relationship between these two different tasks. Thus more precise body information can be used to improve the performance of attribute classification.

3. HP-CNN ARCHITECTURE

3.1. Network design for simultaneous object detection and attribute classification

In this section, we present the proposed end-to-end network for simultaneous pedestrian body parts detection and multi-attribute classification. We use YOLOv2 [19], a real-time single object detector, as our backbone. Thirteen layers from Darknet19 were reformed to extract low-level features. Then three cascading convolutional layers were added to capture context features. So that low semantic information of shallower layers and contextual information of deeper layers can be integrated to precisely predict body parts of pedestrian. After that, squeeze-and-excitation blocks (SE-block) [20] structure is adapted to recalibrate important features for pedestrian attribute classification. The framework of HP-CNN is illustrated in Fig. 1.

We use zero padding at each convolutional layer so that the size of the feature maps is only changed by stride convolutions. The PReLU activation function [21] used in the network is defined in Eq. (1), where α is a learnable parameter. Each convolutional layer in the network is followed by an activation layer and batch normalization layer.

$$f(x) \quad = \quad \max\{0, x\} + \alpha * \min\{0, x\} \qquad (1)$$

SE-block: As shown in Fig. 2, SE-block is applied to improve the representational power of network by using global information to selectively emphasise informative features and suppress less useful ones. In each block, the scaled weighted

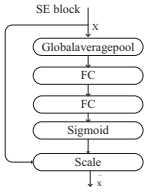

Fig. 2. The architecture of SE-block.

features of SE-layer and the original input features are fused by element-wise sum, and then pass mixed features to the next SE-Block to re-calibrates the original features in the channel dimension.

Multi-scale feature fusion (MSFF): As shown in Fig. 1, MSFF consists of two different operation of feature fusion. Firstly, the low-level features extracted with thirteen layers from Darknet19 are shared by related detection and classification tasks. Concatenation module uses learned weights to extract more useful target and the contextual features. In this paper, we concatenate the $conv5$, $conv8$, $conv13$ to obtain additional context information and fine-grained information for fine-grained attributes. After concatenation, a SE-block is also used to refine feature channels.

Furthermore, three 3×3 cascading convolutional layers are fused to incorporate global contextual information. After that, a kernel size of 1×1 convolutional layer is adopted to reduce the number of feature maps. Followed by 1×1 convolutional layer, we add a SE-block and use skip connection to pass multi-level features to higher layers, which can enhance the importance of the contextual information and integrate information flow. In our case, element-wise sum works better than concatenation model, because the latter may not learn the relationship between the targets and context well.

In the end, for attribute recognition, we add a fully connected layer (FC) and split the network into 51 separate branches for each attribute classification task.

It is obvious that some attributes are strongly relevant with location of body parts, e.g., female and long hair are relevant to head and shoulder areas; jacket and long-trouser usually appear in upper and lower body parts, respectively. Thus the parts object detected regions could enhance pedestrian attributes by joint learning with a unified network framework.

3.2. Multiple Loss Function

In this work, multiple loss functions are used for training the tasks of pedestrian body parts detection and multi-attribute classification simultaneously. The loss function of each task is described in detail as follows.

Pedestrian multi-attribute classification: The task of pedestrian multi-attribute classification uses binary cross-entropy loss function for each attribute, which is shown in Eq. (2). The total loss is the sum of binary cross-entropy losses of the 51 attributes.

$$L_{attr} = \frac{1}{N}\sum_{i=1}^{N}(-y_i^{truth} log y_i^{pred} - (1-y_i^{truth}) log(1-y_i^{pred})) \quad (2)$$

where N is the number of samples ($N = 8317$), y_i^{truth} is the ground truth of each attribute, y_i^{pred} is the probability of each attribute prediction.

Pedestrian body parts detection: We use L_1-smooth loss function [22] as the loss function of coordinate error and Intersection-over-Union (IoU) error, which is shown in Eq. (3), and use the multi-class cross-entropy as the loss function of classification error.

$$L_{loc}(t,p) = \sum_{i\in\{x,y,w,h\}} smooth_{L_1}(t_i, p_i) \quad (3)$$

in which

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, & if|x| < 1 \\ |x| - 0.5, & otherwise \end{cases} \quad (4)$$

The total loss is the sum of the losses of these two different tasks above. Here we set the weight ratio to 3:1 in order to force the network pay more attention to the detection task, which is more difficult to converge.

4. EXPERIMENTS

4.1. Dataset

To our knowledge, the Richly Annotated Pedestrian (RAP) dataset is the largest pedestrian attribute dataset so far, containing 41,585 samples annotated with 72 attributes. The reason we adopt the RAP dataset is that it is one of the few large-scale datasets provided both labeled body parts and attribute information. Following the same protocol as defined in [17], 51 binary attributes and three body parts are used for training and evaluating in our experiment.

4.2. Evaluation Protocols

Pedestrian multi-attribute classification: We adopt the example-based criteria [17] as evaluation metrics, which are defined as:

$$Accuracy_{exam} = \frac{1}{N}\sum_{i=1}^{L}\frac{|Y_i \bigcap f(x_i)|}{|Y_i \bigcup f(x_i)|} \quad (5)$$

$$Precision_{exam} = \frac{1}{N}\sum_{i=1}^{L}\frac{|Y_i \bigcap f(x_i)|}{|f(x_i)|} \quad (6)$$

$$Recall_{exam} = \frac{1}{N}\sum_{i=1}^{L}\frac{|Y_i \bigcap f(x_i)|}{|Y_i|} \quad (7)$$

$$F1 = \frac{2 * Precision_{exam} * Recall_{exam}}{Precision_{exam} + Recall_{exam}} \quad (8)$$

where N is the number of samples, Y_i is the ground-truth labels of attributes of the i^{th} sample, $f(x_i)$ is the predicted positive labels of attributes of the i^{th} sample, respectively. Compared to mean accuracy (mA) regarding as label-based solution [13], which is usually adopted to evaluate each attribute independently, example-based criteria reveals better inter-attribute correlation in our multi-attribute classification problem.

Pedestrian body parts detection: Standard metrics are used to evaluate our approach, including averaged precision (AP) over IoU thresholds which uses different IoU thresholds. IoU is the overlap ratio between the predicted bounding box and the ground truth or the ratio of their intersection and union. AP is the area under the precision-recall curve and means how many objects are correctly predicted. MAP is the average AP value for each category, which is between 0 and 1. The larger mAP value means better performance of detection. In this paper, IoU value is set to 0.5 to calculate the mAP.

4.3. Implementation Details and Parameter Settings

In the experiments, all the tested models were implemented by keras framework and python packages, and initialized with ImageNet pre-trained weights. Also, models were trained and tested on a NVIDIA GeForce GTX 1080Ti with 11GB memory.

Training strategies: For joint learning of pedestrian body parts detection and multi-attribute classification, HP-CNN was optimized by Adam with a weight decay of 1e-4, $beta_1$ of 0.9, $beta_2$ of 0.999 and epsilon of 1e-8. For each image in the dataset is zoomed to a fixed-size of 224×224. In order to make the training process stable, a small learning rate is used to train the network, which is 0.0001 at the beginning of training and will be exponentially degraded every 5200 iterations at a decay rate of 0.9. Batch size is set to 64.

Inference: In the testing stage, for each input image, the network produces multiple outputs including bounding boxes of body parts and the corresponding label for each attribute. For the detection task, we use non-maximum suppression (NMS) to refine the results, which select the corresponding bounding boxes meeting the maximum confidence rate higher than threshold 0.5.

4.4. Comparison with State-of-the-art Automatic Methods

In this subsection, we compare the results from HP-CNN with that from the state-of-the-art models on the RAP dataset. To better understand the effectiveness of proposed model, we trained a variant of HP-CNN such as: 1)HP-CNN without multi-task learning only for multi-label attribute classification (MAC) or body parts detection (BPD), denoted as HP-MAC and HP-BPD, respectively; 2)HP-CNN without the SE-block or multi-scale feature fusion mentioned in Section 3, denoted as HP-CNN (no SE-block) and HP-CNN (no MSFF), respectively; 3)HP-CNN combines both classification and detection tasks by sharing the same backbone networks after $conv20$ in Fig. 1, denoted as HP-CNN (Conv20).

The performance of classification: We report the attribute classification results by testing the state-of-the-art models: 1) SVM classifiers for each attribute with CNN features, including ELF-mm, FC7-mm, FC6-mm [23, 24]; 2) The end-to-end DCNNs based on joint multi-label attributes learning, including ACN [6], DeepMAR [14], DeepMAR* [17]; 3) DCNNs based on multi-label attributes learning with body parts information generated by weakly-supervised localization or attention modules, including WPAL-network [10], HPlus-Net [8], and PGDM-Fusion [18]. The details of results are shown in Table 1.

As shown in Table 1, for the group of a variant of HP-CNN, the multi-tasks based HP-CNN methods generally outperform HP-MAC models. For instance, HP-CNN significantly outperforms HP-MAC by 4.69% and 3.59% in terms of Acc and F1, respectively. It gives the evidence that the additional detection task provides helpful information for identifying pedestrian attributes. It is also noted that HP-CNN (Conv20) produced slightly worse results, which demonstrated that classification and detection tasks would interfere with each other when sharing too much network parameters. It is worth pointing out that HP-CNN with both SE-block and MSFF modules achieved the best results.

Comparing with the state-of-the-art models, We can find the multi-attribute recognition models generally performed much better than the single-label trained methods such as ELF-mm, FC7-mm and FC6-mm, and DCNNs assembled with attention based body parts information, including HPlus-Net and PGDM-Fusion, produced relatively better recognition results. Moreover, the proposed HP-CNN integrates multi-attribute recognition and part detection, which can obtain more precise body information, achieved the highest accuracy and precision values along with the comparable recall, thus yield the best comprehensive F1 value.

The performance of detection: Turning to the detection task, the experimental results of detection accuracy are shown in Table 2 in terms of mAP with IoU thresholds (0.5), where bold faces indicate the highest results. The baseline model is YOLOv2 [19], which is the backbone network in our model.

It can be seen from Table 2 the results are all promise for body parts detection, thus leading to a narrow gap among the tested models. It is noted that HP-BPD can work well as a detector with a competitive mAP of 97.62%, beating the YOLOv2 baseline. When combining with attribute classification tasks, HP-CNN achieved the highest mAP of 97.88%, which has a sight improvement over the HP-CNN (no SE-block) and HP-CNN (no MSFF). It demonstrates that multi-scale feature fusion and SE-block can boost the performance of detection.

Fig. 3 shows several qualitative results of our method on

the RAP dataset. The content below each image is the prediction results of corresponding attributes, such as glasses, hats, hair, etc., while different colors of bounding boxes show the body parts detection results of pedestrian. As we can see from Fig. 3, HP-CNN performs well on both attribute classification and body parts detection.

Table 1. The pedestrian attribute classification performance of 10 benchmark algorithms list in [8] and single-class model mentioned, multi-task model and multi-task model without the innovations mentioned in this work, which evaluated on RAP dataset using example-based evaluation criteria.

Methods	Acc	Prec	Rec	F1
ELF-mm [17]	29.29	32.84	71.18	44.95
FC7-mm [17]	31.72	35.75	71.78	47.73
FC6-mm [17]	33.37	37.57	73.23	49.66
ACN [6]	62.61	80.12	72.26	75.98
DeepMAR [14]	63.67	76.53	77.47	77.00
DeepMAR* [17]	62.02	74.92	76.21	75.56
WPAL-FSPP [10]	53.30	60.82	**78.80**	68.65
HPlus-Net [8]	65.39	77.33	78.79	78.05
PGDM-Fusion [18]	64.57	78.86	75.90	77.35
HP-MAC(no SE-block)	60.03	77.90	70.69	74.12
HP-MAC	61.18	78.89	71.24	74.87
HP-CNN(no SE-block)	65.38	82.15	74.45	78.11
HP-CNN(no MSFF)	65.75	82.16	74.91	78.37
HP-CNN(Conv20)	63.88	81.29	73.22	77.05
HP-CNN	**65.87**	**82.51**	74.79	**78.46**

Table 2. Pedestrian body parts detection evaluation results of the proposed single detection network, multi-task model, multi-task model without the innovations mentioned, and YOLOv2 on the RAP dataset.

Methods	mAP	Head	Top	Down
YOLOv2 [19]	97.59	96.11	**99.45**	97.20
HP-BPD	97.62	96.62	99.27	96.98
HP-CNN(no SE-block)	97.73	96.63	99.38	97.19
HP-CNN(no MSFF)	97.68	96.65	99.26	97.12
HP-CNN(Conv20)	96.44	95.41	98.08	95.85
HP-CNN	**97.88**	**96.69**	99.32	**97.62**

5. CONCLUSION

In this paper, a novel multi-task network called HP-CNN is proposed for integrating the pedestrian attribute classification and body parts detection. Also, two different operations of feature fusion and SE-Blocks are applied to fuse semantic information of different layers and strengthen the sensitivity of CNN features. Experimental evaluations on the large-scale attribute dataset demonstrate that the effectiveness of the proposed HP-CNN compared with the state-of-the-art multi-label DCNN methods. Especially, the auxiliary task of body parts detection provided a great advantage in improving the performance of attribute classification.

In the future, we will investigate more additional pedestrian information in our model such as body joints and keypoints, along with further validation of this study by testing more attribute classification models on more real-world datasets.

6. REFERENCES

[1] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang, "Svdnet for pedestrian retrieval," in *IEEE International Conference on Computer Vision*, 2017, pp. 3820–3828.

[2] Annan Li, Luoqi Liu, Kang Wang, Si Liu, and Shuicheng Yan, "Clothing attributes assisted person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 869–878, 2015.

[3] Peixi Peng, Yonghong Tian, Tao Xiang, Yaowei Wang, and Tiejun Huang, "Joint learning of semantic and latent attributes," in *European Conference on Computer Vision*, 2016, pp. 336–353.

[4] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian, "Multi-type attributes driven multi-camera person re-identification," *Pattern Recognition*, vol. 75, pp. 77–89, 2018.

[5] David Hall and Pietro Perona, "Fine-grained classification of pedestrians in video: Benchmark and state of the art," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 5482–5491.

[6] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe, "Person attribute recognition with a jointly-trained holistic cnn model," in *IEEE International Conference on Computer Vision Workshop*, 2015, pp. 329–337.

[7] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev, "Panda: Pose aligned networks for deep attribute modeling," in *IEEE conference on computer vision and pattern recognition*, 2014, pp. 1637–1644.

[8] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2017, pp. 350–359.

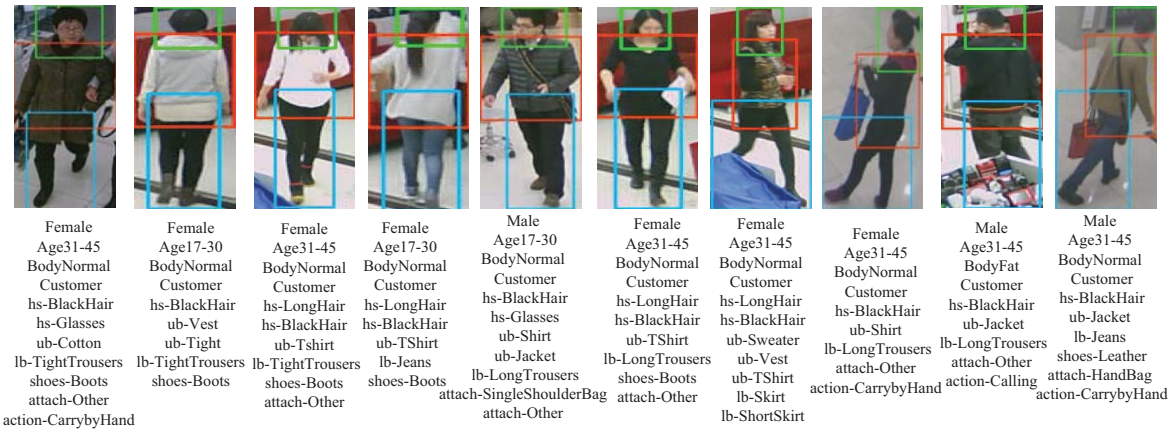

Fig. 3. Some pedestrian images from the RAP dataset with detection and classification results by HP-CNN.

[9] Jianqing Zhu, Shengcai Liao, Zhen Lei, and Stan Z. Li, "Multi-label convolutional neural network based pedestrian attributeclassification," *Image and Vision Computing*, vol. 58, no. C, pp. 224–229, 2017.

[10] Kai Yu, Biao Leng, Zhang Zhang, Dangwei Li, and Kaiqi Huang, "Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization," *arXiv preprint arXiv:1611.05603*, 2016.

[11] Ryan Layne, Timothy Hospedales, and Shaogang Gong, "Towards person identification and re-identification with attributes," in *European Conference on Computer Vision*. Springer, 2012, pp. 402–412.

[12] Jianqing Zhu, Shengcai Liao, Zhen Lei, Dong Yi, and Stan Z. Li, "Pedestrian attribute classification in surveillance: Database and evaluation," in *IEEE International Conference on Computer Vision Workshops*, 2013, pp. 331–338.

[13] 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*. ACM, 2014, pp. 789–792.

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

[15] G. Sharma, F. Jurie, and C. Schmid, "Expanded parts model for human attribute and action recognition in still images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 652–659.

[16] Jianqing Zhu, Shengcai Liao, Dong Yi, Zhen Lei, and Stan Z. Li, "Multi-label cnn based pedestrian attribute learning for soft biometrics," in *International Conference on Biometrics*, 2015, pp. 535–540.

[17] 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.

[18] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang, "Pose guided deep model for pedestrian attribute recognition in surveillance scenarios," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2018, pp. 1–6.

[19] Joseph Redmon and Ali Farhadi, "Yolo9000: Better, faster, stronger," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6517–6525.

[20] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, 2017.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *International Conference on Neural Information Processing Systems*, 2015, pp. 91–99.

[23] Douglas Gray and Hai Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European conference on computer vision*. Springer, 2008, pp. 262–275.

[24] Bryan James Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q Mary, "Person re-identification by support vector ranking.," in *BMVC*, 2010, vol. 2, pp. 687–691.