PENTADENT-NET: PEDESTRIAN ATTRIBUTE RECOGNITION WITH DISTANCE REFINEMENT AND CORRELATION MINING

Yuan Liu[1,2*] *Maoqing Tian*[2] *Jun Hou*[2] *Shuai Yi*[2] *Zhiping Lin*[1]

[1]School of EEE, Nanyang Technological University, Singapore [2]SenseTime Group Limited

ABSTRACT

Pedestrian attribute recognition aims to accurately identify attributes from pedestrian images. Visually similar pedestrian images tend to share similar attributes, and the upper and lower clothing attributes normally match in style. However, how to effectively capture these relations remains a challenging topic. In this paper, we propose a novel network Pentadent-Net (PD-Net). The network consists of a "Representation Refining Branch (RRB)" and an "Attribute Fusion Branch (AFB)". The RRB reduces the feature distances among visually similar images, and the AFB generates a joint-representation for each attribute based on its correlation with other attributes. The polished features from the two branches are then applied to capture the inter-sample relations and attribute dependencies respectively. We also further associate features from the two branches to enrich feature representations. Experimental results on two benchmark datasets (PETA and PA-100K) by our proposed approach demonstrate its advantages over state-of-the-art methods.

Index Terms— Pedestrian Attribute Recognition, Convolutional Neural Network, Graphical Model, Computer Vision, Deep Learning

1. INTRODUCTION

Pedestrian attribute recognition has attracted significant attention from both research and industry due to its wide applications in surveillance, retrieval and behavioural analysis. Many large-scale annotated datasets such as PETA [1], PA-100K [2], PARSE27K [3] and RAP [4] have been released to facilitate the research [5].

With the advancement of Deep Learning in recent years, features extracted from the Convolutional Neural Networks (CNN) have been widely adopted for the task. Li *et al.* [6] extracted pedestrian features with CaffeNet and proposed a new loss function for unbalanced attribute contents. Sudowe *et al.* [3] used a jointly trained CNN model to learn different attribute features simultaneously. The work in [7] proposed to jointly perform attribute localisation and recognition by spatial pyramid pooling. Zhao *et al.* [2] proposed to adopt multi-scale attention masks to acquire abundant attribute features. Part based models [8] [9] [10] have been proposed to predict attributes with detailed features in local parts. He *et al.* [11] split attributes into groups and predicted each group in a multi-task manner with a weighted loss depending on attribute performance. While these CNN-based methods generally provided better performance, they did not capture the relations among visually similar samples nor explicitly explore dependencies among attributes.

To explore the relations among attributes, works [12] [13] [14] have been proposed to adopt Recurrent Neural Networks (RNN) to predict attributes in sequence. However, these RNN-based methods heavily rely on the human-defined prediction sequence which could be challenging to optimise. Some graphical methods have been proposed to mine the relations between attribute pairs. Chen *et al.* [15] estimated the pair-wise conditional probabilities and predicted scores with SVM classifiers. However, conditional probabilities might encounter bias issue with low frequency attributes. Park *et al.* [16] proposed an and-or structured graph to capture attribute dependencies. VSGR [17] applied Graph Convolutional Network (GCN) to capture spatial and semantic information and predict attributes in a strictly defined sequence. However, both methods ignored relations among certain attributes.

Fig. 1. Examples of failure cases.

Intuitively, visually similar images normally share similar attribute contents. Hence, these image feature representations should also be similar to facilitate more accurate attribute prediction. However, existing methods normally ignore this relationship and only focus on the input image itself. As shown in Figure 1 (left), two similar pedestrian images contain similar attributes. However, the recent "Inception-enhanced DeepMAR" [12] still incorrectly yields different predictions for two attributes. In addition, attributes have natural correlations with each other. For example, formal shirts normally match with formal pants and short-sleeves normally match with short pants. However, as shown in Figure 1 (right), the "Inception-enhanced DeepMAR" method ignores this correlation and produces unsatisfying results.

To effectively reduce representation distance among visually similar images and better explore the correlation between attributes, we propose a novel network Pentadent-Net (PD-Net). PD-Net has two parallel branches, namely the "Representation Refining Branch" (RRB) and the "Attribute Fusion Branch" (AFB). The RRB reduces the representation distance among visually similar images. The AFB generates a joint-representation for each attribute based on its correlation with other attributes. The whole network leverages on the information from both branches and is able to predict attributes with better features and abundant details. The contributions of this

*The first author performed the work while interning at SenseTime Group Limited.

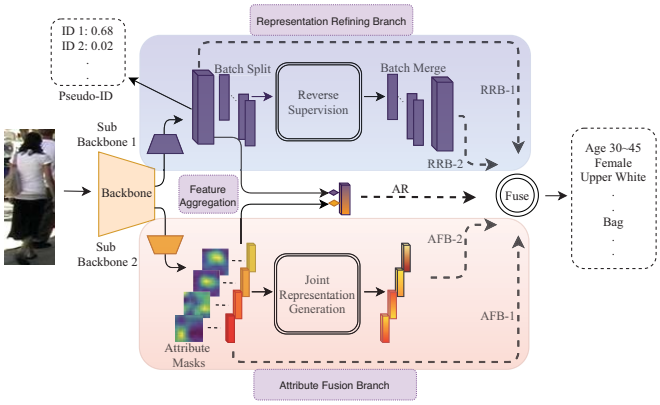

Fig. 2. The overall architecture of the proposed PD-Net. We adopt the Inception V3 [18] architecture as our backbone, and split the backbone from layer Mixed 7a onwards to form two sub-backbones with the same structure following the original architecture. The confidence vectors from RRB, AFB and Feature Aggregation are fused to generate the final prediction.

work can be summarised as follows:

- A novel "Representation Refining Branch" (RRB) to reduce representation distance among visually similar samples, and hence increase the likelihood of their attribute prediction probabilities.
- A novel "Attribute Fusion Branch" (AFB) to mine natural dependencies among attributes. When predicting an attribute, other attributes that are highly correlated to it are taken into consideration to jointly generate the result.
- We demonstrate competitive performance against other existing methods and achieve state-of-the-art results on two large scale pedestrian attribute datasets PETA and PA-100K.

2. PROPOSED APPROACH

2.1. Overall Architecture

The structure of the proposed PD-Net is shown in Figure 2. The RRB branch and the AFB branch are connected on top of the backbone feature extractor to refine representation distance and mine attribute correlation. We also explicitly merge the information from the two branches by introducing a "Feature Aggregation Module" (FAM) to further improve the performance. Both RRB and AFB produce two confidence vectors at different stages and FAM generates another confidence vector. The five vectors are fused together at the end to yield the final prediction. The details of RRB, AFB and FAM are explained in the following subsections.

2.2. Representation Refinement for Visually Similar Images

Visually similar images normally share similar attribute contents. By enforcing a similar feature representation among these images, the classifier is able to yield closer attribute predictions to improve the model performance. In the RRB, we reduce similar samples' representation distance via two steps. Firstly, we generate "Pseudo ID" and perform a Person Re-ID subtask to coarsely refine representations. Secondly, we further fine-tune the representation by "reverse supervision".

The learning objective of Re-ID is to reduce distance among representations of the same identity. Hence, if we consider visually similar samples as the same identity, performing Re-ID and attribute recognition with the same features will help to reduce their distance effectively. However, as we do not have ID annotations, we propose to generate Pseudo ID by clustering. Specifically, we apply a Re-ID model pre-trained with major Re-ID datasets [19] [20] [21] [22] and loss in [23] to extract a 256 dimensional vector from each image in the training dataset, then perform clustering to generate Pseudo ID. Some images with the same Pseudo ID on PA-100K and PETA are shown in Figure 3. Note that we adopt different clustering algorithm for the two datasets, K-Means on PETA and K-NN on PA-100K, which shows greater performance empirically. During training of the attribute model, we connect an ID classifier to the representation extracted from the sub-backbone 1 and train it with the generated Pseudo ID to coarsely refine representation distance.

As the Pseudo ID information is not ground-truth, it might not be sufficient to accurately reduce similar image feature distance. Hence, as the second step, we introduce a graphical approach to reversely supervise the image representations. Given an image I_a, assume there is another image I_b which is coarsely deemed similar to I_a. If we add I_b's feature onto I_a weighted by their cosine similarity and generate a new feature for I_a, there could be two possible scenarios. In the first case, the two images are indeed similar, by adding their coarsely refined features together the attribute prediction for I_a should be improved as they complement each other. In the second case, the two images are not similar and the coarse refinement is incorrect, which could incur large noise when adding the features and degrade attribute performance. To reduce training loss during back-propagation, the network will further reduce the distances between their features in the first case and further increase their feature distance in the second case. We denote this operation as "reverse supervision" that fine-tunes the image representations.

To implement the discussion above and perform reverse supervision, we split the features from sub-backbone 1 batch-wise and feed the whole batch into a Graph Convolution Network (GCN) with propagation rule defined in [24]. Note that each sample's representation within a batch now becomes one of the graph nodes. The cosine similarities between these nodes defined in Equation (1) are calculated online for each batch as the edge weight of the graph.

$$cos\_sim = \frac{(F_{r1})^T F_{r2}}{max(\|F_{r1}\| \times \|F_{r2}\|, \epsilon)}, \tag{1}$$

where F_{r1} and F_{r2} are a pair of RRB feature vectors in a batch, T means the transposition of the matrix, $\|\|$ is vector magnitude and ϵ is a small positive value used to avoid division by zero.

We set a threshold to filter out graph edges with small values to avoid noisy graph operation, as well as performing softmax to all rows after filtering. We adopt two classifiers to generate attribute predictions using features before and after reverse supervision, and name their outputs as "RRB-1" and "RRB-2" accordingly. The two attribute predictions are both supervised by attribute recognition loss L_{attr} which adopts the improved sigmoid cross entropy loss in [6].

$$L_{attr} = -\frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{M} w_i(y_{ki}log(p_{ki}) + (1-y_{ki})log(1-p_{ki})), \tag{2}$$

$$w_i = exp(-p_i/\rho^2), \tag{3}$$

where N, M represent the batchsize and the total number of attributes, p_{ki} and y_{ki} are the predicted probability and ground-truth

Fig. 3. Examples of the images that are assigned the same ID label through clustering for PA-100K (left) and PETA (right)

label for the i^{th} attribute in sample k respectively, w_i is the loss weight for the i^{th} attribute, p_i is the positive ratio of the i^{th} attribute calculated from the training set and ρ is a parameter which we set to 1 according to the original work. We adopt cross entropy loss denoted by L_{id} for the Re-ID sub-task.

To avoid dependency on batch, edges of the reverse supervision graph are replaced by an identity matrix during inference, where each node only connects to itself.

2.3. Attribute Fusion via Joint Representation

The AFB mines the relations among attributes, and generates a joint-representation for each attribute based on its correlation with other attributes. In this way, even in corner-cases where a specific attribute feature is affected by occlusion or poor lighting conditions, the model is still able to yield robust prediction results by exploring its dependencies on other attributes. In AFB, we firstly extract a feature representation for each attribute, then connects the highly-correlated pairs using prior knowledge calculated from the training data to generate joint-representations. The details of each step are described as follows.

We adopt a simple yet effective mechanism in [2] to extract feature for each attribute, which uses 1×1 convolution to generate activation masks for all attributes. The equation for activation mask generation is formulated as

$$F'_{Ai} = GAP(F \odot \{S(C_i(F, \theta_i))\}), i = 1, 2, ..., M, \qquad (4)$$

where F represents the feature map extracted from the sub-backbone 2, C_i and θ_i represent the i^{th} channel of a 1×1 convolution layer and its parameters, S represents the sigmoid activation function, $\odot$ represent the element wise multiplication operation, GAP represents the Global Average Pooling, F'_{Ai} represents the feature for the i^{th} attribute and M is the total number of attributes. We generate an activation mask for each attribute, and perform element-wise masking followed by pooling to obtain a feature representation for each attribute.

After obtaining the attribute representations, we explore the correlation of all attributes by calculating pair-wise Point-biserial Correlation Coefficients offline using training data. A relatively large positive correlation value indicates a strong dependency, which suggests the two attributes tend to appear simultaneously. To generate the joint-representation for an attribute, we select all its highly-corrected attributes and add their features element-wisely to the original attribute representation, weighted by the calculated correlation. In this way, when predicting this attribute, the classifier is able to not only focus on its feature, but also considering the dependencies of it on other relevant attributes.

Specifically, we implement the joint-representation generation process using another GCN with the pair-wise attribute correlation as edge weights. We filter out weak connections on the graph and only keep highly positive-correlated attributes connected. A softmax operation is applied to each row of the edge weight matrix to balance the numbers.

Similar to RRB, attribute prediction is performed using both the original attribute features and the joint-representations. We name the attributes predicted from the original attribute features as "AFB-1" and the joint-representation as "AFB-2". Note that as each attribute has its own feature representation, we have M fully-connected classifiers within a set for M attributes. We adopt two sets of classifiers to predict attributes from these two sets of features. Similar to the RRB, attribute recognition loss in Equation (2) and (3) is adopted. Note that the edge weight of the graph remains unchanged during inference as it is calculated offline.

2.4. Feature Aggregation Module

To further boost the performance of the PD-Net, we blend the information from the RRB and AFB through feature aggregation. As the representation distributions differ significantly from the two parallel branches, directly adding them yields unstable performance empirically. Hence, we transform these features linearly before adding them together. Specifically, we firstly pass the feature vectors from both branches through fully-connected layers to perform transformation. Then, we aggregate all transformed features from AFB element-wisely to form a single vector, and perform element-wise summation again with the transformed feature vector from RRB, which effectively generates the aggregated representation,

$$\begin{cases} F_{Ai_t} = W_{Ai}F_{Ai} + B_{Ai}, \\ F_{R_t} = W_R F_R + B_R, \\ F_a = F_{R_t} + \sum_{i=1}^{M} F_{Ai_t}, \end{cases} \qquad (5)$$

where F_{Ai}, W_{Ai}, B_{Ai} represent the i^{th} original attribute feature from AFB and its transformation layer's parameters. F_R, W_R, B_R represent the coarsely refined feature from RRB and its transformation layer's parameters. F_{Ai_t} represents the transformed attribute representation for the i^{th} attribute, F_{R_t} represents the transformed coarsely refined feature and F_a represents the Aggregated Representation (AR). We use a fully-connected classifier to estimate attribute confidences for all attributes from the AR and adopt the same attribute loss L_{attr} to supervise its training.

2.5. Training and Inference

The whole network is trained in an end-to-end manner, the total loss during training can be formulated as,

$$L = L_{id} + \sum_{l} L^l_{attr}, \qquad (6)$$

where L^l_{attr} represents the loss from attributes predicted with features l, and l includes the five feature representations from RRB, AFB and FAM. During inference, we abandon the Re-ID sub-task and only compute the attribute recognition results. We take the five prediction probabilities and average them as the final recognition result.

3. EXPERIMENTS

3.1. Datasets

We perform experiment on two large-scale datasets for pedestrian attribute recognition.

Table 1. Comparison with State-of-the-Art methods on PA-100K and PETA datasets. * Denotes ensemble-based methods.

	PA-100K					PETA				
Method	Ins Acc	Ins Pre	Ins Rec	Ins F1	Label Acc	Ins Acc	Ins Pre	Ins Rec	Ins F1	Label Acc
PG-Net	73.08	84.36	82.24	83.29	74.95	78.08	**86.86**	84.68	85.76	82.97
DeepMAR	70.39	82.24	80.42	81.32	72.70	75.07	83.68	83.14	83.41	82.60
GRL	-	-	-	-	-	-	84.34	**88.82**	86.51	**86.70**
M-Net	70.44	81.70	81.05	81.38	72.30	-	-	-	-	
HP-Net	72.19	82.97	82.09	82.53	74.21	76.13	84.92	83.24	84.07	81.77
VAA	-	-	-	-	-	**78.56**	86.79	86.12	86.46	84.59
LG-Net	**75.55**	**86.99**	**83.17**	**85.04**	**76.96**	-	-	-	-	
RCRA(RC)	-	-	-	-	-	-	85.42	88.02	**86.70**	85.78
RCRA(RA)	-	-	-	-	-	-	84.69	**88.51**	86.56	**86.11**
PD-Net(Ours)	**78.80**	**87.50**	**86.91**	**87.20**	**80.40**	**79.79**	**87.99**	86.16	**87.07**	84.85
JRL*	-	-	-	-	-	-	86.03	85.34	85.42	**85.67**
VSGR*	80.58	89.40	87.15	88.26	79.52	**81.82**	88.43	**88.42**	88.42	85.21
PD-Net(Ours)*	**81.21**	**90.05**	**87.62**	**88.82**	**80.11**	**81.82**	**90.17**	87.04	**88.58**	85.28

PEdesTrian Attribute (PETA) dataset contains 19000 images and 35 binary labels to perform the task. The dataset is split into three non-overlapping partitions for training, validation and testing, which contain 9500, 1900 and 7600 images, respectively.

PA-100K is the current largest pedestrian attribute dataset, which contains 100000 images collected from outdoor scenes. It contains 26 binary labels and is split into 80000, 10000 and 10000 images for training, validation and testing.

3.2. Implementation Details

We resize all input images to 299×299 pixels. Adam optimisation [25] is adopted with an initial learning rate of 0.0001. Data augmentation techniques including flipping, random cropping and rotation are adopted during training. We apply learning rate decay of 0.5 for every 5000 iterations. The batch size is set to 32 and we train our model for 25000 iterations. The network is implemented using PyTorch and trained on Nvidia V100 Graphic Card.

3.3. Comparison with the State-of-the-Arts

We compare our method against 6 competitors including M-Net [2], DeepMAR [6], HP-Net [2], PG-Net [9], LG-Net [10] and VSGR [17] on PA-100K dataset and 8 competitors including DeepMAR, HP-Net, JRL [14], PG-Net, VAA [26], GRL [12], RCRA [13] and VSGR on PETA dataset. We follow the commonly applied evaluation metrics in [2] to compare the performances. The results are shown in Table 1, where we bold the top two performance for single model and top one for ensemble methods. To showcase the improvement of the network design alone, we do not compare with works like [11] that adopt backbone tricks such as [27] and [28].

As shown in the table, our method outperforms its competitors by a large margin on all five metrics on the PA-100K dataset, which is the current largest pedestrian attribute dataset consists of many hard-cases including occlusion and view-point limitations. For PETA dataset, the PD-Net outperforms all its competitors on Instance-based metrics, which is commonly considered as the primary evaluation criterion as they describe the single-image level prediction performance [29].

Although multi-model ensemble methods are not our primary interest due to their huge computation costs during inference, we still compare our proposed PD-Net with these methods via a similar approach. Specifically, in order to perform fair comparisons with 10-model-ensemble methods, we train 10 PD-Nets with slight variations in input sizes, graph depths and backbones to perform voting, which surpasses all its ensemble-based competitors as shown in the table highlighted with *.

3.4. Ablation Study on Network Components

We re-train DeepMAR model with Inception-V3 backbone as the baseline (B). To showcase the effectiveness of each proposed module, we add our network components one by one onto the baseline and present their performances on PETA in Tables 2. Note that we do not compare precision and recall as they normally share a negative correlation with each other. A represents AFB, R represents RRB and F represents FAM. The results present steady improvement when adding network branches onto the baseline, which shows the effectiveness of individual branches quantitatively.

Table 2. Performance of branches on PETA

Method	Ins Acc	Ins F1	Label Acc
B (DeepMAR)	78.42	86.06	83.94
B + A	79.15	86.59	84.08
B + R	79.37	86.73	84.69
B + A + R	79.45	86.82	84.20
B + A + R + F (PD-Net)	**79.79**	**87.07**	**84.85**

4. CONCLUSION

In this paper, we present a novel parallel structured PD-Net that improves pedestrian attribute recognition performance by reducing feature distance among visually similar images and mining the natural dependencies between attributes. Extensive experiments and ablation studies are conducted to showcase the effectiveness of our network as well as the individual components. Our method outperforms other state-of-the-art methods on two large-scale pedestrian attribute datasets PETA and PA-100K.

5. REFERENCES

[1] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Pedestrian attribute recognition at far distance," *Proceedings of the ACM International Conference on Multimedia - MM 14*, 2014.

[2] 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," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[3] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe, "Person attribute recognition with a jointly-trained holistic CNN model," *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015.

[4] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang, "A richly annotated dataset for pedestrian attribute recognition," *ArXiv*, vol. abs/1603.07054, 2016.

[5] Xiao Wang, Shaofei Zheng, Rui Yang, Bin Luo, and Jin Tang, "Pedestrian attribute recognition: A survey," *ArXiv*, vol. abs/1901.07474, 2019.

[6] Dangwei Li, Xiaotang Chen, and Kaiqi Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015.

[7] Yang Zhou, Kai Yu, Biao Leng, Zhang Zhang, Dangwei Li, and Kaiqi Huang, "Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization," *Procedings of the British Machine Vision Conference 2017*, 2017.

[8] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir D. Bourdev, "Panda: Pose aligned networks for deep attribute modeling," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1637–1644, 2013.

[9] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang, "Pose guided deep model for pedestrian attribute recognition in surveillance scenarios," *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 2018.

[10] Pengze Liu, Xihui Liu, Junjie Yan, and Jing Shao, "Localization guided learning for pedestrian attribute recognition," in *BMVC*, 2018.

[11] X. He, Q. Shi, F. Su, Z. Zhao, and B. Zhuang, "Pedestrian attribute recognition based on mtcnn with online batch weighted loss," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sep. 2019, pp. 2461–2465.

[12] Xin Zhao, Liufang Sang, Guiguang Ding, Yuchen Guo, and Xiaoming Jin, "Grouping attribute recognition for pedestrian with joint recurrent learning," *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.

[13] Xin Zhao, Liufang Sang, Guiguang Ding, Jungong Han, Na Di, and Chenggang Yan, "Recurrent attention model for pedestrian attribute recognition," *AAAI*, 2019.

[14] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li, "Attribute recognition by joint recurrent learning of context and correlation," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[15] Huizhong Chen, Andrew Gallagher, and Bernd Girod, "Describing clothing by semantic attributes," *Computer Vision – ECCV 2012 Lecture Notes in Computer Science*, p. 609–623, 2012.

[16] Seyoung Park, Bruce Xiaohan Nie, and Song-Chun Zhu, "Attribute and-or grammar for joint parsing of human pose, parts and attributes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 7, pp. 1555–1569, 2018.

[17] Qiaozhe Li, Xin Zhao, Ran He, and Kaiqi Huang, "Visual-semantic graph reasoning for pedestrian attribute recognition," *AAAI*, 2019.

[18] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2015.

[19] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1116–1124.

[20] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[21] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014.

[22] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian, "Person transfer GAN to bridge domain gap for person re-identification," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 79–88, 2017.

[23] H. J. Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu, "Cosface: Large margin cosine loss for deep face recognition," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, 2018.

[24] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[25] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[26] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *ECCV*, 2018.

[27] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2017.

[28] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5987–5995, 2016.

[29] M. Zhang and Z. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, Aug 2014.