ENHANCE VIA DECOUPLING: IMPROVING MULTI-LABEL CLASSIFIERS WITH VARIATIONAL FEATURE AUGMENTATION

Ming Chen[1], *Guijin Wang*[1,2,*], *Jing-Hao Xue*[3], *Zijian Ding*[1], *Li Sun*[1]

[1]Department of Electronic Engineering, Tsinghua University,
[2]Beijing National Research Center for Information Science and Technology,
[3]Department of Statistical Science, University College London

ABSTRACT

Multi-label classification remains a challenging problem due to the inherent label imbalance issue, which brings over-fitting of minor categories to modern deep models. In this paper, to tackle this issue, we propose a novel method named Variational Feature Augmentation (VFA) to enhance the deep neural networks for multi-label classification. Our method decouples the feature vectors extracted by the backbone network into multiple low-dimensional spaces via a novely proposed Variational Feature Decoupling Module. The decoupled feature vectors are then re-combined with a shuffle operation and a Feature Augmentation Layer to enrich the minor co-occurrence relations, mitigating the label imbalance. Different from most other methods, VFA does not modify the network architecture or introduce extra computation cost in inference phase. We conduct comprehensive experiments on four benchmarks of two visual multi-label classification tasks, pedestrian attribute recognition and multi-label image recognition, and the results demonstrate the effectiveness and generality of the proposed VFA.

Index Terms— Deep Learning, Multi-Label Classification, Pedestrian Attribute Recognition

1. INTRODUCTION

Multi-label classification is the task of assigning multiple concepts or attributes to a instance. For example, a natural image generally contains multiple objects. One essential issue of multi-label classification is the inherent label imbalance, which causes over-fitting of minor categories and cannot be well handled by common resampling strategies [1] due to the label co-occurrence.

Many deep networks have made progress for multi-label classification tasks [2, 3, 4, 5, 6]. Impressive success was achieved by exploting label correlation via graph-based models [3, 4]. Other approaches extracted attentional regions in image recognition and pedestrian attribute recognition (PAR) problems [5, 6]. However, most of these methods utilize heavy architectures and ignore the essential label imbalance issue. The most general approach in recent multi-label models is to use the binary cross-entropy (BCE) loss function with class-specific re-weighting to balance the conributions of different classes. However, such simple methods often result in limited improvement due to the label co-occurrence and the dominance of negative labels [7]. To address these problems, several recent works [7, 8] have attempted to modify the commonly used BCE loss function. [7] suggested a Distribution-Balanced loss to re-balance the weights taking into account label co-occurrence. [8] proposed Asymmetric Loss (ASL) to asymmetrically focusing on positive and negative samples considering the high negative-positive imbalance in the case of multi-label classification. These methods have achieved the state-of-the-art performances on the mainstream benchmarks like MS-COCO [9]. However, there are still some drawbacks for these methods. E.g, the hyper-parameters in these methods require elaborate tuning for different datasets or tasks. Additionally, most methods have not verified their generality to other tasks. Some methods use the idea of decoupling features. [6] equipped a attention module for each attribute and conducted binary classification independently in PAR task. However, they introduced extra attention modules for all attributes, increasing the amount of parameters and computation cost.

To tackle the label imbalance issue, in this paper, we propose a novel Variaional Feature Augmentation (VFA) method to improve the performance of multi-label classifiers, which performs label re-balancing in the class-wise decoupled feature spaces. VFA is composed of three components. First, a Variational Feature Decoupling Module decouples the feature vectors extracted by backbones into multiple low-dimentional groups of sub-feature vectors, each group responding to one category. The conventional resampling strategies fail in multi-label cases due to the label co-occurrence issue. In this step, the co-occurrence of labels is decoupled for further operations. Secondly, for each group of sub-feature vectors, we independently shuffle the permutation of the batch and reconsititute the co-occurrence of labels to mitigate the label imbalance. Thirdly, a Feature Aggregation Module combines the sub-feature vectors and reproject them into the original

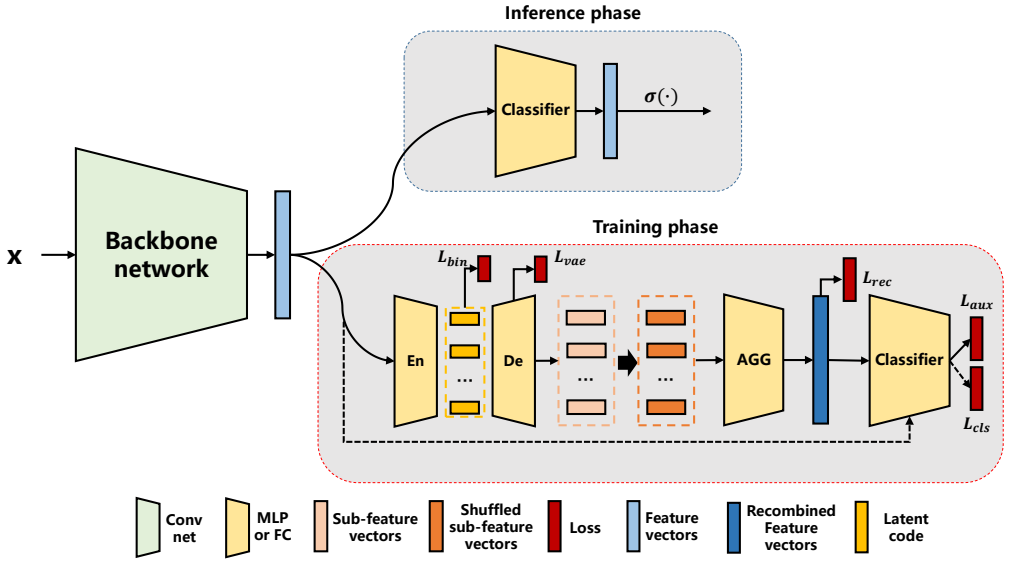

Fig. 1. The framework of Decoupled Feature Augmentation (best viewed in color). In the training phase, the feature vectors extracted by the backbone network are decoupled into multiple sub-feature spaces by the Feature Decoupling (FD) module. The decoupled sub-feature vectors are then shuffled and re-combined to constitute the augmented feature vectors and refine the multi-label classifier. In the inference phase, all the extra modules and opperations are removed to maintain the original architecture.

feature space. The re-combined feature vectors are used to update the classifiers together with the original ones. VFA only introduces extra structures in the training phase and is cost-free in the inference phase comparing with baseline models. We evaluate VFA on four benchmarks of two visual multi-label classification tasks, pedestrian attribute recognition and multi-label image classification. VFA gains consistent improvement and achieves competitive or surpassing performances comparing with other state-of-the-arts.

2. METHOD

For multi-label classification, we first illustrate the symbols and variables used in the following part of the paper. Let $\mathcal{X}$, $\mathcal{V} = \mathbb{R}^D$ and $\mathcal{Y} = \{0, 1\}^C$ denote the instance space, feature space and label space, where D and C are the dimensionality of feature space of the output of backbones and label space, respectively. Given instance-label pairs $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$, the backbone project the instances $\{\mathbf{x}_i\}_{i=1}^N$ into feature vectors $\{\mathbf{v}_i\}_{i=1}^N$, which the classifier predict the labels according to. The end-to-end deep network can be represented as two stages: $f : \mathcal{X} \to \mathcal{V}$ and $h : \mathcal{V} \to \mathcal{Y}$, and the goal of multi-label learning is to learn the backbone f and classifier h.

2.1. Overview of Variational Feature Augmentation

Different from previous methods focusing on enhancing the feature learning with additional modules, our goal is to improve the classifier with augmented feature vectors without

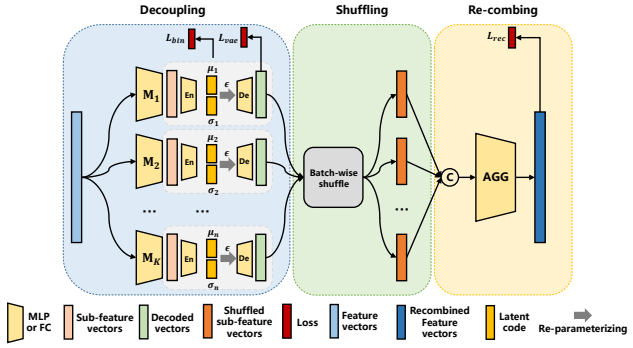

Fig. 2. The details of Variational Feature Decoupling and Feature Aggregation.

introduing extra computation cost in the inference phase. The framework of the proposed VFA is shown in Fig.1. The motivation is to augment the feature vectors extracted by the backbone via decoupling and reconstituting the co-occurrence relations of different labels in the training set. To achieve this purpose, the feature vectors $\mathbf{v}$ is first decoupled into multiple sub-feature vectors $\{\mathbf{v}^j\}_{j=1}^C$ by a Feature Decoupled module consisting of C variational auto-encoders (VAEs)[10], where C is the number of categories. After decoupling, we shuffle the permutation of the sub-feature vectors of each category j in a batch and re-combine them into unitive feature vectors. The re-combination procedure can enrich the co-occurence relations, e.g, the component of category A in instance i could be seamed with the component of category B

in instance j, which does not appear in the original training set. The re-combined features will be re-projected into the feature space and used for refining the multi-label classifiers. In the inference phase, all modules used in the training phase are removed to maintain the same architecture as the baseline model.

2.2. Variational Feature Decoupling

In this work, we decouple the feature into sub-feature spaces discriminatively, which is implemented with C variational auto-encoders and binary classifiers. The detailed structure of Variational Feature Decoupling module is illustrated in Fig.2. Specifically, we first transform the feature vectors $\mathbf{v}$ into C lower-dimensional sub-feature vectors $\{\mathbf{v}^j\}_{j=1}^C$ by C two-layer MLPs $\{f_j\}_{j=1}^C$ with the same structure. For each $\mathbf{v}^j$, we set a VAE composed of an encoder e_j and a decoder d_j, which first encodes the sub-space feature vectors into latent codes $\{\mu_j, \sigma_j\}$ with $[\mu_j, \sigma_j] = e_j(\mathbf{v}^j)$ and construct it with re-parameterization technique [10]

$$\tilde{\mathbf{v}^j} = d_j(\sigma_j \times \epsilon + \mu_j) \tag{1}$$

where μ_j and σ_j represent the first and second moment vector, and ϵ is a standard Gaussian random vector of the same shape with μ_j. The encoding and decoding process is supervised by common VAE losses [10]

$$\mathcal{L}_{vae} = \frac{1}{2}\sum_{j=1}^{C}\sum_{k=1}^{D_z}(1 + log((\sigma_j^{(k)})^2) - (\mu_j^{(k)})^2 - (\sigma_j^{(k)})^2) + \left\|\mathbf{v}^j - \tilde{\mathbf{v}}^j\right\|_2. \tag{2}$$

Additionally, similar to [11], we set a binary classifier h_j on the latent codes of each category to impel the sub-space vectors to be discriminative. The loss function is formulated by

$$\mathcal{L}_{bin} = \frac{1}{C}\sum_{j=1}^{C} l(h_j(\mathbf{z}_j), \mathbf{y}^j), \tag{3}$$

where $l(\cdot,\cdot)$ is standard binary cross entropy loss and $\mathbf{z}_j = \sigma_j \times \epsilon + \mu_j$ is the re-parameterized vector.

2.3. Shuffle and Feature Aggregation

As mentioned earlier, the purpose of feature decoupling is to make feature augmentation more flexible for multi-label classification. Once we obtain the deoupled sub-feature vectors, we shuffle the them in the batch dimension. The motivation of shuffle is to enrich the co-occurrence relations of labels. A schematic diagram is shown in Figure 3. After shuffling, co-occurence relations with low frequency will be enriched and new relations might appear, which could balance the distributions of the original data and benefit the learning of multi-label classifiers. Given the sub-feature vectors $\mathbf{V}^j = [\mathbf{v}_1^j, \mathbf{v}_2^j, ..., \mathbf{v}_n^j] \in \mathbb{R}^{n\times d}$ of category j, where n is the batch size and d is the dimensionality of sub-space feature vectors, this process can be described as

$$[\mathbf{v}_1^j, \mathbf{v}_2^j, ..., \mathbf{v}_n^j] \overset{shuffle}{\rightarrow} [\mathbf{v}_{a_1}^j, \mathbf{v}_{a_2}^j, ..., \mathbf{v}_{a_n}^j] := \tilde{\mathbf{V}}^j, \tag{4}$$

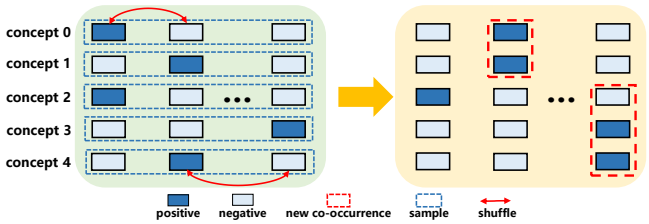

Fig. 3. The semantic diagram of shuffle (best viewed in color). The shuffle is conducted in the batch dimension (the blue dotted boxes). After shuffling, unprecedented co-occurrence relations might appear in the training set (the red dotted boxes).

where $\{a_i\}$ is the new indices of the batch. For all categories, the shuffle is conducted independently. After shuffled, the sub-feature vectors are concatenated and re-projected into the original feature space by a Feature Aggregation module g as:

$$\tilde{\mathbf{V}} = g([\tilde{\mathbf{V}}^1; \tilde{\mathbf{V}}^2; ...; \tilde{\mathbf{V}}^C]) \tag{5}$$

The Feature Aggregation module is implemented with a two-layer MLP. To ensure that the augmented feature vectors $\tilde{\mathbf{V}}$ can be correctly re-projected into the original feature space, we introduce a L2 reconstruction loss

$$\mathcal{L}_{rec} = \left\|\tilde{\mathbf{V}}_u - \mathbf{V}\right\|_2, \tag{6}$$

where $\tilde{\mathbf{V}}_u = g([\mathbf{V}^1; \mathbf{V}^2; ...; \mathbf{V}^C])$ is the reconstructed feature vectors without applying shuffle and $\mathbf{V}$ is the original feature vectors extracted by the backbone network. Finally, the original feature vectors $\mathbf{V}$ and augmented feature vectors $\tilde{\mathbf{V}}$ are jointly used for fine-tuning the classifiers with binary cross entropy losses

$$\mathcal{L}_{cls} = \frac{1}{N}\sum_{i=1}^{N} l(h(\mathbf{v}_i), \mathbf{y}_i) + \frac{\beta}{N}\sum_{i=1}^{N} l(h(\tilde{\mathbf{v}}_i), \tilde{\mathbf{y}}_i), \tag{7}$$

where β is the weight that balances the original loss and auxiliary loss.

3. EXPERIMENTS

In this section, we first introduce the datasets and evaluation metrics. Then, we compare our VFA based on different baseline models with other existing state-of-the-art methods on each public dataset.

3.1. Datasets and Metrics

Pedestrian Attribute Recognition. We conduct experiments for PAR on three benchmarks: PETA[12], RAP[13] and PA100k[14] datasets. **PETA** is a widely used dataset for PAR. It contains 19000 outdoor images and 35 selected attributes for evaluation. **RAP** is the largest PAR dataset of indoor scenes, and it contains 41585 images. We follow [13] to select 51 attributes for evaluation. **PA100k** is the largest

Method	mA	Accu	Prec	Recall	F1
GRL	**86.70**	-	84.34	**88.82**	86.51
HP-Net	81.77	76.13	84.92	83.24	84.07
VeSPA	83.45	77.73	86.18	84.81	85.49
DIAA	84.59	78.56	86.79	86.12	86.46
ALM	86.30	79.52	85.65	**88.09**	86.85
JLAC	**86.96**	**80.38**	**87.81**	87.09	**87.45**
ResNet50	85.19	79.14	87.11	86.18	86.36
VFA(Ours)	**86.47**	**80.48**	**87.35**	**87.74**	**87.31**

Table 1. The comparisons on PETA dataset. The results in **red** and **blue** represent the best and second best performances.

Method	mA	Accu	Prec	Recall	F1
VeSPA	77.70	67.35	79.51	79.67	79.59
HP-Net	76.12	65.39	77.33	78.79	80.09
LGNet	78.68	68.00	**80.36**	79.82	80.09
GRL	81.20	-	77.70	80.90	79.29
ALM	**81.87**	68.17	74.71	**86.48**	**80.16**
JLAC	**83.69**	**69.15**	79.31	**82.40**	**80.82**
ResNet50	80.52	68.44	**79.91**	80.64	79.89
VFA(Ours)	**81.76**	**68.49**	79.09	**81.74**	**80.01**

Table 2. The comparisons on RAP dataset. The results in **red** and **blue** represent the best and second best performances.

PAR dataset with 100000 images from outdoor scenes. It provides 26 commonly used attributes. We adopt five criteria to evaluate the model, including a label-based criterion mean accuracy (mA), accuracy (Accu), precision (Prec), Recall and F1.

Multi-label Image Classification. We conduct experiments for multi-label image classification on MS-COCO[9] dataset. **MS-COCO** is widely used for multi-label recognition recently, and it contains 82081 images for training and 40137 images for validation. The dataset covers 80 commin object categories. We adopt the average of overall/categorial F1-score (OF1/CF1) and mean Average Precision (mAP) as evaluation metrics.

3.2. Comparison with State-of-The-Arts

Pedestrian Attribute Recognition. We take the methods of GRL[15], HP-Net[14], LGNet[16], VeSPA[17], DIAA[18], ALM[6] and JLAC[19] for comparisons. We use a strong baseline (ResNet50) proposed in [20] for all the experiments on pedestrian attribute recognition. Table 1, Table 2 and Table 3 shows the comparisons with other state-of-the-arts. Our VFA gains consistent improvement over baseline models and maintain the same computation and amount of parameters, and also achieve comparable performance with many other state-of-the-arts. For example, on PETA dataset, VFA outperforms more complicated and heavy models such as ALM[6] in both mA and F1 criterion. Morever, we close the gap between the baseline model and The SOTA graph-based JLAC[19] in all five criteria by large steps. The experimental results on PAR datasets demonstrate the effectiveness and efficiency of our VFA.

Method	mA	Accu	Prec	Recall	F1
HP-Net	74.21	72.19	82.97	82.09	82.53
LGNet	76.96	75.55	86.99	83.17	85.04
VSGR	79.52	**80.58**	**89.40**	87.15	**88.26**
ALM	80.68	77.08	84.21	**88.84**	86.46
JLAC	**82.31**	**79.47**	**87.45**	87.77	**87.61**
ResNet50	80.50	78.84	**87.24**	87.12	86.78
VFA(Ours)	**81.30**	**79.01**	86.66	**88.08**	**86.95**

Table 3. The comparisons on PA100k dataset. The results in **red** and **blue** represent the best and second best performances.

Method	Backbone	Input Size	mAP	CF1	OF1
CNN-RNN	VGG16	-	61.2	-	-
SRN	ResNet-101	224	77.1	71.2	75.8
ML-GCN	ResNet-101	448	83.0	78.0	80.3
SSGRL	ResNet-101	576	83.8	76.8	79.7
SSGRL	ResNet-101	448	81.9	76.6	78.9
MCAR	ResNet-101	576	84.5	78.7	81.1
MCAR	ResNet-101	448	83.8	78.0	80.3
TResNet-M	TResNet-M	224	76.6	70.7	73.7
VFA(Ours)	TResNet-M	224	**77.4**	**71.5**	**74.7**
TResNet-L	TResNet-L	448	84.1	77.3	79.5
VFA(Ours)	TResNet-L	448	**84.5**	**78.8**	**80.8**
ASL*	TResNet-L	448	**86.3**	**81.4**	**81.8**
VFA(ours)	TResNet-L	448	**86.5**	**80.4**	**82.4**

* Re-implemented with the open source codes and pre-trained parameters.

Table 4. The comparisons on MS-COCO. The results in **red** and **blue** represent the best and second best performances.

MS-COCO. For multi-label image recognition, we take several representative methods for comparisons[21, 2, 3, 4, 5, 8]. We employ TResNet[8] as our baselines with two settings of input resolution. Table 4 shows the comparisons on MS-COCO dataset. With VFA, all settings of baselines gain consistent improvement. Specifically, VFA improves TResNet-M by 0.8% in mAP and improves TResNet-L by 0.4%. The experimental results on MS-COCO prove that our VFA can effectively improve multi-label classifiers without extra burden of computation and storage.

4. CONCLUSION

In this work, we propose a novel Variational Feature Augmentation method improving the multi-label classification models in a costless way. The proposed method increase co-occurrence relations of labels by decoupling and recombining the feature vectors to refine classifiers. Extensive experiments on three multi-label tasks and six benchmarks including PETA, RAP, PA100k and MS-COCO demonstrate the effectiveness of the proposed method.

5. REFERENCES

[1] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[2] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5513–5522.

[3] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5177–5186.

[4] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 522–531.

[5] Bin-Bin Gao and Hong-Yu Zhou, "Multi-label image recognition with multi-class attentional regions," *arXiv preprint arXiv:2007.01755*, 2020.

[6] Chufeng Tang, Lu Sheng, Zhaoxiang Zhang, and Xiaolin Hu, "Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4997–5006.

[7] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin, "Distribution-balanced loss for multi-label classification in long-tailed datasets," in *European Conference on Computer Vision*. Springer, 2020, pp. 162–178.

[8] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor, "Asymmetric loss for multi-label classification," *arXiv preprint arXiv:2009.14119*, 2020.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[10] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[11] Zheng Ding, Yifan Xu, Weijian Xu, Gaurav Parmar, Yang Yang, Max Welling, and Zhuowen Tu, "Guided variational autoencoder for disentanglement learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7920–7929.

[12] 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*, 2014, pp. 789–792.

[13] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE transactions on image processing*, vol. 28, no. 4, pp. 1575–1590, 2018.

[14] 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*, 2017, pp. 350–359.

[15] Xin Zhao, Liufang Sang, Guiguang Ding, Yuchen Guo, and Xiaoming Jin, "Grouping attribute recognition for pedestrian with joint recurrent learning.," in *IJCAI*, 2018, pp. 3177–3183.

[16] Pengze Liu, Xihui Liu, Junjie Yan, and Jing Shao, "Localization guided learning for pedestrian attribute recognition," *arXiv preprint arXiv:1808.09102*, 2018.

[17] M Saquib Sarfraz, Arne Schumann, Yan Wang, and Rainer Stiefelhagen, "Deep view-sensitive pedestrian attribute inference in an end-to-end model," *arXiv preprint arXiv:1707.06089*, 2017.

[18] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 680–697.

[19] Zichang Tan, Yang Yang, Jun Wan, Guodong Guo, and Stan Z Li, "Relation-aware pedestrian attribute recognition with graph convolutional networks.," in *AAAI*, 2020, pp. 12055–12062.

[20] Jian Jia, Houjing Huang, Wenjie Yang, Xiaotang Chen, and Kaiqi Huang, "Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method," 2020.

[21] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2285–2294.