Learning Disentangled Attribute Representations for Robust Pedestrian Attribute Recognition

Jian Jia[1, 2], **Naiyu Gao**[1, 2], **Fei He**[1, 2], **Xiaotang Chen**[1, 2], **Kaiqi Huang**[1, 2, 3]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[2] CRISE, Institute of Automation, Chinese Academy of Sciences, Beijing, China,
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China
{jiajian2018, gaonaiyu2017, hefei2018}@ia.ac.cn, {xtchen, kqhuang}@nlpr.ia.ac.cn

Abstract

Although various methods have been proposed for pedestrian attribute recognition, most studies follow the same feature learning mechanism, *i.e.*, learning a shared pedestrian image feature to classify multiple attributes. However, this mechanism leads to low-confidence predictions and non-robustness of the model in the inference stage. In this paper, we investigate why this is the case. We mathematically discover that the central cause is that the optimal shared feature cannot maintain high similarities with multiple classifiers simultaneously in the context of minimizing classification loss. In addition, this feature learning mechanism ignores the spatial and semantic distinctions between different attributes. To address these limitations, we propose a novel disentangled attribute feature learning (DAFL) framework to learn a disentangled feature for each attribute, which exploits the semantic and spatial characteristics of attributes. The framework mainly consists of learnable semantic queries, a cascaded semantic-spatial cross-attention (SSCA) module, and a group attention merging (GAM) module. Specifically, based on learnable semantic queries, the cascaded SSCA module iteratively enhances the spatial localization of attribute-related regions and aggregates region features into multiple disentangled attribute features, used for classification and updating learnable semantic queries. The GAM module splits attributes into groups based on spatial distribution and utilizes reliable group attention to supervise query attention maps. Experiments on PETA, RAPv1, PA100k, and RAPv2 show that the proposed method performs favorably against state-of-the-art methods.

Introduction

Pedestrian attribute recognition aims to predict multiple attributes for one pedestrian image. Due to the wide range of applications in person re-identification (Lin et al. 2019), person retrieval (Li et al. 2018b), and scene understanding (Jaderberg et al. 2015), pedestrian attribute recognition has attracted increasing attention from industry and academia. From the perspectives of exploiting additional human knowledge and adopting attention mechanisms, numerous methods have been proposed and got significant performance improvements. However, recent works often neglect essential characteristics of pedestrian attribute recognition that distinguish it from other classification tasks (single-label classification and general multi-label classification). We believe that the essence of pedestrian attribute recognition is two-folds.

On the one hand, unlike the single-label classification, as a sub-task of multi-label classification, pedestrian attribute recognition requires a disentangled and discriminative feature for each attribute to predict the corresponding attribute. However, almost all existing methods adopt the same feature learning manner as the single-label classification task, such as classification in ImageNet (Deng et al. 2009). For example, most works (Liu et al. 2017; Wang et al. 2017; Li et al. 2018c, 2019a; Tan et al. 2020) use the identical average-pooled image feature to classify different attributes, and some works (Zhao et al. 2018, 2019) adopt a shared attribute-group feature to classify attributes in the same group. Although these works propose various approaches and achieve promising results, we argue that it is inferior to represent different attributes with one shared and entangled feature, named by the One-shared-Feature-for-Multiple-Attributes (OFMA) mechanism. The fundamental reason for the inferiority is that, given fixed feature channel dimension, angles between the shared optimal feature learned by the OFMA mechanism and individual attribute classifiers converge to 90 degrees as the number of attributes increases, which impairs the robustness of the model. The detailed analysis is introduced in Section . As far as we know, our work is the first to reveal the limitations of this mechanism mathematically. Besides pedestrian attribute recognition, our analysis of limitations of the OFMA mechanism also applies to general multi-label classification.

On the other hand, different from general multi-label classification in COCO (Lin et al. 2014) and PASCAL-VOC (Everingham et al. 2010), in which samples of the same category can appear at any location in the image, each pedestrian attribute has a relatively consistent spatial distribution across the samples, and some attributes share similar spatial regions. For example, “Hat” and “Glasses” attributes appear at the top of the image, while “Boots” and “Sandals” locates at the bottom of the image.

However, how to exploit the two essential characteristics and incorporate them into model construction is nontrivial. To alleviate the limitations of the OFMA mechanism, we introduce the One-specific-Feature-for-One-Attribute (OFOA) mechanism and propose a Disentangled Attribute Feature Learning (DAFL) framework as illustrated in Fig. 2. Instead of using a shared image feature, the DAFL framework extracts attribute-specific features based on precise spatial localization and representative semantic queries. Specifically, we construct learnable semantic queries and propose two complementary modules, *i.e.*, the cascaded Semantic-Spatial Cross-Attention (SSCA) module and the group attention merging (GAM) module. The learnable semantic queries learns the unique semantic characteristic of each attribute from all samples. The cascaded SSCA module iteratively locates attribute-related spatial regions based on semantic queries and outputs query attention maps. Meanwhile, taking image feature maps as inputs and query attention maps as the affinity matrix, the cascaded SSCA module integrates spatial region features into disentangled attribute features, used for classification and updating semantic queries. To supervise the query attention map and achieve accurate localization, we divide attributes into several groups based on spatial distribution and propose the GAM module to merges qualified query attention maps into group attention memory, which is utilized as the pseudo-label. To supervise the attribute features, besides the classification loss, we construct four triplets for each attribute and apply the semantic triplet loss to achieves the compactness between positive features and discrepancy between positive and negative features [1].

The main contributions of our work are as follows:

- We expose the limitations of the one-shared-feature-for-multiple-attributes mechanism adopted in most existing works and propose the disentangled attribute feature learning framework, an instance of the one-specific-feature-for-one-attribute mechanism.
- We propose a cascaded semantic-spatial cross-attention module to learn discriminative feature for each attribute, which is assisted by a group attention merging module and a semantic triplet loss to improve the localization ability and robustness of the model.
- We confirm the efficacy of the proposed method by achieving state-of-the-art performance on PETA, RAPv1, PA100k, and RAPv2.

Related Work

Pedestrian attribute recognition has recently undergone rapid development. Since Li (Li, Chen, and Huang 2015) introduced deep learning into pedestrian attribute recognition, various methods have been proposed and significant progress have been made. According to the feature used to classify attributes, we divide current works into three categories.

The first category of methods (Li, Chen, and Huang 2015; Yu et al. 2017; Liu et al. 2017; Li et al. 2018a; Liu et al. 2018; Sarafianos, Xu, and Kakadiaris 2018; Han et al. 2019; Guo et al. 2019; Tan et al. 2020) extracted a shared global feature to classify all attributes. These methods usually adopted attention mechanisms and took the average-pooled image feature as the global feature. Liu (Liu et al. 2017) proposed a multi-direction attention network, HydraPlus-Net, to utilize diverse semantic attention from different layers. Sarafianos (Sarafianos, Xu, and Kakadiaris 2018) constructed a Visual Attention and Aggregation (VAA) module and applied it on multi-scale feature maps. Guo (Guo et al. 2019) proposed attention consistency loss to align the attention regions of augmentations of the same image. Tan (Tan et al. 2020) proposed a two-branch network JLAC, where the ARM branch generated attribute features based on the average-pooled image feature and the CRM branch concatenated graph node features to make predictions.

The second category of methods divided attributes into several groups based on the spatial distribution and adopted one group feature to classify multiple attributes in the same group. Zhao (Zhao et al. 2018) introduced human key points to generate body proposals and adopt RoI average pooling layers to extract proposal features as the group features. Instead of utilizing human key points, Li (Li et al. 2019a) exploited the human parsing model to locate body regions and adopted graph convolution networks to obtain corresponding group features.

The last category of methods attempted to extract one specific feature for each attribute. Lin (Wang et al. 2017) took horizontally divided features as input and introduced LSTM (Hochreiter and Schmidhuber 1997) to decode individual features for each attribute. Li (Li et al. 2019c) proposed a visual-semantic graph reasoning framework to capture spatial region relations and attribute semantic relations. Node features of the graph networks are used as the specific features for each attribute. Tang (Tang et al. 2019) constructed multi-scale feature maps by Feature Pyramid Network (FPN) (Lin et al. 2017) and proposed the attribute location module (ALM) to extracted features for each attribute in each scale feature map.

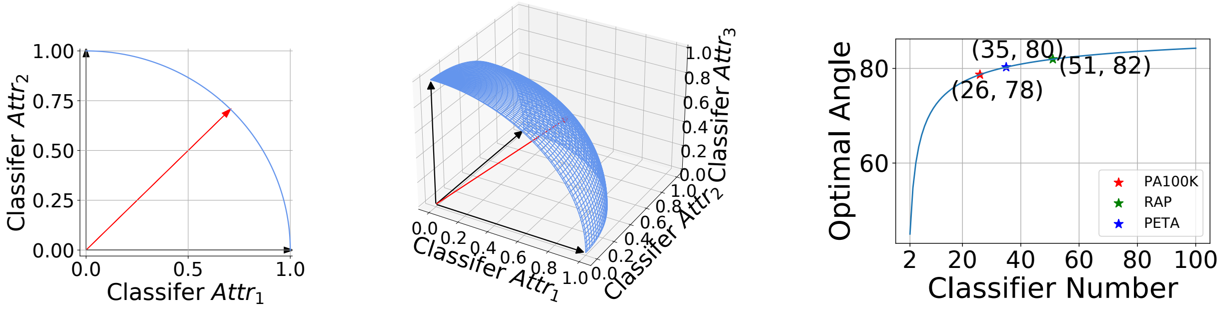
Proposed Approach

In this section, we first introduce the attribute prediction process and point out that the only determinant for attribute prediction is the angle between the shared feature vector and the classifier weight. Then, we expose the limitations of the One-shared-Feature-for-Multiple-Attributes (OFMA) mechanism adopted by most methods and propose the One-specific-Feature-for-One-Attribute (OFOA) mechanism to tackle the deficiencies in OFMA. Finally, following the OFOA mechanism, we construct the Disentangled Attribute Feature Learning (DAFL) framework and design the spatial group attention loss and semantic triplet loss.

Attribute Prediction Process

The prediction results of pedestrian attributes depend on the choice of probability thresholds. To make an intuitive and fair comparison, all existing methods set the probability threshold $p_t = 0.5$. Given a dataset $D = \{x_i, y_i, | i = 1, \ldots, N\}$ and $y_i \in \{0, 1\}^M$, for the j-th attribute of the i-th

[1] We use the positive feature to indicate the feature of the sample with the corresponding attribute. We use the negative feature to indicate the feature of the sample without the corresponding attribute.



(a) Classification on two attributes.

(b) Classification on three attributes.

(c) Classification on M attributes.

Figure 1: Optimal angle between a shared feature vector and multiple classifier weights. Assuming that the classifier weights are orthogonal and the classifier norms are identical, the optimal angle is the angle that minimizes the binary cross-entropy loss. We normalize the feature vector (red arrow) and attribute classifier weights (black arrows) for simplification. In (a) and (b), we depict the optimal angles between a shared feature vector and two classifiers as well as three classifiers, respectively. In (c), we plot the curve of the optimal angle as the number of classifiers increases and give the optimal angles on three popular datasets.

sample x_i, the prediction result $\hat{y}_{i,j}$ is decided as follows:

$$\hat{y}_{i,j} = \begin{cases} 1, & p_{i,j} >= p_t \\ 0, & p_{i,j} < p_t \end{cases}, \tag{1}$$

$$p_{i,j} = \sigma(logits_{i,j}), \tag{2}$$

where $p_{i,j}$ is the predicted probability and $\sigma(\cdot)$ is the sigmoid function. The output of classifier layer is denoted as $logits$, which is computed as:

$$logits_{i,j} = w_j^T f_i = |w_j| \cdot |f_i| \cdot \cos\theta, \tag{3}$$

where $f_i \in R^C$ is the shared feature vector of the sample x_i in OFMA machenism and $W = \{w_j | j = 1, \ldots, M\} \in R^{C \times M}$ is the classifier weight matrix. Taking Eq. 2 and Eq. 3 into Eq. 1, we conclude that the prediction of attributes only depends on the sign of $\cos\theta$ in Eq. 3, *i.e.*, the angle θ between the feature vector f and the classifier weight w:

$$\hat{y}_{i,j} = \begin{cases} 1, & 0^\circ <= \theta <= 90^\circ \\ 0, & 90^\circ < \theta < 180^\circ \end{cases}. \tag{4}$$

Therefore, for a target attribute, a well-trained model should make angles between positive sample features and the corresponding classifier weights as small as possible, or even close to 0°, which means high-confidence prediction.

Limitations of the OFMA Mechanism

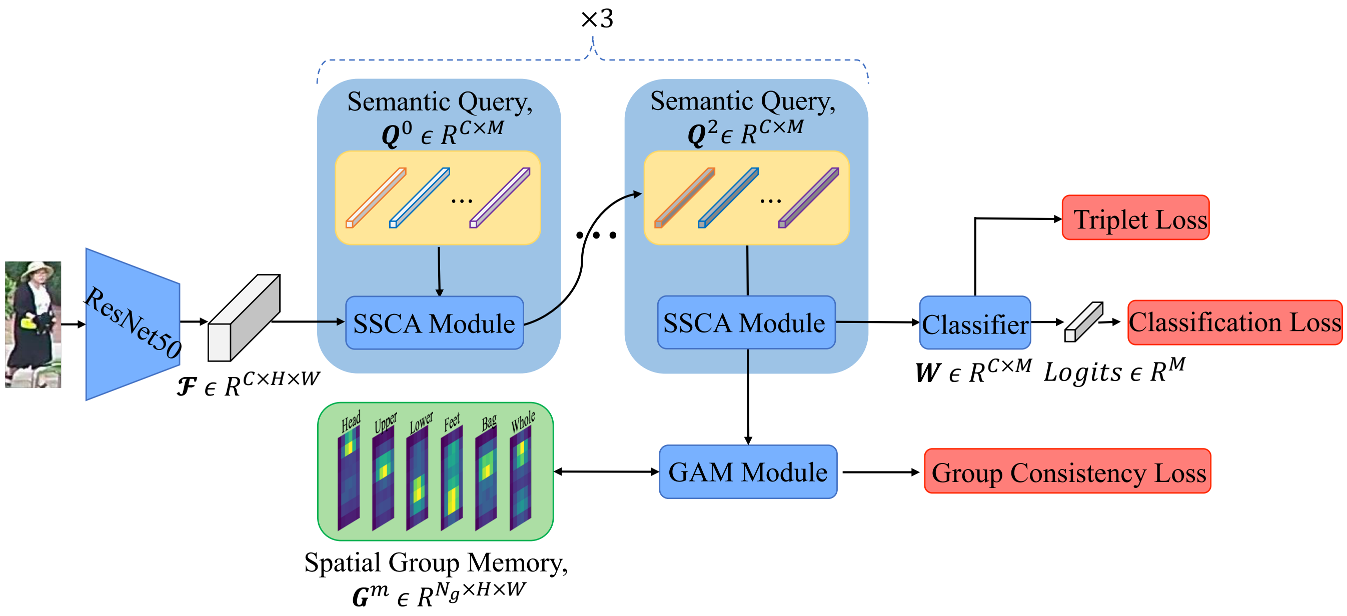
For a well-trained model that follows the OFMA mechanism, we have two critical experimental observations, which is illustrated in the supplementary material. One is that the classifier weights of attributes are mostly orthogonal. The other is that the classifier norms are approximately the same. Considering the orthogonality between attribute classifiers, for the OFMA mechanism, it is intuitively impossible to make angles between the shared feature vector and multiple attribute classifier weights all close to 0°. However, these angles are expected to be as small as possible, which implies high-confidence prediction of the model. But what is the theoretical optimal angle?

We start from classification on two attributes, *i.e.*, classifying a sample (f_i, y_i) where f_i is the feature vector of the sample and $y_i = \{1, 1\}$ is the ground truth label. We take classification on the positive sample of attributes as an example, and the classification on negative samples can be analyzed in the same way. Following the common practice, binary cross-entropy loss is adopted as the classification loss and other experimental settings are present in the supplementary material. Thus, the problem can be formulated as:

$$\max_{f_i} \log\Big(\sigma(w_1^T f_i)\Big) + \log\Big(\sigma(w_2^T f_i)\Big). \tag{5}$$

Given the two experimental observations, we hypothesize that the classifier weights w_1 and w_2 are orthogonal, and their norms are the same. We prove that the optimal feature f is located in the middle of the two classifiers, *i.e.*, the feature has the same distance from both two classifiers and the optimal angles between the shared feature and two classifiers are both 45°, as demonstrated in Fig. 1(a). The proof is available in the supplementary material. For classification on three attributes, the optimal feature has the same distance from all three classifiers, *i.e.*, the optimal angles are 54.74°, as shown in Fig. 1(b). Furthermore, for classification on M attributes, we conclude that the optimal feature achieves a trade-off in the distance between itself and multiple classifiers to minimize the binary cross-entropy loss. This property makes the optimal angle converge to 90°as the number of attributes M increases. Specifically, on existing datasets PA100k, PETA, and RAP, the attributes M are 26, 35, 51, and the optimal angles are 78.69°, 80.27°, and 81.95°respectively, as shown in Fig. 1(c). This theoretical conclusion is verified in the experimental statistics in Fig. 3.

However, the optimal angles close to 90°in the training stage are far from our expectations. As a result, a small perturbation can make features of the test set cross the decision boundary, causing angles of the test set to be greater than 90°and yielding wrong predictions. Specifically, the learned features are susceptible to changes in pedestrian pose, illumination, and background, resulting in incorrect classifica-



×3
Semantic Query,
$\boldsymbol{Q}^0 \in R^{C\times M}$
Semantic Query,
$\boldsymbol{Q}^2 \in R^{C\times M}$
…
…
ResNet50
SSCA Module
SSCA Module
$\boldsymbol{\mathcal{F}} \in R^{C\times H\times W}$
Triplet Loss
Classifier
Classification Loss
$\boldsymbol{W} \in R^{C\times M}$ $Logits \in R^M$
Head
Upper
Lower
Feet
Bag
Whole
GAM Module
Group Consistency Loss
Spatial Group Memory,
$\boldsymbol{G}^m \in R^{N_g\times H\times W}$

(a) DAFL framework.

Figure 2: Illustration of the DAFL framework and the SSCA module. The proposed framework consists of two modules and three losses. Taking the image feature map $\mathcal{F}$ and previous semantic query $\boldsymbol{Q}^{s-1}$ as input, the cascaded SSCA module locates attribute-related regions iteratively and outputs attribute feature $\boldsymbol{F}^{s-1}$ as the semantic query $\boldsymbol{Q}^s$ of next step, where s is the step of the cascaded SSCA module. The GAM module merges query attention maps of the last SSCA module into spatial group memory $\boldsymbol{G}$ as the pseudo-label to implement group consistency loss.

tions. Some examples are given in Fig. 4. Thus, we conclude that, the shared features learned by the OFMA mechanism are too far from the classifiers, which reduces the robustness of the model on the test set. In addition, it is worth noting that this deficiency is determined by the OFMA mechanism and independent of the specific method.

Besides to large angles between the shared feature and classifiers, there is another limitation of the OFMA mechanism. The shared global features and group features of the OFMA mechanism, used to classify multiple attributes, are average pooled from the image feature map. Nevertheless, the pooling operation erases the spatial distinction of attributes and only exploits the channel information. Therefore, the OFMA mechanism is not adequate to exploit the differences between attributes to learn discriminative features for each attribute.

Proposed DAFL Framework

To solve the limitations in the OFMA mechanism, we propose a DAFL framework following OFOA mechanism as illustrated in Fig. 2. Specifically, we propose to locate attribute-related spatial regions and aggregate discriminative regional features as a specific feature for each attribute, rather than pooling a shared feature for all attribute classifications, which erases the distinction of spatial distribution. In addition to the difference in spatial distribution, semantic characteristics also facilitate learning discriminative features, especially for attributes with a similar spatial distribution. Considering that the semantic characteristics are consistent across samples, we design learnable semantic queries and introduce the triplet loss on features of each attribute to further improve the discrimination of features.

The DAFL framework mainly consists of the cascaded SSCA module and the GAM module. The SSCA module is proposed to locate precise spatial regions for each attribute and aggregate region features as attribute features. Concretely, the s-th SSCA module takes semantic query $\boldsymbol{Q}^s \in R^{C\times M}$ and image feature map $\mathcal{F} \in R^{C\times H\times W}$ as inputs and outputs query attention map $\boldsymbol{A} \in R^{M\times H\times W}$ and attribute feature $\boldsymbol{F}^s \in R^{C\times M}$, where M is the number of attributes. The channel, height, and width dimension of the feature map is represented by C, H, and W respectively. The feature map $\mathcal{F}$ is output by the backbone network (ResNet50 (He et al. 2016)). The query attention map $\boldsymbol{A}$ is computed as follows:

$$\boldsymbol{A} = Softmax(\frac{\theta(\boldsymbol{Q})^T \phi(\mathcal{F})}{\sqrt{C}}), \tag{6}$$

where $\theta(\boldsymbol{Q}) = \boldsymbol{W}_\theta \boldsymbol{Q}$ and $\phi(\mathcal{F}) = \boldsymbol{W}_\phi \mathcal{F}$ are two linear embedding functions. As shown in Fig. 2, query attention map highlights attribute-related spatial regions and can be used as affinity matrix to aggregate region features as specific spatial feature F^s, which is formulated as:

$$\boldsymbol{F}^s = \boldsymbol{A}\psi(\mathcal{F})^T, \tag{7}$$

where $\psi(F) = W_\psi F$ is a linear embedding function. Inspired by the success of the multi-head self-attention mechanism (Vaswani et al. 2017), we implement the $\boldsymbol{A}$, $\boldsymbol{F}^s$ in a multi-head manner. To further refine the semantic query and locate reliable spatial regions, we cascade multiple SSCA modules and take attribute feature $\boldsymbol{F}^{s-1}$ of the previous SSCA module as the semantic query $\boldsymbol{Q}^s$ of the current SSCA module:

$$\boldsymbol{Q}^s := \boldsymbol{F}^{s-1}, \tag{8}$$

where $s = 0, 1, \ldots S - 1$ and S is the number of cascaded SSCA modules. The $\boldsymbol{Q}^0$ is randomly initialized.

Based on semantic query $\boldsymbol{Q}$, we can extract spatially disentangled attribute features $\boldsymbol{F}$ from single image feature map $\mathcal{F}$. However, due to the imbalanced attribute distribution, some minority attributes do not have enough positive samples to learn the precise query attention map $\boldsymbol{A}$ and effective attribute features $\boldsymbol{F}$. Inspired by the facts that some attributes have a similar spatial locations, such as “Hat” and “Glasses”, “UpperLogo” and “UpperPlaid”, “Front” and “Back”, we propose to merge query attention maps of multiple attributes with the similar spatial distribution and use the merged group attention to supervise minority attributes. Specifically, we first divide attributes into several groups $\mathcal{G}$ based on the spatial distribution. Then, we propose the group attention merging (GAM) module to merge qualified query attention maps $A_{i,m}$ into group attention $G^a = \{G^a_k | k = 1, \ldots, K\} \in R^{K \times H \times W}$ as follows:

$$\boldsymbol{G}^a_k = \frac{1}{|\mathcal{G}_k|} \sum_{m \in \mathcal{G}_k} \frac{1}{|\mathcal{R}|} \sum_{i=1}^{b_t} \mathbb{1}_{\{\mathcal{R}\}} \boldsymbol{A}_{i,m}, \tag{9}$$

where i denotes the sample index, b_t indicates the batch size, $\mathbb{1}(\cdot)$ is the indicator function, and $|\cdot|$ represents the set cardinality. The condition set $\mathcal{R} = \{\sigma(logits_{i,m}) > \tau,\ y_{i,m} = 1\}$ and the hyper-parameter τ are adopted to select qualified query attention maps for each attribute from the current batch. K is a pre-defined group number and set to $K = 6$ as default. For example, given attribute groups of PA100k as listed in Tab. 1, for group attention $\boldsymbol{G}^a_1$, we first sum the qualified query attentions of “Hat” and “Glasses” in a batch respectively to obtain a “Hat” attention and a “Glasses” attention. Then, two attentions are merged into “Head” group attention based on group $\mathcal{G}_1$.

To mitigate the fluctuation caused by limited batch size and random sampling, we maintain a spatial group memory G^m in a momentum updated way to make the group attention reliable and consistent across batches. The G^m is formulated as:

$$\boldsymbol{G}^m_k \leftarrow (1 - \alpha) \times \boldsymbol{G}^m_k + \alpha \times \boldsymbol{G}^a_k, \tag{10}$$

where $\alpha \in (0, 1]$ is the momentum hyper-parameter. We take the group attention $\boldsymbol{G}^m_k$ as the pseudo-label to supervise the inaccurate query attention map $\boldsymbol{A}_{i,m}$ of attributes in the group $\mathcal{G}_k$. Thus, we propose the group consistency loss to rectify the imprecise spatial localization of minority attributes:

$$L_{group} = \frac{1}{b_t} || \sum_{i=1}^{b_t} \sum_{k=1}^{K} \sum_{m \in \mathcal{G}_k} \boldsymbol{G}^m_k - \boldsymbol{A}_{i,m} ||_2, \tag{11}$$

Group	Attribute
Head ($\mathcal{G}_1$)	Hat, Glasses
UpperBody ($\mathcal{G}_2$)	ShortSleeve, LongSleeve, UpperStripe, UpperLogo, UpperPlaid, UpperSplice
LowerBody ($\mathcal{G}_3$)	LowerStripe, LowerPattern, LongCoat, Trousers, Shorts, Skirt&Dress
Feet ($\mathcal{G}_4$)	Boots
Bag ($\mathcal{G}_5$)	HandBag, ShoulderBag, Backpack, HoldObjectsInFront
Whole ($\mathcal{G}_6$)	AgeOver60, Age18-60, AgeLess18, Female, Front, Side, Back

Table 1: The six spatial groups of attributes in PA100k.

Although the cascaded SSCA and GAM module achieve accurate spatial localization and obtain discriminative attribute features, we find that the distance between positive samples is greater than the distance between positive and negative samples, which undermines the robustness of the model and is illustrated in the supplementary material. Therefore, we apply the triple loss (Hermans, Beyer, and Leibe 2017) to features of each attribute to achieve the compactness of positive features and discrepancy between positive and negative features.

Specifically, for each attribute, we select features from the current batch and construct two triplets. For a positive feature a^p_m of m-th attribute, we select the hardest positive feature f^p_m (with the lowest prediction probability) from other positive features and the hardest negative feature f^n_m (with the highest prediction probability) from all negative features to construct the positive triplet (a^p_m, f^p_m, f^n_m). To minimizes the distance between a^p_m and f^p_m and maximizes the distance between a^p_m and f^n_m, the positive triplet loss is computed as:

$$L_{pos,m} = \sum_{j \in N^p_m} max(0, D(a^p_j, f^p_j) - D(a^p_j, f^n_j)), \tag{12}$$

where N^p_m is the number of positive samples of the m-th attribute in the current batch and $D(\cdot)$ is the distance function. For a negative feature a^n_m, the hardest negative feature f^n_m and the hardest positive feature f^p_m are selected to construct the triplet (a^n_m, f^n_m, f^p_m). The negative triplet loss is computed as:

$$L_{neg,m} = \sum_{j \in N^n_m} max(0, D(a^n_j, f^n_j) - D(a^n_j, f^p_j)), \tag{13}$$

where N^n_m is the number of negative samples of the m-th attribute in the current batch. However, due to the imbalanced distribution of attributes and the random sampling of training batches, positive samples of minority attributes in a batch may be too few to construct effective triplets. Thus, we borrowed the Queue Dictionary mechanism from MOCO (He et al. 2020) to dynamically store positive and negative samples for each attribute. Besides the triplets from the current batch, we also construct two additional triplets from the stored features and apply the triplet loss in the same way as Eq. 12 and Eq. 13.

Experiments and Discussion

In this section, we first conduct experiments on four datasets PETA (Deng et al. 2014), RAPv1 (Li et al. 2016), PA100k (Liu et al. 2017), and RAPv2 (Li et al. 2018b), to make a

Method	PETA		RAPv1		PA100k		RAPv2	
	mA	F1	mA	F1	mA	F1	mA	F1
DeepMAR (ACPR15)	82.89	83.41	73.79	75.56	72.70	81.32	–	–
HPNet (ICCV17)	81.77	84.07	76.12	78.05	74.21	82.53	–	–
JRL (ICCV17)	82.13	82.02	74.74	74.62	–	–	–	–
PGDM (ICME18)	82.97	85.76	74.31	77.35	74.95	83.29	–	–
GRL (IJCAI18)	86.70	86.51	81.20	79.29	–	–	–	–
MsVAA (ECCV18)	84.59	86.46	–	–	–	–	78.34	78.26
RA (AAAI19)	86.11	86.56	81.16	79.34	–	–	–	–
VRKD (IJCAI19)	84.90	**87.91**	78.30	**81.23**	77.87	87.24	–	–
AAP (IJCAI19)	86.97	87.65	81.42	80.65	80.56	86.85	–	–
VAC (CVPR19)	–	–	–	–	79.16	87.59	79.23	77.10
ALM (ICCV19)	86.30	86.85	81.87	80.16	80.68	86.46	79.79	77.77
JLAC (AAAI20)	86.96	87.45	83.69	80.82	82.31	87.61	79.23	77.40
Baseline	86.18	86.43	81.08	79.73	80.71	87.04	80.10	77.35
DAFL	**87.07**	86.40	**83.72**	80.29	**83.54**	**88.09**	81.04	79.13

Table 2: Performance comparison of state-of-the-art methods on the PETA, RAPv1, PA100k, and RAPv2 datasets. Performance in five metrics, including mean Accuracy (mA), accuracy (Accu), precision (Prec), recall, and F1, is evaluated. The first and second highest scores are represented by bold font and underline respectively.

fair comparison with state-of-the-art (SOTA) methods. The dataset information and experimental settings are given in the supplementary material. Then, we conduct exhaustive ablation studies on the largest dataset PA100k to validate the contribution of each component. Finally, we quantitatively analyze our proposed approach and verify our motivation.

Comparison to the State of the Arts

In Tab.2, we compare the performance between our proposed methods and recent SOTA methods on PETA, RAPv1, PA100k, and RAPv2 to show the superiority of our methods. Since some methods (Wang et al. 2017; Li et al. 2019b) adopt the model ensemble policy in the inference stage, we do not list their performance to make a fair comparison. In the label-based metric mA, we achieve SOTA performance and outperform the JLAC method by 0.11%, 0.03%, 1.23%, 1.81% on four datasets. In four instance-based metrics, our method DAFL achieves SOTA performance on the two largest datasets PA100k and RAPv2, and comparable performance on small datasets RAP and PETA. Our method achieves better performance on the label-based metric than those on the instance-based metrics for two reasons. One is that the bottleneck of the mA metric lies in the attributes with a small number of positive samples. The other is that the GAM module and semantic triplet loss of our method can alleviate the imbalance of attribute distribution. In addition, we find that our proposed method achieves better performance on large datasets (90,000 and 67,943 training images on PA100k and RAPv2 respectively) than on small datasets (33,268 and 11,400 training images on PETA and RAPv1 respectively). This phenomenon due to the fact that the effective semantic queries and reliable spatial group memory require more samples to learn.

Method			PA100k	
SSCA	GAM	TripletLoss	mA	F1
–	–	–	80.71	87.04
✓	–	–	82.21	87.44
✓	✓	–	82.65	87.40
✓	–	✓	82.41	87.22
✓	⋆	✓	83.18	87.87
✓	✓	✓	**83.54**	**88.09**

Table 3: Experiments on each component of our method on PA100k. Performance which is improved by each component validates the effectiveness of our method. The first highest scores are represented by bold font. The variant of the GAM module is represented by ⋆ symbol.

Ablation Study

As shown in Tab. 3, we investigate the effect of the cascaded SSCA module, GAM module (with spatial consistency loss), and semantic triplet loss. To make a fair and convincing evaluation, we conduct ablation experiments on the largest dataset PA100k (Liu et al. 2017), which follows the zero-shot pedestrian settings and can truly validate the generalization of the model (Jia et al. 2021). For the mA and F1 metric, the cascaded SSCA module alone can achieve 1.50% and 0.40% performance improvement separately. Combining the GAM and TripletLoss into the cascaded SSCA module can further improve the performance by 1.33%, 0.65% in mA and F1.

To further validate the rationality of proposed spatial consistency loss, we implement a variant GAM* of our method.

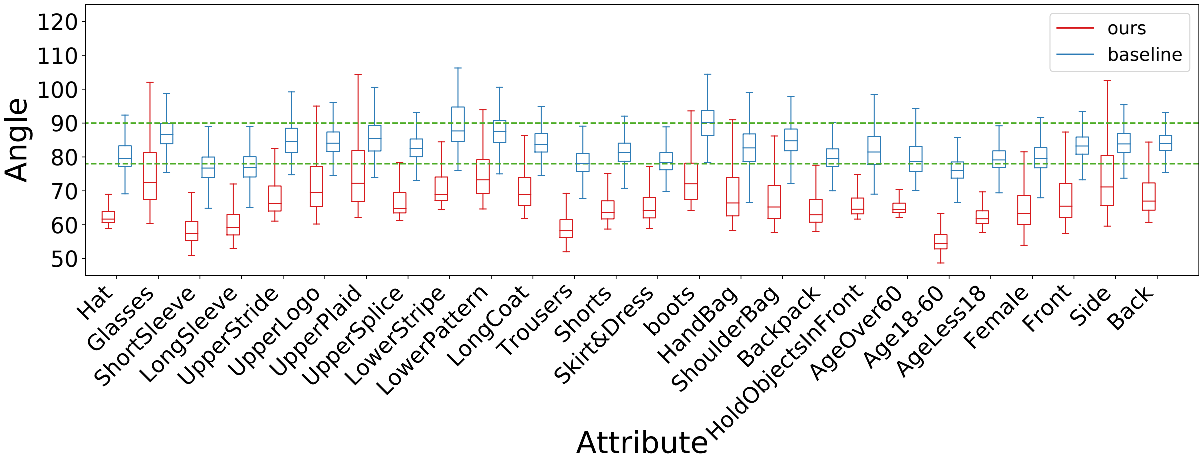


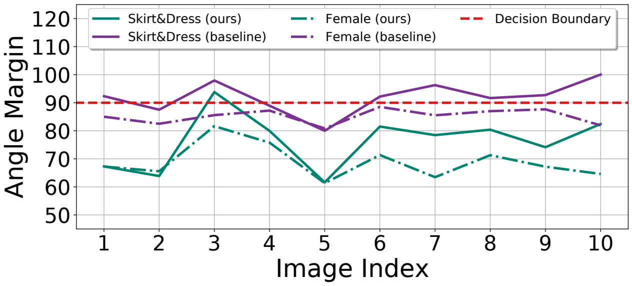
Figure 3: Angle distribution of the baseline model and our proposed method on each attribute. We show angle distribution on each attribute of the baseline model (blue box) and our proposed method (red box). The upper and lower green dashed lines mark the theoretical optimal angle (78°on PA100k) and the angle decision boundary (90°), respectively.

To supervise the query attention map $\boldsymbol{Q}$ of each attribute, the GAM* module maintains a specific spatial memory for each attribute instead of using the group memory shared by multiple attributes. We report the performance of the variant in the second last row of Tab. 3. On the one hand, the performance improvement achieved by the GAM and GAM* module proves the effectiveness of spatial consistency loss. On the other hand, since the GAM module considers the insufficiency in positive samples of minority attributes, it can alleviate the distribution imbalance of attributes and achieve better performance than the GAM* module.

Experiments results on the number of the cascaded SSCA module are listed in Tab 4. As S increases, the performance first increases when $S < 3$ and then decreases when $S > 3$. We argue that the cascaded SSCA module is beneficial for discriminative semantic queries and accurate query attention map when S is small. However, the semantic query $\boldsymbol{Q}^s$ of the current step is the attribute feature $\boldsymbol{F}^{s-1}$ of the previous step, which is aggregated from the feature map $\mathcal{F}$. Thus, when S is large, the number of pixels in the feature map similar to the semantic query gradually increases, and the degree of similarity also increases. As a result, the query attention map $\boldsymbol{A}$ cannot highlight the attribute-related regions but focuses on the whole foreground regions, excluding the background.

Cascade Number	mA	Accu	Prec	Recall	F1
$S = 1$	82.82	79.89	**87.29**	88.50	87.63
$S = 2$	83.31	79.46	86.21	89.10	87.89
$S = 3$	**83.54**	**80.13**	87.01	**89.19**	**88.09**
$S = 4$	83.01	79.49	86.40	88.95	87.66
$S = 5$	83.13	79.60	86.83	88.56	87.69
$S = 6$	82.73	79.44	86.64	88.69	87.65

Table 4: Experiments on the number S of the cascaded SSCA modules.



(a) Predictions on “Skirt&Dress” and “Female” attributes.

Figure 4: Prediction on images with marginal pose variation. Due to the small angles between features and classifiers, our method shows superior robustness than the baseline model.

Conclusion

This paper exposes the limitations in the OFMA mechanism and proposes the discriminative and robust attribute feature learning framework for pedestrian attribute recognition, which follows the OFOA mechanism. The proposed framework makes full use of distinctions between attributes from the spatial distribution and semantic characteristics to extract a specific feature for each attribute. Our proposed method achieves outstanding performance consistently on the PETA, RAPv1, PA100K, and RAPv2.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No.61721004 and Grant No.61876181), the Projects of Chinese Academy of Science (Grant QYZDB-SSW-JSC006), the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA27000000), and the Youth Innovation Promotion Association CAS.

References

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255.

Deng, Y.; Luo, P.; Loy, C. C.; and Tang, X. 2014. Pedestrian attribute recognition at far distance. In *Proceedings of the ACM Int. Conf. Multimedia*, 789–792. ACM.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis*, 88(2): 303–338.

Guo, H.; Zheng, K.; Fan, X.; Yu, H.; and Wang, S. 2019. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 729–739.

Han, K.; Wang, Y.; Shu, H.; Liu, C.; Xu, C.; and Xu, C. 2019. Attribute aware pooling for pedestrian attribute recognition. In *Proceedings of the Int. Joint Conf. Artif. Intell.*

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. In *Advances in neural information processing systems*, 2017–2025.

Jia, J.; Huang, H.; Chen, X.; and Huang, K. 2021. Rethinking of Pedestrian Attribute Recognition: A Reliable Evaluation under Zero-Shot Pedestrian Identity Setting. arXiv:2107.03576.

Li, D.; Chen, X.; and Huang, K. 2015. Multi-attribute Learning for Pedestrian Attribute Recognition in Surveillance Scenarios. In *Proceedings of the IEEE Asia Conf. Pattern Recognit.*, 111–115.

Li, D.; Chen, X.; Zhang, Z.; and Huang, K. 2018a. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *Proceedings of the IEEE Int. Conf. Multimedia Expo.*, 1–6. IEEE.

Li, D.; Zhang, Z.; Chen, X.; and Huang, K. 2018b. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE Transactions on Image Processing*, 28(4): 1575–1590.

Li, D.; Zhang, Z.; Chen, X.; Ling, H.; and Huang, K. 2016. A Richly Annotated Dataset for Pedestrian Attribute Recognition. arXiv:1603.07054.

Li, Q.; Zhao, X.; He, R.; and Huang, K. 2018c. Pedestrian Attribute Recognition by Joint Visual-semantic Reasoning and Knowledge Distillation. In *Proceedings of the Int. Joint Conf. Artif. Intell.*, 3177–3183.

Li, Q.; Zhao, X.; He, R.; and Huang, K. 2019a. Pedestrian Attribute Recognition by Joint Visual-semantic Reasoning and Knowledge Distillation. In *Proceedings of the Int. Joint Conf. Artif. Intell.*, 833–839.

Li, Q.; Zhao, X.; He, R.; and Huang, K. 2019b. Recurrent Prediction With Spatio-Temporal Attention for Crowd Attribute Recognition. *IEEE Transactions Circuits Syst. Video Technol.*, 30(7): 2167–2177.

Li, Q.; Zhao, X.; He, R.; and Huang, K. 2019c. Visual-semantic graph reasoning for pedestrian attribute recognition. In *Proceedings of the AAAI Conf. Artif. Intell.*, volume 33, 8634–8641.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 740–755. Springer.

Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Hu, Z.; Yan, C.; and Yang, Y. 2019. Improving person re-identification by attribute and identity learning. *Pattern Recognit.*

Liu, P.; Liu, X.; Yan, J.; and Shao, J. 2018. Localization guided learning for pedestrian attribute recognition. In *Proceedings of the Brit. Mach. Vis. Conf.*

Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; and Wang, X. 2017. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, 350–359.

Sarafianos, N.; Xu, X.; and Kakadiaris, I. A. 2018. Deep imbalanced attribute classification using visual attention aggregation. In *Proceedings of the European Conference on Computer Vision*, 680–697.

Tan, Z.; Yang, Y.; Wan, J.; Guo, G.; and Li, S. Z. 2020. Relation-aware pedestrian attribute recognition with graph convolutional networks. In *Proceedings of the AAAI Conf. Artif. Intell.*, volume 34, 12055–12062.

Tang, C.; Sheng, L.; Zhang, Z.; and Hu, X. 2019. Improving Pedestrian Attribute Recognition With Weakly-Supervised Multi-Scale Attribute-Specific Localization. In *Proceedings of the IEEE International Conference on Computer Vision*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*, 5998–6008.

Wang, J.; Zhu, X.; Gong, S.; and Li, W. 2017. Attribute recognition by joint recurrent learning of context and correlation. In *Proceedings of the IEEE International Conference on Computer Vision*, 531–540.

Yu, K.; Leng, B.; Zhang, Z.; Li, D.; and Huang, K. 2017. Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. In *Proceedings of the Brit. Mach. Vis. Conf.*

Zhao, X.; Sang, L.; Ding, G.; Guo, Y.; and Jin, X. 2018. Grouping Attribute Recognition for Pedestrian with Joint Recurrent Learning. In *Proceedings of the Int. Joint Conf. Artif. Intell.*, 3177–3183.

Zhao, X.; Sang, L.; Ding, G.; Han, J.; Di, N.; and Yan, C. 2019. Recurrent attention model for pedestrian attribute recognition. In *Proceedings of the AAAI Conf. Artif. Intell.*, volume 33, 9275–9282.