Spatial and Semantic Consistency Regularizations for Pedestrian Attribute Recognition

Jian Jia [1,2], Xiaotang Chen [1,2], Kaiqi Huang [1,2,3] *
[1] the School of Artificial Intelligence, University of Chinese Academy of Sciences
[2] CRISE, Institute of Automation, Chinese Academy of Sciences
[3] CAS Center for Excellence in Brain Science and Intelligence Technology
jiajian2018@ia.ac.cn, {xtchen,kqhuang} @nlpr.ia.ac.cn

Abstract

While recent studies on pedestrian attribute recognition have shown remarkable progress in leveraging complicated networks and attention mechanisms, most of them neglect the inter-image relations and an important prior: spatial consistency and semantic consistency of attributes under surveillance scenarios. The spatial locations of the same attribute should be consistent between different pedestrian images, e.g.*, the "hat" attribute and the "boots" attribute are always located at the top and bottom of the picture respectively. In addition, the inherent semantic feature of the "hat" attribute should be consistent, whether it is a baseball cap, beret, or helmet. To fully exploit inter-image relations and aggregate human prior in the model learning process, we construct a Spatial and Semantic Consistency (SSC) framework that consists of two complementary regularizations to achieve spatial and semantic consistency for each attribute. Specifically, we first propose a spatial consistency regularization to focus on reliable and stable attribute-related regions. Based on the precise attribute locations, we further propose a semantic consistency regularization to extract intrinsic and discriminative semantic features. We conduct extensive experiments on popular benchmarks including PA100K, RAP, and PETA. Results show that the proposed method performs favorably against state-of-the-art methods without increasing parameters.*

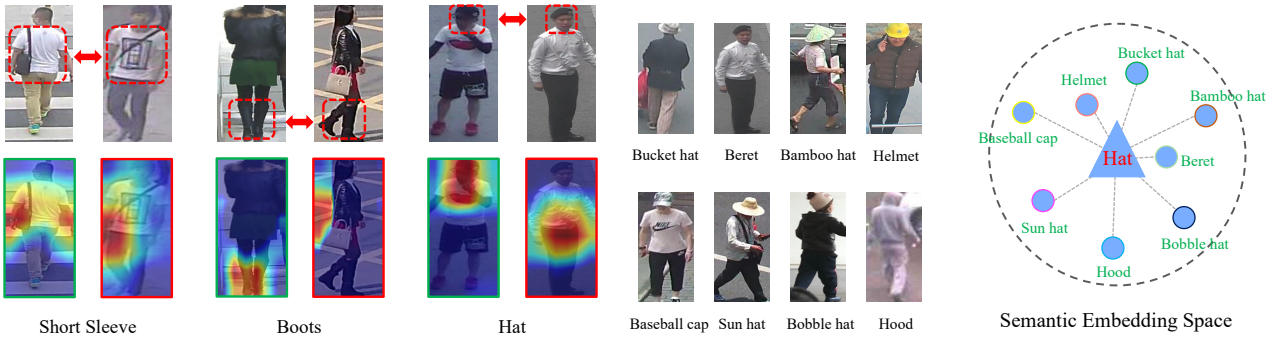
1. Introduction

Pedestrian attribute recognition [26, 20] aims to predict multiple human attributes, such as age, gender, and clothing, as semantic descriptions for a pedestrian image. Due to the ubiquitous application in surveillance scenarios [21], scene understanding [18], and human perception [7], numerous methods [1, 10, 13, 20, 12, 15, 17, 2, 19] have been proposed and significant progress has been made in the last decade.

Existing methods [16, 15, 17, 19] mainly utilize the complicated network, such as Feature Pyramid Network (FPN), to enrich attribute representation from multi-level feature maps, and combine the attention mechanisms to precisely locate attribute-related regions. Recently, VAC [2] utilizes a human prior, that attention regions of random augmentations of the same image are consistent, to improve model robustness. The above methods [16, 15, 17, 19] mainly emphasize learning discriminative attribute features from an individual image, instead of exploiting the relation between different pedestrian images of the same attribute. In contrast, our methods show that mining the inter-image relations between different images of the same attribute can significantly help the model locate attribute-related regions and extract inherent semantic features. We exploit inter-image relations from the perspective of spatial relation and semantic relation.

For the inter-image spatial relation, we hypothesize that the spatial location of the same attribute is basically consistent between different pedestrian images, which is called SPAtial Consistency (SPAC) in this work. For example, the "hat" attribute and the "boots" attribute mostly appears at the top and bottom of the picture, respectively, which is shown in the first row of Figure 1(a). However, we observe that Class Activation Maps (CAMs) [25] of the same attribute of the baseline method have significant location variations. Some examples are shown in the second row of Figure 1(a). These CAMs of the same attribute between different pedestrians are inconsistent, some of which (with red boundary) deviate seriously from attribute-related areas, no matter for the "short sleeve", "boots", or "hat" attributes. This phenomenon contradicts our spatial consistency hypothesis, and indicates that the baseline model easily inclines to focus on the background, irrelevant foreground, or a small part of attribute-related regions, which is called

*Corresponding author.



(a) Spatial consistency on the "short sleeve", "boots", and "hat" attributes. (b) Semantic consistency of different samples on the"hat" attribute.

Figure 1: **Illustration of our main hypothesis on the spatial and semantic consistency**. In (a), CAMs of the baseline method in "short sleeve", "boots", and "hat" attributes of the PA100K are visualized in the second row. Attribute-related regions of each attribute are plotted by the red dotted frame in the first row. Highlighted regions of the second CAM (with red boundary) of each attribute deviate from attribute-related regions severely, which are inconsistent with counterparts of the first CAM (with green boundary). In (b), we present several samples of the "hat" attribute. Although these samples differ greatly in shape, size, and color, the intrinsic semantic features of the "hat" attribute extracted by the model should remain unchanged. Best viewed in color.

the "spatial attention deviation problem" in this work.

For the inter-image semantic relation, inherent semantic features of the same attribute between different images should be consistent, which is called SEMantic Consistency (SEMC) in this work. For example, as illustrated in Figure 1(b), regardless of the difference in shape, size, and color between various samples, the intrinsic semantic features of the "hat" attribute should remain basically unchanged. This property is also indispensable for learning discriminative features and obtaining a robust model.

To achieve the spatial and semantic consistency between pedestrian images of the same attribute, we propose a novel framework composed of the SPAC and SEMC module. Specifically, the SPAC module generates reliable spatial locations for each attribute and maintains a stable spatial memory to suppressing location shift, which is caused by overfitting or label noise. Based on precise spatial locations, the SEMC module extracts intrinsic semantic features and maintains a stable semantic memory to suppress the influence of irrelevant characteristics, such as shape, color, and size for the "hat" attribute.

We make the following three contributions in this work:

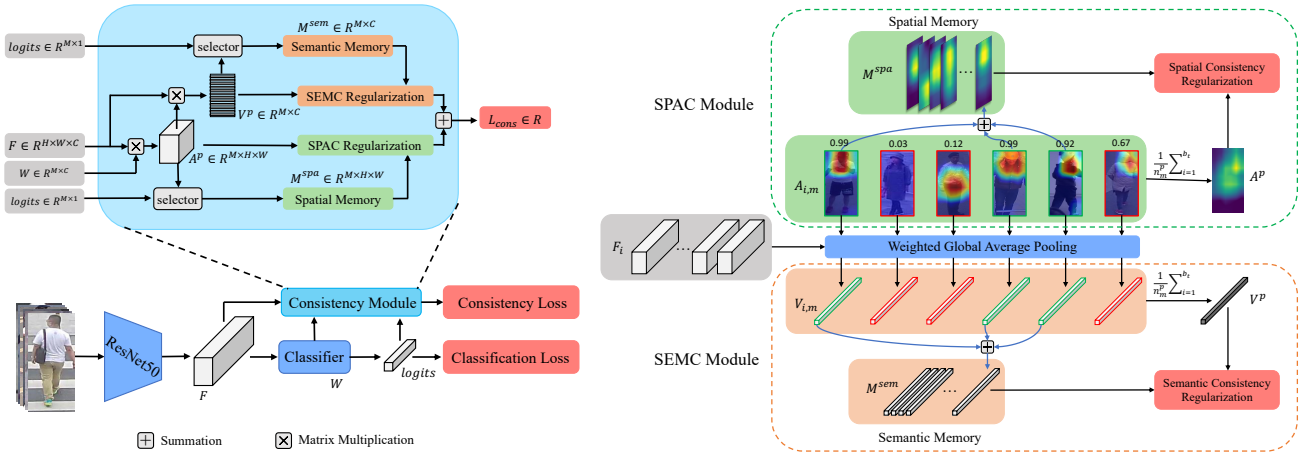
- We establish an effective consistency framework for pedestrian attribute recognition, which makes full use of inter-image spatial and semantic relations between images of the same attribute.
- We design spatial and semantic consistency modules to generate precise spatial attention regions and extract discriminative semantic features for each attribute.
- We confirm the efficacy of the proposed method by achieving state-of-the-art performance on three popular datasets including PA100K, PETA, and RAP.

2. Related Work

Pedestrian attribute recognition has witnessed a fast-growing development recently. Li *et al.* [10] first formulated pedestrian attribute recognition as a multi-label classification task and proposed the weighted sigmoid cross-entropy loss to alleviate the serious imbalance between positive samples and negative samples. To explore attribute context and correlation, the JRL network [20] adopted Long-Shot-Term-Memory [4] to take the pedestrian attribute recognition task as a sequence prediction problem.

Attention mechanism [16, 15, 11, 23, 5] has been widely used in pedestrian attribute recognition to locate attribute-related regions and learn discriminative feature representations. HydraPlus-Net [16] with multi-directional attention modules was introduced to extract pixel-level features and semantic-level features, which were beneficial to locate fine-grained attributes. Based on CAM [25] and EdgeBox [27], Liu *et al.* [15] proposed a Localization Guided Network to extract attribute-related local features. PGDM [11] framework utilized a pre-trained human hose estimator and Spatial Transformer Networks (STNs) [8] to generate reliable attribute-related regions.

Considering the discrimination of multi-scale feature maps and the effectiveness of deep supervisions, WPAL [24], MsVAA [17], and ALM [19] networks are proposed. Yu *et al.* [24] proposed the WPAL network, which introduced a weakly-supervised object detection technique into pedestrian attribute recognition. Sarafianos *et al.* [17] integrated attention mechanism into multi-scale feature maps



(a) Overview of the proposed framework.

(b) Illustration of spatial and semantic consistency on "hat" attribute.

Figure 2: **Illustration of the proposed framework and consistency regularizations**. In (a), we describe two branch structures of the proposed framework and present the pipeline of the consistency module. In (b), we visually demonstrate how to construct spatial and semantic consistency regularizations from the SPAC module (green shading) and SEMC module (orange shading) for the "hat" attribute. For the SPAC module, only reliable CAMs of the "hat" attribute are aggregated into the spatial memory $\boldsymbol{M}^{spa}$ as the supervision of spatial attention regions, but all CAMs $\boldsymbol{A}_{i,m}$ of the "hat" attribute are utilized to compute the SPAC regularization. For the SEMC module, the semantic feature $\boldsymbol{V}_{i,m}$ is firstly extracted by the weighted global average pooling of the feature map F^i, and the weight parameters are the corresponding CAMs. After obtaining the semantic feature $\boldsymbol{V}_{i,m}$, SEMC memory $\boldsymbol{M}^{sem}$ and regularization are constructed as the same as that of the SPAC module. Prediction probabilities of the "hat" attribute are listed above the CAMs. Best viewed in color.

and adopted a variant of focal loss to solve the imbalance between positive and negative samples of the attribute. ALM module [19], which was composed of a Squeeze-and-Excitation (SE) block [6] and a STN [8], was applied to each layer of the Feature Pyramid Network (FPN) [14] to enhance attribute localization. Considering the visual attention regions were consistent between multiple augmentations of the same image, Guo *et al.* [2] proposed an attention consistency loss to get robust attribute locations. In addition, a hierarchical feature embedding (HFE) [22] framework was proposed to learn fine-grained feature embeddings by combining attribute and ID information. Different from previous methods, person ID information was utilized in the HFE framework, which was not provided on the pedestrian attribute recognition task.

Previous methods [10, 11, 16, 20, 17, 19, 2] mainly concentrated on generating precise attribute-related regions and learning to classify attributes from a single image individually. They neither considered the prior spatial structure knowledge of pedestrian attribute, nor exploited the inter-image relation between different pedestrian images of the same attribute. Whereas, both aspects are considered in our proposed method and introduced in Section 3.2 and 3.3.

From the perspective of using inter-image information, the most related method is the JRL network [20]. Based on global feature similarities, the JRL network utilized inter-image information by aggregating several similar pedestrian features to get the final prediction. Different from JRL, our method utilizes spatial and semantic local features of each attribute, and exploits the inter-image relation to construct consistency regularizations as supervision signals of the training process. From the perspective of consistency constraints, the most related method is the VAC model [2], which aimed to make the global attention regions consistent between random augmentations of the same image. However, our method aligns the local attention regions between different pedestrian images of the same attribute. In addition, we also introduce the semantic consistency module to extract discriminative attribute features.

3. Methods

In this section, we first introduce the baseline method. Then, we present the proposed consistency framework, which consists of a classification branch and a consistency branch. The classification branch is completely the same as the baseline network. The consistency branch is divided into spatial consistency module and semantic consistency module, which are introduced separately. The overview of the proposed framework is illustrated in Figure 2(a), and intuitive elaborations of two consistency modules are shown in Figure 2(b). Compared with the baseline method, the proposed method does not introduce extra learnable param-

eters.

3.1. Baseline Method

Given a dataset $\mathbb{D} = \{(\boldsymbol{X}_i, \boldsymbol{y}_i) \mid i = 1, 2, \ldots, N\}$, pedestrian attribute recognition aims to predict multiple attribute $\boldsymbol{y}_i \in \{0, 1\}^M$ to i-th pedestrian image, where N, M denotes the number of images and attributes respectively. The zeros and ones in the attribute vector $\boldsymbol{y}_i$ indicate the absence and presence of the corresponding attributes in the pedestrian image.

Following [12, 19, 17, 2], we formulate pedestrian attribute recognition as a multi-label classification task, and multiple binary classifiers with sigmoid functions [10, 12] are adopted. Binary cross-entropy loss is used as the optimization target:

$$L_{cls} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{i,j} \log\big(p_{i,j}\big) + (1 - y_{i,j}) \log\big(1 - p_{i,j}\big), \tag{1}$$

where $p_{i,j} = \sigma(z_{i,j})$ is the prediction probability of the classifier output logits $z_{i,j}$, and $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function.

3.2. Spatial Consistency Module

In this section, we propose the SPAtial Consistency (SPAC) module combined with spatial consistency regularization to tackle the spatial attention region deviation problem.

SPAC module takes feature map $\boldsymbol{F}_i \in \mathbb{R}^{H \times W \times C}$, classifier weight $\boldsymbol{W} \in \mathbb{R}^{M \times C}$, and logits $\boldsymbol{z}_i \in \mathbb{R}^{M \times 1}$ as inputs, where $\boldsymbol{F}_i$ is the output of the backbone network (ResNet-50 [3] used in our work) for image $\boldsymbol{X}_i$, and H, W, C represent height, width, and channel dimension of feature map respectively. Inspired by the Class Activation Map (CAM) [25], we first obtain spatial attention maps $\boldsymbol{A}_{i,m} \in \mathbb{R}^{H \times W}$ of m-th attribute for image $\boldsymbol{X}_i$ as follows:

$$\boldsymbol{A}_{i,m}(x, y) = \sum_{c=1}^{C} \boldsymbol{W}_{m,c} \boldsymbol{F}_{i,c}(x, y), \quad m \in \{1, 2, \ldots, M\}, \tag{2}$$

where $\boldsymbol{W}_{m,c}$ denotes the c-th element of m-th classifier weight, and $\boldsymbol{F}_{i,c}(x, y)$ indicates the spatial location (x, y) of the c-th channel in the feature map $\boldsymbol{F}_i$. After getting the spatial attention regions of each attribute for every image in a random batch b_t, we adopt the selector — an indicator function takes logits $z_{i,m}$ and ground truth label $y_{i,m}$ as inputs — to aggregate the attention maps of qualified positive samples [1] of the m-th attribute by:

$$\boldsymbol{A}^q_m(x, y) = \frac{1}{n^q_m} \sum_{i=1}^{b_t} \mathbb{1}_{\{\sigma(z_{i,m}) > \tau,\ y_{i,m}=1\}} \boldsymbol{A}_{i,m}(x, y), \tag{3}$$

where $\boldsymbol{A}^q = \{\boldsymbol{A}^q_m | m \in 1, 2, \ldots, M\} \in \mathbb{R}^{M \times H \times W}$ denotes the attention map aggregation of qualified positive samples for each attribute, and $n^q_m = \sum_{i=1}^{b_t} \mathbb{1}_{\{\sigma(z_{i,m}) > \tau,\ y_{i,m}=1\}}$ indicates the number of qualified positive samples of m-th attribute in a random batch b_t. Prediction probabilities of these qualified positive samples are required to be higher than a confidence threshold τ (default as 0.9). The robustness of hyper-parameter τ is validated by the experiments in Figure 3 .

Through strict selection, $\boldsymbol{A}^q_m$ can be regarded as reliable spatial locations for m-th attribute on current batch. To save the reliable spatial location in every batch, the spatial attention maps $\boldsymbol{A}^q_m$ of each attribute are normalized and aggregated into spatial memory $\boldsymbol{M}^{spa} = \{\boldsymbol{M}^{spa}_m | m \in 1, 2, \ldots, M\} \in \mathbb{R}^{M \times H \times W}$ in a momentum updated way to decrease spatial location variation, *i.e.*,

$$\boldsymbol{M}^{spa}_m \leftarrow (1 - \alpha) \times \bar{\boldsymbol{M}}^{spa}_m + \alpha \times \bar{\boldsymbol{A}}^q_m, \tag{4}$$

where $\bar{\boldsymbol{M}}^{spa}_m = \boldsymbol{M}^{spa}_m / \|\boldsymbol{M}^{spa}_m\|_2$, $\bar{\boldsymbol{A}}^q_m = \boldsymbol{A}^q_m / \|\boldsymbol{A}^q_m\|_2$, and $\alpha \in (0, 1]$ is a momentum coefficient. The effect of momentum coefficient α is demonstrated in Figure 4.

As shown in Figure 2(b), due to overfitting and label noise, spatial attention regions of the "hat" attribute deviate from attribute-related regions severely. Model inclines to focus on the background, irrelevant foreground, and a small part of the attribute-related areas. Thus, spatial memory $\boldsymbol{M}^{spa}$, which retains reliable and stable spatial location regions of each attribute, can be taken as the supervision of attribute-related regions to correct the spatial attention deviation. Therefore, based on SPAC module, we propose a spatial consistency regularization L_{spac} by calculating the l_1-distance between the spatial memory $\boldsymbol{M}^{spa}_m$ and the spatial attention map $\boldsymbol{A}^p_m$:

$$L_{spac} = \frac{1}{M} \sum_{m=1}^{M} \|\bar{\boldsymbol{A}}^p_m - \bar{\boldsymbol{M}}^{spa}_m\|_1, \tag{5}$$

$$\boldsymbol{A}^p_m(x, y) = \frac{1}{n^p_m} \sum_{i=1}^{b_t} \mathbb{1}_{\{y_{i,m}=1\}} \boldsymbol{A}_{i,m}(x, y), \tag{6}$$

where $\bar{\boldsymbol{A}}^p_m = \boldsymbol{A}^p_m / \|\boldsymbol{A}^p_m\|_2$, $n^p_m = \sum_{i=1}^{bt} \mathbb{1}_{\{\ y_{i,m}=1\}}$, and b_t indicates the batch size. To take all positive samples into consideration, $\boldsymbol{A}^p_m$ is formulated by averaging spatial attention regions of all positive samples of the m-th attribute in a

[1] We use "positive samples" to represent images that contain target attribute, and "negative samples" to represent images that do not contain target attribute.

random batch. Please note the difference between indicator functions of $\boldsymbol{A}_m^q$ and $\boldsymbol{A}_m^p$.

Overall, to fully utilize the inter-image spatial relation and address the spatial attention region deviation problem, the SPAC module is proposed to extract reliable attribute attention regions $\boldsymbol{A}_m^q$ to update spatial memory and adopt the l_1-distance to align spatial attention regions $\boldsymbol{A}_m^p$ with spatial memory $\boldsymbol{M}_m^{spa}$. Considering the soft weights used in $\boldsymbol{A}_m^p$ and $\boldsymbol{M}_m^{spa}$, we name this method as SSC_{soft}.

3.3. Semantic Consistency Module

Although the SPAC module considers the inter-image spatial relation that attention regions of different images of the same attribute are consistent, inter-image semantic relation has not been utilized, *i.e.*, intrinsic semantic features of the same attribute are consistent between different images. For example, whether the sample is a beret, helmet, bucket hat, or baseball cap, intrinsic semantic features of the "hat" attribute should be consistent. Thus, based on the SPAC module, we propose SEMantic Consistency (SEMC) module to extract intrinsic and discriminative semantic features for each attribute.

According to Equation 2, we first compute the spatial attention map $\boldsymbol{A}_{i,m}$ of m-th attribute for image $\boldsymbol{X}_i$ to obtain attribute-related regions. Then semantic feature vector $\boldsymbol{V}_{i,m} \in \mathbb{R}^{C\times 1}$ can be constructed by weighted global average pooling as:

$$\boldsymbol{V}_{i,m} = \frac{1}{H\times W}\sum_{x=1}^{H}\sum_{y=1}^{W} A_{i,m}(x,y)F_i(x,y), \tag{7}$$

where $F_i(x,y) \in \mathbb{R}^{C\times 1}$ is a spatial feature vector of position (x,y). To provide a consistent supervision for semantic features, we maintain a stable and discriminative semantic memory $\boldsymbol{M}^{sem} = \{\boldsymbol{M}_m^{sem} | m \in 1,2,\ldots,M\} \in \mathbb{R}^{M\times C}$ for each attribute. The selector is adopted in the same way as the SPAC module to aggregate reliable semantic features $V_m^q \in \mathbb{R}^{C\times 1}$ into $\boldsymbol{M}_m^{sem}$ in a momentum updated way:

$$\boldsymbol{V}_m^q = \frac{1}{n_m^q}\sum_{i=1}^{b_t} \mathbb{1}_{\{\sigma(z_{i,m})>\tau,\ y_{i,m}=1\}}\boldsymbol{V}_{i,m}, \tag{8}$$

$$\boldsymbol{M}_m^{sem} \leftarrow (1-\alpha)\times\bar{\boldsymbol{M}}_m^{sem} + \alpha\times\bar{\boldsymbol{V}}_m^q, \tag{9}$$

where $\bar{\boldsymbol{V}}_m^q = \boldsymbol{V}_m^q / \|\bar{\boldsymbol{V}}_m^q\|_2$, $n_m^q = \sum_{i=1}^{b_t}\mathbb{1}_{\{\sigma(z_{i,m})>\tau,\ y_{i,m}=1\}}$, and α is momentum coefficient as same as that of the SPAC module in Equation 4 .

Finally, we design a semantic consistency regularization by computing the l_1-distance between the semantic memory $\boldsymbol{M}_m^{sem}$ and attribute semantic feature $\boldsymbol{V}_m^p$ of all positive samples, which is defined as:

$$L_{semc} = \frac{1}{M}\sum_{m=1}^{M}\|\bar{\boldsymbol{V}}_m^p - \bar{\boldsymbol{M}}_m^{sem}\|_1, \tag{10}$$

$$\boldsymbol{V}_m^p = \frac{1}{n_m^p}\sum_{i=1}^{b_t}\mathbb{1}_{\{y_{i,m}=1\}}\boldsymbol{V}_{i,m}, \tag{11}$$

where $\bar{\boldsymbol{V}}_m^p = \boldsymbol{V}_m^p / \|\boldsymbol{V}_m^p\|_2$, $\bar{\boldsymbol{M}}_m^{sem} = \boldsymbol{M}_m^{sem} / \|\boldsymbol{M}_m^{sem}\|_2$, $n_m^q = \sum_{i=1}^{b_t}\mathbb{1}_{\{y_{i,m}=1\}}$, and the semantic consistency regularization is imposed on semantic features of all positive samples of the m-th attribute.

By bridging the gap between semantic features of different samples of the same attribute, the SEMC module can extract intrinsic and discriminative semantic features for each attribute and eliminate the interference of attribute-irrelevant characteristics (such as shape, size, and color in the "hat" attribute).

3.4. Loss Function

As commonly adopted in most existing methods [17, 2, 19], the weighted binary cross-entropy loss is also utilized in the classification branch of the proposed method as classification loss, which is formulated as :

$$L_{cls} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\omega_{i,j}(y_{i,j}\log(p_{i,j}) + (1-y_{i,j})\log(1-p_{i,j})), \tag{12}$$

$$\omega_{i,j} = y_{i,j}e^{1-r_j} + (1-y_{i,j})e^{r_j}, \tag{13}$$

where r_j is the positive sample ratio of j-th attribute in the training set.

The final loss function L is a weighted summation of the classification loss, SPAC regularization, and SEMC regularization:

$$L = L_{cls} + \mathbb{1}_{\{e>i_e\}}(\lambda_1 L_{spac} + \lambda_2 L_{semc}), \tag{14}$$

where $\lambda_1 = 1$, $\lambda_2 = 0.1$ is set as default in all experiments if not specially specified. Current epoch number in the training stage is indicated by $e \in \{0,\cdots 30\}$, and initial epoch i_e is used to ensure reliable consistency memory and effective consistency regularization.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. We perform experiments on the PETA [1], RAP [13], and PA100K [16]. The PEdesTrian Attribute (PETA) dataset [1] is collected from 10 small-scale person datasets and consists of 19,000 person images, which is divided into 9500 images for the training set, 1900 for the validation set, and 7600 for the test set. Each image is labeled with 61 binary attributes and 4 multi-class attributes.

Table 1: **Performance comparison with state-of-the-art methods on the PETA, RAP, and PA100K datasets.** Five metrics, mean accuracy (mA), accuracy (Accu), precision (Prec), recall (Recall), F1 are evaluated. To make a fair comparison, we also report our reimplementation performance for the MsVAA, VAC, and ALM methods. The first and second highest scores are represented by red font and blue font respectively. The difference between SSC_{soft}, SSC_{hard}, and SSC_{fix} lies in the implementation of M^{spa} and A^q, and we detail it in Section 4.4 .

Method	Backbone	PETA					PA100K					RAP				
		mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1
DeepMAR [10]	CaffeNet	82.89	75.07	83.68	83.14	83.41	72.70	70.39	82.24	80.42	81.32	73.79	62.02	74.92	76.21	75.56
HPNet[16]	InceptionNet	81.77	76.13	84.92	83.24	84.07	74.21	72.19	82.97	82.09	82.53	76.12	65.39	77.33	78.79	78.05
JRL [20]	AlexNet	85.67	–	86.03	85.34	85.42	–	–	–	–	–	77.81	–	78.11	78.98	78.58
LGNet [15]	Inception-V2	–	–	–	–	–	76.96	75.55	86.99	83.17	85.04	78.68	68.00	80.36	79.82	80.09
PGDM [11]	CaffeNet	82.97	78.08	86.86	84.68	85.76	74.95	73.08	84.36	82.24	83.29	74.31	64.57	78.86	75.90	77.35
MsVAA[17]	ResNet101	84.59	78.56	86.79	86.12	86.46	–	–	–	–	–	–	–	–	–	–
VAC [2]	ResNet50	–	–	–	–	–	79.16	**79.44**	**88.97**	86.26	**87.59**	–	–	–	–	–
ALM [19]	BN-Inception	**86.30**	**79.52**	85.65	88.09	**86.85**	80.68	77.08	84.21	88.84	86.46	81.87	68.17	74.71	86.48	80.16
MsVAA[17] *	ResNet50	84.35	78.69	87.27	85.51	86.09	80.10	76.98	86.26	85.62	85.50	79.75	65.74	77.69	78.99	77.93
VAC [2] *	ResNet50	83.63	78.94	**87.63**	85.45	86.23	79.04	78.95	88.41	86.07	86.83	78.47	68.55	**81.05**	79.79	80.02
ALM [19] *	ResNet50	85.50	78.37	83.76	**89.13**	86.04	79.26	78.64	87.33	86.73	86.64	81.16	67.35	74.97	85.36	79.39
Baseline	ResNet50	81.15	77.96	**88.19**	83.77	85.56	78.53	78.87	**88.99**	85.38	86.34	76.09	**68.66**	**83.74**	77.44	79.50
SSC_{soft}	ResNet50	**86.52**	78.95	86.02	87.12	**86.99**	**81.87**	**78.89**	85.98	**89.10**	86.87	**82.77**	**68.37**	75.05	**87.49**	**80.43**
SSC_{hard}	ResNet50	85.92	78.53	86.31	86.23	85.96	81.02	78.42	86.39	87.55	86.55	82.14	68.16	77.87	82.88	79.87
SSC_{fix}	ResNet50	86.07	**79.23**	84.58	**89.26**	86.54	**81.70**	78.85	85.80	**88.92**	**86.89**	**82.83**	68.16	74.74	**87.54**	**80.27**

We follow the common experimental protocol [12, 17, 19], and only 35 attributes whose positive ratios are higher than 5% are used for evaluation. The Richly Annotated Pedestrian (RAP) attribute dataset [13] consists of 33,268 images for training and 8,317 images for testing, a total of 41,585 images extracted from 26 indoor surveillance cameras. Each image is labeled with 69 binary attributes and 3 multi-class attributes. Following the official protocol [13], 51 binary attributes are adopted to evaluate the recognition performance. The PA100K dataset [16] consists of 100,000 pedestrian images and is split into training, validation, and test sets with a ratio of 8:1:1. Each image is described with 26 commonly used attributes. Considering the identical pedestrian identities between training set and test set [9] on the RAP and PETA, performance on the largest dataset PA100K is more convincible.

Evaluation Protocal. Two types of metrics, *i.e.*, a label-based metric and four instance-based metrics, are adopted to evaluate attribute recognition performance [12]. For the label-based metric, we compute the mean value of classification accuracy of positive samples and negative samples as the metric for each attribute. Then we take an average over all attributes as mean accuracy. For instance-based metrics, accuracy, precision, recall, and F1-score are used.

4.2. Implementation Details

The proposed method is implemented with PyTorch and trained in an end-to-end manner. We adopt ResNet50 [3] as the backbone network to extract pedestrian image features for a fair comparison. Pedestrian images are resized to 256×192 as inputs. Random horizontal mirroring, padding, and random crop are used as augmentations. Adam is employed for training with the weight decay of 0.0005. The initial learning rate equals 0.0001, and the batch size is set to 64. Plateau learning rate scheduler is used with reduction factor 0.1 and loss patience 4. The total epoch number of the training stage is 30. Momentum coefficient $\alpha = 0.9$, confidence threshold $\tau = 0.9$ by default. To obtain the stable and reliable spatial memory M_m^{spa} and semantic memory M_m^{sem}, consistency regularizations are added to the classification loss after epoch 4, *i.e.*, $i_e = 4$ in Equation 14 .

4.3. Comparison to the State of the Arts

In Table 1, we compare the performance of the proposed methods with several existing algorithms on the PETA, RAP, and PA100K. For a fair comparison, besides the performance reported by the papers [17, 2, 19], we also report the performance of our reimplements based on the same setting described in Section 4.2.

Compared with the performance reported by the paper of MsVAA [17], VAC [2], and ALM [19] methods, the SSC_{soft} model achieves better performance on the PETA, PA100K, and RAP without increasing learnable parameters. Compared to the MsVAA model adopted ResNet101 as the backbone network, we achieve 1.93% and 0.53% performance improvements in mA and F1 on the PETA dataset. Compared to the ALM model, which utilizes the complicated combination of FPN, STN and SE modules introducing extra 17% parameters, the SSC_{soft} method achieves 0.22%, 1.19%, and 0.9% performance improvements in mA on three popular datasets. Besides, compared with the performance achieved by our reimplementation of the MsVAA, VAC, and ALM methods, the performance of the SSC_{soft}

*Results are reimplemented in the same setting for a fair comparison.

Table 2: **Ablation study of each component of our method on the PETA, PA100K, RAP**. Performance improved by spatial consistency (SPAC) and semantic consistency (SEMC) regularizations validates the effectiveness of our methods. We use SSC_{soft} model as default.

Method			PETA					PA100K					RAP				
SEMC	SPAC	Weighted Loss	mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1
-	-	-	81.15	77.96	88.19	83.77	85.56	78.53	78.87	**88.99**	85.38	86.75	76.09	68.66	**83.74**	77.44	80.06
✓	-	-	82.34	78.51	**88.29**	84.36	85.94	78.63	78.68	88.63	85.52	86.64	76.45	68.58	83.08	77.81	79.97
-	✓	-	84.08	78.85	87.66	85.32	86.19	79.15	78.62	88.24	85.71	86.57	78.55	68.60	82.09	79.34	79.99
-	-	✓	84.17	78.81	87.30	85.58	86.15	79.59	78.86	87.70	86.65	86.77	79.46	66.55	78.39	79.62	78.58
✓	✓	-	84.90	78.49	86.44	85.90	85.91	80.09	**79.11**	88.37	86.33	**86.95**	80.26	**68.77**	80.64	80.07	80.29
✓	✓	✓	**86.52**	**78.95**	86.02	**87.12**	**86.99**	**81.87**	78.86	85.98	**89.10**	86.87	**82.77**	68.37	75.05	**87.49**	**80.43**

method has significant improvements from 1.02% to 4.30% in mA on the PETA, PA100K, and RAP, which fully demonstrates the effectiveness of our method.

It can be noticed that the proposed spatial and semantic consistency method substantially outperforms the visual attention consistency (VAC) method [2]. The VAC method hypothesis that global attention regions of random augmentations of the same image are consistent. However, the VAC method focuses on global attention regions of an individual image and cannot generate precise local attention regions for each fine-grained attribute. In addition, for a pair of augmentations of the same image, if global attention regions of one augmentation are precise, the VAC method can improve the performance by aligning the global attention regions of another augmentation with these of the current one. However, if attention regions of both augmentations are inaccurate, the VAC method cannot solve the attention region deviation problem, which can be addressed by our proposed method.

4.4. Ablation Study and Discussion

In this section, we first investigate the effect of the SPAC and SEMC module by conducting analytical experiments on all three datasets. We then introduce two variants of our methods to demonstrate the effectiveness of spatial and semantic consistency regularizations. Quantitative performance improvements of each attribute on three datasets are presented in the supplementary material.

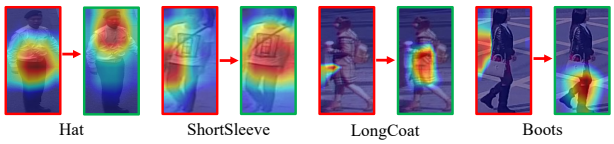
As shown in Table 2, compared to the baseline method, we have the following observations. First, adopting the SEMC module alone can hardly bring performance improvement. The results prove that, without correct attention regions, attribute semantic features lack discrimination and contain more noise, which is in line with the intuitive hypothesis. Second, adopting the SPAC module can directly bring 2.93%, 0.62%, 2.46% performance improvements in mA on the PETA, PA100K, and RAP, respectively. This improved performance demonstrates that spatial consistency regularization is beneficial for locating the attribute-related regions. Third, when the SPAC module and SEMC module are jointly adopted, our method improves the performance over the baseline model by 3.75%, 1.56%, 4.17% in mA on the PETA, PA100K, and RAP.

To further validate the reasonableness of proposed spatial and semantic consistency regularizations, we implement our method with two variants SSC_{hard} and SSC_{fix}. For the SSC_{hard} method, we change the $\boldsymbol{A}^q$ in Equation 3 and $\boldsymbol{M}^{spa}$ in Equation 4 of SPAC module from soft attention maps to binary (hard) attention maps based on a threshold $th_{hard} = 0$. For the SSC_{fix} method, we first train a baseline model and obtain the qualified CAMs $\boldsymbol{A}^q$ of positive samples for each attribute according to Equation 3. Then, we fix $\boldsymbol{M}^{spa}$ as $\boldsymbol{A}^q$ instead of momentum updating to training a new model SSC_{fix}. The experimental results of two variants are listed in Table 1. Although method SSC_{hard} assigns the same weight to each pixel of the region of interest, which is not as flexible as SSC_{soft} and achieves slightly reduced performance, it still achieves competitive performance on PA100k and RAP. Since the SSC_{soft} and SSC_{fix} method can get reliable and accurate spatial attention regions $\boldsymbol{M}^{spa}$, they both achieve the state-of-the-art performance. However, compared to the SSC_{fix} method, SSC_{soft} with the momentum updated memory can avoid a two-stage training process and is more suitable for industry application.

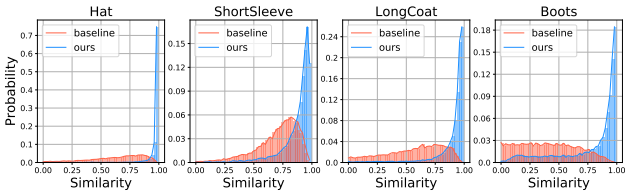
4.5. Effects of SPAC and SEMC Module

Spatial and semantic consistency regularizations are two complementary and indispensable parts of a powerful model. The SPAC module can enhance the localization capability of the backbone network without being disturbed by overfitting and label noise. Based on the precise spatial attention regions of attributes, the backbone network further benefits from the SEMC module to extract intrinsic and discriminative semantic features.

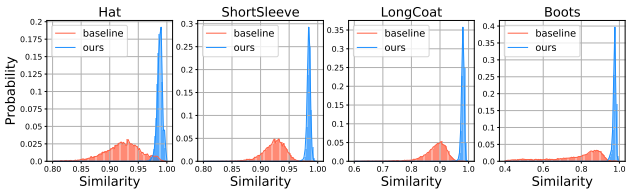
To validate the effectiveness of the proposed SPAC module and SEMC module, we visualize the spatial attention regions and the similarity distributions of spatial and semantic features in Figure 3. The similarities are computed between each pair of images of the same attribute on PA100K. The higher the similarity, the more consistent the attention regions and semantic features of the two images with the same attribute. Compared to the baseline method, as shown in Figure 3(b) and Figure 3(c), we observe that plenty of



(a) Visualization of spatial attention regions $\boldsymbol{A}_{i,m}$ between the baseline method (red boundary) and the proposed method (green boundary).



(b) Comparison of the similarity distribution of spatial attention regions $\boldsymbol{A}_{i,m}$.



(c) Comparison of the similarity distribution of semantic features $\boldsymbol{V}_{i,m}$.

Figure 3: **Illustration of the effect of the SPAC and SEMC module.** We take the "hat", "short sleeve", "long-coat", and "boots" attributes as examples to show (a) the spatial attention regions, (b) similarity distribution of spatial attention regions, and (c) similarity distribution of semantic features between different images of the same attribute. Compared with the baseline method, most similarities of the proposed method are concentrated near 1, which proves the consistency of spatial attention regions and semantic features for each attribute.

similarities concentrate on 1, making the probability curve rise rapidly near 1. The same phenomenon can also be observed in other attributes of the PA100K, RAP, and PETA as shown in supplementary material.

4.6. Hyperparameter Evaluation

There are mainly three key hyperparameters in our method, which are confidence threshold τ, initial epoch i_e, and momentum coefficient α. We set $\tau = 0.9$, $i_e = 4$, $\alpha = 0.9$ if not specially specified. To fully demonstrate the effect of hyperparameters, the following experiments are all conducted on the largest pedestrian attribute dataset PA100K.

Confidence threshold τ is used in Equation 3 and Equation 8 to select reliable spatial attention feature maps $\boldsymbol{A}^q$ and semantic feature vectors $\boldsymbol{V}^q$, which are aggregated to spatial memory $\boldsymbol{M}^{spa}$ and semantic memory $\boldsymbol{M}^{sem}$. As shown in Table 3, with the increase of the confidence threshold τ, there is an obvious performance improvement in mA from 79.63 to 81.28. It is easy to infer that higher confidence threshold τ can select more precise spatial attention feature maps and more discriminative semantic feature vectors. Little performance fluctuation in the other four metrics shows the robustness of the threshold τ.

Table 3: **Experiments on the confidence threshold τ.**

Confidence Threshold	mA	Accu	Prec	Recall	F1
$\tau = 0$	79.63	78.61	**86.89**	87.19	86.63
$\tau = 0.3$	80.08	78.21	86.65	86.88	86.35
$\tau = 0.5$	80.90	78.40	86.49	87.36	86.51
$\tau = 0.7$	80.79	78.20	86.37	87.14	86.35
$\tau = 0.9$	**81.87**	**78.86**	85.98	**89.10**	**86.87**

Table 4: **Experiments on the momentum coefficient α.**

Momentum Coefficient	mA	Accu	Prec	Recall	F1
$\alpha = 0.1$	80.89	78.23	86.45	87.12	86.37
$\alpha = 0.3$	81.04	78.19	86.37	87.05	86.32
$\alpha = 0.5$	81.10	78.17	86.31	87.19	86.33
$\alpha = 0.7$	81.15	78.34	**86.46**	87.35	86.49
$\alpha = 0.9$	**81.87**	**78.86**	85.98	**89.10**	**86.87**

Momentum coefficient α is adopted in Equation 4 and Equation 9 to determine the degree of integration of historical features and current batch features. The larger α is, the fewer historical features are retained. As shown in Table 4, more historical features can bring a few performance improvements.

5. Conclusion

This paper proposes the consistency framework for pedestrian attribute recognition, which makes full use of the inter-image relation of the same attribute and tackles the spatial attention region deviation problem. Specifically, we propose the SPAC module to pay attention to specific attribute-related spatial regions. We also propose the SEMC module to extract intrinsic and discriminative semantic features for each attribute. Moreover, we implement two variants of our method to demonstrate the efficacy of consistency regularizations. The ablation experiments show that two consistency modules can both bring performance improvements. Our proposed method achieves outstanding performance consistently on the PA100K, RAP, and PETA.

6. 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

[1] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792, 2014. 1, 5

[2] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 729–739, 2019. 1, 3, 4, 5, 6, 7

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4, 6

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[5] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *Proceedings of the European Conference on Computer Vision*, 2020. 2

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 3

[7] Houjing Huang, Wenjie Yang, Jinbin Lin, Guan Huang, Jiamiao Xu, Guoli Wang, Xiaotang Chen, and Kaiqi Huang. Improve person re-identification with part awareness learning. *IEEE Transactions on Image Processing*, 29:7468–7481, 2020. 1

[8] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015. 2, 3

[9] Jian Jia, Houjing Huang, Wenjie Yang, Xiaotang Chen, and Kaiqi Huang. Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method. *arXiv preprint arXiv:2005.11909*, 2020. 6

[10] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *2015 3rd IAPR Asian Conference on Pattern Recognition*, pages 111–115. IEEE, 2015. 1, 2, 3, 4, 6

[11] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *2018 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2018. 2, 3, 6

[12] 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*, 28(4):1575–1590, 2018. 1, 4, 6

[13] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016. 1, 5, 6

[14] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 3

[15] Pengze Liu, Xihui Liu, Junjie Yan, and Jing Shao. Localization guided learning for pedestrian attribute recognition. *arXiv preprint arXiv:1808.09102*, 2018. 1, 2, 6

[16] 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*, pages 350–359, 2017. 1, 2, 3, 5, 6

[17] 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*, pages 680–697, 2018. 1, 2, 3, 4, 5, 6

[18] Jing Shao, Kai Kang, Chen Change Loy, and Xiaogang Wang. Deeply learned attributes for crowded scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4657–4666, 2015. 1

[19] 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*, pages 4997–5006, 2019. 1, 2, 3, 4, 5, 6

[20] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Attribute recognition by joint recurrent learning of context and correlation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 531–540, 2017. 1, 2, 3, 6

[21] Xiao Wang, Shaofei Zheng, Rui Yang, Bin Luo, and Jin Tang. Pedestrian attribute recognition: A survey. *arXiv preprint arXiv:1901.07474*, 2019. 1

[22] Jie Yang, Jiarou Fan, Yiru Wang, Yige Wang, Weihao Gan, Lin Liu, and Wei Wu. Hierarchical feature embedding for attribute recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13055–13064, 2020. 3

[23] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1389–1398, 2019. 2

[24] Kai Yu, Biao Leng, Zhang Zhang, Dangwei Li, and Kaiqi Huang. Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. *arXiv preprint arXiv:1611.05603*, 2016. 2

[25] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 1, 2, 4

[26] Jianqing Zhu, Shengcai Liao, Zhen Lei, Dong Yi, and Stan Z Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 331–338. IEEE, 2013. 1

[27] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *Proceedings of the European Conference on Computer Vision*, pages 391–405, 2014. 2