Selective and Orthogonal Feature Activation for Pedestrian Attribute Recognition

Junyi Wu[1,2,3], **Yan Huang**[4*], **Min Gao**[5], **Yuzhen Niu**[3], **Mingjing Yang**[5], **Zhipeng Gao**[1,2], **Jianqiang Zhao**[1,2]

[1] AI Research Center, Xiamen Meiya Pico Information Company Ltd., Xiamen, China
[2] Xiamen Meiya Pico Information Security Research Institute Company Ltd., Xiamen, China
[3] College of Computer and Data Science, Fuzhou University, Fuzhou, China
[4] Institute of Automation, Chinese Academy of Sciences, Beijing China
[5] College of Physics and Information Engineering, Fuzhou University, Fuzhou, China
{Junyi.Wu-1, huangyan.750, Min.Gao-1}@outlook.com, yuzhenniu@gmail.com

Abstract

Pedestrian Attribute Recognition (PAR) involves identifying the attributes of individuals in person images. Existing PAR methods typically rely on CNNs as the backbone network to extract pedestrian features. However, CNNs process only one adjacent region at a time, leading to the loss of long-range inter-relations between different attribute-specific regions. To address this limitation, we leverage the Vision Transformer (ViT) instead of CNNs as the backbone for PAR, aiming to model long-range relations and extract more robust features. However, PAR suffers from an inherent attribute imbalance issue, causing ViT to naturally focus more on attributes that appear frequently in the training set and ignore some pedestrian attributes that appear less. The native features extracted by ViT are not able to tolerate the imbalance attribute distribution issue. To tackle this issue, we propose two novel components: the Selective Feature Activation Method (SFAM) and the Orthogonal Feature Activation Loss. SFAM smartly suppresses the more informative attribute-specific features, compelling the PAR model to capture discriminative features from regions that are easily overlooked. The proposed loss enforces an orthogonal constraint on the original feature extracted by ViT and the suppressed features from SFAM, promoting the complementarity of features in space. We conduct experiments on several benchmark PAR datasets, including PETA, PA100K, RAPv1, and RAPv2, demonstrating the effectiveness of our method. Specifically, our method outperforms existing state-of-the-art approaches by GRL, IAA-Caps, ALM, and SSC in terms of mA on the four datasets, respectively.

Introduction

Pedestrian attribute recognition (PAR), as a multi-label classification problem, aims to predict a set of semantic attributes (*e.g.,* age, gender, hair style, *etc.*) (Wu et al. 2022a; Tan et al. 2020). Due to its ubiquitous applications in surveillance and public security, many efforts have been made to promote its performance in real-world scenarios (Feris et al. 2014; Lin et al. 2019).

CNNs have demonstrated extraordinary capabilities in various human-centric vision tasks (Wu et al. 2022b; Huang et al. 2018, 2019b, 2021a; Zhang and Wang 2023; Zhang et al. 2021), prompting numerous efforts in PAR (Sarfraz et al. 2017; Lin et al. 2019; Zhao et al. 2018; Xiang et al. 2019; Yang et al. 2021). These PAR works employ CNN-based architecture (*e.g.,* the commonly used ResNet (He et al. 2016) and InceptionNet (Szegedy et al. 2016)) to exploit attribute-related features from specific body regions. Despite the promising performance achieved in PAR due to the success of CNN architectures, the task is far from being solved due to challenges posed by poor pedestrian image quality (*e.g.,* low resolution, human pose, and lighting changes). Furthermore, PAR differs from general single-label image classification as it involves predicting the presence of multiple attributes in a single image, making the contextual relationship between different attributes crucial. For instance, when the model determines a high probability of a person wearing a skirt, the contextual relation may indicate a higher likelihood of attributes such as being female or having long hair, which are often associated with wearing a skirt. However, exploiting the contextual relation between attributes remains challenging for CNNs, as highlighted in existing works (He et al. 2021; Tan et al. 2020; Zhao et al. 2018; Yang et al. 2021). This limitation primarily arises from the fact that the convolutional and downsampling operators process one adjacent region at a time (He et al. 2021), hindering the learning of global context and capturing the long-range dependencies between regions in close proximity, even when they may contain attributes with contextual relations.

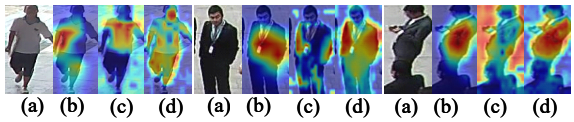

Figure 1: Visualization of heat maps: (a) Original Images, (b) CNN-based method, (c) Transform-based method, and (d) our SOFAFormer which captures more robust features.

Recently, the Vision Transformer (ViT) (Dosovitskiy et al. 2021) has gained popularity as a preferred choice for a variety of computer vision tasks due to its impressive performance and versatility. In contrast to CNNs, which rely on convolutional and downsampling operators, ViT utilizes multi-head self-attention modules to process different re-

*Corresponding author.

gions simultaneously. This allows ViT to capture long-range dependencies and preserve fine-grained details. As depicted in Fig. 1, CNNs tend to focus on learning local feature information across the entire image, while the feature extraction process of ViT can attend to different regions more effectively. Although ViT has the capability to establish the contextual relations, it is still necessary to ensure that the extracted features represent all attributes accurately.

Despite the capabilities of learning contextual relations using ViT, achieving a comprehensive attribute representation for transformer-based PAR remains a challenging and inadequate task. This is primarily due to the issue of imbalanced attribute distribution, which adversely affects the performance of PAR. Specifically, the PAR model may tend to prioritize capturing the attributes that are annotated more frequently, such as hair style, clothing type, *etc.* (referred to as major-annotated attributes[1]), while the attributes that are annotated less frequently such as glasses, backpack, *etc.* (referred to as minor-annotated attributes[2]), may receive less attention or overlooked. In other words, major-annotated attributes may be overemphasized, leading to overfitting, while minor-annotated attributes may be neglected. Consequently, another key problem in PAR is how to further mine more comprehensive attribute representations based on ViT.

One intuitive solution to enhance feature robustness is to incorporate body part information (Li et al. 2018a; Yang et al. 2016; Huang et al. 2019c,a; Zhang, Ren, and Li 2020; Liu et al. 2018; Zhang et al. 2014; Zhao et al. 2018) and extract corresponding fine-grained features. In previous methods, auxiliary body information extractors (*e.g.*, key points and human parsing) are introduced to divide a pedestrian into different semantic regions. Subsequently, these different body parts are aggregated to predict a series of attributes. Although incorporating body part information has empirically proven to be effective in enhancing robustness against appearance variations, the performance of PAR heavily relies on accurate localization of body regions. However, these body regions extractors are not specifically designed for PAR, making it challenging to ensure precise localization and requiring additional computational resources for sophisticated part localization. Another solution is to introduce attention mechanisms to enforce the PAR model to capture more discriminative features (Guo et al. 2019; Sarafianos, Xu, and Kakadiaris 2018; Jia, Chen, and Huang 2021). However, these attention modules introduce additional parameters, which can potentially increase the risk of overfitting (Wu et al. 2023a; Huang et al. 2021b).

In this paper, we propose a ViT framework with Selective and Orthogonal Feature Activation (SOFAFormer) to address the limitations mentioned above and exploit more robust representations from all attribute-specific regions. In our SOFAFormer, we leverage ViT to capture the contextual relations between different attribute-specific regions. However, relying solely on ViT may not provide sufficient robustness to represent all pedestrian attributes simultaneously due to the imbalanced attribute distribution.

To address this issue, we propose a Selective Feature Activation Method (SFAM) in order to prevent the model from overfitting to major-annotated attributes and exploit minor-annotated attribute features. SFAM calculates the feature activation mapping of the final output and sorts it from minimum to maximum. We then apply a threshold (referred to τ in Eq. 7) to select the first few largest feature activations and scale them using a suppression ratio (referred to α in Eq. 7). This approach directs attention in the PAR model towards the lower feature activations, which correspond to minor-annotated attribute-specific regions. By combining the output from ViT with the output from SFAM, we obtain two features that complement each other, resulting in a more comprehensive representation. In our SOFAFormer, both features share a classifier layer, which promotes high correlation to facilitate quick convergence. However, this high correlation reduces the complementarity between the two features. To address this issue, we propose an Orthogonal Feature Activation Loss ($\mathcal{L}_{\mathcal{OFA}}$) that aims to minimize the feature correlation between the two features. $\mathcal{L}_{\mathcal{OFA}}$ encourages the exploration of different feature spaces, potentially leading to a more comprehensive attribute representation.

In summary, main contributions of this paper can be summarized in four-fold:

- We propose a novel SOFAFormer method for PAR, which enables the learning of more comprehensive feature representations. In contrast to previous CNN-based methods, we leverage ViT as the backbone for PAR to capture the contextual relations between different attribute-specific regions.
- To obtain more robust representations, we introduce SFAM to enable the model to focus on the lower informative attribute-specific regions that contain minor-annotated attributes. By combining the features extracted from ViT with the features obtained from SFAM, we achieve a comprehensive representation that captures both major-annotated and minor-annotated attributes. Features from ViT and SFAM can complement each other to obtain a comprehensive representation.
- We design $\mathcal{L}_{\mathcal{OFA}}$ to decrease feature correlation. $\mathcal{L}_{\mathcal{OFA}}$ encourages the PAR model to learn attribute representation from different feature spaces (*i.e.,* from ViT and SFAM), further enhancing the overall performance of the PAR model.
- We perform conduct comprehensive experiments on several PAR datasets, including PETA (Deng et al. 2014), PA100K (Liu et al. 2017), RAPv1 (Li et al. 2016), and RAPv2 (Li et al. 2018b). Our SOFAFormer outperforms existing methods and achieves a new state-of-the-art (SOTA) performance.

Related Work

CNN-Based Methods for PAR

PAR has two key steps: obtaining a feature representation and enhancing robustness of that representation. To extract

[1]Attributes that must be present on the pedestrian, such as hair style, clothing style, *etc.*.

[2]Attributes that may be present on the pedestrian, such as glasses, backpack, *etc.*.

the feature representation, many methods mainly rely on CNNs as the backbone network. Among them, there are approaches that incorporate body region information or attention modules to capture more discriminative feature, which is robust to multiple pedestrian variations.

CNN-Based Methods Embedded Body Information Li *et al.* (Li et al. 2018a) proposed a pose guided deep model (PGDM), which leverages a pre-trained pose estimate model to localize body parts and extract corresponding body features. Zhang *et al.* (Zhang, Ren, and Li 2020) used human pose keypoints as auxiliary information to supervise a deep template matching network, ensuring proper attribute specific region alignment. Yang *et al.* (Yang et al. 2016) incorporated key point estimation and PAR into a multi-task training framework, using key point to obtain body parts prior for feature learning. Zhang *et al.* (Zhang et al. 2014) first detected the poselets and then employed CNNs to extract feature representations from local patches and the entire pedestrian image. Liu *et al.* (Liu et al. 2018) introduced EdgeBoxes (Zitnick and Dollár 2014) to localize attributed-related regions and extract discriminative features from different body parts.

The aforementioned methods employ CNNs as the backbone network, which may lack the ability to capture long-range contextual relations. Additionally, while incorporating body parts can reveal more fine-grained features, it often requires additional computational resources. Moreover, the robustness of the feature representation becomes dependent on the accurate division of the body extractor, which is challenging to guarantee.

CNN-Based Methods Applied Attention Mechanism Guo *et al.* (Guo et al. 2019) formulated a two-branch network that takes an original image and its transformed version as inputs. They introduced a new attention consistency loss to measure the consistency of attention heat maps between the two branches. Tan *et al.* (Tan et al. 2019) proposed three types of attention mechanisms (*e.g.,* parsing attention, label attention and spatial attention) to explore correlated and complementary information. Sarafianos *et al.* (Sarafianos, Xu, and Kakadiaris 2018) presented a visual attention mechanism that extracts and aggregates visual attention masks at different scales. Jia *et al.* (Jia, Chen, and Huang 2021) designed a spatial and semantic consistency (SSC) framework, which incorporates two complementary regularizations to explore inter-image relations from the perspectives of spatial and semantic relations for each attribute.

The attention mechanisms mentioned above aim to guide the PAR model to focus on attribute-specific regions and enhance the overall representation of all attributes. However, these attention mechanisms may still struggle to capture long-range contextual relations and often introduce additional training parameters. The inclusion of new parameters not only requires increased computational resources but also carries the risk of overfitting the model to the training data.

In contrast to previous PAR methods, our SOFAFormer framework takes a different approach by leveraging ViT as the backbone network. This allows the network to effectively capture long-range contextual relations between different attribute-specific regions. To address the issue of neglecting minor-annotated attributes, we introduce a parameter-free SFAM module that encodes these attributes with more fine-grained features, enhancing the overall representation. Additionally, we incorporate a novel loss function called $L_{\mathcal{OFA}}$, which encourages the features to be more comprehensive and discriminative.

Transformer in Vision

Transformers have gained significant attention from researchers and have been widely studied in various computer vision tasks, including image classification (Dosovitskiy et al. 2021), object tracking (Chen et al. 2021), and person re-identification (He et al. 2021; Wu et al. 2023b). He *et al.* (He et al. 2021) were the first to explore the use of transformers as the backbone network for object re-identification and demonstrated that transformers are more suitable for retrieval tasks compared to CNNs. Ma *et al.* (Ma, Zhao, and Li 2021) introduced transformers to capture part-aware long-term correlations and extract robust feature representations for occluded person re-identification. Lanchantin *et al.* (Lanchantin et al. 2021) utilized transformers to exploit complex dependencies among visual features and labels for multi-label image classification.

Inspired by these works, we propose utilizing transformers to capture long-range contextual relations for the PAR task. Transformers have shown their effectiveness in modeling complex relationships, making them a promising choice for capturing contextual information among attribute-specific regions in PAR.

Architecture of SOFAFormer

Our SOFAFormer is proposed to learning robust feature representations without introducing additional parameters, except for the parameters of ViT. As shown in Fig. 2, our SOFAFormer comprises three main modules: 1) Attribute Feature Extraction Module, 2) Selective Feature Activation Method (SFAM), and 3) Orthogonal Feature Activation Loss ($\mathcal{L}_{\mathcal{OFA}}$). We utilize ViT to extract pedestrian attributes from input images, enabling the model to establish long-range contextual relations between different attribute-specific regions. To ensure that all attribute-specific regions are effectively encoded, we propose a parameter-free SFAM that suppresses the largest feature activations in order to encode the lower informative attribute-specific regions. Finally, we introduce incorporate the $\mathcal{L}_{\mathcal{OFA}}$ to decrease the low correlation between different features, thereby promoting a more comprehensive feature space. This section will introduce each module in detail. The detailed process of our SOFAFormer can be found in supplementary material.

Attribute Feature Extraction Module

Given a PAR dataset $\mathcal{D} = \{(\boldsymbol{X}_i, y_i) \mid i = 1, 2, \ldots, N\}$, PAR aims to predict a series of attributes $y_i \in \{0, 1\}^M$ from an image *e.g., i*-th pedestrian image in the dataset, where N, M represents the number of images and attributes, respectively. As depicted in Fig. 2, the attribute feature extractor of

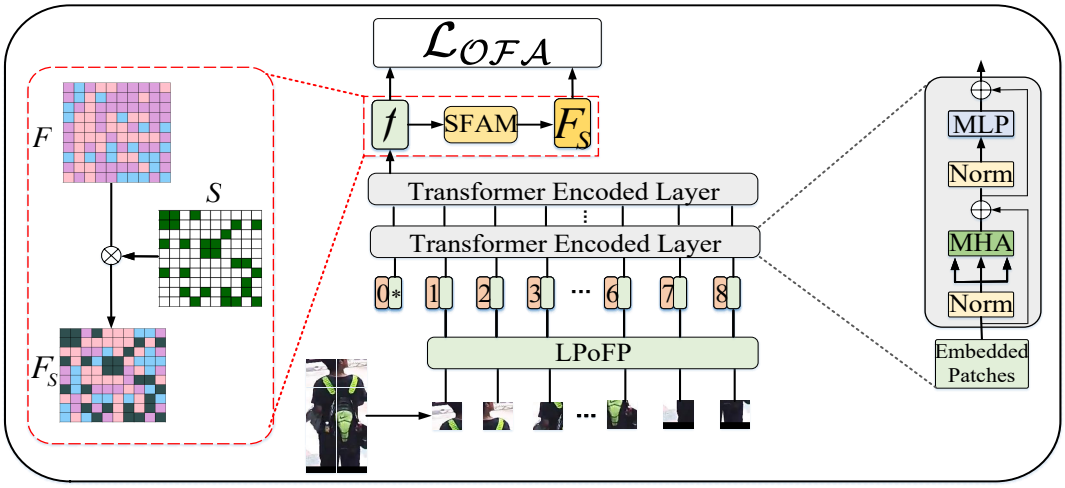

Figure 2: Architecture of the proposed SOFAFormer. $\boldsymbol{f}$ represents the feature output of ViT. $\boldsymbol{F}$ is defined as converting $\boldsymbol{f}$ from 2D Sequence to 4D tensor. $\boldsymbol{S}$ represents the suppression tensor that is used to select which features should be suppressed and which should not. $\boldsymbol{F}_{\mathrm{S}}$ is the output of SFAM, which is obtained by dot product $\boldsymbol{F}$ and $\boldsymbol{S}$. MHA and MLP represent multi-head self-attention and multi-layer perceptron, respectively. LPoFP represents linear projection of flattened patches.

our SOFAFormer follows the pure ViT architecture. Firstly, we uniformly split the pedestrian image into Z fixed-sized $(P \times P \times C)$ patches denoted as $\left\{\boldsymbol{X}_p^z \mid z = 1, 2, \cdots, Z\right\}$, where $Z = (H \times W)/P^2$. Then, an additional learnable embedding token $\boldsymbol{X_{cls}}$ is introduced to the input sequences. This embedding token collects from all patches to serves as the final feature representation output $\boldsymbol{f}$. The input sequence Z_{in} is fed into transformer encoded layer, which can be expressed as:

$$Z_{in} = \left[\boldsymbol{X_{cls}}; \boldsymbol{E}\left(\boldsymbol{X}_p^1\right); \boldsymbol{E}\left(\boldsymbol{X}_p^2\right); \cdots; \boldsymbol{E}\left(\boldsymbol{X}_p^Z\right)\right] + \boldsymbol{E}_{pos}, \tag{1}$$

where $\boldsymbol{E}_{pos} \in \mathbb{R}^{(Z+1)\times C}$ represents position embedding. All patches is mapped to C dimensions by a linear projection $\boldsymbol{E}$.

ViT consists of twelve transformer encoded layers to extract attribute feature representations. As illustrated in Fig. 2, each transformer encoded layer includes a multi-head self-attention module, a multi-layer perceptron module, and two layerNorm layers. On top of ViT we introduce a classification layer to predict the probability each attribute. Finally, we use the binary cross-entropy loss (BCELoss) with sigmoid function as the optimization target, which can be expressed as follows:

$$\begin{aligned}\mathcal{L}_{bce} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\omega_j(y_{i,j}\log(\sigma(\hat{y}_{i,j})) \\ + (1-y_{i,j})\log\left(1-\sigma(\hat{y}_{i,j})\right)),\end{aligned} \tag{2}$$

where

$$\sigma(\hat{y}_{i,j}) = \frac{1}{1+e^{-\hat{y}_{i,j}}}, \tag{3}$$

and $\hat{y}_{i,j}$ is the prediction probability of the classification output, representing the the prediction probability for j-th attribute being present in image x_i. ω_j is a weight proposed in (Li et al. 2018a), which is used to alleviate the distributed imbalance between attributes. ω_j can be expressed as follows:

$$\omega_j = \begin{cases} e^{1-r_j} & y_{i,j} = 1 \\ e^{r_j} & y_{i,j} = 0 \end{cases}, \tag{4}$$

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

The Proposed SFAM Method

PAR indeed involves recognizing a series of attributes from pedestrian images, and due to varying attribute frequencies, the task is prone to the common challenge of imbalanced data. As a result, the recognition performance of minor-annotated attributes may be unsatisfactory. To tackle this issue, our SOFAFormer incorporates a parameter-free SFAM method, which aims to learn more robust feature representations and effectively encode all attribute-specific regions in a balanced way. By doing so, our model strives to improve attribute recognition across all attributes, including those that occur less frequently.

Given a training batch of $\boldsymbol{B}$ images, the final feature output $\boldsymbol{f} \in \mathbb{R}^{B\times(Z+1)\times C}$ (refer to Fig. 2) from ViT is fed into SFAM. In our SFAM, we first transform 2D sequences feature $\boldsymbol{f}$ into a 4D tensor feature map $\boldsymbol{F} \in \mathbb{R}^{B\times C\times H\times W}$ (refer to Fig. 2), where C, H, and W indicate the channel, height, and width of the feature map, respectively. Then, we adopt the implicit assumption proposed by Zagoruyko *et al.* (Zagoruyko and Komodakis 2017) to compute the feature activation mapping. That is, the absolute value of a hidden neuron activation can be used an indication of the importance of that neuron. Therefore, the activation mapping of $\boldsymbol{F}$

can be expressed as:

$$\boldsymbol{F}_{\mathrm{A}} = \sum_{i=1}^{C} |\boldsymbol{F}_i|^p , \tag{5}$$

where $\boldsymbol{F}_i = \boldsymbol{F}(i, :, :)$. That is, $\boldsymbol{F}_{\mathrm{A}}$ is computed by squaring each tensor slice (*i.e.*, $h \times w$) along the channel C of $\boldsymbol{F}$, followed by a summation operation across the channel. Our SFAM follows the definition of (Zagoruyko and Komodakis 2017), with the parameter p is set to 2.

Then, we calculate the average of the activation values r_j on j-th row, which is calculated by averaging the tensor (*i.e.*, $H \times W$) obtained from Eq. 5 along H for each row. It can be expressed as:

$$r_j = \frac{\sum_{w=1}^{W} \boldsymbol{F}_{\mathrm{A}(j,w)}}{W}, \tag{6}$$

When achieving Eq. 6, all r_j are sorted from minimum to maximum. A suppression range threshold τ is set to select the feature values with larger activation. That is, the top ($\tau \times H$) largest ranked feature values are selected for suppression. Next, we suppress these selected values by a factor α, obtaining a suppressed feature $\boldsymbol{F}_{\mathrm{S}}$.

To be specific, we create a suppression tensor $\boldsymbol{S}$ (refer to Fig. 2) with the same size of $\boldsymbol{F}$, $\boldsymbol{S}$ can be expressed as:

$$\boldsymbol{S}_{i,j,w} = \begin{cases} \alpha & r_j > \tau \\ 1 & r_j < \tau \end{cases}, \tag{7}$$

where $i \in (1, C)$, and $w \in (1, W)$. α denotes the suppression ratio. We apply the dot product between $\boldsymbol{S}$ and $\boldsymbol{F}$ to obtain the suppressed feature $\boldsymbol{F}_{\mathrm{S}}$ (refer to Fig. 2). $\boldsymbol{F}_{\mathrm{S}}$ and $\boldsymbol{f}$ share a linear classification layer to predict pedestrian attributes.

Orthogonal Feature Activation Loss

Our SOFAFormer produces two features, one ($\boldsymbol{f}$) is the output of the backbone network and the other ($\boldsymbol{F}_{\mathrm{S}}$) is the output after SFAM. These two features are sent to a linear classification layers for attribute probability prediction, and two probability predictions are added together to obtain the final prediction result.

We expect the two attribute predictions to complement each other, meaning that $\boldsymbol{f}$ and $\boldsymbol{F}_{\mathrm{S}}$ can encode attributes from more diverse feature spaces. Since $\boldsymbol{f}$ and $\boldsymbol{F}_{\mathrm{S}}$ share the final classification layer, the PAR model may encourage the same distribution of two features to quickly minimize the final loss. We believe that two highly correlated features cannot complement each other. Therefore, $\mathcal{L}_{\mathcal{OFA}}$ is proposed to decrease the feature correlation, potentially making feature space more comprehensive. Our $\mathcal{L}_{\mathcal{OFA}}$ encourages the maximization of the difference between two features while also supervising them with BCELoss. In this way, it can promote the mutual enhancement of the two features, further increasing their complementarity, and ultimately improving the overall efficacy and adaptability of our SOFAFormer.

$\mathcal{L}_{\mathcal{OFA}}$ forces two features to be orthogonal to each other and can be expressed as:

$$\mathcal{L}_{\mathcal{OFA}} = \frac{1}{\boldsymbol{B}} \sum_{i=1}^{\boldsymbol{B}} \left[\frac{1}{\boldsymbol{C}} \sum_{j=1}^{\boldsymbol{C}} \left\langle f_{ij} \cdot F_{s_{ij}} \right\rangle \right] \tag{8}$$

where $\boldsymbol{B}$ and $\boldsymbol{C}$ represent training batch size and the dimension of the final feature, respectively.

The final loss of our SOFAFormer is the combination of $\mathcal{L}_{bce}$ and $\mathcal{L}_{\mathcal{OFA}}$, which can be expressed as follows:

$$\mathcal{L} = \mathcal{L}_{bce} + \beta * \mathcal{L}_{\mathcal{OFA}} \tag{9}$$

where β controls contribution of $\mathcal{L}_{\mathcal{OFA}}$.

Experiments

Comprehensive evaluations are conducted to verify the effectiveness of the proposed SOFAFormer. The experiments are conducted on four PAR benchmark datasets, including PETA (Deng et al. 2014), PA100K (Liu et al. 2017), RAPv1 (Li et al. 2016), and RAPv2 (Li et al. 2018b). Details about these datasets and evaluation protocols can be found in the supplementary material. As PA100K is the largest dataset in PAR, we conduct exhaustive ablation studies on this dataset. More ablation studies are shown in the supplementary material for further analysis.

Comparison with State-of-the-Art PAR Methods

In Tab. 1, we show the performance comparison between our SOFAFormer and several recently SOTA methods (Weng et al. 2023; Cao et al. 2023; Wu et al. 2020; Zhao et al. 2019; Li et al. 2019; Wang et al. 2017; Tang et al. 2019) on PETA, PA100K, RAPv1, and RAPv2. It is evident that our method achieves the best performance in terms of mA and accuracy on four datasets, respectively. In mA, our SOFAFormer outperforms the second-best performance by 0.4% (GRL (Zhao et al. 2018)) on PETA, 1.2% (Label2Label (Li et al. 2022)) on PA100K, 1.3% (SSC$_{hard}$ (Jia, Chen, and Huang 2021)) on RAPv1, and 1.9% (IAA-Caps (Wu et al. 2022a)) on RAPV2, respectively. Some methods (*e.g.,* VRKD (Li et al. 2019), JLAC (Tan et al. 2020), and IAA-Caps (Wu et al. 2022a)) introduce additional human semantic parsing or complex attention modules to achieve performance improvement, but this comes at the cost of requiring more computing resources. In contrast, our SOFAFormer, excluding the backbone network, is a parameter-free method, which demonstrates its efficiency in achieving superior performance without the need for additional parameters.

It can be noticed that the proposed method substantially outperforms JLAC on PETA (87.0% *vs.* 87.1%), PA100K (82.3% *vs.* 83.4%), and RAPv2 (79.2% *vs.* 81.9%). JLAC relies on the hypothesis that graph convolutional networks (GCN) can explore all correlations among multiple attributes. However, it relies on CNN to extract feature, and the inherent convolutional and downsampling operations of CNNs limit the ability of the extracted features to model all correlations effectively. Additionally, GCN brings introduces more parameters, which may and potentially increases the risk of overfitting. In contrast, our SOFAFormer adopts

Method	PETA					PA100K					RAPv1					RAPv2				
	mA	Accu	Prec	Rec	F1	mA	Accu	Prec	Rec	F1	mA	Accu	Prec	Rec	F1	mA	Accu	Prec	Rec	F1
JRL	82.1	-	82.6	82.1	82.0	-	-	-	-	-	74.7	-	75.1	75.0	74.6	-	-	-	-	-
PGDM	83.0	78.1	86.9	84.7	85.8	75.0	73.1	84.4	82.2	83.3	74.3	64.6	78.9	75.9	77.4	-	-	-	-	-
GRL	86.7	-	84.3	**88.8**	86.5	-	-	-	-	-	81.2	-	77.7	80.9	79.3	-	-	-	-	-
MsVAA	84.6	78.6	86.8	86.1	86.5	-	-	-	-	-	-	-	-	-	-	78.3	65.6	77.4	79.2	78.3
RA	86.1	-	84.7	88.5	86.6	-	-	-	-	-	81.2	-	79.5	79.2	79.3	-	-	-	-	-
VRKD	84.9	81.0	**88.4**	87.5	**87.9**	77.9	78.5	88.4	86.1	87.2	78.3	69.8	**82.1**	80.4	**81.2**	-	-	-	-	-
VAC	-	-	-	-	-	79.2	79.4	**89.0**	86.3	87.6	-	-	-	-	-	79.2	64.5	75.8	79.4	77.1
ALM	86.3	79.5	85.6	88.1	86.9	80.7	77.1	84.2	88.8	86.5	81.9	68.2	74.7	**86.5**	80.2	79.8	64.8	73.9	82.0	77.8
Da-HA	-	-	-	-	-	79.4	68.9	80.1	81.3	80.7	-	-	-	-	-	-	-	-	-	-
SSC_{hard}	85.9	78.5	86.3	86.2	86.0	81.0	78.4	86.4	87.6	86.6	82.1	68.2	77.9	82.9	79.9	-	-	-	-	-
IAA-Caps	85.3	78.0	86.1	85.8	85.6	81.9	80.3	88.4	88.0	87.8	81.7	68.5	79.6	82.1	80.4	80.0	68.0	**78.8**	81.4	79.7
Label2Label	-	-	-	-	-	82.2	79.2	86.4	88.6	87.1	-	-	-	-	-	-	-	-	-	-
FEMDAR	84.7	78.5	86.8	85.7	85.9	81.0	79.7	88.0	87.5	87.3	79.7	66.9	79.1	79.2	78.8	-	-	-	-	-
$EALC_{w.ACM}$	85.9	80.6	87.5	87.4	87.4	80.5	80.1	87.2	88.6	87.9	82.1	69.3	79.6	82.8	**81.2**	-	-	-	-	-
Baseline	85.3	78.7	86.9	85.9	86.4	81.4	79.1	87.0	87.9	87.4	81.1	68.3	78.6	81.9	80.2	79.6	67.5	78.6	80.6	79.6
SOFAFormer	**87.1**	**81.1**	87.8	88.4	87.8	**83.4**	**81.1**	88.4	**89.0**	**88.3**	**83.4**	**70.0**	80.0	83.0	**81.2**	**81.9**	**68.6**	78.0	**83.1**	**80.2**

Table 1: Performance comparison of SOTA methods on the PETA, PA100K, RAPv1, and RAPv2 datasets. Performance in five metrics, including mean Accuracy (mA), accuracy (Accu), precision (Prec), Recall (Rec), and F1, is evaluated. The first and second highest scores are represented by bold font and underline respectively.

Dataset	Component		mA	Accu	Prec	Rec	F1
	SFAM	$\mathcal{L}_{\mathcal{OFA}}$					
PA100K	-	-	81.4	79.1	87.0	84.8	87.9
	✓	-	82.9	80.8	88.5	88.4	88.1
	✓	✓	83.4	81.1	88.4	89.0	88.3
RAPv2	-	-	79.6	67.5	78.6	80.6	79.6
	✓	-	81.0	68.0	78.1	82.0	79.1
	✓	✓	81.9	68.6	78.0	83.1	80.2

Table 2: Ablation study of each component of our method on the PA100K and RAPv2 datasets.

ViT to capture attribute-specific features and maintain the contextual relation between different regions. By leveraging SFAM, we are able to exploit features from minor-annotated regions without introducing any additional parameters. This design allows the extracted features to cover all attribute-specific regions comprehensively and leads to the superior performance of our method compared to JLAC.

Ablation Study

In the ablation study, we observe notable improvements in performance when incorporating SFAM and $\mathcal{L}_{\mathcal{OFA}}$ into our SOFAFormer. The experiments are given in Tab. 2.

(1) It is observed that with SFAM, the performance increases from 81.4% to 82.9% in term of mA on PA100K. This demonstrates that the combination of SFAM enhances the robustness of the extracted feature. For RAPv2, a clear performance improvement (1.4%) is achieved by directly using SFAM. This is because that our SFAM suppresses major-annotated attribute features to some extent and strengthens the learning of minor-annotated attributes. As result, our SOFAFormer with SFAM is better able to consider each attribute-specific regions and exploits more discriminative feature. (2) Introducing $\mathcal{L}_{\mathcal{OFA}}$ to make the feature space more comprehensive leads to a significant performance improvement. The mA on the PA100K and RAPv2 datasets increased from 82.9% to 83.4%, and 81.0% to 81.9%, respectively. In our SOFAFormer, the output feature $\boldsymbol{f}$ of pure ViT is fed into SFAM and gain a new feature $\boldsymbol{F}_{\mathrm{S}}$. Both $\boldsymbol{f}$ and $\boldsymbol{F}_{\mathrm{S}}$ share the final classifier layer for predicting attribute probabilities. However, sharing the final classifier layer may cause the model to force the two feature distributions to be similar, limiting their complementarity. Therefore, $\mathcal{L}_{\mathcal{OFA}}$ is introduced to encourage lower feature correlation between $\boldsymbol{f}$ and $\boldsymbol{F}_{\mathrm{S}}$. By doing so, $\mathcal{L}_{\mathcal{OFA}}$ pushes attribute-specific regions in different feature spaces, thereby achieving better complementarity between them. This ultimately leads to improved overall performance in attribute prediction.

Visualizations Analysis In Fig. 3, we leverage Grad-CAM (Selvaraju et al. 2017) to generate heat maps, providing an intuitively analysis of the advantages of our method. In Fig. 3 (b), it is evident that the pure ViT (without SFAM and $\mathcal{L}_{\mathcal{OFA}}$) can capture most of the attribute-specific regions but tends to overlook certain attributes. This observation aligns consistent with our previous analysis, where the imbalanced attribute distribution prevents the pure ViT from adequately covering all attributes. However, with the inclusion of SFAM, our SOFAFormer is able to extract features from more attribute-specific regions. For instance, in the sixth image, the pure ViT predominantly focuses on the major-annotated attribute (*e.g.,* clothing style) and neglects the minor-annotated attribute (*e.g.,* backpack). In contrast, Fig. 3 (b) and Fig. 3 (c) demonstrate that SFAM contributes to picking up the minor-annotated attribute (*i.e.,* backpack) that was previously overlooked by the pure ViT. This demonstrates that SFAM effectively enhances the model's ability to capture features from minor-annotated attribute-specific regions, thereby addressing the issue of attribute imbalance and improving the overall attribute recognition performance.

$\mathcal{L}_{\mathcal{OFA}}$ is proposed to enhance the attributes' representa-

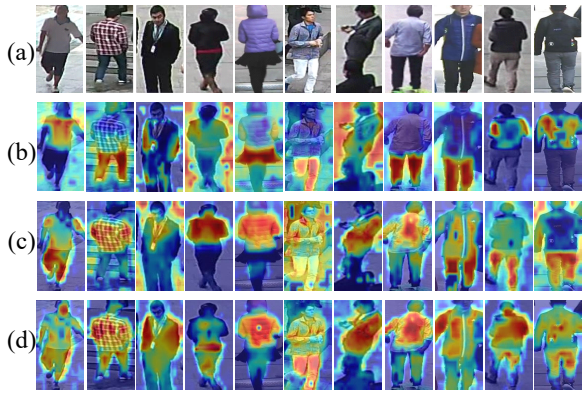

Figure 3: Visualization analysis. (a) original images from PA100K, (b) w/o. SFAM and $\mathcal{L}_{\mathcal{OFA}}$, (c) w. SFAM and w/. $\mathcal{L}_{\mathcal{OFA}}$, (d) w. SFAM and $\mathcal{L}_{\mathcal{OFA}}$.

Dataset	Method	mA	Accu	Prec	Rec	F1
PETA$_{zs}$	MsVAA	71.0	59.4	74.8	70.1	72.4
	VAC	71.1	58.9	**75.0**	70.5	72.1
	ALM	70.7	58.6	73.0	71.3	71.7
	IAA-Caps	72.5	60.1	74.1	73.1	73.1
	SOFAFormer	**74.7**	**62.1**	**75.0**	**75.1**	**74.6**
RAPv2$_{zs}$	MsVAA	71.3	63.6	77.2	76.6	76.4
	VAC	70.2	65.5	**79.9**	76.7	77.1
	ALM	72.0	64.5	77.3	77.7	77.1
	IAA-Caps	72.0	64.6	78.1	77.1	77.2
	SOFAFormer	**73.9**	**66.3**	78.2	**79.4**	**78.4**

Table 3: Performance comparison of four methods on PETA$_{zs}$ and RAPv2$_{zs}$ datasets.

tion by exploring different feature spaces. As shown in Fig. 3 (c), $\mathcal{L}_{\mathcal{OFA}}$ effectively promotes the robustness and comprehensiveness of attribute representations, allowing each attribute region can be considered more effectively. With $\mathcal{L}_{\mathcal{OFA}}$, our SOFAFormer is capable of focusing on the entire pedestrian region and extracting more robust features. The heat maps in Fig. 3 illustrate that our SOFAFormer demonstrates stronger representation ability in discovering more attentive features, leading to more accurate attribute recognition.

Generalization Analysis on PETA$_{zs}$ and RAPv2$_{zs}$ The newly proposed datasets, PETA$_{zs}$ and RAPv2$_{zs}$, have non-overlapping pedestrian IDs between the training and testing sets, which align more closely with real-world scenarios and provide a better measure of the model's generalization ability. In Tab. 3, it is observed that our SOFAFormer achieves the best performance in terms of mA, accuracy, recall, and F1-score across both newly proposed datasets. For instance, our method outperforms the best method IAA-Caps by 2.2% and 1.9% in terms of mA on the two datasets, respectively. The performance improvements on these datasets demonstrate that the proposed SOFAFormer can be better applied to real-world scenarios in the PAR task.

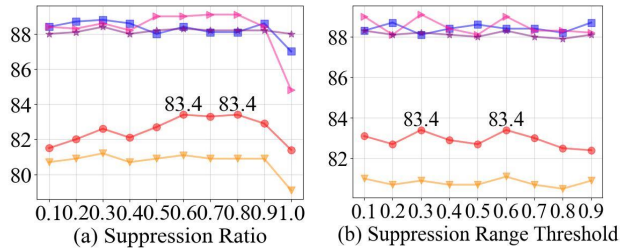

Figure 4: Sensitivity to α and τ on PA100K. Red, orange, blue, pink, and purple line represent mA, Accu, Prec, Rec, and F1, respectively. The maximum value is printed.

Parameter Analysis

In our SFAM, we generate a suppression tensor $\boldsymbol{S}$ with the same size as $\boldsymbol{F}$. We use $\boldsymbol{S}$ to suppress the selected largest activation regions by dot product, which scales the selected feature values using the suppression ratio α. The variation of α can affect the performance of the PAR model. As shown in Fig. 4 (a), the performance of our SOFAFormer improves to varying degrees with different values of α. When α is set to 0.0, it means that the largest activation attribute-specific regions are directly erased. During the training process, with the same parameter setting, the training loss becomes NaN. This is likely due to the erased feature activations, which leads to the inability to correspond with the attribute labels and affects the convergence of the PAR model. The best performance is achieved when α is set to 0.6 or 0.8. We leverage the mFive score proposed in (Yang et al. 2021) to choose more appropriate parameter setting. we set α to 0.6 on the PA100K dataset to yield the best performance

We investigate the effect of varying τ on the performance, and the result is shown in Fig. 4 (b). It can be seen that the performance decreases when $\tau \in (0.3, 0.6)$ or $\tau \in (0.6, 1.0)$. The best performance in term of mA is achieved with τ = 0.3 or 0.6. We also use the mFive score to select the best parameters (*i.e.,* τ = 0.6).

Conclusion

This paper introduces SOFAFormer, a novel and effective approach for the PAR task. Unlike existing PAR methods, SOFAFormer leverages ViT to capture long-range contextual relations between different attribute-specific regions. We propose a parameter-free SFAM module that identifies and focuses on the lower informative minor-annotated attribute regions, complementing the features extracted by ViT. Additionally, we introduce the $\mathcal{L}_{\mathcal{OFA}}$ loss to enhance the robustness and diversity of feature representations from different feature spaces. Experimental results demonstrate the superior performance of SOFAFormer on multiple benchmark PAR datasets. The proposed method consistently outperforms existing SOTA methods, showing its effectiveness in capturing comprehensive attribute representations. The combination of ViT, SFAM, and $\mathcal{L}_{\mathcal{OFA}}$ contributes to substantial performance improvements, making SOFAFormer a promising approach for real-world PAR applications.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62306311, in part by the Fellowship of China Post-Doctoral Science Foundation under Grant 2022T150698, in part by the International Post-Doctoral Exchange Fellowship Program (Talent-Introduction Program) of China under Grant YJ20210324, in part by the Special Research Assistant Program of the Chinese Academy of Sciences under Grant E2S9180301, in part by the Natural Science Foundation of Fujian Province under Grant 2023J01067, and in part by the Public Security Artificial Intelligence Infrastructure Support Platform.

References

Cao, Y.; Fang, Y.; Zhang, Y.; Hou, X.; Zhang, K.; and Huang, W. 2023. A novel self-boosting dual-branch model for pedestrian attribute recognition. *Signal Processing: Image Communication*, 115: 116961.

Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8126–8135.

Deng, Y.; Luo, P.; Loy, C. C.; and Tang, X. 2014. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM International Conference on Multimedia*, 789–792.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*.

Feris, R.; Bobbitt, R.; Brown, L.; and Pankanti, S. 2014. Attribute-based people search: Lessons learnt from a practical surveillance system. In *Proceedings of International Conference on Multimedia Retrieval*, 153–160.

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.

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

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. TransReID: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15013–15022.

Huang, Y.; Wu, Q.; Xu, J.; and Zhong, Y. 2019a. Celebrities-ReID: A Benchmark for Clothes Variation in Long-Term Person Re-Identification. In *International Joint Conference on Neural Networks*, 1–8.

Huang, Y.; Wu, Q.; Xu, J.; and Zhong, Y. 2019b. SBSGAN: Suppression of Inter-Domain Background Shift for Person Re-Identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 9527–9536.

Huang, Y.; Wu, Q.; Xu, J.; Zhong, Y.; and Zhang, Z. 2021a. Clothing Status Awareness for Long-Term Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11895–11904.

Huang, Y.; Wu, Q.; Xu, J.; Zhong, Y.; and Zhang, Z. 2021b. Unsupervised domain adaptation with background shift mitigating for person re-identification. *International Journal of Computer Vision*, 129(7): 2244–2263.

Huang, Y.; Xu, J.; Wu, Q.; Zheng, Z.; Zhang, Z.; and Zhang, J. 2018. Multi-pseudo regularized label for generated data in person re-identification. *IEEE Transactions on Image Processing*, 28(3): 1391–1403.

Huang, Y.; Xu, J.; Wu, Q.; Zhong, Y.; Zhang, P.; and Zhang, Z. 2019c. Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10): 3459–3471.

Jia, J.; Chen, X.; and Huang, k. 2021. Spatial and Semantic Consistency Regularizations for Pedestrian Attribute Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 962–971.

Lanchantin, J.; Wang, T.; Ordonez, V.; and Qi, Y. 2021. General Multi-label Image Classification with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16478–16488.

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

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 preprint arXiv:1603.07054*.

Li, Q.; Zhao, X.; He, R.; and Huang, K. 2019. Pedestrian Attribute Recognition by Joint Visual-semantic Reasoning and Knowledge Distillation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 833–839.

Li, W.; Cao, Z.; Feng, J.; and Lu, J. 2022. Label2Label: A Language Modeling Framework for Multi-Attribute Learning. In *European Conference on Computer Vision*, 562–579.

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 Recognition*, 95: 151–161.

Liu, P.; Liu, X.; Yan, J.; and Shao, J. 2018. Localization guided learning for pedestrian attribute recognition. In *Proceedings of the British Machine Vision Conference*.

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.

Ma, Z.; Zhao, Y.; and Li, J. 2021. Pose-guided Inter- and Intra-part Relational Transformer for Occluded Person Re-Identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, 8126–8135.

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.

Sarfraz, M. S.; Schumann, A.; Wang, Y.; and Stiefelhagen, R. 2017. Deep view-sensitive pedestrian attribute inference in an end-to-end model. In *Proceedings of the British Machine Vision Conference*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.

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 Conference on Artificial Intelligence*, volume 34, 12055–12062.

Tan, Z.; Yang, Y.; Wan, J.; Hang, H.; Guo, G.; and Li, S. Z. 2019. Attention-based pedestrian attribute analysis. *IEEE transactions on Image Processing*, 28(12): 6126–6140.

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/CVF International Conference on Computer Vision*, 4997–5006.

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.

Weng, D.; Tan, Z.; Fang, L.; and Guo, G. 2023. Exploring attribute localization and correlation for pedestrian attribute recognition. *Neurocomputing*, 531: 140–150.

Wu, J.; Huang, Y.; Gao, M.; Gao, Z.; Zhao, J.; Shi, J.; and Zhang, A. 2023a. Exponential Information Bottleneck Theory Against Intra-Attribute Variations for Pedestrian Attribute Recognition. *IEEE Transactions on Information Forensics and Security*, 5623–5635.

Wu, J.; Huang, Y.; Gao, M.; Gao, Z.; Zhao, J.; Zhang, H.; and Zhang, A. 2023b. A Two-Stream Hybrid Convolution-Transformer Network Architecture for Clothing-Change Person Re-Identification. *IEEE Transactions on Multimedia*, 1–15.

Wu, J.; Huang, Y.; Gao, Z.; Hong, Y.; Zhao, J.; and Du, X. 2022a. Inter-Attribute awareness for pedestrian attribute recognition. *Pattern Recognition*, 131: 108865.

Wu, J.; Huang, Y.; Wu, Q.; Gao, Z.; Zhao, J.; and Huang, L. 2022b. Dual-Stream Guided-Learning via a Priori Optimization for Person Re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17(4): 1–22.

Wu, M.; Huang, D.; Guo, Y.; and Wang, Y. 2020. Distraction-aware feature learning for human attribute recognition via coarse-to-fine attention mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12394–12401.

Xiang, L.; Jin, X.; Ding, G.; Han, J.; and Li, L. 2019. Incremental few-shot learning for pedestrian attribute recognition. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 3912–3918.

Yang, L.; Zhu, L.; Wei, Y.; Liang, S.; and Tan, P. 2016. Attribute recognition from adaptive parts. In *Proceedings of the British Machine Vision Conference*, 81.1–81.11.

Yang, Y.; Tan, Z.; Tiwari, P.; Pandey, H. M.; Wan, J.; Lei, Z.; Guo, G.; and Li, S. Z. 2021. Cascaded Split-and-Aggregate Learning with Feature Recombination for Pedestrian Attribute Recognition. *International Journal of Computer Vision*, 1–14.

Zagoruyko, S.; and Komodakis, N. 2017. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proceedings of the International Conference on Learning Representations*.

Zhang, J.; Ren, P.; and Li, J. 2020. Deep Template Matching for Pedestrian Attribute Recognition with the Auxiliary Supervision of Attribute-wise Keypoints. *arXiv preprint arXiv:2011.06798*.

Zhang, N.; Paluri, M.; Ranzato, M.; Darrell, T.; and Bourdev, L. 2014. Panda: Pose aligned networks for deep attribute modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1637–1644.

Zhang, Y.; and Wang, H. 2023. Diverse Embedding Expansion Network and Low-Light Cross-Modality Benchmark for Visible-Infrared Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2153–2162.

Zhang, Y.; Yan, Y.; Lu, Y.; and Wang, H. 2021. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, 788–796.

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 International Joint Conference on Artificial Intelligence*, 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 Conference on Artificial Intelligence*, volume 33, 9275–9282.

Zitnick, C. L.; and Dollár, P. 2014. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, 391–405.