Diverse Features Discovery Transformer for Pedestrian Attribute Recognition

Aihua Zheng[a,b], Huimin Wang[b,c,d], Jiaxiang Wang[b], Huaibo Huang[c,d,*], Ran He[c,d], Amir Hussain[e]

[a]*Information Materials and Intelligent Sensing Laboratory of Anhui Province, School of Artificial Intelligence, Anhui University, HeFei, 230601, China*
[b]*Institute of Artificial Intelligence, Hefei Comprehensive National Science Center., HeFei, 230088, China*
[c]*Center for Research on Intelligent Perception and Computing (CRIPAC), BeiJing, 100089, China*
[d]*Institute of Automation, Chinese Academy of Sciences., BeiJing, 100089, China*
[e]*School of Computing, Merchiston Campus, Edinburgh Napier University, Edinburgh, EH10 5DT, U.K*

Abstract

Recently, Swin Transformer has been widely explored as a general backbone for computer vision, which helps to improve the performance of vision tasks due to the ability to establish associations for long-range dependencies of different spatial locations. By implementing the pedestrian attribute recognition with Swin Transformer, we observe that Swin Transformer tends to focus on a relatively small number of local regions within which attributes may be correlated with other attributes, which leads Swin Transformer to predict attributes in those neglected regions based on such correlation. In fact, discriminative information may exist within these neglected regions, which is crucial for attribute identification. To address this problem, we propose a novel diverse features discovery transformer (DFDT) which can find more attribute relationship regions for robust pedestrian attribute recognition. First, Swin Transformer is used as a feature extraction network to acquire attribute features with the long-distance association, which predicts the corresponding attribute information. Second, we propose a diverse features suppression module (DFSM) to obtain semantic features directly associated with attributes by suppressing the peak locations of the most discriminative features and randomly selected feature regions to spread the feature regions that Swin Transformer is interested in. Third, we plug the diverse features suppression module into different stages of Swin Transformer to learn detailed texture features to help recognition. In addition, we have divided the attribute features into multiple vertical feature regions to improve the focus on local attribute features. Experiments on three benchmark datasets validate the effectiveness of the proposed algorithm.

Keywords: pedestrian attribute recognition, vision transformer, features suppression

*Corresponding author

Email addresses: ahzheng214@foxmail.com (Aihua Zheng), hemion_whm@foxmail.com (Huimin Wang), Netizenwjx@foxmail.com (Jiaxiang Wang), huaibo.huang@cripac.ia.ac.cn (Huaibo Huang), rhe@nlpr.ia.ac.cn (Ran He), hussain.doctor@gmail.com (Amir Hussain)

[1]A. Zheng is with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, School of Artificial Intelligence, Anhui University, Hefei, 230601, China, and also with the Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, 230088, China (e-mail: ahzheng214@foxmail.com).

[2]H. Wang is with the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China, also with the Center for Research on Intelligent Perception and Computing (CRIPAC), BeiJing, 100089, China, and also with the Institute of Automation, Chinese Academy of Sciences, BeiJing, 100089, China (e-mail: hemion_whm@foxmail.com).

[3]J. Wang is with the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China (e-mail: Netizenwjx@foxmail.com).

[4]R. He and H. Huang are with the Center for Research on Intelligent Perception and Computing (CRIPAC), BeiJing, 100089, China, and also with the Institute of Automation, Chinese Academy of Sciences, BeiJing, 100089, China (e-mail: rhe@nlpr.ia.ac.cn; huaibo.huang@cripac.ia.ac.cn).

[5]A. Hussain is with the School of Computing, Merchiston Campus, Edinburgh Napier University, Edinburgh, EH10 5DT, U.K (e-mail: hussain.doctor@gmail.com).

1. Introduction

Pedestrian attribute recognition aims to learn predefined attributes from any given pedestrian image, which are a specific predefined set of attributes, such as age, gender, long/short hair, long/short sleeves, etc. Pedestrian attribute recognition is a fundamental task in computer vision, which has a wide range of applications in many practical applications such as pedestrian re-identification (Layne et al., 2012; Wang et al., 2018), face verification (Kumar et al., 2009; Vo et al., 2021), and pedestrian retrieval (Feris et al., 2014; Siddiquie et al., 2011). Recently, attributes

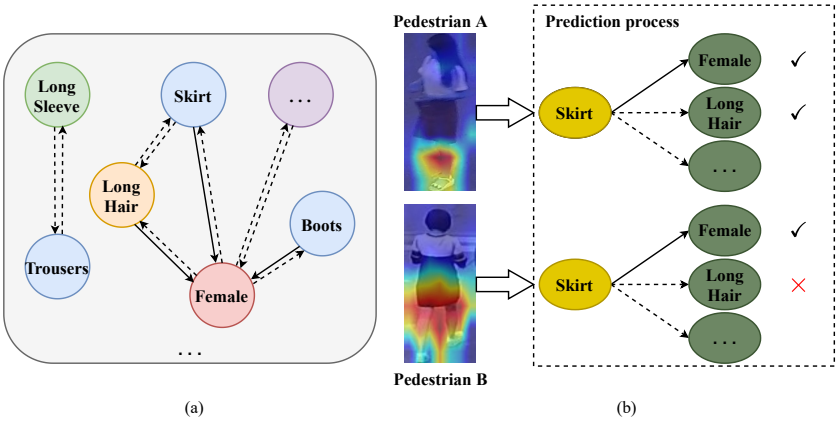

Figure 1: The problem of prediction by inter-attribute correlation. (a) represents that there is some correlation between pedestrian attributes, in which the dotted line indicates a weak correlation and the solid line indicates a strong correlation. For example, when a pedestrian has long hair, there is a high probability that the pedestrian is a female, so there is a strong correlation between long hair and female. When a pedestrian wears long sleeves, there is a certain probability that the pedestrian is wearing trousers, so there is a weak correlation between long sleeves and trousers. (b) mainly expresses the process of Swin Transformer utilizes correlation between attributes for prediction, where the dotted and solid lines represent the same meaning as (b), the yellow attributes indicate the attributes that Swin Transformer focus on, and the green attributes indicate the attributes that Swin Transformer predicts based on the correlation.

recognition has attracted many studies by researchers and has made great developments. Due to the presence of some objective factors, such as varying illumination, local occlusion, low resolution, background clutter, etc., pedestrian attribute recognition is still a very challenging task.

In the past few years, Convolutional Neural Networks (CNNs) (Long et al., 2015; Tran et al., 2015; Zhang and Suganthan, 2016) have been the dominant architectures of pedestrian attribute recognition. Li et al. (2015) propose two networks where one identifies each attribute independently and the other learns all attributes jointly. However, these two networks extract features from an entire image without considering the fine-grained information in the image. To overcome the above-mentioned problem, Zhu et al. (2015), Li et al. (2018a), and Liu et al. (2018b) utilize auxiliary techniques such as part segmentation (Li et al., 2021), pose estimation, and region proposals, respectively, to capture local area features which are combined with global features to jointly predict pedestrian attributes. Although these methods can further improve recognition performance, the computation of localized parts is complex. In addition, they ignore the intrinsic connections between pedestrian attributes. Wang et al. (2017) combine CNN and Long Short-Term Memory (LSTM) to establish the dependence of labels. Zhao et al. (2019) further employ Convolutional Long Short-Term Memory (ConvLSTM) network to establish the spatial correlation of attributes. However, these networks fail to perform parallel computation which is less efficient in practical applications. Some works devote to integrating Graph Convolutional Network (GCN) into the pedestrian attribute recognition task. (Li et al., 2019a) exploit the human parsing model to locate body regions and adopted GCN to obtain corresponding group features. Tan et al. (2020) propose an end-to-end unified framework, which employs GCN to capture both the attribute and contextual relations for pedestrian attribute recognition. However, GCN-based methods need to pre-define the graph structure, which is difficult to apply flexibly in practical situations.

CNN-based models (Peng et al., 2021) have advantages in local feature extraction by collecting local features in a hierarchical manner for better image representation, but most existing CNN-based pedestrian attribute recognition methods have the following two limitations. First, the CNN-based models have still difficulty in modeling global

content-dependent interactions among different image regions, since the coverage of the receptive field of the CNN model is limited and narrow. The lack of global relations among pedestrian attribute features may weaken the ability of representation learning. Second, the modeling of attributes inter-relationships is not flexible enough, since the sequence-based algorithms only establish one-way relationships, and GCN-based algorithms require predefined graph structure.

Recently, Transformer (Dosovitskiy et al., 2020; Carion et al., 2020; Zheng et al., 2021; Wei et al., 2022; Guo et al., 2021) has attracted extensive interest in the computer vision domain. In contrast to CNN, vision Transformers use self-attention layers to capture global interactions between contexts, and is able to learn semantic correlations between different spatial locations. Especially, Swin Transformer (Liu et al., 2021) has shown great promise as it integrates the advantages of both CNN and Transformer. It can learn both local feature information and long-range dependencies of different localities in large-size images. Inspired by the advantages of Swin Transformer, this paper devotes itself to exploring Swin Transformer to adaptively model the correlation between pedestrian attributes, and obtain representations with attribute semantic interactions for attribute classification. However, one can not directly employ Swin Transformer for pedestrian attribute recognition. The main reason is, compared to learning richer attribute features, Swin Transformer prefers to utilize correlations between attributes to identify all attributes. When such correlation does not exist in a pedestrian, Swin Transformer results in dramatic prediction errors. As shown in Fig. 1 (b), for pedestrian A, Swin Transformer focuses on the skirt attribute region and can identify pedestrian A as a female with long hair based on their correlations to the skirt. However, for pedestrian B with short hair instead of long hair, the correlation between the skirt and long hair does not exist. In this case, Swin Transformer still tends to use the prior correlation to predict via focusing on the skirt attribute region, which results in incorrectly predicting the short hair as long hair attribute for pedestrian B.

To alleviate the problem that Swin Transformer tends to rely on correlations between attributes when predicting attributes, we propose a diverse features suppression module, which forces Swin Transformer to learn more attribute features for classification rather than relying only on a local region and correlations. Therein, the diverse features suppression module contains two types of feature suppression, namely peak regions suppression and random regions suppression. First, the peak regions are the most discriminative feature regions for classification, and the attributes of this region often have a strong correlation with those of other regions. By suppressing the peak regions, Swin Transformer is forced to learn the features of the attribute being correlated and thus mitigating to some extent the negative impact of misuse of correlation in Swin Transformer. Second, there are other discriminative regions that contain subtle features for attribute prediction that help Swin Transformer distinguish the difference between two similar attributes, such as a stripe coat and a plaid coat. In order to mine more subtle features, we employ random regions suppression to enforce Swin Transformer to randomly learn detailed information in pedestrian images. In addition, to learn more shallow detail attribute features simultaneously, we plug the diverse features structure module on both the shallow and deep features of Swin Transformer.

Our contribution can be summarized as follows:

(1) We propose an end-to-end framework based on Swin Transformer for pedestrian attribute recognition, which can adaptively learn the correlation between attributes without complex modules to model this correlation.

(2) We propose a plug-and-play diverse features suppression module that drives Swin Transformer to learn more attribute features and weakens the reliance of Swin Transformer on the correlation between attributes. To enable Swin Transformer learns detailed and global features separately, we inserted the module into different stages of Swin Transformer.

(3) We divide the attribute features into multiple vertical regions to improve Swin Transformer's focus on local attribute feature regions, which is due to the fact that attributes are distributed from top to bottom on the entire body of the pedestrian.

(4) Experiments show the superiority of the proposed method over recent methods and the effectiveness of our framework for pedestrian attribute recognition.

The rest of this paper is organized as follows. Section 2 provides an overview of the works related to pedestrian attribute recognition. Section 3 describes the backbone for pedestrian attribute recognition. Section 4 systematically elaborates on the proposed DFDT, including peak region suppression mechanism and random region suppression mechanism. Section 5 shows the comprehensive experimental results of DFDT. Finally, Section 6 concludes the paper together with the future directions.

2. Related Works

2.1. Pedeatrian Attribute Recognition

Table 1: Comparison of seventeen state-of-the-art pedestrian attributes recognition methods in five categories.

Category	Method	Advantages	Disadvantages
Global based	ACN (Sudowe et al., 2015) DeepMar (Li et al., 2015) MTCNN (Abdulnabi et al., 2015)	Sharing network features, simple and effective.	Lack of consideration of fine-grained features, limited performance.
Part-based	DeepCAMP (Diba et al., 2016) PGDM (Li et al., 2018a) LGNet (Liu et al., 2018b) AR-BiFPN (Moghaddam et al., 2021)	fine-grained information, both global and local features.	Dependence on the accuracy of local localization.
Attention-based	HPNet (Liu et al., 2017) VeSPA (Sarfraz et al., 2017) JLPLS_PAA (Tan et al., 2019)	Multiple scales features on , Multi-channel features.	Limited performance, additional design of new attention model.
Sequence-based	RNN-CNN (Wang et al., 2016) JRL (Wang et al., 2017)	Transforming attribute categories into sequential models, exploring the constraints between attributes.	Difficult to establish bi-directional relations of attributes, cannot perform parallel computation.
GCN-based	DCSA (Wang et al., 2017) A-AOG (Park et al., 2017) VSGR (Li et al., 2019b) JLAC (Tan et al., 2020) MTSA (Ji et al., 2020)	Modeling Semantic Relationships between attributes, modeling spatial relationships between image regions.	Complex, difficult to apply in practical scenarios.

Pedestrian attribute recognition is a popular field of study in computer vision and has been widely employed in a variety of vision tasks, such as person retrieval (Siddiquie et al., 2011) and person re-identification (Layne et al., 2012; Hadjkacem et al., 2020; Ruiz et al., 2020). We research seventeen state-of-the-art pedestrian attribute recognition methods in terms of five categories and analyze the advantages and disadvantages of each category of methods as shown in Table 1.

Early methods (Sudowe et al., 2015; Li et al., 2015; Abdulnabi et al., 2015) take the whole image as input and try to learn global attribute representation. However, those methods neglect a focus on fine-grained information. Later, some methods are successively proposed to alleviate the insufficiency of fine-grained information extraction. DeepCAMP (Diba et al., 2016) chunks the images to learn the attribute features of each block of images. Li et al. (2018a) first utilize a human pose to guide the network to locate key points in pedestrian images, then extract local region features depending on these key points. Liu et al. (2018b) propose a Localization Guided Network to extract attribute-related local features. Moghaddam et al. (2021) combine human semantic parsing and pedestrian attribute recognition to mine semantic and spatial information.

Some other methods (Liu et al., 2017; Sarfraz et al., 2017; Guo et al., 2017; Sarafianos et al., 2018; Tan et al., 2019) utilize attention mechanisms to improve the performance of attribute recognition. Liu et al. (2017) propose a multi-directional attention model that consists of a CNN and an attention feature network. Sarfraz et al. (2017) incorporate a view predictor in attribute recognition networks to estimate the weights of views. DIAA (Guo et al., 2017) framework aggregates multi-scale visual attention and weighted focal loss for deep imbalanced classification as a way to improve recognition performance. Tan et al. (2019) take a multi-task-like way to simultaneously learn various attentional mechanisms, i. e., parsing attention, labeled attention, and spatial attention, to explore relevant and complementary information.

Despite the great improvement in recognition performance, the aforementioned methods fail to model potential relations between attributes. Other works (Wang et al., 2016, 2017; Zhao et al., 2018; Liu et al., 2018a; Zhao et al., 2019) analyze pedestrian attribute identification tasks starting from relations between attributes. Wang et al. (2016) first employ Recurrent Neural Network (RNN) to model the dependency between labels. In order to better mine relevant information, JRL (Wang et al., 2017) employ RNN based recurrent sequential prediction model to capture

high-order dependencies of attributes. On the other hand, some methods introduce GCN to mine relationships in multiple attributes. DCSA (Chen et al., 2012) model utilizes a conditional random field to model the correlation between human attributes. A-AOG (Park et al., 2017) is proposed explicitly to represent the decomposition and articulation of body parts, and account for the correlations between poses and attributes. Li et al. (2019b) consider the existence of complex relationships between attributes and different regions and propose a graph reasoning network to jointly model the spatial and semantic relationships of region-region, attribute-attribute, and region-attribute. JLAC (Tan et al., 2020) framework consists of two graph modules, called the attribute-relationship module and contextual relationship module, which are used to discover and capture attribute and contextual relationships, respectively. Ji et al. (2020) propose a new multiple time steps attention mechanism to boost the modeling of the relations between images and attributes.

2.2. Vision Transformer

With the development of deep learning in recent years, more and more creative works have been proposed in which the Transformer (Vaswani et al., 2017) is one of the typical representatives. The Transformer (Vaswani et al., 2017) is a deep neural network mainly based on the self-attention mechanism, which is initially applied in the field of Natural Language Processing (NLP). Inspired by the powerful representation ability of Transformer, researchers propose to extend Transformer to computer vision tasks. Vision Transformer (ViT) (Dosovitskiy et al., 2020) utilizes a pure Transformer for image block sequences directly as well as achieves superior performance on multiple image recognition benchmarks. Afterward, Data-efficient image Transformer (DeiT) (Touvron et al., 2021) distill ViT (Dosovitskiy et al., 2020) by using a teacher-student network to obtain a more lightweight model. Although these Transformers can capture remote dependencies between patches, they ignore local feature extraction. Tokens-to-Token ViT (T2T) (Yuan et al., 2021) proposes a progressive tokenization module that can gradually model local structural information in the process of reducing the length of tokens by aggregating adjacent tokens into one token. Han et al. (2021) propose a Transformer in Transformer (TNT) architecture that further divides patchs into multiple sub-patchs based on ViT, which extracts local features from pixel embeddings through internal transformer blocks. Swin Transformer (Liu et al., 2021) introduce the hierarchical structure which is frequently used in CNN to build hierarchical Transformer based on ViT. In addition, it adopts shifted windows approach to computing attention which greatly reduces computational complexity. By aggregating the advantages of CNN and Transformer, Swin Transformer has more potential to become a general-purpose backbone for computer vision compared to other Transformer architectures. More surprisingly, Swin Transformer continues to dominate in several downstream sub-tasks, which once again proves that Transformer structure is more suitable than CNN to solve computer vision problems.

3. Backbone

To learn attributes representations with relationships between attributes adaptively, we use Swin Transformer (Liu et al., 2021) as the attribute feature extractor. Swin Transformer employs a hierarchical transformer to extract hierarchical feature maps and uses the shifted windows approach to calculate the relationship between patches in the whole feature map. The shifted windowing scheme brings linear computational complexity by limiting self-attention computation to non-overlapping local windows while also allowing the cross-window connection.

In this work, Swin Transformer consists of four stages, each of which has multiple Swin Transformer blocks. Firstly, the pedestrian image I is divided into a set of non-overlapping image patches. And then these image blocks are fed into the linear embedding layer and the Swin Transformer block to extract features. As shown in the Fig. 2, a Swin Transformer block consists of a window based multi-head self-attention (W-MSA) module or a shifted window based multi-head self-attention (SW-MSA) module and a 2-layer multi-layer perceptron (MLP) with Gaussian error linear unit (GELU). A LayerNorm (LN) layer is inserted before each multi-headed self-attention (MSA) module and each MLP module. In addition, a residual connection is applied after each module. Successive Swin Transformer blocks are computed as follows:

$$\hat{z}^{l} = \text{W-MSA}(\text{LN}(\hat{z}^{l-1})) + z^{l-1}, \tag{1}$$

$$z^{l} = \text{MLP}(\text{LN}(\hat{z}^{l})) + \hat{z}^{l}, \tag{2}$$

$$\hat{z}^{l+1} = \text{SW-MSA}(\text{LN}(z^{l})) + z^{l}, \tag{3}$$

$$z^{l+1} = \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1}, \tag{4}$$

where $\hat{z}^l$ and z^l denote the output features of the (S)W-MSA module and the MLP module for block l, respectively; W-MSA and SW-MSA denote window based multi-head self-attention using regular and shifted window partitioning configurations, respectively.

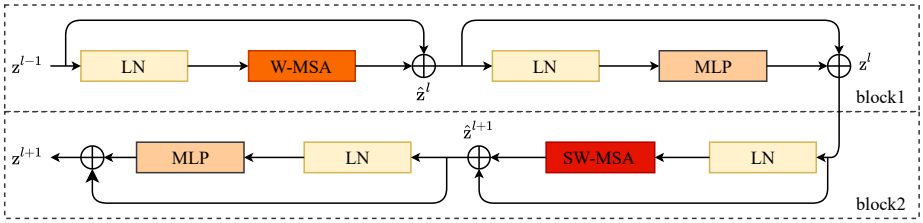

Figure 2: Two successive Swin Transformer blocks structure.

4. Approaches

In this paper, we propose an attribute recognition framework for mining feature diversity called *Diverse Feature Discovery Transformer* (DFDT).

As shown in Fig. 3, our framework consists of a backbone in Section 3 and one main module, namely Diverse Features Suppression Module (DFSM). DFSM contains two types of suppression, namely peak regions suppression (PRS) and random regions suppression (RRS).

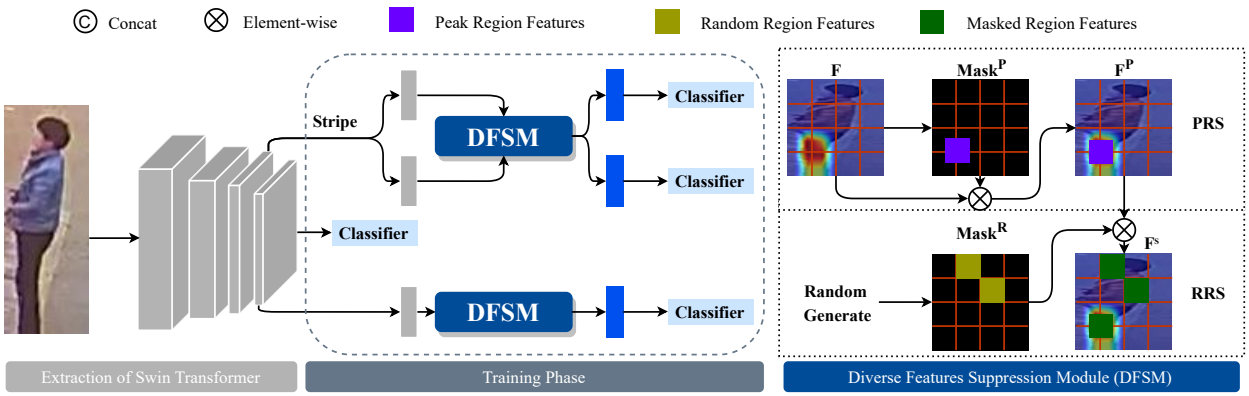

Figure 3: Overview of our overall architecture. The input instance is fed into the gray backbone which has four stages. Then we perform feature suppression on the features in different stages to mine diversity features. The structure of DFSM is shown in the blue part on the right, which contains two suppression mechanisms, namely peak regions suppression (PRS) and random regions suppression (RRS).

4.1. Diverse Features Suppression Module

To alleviate the limitation of Swin Transformer which only focuses on a local region to predict attributes, we propose a diverse features suppression module (DFSM) inspired by Sun et al. (2020), which drives Swin Transformer to pay more attention to the other informative regions and obtain more diverse expressions. First, in order to reduce computational complexity, we compress the channel dimension of the features sent into the DFSM module, and the obtained features are represented by $\mathbf{F} = \{\mathbf{F}_c : c \in [1, C]\}$, where $\mathbf{F}_c \in \mathbb{R}^{H \times W}$. Then, the DFSM can generate binary masks $\mathbf{M} = \{\mathbf{M}_c : c \in [1, C]\}$, where $\mathbf{M}_c \in \mathbb{R}^{H \times W}$ based on $\mathbf{F}$ locating the feature regions to be suppressed. Each element in mask $\mathbf{M}_c$ is in the domain {0,1}, where 1 means the corresponding position is suppressed and 0 indicates that no suppression has been performed. As can be seen from Fig. 3, the mask localization is determined by both peak regions and random regions.

4.1.1. Peak Regions Suppression

To enable Swin Transformer to learn as many features as possible instead of using the correlation between attributes as the dominant factor for recognition, we propose a peak regions suppression mechanism, which enables Swin Transformer to find alternative feature regions that are related to recognition. The peak region features represent the regions that Swin Transformer pays the most attention to, which typically has correlations between attributes in other regions. By suppressing the peak regions, Swin Transformer learns the features of attributes predicted by the correlation, thereby mitigating the effect of this correlation on attribute recognition to some degree. Let $\mathbf{M}_c^P$ be the location of the peak maps from the feature map $\mathbf{F}_c$ and denoted as:

$$\mathbf{M}_c^P(i,j) = \begin{cases} 1, & if \ \ \mathbf{F}_c(i,j) = max(\mathbf{F}_c), \\ 0, & otherwise. \end{cases} \tag{5}$$

where $max(\mathbf{F}_c)$ denotes the maximum of activation maps matrix $\mathbf{F}_c$. The i and j correspond to the rows and columns of the index matrix. We perform peak regions suppression on the features of each channel dimension and let $\mathbf{M}^P$ denote the final obtained peak regions suppression mask which is denoted as follows:

$$\mathbf{M}^P = \{\mathbf{M}_c^P : c \in [1, C]\}, \quad where \ \ \mathbf{M}_c^P \in \mathbb{R}^{H \times W}. \tag{6}$$

4.1.2. Random Regions Suppression

In addition to suppressing the most discriminative regions that the network focuses on, some other local features should also be appropriately suppressed. These regions often contain subtle features that are crucial for attribute prediction, which can help Swin Transformer distinguish the difference between two similar attributes. To mine more subtle features, we propose a random regions suppression approach, which forces Swin Transformer to randomly learn detailed information in pedestrian images. Next, we describe how to randomly select the local area to be suppressed on the feature map $\mathbf{F}_c$. First, we define the size of the region to be suppressed in the feature map $\mathbf{F}_c$ based on the random masking rate r. Based on our experimental setup, we set $r = 1/4$ which means that 1/4 of the region in $\mathbf{F}_c$ is to be suppressed. Then we randomly select 1/4 of the pixels on the all-zero mask with the same size as $\mathbf{F}_c$ and set their value to 1 as the random mask $\mathbf{M}_c^R$. We perform random regions suppression on the features of each channel dimension and let $\mathbf{M}^R$ denote the total random region suppression mask which is denoted as follows:

$$\mathbf{M}^R = \{\mathbf{M}_c^R : c \in [1, C]\}, \quad where \ \ \mathbf{M}_c^R \in \mathbb{R}^{H \times W}. \tag{7}$$

4.1.3. Joint Suppression

During training, we jointly employ peak regions suppression and random regions suppression to mine various features. The final mask position $\mathbf{M}$ corresponding to the feature $\mathbf{F}$ is derived as:

$$\mathbf{M} = \mathbf{M}^P + \mathbf{M}^R. \tag{8}$$

By suppressing the feature $\mathbf{F}$ corresponding to the mask region $\mathbf{M}$, we obtain the joint suppressed features $\mathbf{F}^s$, which is calculated as follows:

$$\mathbf{F}^s = \mathbf{F} - \alpha(\mathbf{M} \odot \mathbf{F}), \tag{9}$$

where α indicates the suppressing factor and $\odot$ refers to the element-wise product. In practice, we set α as a higher number. In our experimental setup, we set α to 0.9 for the best performance.

4.2. Training Phase

As the network deepens, the receptive field of neurons gradually increases, which allows more global information to be contained in the features. However, only the feature suppression operation on the last layer of features will ignore the local details mining. In order to utilize more diverse features, we insert DFSM into the shallow and deep stages of Swin Transformer to enforce Swin Transformer to learn global and local features respectively.

First, in order to learn more detailed information, we perform a suppression operation on $\boldsymbol{f}_3 \in \mathbb{R}^{N \times H \times W}$, which is the feature of the third stage of Swin Transformer. Due to the rich structural prior knowledge of pedestrians, pedestrian

attributes can usually be divided into different parts based on the human structure to learn the features of each part separately. Accordingly, we chunk the features to learn the attribute information specific to each part, and the number of parts is denoted as n. Generally, there is a strong correlation between upper body region attributes and lower body region attributes of a pedestrian, such as long sleeves and long pants and shorts and short sleeves. To disentangle this entanglement between upper and lower local region attributes and to learn more subtle features, we set $n = 2$ which is to divide the shallow feature $\boldsymbol{f}_3$ into two parts $\boldsymbol{p}_1 \in \mathbb{R}^{N\times\frac{H}{2}\times W}$, $\boldsymbol{p}_2 \in \mathbb{R}^{N\times\frac{H}{2}\times W}$ in H dimension. Then, we perform feature suppression in each part to learn the detailed information. We denote the calculation function of the Diverse Features Suppression Module by $\mathbf{DFSM}(\cdot)$ and the two suppressed part features are denoted as $\boldsymbol{p}_1^s$ and $\boldsymbol{p}_2^s$ respectively, which are calculated as follows:

$$\boldsymbol{p}_1^s = \mathbf{DFSM}(\boldsymbol{p}_1), \tag{10}$$

$$\boldsymbol{p}_2^s = \mathbf{DFSM}(\boldsymbol{p}_2). \tag{11}$$

Then, we use two classifiers ϕ_1 and ϕ_2 to classify $\boldsymbol{p}_1^s$ and $\boldsymbol{p}_2^s$ separately, and the two prediction vectors are denoted as:

$$\hat{y}_1 = \phi_1(\boldsymbol{p}_1^s), \tag{12}$$

$$\hat{y}_2 = \phi_2(\boldsymbol{p}_2^s). \tag{13}$$

Second, in order to obtain more global information, We apply the Diverse Features Suppression Module to the output feature $\boldsymbol{f}_4$ of the final stage of Swin Transformer. In addition, to ensure that the global information of the deep features is not broken, we do the suppression directly on the whole features instead of chunking them and then suppressing them separately. We denote the suppressed features as $\boldsymbol{f}_4^s$ with the following formula:

$$\boldsymbol{f}_4^s = \mathbf{DFSM}(\boldsymbol{f}_4). \tag{14}$$

Then we feed $\boldsymbol{f}_4^s$ into the classifier ϕ_3 and the prediction vector is denoted by $\hat{y}_3$. The formula is as follows:

$$\hat{y}_3 = \phi_3(\boldsymbol{f}_4^s). \tag{15}$$

Finally, to maintain the ability of Swin Transformer that model associations between attributes, We directly classify the feature maps $\boldsymbol{f}_4$ with the classifier ϕ_4, and the corresponding prediction vector $\hat{y}_4$ is denoted by:

$$\hat{y}_4 = \phi_4(\boldsymbol{f}_4). \tag{16}$$

4.3. Loss Function

The entire framework is end-to-end trained with the binary cross-entropy loss function, which is defined as follows:

$$\mathcal{L}_k = -\frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{N} y^{ij}\log(p_k^{ij}) + (1-y^{ij})\log(1-p_k^{ij}), \tag{17}$$

where $p_k = \delta(\hat{y}_k)$ is the prediction probability of prediction vector $\hat{y}_k$, $\delta(x) = 1/(1+e^{-x})$ is the sigmoid function, $p_k^{ij} \in [0, 1]$ indicates the probability that the i-th attribute appears in the j-th image, and y^{ij} represents the ground truth for the i-th attribute that appears in the j-th image.

Finally, the total training loss is calculated by summing over the four individual loss:

$$\mathcal{L}_{loss} = \sum_{k=1}^{4}\mathcal{L}_k. \tag{18}$$

5. Experiments

5.1. Datasets

To evaluate the effectiveness of our framework, we conduct experiments on three general datasets as PETA (Deng et al., 2014), PA100k (Liu et al., 2017), and RAPv1 (Li et al., 2016) and a newer dataset as RAPv2 (Li et al., 2018b).
The PETA (Deng et al., 2014) dataset contains 8,705 pedestrians with 19,000 pedestrian images, which is divided into 9,500 images as the training set, 1,900 as the validation set, and 7,600 as the test set. Each pedestrian is labeled with 65 attributes, including binary and multi-valued attributes. For the evaluation, we follow the common experimental protocol in (Deng et al., 2014) of using only the 35 attributes with a positive rate greater than 5%.
The PA100k (Liu et al., 2017) dataset contains 100,000 pedestrian images from 598 real outdoor surveillance cameras, which is the largest pedestrian attribute recognition dataset. It is randomly divided into a training set, a validation set and a test set in the ratio of 8:1:1. Each image is labeled with 26 binary attributes.
The RAPv1 (Li et al., 2016) attribute dataset contains 41,585 pedestrian images extracted from 26 indoor surveillance cameras, which is divided into 33,268 images for the training set and 8,317 for the test set. Each image is labeled with 69 binary attributes and 3 multi-class attributes. According to the protocol in (Li et al., 2016), 51 binary attributes are used to evaluate the recognition performance.
The RAPv2 (Li et al., 2018b) attribute dataset consists of 84,298 images extracted from 25 cameras, in which 50,957 images are used for training, 16,986 images for verifying, and 16,985 images for testing. In consistence with RAPv1, it has 72 attribute labels.

Table 2: Quantitative results. Comparison results with state-of-the-art methods on the PETA dataset, the PA100k dataset, and the RAPv1 dataset. The first and second results are highlighted in bold fonts and underlined, respectively.

Methods	Venue	PETA					PA100k					RAPv1				
		mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1
DeepMar	IAPR2015	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
PGDM	ICME2018	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
LGNet	BMVC2018	-	-	-	-	-	76.96	75.55	86.99	83.17	85.04	78.68	68.00	80.36	79.82	80.09
ALM	ICCV2019	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
JRL	ICCV2017	85.67	-	86.03	85.34	85.42	-	-	-	-	-	77.81	-	78.11	78.98	78.58
MTSA	PRL2020	84.62	78.80	85.67	86.42	86.04	-	-	-	-	-	77.62	67.17	79.72	78.44	79.07
VRKD	IJCAI2019	84.90	80.95	**88.37**	87.47	87.91	77.87	78.49	88.42	86.08	87.24	78.30	69.79	**82.13**	80.35	81.23
JLAC	AAAI2020	86.96	80.38	87.81	87.09	87.45	82.31	79.47	87.45	87.77	87.61	**83.69**	69.15	79.31	82.40	80.82
HPNet	ICCV2017	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
JLPLS-PAA	TIP2019	84.88	79.46	87.42	86.33	86.87	81.61	78.89	86.83	87.73	87.27	81.25	67.91	78.56	81.45	79.98
CoCNN	IJCAI2019	86.97	79.95	87.58	87.73	87.65	80.56	78.30	**89.49**	84.36	86.85	81.42	68.37	81.04	80.27	80.65
SSC	ICCV2021	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
DFDT	Ours	**87.44**±0.12	**81.17**±0.09	87.44±0.09	**88.96**±0.10	**88.19**±0.06	**83.63**±0.16	**81.24**±0.10	88.02±0.09	**89.48**±0.10	**88.74**±0.11	82.34±0.8	**70.89**±0.11	80.36±0.08	84.32±0.07	**82.15**±0.09

Table 3: Quantitative results. Comparison results with state-of-the-art methods on the RAPv2 dataset. The first and second results are highlighted in bold fonts and underlined, respectively.

Methods	Venue	RAPv2				
		mA	Accu	Prec	Recall	F1
ALM	ICCV2019	79.79	64.79	73.93	82.03	77.77
JLAC	AAAI2020	79.23	64.42	75.69	79.18	77.40
DFDT	Ours	**79.96**±0.11	**69.30**±0.10	**79.38**±0.08	**82.62**±0.10	**80.97**±0.09

5.2. Evaluation Metrics

To evaluate the performance of pedestrian attribute recognition, two types of metrics are adopted. (1) Class-based metric: The mean Accuracy (mA) is commonly used as a class-based metric (Deng et al., 2014). We calculate the average of the classification accuracy of positive samples and negative samples for each attribute label as the metric for each attribute. Then we take the average of all attributes as the mean accuracy. (2) Instance-based metric: The instance-based metrics (Li et al., 2016) include accuracy, precision, recall rate, and F1-score. For accuracy, precision and recall, we first compute the scores of predicted attributes against the ground truth for each test image and then

average the scores for overall test images. The F1-score is computed based on precision and recall, therefore F1-score is a comprehensive metric of precision and recall. Compared to mA, which assumes independence between attributes, instance-based metrics take into account the inter-attribute correlation.

5.3. Implementation Details

The baseline model of Swin Transformer (Liu et al., 2021) is used as the backbone which uses pre-training on ImageNet (Krizhevsky et al., 2012) as the initialization. The input shape of images is reshaped to 224×224 with the data augmentations of randomly flip and crop. The network is optimized by stochastic gradient descent algorithm with a batch size of 16, a momentum of 0.9, and a weight decay of 0.0005. The initial learning rate is set to 0.0001. The network is trained for 80 epochs. All experiments were implemented in Pytorch with one NVIDIA RTX 3090.

5.4. Quantitative Results

We compare the performance of the proposed method with 12 state-of-the-art methods on three general datasets such as the PETA dataset, the PA100k dataset, and the RAPv1 dataset, as shown in Table 2. In addition, we compare the performance of the proposed method with the current state-of-the-art methods on a newer dataset called RAPv2, as shown in Table 3. These methods can be divided into four categories: (1) The conventional methods based on whole image or part, such as DeepMar (Li et al., 2015), PGDM (Li et al., 2018a), LGNet (Liu et al., 2018b), and ALM (Tang et al., 2019). (2) The methods based on the Sequence model, such as JRL (Wang et al., 2017) and MTSA(Ji et al., 2020). (3) The methods based on GCN, such as VRKD(Li et al., 2019a) and JLAC (Tan et al., 2020). (4) The methods based on attention and prior knowledge, like HPNet (Liu et al., 2017), JLPLS-PAA (Tan et al., 2019), CoCNN (Han et al., 2019) and SSC (Jia et al., 2021).

The results in Table 2 show that our method achieves the best results for the instance-based metrics on three general datasets, which demonstrates that our method can better model the correlation between attributes than GCN-based and sequence-model-based algorithms. For the class-based metric, our method achieves the best performance on the PETA dataset, the PA100k dataset, and comparable performance on the RAPv1 dataset. Compared to the PETA dataset and the PA100k dataset, there are 51 attributes to be evaluated on the RAPv1 dataset, which means the network needs to learn more fine-grained information to predict these 51 attributes. JLAC and SSC achieve the best and second-best performance in terms of mA on the RAPv1 dataset. JLAC focuses on the significant feature regions which are related to the attributes by modeling the contextual relation of the image with GCN. SSC extracts more accurate attribute features by designing a spatial consistency regularization to locate the spatial location of attributes exactly. Although both methods achieve excellent results on mA, these methods can not fully utilize the association between attributes resulting in the performance of instance-based metrics is not very high. In a comprehensive comparison, our method achieves relatively superior performance. The comparison of two classical and latest methods on RAPv2 dataset is shown in Table 3. Consistently, our method achieves the best performance on all metrics.

Specifically, on the PETA dataset, our method improves 0.47%, 0.22%, and 0.28% than the second-best method on mA, Accu, and F1 respectively; which improves 1.32%, 1.77%, and 1.13% on the PA100k dataset and 0.17%, 4.51%, and 3.2% on RAPv2 dataset. Although the mA measure of our method is slightly overshadowed on the RAPv1 dataset, our method still improves by 1.1% and 0.92% on Accu and F1 respectively. Our method substantially outperforms the state-of-the-art methods on the large-scale PA100k and RAPv2 datasets compared to the PETA dataset and the RAPv1 dataset, indicating that the proposed method is more adequately trained on larger datasets. In addition, we have evaluated the sensitiveness of the proposed model by 5 random trials as updated in Table 2 and Table 3. We can observe that the fluctuation of each metric is around 0.10, which demonstrates that the proposed model is robust.

5.5. Qualitative Results

To highlight the performance of the attribute-specific result of the proposed method, we compare the mean accuracy of 35 attributes in the PETA dataset between our method and baseline, as shown in Fig. 4. The bars are sorted in descending order according to the mean accuracy between the two methods at one attribute. It is evident that our method achieves different degrees of improvement on most attributes of the PETA dataset compared to the baseline. For some attributes which either require detailed textures to assist in recognition (”UpperBodyThinStripe” and ”UpperBodyLogo”) or only cover small parts of the images (“V-neck” and “Sandals”), the improvement is particularly prominent. This evidences the effectiveness of the diverse features suppression module in attribute recognition.

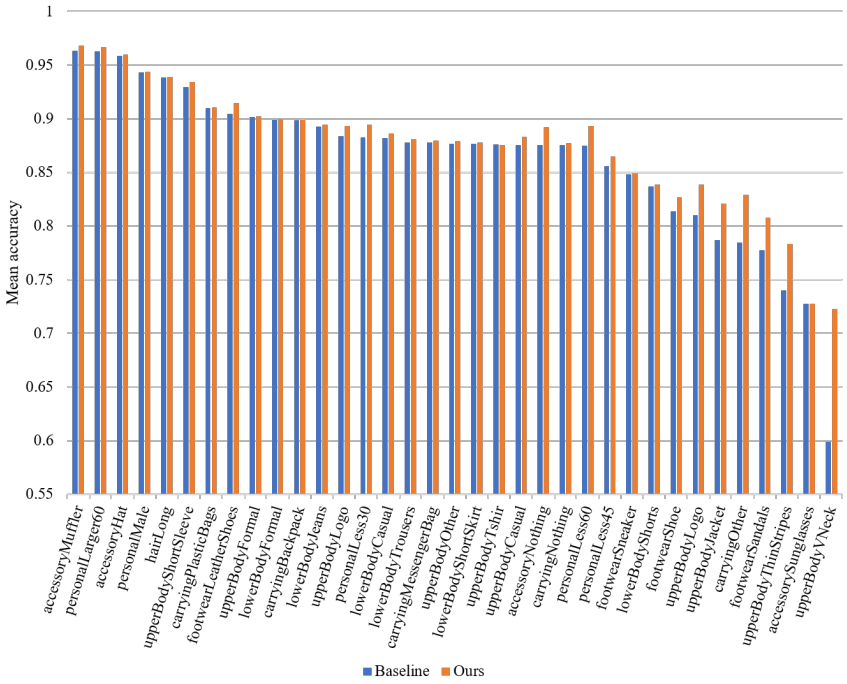

Figure 4: Qualitative results. The mean accuracy comparison results between baseline and our method of all attributes on PETA dataset.

5.6. Ablation Study

Table 4: Ablation study results. The validation of each component of the model was performed on PETA dataset, PA100k dataset, and RAPv1 dataset. The first and second results are highlighted in bold fonts and underlined, respectively.

Methods	Params	FLOPs	PETA					PA100K					RAPv1				
			mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1
Baseline	86.78M	15.12G	86.42	80.31	87.12	88.38	87.74	82.27	80.73	88.01	89.19	88.60	80.20	70.61	**80.51**	83.65	82.05
+ PRS	87.20M	15.13G	86.80	80.87	87.32	88.62	87.96	82.84	81.00	88.08	89.29	88.68	81.87	70.81	80.17	84.17	82.12
+ RRS	87.15M	15.13G	87.24	80.78	87.28	88.50	87.89	83.20	81.16	**88.16**	89.26	88.71	81.65	70.71	80.37	83.74	82.02
+ PRS + RRS (Ours)	87.59M	15.14G	**87.44**	**81.17**	**87.44**	**88.96**	**88.19**	**83.63**	**81.24**	88.02	**89.48**	**88.74**	**82.34**	**70.89**	80.36	**84.32**	**82.15**

To verify how each module proposed in the network performance, we perform ablation experiments on PETA dataset, PA100k dataset, and RAPv1 dataset. The experimental results are shown in Table 4. We first introduce a baseline model without using any suppression method, then separately add Peak Regions Suppression (i.e., PRS) and Random Regions Suppression (i.e., RRS) to the baseline model. Finally, the two suppression methods are jointly integrated (i.e., Ours) in the baseline.

PRS and RRS boost the performance of attribute recognition on all three datasets compared to baseline, especially in the metrics of mA and acc, which shows the effectiveness of these two suppression mechanisms. Integrating both PRS and RRS further improves the recognition performance on all three datasets, which shows that the complementary and correlated features are learned. In addition, The proposed model has 87.59M parameters and 15.14G computational complexity, which only introduces 0.81M more parameters and 0.02G more computational complexity compared with the baseline. This demonstrates that the proposed model achieves superior performance with only a slight increase in the number of parameters and the computational complexity. We also observe that the recognition performance of baseline is impressive, reaching comparable results with previous state-of-the-art methods. This is because there are different degrees of correlations between pedestrian attributes, as shown in the Fig. 5, and Swin Transformer has an excellent spatial semantic interaction modeling capability for the pedestrian attribute recognition

task. Note that the proposed method in this paper generally achieves low precision and high recall, with an occasional increase in one metrical leading to a decrease in another. However, as a comprehensive measure between the precision and recall, F1 is significantly improved by introducing two components.

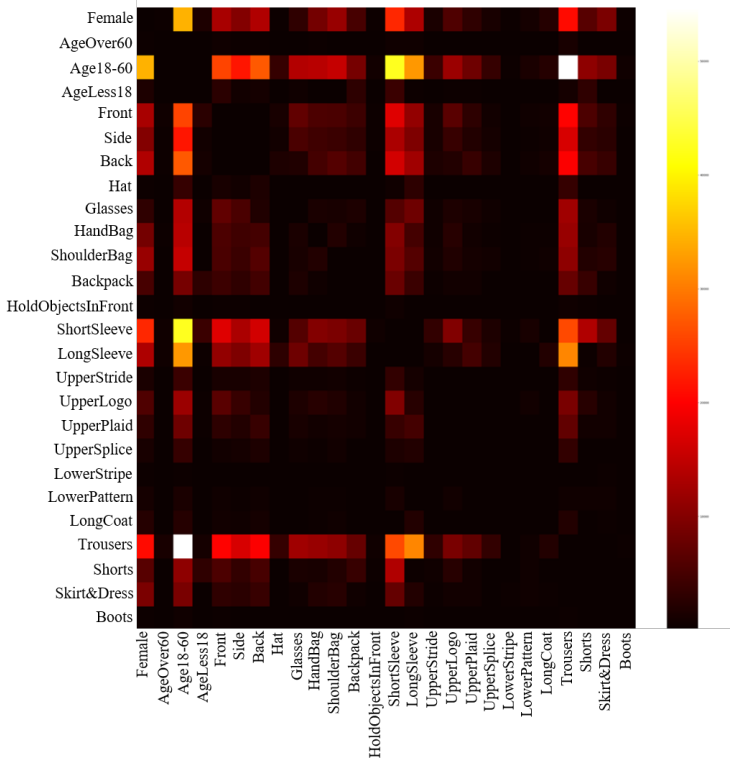

Figure 5: The visualization of the correlation between attributes on PA100k dataset. The warm and dark colors means the strong and the weak relations respectively.

Table 5: Hyperparameter evaluation results. Experiments on the suppressing factor α, the number of part n, and the random masking rate r. The first and second results are highlighted in bold fonts and underlined, respectively.

	α					n				r			
	1	0.9	0.8	0.7	0	1	2	3	4	1/20	1/10	1/4	1/2
mA	87.32	**87.44**	87.28	87.09	86.42	87.04	87.44	87.21	**87.50**	87.08	87.25	**87.44**	87.14
Accu	81.00	**81.17**	81.05	80.97	80.31	80.99	81.17	80.45	**81.27**	81.14	80.85	**81.17**	81.15
Prec	87.33	87.44	**87.72**	87.50	87.12	87.45	87.44	86.82	**87.70**	87.62	87.17	87.44	**87.74**
Recall	88.79	**88.96**	88.44	88.49	88.38	88.60	**88.96**	88.46	88.76	88.65	88.68	**88.96**	88.55
F1	88.06	**88.19**	88.08	87.99	87.74	88.02	88.19	87.63	**88.23**	88.13	87.91	**88.19**	88.14

5.7. Hyperparameter Evaluation

There are mainly three key hyperparameters in our method, suppressing factor α, the number of parts n, and random masking rate r. We set $\alpha = 0.9$, $n = 2$, and $r = 0.25$ for the best performance. To demonstrate the effect of hyperparameters, we adopt the control variable method to obtain the optimal value of one parameter by adjusting the value of this parameter while fixing the other parameters. We conduct the evaluation on the PETA dataset as reported in Table 5.

Suppressing factor α in Eq. (9) indicates the degree of feature suppression. With the increase of the α, there is a significant performance improvement in mA from 86.42 to 87.44. This indicates that setting α to a larger value has a greater improvement than without the diverse feature suppression module ($\alpha = 0$). Especially, $\alpha = 0.9$ for the best performance.

The number of parts n denotes the number of shallow feature chunks in the training phase. To be able to learn the detailed features that assist in attribute classification, we chunk the features in the H dimension, and then each chunk feature is trained separately. Compared to no chunking in the shallow features ($n = 1$), chunking the features can boost the performance. The best performance for attribute recognition is achieved when the features are chunked into 4 parts ($n = 4$). Considering the computational complexity, we set $n = 2$ to the default value, which achieves comparable performance with $n = 4$.

Random masking rate r indicates the percentage of suppressed regions in the feature. The experimental results show that the performance gained by larger or smaller r is not very desirable. A larger r means that a larger number of feature regions are suppressed, and a relatively limited number of attribute features are left for the network to learn, which makes it difficult to further improve the attribute recognition performance. Smaller r means that only a small portion of feature regions are suppressed, and these suppressed regions may not be relevant to attribute recognition, which may not drive the network to learn other attribute features. When $r = 1/4$, the best performance of attribute recognition is achieved.

5.8. Visualization

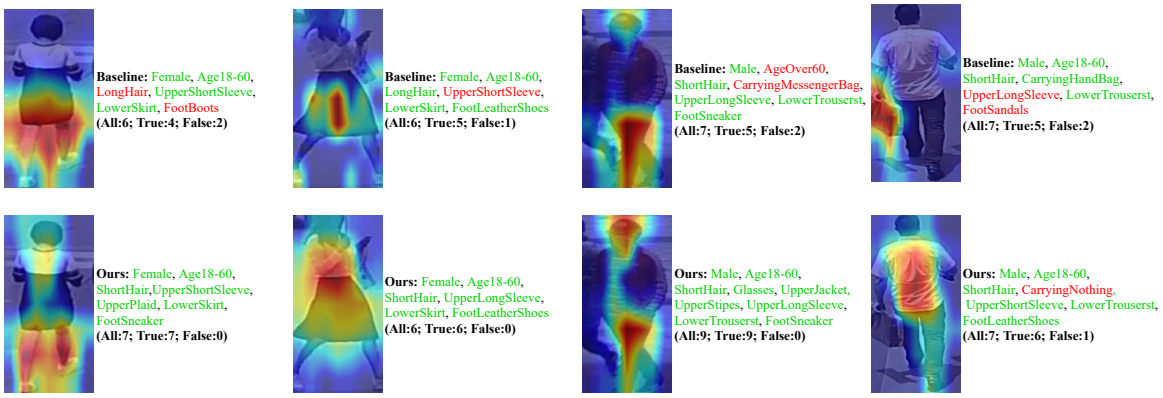

Figure 6: Visualization. We use Grad-Cam (Selvaraju et al., 2017) to show the regions that network pays attention to. The parts with warm colors represent where the network pays attention to. The first and the second line show the results on the input images with baseline and our framework respectively. We also show the prediction results for these images using both the baseline and our method. Where the green attributes indicate attributes that were predicted correctly and the red attributes indicate attributes that were predicted incorrectly.

We demonstrate the features extracted by baseline and our method for visualization as shown in Fig. 6. Obviously, compared to baseline, our method focuses on a larger number of attribute regions, which indicates that our proposed diverse feature suppression module is able to mine more effective features for attribute recognition. In addition, we display the prediction results of the baseline and our method for these images. The first pedestrian with short hair and a skirt was predicted reliably by baseline to be "Female", "Age18-60", and "UpperShortSleeve", whilst the baseline only focuses on the skirt attribute region. This illustrates that Swin Transformer is capable of adaptively learning the correlation between attributes, and can use this correlation to help predict attributes. But the baseline incorrectly predicts pedestrian "ShortHair" attributes to "LongHair" attributes, which means that predictions by using the correlation after focusing on an attribute region are sometimes unreliable, especially when such correlations do not exist for some pedestrians. In contrast, our method accurately predicts all attributes of the pedestrian by focusing on more attribute regions. This demonstrates that our proposed diverse features suppression module can effectively alleviate the problem that Swin Transformer relies excessively on attribute correlations for prediction. Therefore, our method enables Swin Transformer to learn meaningful attribute features. In addition, for the third pedestrian, the baseline cannot predict the style of the pedestrian's upper clothes, while our method accurately predicts that the

pedestrian's upper clothes are stripped by mining into the detailed features of the upper clothes. This illustrates that our method mines more detailed features to help Swin Transformer distinguish the attributes with high similarity.

5.9. Experiment on Vehicle Attribute Recognition

Table 6: Experimental results on the VeRi776 dataset. The first and second results are highlighted in bold fonts and underlined, respectively.

Methods	mA	Accu	Prec	Recall	F1
Baseline	50.55	20.46	39.10	21.28	27.56
+PRS	<u>52.09</u>	<u>21.78</u>	<u>44.55</u>	<u>22.27</u>	<u>29.70</u>
+RRS	51.67	21.34	44.05	<u>22.27</u>	29.59
+PRS+RRS(Ours)	**52.98**	**22.77**	**45.54**	**22.77**	**30.36**

To verify the generalization of our method, we conducted experiments on the non-pedestrian attribute dataset called VeRi776 (Liu et al., 2016). The VeRi776 (Liu et al., 2016) dataset is a vehicle dataset containing over 50,000 images of 776 vehicles, with the training set containing 37,778 images and the test set including 11,579 images. Each image was captured in a real-world unconstrained surveillance scene and labeled with a different attribute. We predict 19 attributes on the VeRi776 (Liu et al., 2016) dataset, including 10 color attributes and 9 vehicle type attributes. The experimental results are reported in Table 6. In consistent with the human datasets, our method improves the performance a lot compared to baseline and achieves promising performance on vehicle dataset, which further evidences the generality of the proposed method in attribute recognition.

6. Conclusion

To our best knowledge, this is the first work to resolve the Transformer applied to the pedestrian attribute recognition problem by feature suppression. In this paper, we first argue that the challenging factor of Transformer applied to PAR is that Transformer relies excessively on inter-attribute correlations to classify attributes. We have contributed an end-to-end network (DFDT) based on diverse feature mining, followed by two feature suppression mechanisms: peak region suppression and random region suppression. Compared with state-of-the-art pedestrian attribute recognition methods, extensive experiments demonstrate the promising performance of the proposed method. In addition, the proposed DFDT can be extended to other multi-label learning tasks, such as face attribute recognition and multi-object classification. In the future, we will consider combining DFDT with the idea of mining fine-grained attribute information for pedestrian re-identification tasks.

7. Acknowledgements

This research is supported in part by the National Natural Science Foundation of China (61976002), the University Synergy Innovation Program of Anhui Province (GXXT-2022-036) and the Anhui University (210341). Amir Hussain would also like to acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC) - Grants Ref. EP/M026981/1, EP/T021063/1, EP/T024917/1.

References

Abdulnabi, A.H., Wang, G., Lu, J., Jia, K., 2015. Multi-task cnn model for attribute prediction. IEEE Transactions on Multimedia 17, 1949–1959.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer. pp. 213–229.

Chen, H., Gallagher, A., Girod, B., 2012. Describing clothing by semantic attributes, in: European Conference on Computer Vision, Springer. pp. 609–623.

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

Diba, A., Pazandeh, A.M., Pirsiavash, H., Van Gool, L., 2016. Deepcamp: Deep convolutional action & attribute mid-level patterns, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3557–3565.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

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

Guo, H., Fan, X., Wang, S., 2017. Human attribute recognition by refining attention heat map. Pattern Recognition Letters 94, 38–45.

Guo, Y., Zheng, Y., Tan, M., Chen, Q., Li, Z., Chen, J., Zhao, P., Huang, J., 2021. Towards accurate and compact architectures via neural architecture transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Hadjkacem, B., Ayedi, W., Ayed, M.B., Alshaya, S.A., Abid, M., 2020. A novel gait-appearance-based multi-scale video covariance approach for pedestrian (re)-identification. Engineering Applications of Artificial Intelligence 91, 103566.

Han, K., Wang, Y., Shu, H., Liu, C., Xu, C., Xu, C., 2019. Attribute aware pooling for pedestrian attribute recognition. arXiv preprint arXiv:1907.11837 .

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y., 2021. Transformer in transformer. Advances in Neural Information Processing Systems 34.

Ji, Z., Hu, Z., He, E., Han, J., Pang, Y., 2020. Pedestrian attribute recognition based on multiple time steps attention. Pattern Recognition Letters 138, 170–176.

Jia, J., Chen, X., Huang, K., 2021. Spatial and semantic consistency regularizations for pedestrian attribute recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 962–971.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25.

Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K., 2009. Attribute and simile classifiers for face verification, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE. pp. 365–372.

Layne, R., Hospedales, T.M., Gong, S., Mary, Q., 2012. Person re-identification by attributes., in: British Machine Vision Conference, p. 8.

Li, D., Chen, X., Huang, K., 2015. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios, in: 2015 3rd Asian Conference on Pattern Recognition, IEEE. pp. 111–115.

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

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

Li, D., Zhang, Z., Chen, X., Ling, H., Huang, K., 2016. A richly annotated dataset for pedestrian attribute recognition. arXiv preprint arXiv:1603.07054 .

Li, Q., Zhao, X., He, R., Huang, K., 2019a. Pedestrian attribute recognition by joint visual-semantic reasoning and knowledge distillation., in: International Joint Conference on Artificial Intelligence, pp. 833–839.

Li, Q., Zhao, X., He, R., Huang, K., 2019b. Visual-semantic graph reasoning for pedestrian attribute recognition, in: Proceedings of the Association for the Advance of Artificial Intelligence, pp. 8634–8641.

Li, Z., Sun, Y., Zhang, L., Tang, J., 2021. Ctnet: Context-based tandem network for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Liu, H., Wu, J., Jiang, J., Qi, M., Ren, B., 2018a. Sequence-based person attribute recognition with joint ctc-attention model. arXiv preprint arXiv:1811.08115 .

Liu, P., Liu, X., Yan, J., Shao, J., 2018b. Localization guided learning for pedestrian attribute recognition. arXiv preprint arXiv:1808.09102 .

Liu, X., Liu, W., Ma, H., Fu, H., 2016. Large-scale vehicle re-identification in urban surveillance videos, in: 2016 IEEE international conference on multimedia and expo (ICME), IEEE. pp. 1–6.

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

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

Moghaddam, M., Charmi, M., Hassanpoor, H., 2021. Jointly human semantic parsing and attribute recognition with feature pyramid structure in efficientnets. IET Image Processing 15, 2281–2291.

Park, S., Nie, B.X., Zhu, S.C., 2017. Attribute and-or grammar for joint parsing of human pose, parts and attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 1555–1569.

Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., Ye, Q., 2021. Conformer: Local features coupling global representations for visual recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 367–376.

Ruiz, I., Raducanu, B., Mehta, R., Amores, J., 2020. Optimizing speed/accuracy trade-off for person re-identification via knowledge distillation. Engineering Applications of Artificial Intelligence 87, 103309.

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

Sarfraz, M.S., Schumann, A., Wang, Y., Stiefelhagen, R., 2017. Deep view-sensitive pedestrian attribute inference in an end-to-end model. arXiv preprint arXiv:1707.06089 .

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

Siddiquie, B., Feris, R.S., Davis, L.S., 2011. Image ranking and retrieval based on multi-attribute queries, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 801–808.

Sudowe, P., Spitzer, H., Leibe, B., 2015. Person attribute recognition with a jointly-trained holistic cnn model, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 87–95.

Sun, G., Cholakkal, H., Khan, S., Khan, F., Shao, L., 2020. Fine-grained recognition: Accounting for subtle differences between similar classes, in: Proceedings of the Association for the Advance of Artificial Intelligence, pp. 12047–12054.

Tan, Z., Yang, Y., Wan, J., Guo, G., Li, S.Z., 2020. Relation-aware pedestrian attribute recognition with graph convolutional networks, in:

Proceedings of the Association for the Advance of Artificial Intelligence, pp. 12055–12062.
Tan, Z., Yang, Y., Wan, J., Hang, H., Guo, G., Li, S.Z., 2019. Attention-based pedestrian attribute analysis. IEEE Transactions on Image Processing 28, 6126–6140.
Tang, C., Sheng, L., Zhang, Z., 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, pp. 4997–5006.
Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, PMLR. pp. 10347–10357.
Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, pp. 4489–4497.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in Neural Information Processing Systems 30.
Vo, D.M., Nguyen, D.M., Lee, S.W., 2021. Deep softmax collaborative representation for robust degraded face recognition. Engineering Applications of Artificial Intelligence 97, 104052.
Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W., 2016. Cnn-rnn: A unified framework for multi-label image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285–2294.
Wang, J., Zhu, X., Gong, S., Li, W., 2017. Attribute recognition by joint recurrent learning of context and correlation, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 531–540.
Wang, J., Zhu, X., Gong, S., Li, W., 2018. Transferable joint attribute-identity deep learning for unsupervised person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2275–2284.
Wei, Y., Wu, C., Li, G., Shi, H., 2022. Sequential transformer via an outside-in attention for image captioning. Engineering Applications of Artificial Intelligence 108, 104574.
Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S., 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 558–567.
Zhang, L., Suganthan, P.N., 2016. Visual tracking with convolutional random vector functional link network. IEEE transactions on cybernetics 47, 3243–3253.
Zhao, X., Sang, L., Ding, G., Guo, Y., Jin, X., 2018. Grouping attribute recognition for pedestrian with joint recurrent learning., in: International Joint Conference on Artificial Intelligence, p. 27th.
Zhao, X., Sang, L., Ding, G., Han, J., Di, N., Yan, C., 2019. Recurrent attention model for pedestrian attribute recognition, in: Proceedings of the Association for the Advance of Artificial Intelligence, pp. 9275–9282.
Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al., 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890.
Zhu, J., Liao, S., Yi, D., Lei, Z., Li, S.Z., 2015. Multi-label cnn based pedestrian attribute learning for soft biometrics, in: International Conference on Biometrics, IEEE. pp. 535–540.