Machine Intelligence Research
www.mi-research.net

22(4), August 2025, 783-796
DOI: 10.1007/s11633-024-1504-6

AdaGPAR: Generalizable Pedestrian Attribute Recognition via Test-time Adaptation

Da Li[1] Zhang Zhang[1,2] Yifan Zhang[1,2] Zhen Jia[1] Caifeng Shan[3]

[1] New Laboratory of Pattern Recognition (NLPR), Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, China

[2] University of Chinese Academy of Sciences, Beijing 100049, China

[3] School of Intelligence Science and Technology, Nanjing University, Nanjing 210023, China

Abstract: Generalizable pedestrian attribute recognition (PAR) aims to learn a robust PAR model that can be directly adapted to unknown distributions under varying illumination, different viewpoints and occlusions, which is an essential problem for real-world applications, such as video surveillance and fashion search. In practice, when a trained PAR model is deployed to real-world scenarios, the unseen target samples are fed into the model continuously in an online manner. Therefore, this paper proposes an efficient and flexible method, named AdaGPAR, for generalizable PAR (GPAR) via test-time adaptation (TTA), where we adapt the trained model through exploiting the unlabeled target samples online during the test phase. As far as we know, it is the first work that solves the GPAR from the perspective of TTA. In particular, the proposed AdaGPAR memorizes the reliable target sample pairs (features and pseudo-labels) as prototypes gradually in the test phase. Then, it makes predictions with a non-parametric classifier by calculating the similarity between a target instance and the prototypes. However, since PAR is a task of multi-label classification, only using the same holistic feature of one pedestrian image as the prototypes of multiple attributes is not optimal. Therefore, an attribute localization branch is introduced to extract the attribute-specific features, where two kinds of memory banks are further constructed to cache the global and attribute-specific features simultaneously. In summary, the AdaGPAR is training-free in the test phase and predicts multiple pedestrian attributes of the target samples in an online manner. This makes the AdaGPAR time efficient and generalizable for real-world applications. Extensive experiments have been performed on the UPAR benchmark to compare the proposed method with multiple baselines. The superior performance demonstrates the effectiveness of the proposed AdaGPAR that improves the generalizability of a PAR model via TTA.

Keywords: Pedestrian attribute recognition, domain generalization, test-time adaptation, attribute localization, non-parametric classifier.

Citation: D. Li, Z. Zhang, Y. Zhang, Z. Jia, C. Shan. AdaGPAR: Generalizable pedestrian attribute recognition via test-time adaptation. *Machine Intelligence Research*, vol.22, no.4, pp.783–796, 2025. http://doi.org/10.1007/s11633-024-1504-6

1 Introduction

Pedestrian attribute recognition (PAR) aims at parsing a pedestrian image into semantic descriptions with multiple visual attributes, e.g., age, gender and clothing styles. It has attracted extensive attention due to its great application potentials, such as person retrieval[1, 2], person re-identification[3, 4], and fashion search[5, 6]. With the success of deep-learning-based methods, the performance of PAR has been significantly improved.

However, in current studies on PAR, researchers typically rely on the independent and identically distributed (i.i.d.) assumption, meaning that both the training set and test set share the same underlying distribution. Such an assumption is far from the real-world application scenarios, where the trained model often faces a significant domain gap between the training set (source domain) and the deployed environment (target domain). For example, a PAR model trained on a dataset collected in a shopping mall (indoor) may fail to work in a college campus (outdoor) where the data distribution is much different with the change of location. Moreover, it is very challenging and expensive to collect an all-encompassing dataset for training a unified model due to the following reasons:

1) The target domain is complex and dynamic, with variations in climates, illuminations, occlusions, and other factors. Collecting enough training samples for all possible situations within an acceptable time span is challenging.

2) Considering the factors related to privacy, security and safety, it is sometimes impractical to obtain the prior knowledge about the distribution of pedestrian attrib-

Research Article
Manuscript received on November 16, 2023; accepted on March 25, 2024; published online on January 7, 2025
Recommended by Associate Editor Wanli Ouyang
Colored figures are available in the online version at https://link.springer.com/journal/11633

utes in the target domain.

Fig. 1 illustrates the adverse impact of domain gap on the PAR performance. Because of the large discrepancy between Market1501[7] and RAPv2[1], the mean accuracy (mA) and F_1 score drop by 22.7% and 22.5% respectively when the model trained on Market1501 is directly tested on RAPv2; And they drop by 15.5% and 23.6% respectively vice versa. Therefore, it is a challenging and promising direction to improve the capability of the PAR model trained on source domains to recognize samples from unseen domains directly, i.e., generalizable PAR (GPAR).

Recently, Specker et al.[8] propose the first public benchmark termed UPAR, which unifies four popular PAR datasets with 40 binary attributes and defines standard training/test settings with the requirements of GPAR. In addition, a GPAR baseline is also proposed, which aggregates a number of tricks, such as data augmentation, dropout and label smoothing, in the training process to improve the model′s ability of generalization across different domains. Although the proposed baseline enhances the generalizability of the model to some extent, it overlooks the domain information carried by the target samples. Recent studies[9, 10] show that generalizing a model to any unknown distribution is almost impossible without exploiting the target samples during inference. Thus, it is essential to explore test-time adaptation[11] in GPAR tasks.

In this work, we propose a novel method, named AdaGPAR, for generalizable PAR via test-time adaptation (TTA) to exploit the online test samples from target domain during the inference stage without requiring any annotation information. Noted that, the setting of our work is significantly different from the settings of unsupervised domain adaption (UDA)[12] and source-free domain adaptation (SFDA)[13]. The UDA aims at improving the model′s performance in unseen target domain by updating the model based on both the labeled training samples from source domain and unlabeled training samples from target domain. Though the SFDA mitigates the dependence on the training samples from source domain, it still needs to access all the training samples (unlabeled or par-

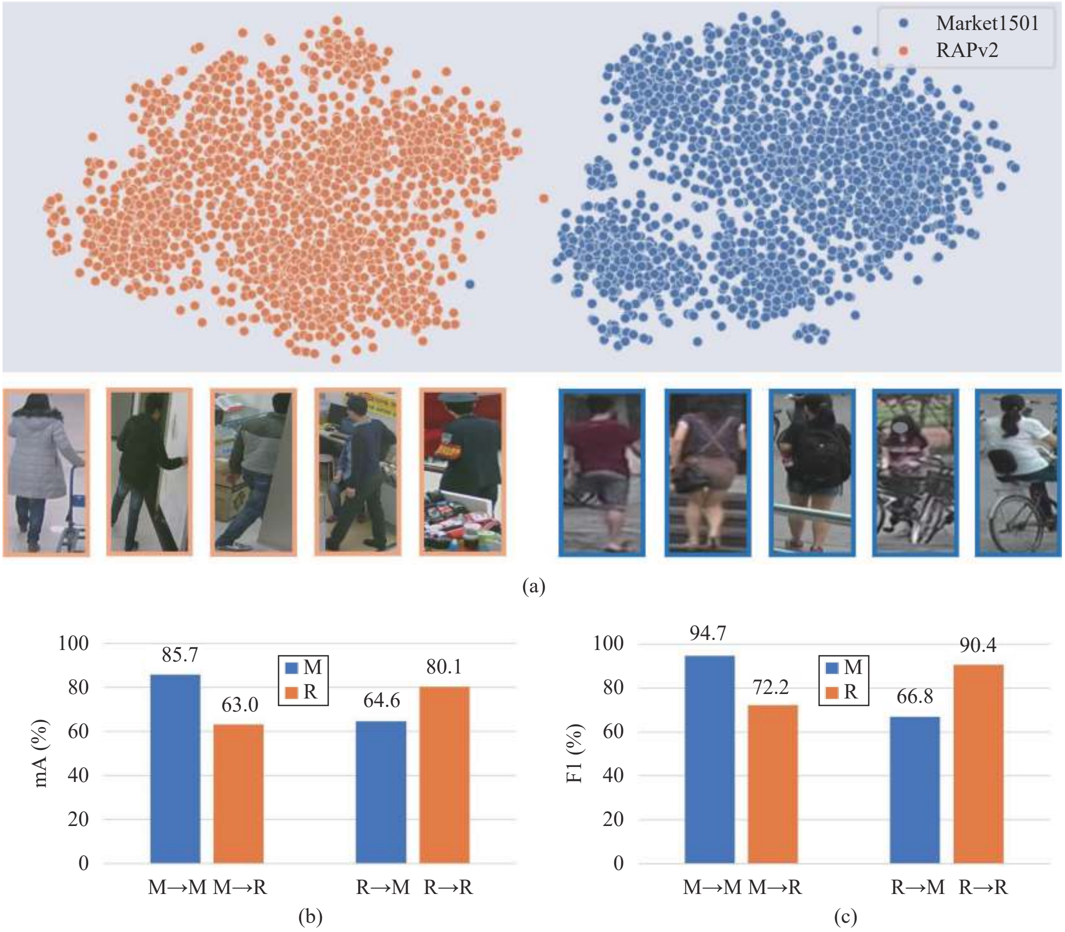


Fig. 1 An illustration for the negative influence of domain gap on the performance of PAR. (a) illustrates the data distributions of Market1501[7] and RAPv2[1] in the feature space. (b) and (c) display the recognition results of mA and F_1, respectively. M is short for Market1501 and R is short for RAPv2. M→M: Train and test on Market1501. M→R: Train on Market1501 and test on RAPv2. R→R: Train and test on RAPv2. R→M: Train on RAPv2 and test on Market1501. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

tial labeled) of target domain to update the trained source model. Such setting is impractical for the real-world applications due to the limitations on privacy and safety. However, from the perspective of TTA, only the model trained in the source domain and the unlabeled test samples from target domain in an online manner are available in the inference phase. It is much closer to the requirements of real-world applications.

Inspired by the work AdaNPC[14], we store the reliable target sample pairs (feature and predicted pseudo-label) to memory as the prototypes, which will be utilized to predict the results of a target sample based on the non-parametric classifier. However, as a problem of multi-label classification, a pedestrian image usually has multiple attribute labels. A sample that is reliable for one attribute may not be reliable for the others. Therefore, only retaining the holistic feature of one pedestrian image as different attribute prototypes will not be the optimal choice for attribute prediction in an unseen domain. To tackle this challenge, we construct the network in training phase through extending the popular PAR strong-baseline[15] with an attribute localization branch to extract the attribute-specific features. Specifically, two kinds of memory banks are built to store both the global and attribute-specific features as attribute prototypes at different granularity levels. Then, the classification results based on the two kinds of prototypes are fused to enhance the robustness of attribute prediction in unseen domain. Note that the proposed AdaGPAR is training-free in the inference phase and predicts the attributes on target domain in an online manner. Thus, the proposed method is more efficient for the real-world applications.

In summary, the contributions of this paper are as follows:

1) A new paradigm is proposed to solve the task of GPAR with TTA by exploiting the domain information contained in the target samples. As far as we know, it is the first work that improves the generalization ability of a PAR model from the perspective of TTA.

2) A TTA-based method, named AdaGPAR, is presented to enhance the PAR accuracy in target domains without backward gradient updating. It predicts the attributes of target samples in an online manner, which is time efficient and flexible for real-world applications.

3) Extensive experiments are performed on the UPAR[8] benchmark. In comparison with the baseline in [8] and the other two TTA-based methods, the proposed AdaGPAR obtains superior performance.

2 Related work

2.1 Pedestrian attribute recognition

Most of current studies on PAR usually concentrate on enhancing recognition accuracy on specific datasets. Deep learning is commonly utilized in these approaches to acquire robust feature representations. The early deep-learning-based methods[16, 17] typically train a holistic CNN model for joint multi-attribute classification. A number of later approaches use auxiliary information, such as pose[18] or body parts[19], to improve attribute localization. Instead of using the auxiliary part-based information, many recent approaches[20–22] design various attention modules to enhance the performance of attribute recognition. Yang et al.[20] design a cascaded split-and-aggregate model that learns both the individuality and commonality among attributes. This model uses an attribute-specific attention module to locate the most informative region for each attribute. Guo et al.[21] introduce two types of attention-consistency losses. These losses ensure that a pedestrian image maintains an equivalent attention map across various spatial transforms as well as consistent attention maps across different networks. Liu et al.[22] propose a dual-branch self-attention network for PAR that learns the attribute-specific features and regional contextual features simultaneously. Due to the difficulties in localizing the fine-grained attributes, some researchers[23–25] resort to explore the relations among multiple attributes. Tan et al.[23] propose a unified graph convolutional network (GCN) that jointly models both the semantic and contextual relations. Fan et al.[24] construct a relationship framework using GCN to model various types of relations among attributes, including inner-relations, hierarchical-relations, and spatial-relations. Cheng et al.[25] incorporate the textual modality to explore the textual correlations among attribute annotations and utilize the transformer encoder to capture both the intra-modal and cross-modal correlations. In addition, Weng et al.[26] enhance PAR performance by delving into both the attribute localization and correlation, where the authors exploit the attention mechanism to extract attribute-specific features and employ the transformer encoder to model the attribute correlations.

Currently, fine-tuning the foundation model on the downstream tasks has become a prevalent paradigm in computer vision. A number of human-centric foundation models[27–29] have also been released. All of them obtain impressive recognition accuracy on the downstream task of PAR.

However, all the aforementioned studies neglect the challenges encountered in the real-world application, i.e., category shift and domain shift. Different from the traditional PAR work, Li et al.[30] formulate the incremental PAR as a problem of multi-label continual learning with incomplete labels. To tackle the category shift, the authors propose a self-training based approach via dual uncertainty-aware pseudo-labeling to transfer the knowledge learned in previous tasks to novel tasks. For the issue of domain shift, Specker et al.[8] present a novel benchmark, named UPAR, for GPAR through unifying four popular PAR datasets (domains) with 40 binary attributes. The authors also propose a baseline method that aggregates several techniques, such as dropout, data augmentation and label smoothing, in the training process to enhance the generalizability of PAR model across different domains. However, it overlooks the domain information carried by the target samples. Instead, our work explores the GPAR from the perspective of TTA[11] to mitigate the above issue.

2.2 Test-time adaptation

Test-time adaptation (TTA) provides a flexible strategy for domain generalization (DG), which aims to adapt the model trained in source domain (source model) to unseen domains using the online unlabeled target samples solely before making prediction[11]. Numerous approaches for TTA have been proposed in recent years, which can be typically divided into two categories, i.e., test-time training (TTT) methods and fully TTA methods.

TTT methods[31, 32] fine-tune the source model via auxiliary self-supervised learning task during the test phase. Sun et al.[31] design a task for rotation classification to predict the rotation angle of rotated images. Liu et al.[32] propose to learn an extra self-supervised branch based on contrastive learning in the source model. The main characteristic of TTT is that the source model is trained with both the supervised loss (main task) and self-supervised loss (auxiliary task) simultaneously based on a multi-task architecture. It modifies the original training procedure that may be not feasible in the real-world applications.

Different from the TTT methods, fully TTA methods need not change the training procedure of source model. They typically adapt the trained model solely with target samples using normalization-based methods[33, 34], entropy minimization[35, 36], or prototype-based method[37]. Besides the above approaches, Jang et al.[38] propose a TTA method via self-training with nearest neighbor information to mitigate the confirmation bias. Wang et al.[39] address TTA as a problem of feature revision. The authors propose a self-distillation strategy to ensure the feature uniformity at test time and also present a memorized spatial local clustering strategy to align the representations among the neighborhood target samples. In addition, Zhang et al.[14] propose a novel TTA method under a non-parametric paradigm by storing features and predicted pseudo-labels of the target samples. This method is parameter-free and can also effectively alleviate the knowledge forgetting in the continual adapting. It is worth noting that the aforementioned approaches are validated for conventional classification tasks. Our work devotes to adapt the TTA method to GPAR which is a problem of multi-label classification.

Moreover, some studies[11] also classify the source-free domain adaptation (SFDA)[13] as a type of TTA. However, SFDA can access all the unlabeled target samples to adapt the source model, which may be impractical in real-world applications. Therefore, our work mainly focuses on improving the generalizability of a PAR model with the target samples in an online manner.

3 Problem definition

We formulate the GPAR from the perspective of TTA. In particular, we consider a source domain dataset $\mathcal{D}^s = \{(\boldsymbol{x}_i^s, \boldsymbol{y}_i^s)\}_{i=1}^{n_s}$, where $(\boldsymbol{x}_i^s, \boldsymbol{y}_i^s)$ is sampled i.i.d. from a distribution $\mathbf{D}^s$, n_s is the total number of samples in $\mathcal{D}^s$ and $\boldsymbol{y}_i^s \in \{0,1\}^{m_s}$ is a multi-hot label vector for the training sample $\boldsymbol{x}_i^s$ in which m_s is the number of annotated attributes. The GPAR aims to train a model Θ on the source domain $\mathcal{D}^s$ which can perform well on the unseen target domain $\mathcal{D}^t = \{\boldsymbol{x}_i^t\}_{i=1}^{n_t}$, where the target sample $\boldsymbol{x}_i^t$ is sampled from a distribution $\mathbf{D}^t$ $(\mathbf{D}^t \neq \mathbf{D}^s)$ and n_t is the number of samples in $\mathcal{D}^t$. According to the definition of TTA, only the trained model Θ on $\mathcal{D}^s$ and the unlabeled target samples $\boldsymbol{x}_i^t$ in an online manner are available when we make predictions on $\mathcal{D}^t$ in the test phase. In summary, the characteristics of GPAR in our work are as follows:

1) In the training phase, the PAR model Θ is trained with $\mathcal{D}^s$ solely.

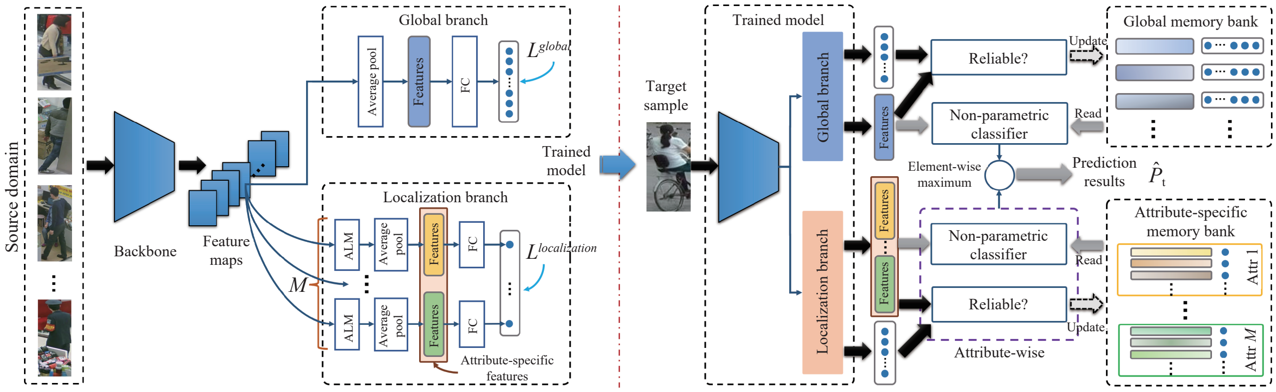
2) $\mathbf{D}^t \neq \mathbf{D}^s$, which is known as domain shift/gap.

3) In the test phase, only Θ and $\boldsymbol{x}_i^t$ are available in an online manner.

Noted that, 1) our work focuses on mitigating the problem of domain shift, so we assume that the source domain and the target domain have the same attribute categories. Thus, the number of attributes in $\mathcal{D}^t$, i.e., m_t, equals m_s. 2) The samples in $\mathcal{D}^s$ are only sampled from a single domain, while the $\mathcal{D}^t$ may consist of multiple datasets.

4 Method

Fig. 2 displays an overview of the proposed AdaGPAR, which consists of two components, i.e., model training on source domain (training phase) and model adaptation on target domains via TTA (test phase). In order to predict the attributes of a target sample in the test phase, one key step in AdaGPAR is to cache the reliable target samples for each attribute category. However, a pedestrian image possesses multiple binary attributes. A pedestrian image that is reliable to one attribute with high prediction confidence may be not reliable to other attribute categories. Thus, if we only memorize the global features that will inevitably interfere the further prediction based on measuring the similarities between the feature of a target sample and the cached features in memory. To alleviate this issue, two strategies are applied in AdaGPAR. For one thing, an attribute localization branch is introduced to the source model to extract the attribute-specific features; For another thing, two



(a) Model training on source domain

(b) Model adaptation via TTA

Fig. 2 An overview of the proposed AdaGPAR. It consists of two components, i.e., (a) model training on source domain and (b) model adaptation on target domains via TTA. M is the number of attribute categories, ALM is short for attribute localization module, FC is short for fully-connection, and $\hat{P}_t$ denotes the prediction results of the target sample. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

types of memory banks are constructed to store the holistic features and attribute-specific features simultaneously. This section will explain them in detail from the following two aspects.

4.1 Model training on source domain

For the source domain, we build a two-branch PAR model through extending the popular PAR method, i.e., strongbaseline[15], with an attribute localization branch. As shown in Fig. 2(a), the global branch inherits the original structure of strongbaseline. Inspired by the work [40], the localization branch contains M[1] attribute localization modules, each of which consists of a squeeze-and-excitation (SE) module[41] to exploit the inner-channel correlations of the input feature maps, followed by a spatial transformer[42] to localize the attribute-specific regions. The two branches share the same backbone. We apply the localization branch to the feature maps of the final layer.

Fig. 3 presents the detailed structure of an attribute localization module. When the feature maps of the final layer, termed as Z_i, is input to the localization module, a weight vector will be first generated by the SE module. The weight vector is multiplied by Z_i channel-wisely to produce the weighted features which are further added to Z_i to preserve the complementary information. The output of SE module, $\tilde{Z}_i$, is then fed into the spatial transformer to generate the attribute-specific features. Specifically, a fully-connected layer takes the input $\tilde{Z}_i$ to estimate the parameters of the transformation matrix $\mathcal{T}_\theta$. $\mathcal{T}_\theta$ is subsequently applied to $\tilde{Z}_i$ and then sampled by bilinear interpolation to generate the attribute-specific feature maps Z_{ij} which are further utilized for attribute prediction.

We optimize the PAR model with weighted binary cross entropy (BCE) loss. For one of the two branches, the BCE loss is defined as follows:

$$L^b = -\frac{1}{n_s}\sum_{i=1}^{n_s}\sum_{j=1}^{M}\omega_j\left(y_{ij}\log\left(\hat{p}_{ij}^b\right)+\right.$$
$$\left.(1-y_{ij})\log\left(1-\hat{p}_{ij}^b\right)\right) \tag{1}$$

where y_{ij} is the ground truth label for the i-th sample in $\mathcal{D}^s$ about the j-th attribute and $\hat{p}_{ij}^b$ is the corresponding prediction result by branch $b \in \{\text{global, localization}\}$. ω_j denotes the loss weight for the j-th attribute to mitigate the distribution imbalance. ω_j is calculated as follows:

$$\omega_j = \begin{cases} \exp\left((1-\gamma_j)/\delta^2\right), & \text{if } y_{ij}=1 \\ \exp\left(\gamma_j/\delta^2\right), & \text{if } y_{ij}=0 \end{cases} \tag{2}$$

where γ_j is the ratio of positive samples for the j-th attribute and δ is the temperature parameter which is set to 1 in this work. Finally, the total loss is formulated as $L = L^{global} + L^{localization}$.

4.2 Model adaptation via TTA

In this work, the proposed AdaGPAR predicts the attributes on target domains only with the source model and unlabeled target samples in an online manner. Inspired by the work AdaNPC[14], AdaGPAR constructs two kinds of memory banks to cache the reliable global features and attribute-specific features, respectively. As displayed in Fig. 2(b), for an unseen target sample $\boldsymbol{x}_i^t$ in the test phase, it is fed into the source model to obtain the predictions of the linear classifiers and capture the related features. Specifically, we can get two kinds of prediction results, i.e., the output of the global branch $\hat{\boldsymbol{p}}_i^g$ and the output of the localization branch $\hat{\boldsymbol{p}}_i^l$. Similarly, the global features $\boldsymbol{z}_i^g$ and a group of attribute-specific

[1] M denotes the number of attribute categories. We assume $m_s = m_t = M$ in this work.

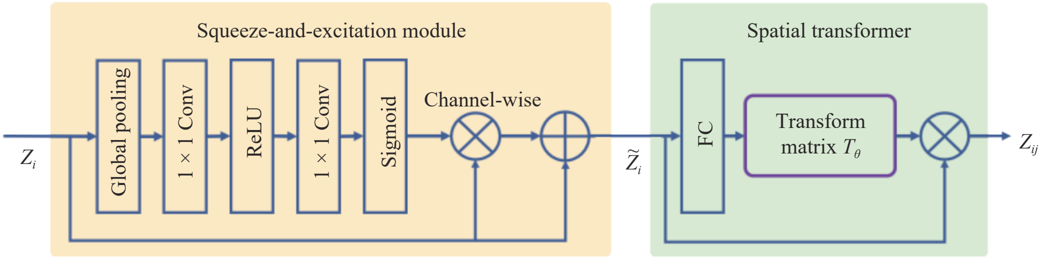


Fig. 3 An illustration for the attribute location module. It consists of a squeeze-and-excitation (SE) module[41] and a spatial transformer[42]. The output Z_{ij} denotes the feature maps for the j-th attribute of sample i.

features $\boldsymbol{z}_i^l=\bigcup_{j=1}^M \boldsymbol{z}_{ij}^l$ are extracted. Then, we predict the attribute categories of $\boldsymbol{x}_i^t$ by the non-parametric classifier via calculating the feature similarity with the vectors stored in the memory banks. Subsequently, $\boldsymbol{z}_i^g$ and $\boldsymbol{z}_i^l$ will be selected to update the memory banks if they are determined as reliable. Algorithm 1 presents the detailed procedure.

Algorithm 1. Attribute prediction via AdaGPAR

```
Input: $\Theta$, $\boldsymbol{x}_i^t$, G-MB, AS-MB
Output: $\hat{\boldsymbol{p}}_i^t$
$\left\{\boldsymbol{z}_i^g, \bigcup_{j=1}^M \boldsymbol{z}_{ij}^l, \hat{\boldsymbol{p}}_i^g, \hat{\boldsymbol{p}}_i^l\right\}=\Theta\left(\boldsymbol{x}_i^t\right)$;
// Attribute prediction with kNN.
Confirm $\eta^G\left(\boldsymbol{x}_i^t\right)$ using kNN based on $\boldsymbol{z}_i^g$ and vectors
  in G-MB;
Confirm $\eta^{AS}\left(\boldsymbol{x}_i^t\right)$ using attribute-wise kNN based on
  $\bigcup_{j=1}^M \boldsymbol{z}_{ij}^l$ and vectors in AS-MB;
$\hat{\boldsymbol{p}}_i^t = elmax\left(\eta^G\left(\boldsymbol{x}_i^t\right), \eta^{AS}\left(\boldsymbol{x}_i^t\right)\right)$;
// Update G-MB.
Calculate $\mathcal{C}_i^g$ using (3);
if $\mathcal{C}_i^g < \mathbb{T}_g$ then
    Update G-MB using $(\boldsymbol{z}_i^g, \hat{\boldsymbol{p}}_i^g)$;
end if
// Update AS-MB.
for $j \in \{1, \cdots, M\}$ do
    // Positive.
    if $\mathcal{R}_{pos}\left(\boldsymbol{z}_{ij}^l, \hat{p}_{ij}^l\right) == 1$ then
        Add $\left(\boldsymbol{z}_{ij}^l, \hat{p}_{ij}^l\right)$ to the block of the $j$-th attribute
        in AS-MB as positive vector;
    end if
    // Negative.
    if $\mathcal{R}_{neg}\left(\boldsymbol{z}_{ij}^l, \hat{p}_{ij}^l\right) == 1$ then
        Add $\left(\boldsymbol{z}_{ij}^l, \hat{p}_{ij}^l\right)$ to the block of the $j$-th attribute
        in AS-MB as negative vector;
    end if
end for
```

4.2.1 Memory bank construction

Global memory bank and attribute-specific memory bank are constructed in this work. The following interprets the two kinds of memory banks from the aspects of reliable data selection and memory bank updating.

Global memory bank (G-MB) is used to store the reliable target features and their prediction scores obtained by the global branch. Since the global features cover knowledge about all the attributes, it may not be optimal to memorize the samples with high predicted scores on just a few categories. To tackle this issue, we calculate the attribute-wise entropy based on the prediction results $\hat{\boldsymbol{p}}_i^g$. Then, the mean value is computed to further compare with a constant threshold. In particular, the criterion for $\boldsymbol{x}_i^t$ is defined as follows:

$$\mathcal{C}_i^g=-\frac{1}{M}\sum_{j=1}^{M}\hat{p}_{ij}^g\log\hat{p}_{ij}^g+\left(1-\hat{p}_{ij}^g\right)\log\left(1-\hat{p}_{ij}^g\right) \qquad (3)$$

where $\hat{p}_{ij}^g$ is the prediction score about the j-th attribute of $\boldsymbol{x}_i^t$ obtained by the global branch. The pair $(\boldsymbol{z}_i^g, \hat{\boldsymbol{p}}_i^g)$ is selected to update the G-MB when $\mathcal{C}_i^g < \mathbb{T}_g$. $\mathbb{T}_g$ is set to 0.4 in the experiments. These reliable pairs in the memory bank act as the support vectors (attribute prototypes) for the subsequent prediction based on the non-parametric classifier. As the memory is large enough to cache the reliable global features in our experiments, we just add the selected $(\boldsymbol{z}_i^g, \hat{\boldsymbol{p}}_i^g)$ to the end of G-MB directly. When AdaGPAR is employed to the real-world scenarios, we can clear the early data with higher value of $\mathcal{C}^g$ gradually to satisfy the constraint on memory size.

The criterion used in G-MB still has two limitations: 1) Equation (3) enforces to choose the samples with lower entropy about all the attributes. It inevitably abandons lots of samples that contain useful knowledge for some attributes. 2) The constraint on the mean value cannot ensure that the entropy of each attribute satisfies the threshold. This will inevitably introduce noisy information.

Attribute-specific memory bank (AS-MB) is constructed to mitigate the above problems. Different from the G-MB, AS-MB includes M blocks (as shown in Fig. 2(b)), in which each block relates to an attribute category. Thus, we can operate the memory in attribute level feasibly. Recognizing a single attribute is indeed a task of binary classification. So, both the positive and negative samples for an attribute are selected to store to the AS-MB. For convenience, we make decision based on the output of localization branch $\hat{\boldsymbol{p}}_i^l$ directly. $\left(\boldsymbol{z}_{ij}^l, \hat{p}_{ij}^l\right)$ is chosen as a reliable positive pair if the output of function $\mathcal{R}_{pos}\left(\boldsymbol{z}_{ij}^l, \hat{p}_{ij}^l\right)$ is 1. The definition of $\mathcal{R}_{pos}\left(\boldsymbol{z}_{ij}^l, \hat{p}_{ij}^l\right)$ is as follows:

$$\mathcal{R}_{pos}\left(\boldsymbol{z}_{ij}^{l},\hat{p}_{ij}^{l}\right)=\begin{cases}1, & \text{if } \hat{p}_{ij}^{l}>\mathbb{T}_{l}^{pos}\\ 0, & \text{if } \hat{p}_{ij}^{l}\le\mathbb{T}_{l}^{pos}\end{cases}\tag{4}$$

where $\mathbb{T}_{l}^{pos}$ is a constant threshold, and it is set to 0.9 in our work. Similarly, $\left(\boldsymbol{z}_{ij}^{l},\hat{p}_{ij}^{l}\right)$ is chosen as a reliable negative pair if the output of function $\mathcal{R}_{neg}\left(\boldsymbol{z}_{ij}^{l},\hat{p}_{ij}^{l}\right)$ is 1. (5) provides the definition of $\mathcal{R}_{neg}\left(\boldsymbol{z}_{ij}^{l},\hat{p}_{ij}^{l}\right)$,

$$\mathcal{R}_{neg}\left(\boldsymbol{z}_{ij}^{l},\hat{p}_{ij}^{l}\right)=\begin{cases}1, & \text{if } \hat{p}_{ij}^{l}<\mathbb{T}_{l}^{neg}\\ 0, & \text{if } \hat{p}_{ij}^{l}\ge\mathbb{T}_{l}^{neg}\end{cases}\tag{5}$$

where $\mathbb{T}_{l}^{neg}$ is also a constant threshold which is set to 0.15 in this work.

Noted that, the number of negative samples is much larger than that of positive samples for one attribute. It may lead to the problem of out-of-memory (OOM) if we store the large-scale negative data continuously. Therefore, we set the maximum size of negative data for each attribute block in AS-MB. When the negative part of a block reaches the maximum size, the previous negative pairs with maximum $\hat{p}_{ij}^{l}$ will be removed.

4.2.2 Attribute prediction via non-parametric classifier

In the proposed AdaGPAR, k-nearest neighbours (kNN) is utilized to predict the attributes of target samples in an online manner based on the data cached in the memory bank.

To avoid confusion, we use $\left(\boldsymbol{v}_{k}^{G},\hat{\boldsymbol{y}}_{k}^{G}\right)$ to represent a pair that is read from G-MB, where $\boldsymbol{v}_{k}^{G}$ is one of the cached global features and $\hat{\boldsymbol{y}}_{k}^{G}$ is the corresponding prediction results (pseudo-labels). Given an unseen sample $\boldsymbol{x}_{i}^{t}$, its prediction score is calculated as follows:

$$\eta^{G}\left(\boldsymbol{x}_{i}^{t}\right)=sigmoid\left(\sum_{k\in\mathcal{K}}w_{ik}\hat{\boldsymbol{y}}_{k}^{G}\right)\tag{6}$$

where w_{ik} is the cosine similarity between $\boldsymbol{z}_{i}^{g}$ and $\boldsymbol{v}_{k}^{G}$. $\mathcal{K}$ is a set for the top K closest vectors to $\boldsymbol{z}_{i}^{g}$ in the G-MB. And $K=5$ in the experiments for the G-MB. Different from AdaNPC[14], which initializes the memory bank with the features of source samples, we only store the features of target samples. When the number of support vectors in the memory bank is less than $K\times\mathbb{L}$, we will enforce the $\eta^{G}\left(\boldsymbol{x}_{i}^{t}\right)$ to equal the prediction scores of the source model. $\mathbb{L}$ is set to 50 for the G-MB.

It has the similar procedure to obtain the prediction results $\eta^{AS}\left(\boldsymbol{x}_{i}^{t}\right)$ based on the vectors stored in AS-MB. The sole difference is that we calculate prediction score attribute by attribute. Concretely, the K is set to 50 and $\mathbb{L}$ is set to 4 for AS-MB empirically.

Finally, the prediction scores of $\boldsymbol{x}_{i}^{t}$ are obtained by performing element-wise maximum between $\eta^{G}\left(\boldsymbol{x}_{i}^{t}\right)$ and $\eta^{AS}\left(\boldsymbol{x}_{i}^{t}\right)$,

$$\hat{\boldsymbol{p}}_{i}^{t}=elmax\left(\eta^{G}\left(\boldsymbol{x}_{i}^{t}\right),\eta^{AS}\left(\boldsymbol{x}_{i}^{t}\right)\right).\tag{7}$$

5 Experiments

5.1 Dataset and evaluation metrics

We evaluate the effectiveness of proposed AdaGPAR on the UPAR[8] dataset, which is built through unifying four popular PAR datasets, i.e., Market1501[7], PA100k[43], PETA[44] and RAPv2[1]. UPAR dataset contains 224 737 pedestrian images with 40 binary attributes over 12 categories. It is now the sole benchmark for GPAR. Four kinds of partitions (as shown in Table 1) are used to measure the performance.

Table 1 Different data partitions in UPAR. "Part." is short for partition. M, PA, P and R denote Market1501, PA100k, PETA and RAPv2, respectively.

Part.	Source domain		Target domain	
	Dataset	# Training samples	Dataset	# Test samples
1	M	16 289	R, P, PA	32 766
2	PA	88 923	R, P, M	35 873
3	P	10 402	R, PA, M	38 896
4	R	63 264	P, PA, M	30 042

Three types of metrics are adopted to evaluate the performance of AdaGPAR, i.e., mean accuracy (mA), instance-based metrics, and mFive. mA[1] is a kind of label-based metric, which intuitively measures the ability of a PAR model to complete one visual attribute recognition task. Instance-based metrics[1] include accuracy, precision, recall rate, and F_1 score. They are used to measure the consistence of all attributes occurring in a given pedestrian image. mFive[20] is the average of five criteria involving mA and the instance-based metrics. It offers a more comprehensive evaluation of different methods, as it avoids the problem of some methods excelling at one specific metric but performing poorly at others.

5.2 Implementation details

ResNet50[45] and ConvNeXt-B[46] are adopted as the backbone in this work. They are pretrained on the ImageNet dataset.

In the training phase, we follow the settings in the UPAR[8] to fine-tune the model with above two backbones. In particular, the input images are resized to 256×192. We augment them with random horizontal mirroring, random cropping, and AugMix. The Adam optimizer is used with initial learning rate of 1×10^{-4} and weight decay of 5×10^{-4}. And the plateau scheduler is applied to reduce the learning rate with a factor of 0.1 when the evaluation results are not enhanced for four epochs. We set the batchsize to 32 for Market1501 and PETA, while it is set to 64 for PA100k and RAPv2. Moreover, label smoothing and exponential moving average (EMA) are also applied to improve the model performance.

In the test phase, we predict the attributes of target samples online using the source model and the data cached in the memory bank without backpropagation. Sequences of the unseen samples with different orders and different batchsizes are adopted to evaluate the effectiveness of AdaGPAR. Without specification, the following experimental results are obtained with the sequence of default order and the batchsize of 8.

5.3 Comparison with different baselines

The final performance of proposed AdaGPAR and three baselines (one is for the baseline of UPAR, and the other two are the TTA-based methods) on UPAR dataset are presented in Table 2, where the mean values and the variances are reported over four different data partitions (shown in Table 1). These experimental results are obtained with two kinds of backbones, i.e., ResNet50[45] and ConvNeXt-B[46]. The detailed results of each data partition are listed in the supplementary material.

Compared to the reproduced results of UPAR baseline[8] (UPAR* in Table 2), the proposed AdaGPAR produces superior performance in terms of mA and F_1 score. In particular, the values of mA are enhanced by 1.0% and 2.0% with ResNet50 and ConvNeXt-B, respectively. T3A[37] and AdaNPC[14] are two TTA-based baseline methods that are reproduced to fulfil the task of PAR, in which only the global features are considered. T3A exceeds the results of UPAR on mA, and AdaNPC obtains impressive performance on F_1 scores. However, the two methods only utilize the global features which inevitably introduce interference information in selecting reliable samples for each category. Thus, none of them can achieve superior performance on both mA and F_1 score. Different from T3A and AdaNPC, both the global and attribute-specific features are considered in the AdaGPAR. Thus, AdaGPAR attains the best performance on mA with comparable results of F_1 scores. We also calculate the mFive[20] as a comprehensive evaluation metric (the last column in Table 2). We can find that AdaGPAR obtains the best performance on the metric mFive with both the backbones of ResNet50 and ConvNeXt-B, which demonstrates its effectiveness.

5.4 Ablation study

5.4.1 Effectiveness of key components

The mechanism of TTA and attribute localization branch are the key components of AdaGPAR. We apply them to the strongbaseline[8] and UPAR baseline[15] respectively to validate their significance in improving the performance on target domain. The contribution of each term is shown in Table 3. Both the mA and F_1 score are enhanced when the TTA is utilized with the backbones of ResNet50 and ConvNeXt-B. The superior performance indicates the effectiveness to improve the generalization by exploiting the domain information contained in unlabeled target samples. However, the improvement on mA is not significant if we only memorize the global features which compound the knowledge about all the attributes. The mA is enhanced by more than 1.0% and 1.5% for ResNet50 and ConvNeXt-B respectively when the localization branch (L.B.) is further introduced to the network to extract the attribute-specific features. Moreover, they also obtain the best performance on mFive when the TTA and L.B. are used simultaneously ("+TTA&L.B."). The superior performance demonstrates the effectiveness of AdaGPAR to mitigate the issue of domain shift. The experimental results also indicate the flexibility of AdaG-

Table 2 Comparison with different baselines on UPAR dataset. "*" denotes that we reproduce these baselines following the recommended setups. The **bold** and underline values demonstrate the best and second-best results, respectively.

Methods	Backbone	mA (%)	Accuracy (%)	Precision (%)	Recall (%)	F_1 score (%)	mFive (%)
UPAR[8]	ResNet50	67.0±2.5	–	–	–	**74.2**±4.5	–
	ConvNeXt-B	70.5±1.9	–	–	–	80.1±2.7	–
UPAR*[8]	ResNet50	67.4±2.6	59.5±5.4	73.8±4.2	72.8±4.9	73.3±4.4	69.3
	ConvNeXt-B	70.4±3.2	67.8±5.0	80.6±4.1	78.9±3.3	79.7±3.7	75.5
T3A*[37]	ResNet50	68.3±3.5	57.5±5.1	70.8±4.7	72.8±3.9	71.8±4.3	68.3
	ConvNeXt-B	71.5±4.0	65.9±4.5	78.7±3.4	78.0±3.4	78.3±3.4	74.5
AdaNPC*[14]	ResNet50	65.8±2.4	**60.4**±5.3	**76.2**±4.2	72.0±4.7	74.0±4.2	69.7
	ConvNeXt-B	69.6±2.9	**68.5**±4.7	**81.9**±3.8	78.5±3.2	**80.2**±3.4	75.7
AdaGPAR (ours)	ResNet50	**68.4**±3.3	60.3±6.3	71.3±4.9	**77.1**±5.5	74.1±5.1	**70.2**
	ConvNeXt-B	**72.4**±2.8	67.9±4.5	78.1±4.0	**81.8**±2.5	79.9±3.3	**76.1**

Table 3 Ablation analysis for the key components in proposed AdaGPAR. "UPAR*" and "strongbaseline*" denote that we reproduced these methods with recommended setups. L.B. is short for localization branch. The values listed in this table are obtained through averaging the results of four different data partitions.

Methods	Backbone	mA (%)	Accuracy (%)	Precision (%)	Recall (%)	F_1 score (%)	mFive (%)
Strong baseline*[15]	ResNet50	66.0	57.0	**73.1**	69.5	71.2	67.3
	ConvNeXt-B	70.3	67.2	**81.4**	77.4	79.3	75.1
+TTA	ResNet50	66.6	57.6	71.0	72.7	71.8	67.9
	ConvNeXt-B	70.8	**67.7**	80.4	79.1	**79.7**	75.5
+TTA&L.B. (AdaGPAR)	ResNet50	**67.1**	**58.6**	70.6	**74.9**	**72.7**	**68.8**
	ConvNeXt-B	**71.9**	67.3	77.7	**81.5**	79.5	**75.6**
UPAR*[8]	ResNet50	67.4	59.5	**73.8**	72.8	73.3	69.3
	ConvNeXt-B	70.4	67.8	**80.6**	78.9	79.7	75.5
+TTA	ResNet50	67.6	60.1	72.5	75.2	73.9	69.9
	ConvNeXt-B	70.9	**68.3**	79.4	80.8	**80.1**	75.9
+TTA&L.B. (AdaGPAR)	ResNet50	**68.4**	**60.3**	71.3	**77.1**	**74.1**	**70.2**
	ConvNeXt-B	**72.4**	67.9	78.1	**81.8**	79.9	**76.1**

PAR that it can cooperate with the conventional PAR model and existing DG methods to improve the performance.

To further verify the effectiveness of the localization branch (L.B.), we replace it by applying the class activation mapping (CAM)[47] on the global feature maps to get the attribute-specific features. The ablation results based on UPAR baseline[8] are shown in Table 4. The experimental results indicate that the performance of mA can be improved by extracting the attribute-specific features with CAM ("+TTA&CAM"). However, the values of mA and mFive are inferior than those of the proposed AdaGPAR ("+TTA&L.B."). It demonstrates that the L.B. in AdaGPAR is more effective than CAM to improve the recognition performance in the target domain.

5.4.2 Sensitivity to sequence properties

In the test phase, the AdaGPAR is applied to the target sample in an online manner, where the target samples compose a sequence. We perform a number of experiments based on multiple sequences with different orders and processing batchsizes to verify whether the two properties will impact the recognition accuracy significantly.

Sequence order. In the real-world application scenarios, test samples in the sequence may be sampled from a variety of distributions. Thus, the model may be applied to the test samples from a seen distribution again after the model is updated with the samples from other different distributions. To verify the effectiveness of proposed AdaGPAR for this challenge, five sequences of target samples with different orders are generated in our experiments. As shown in Table 1, the target domain in UPAR dataset consists of three different datasets. For the default order, the target samples within each dataset are arranged together for ordering. It means that the seen distribution will not appear again when all the samples from this distribution are processed. Besides the default order, we shuffle the target samples for each data partition under four distinct seeds. Thus, a mini-batch in the sequence may consist of target samples from different distributions (seen or unseen). Fig. 4 presents the trends of mA and F_1 score with different orders based on two backbones. Their values are also obtained through aver-

Table 4 Ablation study on the localization branch (L.B.). "*" denotes that the UPAR baseline is reproduced with recommended setups.

Methods	Backbone	mA (%)	Accuracy (%)	Precision (%)	Recall (%)	F_1 score (%)	mFive (%)
UPAR*[8]	ResNet50	67.4	59.5	**73.8**	72.8	73.3	69.3
	ConvNeXt-B	70.4	67.8	**80.6**	78.9	79.7	75.5
+TTA	ResNet50	67.6	60.1	72.5	75.2	73.9	69.9
	ConvNeXt-B	70.9	**68.3**	79.4	80.8	**80.1**	75.9
+TTA&CAM	ResNet50	67.7	60.1	72.3	75.5	73.9	69.9
	ConvNeXt-B	71.1	**68.3**	79.2	81.1	**80.1**	76.0
+TTA&L.B. (AdaGPAR)	ResNet50	**68.4**	**60.3**	71.3	**77.1**	**74.1**	**70.2**
	ConvNeXt-B	**72.4**	67.9	78.1	**81.8**	79.9	**76.1**

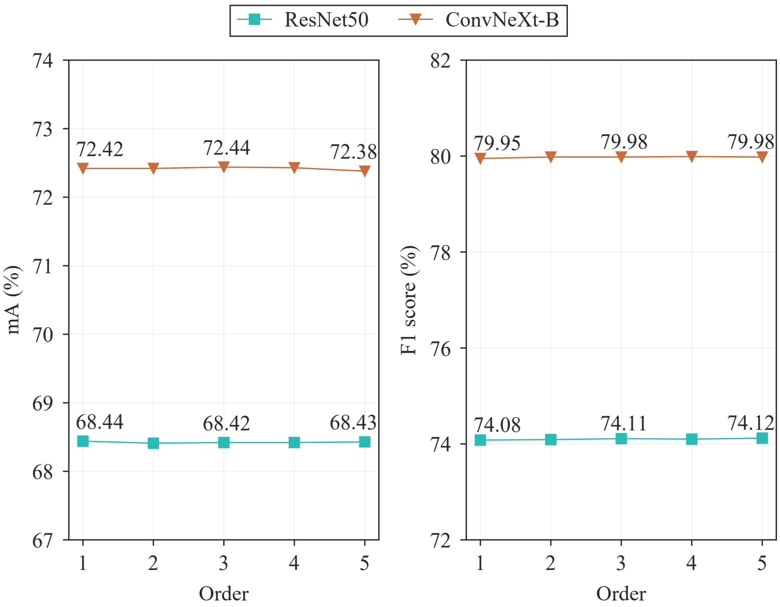


Fig. 4 The trends of mA and F_1 score with different sequence orders. The mA and F_1 score are obtained by averaging the results of four data partitions. The processing batchsize is set to 8. Left: mA; Right: F_1 score. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

aging the results of four data partitions. We can find that the sequence order plays little effect on the mA and F_1 score. It demonstrates the effectiveness of AdaGPAR to alleviate the aforementioned challenge.

Processing batchsize. We set five different processing batchsizes, i.e., {2, 4, 8, 16, 32}, to verify the influence on recognition accuracy. The average results of four data partitions are shown in Fig. 5. We can find that the mA and F_1 score decrease slightly with two different backbones when the batchsize ranges from 2 to 32. However, the influence is also not significant, where the differences between maximum and minimum are less than 0.2% for both the mA and F_1 score. Noted that, the elapsed time in test phase increases gradually along with the reducing of batchsize. Therefore, we set it to 8 as a compromise of accuracy and elapsed time.

5.5 Memory usage and inference time

Memory usage. Two kinds of memory banks, includ-

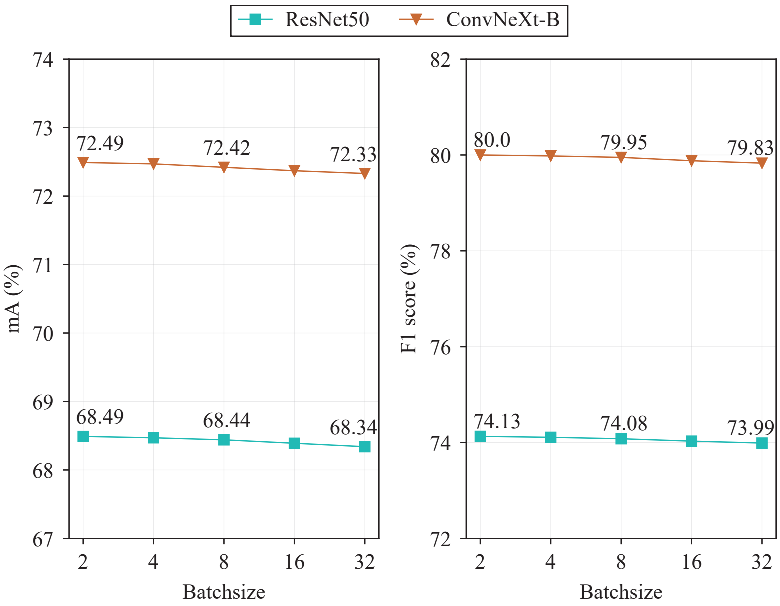


Fig. 5 The trends of mA and F_1 score with different processing batchsizes. The mA and F_1 score are obtained by averaging the results of four data partitions. The sequence follows the default order. Left: mA; Right: F_1 score. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

ing G-MB and AS-MB, are built to implement AdaGPAR. For the G-MB, consider an extreme case that we memorize all the target samples of the third partition in Table 1, the memory cost is about 79.66 MB (38 896 × 2 048 B) for ResNet50. For the AS-MB, we set a fixed size, i.e., 5 000, to cache the reliable negative features for one attribute in our experiments. We also consider an extreme situation that the size for positive samples is set to 5 000 as well (the number of positive samples of each category is usually less than 5 000). Thus, the memory cost of AS-MB for 40 different attributes is about 102.4 MB (40 × (5 000+5 000) × 256 B). It is easy to accommodate the two kinds of memory banks on current servers.

Inference time. We provide a comparison of inference time for the baselines and AdaGPAR. All the experiments are performed with a NVIDIA RTX Titan GPU. As shown in Table 5, though the kNN search in both the G-MB and AS-MB aggravates the time consumption of AdaGPAR compared with the baseline methods, it can complete the inference in real-time. Thus, it is feasible to employ the AdaGPAR to the real-world scenarios.

Table 5 A comparison of inference time for different methods. Each experiment is conducted three times to compute the average inference time.

Methods	Backbone	Inference time (ms)
UPAR	ResNet50	2.40
	ConvNeXt-B	4.50
T3A	ResNet50	7.45
	ConvNeXt-B	8.88
AdaNPC	ResNet50	4.44
	ConvNeXt-B	5.65
AdaGPAR (ours)	ResNet50	15.31
	ConvNeXt-B	16.44

Analysis. The proposed AdaGPAR performs attributes prediction in target domain through kNN searching in the two kinds of memory banks. The memory usage and inference time are increased. However, the proposed AdaGPAR is training-free, only the online test samples from target domain are selected to capture the distribution information. Compared with simply enlarging the number of model parameters and extending the training time in the source domain, is the introducing of TTA a better choice to improve the model's generalizability? To verify the superiority of proposed AdaGPAR, we replace the backbone of UPAR baseline by the ResNet101, whose size is about 1.7 times larger than that of AdaGPAR (163.1 MB VS 94.0 MB). The increasing number of model parameters results in longer training time. For example, the training time is increased by 22.5 minutes approximately based on the third partition of UPAR dataset. Table 6 presents the comparison results between AdaGPAR and UPAR baseline with the ResNet101 backbone. We can find that the proposed AdaGPAR based on ResNet50 can also obtain superior performance than the UPAR baseline with ResNet101. It demonstrates that TTA is more effective than simply expanding the number of model parameters for improving the generalizability of a model.

5.6 Limitations

Our work has some limitations. 1) Choosing the reliable features for different attributes is one of the key steps in AdaGPAR. We now determine the features of a sample as reliable only based on the prediction scores. However, the high prediction score may not be enough to indicate the reliability of one feature belonging to an attribute prototype. Therefore, we will introduce the uncertainty estimation in the future work to alleviate the above issue. 2) Since the kNN classifier is adopted to predict the attributes, the representation ability of the source model plays an important role to get satisfied accuracy. However, the backbones used in our work are all pretrained with ImageNet which has a remarkable domain gap with pedestrian images. This problem will inevitably interfere the robustness of extracted features. So, we will introduce the human-centric foundation model in the future work to further improve the performance in target domain.

6 Conclusions

This paper focuses on the problem of generalizable pedestrian attribute recognition (GPAR), which poses a great challenge in real-world applications. Though numerous approaches of domain generalization have been proposed to tackle the issue of distribution shift, they usually concentrate on the data of source domain solely but overlook the domain information carried by the target samples. Therefore, we propose a novel method for GPAR, named AdaGPAR, from the perspective of test-time adaptation (TTA). The proposed AdaGPAR per-

Table 6 Performance comparison between the proposed AdaGPAR and the UPAR baseline with the ResNet101 backbone. "*" denotes that the UPAR baseline is reproduced with recommended setups.

Methods	Backbone	Model size	mA (%)	Accuracy (%)	Precision (%)	Recall (%)	F_1 (%)	mFive (%)
UPAR*[8]	ResNet50	90.3 MB	67.4	59.5	**73.8**	72.8	73.3	69.3
	ResNet101	163.1 MB	68.2	60.2	73.5	74.3	73.9	70.0
AdaGPAR (ours)	ResNet50	94.0 MB	**68.4**	**60.3**	71.3	**77.1**	**74.1**	**70.2**

forms prediction only based on the source model and unlabeled target samples in an online manner without additional training. Specifically, two kinds of memory banks are constructed to cache the reliable features of both global and attribute-specific types with their prediction scores. The kNN classifier is subsequently adopted to predict attributes based on the memorized data. Thus, the AdaGPAR is training-free in target domain and can be employed to real-world scenarios flexibly. Extensive experiments are conducted on the UPAR dataset, where sequences of the target samples with different orders and batchsizes are adopted to measure the performance of AdaGPAR. The superior performance demonstrates the effectiveness of AdaGPAR in improving the generalizability of a PAR model via TTA.

Acknowledgements

This work was supported in part by the National Science and Technology Major project, China (No. 2022ZD0117901); in part by the National Natural Science Foundation of China (Nos. 62373355, 62276256 and 62106260).

Declarations of conflict of interest

The authors declared that they have no conflicts of interest to this work.

References

[1] D. W. Li, Z. Zhang, X. T. Chen, K. Q. Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1575–1590, 2019. DOI: 10.1109/TIP.2018.2878349.

[2] Q. Y. Peng, L. X. Yang, X. H. Xie, J. H. Lai. Learning weak semantics by feature graph for attribute-based person search. *IEEE Transactions on Image Processing*, vol. 32, pp. 2580–2592, 2023. DOI: 10.1109/TIP.2023.3270741.

[3] S. Z. Li, H. M. Yu, R. Hu. Attributes-aided part detection and refinement for person re-identification. *Pattern Recognition*, vol. 97, Article number 107016, 2020. DOI: 10.1016/j.patcog.2019.107016.

[4] J. Q. Zhu, L. Liu, Y. B. Zhan, X. B. Zhu, H. Q. Zeng, D. C. Tao. Attribute-image person re-identification via modal-consistent metric learning. *International Journal of Computer Vision*, vol. 131, no. 11, pp. 2959–2976, 2023. DOI: 10.1007/s11263-023-01841-7.

[5] W. G. Wang, Y. L. Xu, J. B. Shen, S. C. Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 4271–4280, 2018. DOI: 10.1109/CVPR.2018.00449.

[6] Y. W. Zhang, P. Zhang, C. Yuan, Z. Wang. Texture and shape biased two-stream networks for clothing classification and attribute recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 13535–13544, 2020. DOI: 10.1109/CVPR42600.2020.01355.

[7] Y. T. Lin, L. Zheng, Z. D. Zheng, Y. Wu, Z. L. Hu, C. G. Yan, Y. Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, vol. 95, pp. 151–161, 2019. DOI: 10.1016/j.patcog.2019.06.006.

[8] A. Specker, M. Cormier, J. Beyerer. UPAR: Unified pedestrian attribute recognition and person retrieval. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, USA, pp. 981–990, 2023. DOI: 10.1109/WACV56688.2023.00104.

[9] A. Dubey, V. Ramanathan, A. Pentland, D. Mahajan. Adaptive methods for real-world domain generalization. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, USA, pp. 14335–14344, 2021. DOI: 10.1109/CVPR46437.2021.01411.

[10] Y. F. Zhang, J. D. Wang, J. Liang, Z. Zhang, B. S. Yu, L. Wang, D. C. Tao, X. Xie. Domain-specific risk minimization for domain generalization. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Long Beach, USA, pp. 3409–3421, 2023. DOI: 10.1145/3580305.3599313.

[11] J. Liang, R. He, T. N. Tan. A comprehensive survey on test-time adaptation under distribution shifts, [Online], Available: https://arxiv.org/abs/2303.15361, 2023.

[12] X. F. Liu, C. Yoo, F. X. Xing, H. Oh, G. El Fakhri, J. W. Kang, J. Woo. Deep unsupervised domain adaptation: A review of recent advances and perspectives. *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, Article number e25, 2022. DOI: 10.1561/116.00000192.

[13] C. Eastwood, I. Mason, C. K. I. Williams, B. Schölkopf. Source-free adaptation to measurement shift via bottom-up feature restoration. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[14] Y. F. Zhang, X. Wang, K. X. Jin, K. Yuan, Z. Zhang, L. Wang, R. Jin, T. N. Tan. AdaNPC: Exploring non-parametric classifier for test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, USA, pp. 41647–41676, 2023.

[15] J. Jia, H. J. Huang, X. T. Chen, K. Q. Huang. Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting, [Online], Available: https://arxiv.org/abs/2107.03576, 2021.

[16] D. W. Li, X. T. Chen, K. Q. Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition*, Kuala Lumpur, Malaysia, pp. 111–115, 2015. DOI: 10.1109/ACPR.2015.7486476.

[17] P. Sudowe, H. Spitzer, B. Leibe. Person attribute recognition with a jointly-trained holistic CNN model. In *Proceedings of IEEE International Conference on Computer Vision Workshop*, Santiago, Chile, pp. 329–337, 2015. DOI: 10.1109/ICCVW.2015.51

[18] D. W. Li, X. T. Chen, Z. Zhang, K. Q. Huang. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *Proceedings of IEEE International Conference on Multimedia and Expo*, San Diego, USA, 2018. DOI: 10.1109/ICME.2018.8486604.

[19] P. Z. Liu, X. H. Liu, J. J. Yan, J. Shao. Localization guided learning for pedestrian attribute recognition. In *Proceedings of British Machine Vision Conference*, Newcastle, UK, Article number 142, 2018.

[20] Y. Yang, Z. C. Tan, P. Tiwari, H. M. Pandey, J. Wan, Z.

Lei, G. D. Guo, S. Z. Li. Cascaded split-and-aggregate learning with feature recombination for pedestrian attribute recognition. *International Journal of Computer Vision*, vol. 129, no. 10, pp. 2731–2744, 2021. DOI: 10.1007/s11263-021-01499-z.

[21] H. Guo, X. C. Fan, S. Wang. Visual attention consistency for human attribute recognition. *International Journal of Computer Vision*, vol. 130, no. 4, pp. 1088–1106, 2022. DOI: 10.1007/s11263-022-01591-y.

[22] Z. Y. Liu, Z. Zhang, D. Li, P. Zhang, C. F. Shan. Dual-branch self-attention network for pedestrian attribute recognition. *Pattern Recognition Letters*, vol. 163, pp. 112–120, 2022. DOI: 10.1016/j.patrec.2022.10.003.

[23] Z. C. Tan, Y. Yang, J. Wan, G. D. Guo, S. Z. Li. Relation-aware pedestrian attribute recognition with graph convolutional networks. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, USA, pp. 12055–12062, 2020. DOI: 10.1609/aaai.v34i07.6883.

[24] H. N. Fan, H. M. Hu, S. L. Liu, W. Q. Lu, S. L. Pu. Correlation graph convolutional network for pedestrian attribute recognition. *IEEE Transactions on Multimedia*, vol. 24, pp. 49–60, 2022. DOI: 10.1109/TMM.2020.3045286.

[25] X. H. Cheng, M. X. Jia, Q. Wang, J. Zhang. A simple visual-textual baseline for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6994–7004, 2022. DOI: 10.1109/TCSVT.2022.3178144.

[26] D. F. Weng, Z. C. Tan, L. W. Fang, G. D. Guo. Exploring attribute localization and correlation for pedestrian attribute recognition. *Neurocomputing*, vol. 531, pp. 140–150, 2023. DOI: 10.1016/j.neucom.2023.02.019.

[27] W. H. Chen, X. Z. Xu, J. Jia, H. Luo, Y. H. Wang, F. Wang, R. Jin, X. Y. Sun. Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 15050-15061, 2023. DOI: 10.1109/CVPR52729.2023.01445.

[28] S. X. Tang, C. Chen, Q. S. Xie, M. L. Chen, Y. Z. Wang, Y. Z. Ci, L. Bai, F. Zhu, H. Y. Yang, L. Yi, R. Zhao, W. L. Ouyang. HumanBench: Towards general human-centric perception with projector assisted pretraining. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 21970–21982, 2023. DOI: 10.1109/CVPR52729.2023.02104.

[29] Y. Z. Ci, Y. Z. Wang, M. L. Chen, S. X. Tang, L. Bai, F. Zhu, R. Zhao, F. W. Yu, D. L. Qi, W. L. Ouyang. UniHCP: A unified model for human-centric perceptions. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 17840–17852, 2023. DOI: 10.1109/CVPR52729.2023.01711.

[30] D. Li, Z. Zhang, C. F. Shan, L. Wang. Incremental pedestrian attribute recognition via dual uncertainty-aware pseudo-labeling. *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2622–2636, 2023. DOI: 10.1109/TIFS.2023.3268887.

[31] Y. Sun, X. L. Wang, Z. Liu, J. Miller, A. A. Efros, M. Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, Article number 856, 2020. DOI: 10.5555/3524938.3525794.

[32] Y. J. Liu, P. Kothari, B. van Delft, B. Bellot-Gurlet, T. Mordan, A. Alahi. Ttt++: When does self-supervised test-time training fail or thrive? In *Proceedings of the 35th Conference on Neural Information Processing Systems*, pp. 21808–21820, 2021.

[33] S. Schneider, E. Rusak, L. Eck, O. Bringmann, W. Brendel, M. Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, Canada, Article number 968, 2020.

[34] H. Lim, B. Kim, J. Choo, S. Choi. TTN: A domain-shift aware batch normalization in test-time adaptation. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[35] D. Q. Wang, E. Shelhamer, S. T. Liu, B. A. Olshausen, T. Darrell. Tent: Fully test-time adaptation by entropy minimization. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[36] J. Liang, D. P. Hu, J. S. Feng. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6028–6039, 2020. DOI: 10.5555/3524938.3525498.

[37] Y. Iwasawa, Y. Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, pp. 2427–2440, 2021.

[38] M. Jang, S. Y. Chung, H. W. Chung. Test-time adaptation via self-training with nearest neighbor information. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[39] S. Wang, D. A. Zhang, Z. P. Yan, J. G. Zhang, R. Li. Feature alignment and uniformity for test time adaptation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 20050–20060, 2023. DOI: 10.1109/CVPR52729.2023.01920.

[40] C. F. Tang, L. Sheng, Z. X. Zhang, X. L. Hu. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Seoul, Republic of Korea, pp. 4996–5005, 2019. DOI: 10.1109/ICCV.2019.00510.

[41] J. Hu, L. Shen, G. Sun. Squeeze-and-excitation networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 7132–7141, 2018. DOI: 10.1109/CVPR.2018.00745.

[42] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu. Spatial transformer networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 2017–2025, 2015. DOI: 10.5555/2969442.2969465.

[43] X. H. Liu, H. Y. Zhao, M. Q. Tian, L. Sheng, J. Shao, S. Yi, J. J. Yan, X. G. Wang. HydraPlus-Net: Attentive deep features for pedestrian analysis. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 350–359, 2017. DOI: 10.1109/ICCV.2017.46.

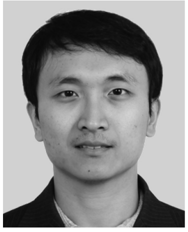
[44] Y. B. Deng, P. Luo, C. C. Loy, X. O. Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, USA, pp. 789–792, 2014. DOI: 10.1145/2647868.2654966.

[45] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Deep residual

learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[46] Z. Liu, H. Z. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, S. N. Xie. A convNet for the 2020s. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, pp. 11966–11976, 2022. DOI: 10.1109/CVPR52688.2022.01167.

[47] B. L. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba. Learning deep features for discriminative localization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 2921–2929, 2016. DOI: 10.1109/CVPR.2016.319.

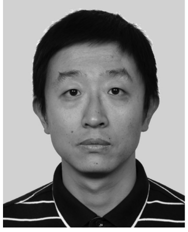


Da Li received the M. Eng. degree in electronics and communication engineering from the Suzhou Institute of Nano-Tech and Nano-Bionics, Chinese Academy of Sciences, China in 2013, and the Ph. D. degree in computer applications technology from the School of Artificial Intelligence, University of Chinese Academy of Sciences, China in 2020. He was a post-doctoral researcher with the New Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China. He is currently an assistant research fellow with the NLPR, CASIA, China.

His research interests include computer vision, big visual data, and video surveillance.

E-mail: da.li@cripac.ia.ac.cn

ORCID ID: 0000-0001-6822-3989

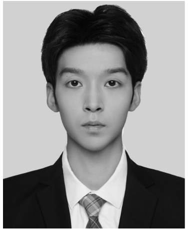


Zhang Zhang received the B. Sc. degree in computer science and technology from the Hebei University of Technology, China in 2002, and the Ph. D. degree in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (CASIA), China in 2009. He is currently an associate professor with the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, CASIA, China. He has published 40 research papers on computer vision and pattern recognition, including some highly ranked journals and conferences, such as the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *IEEE Transactions on Image Processing*, CVPR, and ECCV.

His research interests include action and activity recognition, human attribute recognition, person re-identification, and large-scale person retrieval.

E-mail: zzhang@nlpr.ia.ac.cn (Corresponding author)

ORCID ID: 0000-0001-9425-3065

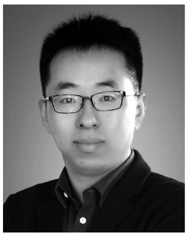


Yifan Zhang received the B. Sc. degree in computer science and technology from South China University of Technology, China in 2021. He is currently a Ph. D. degree candidate with the New Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China. He has published several research papers on computer vision and pattern recognition, including some highly ranked journals and conferences, e.g., IEEE MM, IEEE TIP, CVPR, ICML, NeurIPS, ICLR, and KDD.

His research interests include the development of robust and reliable machine learning (ML) systems that can effectively handle unexpected inputs and distribution shifts.

E-mail: yifanzhang.cs@gmail.com

ORCID ID: 0000-0002-6227-0183



Zhen Jia received the B. Eng. degree in electronic engineering from Yunnan University, China in 2013, and the Ph. D. degree in pattern recognition and intelligent systems from the New Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China in 2020. He is currently a postdoctoral researcher at NLPR, CASIA, China.

His research interests include zero-shot learning, few-shot learning, and generative models.

E-mail: zhen.jia@nlpr.ia.ac.cn

ORCID ID: 0000-0002-6810-2279

Caifeng Shan received the B. Eng. degree in computer science and technology from the University of Science and Technology of China (USTC), China in 2001, the M. Eng. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, China in 2004, and the Ph. D. degree in computer vision from the Queen Mary University of London, UK in 2007. He has authored more than 150 papers and 80 patent applications. He has served as an Associate Editor or the Guest Editor for many scientific journals, including in *IEEE Transactions on Circuits and Systems for Video Technology* and *IEEE Journal of Biomedical and Health Informatics*.

His research interests include computer vision, pattern recognition, image and video analysis, machine learning, bio-medical imaging, and related applications.

E-mail: caifeng.shan@gmail.com

ORCID ID: 0000-0002-2131-1671

MIR | Machine Intelligence Research

Citation: D. Li, Z. Zhang, Y. Zhang, Z. Jia, C. Shan. Adagpar: generalizable pedestrian attribute recognition via test–time adaptation. *Machine Intelligence Research*, vol.22, no.4, pp.783–796, 2025. https://doi.org/10.1007/s11633–024–1504–6