Improved Pedestrian Attribute Recognition Algorithm Based on Image Style Transfer

Senlin Zhang[1], Lingyu Zhao[1], Wenkai Ren[1], Wanwan Wang[2,*], Jiangang Zhang[2]
[1]College of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, Henan, China
[2]Anhui USTC iFLYTEK Co., Ltd., Anhui, Hefei, China
**Corresponding Author*

Abstract: As an important research direction in the field of computer vision, pedestrian attribute recognition has a wide range of value in video surveillance and other applications. However, the existing models have the problem of insufficient generalization ability when testing in new scenes. In this paper, we propose an improved method for pedestrian attribute recognition based on CycleGAN image style transfer. The source domain training data (PA100K) is transformed into the target scene (RAP) style in an unsupervised manner, and an enhanced dataset with both target domain visual features and source domain annotation information is constructed. The experimental results show that the model fused with style transfer data significantly improves the accuracy of attribute recognition on RAP test set, in which gender recognition is improved by 3%, and shirt color recognition is improved by 12%, which verifies the effectiveness of the method in cross-domain adaptation. This study not only avoids the high cost of target scene data labeling, but also provides an effective solution for the scene transfer of pedestrian attribute recognition models.

Keywords: Pedestrian Attribute; Image Style Transfer; CycleGAN; ResNet; Data Augmentation

1. Introduction

With the increasing demand of security services, video surveillance systems are widely used in many fields such as security, criminal investigation and transportation. With the worldwide popularity of video surveillance, it is more and more difficult for manual retrieval to deal with the huge data collected by surveillance cameras [1]. In the analysis, we can only rely on manual screening slowly, but manual monitoring is difficult to concentrate for a long time, and will miss a variety of important security-related events. As the key target object in the surveillance, it is very important and meaningful to use relevant algorithms to analyze pedestrians in real time and identify their appearance attributes, and to focus on tracking and re-identification of special people. However, for a large number of video and image data, the traditional person re-identification method that relies on manual features is no longer applicable.

In recent years, with the wide application of deep learning in the field of computer vision, person attribute re-identification methods based on deep learning can realize automatic analysis in large-scale data processing, which greatly saves time and labor costs. Zheng et al. formulated the person re-identification problem as a multi-classification problem [2], and extracted the global features of pedestrians for retrieval by a two-level fine-tuning strategy on the basis of ImageNet dataset. By decomposing the global features to extract local features, and training each local feature separately, the expression ability of different detailed features is improved [3]. Liu et al. implemented Hydra Plus-Net and proposed a multi-direction attention mechanism module to extract and fuse multi-layer features, which enabled the model to capture attention from shallow to semantic layers and enriched the final feature representation of pedestrian images. Wang et al. proposed the JRL model, which uses a joint recurrent learning method to unify pedestrian attribute correlation and context information in a single model, and constructs an end-to-end codec architecture that can jointly learn image-level context information and attribute-level sequence correlation [4]. Chen et al. proposed SimCLR

(Simple Framework for contrastive Learning of Representations) to enhance the robustness of pedestrian attribute recognition models by contrastive learning. By constructing positive and negative sample pairs, SimCLR encourages the model to output similar features on similar samples and different features on different samples, which can achieve high recognition performance on limited labeled data. Starting from the specific application scenario of PAR, Jia et al. sorted out and summarized the research on PAR in surveillance scenarios, focusing on the deep learning-based PAR method [5]. However, the person re-identification task still faces many problems, among which the key problem is that factors such as illumination, viewpoint change, posture change and human occlusion affect feature extraction, thus affecting the search accuracy. The optimization on the dataset can solve this problem to a certain extent

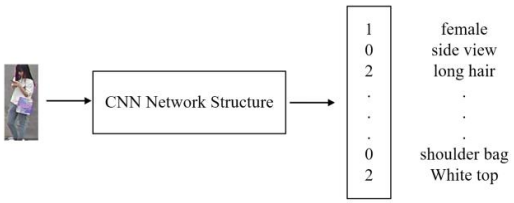

Figure 1. Flowchart of Pedestrian Attribute Recognition

The pedestrian attribute recognition process is shown in Figure 1. Aiming at the problems that the existing pedestrian attribute recognition models have insufficient generalization ability in new scenarios and rely on a large amount of labeled data, which leads to high application costs, this paper innovativly proposes a data augmentation method based on unsupervised style transfer. This method uses the CycleGAN (Cycle Generative Adversarial Networks) image style transfer technology to adaptively convert the image style of the source domain training data to the target scene style, while retaining the original semantic label information [6]. Thus, a large-scale augmented data set with the visual features of the target scene is constructed. This strategy not only effectively solves the performance degradation problem caused by the difference in cross-domain data distribution, but also avoids the expensive cost of target scene data labeling, which provides an effective and feasible solution for improving the recognition accuracy and adaptability of the model in the new environment.

2. Related Work

2.1 Related Work on GAN

In recent years, with the development of deep learning and multi-task learning technology, pedestrian attribute recognition algorithms have made significant progress. However, there are still several key challenges to be solved in this field. The first problem is in cross-domain transfer. Existing models perform poorly when faced with different data sets and scenarios. Secondly, although the research on multi-attribute joint recognition has made certain progress, the existing methods still have shortcomings in the modeling of dependencies between attributes, which directly restricts the further improvement of recognition accuracy.

Since 2017, the introduction of Generative Adversarial Networks (GAN) technology has opened up a new way for image style transfer. In particular, the CycleGAN framework can effectively achieve high-quality style transfer between the source domain and the target domain through its unique cycle consistency constraint mechanism. The algorithm generates new images by combining the content of one image with the style of another image to change the style of an image. CycleGAN is an unsupervised learning method based on generative adversarial networks, which are innovative deep learning models capable of converting images between different styles without pairwise training data. The core idea is to use two generators and two discriminators to train adversarially, and ensure that the content of the image remains consistent through cycle consistency loss. The CycleGAN model inherits the idea of adversarial training of GAN, and realizes the function of generating a mapping between the source domain and the target domain without pairwise relationship in a dual training and learning way, so that CycleGAN can realize the transfer without the need for paired data sets.

The network structure of CycleGAN is composed of two generators (G and F) and two discriminators (Dx and Dy). The generator is responsible for converting the input image from one domain to another, and the generated result is judged to be true or false by the discriminator. This symmetrical architecture

allows converting images back and forth between different domains, and the corresponding network architecture diagram is shown in Figure 2.

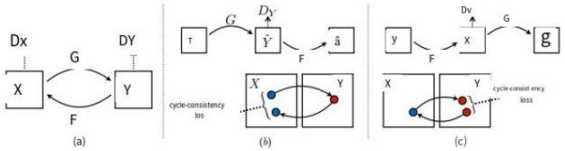

Figure 2. The Network Architecture of CycleGAN

Compared with traditional methods, CycleGAN has the following significant advantages: firstly, the framework does not rely on pairwise training data, which greatly reduces the difficulty of data collection; Secondly, the generated images perform well in terms of visual realism and style consistency. Most importantly, CycleGAN can effectively reduce the distribution difference between domains, and significantly improve the adaptability and generalization performance of the model in new scenes. These characteristics make CycleGAN an effective tool to solve the cross-domain transfer problem in pedestrian attribute recognition, especially in practical application scenarios with scarce data or complex environments.

2.2 Introduction to the Pedestrian Attribute Dataset

With the rapid development of computer vision and deep learning technology, as well as the emergence of massive video image data. Pedestrian attribute recognition has gradually become a research hotspot. In this process, a large number of annotated datasets on pedestrian attributes have emerged for researchers to use.

G. Sharma et al. published HAT dataset, which contains 9344 samples and 27 binary attributes [7]. H. Chen et al. published CAD, which contains 1856 images with 23 binary attributes and 3 multi-class attributes [8]. J. Zhu et al. proposed Apis, a dataset with 3661 images each labeled with 11 binary attributes [9]; Y. Deng et al. proposed PETA, a dataset consisting of 19,000 images, each labeled with 61 binary and 4 multi-class attributes [10]; D. Li et al. release RAP, a large-scale, detailed annotated pedestrian attributes dataset consisting of 41,585 images [11]. The resolution of the images ranges from 200x200 pixels to 800x800 pixels. Lighting conditions are diverse, including natural lighting, indoor lighting, and shadow conditions, among others. Each image is annotated with 69 binary attributes and 3 multi-class attributes. In the same year, Y. Li et al. proposed WIDER, which contains 13,789 images with 14 binary attributes for each person [12]. Liu et al. Published PA100K, a large-scale pedestrian attribute dataset containing 100,000 pedestrian samples [13]. The images of the dataset were collected at different times and places, and the lighting conditions were very diverse, including natural lighting, indoor lighting, and shadow conditions. The images have a wide range of resolutions, ranging from 100x100 pixels to 600x600 pixels. The size of the pedestrian object in the image also varies, usually accounting for a large proportion of the image, and the size of the pedestrian in some images is small. Each image is annotated with 26 binary attributes, including gender, age, hairstyle, clothing color, knapsack information, and so on.

Although there are many datasets for pedestrian attribute recognition, the annotations are not uniform. For example, the PA100K dataset has 26 binary attributes, while the RAP dataset has 69 binary attributes and 3 multi-class attributes. Because the PA100K and RAP datasets are the datasets with large amount of data and relatively complete annotation in many large datasets, the former is for outdoor scenes, and the latter is for indoor scenes, which is more suitable for the task of improving recognition performance in specific scenes by using style transfer in this paper, so this paper chooses PA100K and RAP for experiments. In order to complete the influence of image style transfer on pedestrian attribute recognition. Figure 3 and Figure 4 are the images in PA100K and RAP, respectively.

Figure 3. PA100K Dataset Presentation

Figure 4. RAP Dataset Presentation

3. Experimental Process and Result Analysis

3.1 Environment Introduction

The experiments in this paper are developed and run under the PyTorch framework, the computer operating system is Windows11, the processor model is i7-14675HX, the graphics card is NVIDIA GeForce RTX 4060, the memory is 16GB, the programming environment is Python3.8, and the CUDA version is 12.6.

3.2 Data Set Selection

In this experiment, the PA100K and RAP datasets were utilized for pedestrian attribute recognition research. However, it was observed that the annotation standards and attribute types across these two datasets were inconsistent, and the original annotation categories were insufficient for comprehensive analysis, particularly in fine-grained recognition tasks where detailed attribute discrimination is crucial. To address these limitations, we conducted an in-depth evaluation of multiple existing datasets and manually refined the annotations using LabelMe software, ensuring both high accuracy and standardization across all labeled attributes. The final annotated dataset includes 21 carefully selected attributes (9 binary and 12 multi-class), expanding PA100K into a fine-grained dataset with 106 distinct labels, thereby significantly enhancing its suitability for detailed attribute analysis. Furthermore, we established a unified annotation protocol to maintain consistency across different datasets, which not only improves current research reproducibility but also facilitates future cross-dataset comparisons. The specific attribute categories and their corresponding label distributions are systematically summarized in Table 1 for reference.

As the PA100K dataset provides a large number of pedestrian images, these images not only show a variety of pedestrian poses, but also cover a variety of clothing styles, which contains rich visual information. These images provide our model with a comprehensive sample library of person poses and clothing styles, which is helpful to train the model to recognize various pedestrian attributes. Therefore, in this experiment, 90000 images from PA100K are randomly selected as the training set, called PA100K_9w, and the remaining 10000 images are used as the validation set, called PA100K_1w. The RAP dataset consists of 41,585 images, all of which are used as the test set to verify the recognition effect of the model in new scenes that have not been seen before.

Table 1. an Illustration of the Annotation of Pedestrian Attributes

Attribute	Category	Instruction
gender	2	Male and female
age	6	All ages
orientation	3	Forward, side, back
phone	2	Whether to make a phone call
umbrella	2	Whether to bring an umbrella
children	3	Whether to hold a child
backpack	2	Double backpack if not
shoulder bag	2	Whether it's a single backpack
handbag	2	Whether it is a handbag
items	2	Whether or not to hold an object
hair	7	Hairstyle
hat	3	Hat type
glasses	2	Wear glasses or not
scarf	2	Whether to wear a scarf
type of top	12	Type of top to wear
top color	11	Top color
texture	6	Solid colors, broken flowers, etc
bottom type	7	Wear bottom type
bottom color	11	Bottom color
shoes	8	Leather shoes, etc
gender	11	Shoe color

3.3 Weighted Cross-entropy Loss Function

In the task of pedestrian attribute re-identification, we employed a weighted multi-class cross-entropy loss function to optimize the classification of each attribute. There are a total of 21 attributes, each corresponding to different numbers of classes and class weights. The class weights w_c are used to address the issue of class imbalance, enabling the model to focus more on classes with fewer samples during training, thereby enhancing classification performance.

For each attribute i, the loss function is calculated as follows:

$$L_i = -\sum_{c=1}^{C_i} w_c \cdot y_{ic} \log(\hat{y}_{ic}) \quad (1)$$

In the formula: C_i is the number of classes for the $i-th$ attribute, y_{ic} is the true label for class c of the i-th attribute, and $\hat{y}_{ic}$ is the predicted probability output by the model for class c of the i-th attribute.

The loss function calculates the loss for each class using weighted cross-entropy, where the weight w_c is used to compensate for the imbalance in the number of samples across classes, ensuring that the model performs well across all classes. The total loss of the model is obtained by summing the losses of all 21 attributes:

$$L_{total} = -\sum_{i=1}^{21} L_i \quad (2)$$

The total loss L_{total} summarizes the losses of all 21 attributes and adjusts the contribution of each attribute to the model training based on the class weights of the attributes. This weighted loss function effectively addresses the class imbalance issue, thereby improving the performance of pedestrian attribute re-identification and enhancing the overall classification and generalization capabilities of the model.

3.4 Model Training

In this experiment, PA100K_9w was selected as the training set and PA100K_1w was selected as the validation set. Due to the huge amount of training data, this experiment deployed the model training on the server with 24G video memory, and configured GPU accelerated model training. The main parameter Settings are epoch=200, lr=1e-4, batch_size=100, and workers=4. In this experiment, the ResNet50 network structure is used [14], which effectively solves the gradient disappearance problem of deep networks by introducing skip connections (Figure 5), so that the network can be stably trained to deeper layers.

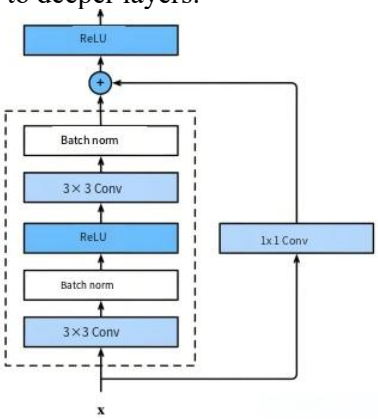

Figure 5. Residual Module Structure

The multi-task classification process based on residual networks is shown in Figure 6. The input pedestrian image is first processed by the initial convolution layer of ResNet50, which uses 7×7 convolution kernels with batch normalization and ReLU activation function, and extracts basic visual features through Max pooling operation. Then, the image features are extracted through four levels of residual modules in turn, and each module is composed of multiple Bottleneck structures. Through the cascading operation of 1×1 convolution dimension reduction, 3×3 convolution feature extraction and 1×1 convolution dimension elevation, feature reuse is realized by combining skip connections, which effectively solves the gradient degradation problem in deep network training.

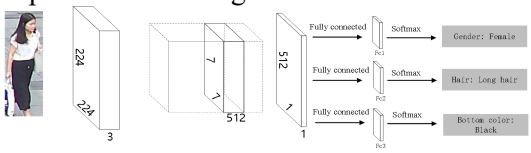

Figure 6. Multi Task Classification Model Based on Residual Network

After training, the best model is selected as baseline_v1 based on the performance on the validation set. The results are shown in Table 2. It can be found that the optimal model trained based on PA100K is not ideal in RAP, that is, the pedestrian attribute recognition algorithm model often does not perform well enough in

the previously unseen target test scene, that is, the generalization ability is not strong enough. The accuracy of baseline_v1 on gender in PK100K_1w is 92.2%, and the accuracy on RAP is 85.5%, which is 6.7% lower. Other attributes also have similar performance. Since RAP does not label texture information, the results are not statistically corresponding to each other.

3.5 Dataset Style Transfer

Aiming at the problem that the baseline_v1 model in Section 3.4 performs poorly on the RAP test set due to domain discrepancy, this study proposes an efficient data augmentation scheme based on unsupervised style transfer to bridge the domain gap. Traditional solutions typically rely on manual labeling and re-training with target scene data, which poses significant challenges in practical applications. Considering that the RAP dataset contains 69 binary attributes and 3 multi-class attributes, with each attribute involving multiple fine-grained categories (e.g., shirt color containing 11 distinct variations), this manual labeling approach is not only prohibitively time-consuming and labor-intensive but also difficult to scale across different real-world scenarios. To overcome these limitations, we innovatively adopt the CycleGAN framework to automatically adapt the style of source domain data (PA100K_9w) to match the visual characteristics of the target domain (RAP), while rigorously preserving the original semantic label information to ensure annotation consistency.

The method employs adversarial training with dual generators and discriminators, combined with cycle consistency loss, to achieve unpaired cross-domain style transfer. This effectively addresses domain shift while preserving attribute integrity, enhancing target domain generalization. As shown in Figure 7, the process successfully generates target-like images while retaining original labels, reducing manual annotation costs and offering a scalable domain adaptation solution for attribute recognition.

3.6 Fusion Training

To enhance the model's learning capability and generalization performance, we combined the original PA100K_9w dataset with the newly annotated PA100K_9w_new data to form an augmented training set containing 180,000 images. This substantial expansion of training samples effectively addresses potential data scarcity issues while providing richer feature representations for model optimization. The validation set remained unchanged as PA100K_1w to ensure consistent evaluation criteria across different experimental phases. For rigorous experimental control, all training hyperparameters and network architectures were maintained identical to baseline_v1, with the only variable being the expanded training data scale. This standardized approach allows for direct comparison of model improvements attributable solely to the enhanced dataset. After comprehensive training, baseline_v2 was selected as the optimal model based on its superior validation set performance, demonstrating the effectiveness of data augmentation in boosting model accuracy.

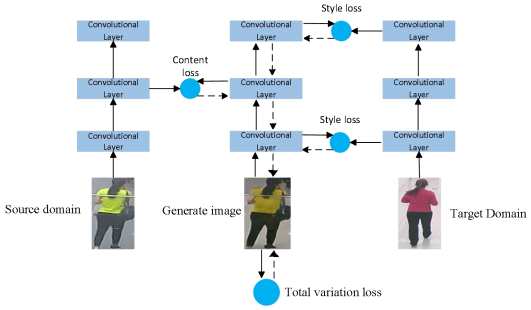

Figure 7. Pedestrian Attribute Recognition Style Transfer Process Diagram

3.7 Performance Testing

In the testing phase, baseline_v1 and baseline_v2 were used to test on RAP, and the accuracy of each attribute was calculated. The experimental results show that the CycleGAN-based style transfer method effectively improves the recognition performance of the model in new scenarios. The accuracy of baseline_v2 on the RAP test set is significantly improved compared with baseline_v1. The accuracy of gender recognition is increased by 3 percentage points, age recognition is increased by 1.2%, top color recognition is increased by 12%, and bottom color recognition is increased by 3.7%. It verifies the effectiveness of the proposed method in alleviating the cross-domain data distribution differences. By simulating the visual features of the target scene (such as indoor lighting conditions and pedestrian pose distribution), the style transfer data enhances the adaptability

of the model to the target domain features, especially in the key attributes such as clothing texture and light reflection. At the same time, this method avoids the expensive annotation cost of the target domain. Although the absolute improvement of such attributes is small (1.5%-1.8%), it shows that style transfer can effectively alleviate the problem of local feature distribution shift, which provides an efficient and low-cost solution for the deployment of models in new scenes in practical applications.

Table 2 Accuracy on RAP before and after Data Fusion

Attribute	baseline_v1		baseline_v2
	PA100K_1w	RAP	RAP
gender	0.922	0.855	0.885
age	0.728	0.575	0.587
orientation	0.883	0.824	0.839
phone	0.910	0.950	0.962
umbrella	0.996	0.998	0.997
children	0.995	0.999	0.999
backpack	0.944	0.968	0.967
shoulder bag	0.890	0.909	0.917
handbag	0.906	0.816	0.851
items	0.999	0.999	0.999
hair	0.894	0.818	0.849
hat	0.984	0.977	0.978
glasses	0.913	0.890	0.902
scarf	0.996	0.988	0.989
type of top	0.783	0.333	0.342
top color	0.823	0.563	0.683
texture	0.919	0.941	0.952
bottom type	0.889	0.587	0.624
bottom color	0.721	0.289	0.304
shoes	0.775	0.728	0.746

4. Closing Remarks

Aiming at the problem that the performance of pedestrian attribute recognition model decreases in new scenarios, this paper innovativelyintroduces CycleGAN style transfer technology into the data augmentation process. Through experimental verification, this method effectively improves the attribute recognition accuracy of the model in the target scene, especially in the key attributes such as color and gender. The research breaks through the dependence of traditional methods on the labeled data of the target domain, and provides a feasible technical path for cross-scene deployment in practical applications. Future work can further explore the collaborative optimization of style transfer and fine-grained feature recognition, and expand the application potential of this method in other vision tasks. This study provides an effective reference for improving the practicability and adaptability of pedestrian attribute recognition models.

References

[1] Cao Yuran, Lu Weiqing, Yu Jinzuo, et al. Review and Prospect of Pedestrian Attribute Recognition Methods Based on Single Frame and Video Data in Surveillance Scenarios. Journal of Computer-Aided Design & Computer Graphics, 2024, 36(03): 336-356.

[2] Zheng L, Zhang H, Sun S, et al. Person re-identification in thewild//Proceedings of the IEEE conference on computer vision andpattern recognition,2017:1367-1376.

[3] Liu X, Zhao H, Tian M, et al. Hydraplus-net: Attentive deep features forpedestrian analysis//Proceedings of the IEEE international conferenceon computer vision,2017:350-359.

[4] Wang J, Zhu X, Gong S, et al. Attribute Recognition by Joint Recurrent Learning of Context and Correlation//2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 531-540.

[5] Jian Jia, Xiao-Tang Chen & Kai-qi Huang. (2022). Pedestrian attribute recognition in surveillance scenes: a survey. Acta Computerica Sinica,45(08),1765-1793.

[6] Xu C, Li W, Chen Z B. Scarcity-GAN: Scarce data augmentation for defect detection via generative adversarial nets. Neurocomputing, 2024, 566 (Jan. 21): 127061.1 to 127061.12.

[7] Sharma, G., & Jurie, F. (2011). Learning discriminative spatial representation for image classification. Procedings of the British Machine Vision Conference 2011, 6.1-6.11.

[8] CHEN H, GALLAGHER A, GIROD B. Describing Clothing by Semantic Attributes//Computer Vision – ECCV 2012, Lecture Notes in Computer Science. 2012: 609-623.

[9] Zhu J, Liao S, Yi D, et al. Multi-label CNN based pedestrian attribute learning for soft biometrics//International Conference on Biometrics. IEEE,

2015:535-540.

[10]Deng Y, Luo P, Loy C C, et al. Pedestrian Attribute Recognition at Far Distance. ACM, 2014, 22(ACM MM '14):789-792.

[11]Li D, Zhang Z, Chen X, et al. A Richly Annotated Dataset for Pedestrian Attribute Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(10): 2033-2047.

[12]Li Y, Huang C, Loy C C, et al. Human Attribute Recognition by Deep Hierarchical Contexts. Lecture Notes in Computer Science, 2016, 9907: 480-495.

[13]Liu X, Zhao H, Tian M, et al. HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis//2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 5078-5087.

[14]He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778.

