Pedestrian Attribute Recognition Method Based on Multi-source Teacher Model Fusion

Zhengyan Ding(✉) and Yanfeng Shang

Research Center on Internet of Things, The Third Research Institute of the Ministry of Public
Security, Shanghai 201204, China
dzy_wlw@163.com

Abstract. Existing pedestrian attribute recognition (PAR) methods usually use deep models to achieve attribute classification. However, the attribute recognition models trained on the public datasets have poor generalization ability and cannot be applied to complex scenarios. Also, through the traditional multi-label classification framework and a single network model, it is difficult for different attribute features to be effectively represented and fused. Aiming at the above problems, this paper proposes a novel knowledge distillation framework based on the fusion of multi-source prior teacher models. Focusing on the diversity of datasets, architectures and knowledge, different model training schemes are designed. For ensuring the diversity and accuracy of teacher models, this paper selects models through adaptive scoring mechanism and finally adopts active learning mechanism to achieve closed-loop model optimization. Tested on four common PAR benchmark datasets, experimental results show that under the condition that the complexity of the model is unchanged, the mean accuracy is improved by 2% to 5%, compared with the baseline model.

Keywords: Pedestrian attribute recognition · Multi-source · Knowledge distillation · Active learning

1 Introduction

Pedestrian attribute recognition (PAR) is always very important in the field of intelligent video analysis. By identifying the visual attributes of the pedestrian target, including semantic information such as gender, age, clothing, and so on, it is possible to provide a structured description and rapid retrieval of the specific target.

With the continuous development of deep learning technology, especially the widespread application of convolutional neural network models for image classification, researchers also proposed the PAR method based on deep network models [1] and made many improvements as follows. Zeng et al. [2] proposed a collaborative attention sharing mechanism for pedestrian multi-attribute recognition, which is different from the traditional feature linear fusion module to realize the adaptive selection of feature channels and spatial regions. Moghaddam et al. [3] further combined with semantic information to analyze the location of different parts of the target, reducing the interference of unrelated component features on specific attribute recognition tasks, and

G. Zhai et al. (Eds.): IFTC 2022, CCIS 1766, pp. 16–29, 2023.
https://doi.org/10.1007/978-981-99-0856-1_2

using a lightweight backbone network to achieve the improvement of PAR efficiency. In addition, by collecting pedestrian target data in different scenarios and labeling relevant attribute information, the researchers built multiple large-scale PAR datasets, such as PA100k [4], PETA [5], RAPv1 [6] and RAPv2 [7], etc. For example, the RAP-V2 dataset contains 54 attributes of pedestrian targets, covering various global and local attributes. Global attributes mainly include gender, age, etc. and local attributes mainly include head, upper body and lower body attributes.

However, due to the inconsistent region of interest (ROI) between different attribute recognition tasks, it is difficult to effectively and comprehensively characterize the features of various attributes through a single model using traditional multi-label classification framework [8], and the fusion of multiple classification models will lead to a significant increase in computational complexity [2]. On the other hand, due to the complexity and diversity of practical application scenarios, the PAR model trained by the above public dataset has poor generalization ability in actual scenarios, especially for some difficult samples, such as occlusion, truncation, blur, etc. The existing feature optimization method usually adopts a model structure similar to the feature pyramid [9], which will lead to a significant increase in model complexity compared with the original model, and since that no prior knowledge except for scale information is introduced, the feature representation ability of the PAR model still needs to be further improved. Especially for lightweight models, the recognition accuracy is reduced more, resulting in the inability to meet the application requirements.

Focused on the above problems, this paper proposes a novel PAR method based on the fusion of multi-source teacher models, for actual tasks under the video surveillance scenario:

1) For the diversity of datasets, this paper uses sample data of different scenario categories and statistical distributions, to train the teacher model that fits multi-source data.
2) For the diversity of architectures, this paper adopts model architectures of different backbone networks and training tasks, to train the teacher model with multi-source feature representation.
3) For the diversity of knowledge, this paper introduces prior knowledge from metric learning and self-supervised learning respectively, to train the teacher model translated from multi-source knowledge.

Through the knowledge distillation framework, the above teacher models are fused by adaptive scoring mechanism, which guarantees both the diversity and accuracy in the meantime. Then, using the active learning framework, the pseudo-label of massive unlabeled data is generated, in which only a small number of uncertain samples need manual correction and the teacher model is iterated with only a small annotation cost, to realize the closed-loop optimization mechanism of the model in actual surveillance application.

2 Baseline Model

2.1 Datasets

For baseline model, this paper uses the public PAR datasets for training and testing, as shown in Fig. 1, covering various global and local attributes.

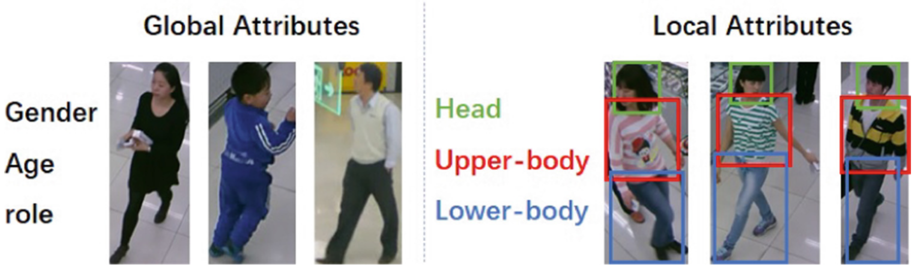

Fig. 1. Global attributes and local attributes in RAP-V2 dataset

2.2 Architecture

The multi-label classification framework is selected as baseline model architecture, in which the backbone network is ResNet50 and the specific training strategy and hyperparameter settings refer to [8].

2.3 Knowledge

In the training phase of baseline model, knowledge from attribute label information is acquired using the traditional supervised learning methods, which depend on dataset quality seriously.

3 Multi-source Datasets

3.1 Datasets of Different Scenario Categories

Considering that datasets of different scenario categories have obvious difference in lighting conditions, shooting angles, clarity, etc., this paper divides the existing training data into three different scenario categories (indoor, outdoor near and outdoor far), and integrates the scene type information as prior information into training different models, denoted as *Teacher_1_X*. The training process is shown in Fig. 2, and finally based on the fusion results of three models, the 1st teacher model is obtained through the knowledge distillation framework [10], which is recorded as *Teacher_1*.

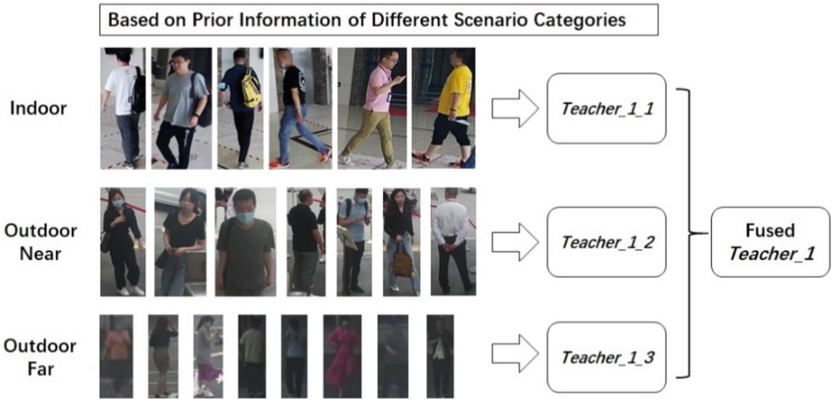

Fig. 2. Fusing multi-source teacher models based on different scenario categories

3.2 Datasets of Different Statistical Distributions

In the actual application, the PAR dataset has the problem of uneven statistical distribution, and different sampling process according to the label information for a specific attribute will lead to sample data with different statistical distributions, resulting in obvious differences in training results. Therefore, this paper takes the statistical distribution of sample data as a prior information, so as to obtain multi-source teacher models, denoted as *Teacher_2_X*. The training process is shown in Fig. 3, taking the two-categories attribute as example, three different sample distributions can be designed. Finally, the knowledge distillation framework is used for model fusion, and the 2nd teacher model is obtained, which is recorded as *Teacher_2*.

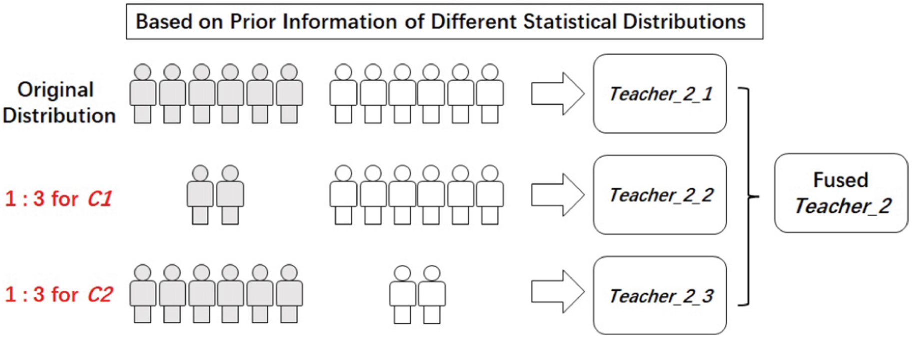

Fig. 3. Fusing multi-source teacher models based on different statistical distributions

4 Multi-source Architectures

4.1 Architectures of Different Backbone Networks

At present, the mainstream CNN networks mainly include ResNet series models [11] and Inception series models [12]. This paper selects the ResNet50 model as the baseline backbone network, and accordingly selects the InceptionV4 model as the second backbone network. In order to improve the difference of model structure, this paper further selects the ResNet50 model as the third backbone network. ResNet series models [13] are different from the traditional convolutional neural network, using Involution to replace the convolution operation. The self-attention mechanism similar to the Vision-Transformer structure [14] is integrated into the learning of visual features, so as to obtain a new type of efficient backbone network. As shown in Fig. 4, by using the model structure of three different backbone networks, a variety of teacher models are trained for heterogeneous feature representation, denoted as *Teacher_3_X*. Finally, the knowledge distillation framework is used for fusion to obtain the 3rd type of teacher model, which is recorded as *Teacher_3*.

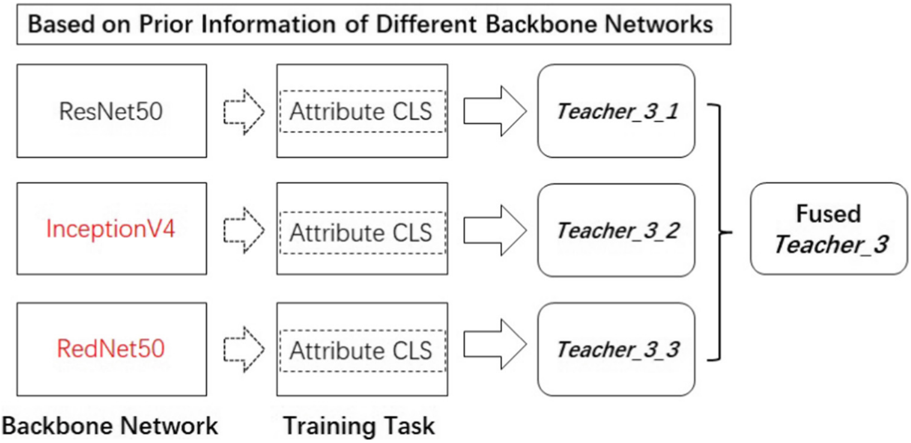

Fig. 4. Fusing multi-source teacher models based on different backbone networks

4.2 Architectures of Different Training Tasks

The PAR task is related to pedestrian key-point detection (denoted as *Extra_Task_1*) and pedestrian part segmentation (denoted as *Extra_Task_2*):

1) The visual features of PAR task rely on the spatial attention mechanism. For example, the vital features of upper body attribute are mainly located in the upper body of pedestrians, and the position information of the upper body can be judged exactly through *Extra_Task_1* and *Extra_Task_2*, thereby improving the relevant attribute recognition results, as shown in Fig. 5.

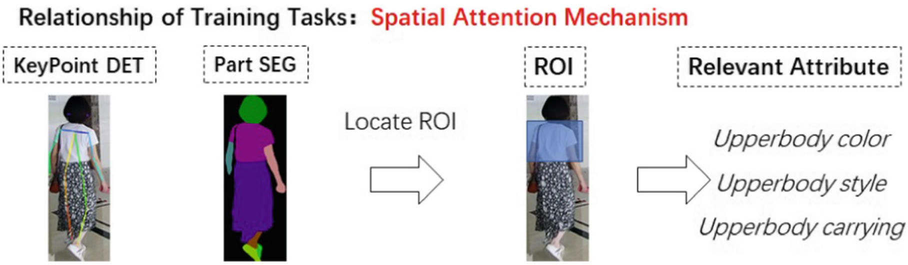

Fig. 5. Relationship of training tasks based on spatial attention mechanism

2) The semantic features of PAR task rely on the spatial relation information between multiple local regions of the human body. For example, judging the lift-hand action through the spatial relation between the wrist and the shoulder, and combining the information of the carrying thing at the wrist position, the call-action attribute will be accurately identified, as shown in Fig. 6.

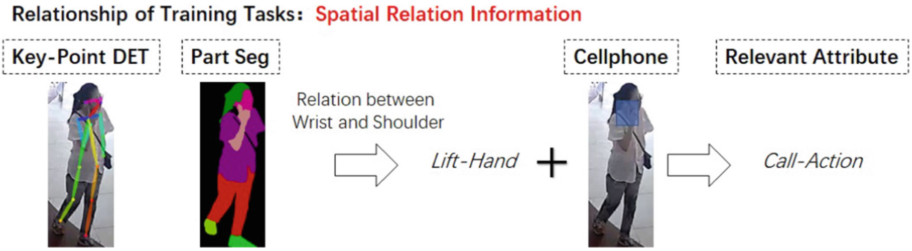

Fig. 6. Relationship of training tasks based on spatial relation information

Therefore, on the basis of the baseline model architecture (only attribute classification task), this paper further combines the two related tasks (key-point detection and part segmentation). As shown in Fig. 7, by using the combination of three different training tasks, a variety of models are trained for heterogeneous feature representation, denoted as *Teacher_4_X*. Finally, the knowledge distillation framework is used for fusion to obtain the 4th type of teacher model, which is recorded as *Teacher_4*.

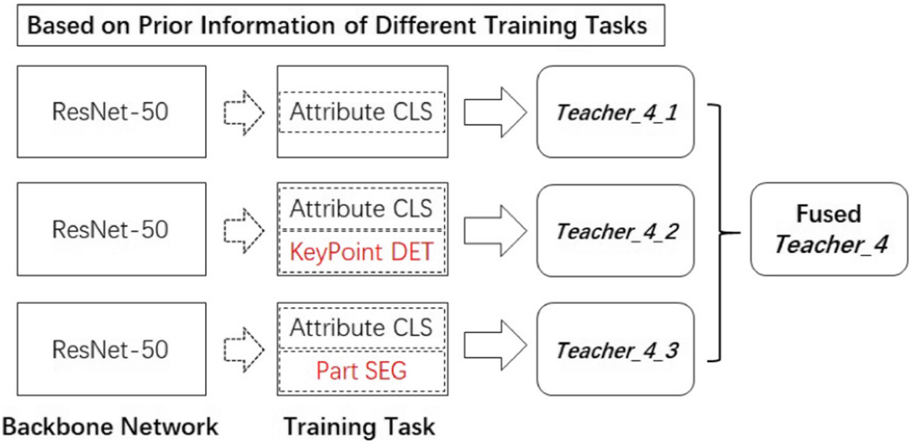

Fig. 7. Fusing multi-source teacher models based on different training tasks

5 Multi-source Knowledge

5.1 Knowledge from Metric Learning

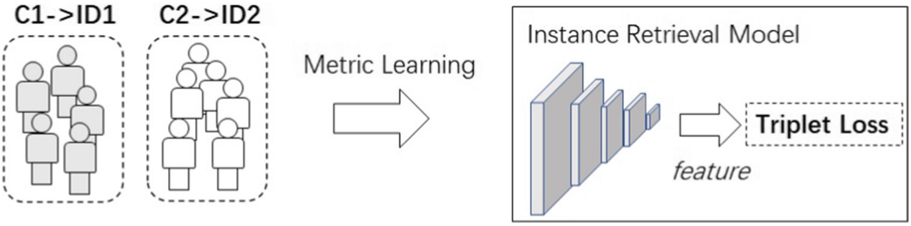

Fig. 8. Training Instance Retrieval Model Based on Metric Learning

As shown in Fig. 8, the original attribute categories are regarded as pedestrian ID information, and triplet loss is introduced to train instance retrieval model and calculate the feature similarity between pedestrian target images based on metric learning.

For actual application, the results of attribute classification model are usually limited by category knowledge from the labeled data. This paper uses the instance retrieval model to assist in mining knowledge from unlabeled data, as shown in the Fig. 9. For the massive unlabeled data, we select unsure samples of the classification model, which confidence is below the threshold (τ is 0.7), as query data, and the labeled data is used as gallery dataset. Then we obtain the pseudo-label information by voting of the top-50 results, according to the feature similarity. Finally, the data with pseudo-label information is fused with the original data to train the 5th teacher model, introducing prior knowledge from metric learning, which is recorded as *Teacher_5*.

5.2 Knowledge from Self-supervised Learning

This paper uses the Masked AutoEncoder (MAE) model [15] for data generation, as shown in Fig. 10. Firstly, we mask some regions in the original image randomly according

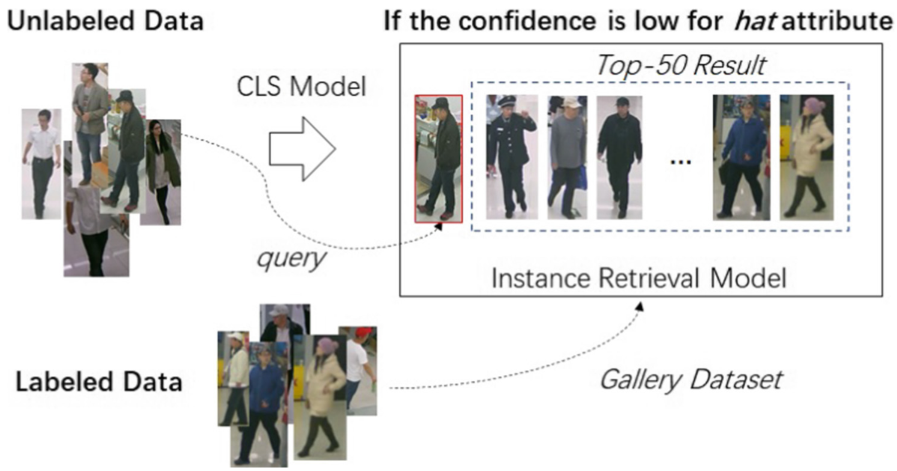

Fig. 9. Generate pseudo-label of unsure data by instance retrieval model

to a certain proportion, and then restore the image through an asymmetric encoder-decoder structure, in which the encoder module adopts the deep network model based on the Transformer structure [14] for feature coding, and the decoder module adopts a lightweight model. For the newly generated sample data, the attribute recognition model is used for classification, and only samples with consistent labels are retained, so that the key features related to the specific attribute are retained. The MAE model is trained by self-supervised learning on massive unlabeled data, so it can effectively achieve general feature representation of pedestrian targets. Finally, the newly generated data is fused with the original data, to train the 6th teacher model, introducing prior knowledge from self-supervised learning, which is recorded as *Teacher_6*.

6 Fusion of Multi-source Teacher Models

6.1 Scoring Mechanism for Model Accuracy

This paper refers to the mainstream evaluation methods for PAR task [8], and combines two metrics to score the model accuracy, as follows:

1) For attribute: mean accuracy of all attributes, denoted as *mA*.
2) For sample: F1 score of all samples, denoted as *F1_score*, representing the harmonic average of the mean accuracy and the mean recall.

In this paper, the *mA* and *F1_score* of the baseline model is used as reference value to select the teacher models. If the *mA* decreases by more than 10% or the *F1_score* decreases by more than 0.05, it is supposed that the teacher model does not meet the fusion requirements for model accuracy.

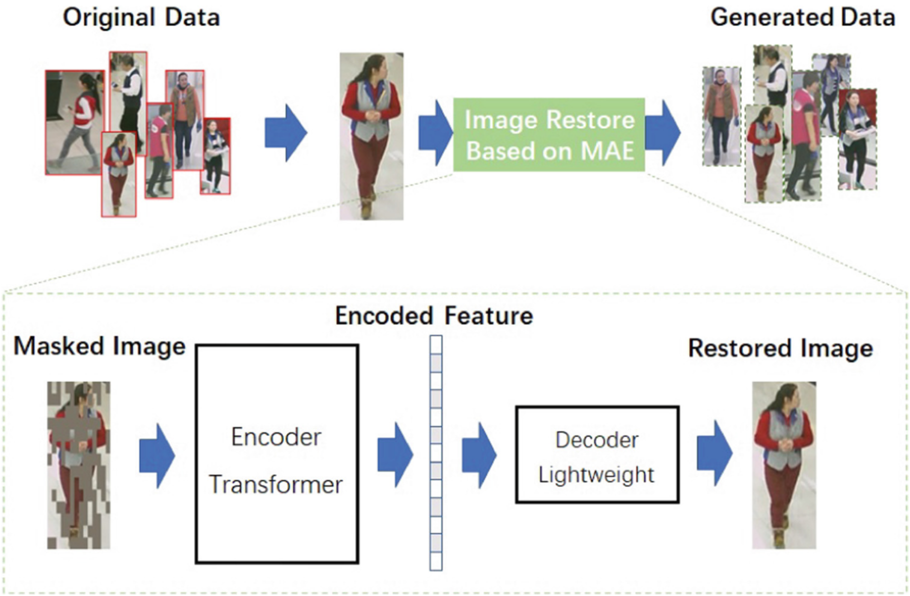

Fig. 10. Generate restored data by MAE model

6.2 Scoring Mechanism for Model Diversity

This paper refers to the mainstream evaluation methods for calculating distribution difference between model outputs [16], and uses JS divergence to score the model diversity. Compared with the baseline model, the teacher models are evaluated on the validation dataset, and the mean JS for all samples and all categories is taken as the metric for model diversity, denoted as *mJS*.

6.3 Adaptive Fusion Mechanism

Considering the accuracy and diversity requirements of the teacher model, this paper implements an adaptive fusion mechanism for multi-source teacher models, and the detailed pipeline is as follows:

Step 1: In the training phase for teacher models, we evaluate each model of different iteration cycles (the value of *epoch* varies from 1 to *max_epoch*), and select the model that meets the accuracy requirements based on *mA* and *F1_score* (see Sect. 6.1). Then the candidate model group is constructed, including all *Teacher_i* models that meet the accuracy requirements, denoted as *Group_i*. The value of *i* varies from 1 to 6, corresponding to the type number of multi-source teacher models.

Step 2: For the six different candidate model groups {*Group_i*}, the teacher models of each group are compared with the baseline model, and the model with the largest diversity is selected, that is, the value of *mJS* is the largest. Then the selected model is denoted as *BestTeacher_i*.

Step 3: For the six different best teacher models {*BestTeacher_i*}, the knowledge distillation framework is used for training, to obtain an adaptive fusion of multi-source teacher models.

6.4 Iteration Based on Active Learning

However, based on the fusion mechanism of multi-source teacher models proposed above, the optimized PAR model still has low accuracy in practical application scenarios, and data optimization needs to be realized, in which massive pseudo-label samples are supplemented. In order to reduce the labeling cost, this paper adopts an iterative optimization mechanism based on active learning, and selects the most informative samples to be labeled manually, so that the difficult samples are more focused on and the iteration efficiency is improved significantly. In this paper, a probabilistic model [17] is used to estimate the probability distribution of the output results of the PAR model and the uncertainty of the sample is calculated by combining the scoring function.

The whole iterative framework is shown in Fig. 11, only a small amount of manually corrected pseudo-label data is supplemented to achieve closed-loop optimization, from model to data.

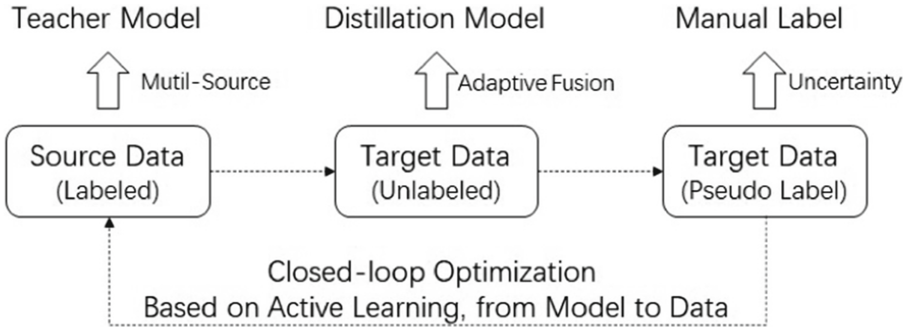

Fig. 11. The whole iterative framework based on active learning

7 Experiments

7.1 Experimental Settings

We conduct experiments on four common PAR benchmark datasets, PA100k [4], PETA [5], RAPv1 [6], RAPv2 [7]. The proposed methods are implemented with PyTorch, referring to the training parameters and evaluation metrics in [8]. As shown in Table 1, the settings of each benchmark dataset are consistent with [8].

Table 1. The settings of each benchmark dataset

Dataset	Attributes	Training Images	Testing Images
PA100k [4]	26	90,000	10,000
PETA [5]	35	11,400	7,600
RAPv1 [6]	51	33,268	8,317
PAPv2 [7]	54	67,943	16,985

7.2 Experimental Results

In the experiments, the metrics for attributes (mA) and samples (Precision, Recall, F1 score) are calculated, as shown in Tables 2, 3, 4 and 5. Compared with the existing PAR methods, our proposed method achieves a better performance, which outperforms the baseline model by 4.59%, 4.23%, 4.52%, 2.47% for mA in PA100k, PETA, RAPv1 and PARv2 respectively, under the condition that the complexity of the model is unchanged.

Table 2. Recognition results of different algorithms on the PA100k dataset (%)

Method	Backbone	mA	Precision	Recall	F1 score
DeepMAR [1]	CaffeNet	72.70	82.24	80.42	82.32
RPAR [8]	ResNet50	79.38	89.41	84.78	86.55
Baseline	ResNet50	79.00	89.17	84.71	86.49
Ours	**ResNet50**	**83.59**	90.58	89.41	**89.65**

Table 3. Recognition results of different algorithms on the PETA dataset (%)

Method	Backbone	mA	Precision	Recall	F1 score
DeepMAR [1]	CaffeNet	82.89	83.68	83.14	83.41
RPAR [8]	ResNet50	85.11	86.99	86.33	86.39
Baseline	ResNet50	83.10	90.72	82.40	85.77
Ours	**ResNet50**	**87.33**	90.01	88.52	**89.02**

Taking the RAPv2 dataset as an example, the attributes with the top-10 accuracy gain are shown in Fig. 12, and it is demonstrated that the accuracy of each attribute is improved by more than 10%, compared with baseline model.

Table 4. Recognition results of different algorithms on the RAPv1 dataset (%)

Method	Backbone	mA	Precision	Recall	F1 score
DeepMAR [1]	CaffeNet	73.79	74.92	76.21	75.56
RPAR [8]	ResNet50	78.48	82.84	76.25	78.94
Baseline	ResNet50	77.26	83.08	71.78	76.36
Ours	**ResNet50**	**81.78**	81.84	81.33	**81.53**

Table 5. Recognition results of different algorithms on the RAPv2 dataset (%)

Method	Backbone	mA	Precision	Recall	F1 score
RPAR [8]	ResNet50	78.28	77.96	79.38	78.30
Baseline	ResNet50	76.83	81.50	77.98	79.35
Ours	**ResNet50**	**79.30**	79.99	79.85	**79.91**

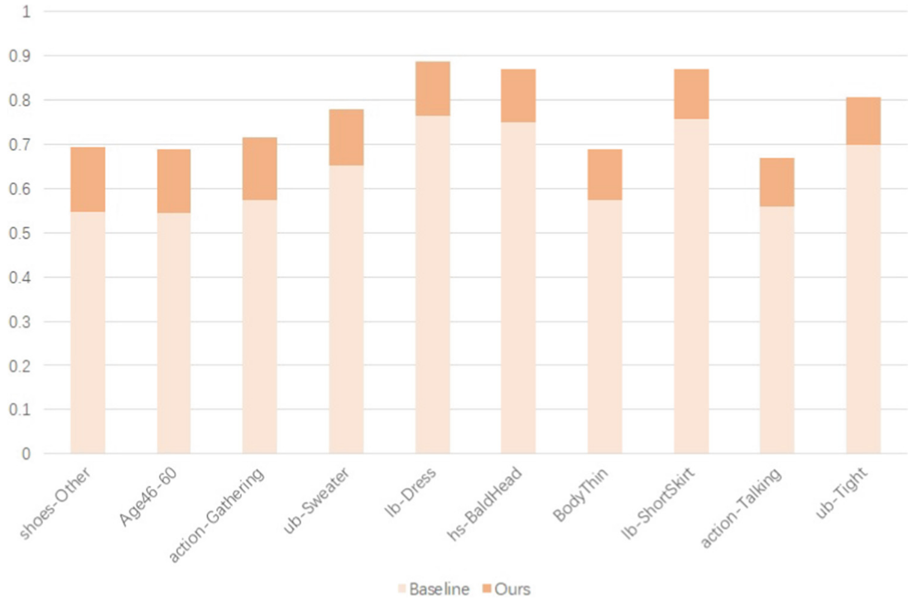

Fig. 12. The top-10 attributes for accuracy gain in the RAPv2 dataset

8 Conclusion

Focusing on the PAR task, this paper realizes the closed-loop optimization of the model by adaptive fusion of the multi-source teacher models and active learning mechanism, so as to effectively improve the feature representation ability of the recognition model.

Compared with the traditional model fusion method, this paper takes into account multi-source datasets, architectures and knowledge, which enhances the interpretability of model fusion. In the next step, this paper will further improve the diversity of model architectures, such as designing new model structures through AutoML methods.

Acknowledgements. This paper is sponsored by Shanghai Sailing Program 20YF1409300.

References

1. Li, D., Chen, X., Huang, K.: Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In: Proceedings of the 2015 Asian Conference on Pattern Recognition. Piscataway: IEEE, pp. 111–115 (2015)
2. Zeng, H., Ai, H., Zhuang, Z., et al.: Multi-task learning via co-attentive sharing for pedestrian attribute recognition. In: 2020 IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp. 1–6 (2020)
3. Moghaddam, M., Charmi, M., Hassanpoor, H.: Jointly human semantic parsing and attribute recognition with feature pyramid structure in EfficientNets. IET Image Processing (2021)
4. Liu, X., Zhao, H., Tian, M., et al.: Hydraplus-net: Attentive deep features for pedestrian analysis. In: Proceedings of the IEEE international conference on computer vision, pp. 350–359 (2017)
5. Deng, Y., Luo, P., Loy, C.C., et al.: Pedestrian attribute recognition at far distance. In: Proceedings of the 22nd ACM international conference on Multimedia, pp. 789–792 (2014)
6. Li, D., Zhang, Z., Chen, X., et al.: A richly annotated dataset for pedestrian attribute recognition (2016). arXiv preprint arXiv:1603.07054
7. Li, D., Zhang, Z., Chen, X., et al.: A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. IEEE transactions on image processing **28**(4), 1575–1590 (2018)
8. Jia, J., Huang, H., Yang, W., et al.: Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method (2020). arXiv preprint arXiv:2005.11909
9. Lin, T.Y., Dollár, P., Girshick, R., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117–2125 (2017)
10. Bagherinezhad, H., Horton, M., Rastegari, M., et al.: Label refinery: Improving imagenet classification through label progression (2018). arXiv preprint arXiv:1805.02641
11. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
12. Szegedy, C., Vanhoucke, V., Ioffe, S., et al.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2818–2826 (2016)
13. Li, D., Hu, J., Wang, C., et al.: Involution: Inverting the inherence of convolution for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12321–12330 (2021)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale (2020). arXiv preprint arXiv:2010.11929
15. He, K., Chen, X., Xie, S., et al.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16000–16009 (2022).

16. Cui, C., Guo, R., Du, Y., et al.: Beyond Self-Supervision: A Simple Yet Effective Network Distillation Alternative to Improve Backbones (2021). arXiv preprint arXiv:2103.05959
17. Choi, J., Elezi, I., Lee, H.J., et al.: Active learning for deep object detection via probabilistic modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10264–10273 (2021)