Research on Multi-scale Pedestrian Attribute Recognition Based on Dual Self-attention Mechanism

He Xiao[1,2(✉)], Wenbiao Xie[1], Yang Zhou[3], Yong Luo[4], Ruoni Zhang[1], and Xiao Xu[1]

[1] School of Software Engineering, Jiangxi University of Science and Technology, NanChang 330013, Jiangxi, People's Republic of China
xiaohe804@gmail.com, vebrun@jxust.edu.cn
[2] Nanchang Key laboratory of Virtual Digital Factory and Cultural Communications, Nanchang 330013, People's Republic of China
[3] Information and Communication Branch, State Grid Jiangxi Electric Power Co, Nanchang 330095, China
[4] School of Software, Jiangxi Normal University, Nanchang 330022, China
luoyong1020@jxnu.edu.cn

Abstract. As one of the important fields of computer vision research, pedestrian attribute recognition has gained increasing attention from domestic and foreign researchers due to its huge potential applications. However, obtaining long-distance pedestrian information in actual scenes poses challenges such as lack of information, incomplete feature extraction, and low attribute recognition accuracy. To address these issues, we propose a multi-scale feature fusion network based on a dual self-attention mechanism. The fusion module merges multi-scale features to enable more complete attribute extraction, while the dual self-attention module focuses the network on important regions. Experimental results on PA-100K, RAP, and PETA datasets achieved mean accuracies of 81.97%, 81.53%, and 86.37%, respectively. Extensive experiments demonstrate that the proposed method is highly competitive in pedestrian attribute recognition.

Keywords: incomplete feature · dual self-attention · multi-scale fusion · pedestrian attribute recognition

1 Introduction

As computer vision technology continues to evolve, pedestrian attribute recognition has emerged as a vital research focus, garnering extensive interest. Pedestrian attribute recognition refers to the recognition and classification of a range of pedestrian attributes in images (such as clothing, hairstyle, gender, age, etc.),

The Jiangxi Province Office of Education provided funding support for this research.

C. Wu et al. (Eds.): MONAMI 2023, LNICST 559, pp. 215–226, 2024.
https://doi.org/10.1007/978-3-031-55471-1_16

which enables quick recognition and judgment of pedestrian identities [1–4]. The technology of pedestrian attribute recognition holds extensive applicability in areas like intelligent monitoring, security detection, and intelligent transportation. Its deployment carries substantial importance for enhancing societal security and safeguarding individuals' lives and assets [5,6].

Over recent years, as deep learning technology has steadily evolved, models for pedestrian attribute recognition based on deep neural networks have made significant advancements. In the realm of existing research, some academics have put forward pioneering modules like multi-scale feature fusion [7,8] and self-attention mechanism [9,10], effectively enhancing the accuracy and stability of pedestrian attribute recognition. However, existing pedestrian attribute recognition models still have some problems, such as poor attribute recognition performance in complex scenes and insufficient processing of the mutual influence relationship between different attributes.

Pedestrian attribute recognition stands as a critical research focus within the computer vision domain and carries substantial importance across a multitude of applications. However, it also faces several challenges, including**viewpoint variations, illumination variations, occlusions, imbalanced attribute distributions, low resolution, and blurriness** [11].

Viewpoint variations occur due to different angles and distances between the shooting device and pedestrians, which cause changes in the viewpoint and size of pedestrians in the image. This can affect the recognition of pedestrian attributes. Illumination variations occur due to different lighting conditions that cause changes in the brightness, contrast, color, and other aspects of pedestrian images, making the recognition of pedestrian attributes more difficult. Occlusions occur when other objects such as pedestrians, vehicles, and buildings obstruct parts of the pedestrian's body or clothing, which can affect the recognition of pedestrian attributes. Imbalanced attribute distributions occur because the distribution of pedestrians with different attributes in the dataset is uneven, which can lead to poor recognition performance for certain attributes. Low resolution occurs owing to the substandard resolution of pedestrian images, making it difficult to capture the details of pedestrian attributes and thus affecting the recognition of pedestrian attributes. Blurriness occurs due to image blur or motion blur, making the recognition of pedestrian attributes more difficult.

2 Related Work

Global-based models. Li [13] et al. put forward two significant models: DeepSAR and DeepMAR. The former, DeepSAR, operates without considering any correlation between attributes while the latter, DeepMAR, is built specifically to consider such correlations. Recognizing a prevalent issue in the field, Li et al. also made strides in addressing the problem of imbalanced attribute distribution. They did this by proposing a novel loss function, a solution that is designed to specifically tackle this imbalance. In an innovative approach, Sudowe [12] et al. incorporated AlexNet, a pre-existing, influential model, as the backbone for

feature extraction in their global-based model. This model stands out due to its unique multi-branch classification layer designed for each attribute. Abdulnabi [14] et al. made a noteworthy contribution to the field by proposing a joint multi-task learning algorithm. This algorithm, used for attribute estimation, is based on Convolutional Neural Networks (CNN), a popular and effective tool in the realm of machine learning.

Part-based models. Tang [7] et al. developed a complex yet efficient approach to attribute analysis. They fused deep semantic features and low-level detail features into the attribute localization module and used an element-by-element voting mechanism to select the optimal value, ensuring a comprehensive and accurate process. Diba [17] et al. proposed an inventive CNN which is specifically engineered for mining mid-level image information. This tool is especially useful for fine-grained pedestrian attribute recognition, enhancing the detail and accuracy of such analyses. In a progressive move, YangDiba [16] et al. employed an end-to-end learning framework. This framework is uniquely tailored for the joint part localization and multi-label classification, thus serving dual purposes in an integrated manner. Li [18] et al. implemented a pose estimation algorithm with a specific purpose in mind - to accurately locate local bodies. This application of the algorithm demonstrates its utility in precise localization tasks. Zhu [15] et al. brought forth a novel approach in the form of a multi-label convolutional neural network. This innovative network is designed to jointly predict multiple attributes, showcasing the potential for simultaneous analysis.

Attention-based models. Guo [22] et al. suggested a path to enhance attribute recognition performance - by improving the attention map. This proposal highlights the critical role of attention maps in attribute recognition, and how their enhancement can potentially lead to superior performance outcomes. Sarfraz [21] et al. put forth a compelling argument for considering viewpoint cues in attribute analysis. They advocated that this consideration would significantly improve the estimation of the correlation between attributes, leading to more accurate and insightful results. Liu [20] et al. introduced a notable proposal to utilize the multi-directional attention (MDA) module. This innovative approach aims to maximize the potential of attention mechanisms in processing and analysing data.

This paper aims to propose a multi-scale pedestrian attribute recognition model based on a dual self-attention mechanism to address the aforementioned challenges and improve the accuracy and robustness of pedestrian attribute recognition. Specifically, this paper first designs a multi-scale feature fusion module to effectively extract multi-scale features from pedestrian images. Then, a dual self-attention mechanism is introduced to model the mutual relationships between different attributes in pedestrian images, thereby improving the accuracy and robustness of pedestrian attribute recognition.

The primary advancements presented in this study can be summarized as follows: 1. An innovative multi-scale feature fusion module has been proposed. The primary function of this module is to effectively extract the correlation between pedestrian attributes from multi-scale features. This approach aims to leverage the potential of multi-scale features and their correlation for in-depth analysis.

2. A dual self-attention mechanism module has been put forward. This module processes features using parallel channel self-attention and spatial self-attention, a unique approach that enables the network to focus more on significant feature regions. As a result, this module significantly improves the accuracy of attribute recognition, marking a step forward in this field. 3. To validate the effectiveness of the proposed methods, extensive experiments were conducted on three public datasets, namely PA-100K [20], RAP [23], and PETA [24]. These experiments included ablation experiments and comparison experiments with classical algorithms. The results demonstrated the superiority of the proposed network framework in the specific task of pedestrian attribute recognition, thereby endorsing its potential for practical application and further research.

3 Methods

This paper proposes an end-to-end multi-scale pedestrian attribute recognition framework. In this method, features with multi-level receptive fields are obtained through a fusion module, and the network focuses on important feature regions through a dual self-attention mechanism. The entire network learns end-to-end in a multi-task setting, with attribute recognition as the primary task. The network framework is illustrated in Fig. 1.

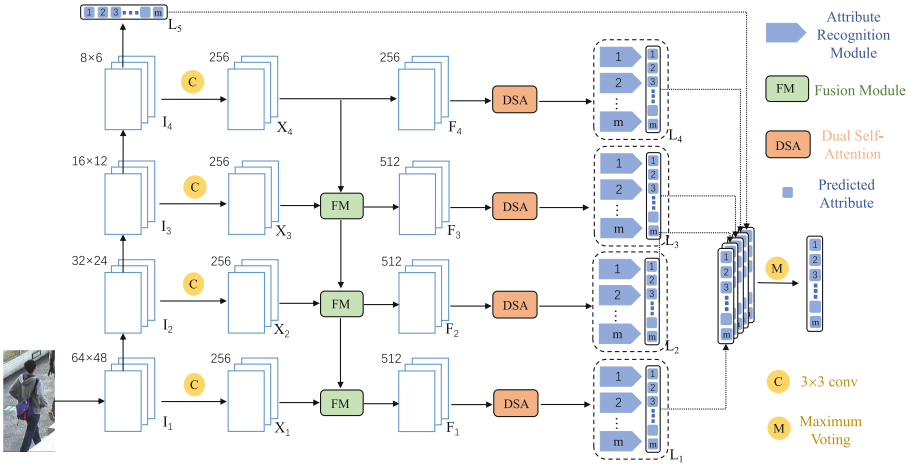

Fig. 1. The comprehensive structure of the framework.

3.1 Network Architecture

The fundamental concept of this study revolves around the use of the dual self-attention module to zero in on the essential details within multi-scale features. As it is well-known, shallow networks have larger receptive fields, which allow them

to capture more details, but they lack contextual information. Deep features and shallow features complement each other. To address this, we employ a feature pyramid structure to glean features across varied scales, thereby making the information more complete.

We adopt ResNeXt50 [25] as the backbone to extract features of four stages, and the specifics of the features are depicted in the far left column of Fig. 1. The feature information of the four scales is represented as $I_i = R^{C_i \times H_i \times W_i}, i \in \{4, 3, 2, 1\}$. The spatial dimensions $H_i \times W_i$ of I_i stand at 8×6,16×12,32×24, 64×48. And the dimensions of the channels C_i correspondingly measure 2048, 1024,512, and 256.

The backbone's pathway can be interpreted as a bottom-up approach, whereas the feature fusion module can be viewed as a top-down strategy. The lateral connections between them play a crucial role in adjusting the feature dimension to a unified d dimension, which is accomplished through convolution. In our experiments, d is set to 256. Then, features from adjacent stages are concatenated by fusion module, which can be articulated as:

$$F_i = FM(C(I_{i+1}), C(I_i)), i \in \{3, 2, 1\} \tag{1}$$

where C represents a 1×1 convolution, FM represents the feature fusion module, and F_i is the fused feature. Since the features of the fourth stage are not fused with other features, they are represented as:

$$F_i = X_i = C(I_i), i \in \{4\} \tag{2}$$

The channel dimension of F_i is $2d$, except for F_4, which has a dimension of d. This is because $F_i(i \in \{3, 2, 1\})$ merge features from two stages, as mentioned earlier, and in our experiments, d is set to 256.

3.2 Fusion Module

Shallow networks with larger receptive fields can capture rich details, while deep networks with smaller receptive fields can capture rich semantic features. To integrate neighboring features and apprehend the attribute correlation amid multi-scale features, we proposed a fusion module, which illustrated in Fig. 2.

For the deep-level feature information, we first upsample it and then perform a sigmoid activation operation. For the shallow-level feature information, we perform a four-layer convolution operation on it. Then, We perform an element-wise addition of the two features. The corresponding formula is articulated as:

$$s' = f(X_i) + \sigma\left(up(X_{i+1})\right), i \in \{3, 2, 1\} \tag{3}$$

where f represents a four-layer convolution, up represents upsampling, and σ represents sigmoid activation operation. Finally, the shallow-level feature $f(X_i)$ and s' are concatenated along the channel dimension to obtain F_i, whose dimension is $2d$. The formula is as follows:

$$F_i = cat(f(X_i), s'), i \in \{3, 2, 1\} \tag{4}$$

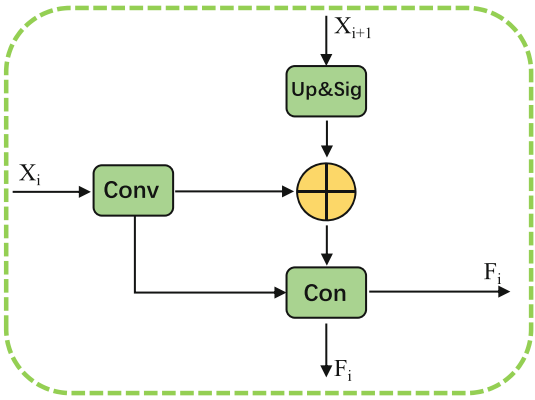

Fig. 2. Fusion Module.

3.3 Dual Self-Attention

In order to glean crucial data from the combined features, we introduce a dual self-attention module that simultaneously employs channel attention and spatial attention. We have termed this the Dual Self-Attention module, and comprehensive details about this module can be seen in Fig. 3.

For the channel self-attention, the feature F_i is split into three branches, each of which undergoes a convolution operation. In branch one, the feature is first average pooled, then softmax activated, and then dot producted with branch two. After being pooled again, the result is dot producted with branch three to obtain $F_{out}^{c_i}$. The equation can be articulated as follows:

$$F_{out}^{c_i} = p\left[\sigma(p(f(F_i))) \cdot f(F_i)\right] \cdot r(f(F_i)), i \in \{1, 2, 3\} \tag{5}$$

For the spatial self-attention, similarly, the feature F is split into three branches, each of which undergoes a convolution operation. However, unlike channel self-attention, branch one does not have a pooling operation, but an additional convolution operation after the first dot product, which is expressed in the formula below:

$$F_{out}^{s_i} = f\left[\sigma(f(F_i)) \cdot f(F_i)\right] \cdot r(f(F_i)), i \in \{1, 2, 3\} \tag{6}$$

where f represents convolution operation, p represents pooling operation, σ represents Sigmoid activation, and r represents ReLU activation. Finally, $F_{out}^{c_i}$ and $F_{out}^{s_i}$ are element-wise added together:

$$F_{out}^{i} = F_{out}^{c_i} + F_{out}^{s_i}, i \in \{1, 2, 3\} \tag{7}$$

4 Experiments

This chapter conducts experiments on the proposed method, which includes datasets, evaluation metrics, experimental equipment and settings, and comparison experiments.

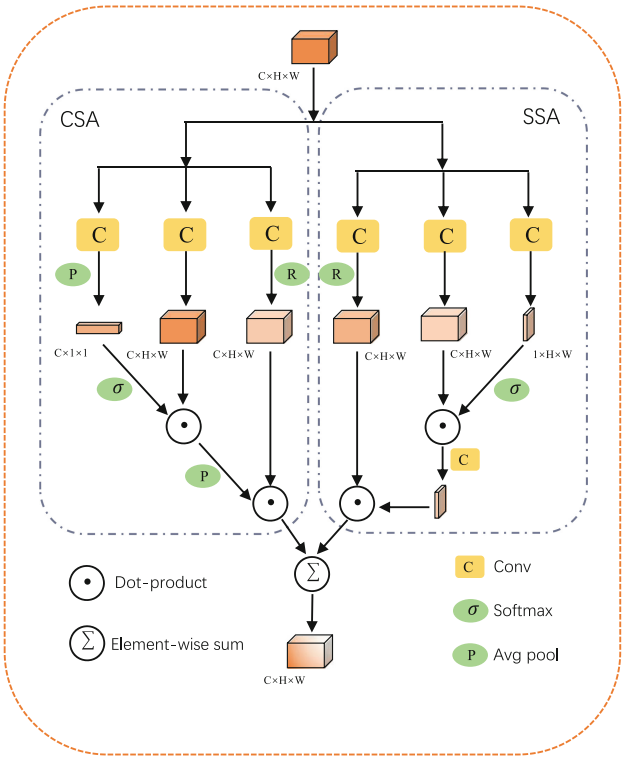

Fig. 3. Dual Self-Attention. The left CSA is Channel Self-Attention, the right SSA is Spatial Self-Attention, and both of them are used in parallel to form the Dual Self-Attention.

4.1 Datasets

The PA-100K [20] dataset, currently the most extensive one utilized for pedestrian attribute recognition, comprises 100,000 pedestrian images. Each of these images is annotated with 26 binary attributes.

The RAP [23] dataset comprises 41,585 images, with each pedestrian being marked with 69 binary attributes that cover both pedestrian characteristics and actions.

The PETA [24] dataset comprises 19,000 images, with their resolutions varying between 17×39 and 169×365. Every pedestrian in the dataset is marked with 61 binary attributes and 4 multi-class attributes.

4.2 Evaluation Metrics

We implement a **label-based** method to determine the mean accuracy (**mA**) of each attribute. This involves computing the classification accuracy for both

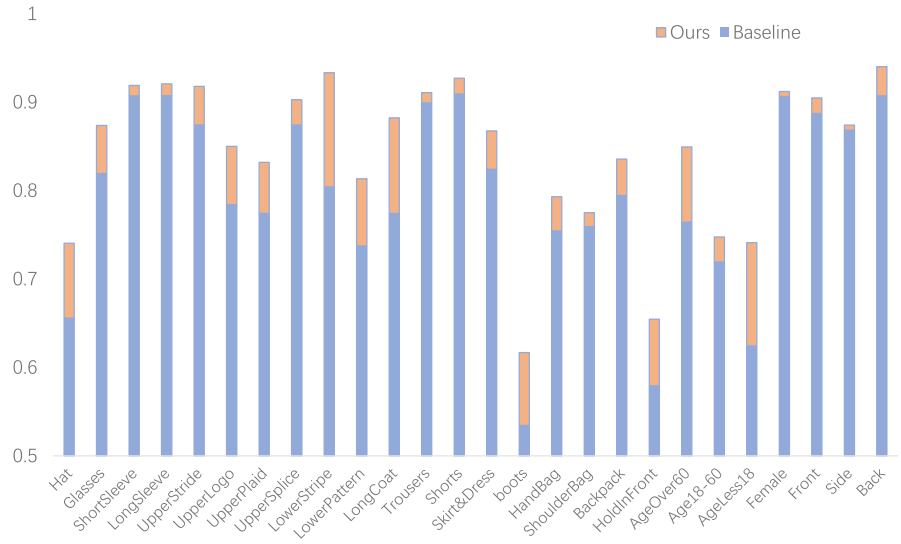

Fig. 4. Attribute comparison on PA-100K. The orange bars represent ours, while the blue bars represent the baseline. (Color figure online)

positive and negative samples and then averaging these results. The formula used for this calculation is

$$mA = \frac{1}{2\,N}\sum_{i=1}^{M}\left(\frac{TP_i}{P_i} + \frac{TN_i}{N_i}\right) \tag{8}$$

We employ a **sample-based** approach and evaluate our method using four commonly used metrics: **accuracy**, **recall**,**precision**, and **F1** score. Given the familiarity of these metrics, we omit their formulas and detail information.

Table 1. The Comparative Analysis on the PA-100K dataset

Method	mA	Rec	Prec	Acc	F1
ALM [7]	80.68	88.84	84.21	77.08	86.46
HPNet [20]	74.21	82.09	82.97	72.19	82.53
PGDM [18]	74.95	82.24	84.36	73.08	83.29
LGNet [26]	76.96	83.17	86.99	75.55	85.04
VeSPA [21]	76.32	81.49	84.99	73.00	83.20
DeepMar [13]	72.70	80.42	82.24	70.39	81.32
CoCNN [27]	80.56	84.36	**89.49**	78.30	86.85
StrongBaseline [29]	80.50	87.12	87.24	78.84	86.78
ours	**81.97**	**90.74**	85.69	**78.88**	**86.89**

Table 2. The Comparative Analysis on the RAP dataset

Method	mA	Rec	Prec	Acc	F1
ALM [7]	**81.87**	86.48	74.71	**68.17**	**80.16**
PGDM [18]	74.31	75.90	78.86	64.57	77.35
LGNet [26]	78.68	79.82	**80.36**	68.00	80.09
HPNet [20]	76.12	78.79	77.33	65.39	78.05
VeSPA [21]	77.70	79.67	79.51	67.35	79.59
DeepMar [13]	73.79	76.21	74.92	62.02	75.56
ours	81.53	**87.85**	76.78	67.96	79.69

Table 3. The Comparative Analysis on the PETA dataset

Method	mA	Rec	Prec	Acc	F1
MTA-Net [28]	84.62	86.42	85.67	78.80	86.04
WPAL [19]	85.50	85.78	84.07	76.98	84.90
ALM [7]	86.30	88.09	85.65	79.52	**86.85**
PGDM [18]	82.97	84.68	**86.86**	78.08	85.76
HPNet [20]	81.77	83.24	84.92	76.13	84.07
VeSPA [21]	83.45	84.81	86.18	77.73	85.49
DeepMar [13]	82.89	83.14	83.68	75.07	83.41
ours	**86.37**	**89.39**	85.62	**79.56**	86.77

4.3 Experimental Equipment and Settings

Our experimental process takes place on the Ubuntu 18.04 operating system, making use of Python 3.8. The network architecture we propose is brought to life using the PyTorch framework, and the training process occurs on a pair of 2080Ti devices. We establish the batch size at 48 and run the model training for a comprehensive 60 epochs.

4.4 Comparison Experiments

Attribute Comparison. Label-based mA is a crucial metric for assessing model performance. In this section, we present a visual representation of the mA values for every attribute present in the PA-100K dataset and compare them with the Baseline [29] in an intuitive manner. As evident from the visualization, our proposed model outperforms the Baseline for all attributes in terms of mA values.

Methods Comparison. There are many classic models that have been experimented on the RAP, PETA,PA-100K and datasets, including DeepMar(Deep

Multi-Attribute Recognization model) [13], VeSPA [21], HPNet(Hydra Plus Net) [20], LGNet (Location Guided Network) [26], CoCNN [27], ALM [7], PGDM(Posed Guided Deep Model) [18], WPAL [19], MTA-Net [28], StrongBaseline [29], etc. The performance comparison of our model with others on the mentioned three datasets is displayed in Table 1, Table 2, and Table 3. It's clear that our approach outperforms others, especially on the PA-100K dataset, with a recall rate improvement of 1.8 over the second place, breaking through 90. The performance on the RAP dataset is slightly inferior, but still has the highest recall rate. There is also good performance on the PETA dataset, with a chance to break through a recall rate of over 90.

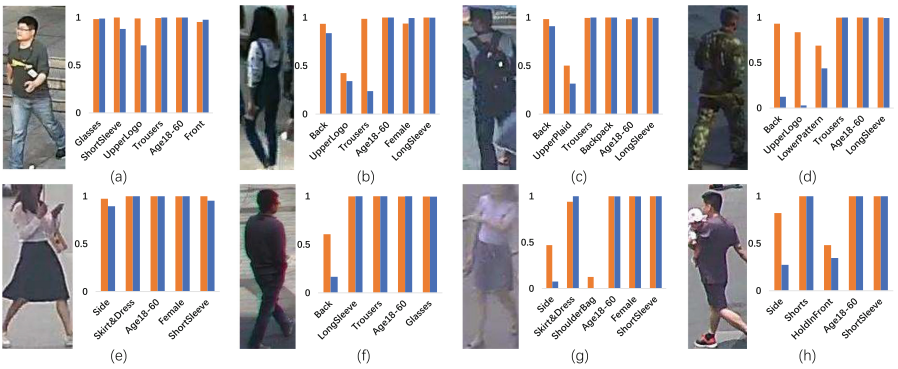

Fig. 5. Sample comparison. The orange bars represent ours, while the blue bars represent the baseline. (Color figure online)

Sample Comparison. Figure 6 displays eight pedestrian images labeled from 'a' to 'h', along with their true labels. The histogram on the right side illustrates the prediction outcomes from our model and the StrongBaseline [29] model. The x-axis of the histogram signifies the predicted attribute, whereas the y-axis denotes the probability of that predicted attribute. It's noticeable that our model outperforms the Baseline model in predicting both coarse-grained and fine-grained attributes. When it comes to attributes associated with blurry and obscured samples, such as the 'ShoulderBag' attribute of sample 'g' which is obscured by the body, the Baseline's prediction probabilities are incredibly low, registering at only 0.0027, which appears as zero on the histogram. In contrast, our model predicts a probability of 0.1258.

5 Conclusion

In this study, we initially employ the feature pyramid structure to extract feature information across multiple scales. Subsequently, our proposed Fusion Module is

applied to enhance the completeness of the feature information. Additionally, we introduce a dual self-attention module that extracts information from both the channel and spatial dimensions of features. Comprehensive experiments carried out on RAP, PETA, and PA-100K datasets serve to underscore the superior performance of our model compared to preceding models.

Acknowledgment. This research received partial support from the National Natural Science Foundation of China(No. 62067003) and the Foundation of Jiangxi Educational Committee (No. GJJ200824).

References

1. Feris, R., Bobbitt, R., Brown, L., Pankanti, S.: Attribute-based people search: lessons learnt from a practical surveillance system. In: Proceedings of the ACM International Conference on Multimedia Retrieval 2014, Glasgow, United Kingdom, pp. 153–160, April 2014
2. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Hu, Z., Yan, C., et al.: Improving person re-identification by attribute and identity learning. Pattern Recognit. **95**, 151–161 (2019)
3. Di, X., Zhang, H., Patel, V.M.: Polarimetric thermal to visible face verification via attribute preserved synthesis. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems, United states, October 2018
4. Di, X., Riggan, B.S., Hu, S., Short, N.J., Patel, V.M.: Multi-scale thermal to visible face verification via attribute guided synthesis. IEEE Trans. Biom. Behav. Identity Sci. 266–280 (2021)
5. Shi, Y., Ling, H., Wu, L., Shen, J., Li, P.: Learning refined attribute-aligned network with attribute selection for person re-identification. Neurocomputing **402**, 124–133 (2020)
6. Li, H., Chen, Y., Tao, D., Yu, Z., Qi, G.: Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification. IEEE Trans. Inf. Forensics Secur. **16**, 1480–1494 (2021)
7. Tang, C., Sheng, L., Zhang, Z.-X., Hu, X.: Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 4996–5005, October2019
8. Zhong, J., Qiao, H., Chen, L., Shang, M., Liu, Q.: Improving pedestrian attribute recognition with multi-scale spatial calibration. In: International Joint Conference on Neural Networks (IJCNN) 2021, pp. 1–8 (2021)
9. Liu, Z., Zhang, Z., Li, D., Zhang, P., Shan, C.: Dual-branch self-attention network for pedestrian attribute recognition. Pattern Recognit. Lett. **163**, 112–120 (2022)
10. Fan, Z., Guan, Y.: Pedestrian attribute recognition based on dual self-attention mechanism. Comput. Sci. Inf. Syst. **20**, 793–812 (2023)
11. Wang, X., Zheng, S., Yang, R., Luo, B., Chen, Z., Tang, J.: Pedestrian Attribute Recognition: A (2019)
12. Sudowe, P., Spitzer, H., Leibe, B.: Person attribute recognition with a jointly-trained holistic CNN model. In: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), Santiago, Chile, pp. 329–337, December 2015. Survey. Pattern Recognit., 121, 108220

13. Li, D., Chen, X., Huang, K.: Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, pp. 111–115, November 2015
14. Abdulnabi, A.H., Wang, G., Lu, J., Jia, K.: Multi-Task CNN model for attribute prediction. IEEE Trans. Multimed. **17**, 1949–1959 (2015)
15. Zhu, J., Liao, S., Yi, D., Lei, Z.: Multi-label CNN based pedestrian attribute learning for soft biometrics. In: 2015 International Conference on Biometrics (ICB), pp. 535–540 (2015)
16. Yang, L., Zhu, L., Wei, Y., Liang, S., Tan, P.: Attribute Recognition from Adaptive Parts (2016). ArXiv, abs/1607.01437
17. Diba, A., Pazandeh, A.M., Pirsiavash, H., Gool, L.V.: DeepCAMP: deep convolutional action & attribute mid-level patterns. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, pp. 3557–3565 (2016)
18. Li, D., Chen, X., Zhang, Z., Huang, K.: Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In: IEEE International Conference on Multimedia and Expo (ICME) 2018, pp. 1–6 (2018)
19. Yu, K., Leng, B., Zhang, Z., Li, D., Huang, K.: Weakly-supervised Learning of Mid-level Features for Pedestrian Attribute Recognition and Localization (2016). ArXiv, abs/1611.05603
20. Liu, X., et al.: HydraPlus-Net: attentive deep features for pedestrian analysis. In: IEEE International Conference on Computer Vision (ICCV) 2017, pp. 350–359 (2017)
21. Sarfraz, M.S., Schumann, A., Wang, Y.: Deep View-Sensitive Pedestrian Attribute Inference in an end-to-end Model (2017). ArXiv, abs/1707.06089
22. Guo, H., Fan, X., Wang, S.: Human attribute recognition by refining attention heat map. Pattern Recognit. Lett. **94**, 38–45 (2017)
23. Li, D., Zhang, Z., Chen, X., Ling, H., Huang, K.: A Richly Annotated Dataset for Pedestrian Attribute Recognition (2016). ArXiv, abs/1603.07054
24. Deng, Y., Luo, P., Loy, C.C.: Pedestrian attribute recognition at far distance. In: Proceedings of the 22nd ACM International Conference on Multimedia (2014)
25. Xie, S., Girshick, R.B., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017, pp. 5987–5995 (2016)
26. Liu, P., Liu, X., Yan, J., Shao, J.: Localization Guided Learning for Pedestrian Attribute Recognition (2018). ArXiv, abs/1808.09102
27. Han, K., Wang, Y., Shu, H., Liu, C., Xu, C., Xu, C.: Attribute Aware Pooling for Pedestrian Attribute Recognition (2019). ArXiv, abs/1907.11837
28. Ji, Z., Hu, Z., He, E., Han, J., Pang, Y.: Pedestrian attribute recognition based on multiple time steps attention. Pattern Recognit. Lett. **138**, 170–176 (2020)
29. Jia, J., Huang, H., Yang, W., Chen, X., Huang, K.: Rethinking of Pedestrian Attribute Recognition: Realistic Datasets with Efficient Method (2020). ArXiv, abs/2005.11909