Improving Pedestrian Attribute Recognition with Dense Feature Pyramid and Mixed Pooling

He Xiao[1,2(✉)], Chen Zou[1], Yaosheng Chen[1], Sujia Gong[1], and Siwen Dong[1]

[1] School of Software Engineering, Jiangxi University of Science and Technology, Nanchang 330013, Jiangxi, People's Republic of China
xiaohe804@gmail.com, chen@jxust.edu.cn, adsw@mail.jxust.edu.cn
[2] Nanchang Key Laboratory of Virtual Digital Factory and Cultural Communications, Nanchang 330013, People's Republic of China

Abstract. In the field of computer vision, pedestrian attribute recognition plays a crucial role in pedestrian detection and pedestrian re-identification. However, this task faces challenges such as blurry images, difficulty in recognizing fine-grained features, and overlooking relationships between pedestrian attributes. To address these challenges, we propose a novel method for pedestrian attribute recognition. Our method is based on convolutional neural networks and incorporates a feature pyramid structure that is specifically designed for the task of pedestrian attribute recognition (PAR). Additionally, we enhance feature information by employing multi-scale feature fusion. Furthermore, our proposed AIIM module facilitates interactions between different attributes by establishing both remote dependencies and short-range dependencies. Through comprehensive experimentation, we have validated the effectiveness of our method and achieved state-of-the-art results. Specifically, our method has achieved impressive average accuracies (mA) of 86.27%, 83.45%, and 81.56% on well-known datasets such as PETA, RAP, and PA100k, respectively.

Keywords: Convolutional neural network · Feature pyramid · Multi-scale fusion · Mixed pooling · Pedestrian attribute recognition

1 Introduction

Pedestrian attribute recognition [1] is a significant area of research in computer vision, with practical applications in video surveillance. This field focuses on identifying semantic descriptions that can be used as soft biometric features for purposes such as pedestrian detection, re-identification, and retrieval. The goal of pedestrian attribute recognition is to predict a set of attributes from a predefined list for a given person image. These attributes provide valuable semantic information to describe and understand pedestrians in images. Incorporating attribute information into computer vision algorithms for tasks like re-identification [2–4], pedestrian detection [5, 6], and retrieval [7, 8] can enhance performance. However, real-world surveillance scenarios present challenges due to factors such as camera angles, lighting conditions, and complexities in pedestrian attributes like pose, occlusion, and blur.

C. Wu et al. (Eds.): MONAMI 2023, LNICST 559, pp. 227–239, 2024.
https://doi.org/10.1007/978-3-031-55471-1_17

The initial approach to pedestrian attribute recognition involved manual feature extraction and separate classification models for each attribute. However, with the emergence of Convolutional Neural Networks (CNN), researchers started exploring multi-task training where all attributes are integrated into a single network, thereby sharing network parameters. This approach proved to be more effective in achieving improved results.

At present, the fundamental methods for pedestrian attribute recognition can be categorized into several groups:

Global image-based models such as DeepSAR [9] and DeepMAR [9] are simpler, more intuitive, and efficient in the context of pedestrian attribute recognition (PAR). However, their performance is constrained by the lack of fine-grained recognition and consideration of correlations between pedestrian attributes. Models that incorporate localization, such as PGDM [10] and LGNet [11], leverage both local and global information to achieve more precise performance. Yet, these models suffer from their heavy reliance on accurate localization, which increases training and inference time and cost. Attention-based models like HydraPlus-Net [12], VeSPA [13], and others selectively focus on specific information but overlook other relevant behavioral and cognitive processes. While these models showcase the effectiveness of attention mechanisms, they shift the research focus towards local features and attention-based features, neglecting the study of attribute correlations.

In recent years, several models have been proposed to address pedestrian attribute identification. These models, including JRL [15], RCRA [16], and other sequence prediction-based models, employ recurrent learning to capture correlations between attributes. Furthermore, some models, such as WPAL [17] and other loss function-based models, have improved upon the optimized loss function for PAR to achieve more precise predictions.

By incorporating multi-scale features into the feature pyramid architecture, convolutional neural networks (CNNs) have proven effective in utilizing the advantages of different feature levels. However, the conventional method of constructing the feature pyramid may not fully capture the intricate characteristics present in natural scenarios, such as pedestrian attribute recognition. In order to address this limitation, we propose an innovative and more dense feature pyramid model.

Our approach aims to enhance the feature pyramid architecture by incorporating additional scales and adding finer details to the feature hierarchy. This allows us to extract and analyze features at multiple levels of granularity, thereby enriching the representation of the input data. By doing so, we can better capture the distinctive attributes and characteristics of pedestrians in real-world scenarios.

Modeling contextual relationships among different regions in an image is advantageous for image recognition. However, previous methods may struggle to capture global context and identify distance dependencies between regions, resulting in high computational complexity. Our proposal introduces hybrid pooling to acquire context information. This approach involves a strip pooling sub-module that establishes long-range dependencies and a lightweight pyramidal pooling sub-module that creates short-range dependencies.

To gain a clearer understanding of the effectiveness and focus of our model, we randomly selected samples from the PETA dataset and conducted a sample heat map analysis. Figure 1 presents four sets of images arranged from left to right as follows: the sample map, the heat representation map of the backbone network output, and the heat map of the final model output.

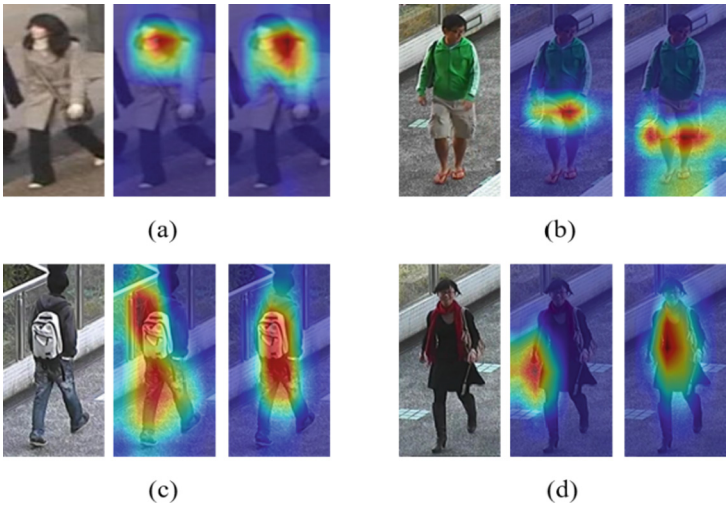

Fig. 1. Sample heat map

To achieve our goal, we present a novel model architecture that utilizes dense feature pyramids to extract and combine features at various scales. Furthermore, we enhance the model's performance by incorporating global contextual information to learn the relationships between different attributes in the attribute set and identifying long-range dependencies in diverse regions. We evaluate our approach on three publicly available datasets (PETA, PA100k, and RAP) to demonstrate its effectiveness. Our paper makes the following contributions:

- We propose an improved feature pyramid structure with a dense information encoding module and feature fusion module for the PAR task.
- We introduce a new PAR network model architecture that employs mixed pooling as our AIIM module, enabling the capture of remote dependencies and the learning of correlations between attributes.
- Through extensive ablation experiments on the PETA, PA100k, and RAP datasets, we validate the effectiveness of our proposed network framework.

2 Methods

2.1 Network Architecture

The paper presents a network architecture, as depicted in Fig. 2. Firstly, we feed images into the ResNeXt50 [23] backbone network and utilize a feature pyramid concept to construct a four-layer network model with a feature pyramid structure. The input image

is resized to 256 × 192, and the output features in the main network are represented as $I_i = R^{C_i \times H_i \times W_i}$, where i ranges from 1 to 4, representing stages from shallow to deep. The output feature I_i are 64 × 48, 32 × 24, 16 × 12 and 8 × 6, respectively. The number of channels for each stage is 256, 512, 1024, and 2048.

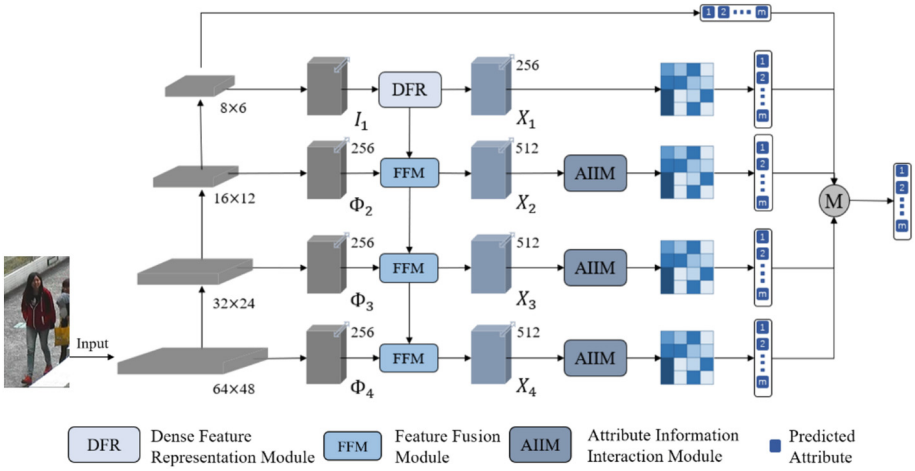

Fig. 2. Overall network architecture

In the construction of feature pyramids [24], existing approaches have primarily focused on inter-layer feature interactions, overlooking intra-layer feature relationships. However, taking inspiration from prior research on dense prediction tasks [25], we propose a denser approach to intra-layer feature conditioning. To accomplish this, we introduce a DFR module to the top layer of the feature pyramid and fuse it with other shallow features. Compared to conventional feature pyramids, our pyramid structure [26] captures global long-range dependencies both within and between layers, thereby producing a comprehensive and distinctive feature representation.

In order to minimize redundant information within the network and prepare for subsequent feature fusion across different scales, we compress the channel number of the previously acquired feature maps I_1, I_2, and I_3 to 256 channels each, denoted as Φ_i, where $i \in \{1, 2, 3\}$. Considering that deeper features typically encompass more abstract representations compared to shallower features, we propose a global concentration rule utilizing a top-down approach. This rule employs the spatial display visual center obtained from the deepest features to simultaneously condition all preceding shallow features. For I_4, we input it into the DFR module to obtain Φ_4, which also has 256 channels.

In order to acquire global information from each feature map, establish correlations between different attributes, we employ the attribute information interaction module. This module facilitates the establishment of correlations and dependencies among global information. Ultimately, the attribute identification module generates multiple prediction outputs, and the final PAR prediction results are obtained through a multi-branch voting mechanism.

2.2 Dense Feature Representation Module

The focus of this module is to propose a lightweight MLP architecture that captures the global long-range dependence of I_4. Instead of using the multi-headed self-attentive module from the standard transformer encoder, we replace it with an MLP layer. Compared to the transformer encoder that relies on the multi-headed attention mechanism, our lightweight MLP architecture is simpler in structure, has a smaller size, and offers higher computational efficiency. Additionally, we incorporate a learnable vision-centric mechanism to aggregate the local corner regions of the input image along with the lightweight MLP. This parallel structured network, which we refer to as spatial DFR (Dense Feature Representation), is placed in the processing of top-level features. Based on the proposed DFR, we construct the feature pyramid using a top-down approach and introduce a feature fusion technique that optimally combines the shallow features of the pyramid with the visually focused information from the deepest features (Fig. 3).

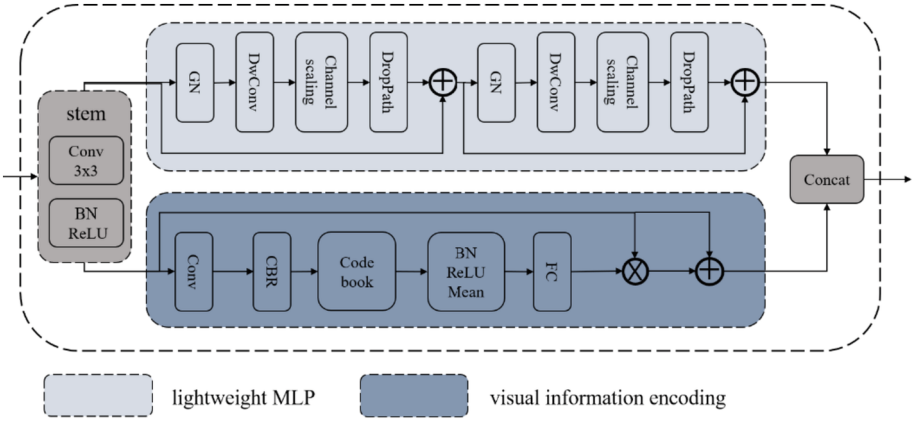

Fig. 3. Dense feature representation module

Our proposed Dense Feature Representation (DFR) module consists of two parallel connected blocks, as depicted in the upper part of Fig. 2. To capture global long-range dependencies, i.e., the global information of the image, we utilize a lightweight MLP. Furthermore, we incorporate a visual information encoding submodule in the lower part of the figure to preserve features in local corner regions, i.e., the local information of the image. This module is applied to the top-level features produced by the backbone network. Instead of directly processing the original feature map, we employ a stem block for feature smoothing between the top-level features and this module. The stem block comprises a 3×3 convolution with an output channel size of 256, a batch normalization layer, followed by a batch normalization layer and an activation function layer. The process can be mathematically expressed using Eq:

$$X = cat(MLP(X_{in}); VIE(X_{in})) \tag{1}$$

where X is the output of DFR. X_{in} refers to the output of the Stem block, which is obtained through the following steps:

$$X_{in} = \sigma(BN(conv_{3\times3}(I_4))) \tag{2}$$

The lightweight MLP incorporates two primary residual modules: the depthwise convolution-based module and the channel MLP-based module. Each module incorporates a channel scaling layer and DropPath operation to enhance the generalization and robustness of features. The utilization of depthwise convolution, rather than traditional convolution, in this segment enhances the capability of feature representation while simultaneously reducing computational costs.

$$X'_{in} = DConv(GN(X_{in})) + X_{in} \tag{3}$$

where X'_{in} represents the output of depthwise convolution-based module.

$$MLP(X_{in}) = CMLP\left(GN\left(X'_{in}\right)\right) + X'_{in} \tag{4}$$

where CMLP(·) represents the channel MLP.

The main function of the visual information encoding submodule is to initially encode the features from the Stem block through a set of convolutional layers. The encoded features are further processed by the CBR block, which comprises a 3 × 3 convolution with a BN layer and ReLU activation function. These steps allow the encoded features to be inputted into the Codebook. To achieve this, a set of scaling factors effectively maps the corresponding positional information. The information of the entire image for the k_{th} codeword can be calculated as follows:

$$e_k = \sum\nolimits_{i=1}^{N} \frac{e^{-s_k \left\| X'_i - b_k \right\|^2}}{\sum_{j=1}^{k} e^{-s_k \left\| X'_i - b_k \right\|^2}} \left(X'_i - b_k\right) \tag{5}$$

where X'_i is i_{th} pixel point, b_k is k_{th} learnable visual codeword, and s_k is k-th scaling factor. $X'_i - b_k$ denotes the information regarding the relative pixel position to a codeword.

$$e = \sum\nolimits_{k=1}^{K} \varnothing(e_k) \tag{6}$$

The process involves incorporating a BN layer with ReLU and mean layer in ∅. Once the codebook output is obtained, we utilize a fully connected layer and a 1 × 1 convolution layer to forecast features that accentuate crucial classes. Subsequently, we apply channel-wise multiplication between the input features from the Stem block Xin and the scaling factor coefficient $\delta(\cdot)$. The aforementioned procedures can be represented as:

$$Z = X_{in} \otimes (\delta(Conv_{1\times1}(e))) \tag{7}$$

The sigmoid function δ(·) is utilized, and channel-wise multiplication ⊗ is applied. Afterwards, the features Xin generated from the Stem block are combined with the local corner region features Z via a channel-wise addition. This process is expressed as:

$$VIE(X_{in}) = X_{in} \oplus Z \tag{8}$$

2.3 Feature Fusion Module

We use the concept of a feature pyramid to combine features from different scales, taking advantage of their unique levels. The deep layer network has a strong representation of semantic information and is capable of obtaining abundant semantic information. Shallow networks have the ability to capture intricate details due to their neurons having a narrower view of the input. This is a result of the smaller receptive fields that arise from fewer layers of processing. By merging the features of adjacent layers, we can capture the correlation between attributes of features at different scales.

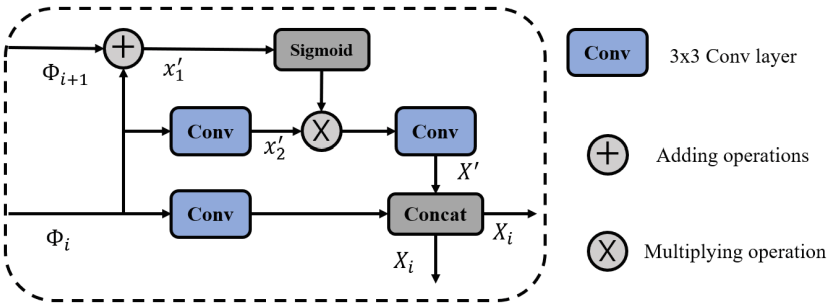

Fig. 4. Feature fusion module

Figure 4 illustrates the Flow-based Feature Module (FFM). Within adjacent layers of scaled features, the deeper feature maps are added and combined with the shallow features, followed by activation through the sigmoid function. The nearby shallow features undergo further processing via a convolutional block to extract information. This block consists of a 3×3 convolution layer, followed by normalization using BatchNorm2d, and activation using ReLU. Finally, the newly obtained features, denoted as X', are integrated with the features extracted from the original shallow feature map in the channel dimension to yield the desired final feature output, denoted as X_i.

2.4 Attribute Information Interaction Module

The methods currently being researched to enhance the remote dependency modeling capability of CNN include introducing attention mechanisms [28–30], using dilation convolution or depthwise separable convolution [31], among others. However, some of these methods require significant memory consumption to compute large affinity matrices for each space, while others may not effectively analyze natural scenes such as pedestrian attribute recognition.

An important characteristic of natural images is that their features typically appear in a non-uniform manner. Previous research [32] has shown that strip pooling can effectively resolve some natural scenes and reduce computational complexity, while maintaining the ability for global modeling.

The pyramid pooling module [23] (PPM) is effective in enhancing scenario resolution networks, but it relies too heavily on standard spatial pooling operations. To address

this, we propose a hybrid pooling approach that combines standard spatial pooling and strip pooling. This approach aggregates different pooling operations with different types of contextual information to create a more discriminative feature representation for pedestrian attribute recognition tasks.

In this paper, we propose using a hybrid pooling approach to establish attribute information correlations for pedestrians. The module consists of two sub-modules: a strip pooling module for establishing long-range dependencies and a sub-module for collecting short-range dependencies using lightweight pyramidal pooling.

3 Experiment Results and Analysis

3.1 Data Set

To evaluate the efficacy of our proposed model, we conducted experiments on three publicly available datasets: RAP, PETA, and PA-100K. The PETA dataset comprises 19,000 images featuring 8,705 pedestrians, with resolutions ranging from 17 × 39 to 169 × 365. The RAP dataset consists of 41,585 images captured by real indoor surveillance cameras, and resolutions varying from 36 × 92 to 344 × 554. PA-100K is a substantial pedestrian attribute dataset, consisting of 100,000 images.

During the experimentation process, we divided the PETA dataset into a training set of 11,400 images and a testing set of 7,600 images. The RAP dataset's 33,268 images were utilized for training, while 8,317 images were allocated for testing. As for the PA-100K dataset, we employed 90,000 images for training purposes and 10,000 images for testing purposes.

3.2 Evaluation Metrics

To evaluate the chosen datasets RAP, PETA, and PA100k, two evaluation metrics were employed: label-based and sample-based. In regard to label-based evaluation, we employed the mean accuracy (mA). We computed the accuracy for each attribute across all samples, and subsequently obtained the average of all attributes to derive mA. The mA evaluation metric is defined as follows:

$$mA = \frac{1}{2N}\sum_{i=1}^{m}\left(\frac{TP_i}{P_i} + \frac{TN_i}{N_i}\right) \tag{9}$$

For evaluating samples, we use four commonly used metrics: accuracy, precision, recall, and F1 value, which are defined as follows.

$$accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|Y_i \cup f(x_i)|} \tag{10}$$

$$precision = \frac{1}{2N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|f(x_i)|} \tag{11}$$

$$recall = \frac{1}{2N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|Y_i|} \tag{12}$$

$$F1 = \frac{2{*}precision{*}recall}{precision + recall} \tag{13}$$

3.3 Comparison Experiments

Comparison with State-of-the-Arts

To demonstrate the effectiveness of our proposed method, we compared the performance of our proposed network with several state-of-the-art networks, such as DeepMar, PGDM, LG-Net, HPNet, VeSPA, JRL, MT-CAS, MTA-Net, and WPAL. Our experiments were conducted on three publicly available datasets: PETA, PA-100K, and RAP datasets. The results of these experiments are presented in Table 1, which includes evaluation metrics such as mA, accuracy, precision, recall, and F1. Upon analyzing the Table 1, it becomes apparent that our model exhibits clear advantages over the other networks.

Table 1. Comparison of our network with state-of-the-arts on PETA and RAP datasets

Methods	PETA					RAP				
	mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1
ACN	81.15	73.66	84.06	81.26	82.64	69.66	62.61	80.12	72.26	75.98
DeepMar	82.89	75.07	83.68	83.14	83.41	73.79	62.06	74.92	76.21	75.56
PGDM	82.97	78.08	86.86	84.68	85.76	74.31	64.57	78.86	75.90	77.35
LG-Net	–	–	–	–	–	78.68	68.00	**80.36**	79.82	80.09
HPNet	81.77	76.13	84.92	83.24	84.07	76.12	65.39	77.33	78.79	78.05
VeSPA	83.45	77.73	86.18	84.81	85.49	77.70	67.35	79.51	79.67	79.59
JRL	85.67	-	86.03	85.34	85.42	77.81	–	78.11	78.98	78.58
MT-CAS	83.17	78.78	**87.49**	85.35	86.41	–	–	–	–	–
MTA-Net	84.62	78.80	85.67	86.42	86.04	77.62	67.17	79.72	78.44	79.07
WPAL	85.50	76.98	84.07	85.78	84.90	81.25	50.30	57.17	78.39	66.12
ours	**86.27**	**79.34**	84.96	**88.86**	**86.59**	**83.45**	**68.94**	77.47	**85.09**	**80.80**

As depicted in the Table 1, our proposed model exhibits superior performance on all three datasets: PETA, RAP, and PA100k. In comparison to other models, our model ranks first in four evaluation metrics on both PETA and RAP datasets, and also ranks first in three evaluation metrics on the PA100k dataset. This indicates that our proposed model has better generalization ability. In addition, the PA100k dataset stands out due to its extensive image collection and relatively low resolution, making it a more representative portrayal of image diversity and motion blur in real-world scenarios. The impressive performance of RMFA module on the PA100k, as demonstrated in Table 2, highlights the effectiveness of integrating multi-scale information. Despite the PETA and RAP datasets having a larger number of labeled attributes compared to PA100k, the results presented in the table further underscore the importance and effectiveness of mining contextual information and attribute correlation.

Table 2. Comparison of our network with state-of-the-arts on PA100k datasets

Methods	PA100k				
	mA	Accu	Prec	Recall	F1
DeepMar	72.70	70.39	82.24	80.42	81.32
PGDM	74.95	73.08	84.36	82.24	83.29
LG-Net	79.96	75.55	86.99	83.17	85.04
HPNet	74.21	72.19	82.97	82.09	82.53
VeSPA	76.32	73.00	84.99	81.49	83.20
MT-CAS	77.20	78.09	**88.46**	84.86	**86.62**
ours	**81.60**	**78.40**	85.16	**88.85**	86.55

Attribute Recognition Comparison

As mA is one of the important metrics for evaluating the effectiveness of a model among all evaluation metrics, we plotted the mA values of 35 attributes in the PETA dataset based on the test results of our model compared to StrongBaseline [38].

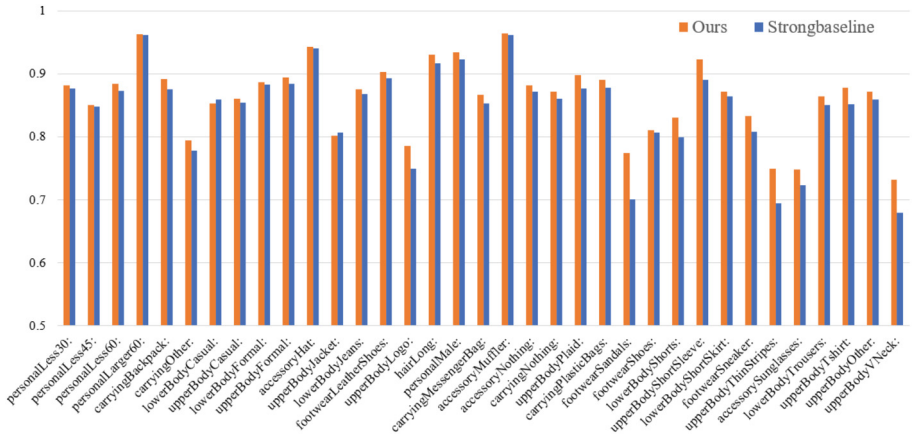

Fig. 5. Comparison of attribute recognition in our model and strongbaseline

The results, as shown in the Fig. 5, indicate that our model improves the identification of almost all attributes, including "upper Body Thin Stripes," "footwear Sandals," and "accessory Sunglasses." This improvement can be attributed to the DFR module, which allows for the construction of a dense feature pyramid and the fusion of multi-scale information. For pedestrian attributes such as "Age," "personal Male," and "upper Body Short Sleeve," the enhancement in high-resolution attributes like "lower Body Short Skirt" can be credited to attribute correlation and contextual information mining.

3.4 Ablation Experiments

In order to assess the effectiveness of the key components in our proposed network architecture and the influence of other factors, we conducted ablation experiments. The data set used in ablation experiments was PETA. The evaluation metrics employed for this experiment encompass mean average precision, accuracy, precision, recall, and F1. Table 3 shows two architectures: one without the DFR module and one without the AIIM module. Analyzing the data, our proposed complete architecture outperforms in four out of five evaluation indicators. This supports the importance and effectiveness of both the DFR and AIIM modules in the overall network architecture.

Table 3. Results of ablation experiments

	mA	Accu	Prec	Recall	F1
-DFR	84.74	76.70	81.34	89.51	84.81
-AIIM	85.24	76.79	81.33	**89.96**	85.00
ours	**86.27**	**79.34**	**84.96**	88.86	**86.59**

4 Conclusion

In this paper, we propose a dense feature pyramid structure for natural scenarios, such as pedestrian attribute recognition. We optimize the feature pyramid using the dense information encoding module and the proposed feature fusion method. Additionally, we propose a novel architecture for pedestrian attribute recognition that leverages the correlation between contextual information and pedestrian attributes, combined with the hybrid pooling-based AIIM module. We conducted extensive experiments on three datasets, namely PEAT, RAP, and PA100k, and concluded that our proposed architecture significantly improves performance while achieving excellent results. Furthermore, we demonstrate the effectiveness of key blocks in our proposed architecture through ablation experiments. However, the performance shown on the PA100K dataset is not satisfactory, indicating that the recognition performance of the model decreases when the dataset becomes larger. This result is possibly due to the increase of disturbing factors in the images caused by the increase of the dataset, which can be investigated in the future to improve the performance of the pedestrian attribute recognition model.

Acknowledgment. This work received support from the Foundation of Jiangxi Educational Committee under grant No. GJJ200824.

References

1. Liu, W., Liao, S., Ren, W., Hu, W., Yu, Y.: High-level semantic feature detection: a new perspective for pedestrian detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5187–5196 (2019)

2. Lin, Y., et al.: Improving person re-identification by attribute and identity learning. Pattern Recognit. **95**, 151–161 (2019)
3. 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)
4. Li, D., Zhang, Z., Chen, X., Huang, K.: A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. IEEE Trans. Image Process. **28**(4), 1575–1590 (2018)
5. Brunetti, A., Buongiorno, D., Trotta, G.F., Bevilacqua, V.: Computer vision and deep learning techniques for pedestrian detection and tracking: a survey. Neurocomputing **300**, 17–33 (2018)
6. Tay, C.P., Roy, S., Yap, K.H.: AANet: attribute attention network for person re-identifications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7134–7143 (2019)
7. Sun, Y., Zheng, L., Deng, W., Wang, S.: SVDNet for pedestrian retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3800–3808 (2017)
8. 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), pp. 111–115. IEEE, November 2015
9. Li, D., Chen, X., Zhang, Z., Huang, K.: Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In: 2018 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE, July 2018
10. Liu, P., Liu, X., Yan, J., Shao, J.: Localization guided learning for pedestrian attribute recognition (2018). arXiv preprint arXiv:1808.09102
11. Liu, X., et al.: HydraPlus-Net: attentive deep features for pedestrian analysis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 350–359 (2017)
12. Sarfraz, M.S., Schumann, A., Wang, Y., Stiefelhagen, R.: Deep view-sensitive pedestrian attribute inference in an end-to-end model (2017). arXiv preprint arXiv:1707.06089
13. Guo, H., Fan, X., Wang, S.: Human attribute recognition by refining attention heat map. Pattern Recognit. Lett. **94**, 38–45 (2017)
14. Wang, J., Zhu, X., Gong, S., Li, W.: Attribute recognition by joint recurrent learning of context and correlation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 531–540 (2017)
15. Zhao, X., Sang, L., Ding, G., Han, J., Di, N., Yan, C.: Recurrent attention model for pedestrian attribute recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 9275–9282, July 2019
16. 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 preprint arXiv:1611.05603
17. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. Adv. Neural Inf. Process. Syst. **29** (2016)
18. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
19. Honari, S., Yosinski, J., Vincent, P., Pal, C.: Recombinator networks: learning coarse-to-fine feature aggregation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5743–5752 (2016)
20. Deng, Y., Luo, P., Loy, C.C., Tang, X.: Pedestrian attribute recognition at far distance. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 789–792, November 2014
21. Li, D., Zhang, Z., Chen, X., Ling, H., Huang, K.: A richly annotated dataset for pedestrian attribute recognition (2016). arXiv preprint arXiv:1603.07054

22. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)
23. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
24. Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K.: DenseASPP for semantic segmentation in street scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3684–3692 (2018)
25. Quan, Y., Zhang, D., Zhang, L., Tang, J.: Centralized Feature Pyramid for Object Detection (2022). arXiv preprint arXiv:2210.02093
26. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation (2017). arXiv preprint arXiv:1706.05587
27. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs (2016). arXiv preprint arXiv:1606.00915
28. Zhao, H., Jia, J., Koltun, V.: Exploring self-attention for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10076–10085 (2020)
29. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
30. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: CCNET: criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 603–612 (2019)
31. Guo, Y., Li, Y., Wang, L., Rosing, T.: Depthwise convolution is all you need for learning multiple visual domains. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 8368–8375, July 2019
32. Hou, Q., Zhang, L., Cheng, M.M., Feng, J.: Strip pooling: rethinking spatial pooling for scene parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4003–4012 (2020)
33. Sudowe, P., Spitzer, H., Leibe, B.: Person attribute recognition with a jointly-trained holistic CNN model. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 87–95 (2015)
34. Zeng, H., Ai, H., Zhuang, Z., Chen, L.: Multi-task learning via co-attentive sharing for pedestrian attribute recognition. In: 2020 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE, July 2020
35. 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 (BTAS), pp. 1–10. IEEE, October 2018
36. Tang, C., Sheng, L., Zhang, Z., Hu, X.: Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4997–5006 (2019)
37. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456. PMLR, June 2015
38. Zhong, J., Qiao, H., Chen, L., Shang, M., Liu, Q.: Improving pedestrian attribute recognition with multi-scale spatial calibration. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE, July 2021
39. Jia, J., Huang, H., Yang, W., Chen, X., Huang, K.: Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method (2020). arXiv preprint arXiv:2005.11909
40. Zou, C., Xie, W., Xie, X., Zhao, K., Liu, Q., Xiao, H.: Pedestrian Attribute Recognition Based on Multi-Scale Feature Fusion Over a Larger Receptive Field and Strip Pooling (2022)