Pedestrian Attribute Recognition Based on Multi-Scale Feature Fusion Over a Larger Receptive Field and Strip Pooling

Chen Zou[1], Wenbiao Xie[1], Xiaomei Xie[1], Kai Zhao[1], Qiuming Liu[1,2] and He Xiao[1,*]

[1]*School of Software Engineering, Jiangxi University of Science and Technology, Jiangxi Nanchang 330013, P. R. China*
[2]*Nanchang Key laboratory of Virtual Digital Factory and Cultural Communications, Jiangxi Nanchang 330013, P. R. China*

Abstract

Pedestrian attribute recognition is a vital task in computer vision, which is widely used in pedestrian detection and pedestrian re-identification, etc. Pedestrian attribute recognition aims to excavate the attributes of the target person from video or images. To solve the specific challenging factors in this task, such as changes in viewpoint, lacking illumination, and low resolution, we propose a brand new pedestrian attribute recognition method. Using ASPP to extend the receptive fields and densely connect the generated features, multi-scale feature fusion merges features from different receptive fields into one feature that is more discriminative than the input features. And the AIIM module is used to establish remote dependencies at different spatial scales. Extensive experiments show that our proposed method achieves state-of-the-art results with average accuracy (mA) of 86.35%, 81.60%, and 83.29% on public datasets such as PETA, PA100k, and RAP, respectively.

Keywords

Convolutional neural network; Atrous spatial pyramid pooling; Multi-scale fusion; Strip pooling; Pedestrian attribute recognition

1. Introduction

Pedestrian attribute recognition[1] aims to predict a set of attributes from a predefined list of personal attributes $A = \{a_1, a_2, \cdots, a_m\}$ from given images and video, and these attributes are employed as soft biometric features in visual surveillance system. While in computer vision tasks such as pedestrian detection[2][3], pedestrian re-identification[4][5][6], and pedestrian retrieval[7][8], the attribute information can be integrated into computer vision algorithms to achieve better performance. Therefore, pedestrian attribute recognition is a vital task in computer vision. In surveillance scenarios, the cameras shoot at different angles and under different lighting conditions, which can lead to point-of-view problems. All these factors make pedestrian attribute recognition a challenging task in specific scenes.

Global Image-based models: Such as DeepSAR[9] and DeepMAR[9], etc., take the whole image as input and perform multi-task learning of PAR. It has been experimentally observed that the performance of these models is still limited by the lack of fine-grained recognition and the consideration of correlation between pedestrian attributes.

Part-based models: These algorithms can jointly use local and global information to obtain more accurate recognition, such as PGDM[10] and LGNet[11]. In PGDM, key points are converted into information regions by the obtained pose information, and feature learning is performed from each key point related region using independent convolutional neural network. LGNet proposes a localization guidance network localizing the regions corresponding to different attributes. The

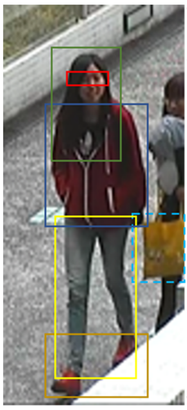
ICBASE2022@3rd International Conference on Big Data & Artificial Intelligence & Software Engineering, October 21-23, 2022, Guangzhou, China
chen@jxust.edu.cn (Chen Zou), vebrun@jxust.edu.cn (Wenbiao Xie), mavia@jxust.edu.cn (Xiaomei Xie), kewitt@jxust.edu.cn (Kai Zhao), liuqiuming@hust.edu.cn (Qiuming Liu), xiaohe804@gmail.com (He Xiao)

shortcoming of such algorithms is that the final recognition performance depends on the accuracy of localization information, and it increases its training time because of the introduction of human part localization.

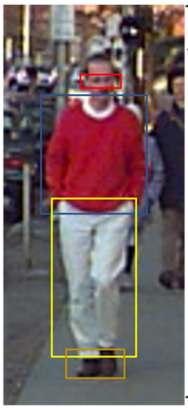
Attention-based models: Methods based on attention mechanisms have become prevalent in recent years, such as HydraPlus-Net[12], VeSPA[13], etc. HydraPlus-Net is introduced to encode multi-scale features from multiple levels for pedestrian attributes analysis using multi-directional attention (MDA) modules. VeSPA take view cues into account and use the attention weights of view predictions multiplied by view-specific coarse attribute predictions to obtain the final multi-class attribute predictions to improve attribute prediction accuracy. Guo et al. [14]emphasize the importance of refining the attentional heat map for each attribute. Although all of these approaches focus on feature learning, shifting the research focus from the earlier global features to local features and attention-based features. However, both ignore the study of attribute relevance.

Sequential prediction-based models: JRL[15], RCRA[16], etc., are proposed in this class of methods that use recurrent learning to obtain the correlation between different attributes to achieve more accurate prediction results.

Models based on loss functions: In recent years, there are also some loss functions optimized for PAR improvement, such as WPAL[17].



	Ours	baseline
Age16-30	1	1
Casual lower	1	1
Casual upper	1	1
Jeans	1	1
Long hair	1	1
No accessory	1	1
No carrying	1	0
Shoes	1	0



	Ours	baseline
Age31-45	1	1
Casual lower	1	1
Casual upper	1	1
Leather Shoes	1	1
Male	1	1
No accessory	1	0
No carrying	1	1
Trousers	1	1

Figure 1. Sample predictions comparisons, with "1" indicating accurate recognition and "0" indicating failure to recognize.

To address the shortcomings of some previous algorithms, we propose a brand new pedestrian attribute recognition model architecture. As shown in figure 1, the realized boxes in the figure are the label boxes to be recognized for this sample, and the attributes that StrongBaseline[38] predicts incorrectly are in the dashed boxes. The attribute labels of the corresponding samples and the results predicted by the model are shown in the table next to them. As can be seen from the first sample, StrongBaseline does not accurately identify "No carrying" and "Shoes", and in the second sample, StrongBaseline does not accurately identify "No accessory", while our model accurately identifies all attribute labels. We believe that the reason why "No carrying" is not accurately recognized is that the original model misidentifies the objects in the dashed boxes in the figure. For another problem shown in the figure, the original model fails to recognize the attributes of "Shoes" and "No accessory", which are small detection targets, we solve the problem by increasing the receptive field and fusing the multi-scale information.

Modeling the contextual relations of different image regions can help to improve recognition accuracy, but earlier methods may fail to collect global context and capture the long-range dependencies of different regions due to the limitation of the receptive field[18]. Larger receptive fields make attribute-specific features easier to recognize. In convolutional neural networks, low-level features can help capture local area information, while high-level features exhibit global semantic information. Therefore, by fusing multi-scale features in the feature pyramid architecture[19][20], the advantages of different levels of features can be exploited.

In this paper， we propose a new model architecture to extract and fuse multi-scale features over a larger receptive field and obtain correlations by obtaining global contextual information and capturing long-range dependencies in different regions. We conducted experiments on three public datasets PETA[21], PA100k[12], and RAP[22] to validate the effectiveness of proposed algorithm. We have made the following contributions to this paper:

- We propose a multi-scale information encoding module (MSEM) for PAR to extract multiscale features and encode multiscale information over a larger receptive field.
- We adopt the feature pyramid model to construct a fusion module (FFM) of multi-scale features.
- We propose a new PAR network model architecture that uses stripe pooling as our attribute information interaction module (AIIM) to capture remote dependencies and learn correlations between attributes.
- Extensive experiments on three public datasets, PETA, PA100k, RAP, and ablation experiments demonstrate the effectiveness of the proposed network framework.

2. Method

2.1. Overall Network Architecture

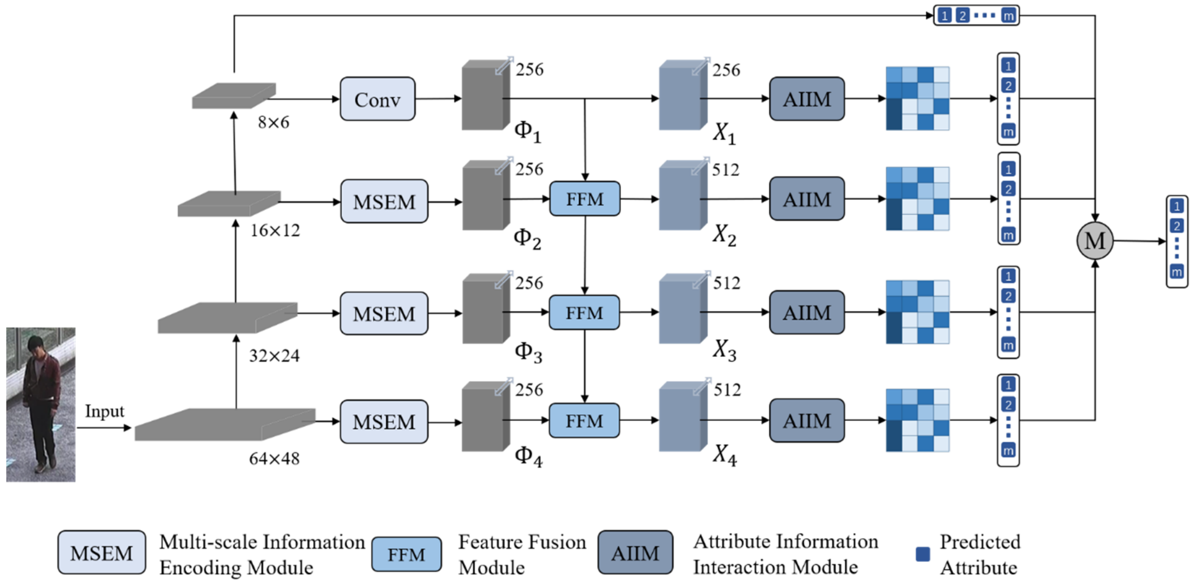


Figure 2. overall network architecture.

The overall network architecture of this paper is shown in figure 2. Inspired by the feature pyramid, we use ResNeXt50[23] as the backbone network and construct a network model with a four-layer feature pyramid structure. The image size is resized to 256 × 192 and input to the backbone network. In the backbone network, the output of the feature by the four stages from shallow to deep are represented as $I_i = R^{C_i \times H_i \times W_i}$, $i \in \{1,2,3,4\}$. The size of the output features I_i is 64×48, 32×24, 16×12, 8×6, and the number of channels is 256, 512, 1024, and 2048, respectively.

The feature maps I_1, I_2, and I_3 obtained above are the input of the multiscale information encoding module, which is composed in the main part of DenseASPP[24]. The module convalesces the generated feature maps with different dilation rates of dilation so that the neurons in the output feature maps contain multiple receptive field sizes that encode multiscale information and feed the output of each dilation convolution layer to all the unexecuted dilation convolution layers using dense connections. To construct the feature pyramid structure later as well as to effectively reduce the network parameters, we extract these feature maps uniformly integrated into 256 channels to obtain Φ_i , $i \in \{1,2,3\}$. For I_4, we only perform a simple dimensionality reduction to obtain Φ_4 , and again the number of channels of Φ_4 is 256.

To aggregate information from different spatial scales and fuse features at different scales, we use a feature fusion module to connect shallow features with features at a deeper level to obtain a more discriminative feature map, denoted by X_i , $i \in \{1,2,3\}$, and X_4 is obtained from Φ_4 without any processing, i.e., $X_4 = \Phi_4$. Subsequently, to obtain the global information in each feature map and the correlation between different attributes, the attribute information interaction module was used to establish the correlation and dependency of the global information. Finally, multiple prediction outputs are made by the attribute identification module as well as a multi-branch voting mechanism to obtain the final PAR prediction results.

2.2. Multi-scale Information Encoding Module

Multiscale information[25][26] helps to resolve ambiguous situations and produces more effective classification. ASPP[26][27] proposes to concatenate feature maps generated by dilation convolution with different dilation rates so that the neurons in the output feature map contain multiple receptive field sizes, and this size information encodes multiscale information and achieves improved performance. ASPP is built on atrous convolution based multi-scale feature generation approach, multiple dilation convolution features with different dilation rates generate the final feature representation. Although ASPP can effectively generate multiscale features, its feature resolution on the scale axis is not dense enough for specific scenes. In addition, in PAR scenes, some attribute feature information of people varies widely in scale, which requires that the multi-scale information needs to be encoded correctly for the high-level feature representation in the scene.

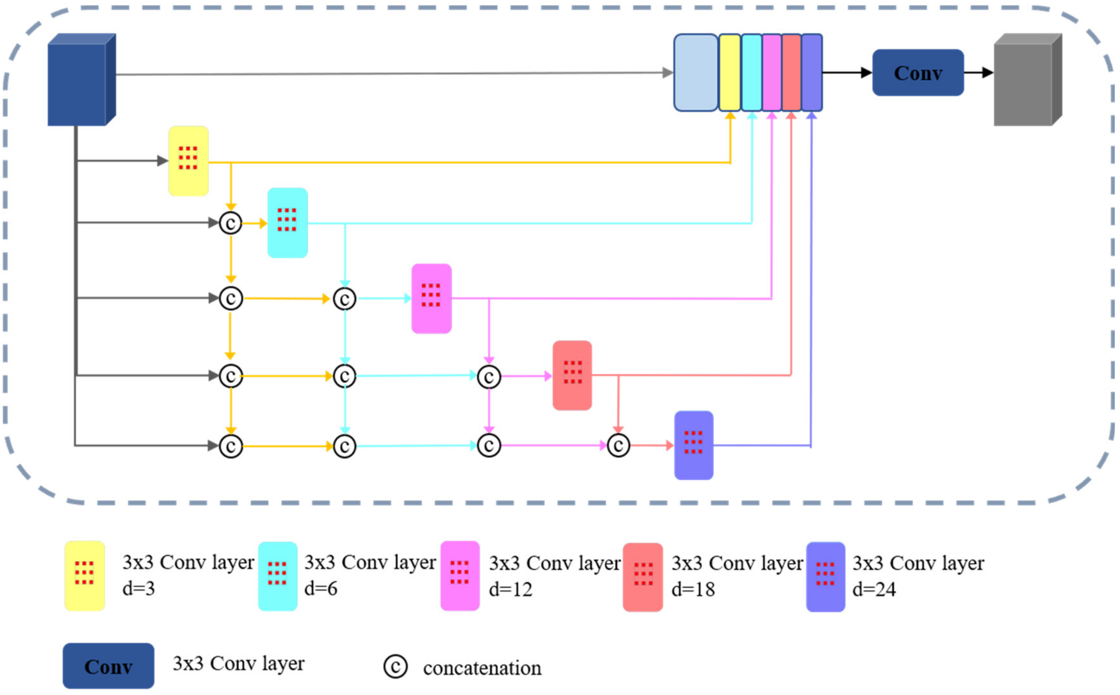


Figure 3. Multi-scale Information Encoding Module.

Consequently, DenseASPP[24] for multi-scale information extraction and processing is beneficial for further feature fusion afterward. As shown in figure 3, DenseASPP consists of multiple dilation convolutional layers with different dilation rates and a channel compression layer. The output of each dilation convolutional layer is deeply joined to all the unexecuted dilation convolutional layers. In MSEM, each dilation convolutional layer uses dilation rates of 3, 6, 12, 18, and 24, respectively. To reduce the redundant information in the network and for the next step of feature fusion on different scale spaces, the feature maps compression channel follows the DenseASPP. The specific processing operation is mainly activated by 3×3 convolution, BatchNorm2d[36], and Relu functions, the number of channels is finally set to 256.

2.3. Feature Fusion Module

In convolutional neural networks, the deep layer network[19] is robust in semantic information representation and can obtain rich semantic information, while the shallow layer network has a smaller receptive domain and can capture rich details. By fusing features from different scale spaces and using different levels of features, attribute associations between features from different scale spaces can be obtained.

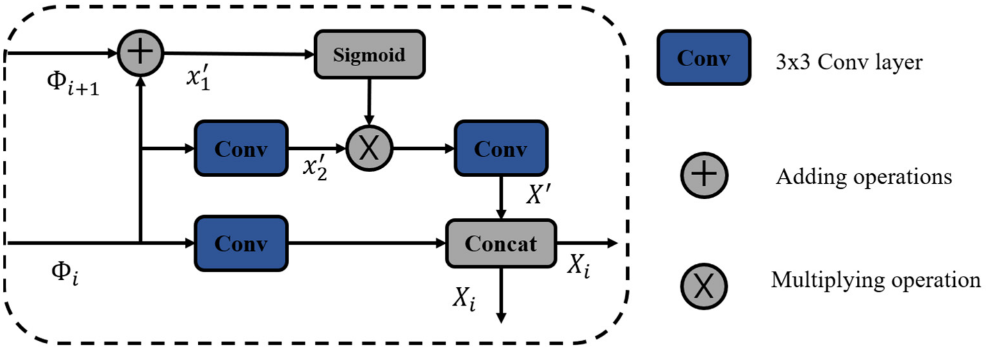


Figure 4. Feature Fusion Module.

FFM is shown in figure 4. In the two adjacent layers of scale features, the deeper feature maps are summed and fused with the shallow features and finally activated by the sigmoid function as follows:

$$x_1' = \sigma(\Phi_i + \Phi_{i+1}) \tag{1}$$

where x_1' is the weight established by feature addition and fusion. σ is the sigmoid activation function. $+$ represents the add operation.

The shallow features adjacent to it are then passed through a convolution block to extract the information, which consists of a 3×3 convolution, a normalization and an activation function, with BatchNorm2d chosen for the normalization and ReLU for the activation function:

$$x_2' = f_2(\Phi_i) \tag{2}$$

where f_i stands for convolutional block operation. The above obtained x_2' is multiplied with the weight vector x_1', followed by a convolution block to obtain the reweighted feature X':

$$X' = f_3(x_1' * x_2') \tag{3}$$

At the end, the newly obtained features X' are stitched and fused with the features extracted from the original information of the shallow feature map in the channel dimension to obtain our final desired feature output X_i:

$$X_i = f_1(\Phi_i) \oplus X' \tag{4}$$

where $\oplus$ is the concatenation.

2.4. Attribute Information Interaction Module

Based on the current research, there are many methods to improve the remote dependency modeling capability in convolutional neural networks. One is to introduce attention mechanisms, such as self-attention[28] mechanism, non-local module[29], or Criss-Cross Attention[30], etc., which have the disadvantage of consuming a large amount of memory to compute a large affinity matrix for each spatial location. From the convolutional viewpoint, it is also possible to use dilation convolution or deep separable convolution. However, since these methods probe the input feature map within a square window, this limits flexibility in capturing contextual information in realistic scenes.

In this paper, the strip pooling method is used to establish the correlation of attribute information for pedestrians. As described in the paper[31], each positioning element of the output tensor establishes a correlation with all the position elements that have the same horizontal or vertical coordinates as it. Based on the proposed aggregation mechanism, the input tensor of each position establishes a relationship with other positions, thus outputting a feature tensor with contextual information established. However, the number of parameters in this method is significantly reduced compared to other methods[29][30].

3. Experiment results and analysis

3.1. Data set

To validate the effectiveness of the proposed model, experiments were conducted on three public datasets, PETA, RAP, and PA-100K, respectively.

There are 19,000 images in the PETA[21] dataset containing 8705 pedestrians with resolutions ranging from 17×39 to 169×365. Each pedestrian in this dataset is annotated with 65 attributes (61 binary attributes and 4 multi-category attributes). In experiments, the 19,000 images were divided into 11400 and 7600 for training and testing respectively.

The RAP[22] dataset contains 41585 images collected from real indoor surveillance cameras with resolutions ranging from 36×92 to 344×554. Each image sample contains 72 attributes (69 binary attributes and 3 multiclass attributes). In experiments, 33,268 images were used for training and 8,317 images were used for testing.

PA-100K[12] is a large pedestrian attribute dataset with a total of 100,000 images collected from outdoor surveillance cameras, which is labeled by 26 binary attributes. In experiments, it is divided into 90,000 images for the training set and 10,000 images for the test set.

3.2. Evaluation metrics

For the evaluation of the selected datasets PETA, RAP, and PA100k, two types of evaluation metrics were used, label-based and sample-based.

For label-based evaluation, the mean accuracy (mA) is adopted. For each attribute, the accuracy is calculated for all samples (both positive and negative) and then the average of all attributes is calculated to obtain the mA. The evaluation metric mA is shown below:

$$mA = \frac{1}{2N}\sum_{i=1}^{m}\left(\frac{TP_i}{P_i} + \frac{TN_i}{N_i}\right) \tag{5}$$

where m is the number of attributes and N is the numbers of samples. P_i and N_i are the numbers of positive and negative cases for the i-th attribute. TP_i and TN_i are the numbers of positive and negative cases correctly predicted for the i-th attribute.

For the sample-based evaluation, we use four widely used metrics, including accuracy, precision, recall, and F1 value, defined as follows.

$$accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|Y_i \cup f(x_i)|} \tag{6}$$

$$precision = \frac{1}{2N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|f(x_i)|} \tag{7}$$

$$recall = \frac{1}{2N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|Y_i|} \tag{8}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{9}$$

where N is the number of samples, Y_i denotes the true positive label of the i-th sample, $f(x_i)$ returns the predicted positive label of the i-th sample, and $|\cdot|$ denotes the set base.

3.3. Comparison experiments

3.3.1. Comparison with state-of-the-art methods.

To demonstrate the effectiveness of RMFA, the performance of RMFA was compared with that of several state-of-the-art networks, such as ACN[32], DeepMar[9], PGDM[10], LG-Net[11], HPNet[12], VeSPA[13], JRL[15], MT-CAS[33], MTA-Net[34], WPAL[17], ALM[35], etc. The experimental results on the PETA and RAP datasets are shown in table 1 and on the PA100k dataset in table 2. The evaluation criteria mA, accuracy, precision, recall, and F1 are listed. The advantages of the RMFA can be clearly understood in table1 and table 2.

Table 1. Comparison with state-of-the-art models on PEAT and RAP datasets.

Methods	PETA					RAP				
	mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1
ACN[35]	81.15	73.66	84.06	81.26	82.64	69.66	62.61	80.12	72.26	75.98
DeepMar[9]	82.89	75.07	83.68	83.14	83.41	73.79	62.06	74.92	76.21	75.56
PGDM[10]	82.97	78.08	86.86	84.68	85.76	74.31	64.57	78.86	75.90	77.35
LG-Net[11]	-	-	-	-	-	78.68	68.00	**80.36**	79.82	80.09
HPNet[12]	81.77	76.13	84.92	83.24	84.07	76.12	65.39	77.33	78.79	78.05
VeSPA[13]	83.45	77.73	86.18	84.81	85.49	77.70	67.35	79.51	79.67	79.59
JRL[15]	85.67	-	86.03	85.34	85.42	77.81	-	78.11	78.98	78.58
MT-CAS[33]	83.17	78.78	**87.49**	85.35	86.41	-	-	-	-	-
MTA-Net[34]	84.62	78.80	85.67	86.42	86.04	77.62	67.17	79.72	78.44	79.07
WPAL[17]	85.50	76.98	84.07	85.78	84.90	81.25	50.30	57.17	78.39	66.12
ALM[35]	86.30	**79.52**	85.65	88.09	**86.85**	81.87	68.17	74.71	**86.48**	80.16
RMFA(ours)	**86.35**	79.28	85.16	**88.55**	86.55	**83.29**	**69.11**	77.58	85.23	**80.93**

While the performance of ALM on the PETA dataset is close to that of RMFA, ALM underperforms RFMA on both the RAP and PA100k datasets, which demonstrates the better generalization capability of RFMA. In addition, due to the relatively lower resolution of the images in the PA100k dataset, it can better present the Motion blur phenomenon in real scenes. As shown in table 2, the excellent performance of RMFA on the PA100k dataset shows the effectiveness of expanding the receptive field and fusing multi-scale information. While the PETA and RAP datasets are labeled with more attributes compared to the PA100k dataset, the data results in the table also show the necessity and effectiveness of excavating global contextual information and feature relevance.

Table 2. Comparison with state-of-the-art models on PA100k datasets.

Methods	PA100k				
	mA	Accu	Prec	Recall	F1
DeepMar[9]	72.70	70.39	82.24	80.42	81.32
PGDM[10]	74.95	73.08	84.36	82.24	83.29
LG-Net[11]	79.96	75.55	86.99	83.17	85.04
HPNet[12]	74.21	72.19	82.97	82.09	82.53
VeSPA[13]	76.32	73.00	84.99	81.49	83.20
MT-CAS[33]	77.20	78.09	**88.46**	84.86	**86.62**
ALM[35]	80.68	77.08	84.21	88.84	86.46
RMFA(ours)	**81.60**	**78.40**	85.16	**88.85**	86.55

3.3.2. Attribute Recognition Comparison.

Among all evaluation metrics, mA is one of the vital metrics in model evaluation. As an example, on the PETA dataset, the mA values of 35 attributes are plotted based on the test results of our model and StrongBaseline, as shown in figure 5. It can be seen from figure 5 that RMFA improves the recognition of almost all attributes. In the case of “upper Body Thin Stripes”, “footwear Sandals”, “accessory Sunglasses”, “upper Body V Neck” and other fine-grained attributes have improved significantly, which may attribute to the expansion of the receptive field and the integration of multi-scale information. The improvement in high-resolution attributes such as “lower Body Casual” may attribute to attribute relevance and contextual information excavating.

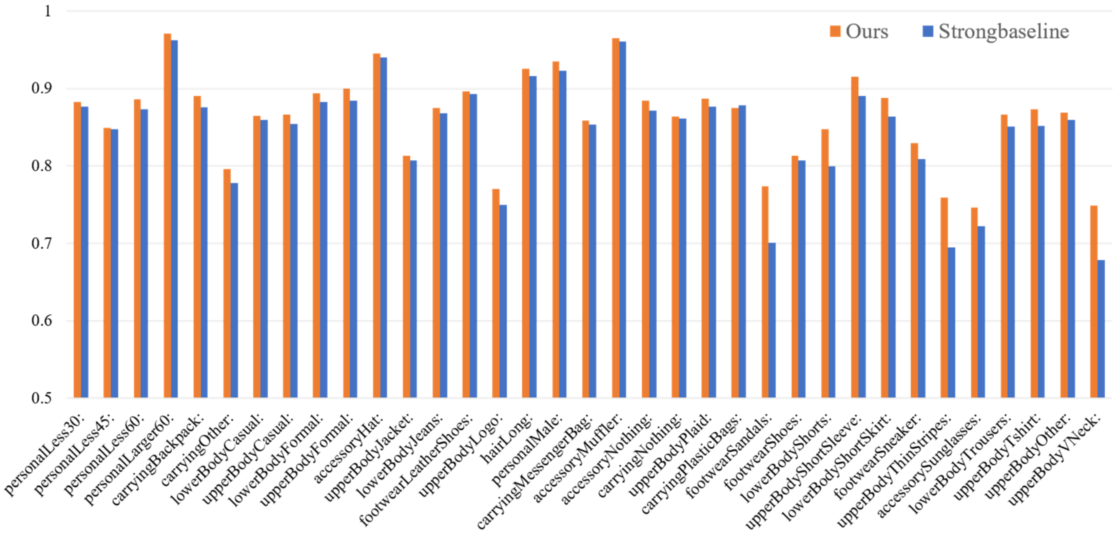


Figure 5. Comparison of attribute recognition.

In addition, four sample maps were randomly selected from the PETA dataset for recognition prediction. As shown in figure 6, where the histogram is a comparison of the prediction results of RMFA with StrongBaseline. The horizontal axis of the histogram is the predicted attribute of this sample, and the vertical axis is the probability value of the predicted attribute. The prediction performance improvement of RMFA can be visually shown in figure 6.

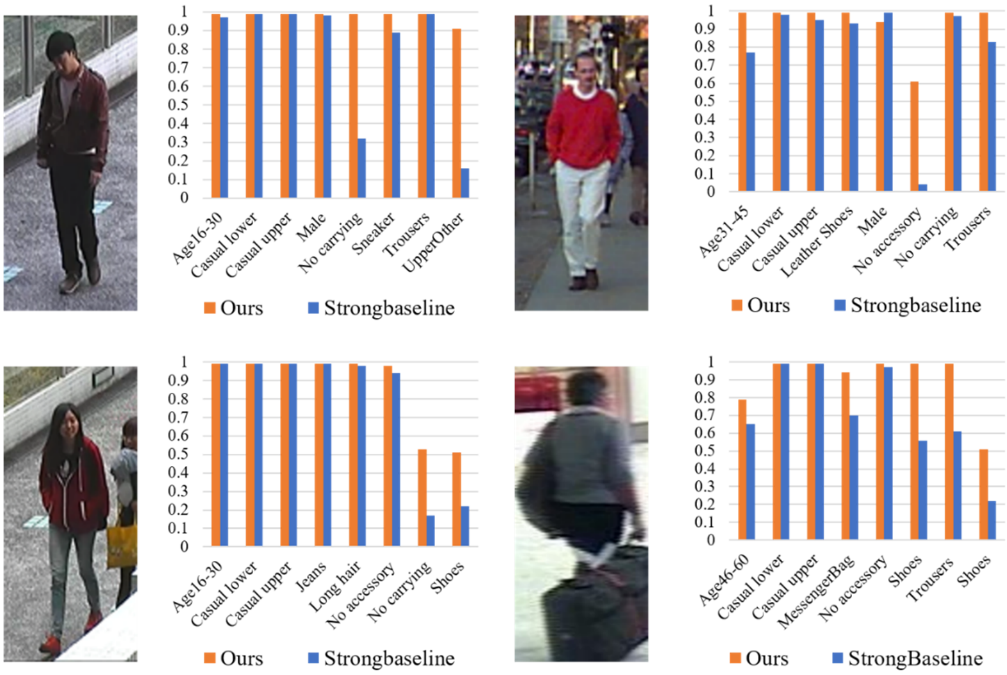


Figure 6. Recognition probability of RMFA and StrongBaseline on samples.

3.4. Ablation experiments

To analyze the effectiveness of key components of the RMFA network architecture, we conduct ablation experiments on the PETA dataset. The evaluation metrics for this experiment still use the five important metrics mentioned above, including mA, accuracy, precision, recall, and F1. As shown in the table, the first row of data in the table indicates the architecture based on our architecture (RMFA) with the MSEM module removed, the second row of data indicates the architecture with the FFM module removed, and the third row of data indicates the architecture with the AIIM removed. As the data in table 3 shows, four of the five evaluation metrics have the highest ranking in proposed complete architecture. This demonstrates the effectiveness and importance of the MSEM module, FFM module, and AIIM module in the overall network architecture.

Table 3. Ablation experiments on the PETA dataset.

	mA	Accu	Prec	Recall	F1
-MSEM	85.90	79.26	85.01	88.68	86.53
-FFM	85.72	73.4	75.84	**93.54**	82.97
-AIIM	84.61	78.17	84.51	87.45	85.68
RMFA	**86.35**	**79.28**	**85.16**	88.55	**86.55**

In addition, in this ablation experiment, a comparison with the original overall architecture was plotted based on the mA values of the properties of the different ablation experimental architectures. As shown in figure 7, figure 8, and figure 9, the comparison plots of attribute recognition with the MSEM module removed, the FFM module removed, and the AIIM module removed are shown respectively. As shown in figure 7, recognition rates of attributes such as “upper Body Logo”, “footwear Sandals”, “upper body pinstripe”, and “upper Body V Neck” decreased significantly after removing the MSEM module, and most other attributes also showed a minor decrease in recognition rates. This shows our previous consideration about the effective impact of expanding the receptive field. As shown in figure 8, after removing the FFM module, the recognition rate decreases significantly for the attributes “accessory Nothing”, “carrying Nothing”, “upper Body Thin Stripes”, etc. This experiment shows that the fusion of multi-scale information is effective for fine-grained attribute recognition. As shown in figure 9, after removing the AIIM module, we can see that the fusion of multi-scale information is effective in identifying attributes such as “Age”, “Casual”, “Formal”, “Personal Male”, etc. “Personal Male” and some other attributes that require global information judgment produces a decrease in recognition. Also, the decrease in the recognition rate of some other attributes shows that obtaining global contextual information and attribute relevance also affects the recognition of other attributes to different degrees, and this shows the effectiveness of the AIIM module for the overall architecture.

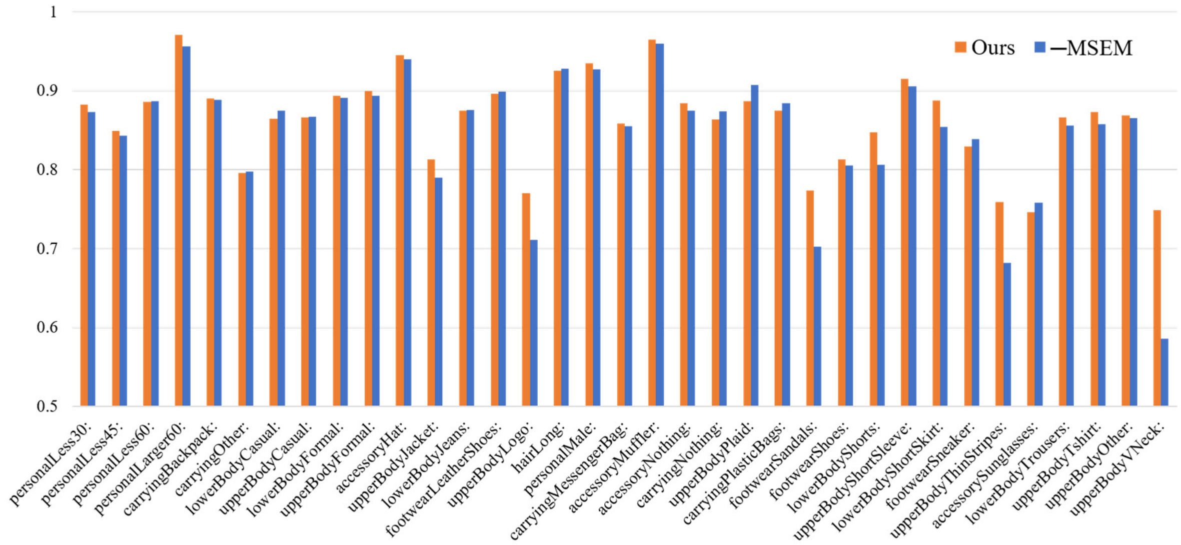


Figure 7. Comparison of attribute predictions on ablation experiments with MESM removed.

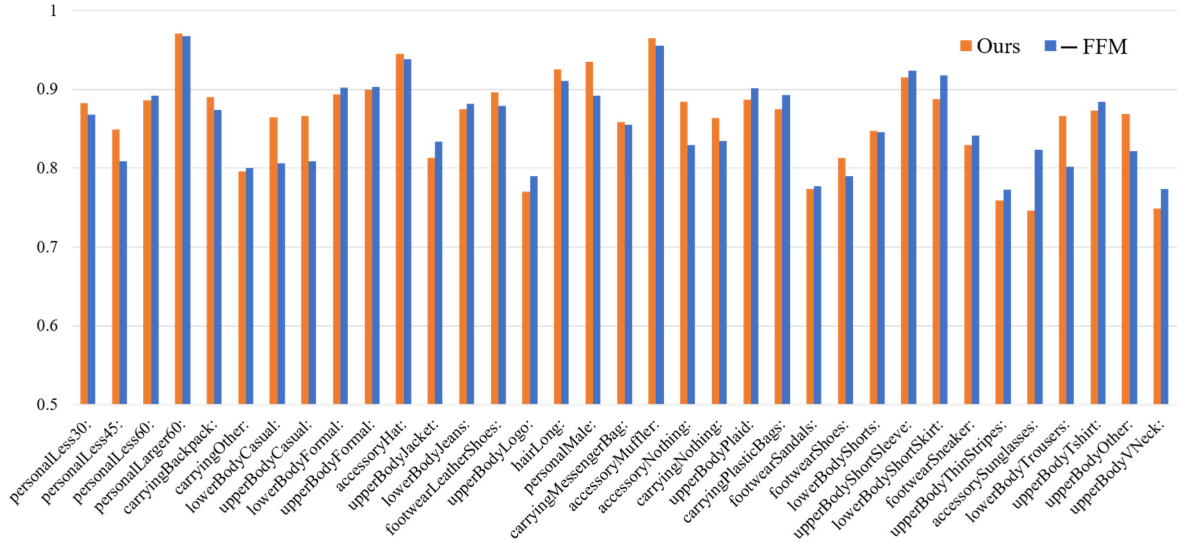


Figure 8. Comparison of property predictions on ablation experiments with FFM removed.

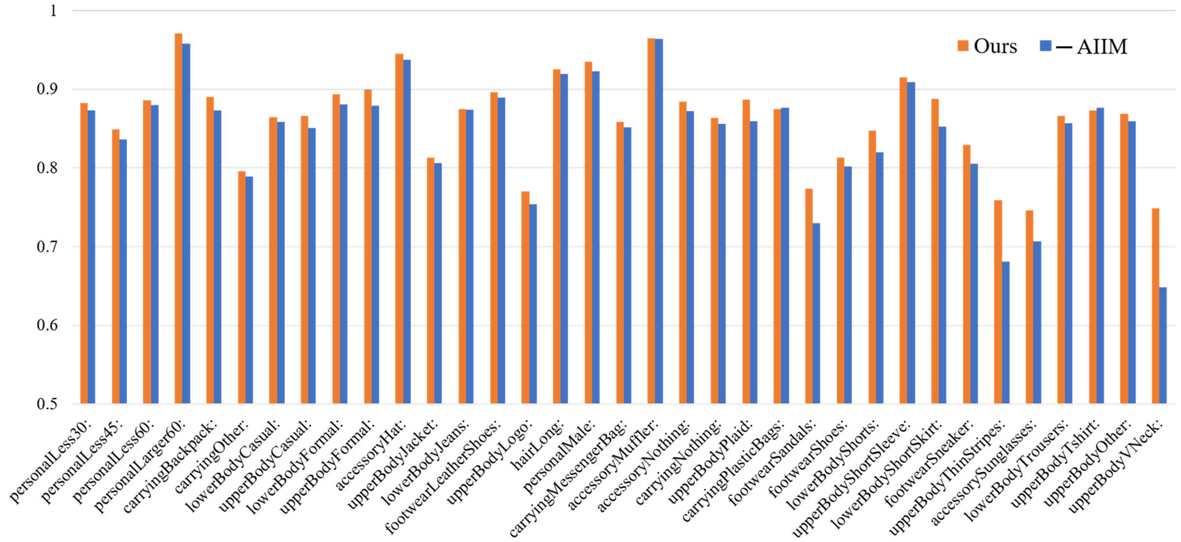


Figure 9. Comparison of property predictions on ablation experiments with AIIM removed.

4. Conclusion

In this paper, a pedestrian attribute recognition algorithm based on a combination of extracting and fusing multi-scale features from an expanded receptive field and extracting attribute correlations is proposed. The MSEM module is used to encode the expanded receptive field and multi-scale information. In the FFM module, the feature pyramid construct is used to fuse different feature information from different scales. And finally, the AIIM module is used to obtain contextual information and attribute relevance. Extensive experiments are conducted on PEAT, RAP, and PA100k datasets and achieved excellent results. In addition, the ablation experiments demonstrate the effectiveness of the key blocks in RMFA. The performance of RMFA is significantly improved due to the cooperation between different modules. In the future, further exploration of the relationship between multiple attributes to optimize the effectiveness of the AIIM modules can be taken into consideration.

5.Acknowledgment

This work was supported by the Foundation of Jiangxi Educational Committee (No. GJJ200824) and the Natural Science Foundation of Jiangxi Province (Grant No. 20202BAB212003).

6.References

[1] Wang X, Zheng S, Yang R, Zheng A, Chen Z, Tang J and Luo B (2022). Pedestrian attribute recognition: A survey. *Pattern Recognition, 121*, 108220.

[2] Liu W, Liao S, Ren W, Hu W and Yu Y (2019). 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).

[3] Brunetti A, Buongiorno D, Trotta G F and Bevilacqua V (2018). Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing, 300,* 17-33.

[4] Tay C P, Roy S and Yap K H (2019). Aanet: Attribute attention network for person re-identifications. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7134-7143).

[5] Lin Y, Zheng L, Zheng Z, Wu Y, Hu Z, Yan C and Yang Y (2019). Improving person re-identification by attribute and identity learning. *Pattern Recognition, 95,* 151-161.

[6] Shi Y, Ling H, Wu L, Shen J and Li P (2020). Learning refined attribute-aligned network with attribute selection for person re-identification. *Neurocomputing, 402,* 124-133.

[7] Li D, Zhang Z, Chen X and Huang K (2018). A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE transactions on image processing, 28*(4), 1575-1590.

[8] Sun Y, Zheng L, Deng W and Wang S (2017). Svdnet for pedestrian retrieval. In *Proceedings of the IEEE international conference on computer vision* (pp. 3800-3808).

[9] Li D, Chen X and Huang K (2015, November). Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)* (pp. 111-115). IEEE.

[10] Li D, Chen X, Zhang Z and Huang K (2018, July). 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.

[11] Liu P, Liu X, Yan J and Shao J (2018). Localization guided learning for pedestrian attribute recognition. *arXiv preprint arXiv:1808.09102.*

[12] Liu X, Zhao H, Tian M, Sheng L, Shao J, Yi S and Wang X (2017). Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision* (pp. 350-359).

[13] Sarfraz M S, Schumann A, Wang Y and Stiefelhagen R (2017). Deep view-sensitive pedestrian attribute inference in an end-to-end model. *arXiv preprint arXiv:1707.06089.*

[14] Guo H, Fan X and Wang S (2017). Human attribute recognition by refining attention heat map. *Pattern Recognition Letters, 94,* 38-45.

[15] Wang J, Zhu X, Gong S and Li W (2017). Attribute recognition by joint recurrent learning of context and correlation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 531-540).

[16] Zhao X, Sang L, Ding G, Han J, Di N and Yan C (2019, July). Recurrent attention model for pedestrian attribute recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 9275-9282).

[17] Yu K, Leng B, Zhang Z, Li D and Huang K (2016). Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. *arXiv preprint arXiv:1611.05603.*

[18] Luo W, Li Y, Urtasun R and Zemel R (2016). Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems, 29.*

[19] Lin T Y, Dollár P, Girshick R, He K, Hariharan B and Belongie S (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).

[20] Honari S, Yosinski J, Vincent P and Pal C (2016). Recombinator networks: Learning coarse-to-fine feature aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5743-5752).

[21] Deng Y, Luo P, Loy C C and Tang X (2014, November). Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 789-792).

[22] Li D, Zhang Z, Chen X, Ling H and Huang, K (2016). A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054.*

[23] Xie S, Girshick R, Dollár P, Tu Z and He K (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492-1500).

[24] Yang M, Yu K, Zhang C, Li Z and Yang K (2018). Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3684-3692).

[25] Zhao H, Shi J, Qi X, Wang X and Jia J (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881-2890).

[26] Chen L C, Papandreou G, Schroff F and Adam H (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587.*

[27] Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915, 2016.*

[28] Zhao H, Jia J and Koltun V (2020). Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10076-10085).

[29] Wang X, Girshick R, Gupta A and He K (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794-7803).

[30] Huang Z, Wang X, Huang L, Huang C, Wei Y and Liu W (2019). Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 603-612).

[31] Hou Q, Zhang L, Cheng M M and Feng J (2020). Strip pooling: Rethinking spatial pooling for scene parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4003-4012).

[32] Sudowe P, Spitzer H and Leibe B (2015). Person attribute recognition with a jointly-trained holistic cnn model. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 87-95).

[33] Zeng H, Ai H, Zhuang Z and Chen L (2020, July). 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.

[34] Di X, Zhang H and Patel V M (2018, October). 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.

[35] Tang C, Sheng L, Zhang Z and Hu X (2019). 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).

[36] Ioffe S and Szegedy C (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). PMLR.

[37] Zhong J, Qiao H, Chen L, Shang M and Liu Q (2021, July). Improving Pedestrian Attribute Recognition with Multi-Scale Spatial Calibration. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

[38] Jia J, Huang H, Yang W, Chen X and Huang K (2020). Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method. *arXiv preprint arXiv:2005.11909.*