Computer Science and Information Systems 21(4):1567–1582 https://doi.org/10.2298/CSIS231015048L

ALFormer: Attribute Localization Transformer in Pedestrian Attribute Recognition

Yuxin Liu, Mingzhe Wang, Chao Li, and Shuoyan Liu*

Institute of Computing Technology, China Academy of Railway Sciences Corporation Limited
Beijing, China
{yxin_l,zzh355997367,lchao1234}@163.com
06112062@bjtu.edu.cn

Abstract. Pedestrian attribute recognition is an important task for intelligent video surveillance. However, existing methods struggle to accurately localize discriminative regions for each attribute. We propose Attribute Localization Transformer (ALFormer), a novel framework to improve spatial localization through two key components. First, we introduce Mask Contrast Learning (MCL) to suppress regional feature relevance, forcing the model to focus on intrinsic spatial areas for each attribute. Second, we design an Attribute Spatial Memory (ASM) module to generate reliable attention maps that capture inherent locations for each attribute. Extensive experiments on two benchmark datasets demonstrate state-of-the-art performance of ALFormer. Ablation studies and visualizations verify the effectiveness of the proposed modules in improving attribute localization. Our work provides a simple yet effective approach to exploit spatial consistency for enhanced pedestrian attribute recognition.

Keywords: spatial attention, attribute localization, contrast loss, random mask.

1. Introduction

Due to the constant demand for intelligent video surveillance and psychosocial behavior research, Pedestrian Analysis Recognition (PAR) is a long-standing research topic.Compared to low-level features, pedestrian attributes can be considered high-level semantic information and are more robust against changes in lighting conditions and viewpoint. Therefore, PAR is widely applied in the tasks of person re-identification, face verification, and human identity recognition.

With the development of deep learning technology, PAR has made significant progress in the past decade. Existing methods[1–4] mainly use complex networks such as Feature Pyramid Network (FPN) to enrich attribute representation from multilevel feature maps and combine attention mechanisms to accurately locate attribute-related regions. However, these CNN-based methods emphasized the extraction of local features, it is difficult to capture long-range dependencies.

In recent years, vision transformers such as ViT [5], DeiT[6]and Swin Transformer[7] have been widely used in various computer vision tasks, which combine the learning methods of natural language processing and computer vision. Using the multi-head self-attention module to focus on multiple areas at the same time, transformer-based models

[*] Corresponding author

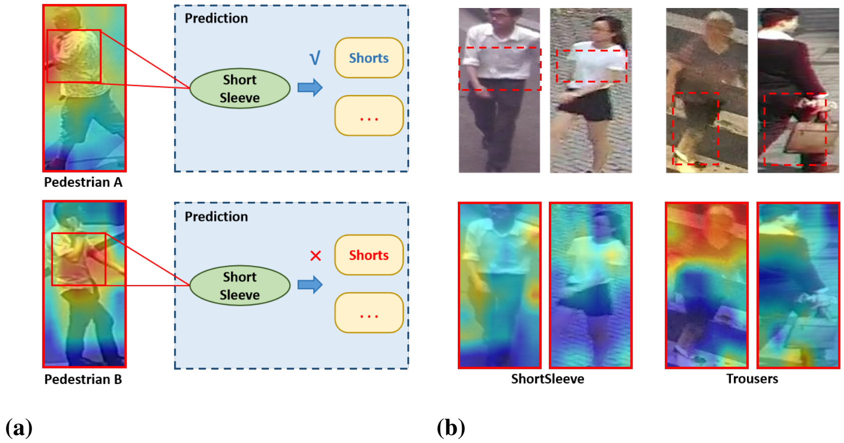

Fig. 1. Illustration of the problem of predicting by inter-attribute correlation and our main hypothesis on attribute localization.In(a), Class Activation Maps (CAM)[8] (with red boundary) of the baseline method in "short sleeve" and "trousers" attributes of the PA100K are visualized in the second row. Attribute-related regions of each attribute are plotted by the red dotted frame in the first row. In (b), it mainly expresses the process of Swin Transformer utilizes correlation between attributes for prediction, where the green attributes indicate the attributes that Swin Transformer focus on, and the yellow attributes indicate the attributes that Swin Transformer predicts based on the correlation

can capture long-range dependencies and retain more detailed information, which makes them suitable for extracting rich structural features. However, they mainly complete feature extraction by modelling the fine-grained relevance of images. In the task of PAR, it is manifested as over-reliance on attribute relevance for prediction, thereby reducing the accuracy of attribute positioning. As shown in Figure 1(a), for pedestrian A, Swin Transformer focuses on the short sleeve attribute region and can identify pedestrian A as a person with shorts based on its correlation to the short sleeve. However, for pedestrian B with trousers instead of shorts, the correlation between the short sleeve and shorts does not exist. In this case, Swin Transformer still tends to use the prior correlation to predict via focusing on the short sleeve attribute region, which results in incorrectly attribute localization and predicting the trousers as shorts attribute for pedestrian B. Furthermore, recognition errors may occur for other attributes such as ’handbag’, ’female’, and ’long hair’. Although there is a significant correlation between these attributes, it is not absolute. In the PA100k [1] dataset, a considerable proportion of males carry handbags. Therefore, conventional Swin Transformer-based methods struggle to achieve satisfactory recognition results for these attributes due to incorrect attribute relevance. This highlights the baseline network’s over-reliance on attribute relevance. So correct semantical correlation (such as gender attributes and skirt attributes) is critical between attributes [9].

Additionally, we hypothesize that the spatial location of the same attribute mostly appears at the similar position between different pedestrian images. For example, the short sleeve attribute and the trousers attribute mostly appear in the upper and lower part of the picture, respectively, which is shown in the first row of Figure 1(b). However, we observe that Class Activation Maps (CAM) [8] of the same attribute generated by the baseline method exhibit significant location variations. Some examples are shown in the second row of Figure 1(b). Considering two aspects, we propose Attribute Localization Transformer (ALFormer). Specifically, ALFormer uses attribute spatial memory(ASM) module that generates reliable and stable attribute-related regions to suppress position shifts caused by over-fitting or label noise of single image. Then ASM extracts inherent information based on accurate attribute positions, which assists to extract more comprehensive and represent attribute spatial features. Experimental evaluations results show that the proposed method achieves 1.79% and 2.34% improvements on widely adopted public datasets PA100k and PETA..

We make the following four contributions in this works:

- The Attribute Localization Transformer (ALFormer), a novel Transformer framework is proposed to extract comprehensive attribute spatial information for pedestrian attribute recognition.
- Mask Contrast Learning (MCL) scheme is introduced to suppress relevance to locate more precise region for each attribute.
- The Attribute Spatial Memory (ASM) module is designed to generate exact spatial attention regions and extract discriminative semantic features for each attribute.
- The experiments on two datasets (PETA and PA100K) show that our framework has excellent performance compared with several other state-of-the-art methods. In addition, we also present visualization results to verify the effectiveness of the proposed framework.

2. Related Works

Due to the constant demand for intelligent video surveillance and psychosocial behavior research, pedestrian attribute recognition is an important and challenging task.

Early methods [10–12] usually rely on hand-crafted features and learn individual classifiers for each attribute. Kumar et al. [11] extracted HOG and color histograms of important facial feature areas and then trained multiple support vector ma-chines for facial attribute recognition. Bourdev et al. [12] developed a three-layer SVM method to extract high-level semantic information. However, it cannot effectively utilize the potential relevance between attributes for separated classifiers.

Recently deep learning is well incorporated into PAR from a variety of studies. Liu et al. [13] designed a deep CNN composed of LNet and ANet to locate faces and predict facial attributes. Abdulnabi et al. [14] proposed a joint multi-task learning method to achieve information sharing between attributes for clothing attribute prediction. Hand and Chellappa [15] adopted a multi-task learning CNN method based on a grouping scheme to classify facial attributes. Han et al. [16] designed a Deep Multi-Task Learning CNN method (DMTL) that learns shared features for all attributes and specific category features for heterogeneous attributes. Li et al. [17] proposed a landmark-free facial attribute prediction method that does not rely on landmark an-notations. Li et al. [18] introduced pedestrian structural knowledge into pedestrian at-tribute recognition and proposed a posture-oriented pedestrian attribute deep prediction model.

Based on global feature similarities, the JRL network [19] uses Long Short-Term Memory [20] to treat the pedestrian attribute recognition task as a sequence prediction problem. Attention mechanisms [1, 2, 18, 21, 22] are widely used in pedestrian attribute recognition to locate attribute-related regions and learn discriminative feature representations. Tan et al. [21] proposed a pedestrian attribute analysis method based on parsing attention, label attention, and spatial attention. The multi-directional attention module HydraPlus-Net [1] extracts pixel-level features and semantic-level features to locate fine-grained attributes. Liu et al. [2], based on CAM [8] and EdgeBox [23], proposed a Localization Guided Network method to extract attribute-related local features. The PGDM [18] framework uses a pre-trained human soft tube estimator and Spatial Transformer Networks (STNs) [24] to generate reliable attribute-related regions.

In recent years, Transformers have played an extremely important role in deep learning. It is a model proposed by Vaswani et al. [25] for natural language processing (NLP) tasks with far-reaching impact. The Vision Transformer (ViT) proposed by Dosovitskiy et al. [5] achieved breakthrough applications in the field of computer vision, combining learning methods from natural language processing and computer vision, using multi-head self-attention modules to simultaneously focus on multiple regions, capture long-range dependencies and retain more detailed information, thereby extracting rich structural features to achieve higher performance improvements. DeiT proposed by Touvron et al. [6] uses knowledge distillation to guide transformers to learn more effectively so that the network no longer needs large-scale pre-training data. To reduce computation of ViT, Liu et al. proposed the Swin Transformer [7], which introduces a shift-window-based attention mechanism to reduce computational costs. Transformer models achieve competitive performance in different computer vision tasks. However, such methods usually rely too much on predicting image region relationships, which leads to biased learning of attribute position information in PAR.

As stated, previous methods generally considered aspects such as global or local pedestrian images, attention mechanisms, and attribute correlation but these methods only considered a single image and could not fully exploit the spatial information of each pedestrian attribute. In contrast, our method makes full use of the potential attribute spatial information among different images. With proposed ASM and MCL modules, framework can better focus on attribute region and in turn obtain robust performance.

3. Attribute Localization Transformer (ALFormer)

This section first describes the overall architecture of Attribute Localization Transformer (ALFormer). After that, we present an attribute joint learning framework composed of two parts: an attribute spatial module (ASM) module and a mask contrast learning (MCL) scheme that incorporates Random Mask and contrast loss. Finally, we present the detail of the employed loss function for ALFormer.

3.1. Overall Architecture

Figure 2 shows the overall structure of the proposed ALFormer framework. Our basic network is identical to the four stages of Swin Transformer. Given a 2D image, it uniformly splits the input image into a sequence of 2D non-overlapping flattened patches. However, studies have shown that basic network over-reliance on attribute relevance for prediction, thereby reducing the accuracy of attribute positioning. This paper proposes MCL

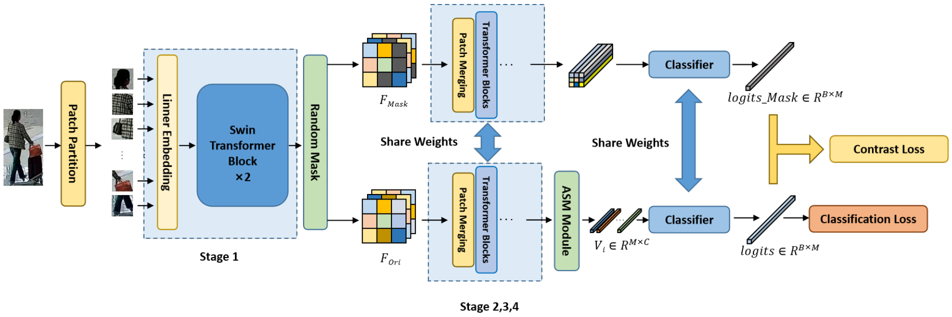

Fig. 2. Overall framework architecture

scheme. It first introduces a random mask module that obtains feature map from stage 1 and generates two results: F_{mask} and F_{Ori} .F_{mask} (masked feature map) is fed into the rest stages and classifier to obtain prediction $logits\_Mask$. F_{Ori}(original input feature map) is put into stage 2,3,4 and ASM module to generate prediction $logits$. ASM module generates spatial attention map and embedded vector V_i for each attribute. ALFormer calculates contrast loss between $logits$ and $logits\_Mask$. In addition, classification loss is calculated by $logits$. Finally, ALFormer outputs the predicted score on attributes.

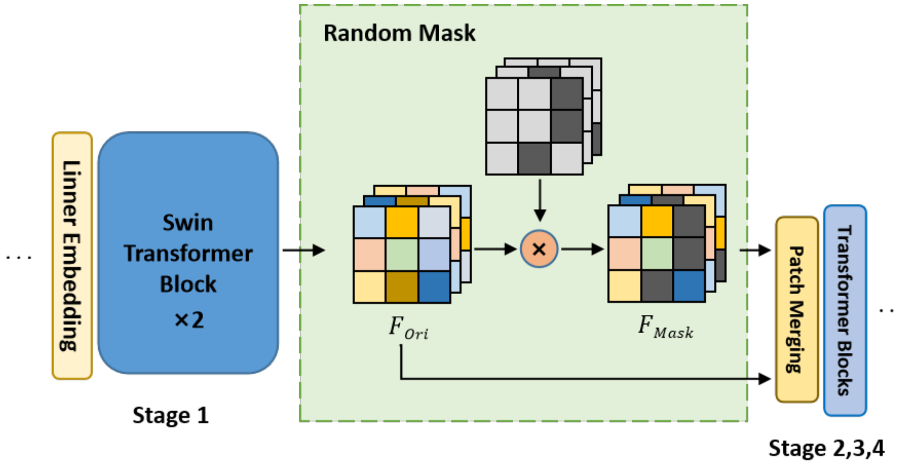

Fig. 3. MCL scheme structure. We introduce Random Mask (RM) on the feature map of backbone stage 1, and calculate the contrast loss with the original feature prediction

3.2. Mask Contrast Learning Scheme

As illustrated in Figure 2, the summation of ASM module and MCL scheme is to capture the spatial information of pedestrian attributes. Different spatial regions in input image are associated with different attributes. It is desirable to focus on related regions when recognizing one attribute. However, we observe that Swin Transformer tends to focus on a relatively small number of local regions within which attributes may be correlated with other attributes, which leads Swin Transformer to predict attributes in those neglected regions based on such correlation. In fact, discriminative information may exist within these neglected regions, which is crucial for attribute identification. To address this problem, we propose MCL scheme which consists of Random Mask and contrast loss. It can force the framework to focus on intrinsic and discriminative attribute spatial regions for PAR tasks. The MCL scheme is illustrated in Fig.3

Different from existing random occlusion schemes, the proposed MCL scheme is performed at the feature level. The advantage is to reduce the correlation among feature regions and force the framework to focus on exact attribute spatial regions.

Specifically, random mask features are generated for each batch of input image at corresponding number. And then, Stage 1 of the framework outputs F_{Ori} in the spatial dimension of the feature map, as shown in the formula:

$$F_{mask} = RM(F_{Ori}) \tag{1}$$

A contrast loss function L_{con} is set at the classifier stage to evaluate the difference be-tween prediction results with and without Random Mask.As shown in the formula:

$$L_{con} = -pred_{Ori} \times \log(pred_{Mask}) \tag{2}$$

3.3. Attribute Spatial Memory Module

In this section, we propose the Attribute Spatial Memory(ASM) module to tackle the spatial attention region deviation problem. As shown in Figure 4, we visually demonstrate how to extract attribute spatial and semantic feature from ASM module. ASM module first takes input feature to generate attention map $A \in R^{B\times M\times H\times W}$, and then selects reliable ones and saves them in Memory module. Finally, it generates exact spatial attention regions and extract discriminative semantic features for each attribute.

In detail, ASM module takes feature map $F \in R^{B\times C\times H\times W}$ and classifier weight $w \in R^{M\times C}$ as inputs, where $F \in R^{B\times C\times H\times W}$ is the output of the backbone framework and H, W, C represent height,width and channel dimension of feature map.

Inspired by CAM[8], we introduce probability values as a criterion to construct spatial attention. As shown in the formula, ASM generates attention maps for each attribute by weighting the classifier:

$$Attention\_Map\ A_{i,m} = w_m \cdot F_i(x, y) \quad m = 0, 1, 2 \ldots, M-1 \tag{3}$$

where M is the number of attributes, F_i is the feature map of the backbone framework, w_m is the classifier weight of the m-th attribute.

After getting the spatial attention regions of each attribute, we adopt the selector which takes logits and ground truth labels as input to aggregate the attention maps of qualified positive samples of$m$-th attribute. The main reason is that some attributes among images are negative or low confidence values. As shown in the formula, ASM filters out negative samples to get high-confidence predictions:

$$Attention\_Map = \{A_{i,m} \mid Lable_{i,m} = 1, Logitis_{i,m} > coeff\_threshold\} \tag{4}$$

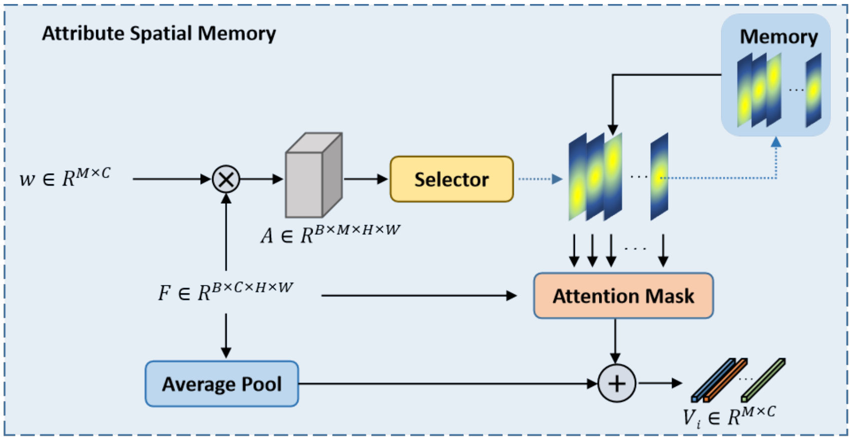

Fig. 4. Attribute Spatial Memory(ASM) module structure

Through strict selection, $A_{i,m}$can be regarded as reliable attribute spatial locations for m-th attribute. A criticism of this experimental design is that in actual training process, there are differences in attribute distribution of images in different batches. Moreover, all images in a batch may be negative on some attributes. Inspired by EMA strategy, we adopt a momentum-weighted scheme to update attention maps and save the reliable spatial location. In addition, for cases where all images in a batch are negative on some attributes, the corresponding attention map will not participate in updates. Specifically shown in the formula:

$$A_{s+1} = l \cdot A_{e+1} + (1-l) \cdot A_{e+1} \tag{5}$$

Where s is the number of training batches step. We save the generated attribute spatial attention map in memory module to assist subsequent feature pooling fusion.

The attention map $A_{i,m}$ indicates the contribution of different spatia regions to the prediction of attributes.. Therefore, ASM generates corresponding embedding vectors for different attributes separately, which weights the spatial location information. Thus improves spatial localization for each attribute. As shown in the formula, we normalize the spatial attention map and then use it as a spatial weighting coefficient to perform weighted pooling on the feature Feature:

$$Embed\_Attr_m = Attention\_Map(F_i) \quad m = 0, 1, 2 \ldots, M-1; i = 0, 1, 2 \ldots, B-1 \tag{6}$$

Where M is the number of attribute types and B is the number of batches. The semantic embedding vector $Embed\_Attr_m$ for each attribute is obtained through weighted pooling.

3.4. Loss Function

As illustrated in Figure 2, contrast loss L_{con} and classification loss L_{cls} are obtained from the ALFormer. The classification loss is calculated according to binary cross entropy loss, which is formulated as follow:

$$L_{cls} = \frac{1}{N}\sum_{i=1}^{N}(y_i \log p_i + (1 - y_i) \log (1 - p_i)) \tag{7}$$

The final loss function L is a weighted summation of the classification loss and contrast loss:

$$Loss = L_{cls} + \lambda \cdot L_{con} \tag{8}$$

Where $\lambda = 1$ is set as default in all experiments if not specially specified.

4. Experiments

4.1. Datasets and Evaluation Metrics

We conduct experiments on public datasets PETA [26] and PA100K [1]. Figure 5 shows some pedestrian image samples. The Pedestrian Attribute (PETA) dataset [26] is collected from 10 small-scale human datasets and consists of 19,000 human images, of which 9,500 images are used for the training set, 1,900 for the validation set, and 7,600 for the test set. Each image is labeled with 61 binary attributes and 4 multi-class attributes. We follow the commonly used experimental scheme [3, 4, 27] and only evaluate 35 attributes with a positive rate greater than 5%.

PA100K dataset is constructed from images taken by 598 real outdoor surveillance cameras and contains 100,000 pedestrian images with resolutions ranging from 50 × 100 to 758 × 454. It is the largest pedestrian attribute recognition dataset to date. The entire dataset is randomly divided into training set, validation set and test set in a ratio of 8:1:1. Each image in the dataset is labeled with 26 attributes, with labels of 0 or 1 indicating the presence or absence of the corresponding attribute.

Two metrics are used to evaluate attribute recognition performance [27], namely label-based metrics and four instance-based metrics. For label-based metrics, Deng et al. [27] use mean accuracy (mA) to evaluate attribute recognition algorithms. For each attribute, mA indicates the classification accuracy of positive and negative samples separately and then takes their average value as the recognition result of that attribute. Finally, the average value of all attributes is taken to obtain the recognition rate. Specifically, the experiment uses this indicator as part of the evaluation criteria as shown in the formula:

$$mA = \sum_{i}^{L}(\frac{TP_i}{P_i} + \frac{TN_i}{N_i}) \tag{9}$$

where L is the number of attributes. TP_i and TN_i are the number of correctly predicted positive and negative samples respectively, and P_i and N_i are the number of positive and negative samples respectively.

For instance-based metrics, we use Accuracy, Precision, Recall and F1-score.

Fig. 5. Public dataset image samples. (a) Pedestrian samples from the PA100k dataset; (b) Pedestrian samples from the PETA dataset

4.2. Implement Details

Our method is implemented based on the PyTorch framework and trained in an end-to-end manner. Since our work mainly focuses on how to use local features defined by attributes, such framework can improve spatial location for each attribute. In the experiment, we adopt Swin Transformer [7] as the backbone network to extract pedestrian image features for a fair comparison. Pedestrian images are uniformly adjusted to 256×192 as input. Random horizontal mirroring, padding and random cropping are used for enhancement. AdamW is used for training with a weight decay of 0.0005. The initial learning rate is 0.0001 and the batch size is set to 64. The total epoch of the training phase is 50. The momentum coefficient is 0.9998. The label smoothing coefficient is set to 0.2.

4.3. Baseline Method

Given a dataset, $D = \{(X_i, y_i) \mid i = 1, 2, \ldots, N\}$, PAR aims to predict multiple attribute $y_i \in \{0, 1\}$ to i-th pedestrian image, where N,M denotes the number of images and attributes respectively, zero and one in the attribute vector y_i indicate the absence and presence of the corresponding attributes of a pedestrian image. Following [3, 4, 27, 28], we formulate pedestrian attribute recognition as a multi-label classification task which adopts multiple binary classifiers with sigmoid functions [27]. Binary cross-entropy loss L_{cls} is used as the optimization target:

$$L_{cls} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{i,j} \log p_{i,j} + (1 - y_{i,j}) \log (1 - p_{i,j})) \tag{10}$$

where $p_{i,j}$ is the prediction probability of the classifier.

4.4. Ablation Analysis

In this section, we first study the impact of the ASM and MCL modules by conducting analysis experiments on both two datasets. The baseline methods use the original Swin Transform set only in section 4.3. As shown in Table 1, compared to the baseline method, we have conducted the following observations. First, when the ASM module is not used, the attribute positioning accuracy is worse and the recognition accuracy decreases. The results show that without correct attention regions, attribute semantic features lack discriminative ability and contain more noise. The ASM module generates reliable attribute space attention maps through attribute prediction scores, guiding attribute space feature fusion and improving attribute positioning accuracy. Second, using the ASM module can directly bring 1.18% and 2.06% performance improvements on PA100K and PETA respectively, indicating that the spatial memory module is beneficial for locating attribute-related regions. In addition, we introduced occlusion contrast loss MCL by generating random occlusions on pedestrian images and evaluating the difference between their predictions and normal input predictions. By suppressing regional relevance to improve the framework's attribute positioning ability, it brings 0.39% and 0.47% performance improvements on PA100K and PETA respectively. Finally, when the ASM module and MCL module are used together, our method's mA on PA100K and PETA is 1.79% and 2.34% higher than the baseline model.

In Figure 6, by comparing the recognition mean accuracy rates of the baseline model Swin Transform and our model in the 26 attributes of the PA100K dataset , it can be seen that the model in this paper has a recognition effect on most attributes. For some attributes that rely on spatial regions (such as ”hat”, ”boots”), the improvement is significant due to ALFormer focus on intrinsic spatial regions. For some attributes that inferred from other related attributes (e.g. ”female”), the improvement is also significant because ALFormer could generate reliable attention maps that capture inherent locations for each attribute.

Table 1. Ablation study of each component of our method on the PETA and PA100K. Performance 328 improved by Mask Contrast Learning (MCL) and Attribute Spatial Memory (ASM) validates the 329 effectiveness of our methods

Model	PA100K					PETA				
	mA	Accu	Pre	Recall	F1	mA	Accu	Pre	Recall	F1
Baseline	82.82	81.47	89.08	88.88	88.98	87.20	80.17	**86.54**	88.73	87.62
Ours(+MCL)	83.21	**81.70**	**89.18**	88.99	**89.09**	87.67	71.65	76.81	89.01	82.46
Ours(+SPM)	84.00	77.00	82.85	90.08	86.32	89.26	79.55	84.83	89.92	87.30
Ours(+SPM,MCL)	**84.61**	78.86	84.11	**91.03**	87.43	**89.54**	**80.75**	86.15	**90.04**	88.05

4.5. Comparison with State-of-the-art Methods

In Table 2, we compare the performance of the proposed methods with several existing algorithms on the PA100K and Table 3 is on the PETA. Compared with the performance reported by the paper of MsVAA [3], VAC [28], and ALM [4] methods, ALFormer achieves better performance on the PETA and PA100K with increasing few learnable parameters.

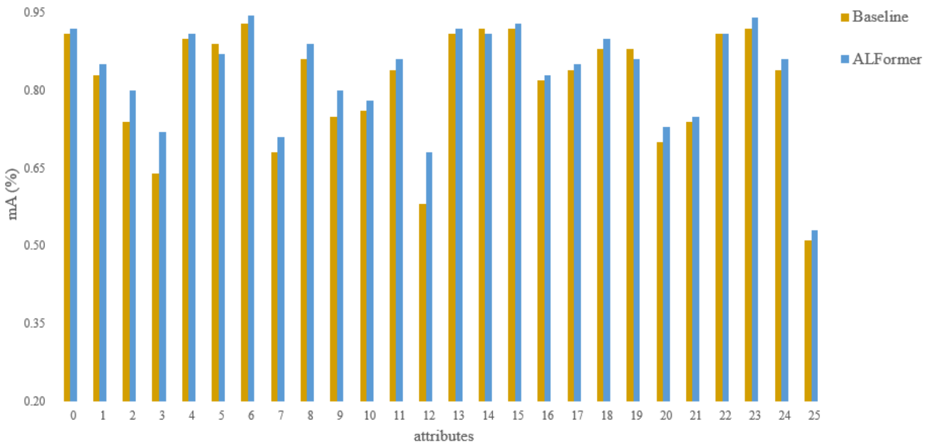

Fig. 6. The mA for each label on PA100k compared to baseline model. Legends in the figure: models are the same as in Table 1. The labels 0 to 25 are: female, ageover60, age18-60, ageless18, front, side, back, hat, glasses, handbag, shoulderbag, backpack, holdobjectsinfront, shortsleeve, longsleeve, upperstride, upperlogo, upperplaid, uppersplice, lowerstripe, LowerPattern, longcoat, trousers, shorts, skirtanddress, boots

Compared to the MsVAA model adopted ResNet101 as the backbone network, we achieve 4.95% and 1.56% performance improvements in mA and F1 on the PETA dataset. Compared with localization-based methods such as the ALM model, which utilizes the complicated combination of FPN, STN and SE modules introducing extra 17% parameters, ALFormer achieves 3.93% and 3.24% performance improvements in mA on both popular datasets. We localize attributes by ASM, which is a simple yet effective attention module to generate reliable and stable attribute-related regions and avoid heavy handcrafted design.

It can be noticed that the proposed method substantially outperforms the visual attention consistency (VAC) method. The VAC method hypothesis that global attention regions of random augmentations of the same image are consistent. The VAC method focuses on global attention regions of an individual image and cannot generate precise local attention regions for each fine-grained attribute. However, for a pair of augmentations of the same image, if attention regions of both augmentations are inaccurate, the VAC method cannot solve the attention region deviation problem, which can be addressed by our proposed method.

4.6. Effects of ALFormer

To better understand the effect of the improved model, we visualize the attribute regions most attended by the framework in Figure 7. It visualizes the attribute heat maps of Swin Transformer on the test set of the PA100K dataset. We take attribute attention maps of four pedestrian images(from P1 to P4) as examples that shows the visualization of spatial attention regions between the baseline method (green boundary) and the proposed AL-

Table 2. Performance (%) of the comparison state-of-the-art methods and the proposed method on the PA100K datasets. Five metrics, mean accuracy (mA), accuracy (Accu), precision (Prec), recall (Recall), F1 are evaluated. The best results are highlighted in red color, while the second bests are in blue.

method	mA	Accu	Pre	Recall	F1
DeepMar[29]	72.70	70.39	82.24	80.24	81.32
HP-Net[1]	74.21	72.19	82.97	82.09	82.53
PGDM[19]	74.95	73.08	84.36	82.24	83.29
JLPIS-PAA[31]	81.61	78.89	86.83	87.73	87.27
ALM[4]	80.68	77.08	84.21	88.84	86.46
JLAC[30]	82.31	79.47	87.45	87.77	87.61
CoCNN[37]	80.56	78.30	89.49	84.36	86.85
LG-Net[32]	76.96	75.55	86.99	83.17	85.04
VAC[9]	79.04	78.95	88.41	86.07	86.83
SSC[33]	81.70	78.85	85.80	88.92	86.89
ALFomer	84.61	78.86	84.11	91.03	87.43

Table 3. Performance (%) of the comparison state-of-the-art methods and the proposed method on the PETA datasets. The best results are highlighted in red color, while the second bests are in blue

method	mA	Accu	Pre	Recall	F1
CNN+SVM[13]	76.65	45.41	51.33	75.14	61.00
ACN[38]	81.15	73.66	84.06	81.26	82.64
DeepMar[29]	82.89	75.07	83.68	83.14	83.41
HP-Net[1]	81.77	76.13	84.92	83.24	84.07
PGDM[19]	82.97	78.08	86.86	84.68	85.76
DIAA[35]	84.59	78.56	86.79	86.12	86.46
JLPIS-PAA[31]	84.88	79.46	87.42	86.33	86.87
WPAL[34]	85.50	76.98	84.07	85.78	84.90
ALM[4]	86.30	79.52	85.65	88.09	86.85
JRL[20]	85.67	-	86.03	85.34	85.42
RCRA[36]	85.78	-	85.42	88.02	86.07
JLAC[30]	89.96	80.38	87.81	87.09	87.45
CoCNN[37]	86.97	79.95	87.58	87.73	87.65
MsVAA[3]	84.59	78.56	86.79	86.12	86.46
VAC[9]	83.63	78.94	87.63	85.45	86.23
SSC[33]	86.07	79.23	84.58	89.26	86.54
ALFomer	89.54	80.75	86.15	90.04	88.05

Former framework (red boundary). Compared with the baseline method, it is noted that the ALFormer helps to locate attribute-related regions for each attribute. In pedestrian image P2 and P3 shown in Figure 7, only head region is considered when recognizing the attribute "Glasses". Our method can effectively improve spatial location for each attribute by generating stable spatial attention and relevance suppression modules.

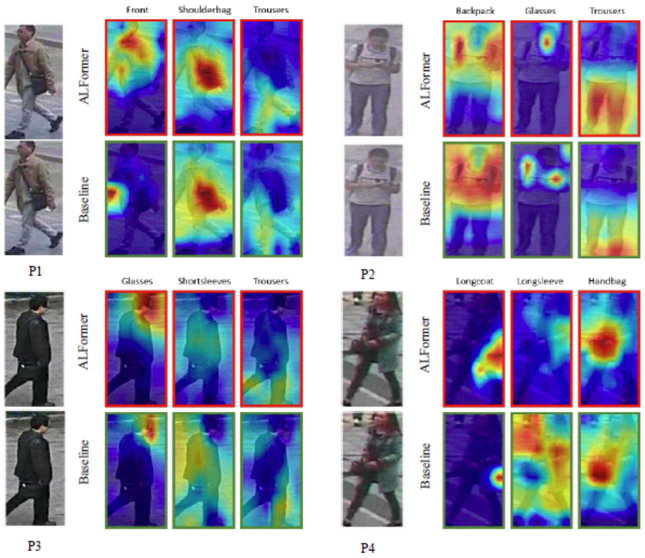

Fig. 7. Illustration of the effect of ALFormer. We take the attributes of four pedestrian images as examples to show the visualization of spatial attention regions between the baseline method (green boundary) and the proposed method (red boundary)

5. Conclusions

In this work, we identified and addressed the limitation of existing methods in accurately localizing discriminative regions for pedestrian attributes. We proposed ALFormer, a transformer-based framework that introduces two novel components Mask Contrast Learning and Attribute Spatial Memory. The former suppresses misleading regional correlations to force focus on intrinsic areas. The latter captures inherent spatial locations for each attribute. Experiments showed state-of-the-art results on two datasets, with ablation studies confirming the efficacy of our modules. The improved localization was further demonstrated through visualizations. Our model provides an effective way to exploit spatial consistency for pedestrian attribute recognition. Future directions include extending the approach to other fine-grained recognition tasks and exploring optimal fusion with global context.

Acknowledgments. This research work was supported by the China State Railway Group CO., Ltd (N2023X005).

References

1. Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., ... and Wang, X. (2017). Hydraplus-net: Attentive deep features for pe-destrian analysis. In Proceedings of the IEEE international conference on computer vision (pp. 350-359).
2. Liu, P., Liu, X., Yan, J., and Shao, J. (2018). Localization guided learning for pedestrian attribute recognition. arXiv preprint arXiv:1808.09102.
3. Sarafianos, N., Xu, X., and Kakadiaris, I. A. (2018). Deep imbalanced attribute classification using visual attention aggregation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 680-697).
4. 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).
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
6. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021, July). Training data-efficient image trans-formers and distillation through attention. In International conference on machine learning (pp. 10347-10357). PMLR.
7. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).
8. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2921-2929).
9. Fan, Z., Guan, Y.: Pedestrian attribute recognition based on dual self-attention Mechanism. Computer Science and Information Systems, Vol. 20, No. 2, 793–812. (2023), https://doi.org/10.2298/CSIS220815016F).
10. Kumar, N., Berg, A., Belhumeur, P. N., and Nayar, S. (2011). Describable visual attributes for face verification and image search. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(10), 1962-1977.
11. Kumar, N., Belhumeur, P., and Nayar, S. (2008). Facetracer: A search engine for large collections of images with faces. In Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Pro-ceedings, Part IV 10 (pp. 340-353). Springer Berlin Heidelberg.
12. Bourdev, L., Maji, S., and Malik, J. (2011, November). Describing people: A poselet-based approach to attribute classification. In 2011 International Conference on Computer Vision (pp. 1543-1550). IEEE.
13. Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision (pp. 3730-3738).
14. Abdulnabi, A. H., Wang, G., Lu, J., and Jia, K. (2015). Multi-task CNN model for attribute prediction. IEEE Transactions on Multimedia, 17(11), 1949-1959.
15. Hand, E., and Chellappa, R. (2017, February). Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).
16. Han, H., Jain, A. K., Wang, F., Shan, S., and Chen, X. (2017). Heterogeneous face attribute estimation: A deep multi-task learning approach. IEEE transactions on pattern analysis and machine intelligence, 40(11), 2597-2609.
17. Li, J., Zhao, F., Feng, J., Roy, S., Yan, S., and Sim, T. (2018). Landmark free face attribute prediction. IEEE Transactions on Image Processing, 27(9), 4651-4662.

18. Li, D., Chen, X., Zhang, Z., and Huang, K. (2018, July). Pose guided deep model for pedestrian attribute recognition in surveil-lance scenarios. In 2018 IEEE international conference on multimedia and expo (ICME) (pp. 1-6). IEEE.
19. 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).
20. Graves, A., and Graves, A. (2012). Long short-term memory. Supervised sequence labelling with recurrent neural networks, 37-45.
21. Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., and Zhang, S. (2019). Towards rich feature discovery with class activation maps augmentation for person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1389-1398).
22. Hou, R., Chang, H., Ma, B., Shan, S., and Chen, X. (2020). Temporal complementary learning for video person re-identification. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16 (pp. 388-405). Springer International Publishing.
23. Zitnick, C. L., and Dollár, P. (2014). Edge boxes: Locating object proposals from edges. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 (pp. 391-405). Springer International Publishing.
24. Jaderberg, M., Simonyan, K., and Zisserman, A. (2015). Spatial transformer networks. Advances in neural information processing systems, 28.
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... and Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
26. 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).
27. 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.
28. Guo, H., Zheng, K., Fan, X., Yu, H., and Wang, S. (2019). Visual attention consistency under image transforms for multi-label image classification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 729-739).
29. D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in Proc. IEEE Asia Conf. Pattern Recognit., 2015, pp. 111–115.
30. Z. Tan, Y. Yang, J. Wan, G. Guo, and S. Z. Li, "Relation-aware pedestrian attribute recognition with graph convolutional networks," in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 07, 2020, pp. 12 055–12 062.
31. Z. Tan, Y. Yang, J. Wan, H. Hang, G. Guo, and S. Z. Li, "Attentionbased pedestrian attribute analysis," IEEE Trans. Image Process., vol. 28, no. 12, pp. 6126–6140, 2019.
32. Pengze Liu, Xihui Liu, Junjie Yan, and Jing Shao. Localization guided learning for pedestrian attribute recognition. arXiv preprint arXiv:1808.09102, 2018.
33. Jia J, Chen X, Huang K. Spatial and semantic consistency regularizations for pedestrian attribute recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 962-971.
34. Kai Yu, Biao Leng, Zhang Zhang, Dangwei Li, and Kaiqi Huang. Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. arXiv preprint arXiv:1611.05603, 2016.
35. Sarafianos, N., Xu, X., Kakadiaris, I.A., 2018. Deep imbalanced attribute classification using visual attention aggregation, in: ECCV, pp. 680–697.
36. Zhao, X., Sang, L., Ding, G., Han, J., Di, N., Yan, C., 2019. Recurrent attention model for pedestrian attribute recognition, in: AAAI, pp. 9275–9282.

37. Han, K., Wang, Y., Shu, H., Liu, C., Xu, C., Xu, C., 2019. Attribute aware pooling for pedestrian attribute recognition. arXiv preprint arXiv:1907.11837.
38. 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).

Yuxin Liu received his M.Sc. degree in computer science from Beijing Jiaotong University, Beijing, China, in 2018. He is currently a Research Associate in China Academy of Railway Sciences Corporation Limited. He is also a researcher of the Computer Vision Laboratory of Institute of Computing Technology (ICT-CVLab). His current research interests include image processing, object detection, computing architecture, and deployment.

Mingzhe Wang received the B.S. degrees in mechanical electronics engineering from the Lanzhou Jiaotong University, Lanzhou, China, in 2002. He is currently a Professor with Institute of Computing Technology, China Academy of Railway Sciences Corporation Limited. His research interests include railway informatization.

Chao Li received the M.S. degrees in transportation information engineering and control from the China Academy of Railway Sciences, Beijing, China, in 2008. He is currently a Professor with Institute of Computing Technology, China Academy of Railway Sciences Corporation Limited. His research interests include informatization and intelligence of railway passenger stations and sections.

Shuoyan Liu received his Ph.D. degree in computer science from the Beijing Jiaotong University, Beijing, China, in 2011. She is currently a Associate Professor in China Academy of Railway Sciences Corporation Limited. She is also a researcher of the Computer Vision Laboratory of Institute of Computing Technology (ICT-CVLab). Her research interests include computer vision, object detection, image analysis, and visualization.

Received: October 15, 2023; Accepted: March 13, 2024.