An Empirical Study of Mamba-based Pedestrian Attribute Recognition

Xiao Wang[1], Weizhe Kong[2], Jiandong Jin[2], Shiao Wang[1], Ruichong Gao[1],
Qingchuan Ma[1], Chenglong Li[2]*, Jin Tang[1]
[1]School of Computer Science and Technology, Anhui University, Hefei 230601, China
[2]School of Artificial Intelligence, Anhui University, Hefei 230601, China
{*xiaowang, tangjin*}*@ahu.edu.cn*, *weizhekong99@gmail.com*, *jdjinahu@foxmail.com*,
wsa1943230570@126.com, {*e02114322, e02114334*}*@stu.ahu.edu.cn*, *lcl1314@foxmail.com*

Abstract

Current strong pedestrian attribute recognition models are developed based on Transformer networks, which are computationally heavy. Recently proposed models with linear complexity (e.g., Mamba) have garnered significant attention and have achieved a good balance between accuracy and computational cost across a variety of visual tasks. Relevant review articles also suggest that while these models can perform well on some pedestrian attribute recognition datasets, they are generally weaker than the corresponding Transformer models. To further tap into the potential of the novel Mamba architecture for PAR tasks, this paper designs and adapts Mamba into two typical PAR frameworks, i.e., the text-image fusion approach and pure vision Mamba multi-label recognition framework. It is found that interacting with attribute tags as additional input does not always lead to an improvement, specifically, Vim can be enhanced, but VMamba cannot. This paper further designs various hybrid Mamba-Transformer variants and conducts thorough experimental validations. These experimental results indicate that simply enhancing Mamba with a Transformer does not always lead to performance improvements but yields better results under certain settings. We hope this empirical study can further inspire research in Mamba for PAR, and even extend into the domain of multi-label recognition, through the design of these network structures and comprehensive experimentation. The source code of this work will be released at https://github.com/Event-AHU/OpenPAR

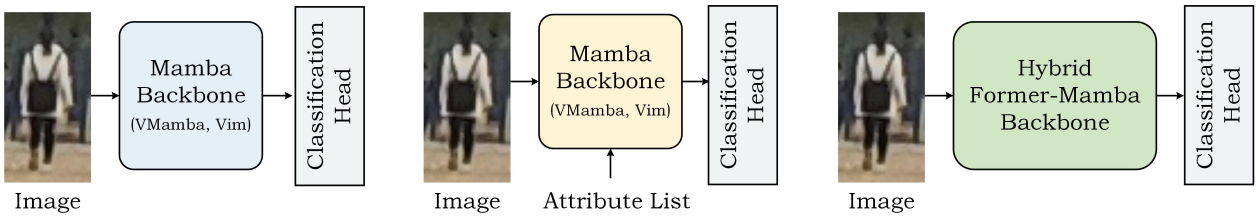
1. Introduction

Pedestrian Attribute Recognition (PAR) [48] is a widely exploited research topic in the computer vision (CV) community. It aims to recognize human attributes from a set of attribute descriptions, such as *short black hair*, *wearing hats*, *with back bag*, etc. On the one hand, pedestrian attribute recognition can be used to describe the appearance and motion characteristics of pedestrians, playing a crucial role in understanding them; on the other hand, pedestrian attributes can serve as mid-level semantic representations, assisting in improving the performance of other visual tasks, including pedestrian detection [55], person re-identification [54, 59], and more.

With the help of deep learning, various deep PAR models are proposed, including CNN (Convolutional Neural Networks), RNN (Recurrent Neural Networks), GNN (Graph Neural Networks), Transformer, etc. Specifically, MTCNN [1] uses CNN to extract features and combine them in a multi-task manner to solve the PAR task. JRL [44] and GRL [56] are solving PAR as a sequence prediction task using LSTM to model the association between attributes. PARformer [7] and VTB [3] use Transformer to solve PAR task, the difference is that VTB introduces text modality. Although existing PAR models work well in simple scenarios, the recognition performance in challenging scenarios is still limited (e.g., low illumination, cross-domain). In addition, the training/inference cost ($\mathcal{O}(N^2)$) is rather high due to the utilization of self-attention in Transformer networks. Based on the observations above and reflections, it is natural to raise the following question: *how can we design a new pedestrian attribute recognition framework that achieves a comparable or even better performance compared to the Transformer-based models, meanwhile, reducing the cost significantly?*

Recently, the State Space Model (SSM) has drawn more and more attention in the artificial intelligence community due to its linear complexity ($\mathcal{O}(N)$) and good performance. It has been widely utilized for visual tracking [17], image-/video-based classification [27, 60], and time series analysis [49]. The experimental results reported in the SSM survey [47] demonstrate that the Vim-S [60] based attribute recognition model performs better than the ViT-S based

*Corresponding author: Chenglong Li (*lcl1314@foxmail.com*)



(a). Image to Multi-label Regression *(Pure Mamba-based)*

(b). Image-Text Fusion based PAR *(Pure Mamba-based)*

(c). Image to Multi-label Regression *(Hybrid Transformer-Mamba based)*

Figure 1. An illustration of our proposed Mamba-based pedestrian attribute recognition framework.

version on the PA100K [31] dataset, but achieves worse results on the PETA [5] dataset. Hence, crafting innovative Mamba-inspired PAR architectures that deliver uniform performance enhancements across benchmarks, while also maintaining low training and testing overhead, poses an ongoing challenge.

Inspired by these works, in this paper, we attempt to conduct extensive experiments to validate the effectiveness of Mamba-based pedestrian attribute recognition. Generally speaking, as illustrated in Fig. 1, we design two representative Mamba-based PAR frameworks, i.e., *image-based multi-label classification PAR framework* and *image-text fusion-based PAR framework*. Our experimental results demonstrate that the pure Mamba-based model performs better than the image-text fusion based model when the VMamba [33] is adopted, but the opposite when the Vim [60] is utilized. Furthermore, we consider the design of introducing Transformers to create a more powerful and efficient hybrid Mamba-Transformer backbone architecture for better feature representation extraction, in order to improve the performance of attribute recognition, as illustrated in Fig. 3. By conducting extensive experiments, we find that when utilizing Mamba for the dense fusion of hierarchical Transformer features, i.e., the Fig. 3 (e), the final results can beat the ViT-B based PAR framework. Also, when distilling from Transformer to Mamba for PAR, i.e., the Fig. 3 (g, h), the performance can also be improved over the Vim-S and ViT-S based version.

To sum up, we conclude the key contributions of this paper with the following three aspects:

1). We exploit the integration of the recently released Mamba architectures with the PAR task by proposing two representative Mamba-based PAR frameworks, i.e., *the image based multi-label classification framework*, and *image-text fusion based PAR*. Our experimental results indicate that different visual Mamba models are suitable for different PAR frameworks.

2). We propose several variations of hybrid Transformer-Mamba networks to improve the final attribute recognition performance further and hope it provides comprehensive guidelines for Mamba-based attribute recognition.

3). We conduct extensive experiments on multiple PAR benchmark datasets, including PA100K [31], PETA [5], RAP-V1 [26], RAP-V2 [25], WIDER [29], PETA-ZS [20], and RAP-ZS [20], to demonstrate the effectiveness and efficiency of the proposed Mamba-based PAR frameworks.

This paper is organized as follows: Firstly, we review the related works on pedestrian attribute recognition and SSM; then, we will introduce the pure Mamba-based and Mamba-based vision-language fusion PAR frameworks; we also propose different variations of hybrid Mamba-Transformer networks for the PAR. After that, we will jump into the experiments, compare them with other state-of-the-art PAR models, and discuss the influence of different settings for Mamba-based PAR. We also give some analysis of the model efficiency and visualizations. Finally, we conclude this paper and propose possible future works worth to be studied.

2. Related Works

In this section, we will introduce the related works on pedestrian attribute recognition, and state space model. More details can be found in the following surveys [45, 47, 48].

2.1. Pedestrian Attribute Recognition

Pedestrian attribute recognition has made significant progress over the years, with many methods proposed by researchers achieving promising results. Early methods primarily involved feeding images into CNN networks for multi-task training, and sharing parameters between networks to obtain richer features. For example, MTCNN [1] adopted this multi-task training approach, using CNN to extract image features and allowing knowledge sharing among different attribute categories. Due to the correlations between different attributes, sequential prediction models were introduced into pedestrian attribute recognition research. JRL [44] modeled the contextual information between pedestrians and the contextual information between attributes using LSTM, transforming pedestrian attribute recognition into a sequence prediction problem. Building

upon JRL, GRL [56] divided the attribute list into multiple attribute groups and modeled the spatial and semantic correlations between groups using LSTM. With the development of deep learning networks, attention models began to be used in pedestrian attribute recognition. Specifically, HydraPlus-Net [32] introduced a multi-directional attention module to extract multi-scale attention features for obtaining more comprehensive pedestrian features. Transformer, a model proposed by Vaswani et al. [43], has been widely used in the field of natural language processing and has achieved great success. Some researchers have also applied it to pedestrian attribute recognition. Fan et al. [7] proposed a pure Transformer-based PAR network, using a Transformer-based backbone network for image feature extraction to obtain more discriminative features. Cheng et al. [3] proposed VTB, encoding pedestrian images into visual features and encoding attribute annotations into text features using pre-trained text encoders, and using Transformer for multimodal fusion. However, due to the global attention mechanism of the Transformer, its complexity grows quadratically with the length of the input sequence. Gu et al. [9] introduced Mamba, a type of State Space Model that optimizes the complexity of sequence processing to linear. Therefore, we introduce a state space modal for optimization, improving computational efficiency while ensuring the accuracy of pedestrian attribute recognition.

2.2. State Space Model

The State Space Model (SSMs) [9, 11–13] is a mathematical framework used to model dynamic systems, which has received widespread attention in recent years due to its excellent performance in handling long sequence data. S4 [12], as a pioneering work in using state space modal for long sequence processing, introduced Hippo [10] to address the issue of long-range dependencies in sequence modeling, significantly improving model performance. S4ND [34] extends the spatial state model to multidimensional signals, enabling it to model large-scale visual data as continuous multidimensional signals, laying the foundation for SSMs's use in multimodal tasks. Mamba [9] introduces a selective mechanism for information processing while balancing training, inference efficiency, and model effectiveness, outperforming Transformer [43] in natural language processing. After Mamba, there appeared works applying Mamba to the visual domain, such as Vim [60] and VMamba [33]. Both of these works show the potential of Mamba in vision tasks. Vim proposes a bidirectional Mamba to model images, and VMamba proposes a four-direction cross-scan module. Through this multi-direction sequence order, it is more effective to model two-dimensional image data and maintain linear computational complexity. Mamba-2 [4] is the improvement of Mamba, and proposes SSD (State Space Duality) framework, which improves Mamba by 2-8 times speedup and obtains better performance. Our work explores the potential of Mamba in handling visual-text fusion features and comprehensively applies Mamba and its derivative modules to pedestrian attribute recognition.

3. Methodology

In this section, we will first give a brief review of the State Space Model (SSM), then, we will introduce two Mamba-based PAR frameworks, including the *pure image-based multi-label classification* and *image-text fusion based PAR framework*. Further, we design eight variations of hybrid Mamba-Transformer networks for the PAR task.

3.1. Preliminary: State Space Model

The State Space Model (SSM) originates from the classic Kalman filter [22] algorithm which introduces linear filtering. It converts a one-dimensional sequence into an N-dimensional hidden state for output. The calculation formula is as follows,

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t) + \mathbf{D}x(t). \end{aligned} \tag{1}$$

where $x(t) \in \mathbb{R}^L$ and $h'(t) \in \mathbb{R}^N$ are the input sequence and the derivative of the hidden state respectively. $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the state matrix, $\mathbf{B} \in \mathbb{R}^{N\times L}$ is the input matrix , $\mathbf{C} \in \mathbb{R}^{L\times N}$ is the output matrix, and $\mathbf{D} \in \mathbb{R}^{L\times L}$ is the feed-through matrix. The above formula is a continuous-time SSM. In order to facilitate the deep learning algorithm's understanding of the input, we need to discretize the matrixes through some methods, such as zero order hold (ZOH), etc. Specifically, Gu et al. [12] propose structured state-space sequence models (S4), which convert continuous parameters into discrete ones by introducing timescale parameter Δ. The formula is as follows,

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}), \\ \overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}, \\ \overline{\mathbf{C}} &= \mathbf{C}. \end{aligned} \tag{2}$$

where $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$, and $\bar{\mathbf{C}}$ are discrete parameters of the system from $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ respectively. Furthermore, the formula can be shown that,

$$\begin{aligned} h_t &= \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t, \\ y_t &= \bar{\mathbf{C}}h_t. \end{aligned} \tag{3}$$

where the $\mathbf{D}$ matrix can sometimes be ignored as a residual. Based on the aforementioned models, in the field of computer vision, Vim [60] and VMamba [33] have been proposed and received considerable attention. This paper will explore SSM-based pedestrian attribute recognition frameworks using these two models.

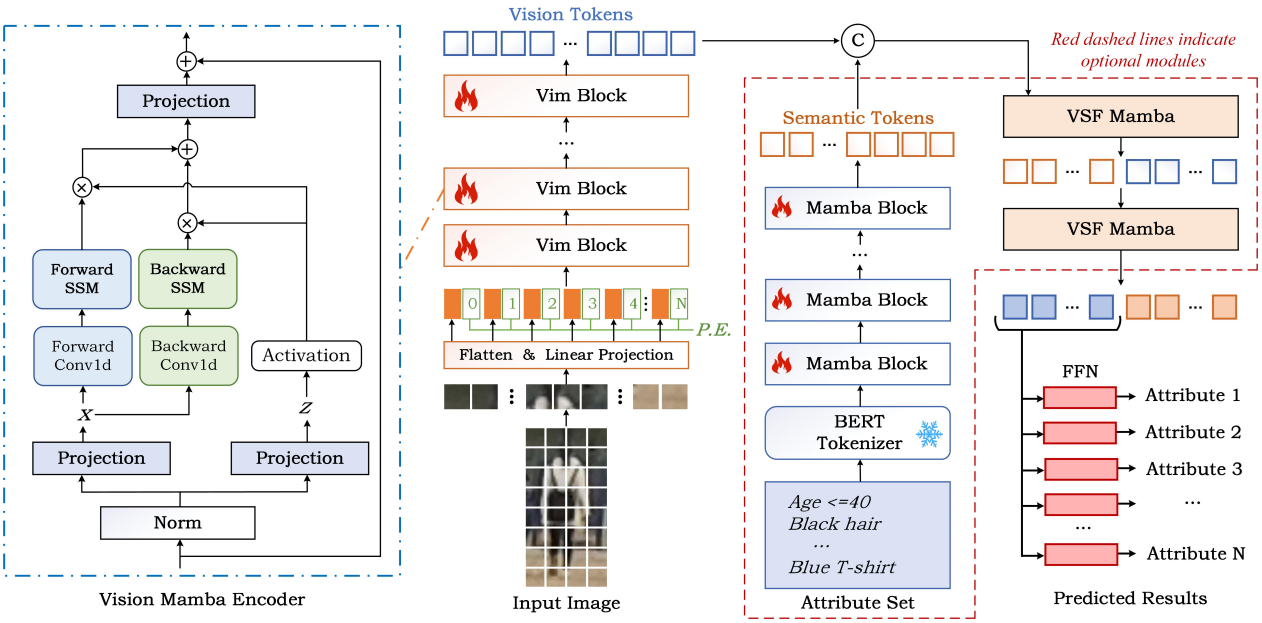


Figure 2. An illustration of our proposed vision-language fusion framework based on Mamba for pedestrian attribute recognition.

3.2. When Mamba Meets PAR

In this section, we mainly focus on two widely used PAR frameworks, i.e., the *image-based multi-label classification framework* [7, 24, 42], and the *image-text fusion based PAR framework* [3, 46], as shown in Fig. 2. Considering that image-based multi-label recognition can be seen as a special case of the image-text integration framework, that is, by removing the attribute label encoding module, this paper will focus on elaborating the PAR process of the image-text integration framework. The details of multi-label recognition will not be further discussed.

Generally speaking, our proposed Mamba-based visual-language fusion PAR architecture consists of three main modules, i.e., the Mamba vision backbone which is composed of a Vim [60] network, a Mamba-based text encoder for extracting the attribute semantic information, and a visual-semantic fusion (VSF) Mamba module for image-text feature interaction and fusion. Finally, we adopt a classification head for pedestrian attribute recognition. We will introduce these modules in the subsequent paragraphs respectively.

Mamba for Input Representation. Given a pedestrian image I, we need to predict its attributes from an attribute set $A = \{a_1, a_2, ..., a_L\}$, L is the number of person attributes, and the ground truth annotations Y are provided for the supervised learning of our PAR framework. In our framework, we adopt the vision Mamba network Vim [60] as an example to demonstrate how to encode the given person image into visual features X_v. Note that, both the Vim [60] and VMamba [33] can be adopted for the encoding of pedestrian images, as validated in our experimental results. Firstly, we divide the image into equally sized image patches based on a specified patch size, then, these patches are flattened and fed into a linear projection to obtain the patch tokens $X_v \in \mathbb{R}^{P^2 \times D}$. Positional embeddings $P.E.$ are added to each token to better capture the spatial positions of each patch.

Next, we feed the tokens into a stack of Vim [60] blocks, $M^V = \{M_0^V, M_1^V, \ldots, M_N^V\}$, where N is the number of Vim blocks, thus, we can obtain the visual tokens X_v. More in detail, as shown in Fig. 2 the input tokens are firstly processed using the normalization layer and then transformed into x and z using projection layers. For the x, the forward and backward processing branches are adopted for the vision feature learning, and each branch contains both Conv1d and SSM layers. For the z, an activation layer $\sigma(\cdot)$ is adopted to enhance the feature and multiply with the output of the forward and backward processing branches respectively. The compute procedure can be simply written as:

$$\sigma(z) * FSSM(Conv1d(x)) + \sigma(z) * BSSM(Conv1d(x)) \tag{4}$$

The output features are added and projected as the output of each Mamba block.

To better help the framework understand what is human attributes, the attribute labels defined on the whole dataset are usually directly integrated into the PAR framework. Given the attribute set A, we use the pre-trained BERT [23] tokenizer to embed the attributes into semantic tokens X_s. These semantic tokens are further enhanced

by the text Mamba [9] blocks $M^T = \{M_0^T, M_1^T, \ldots, M_K^T\}$ and the output attribute tokens are fed into vision-language aggregation module which will be introduced in the following paragraphs.

[Optional] Mamba for Vision-Language Aggregation. Previous purely visual PAR methods struggle to effectively correlate attributes with visual features, and the inconsistency of optimization objectives across multiple attribute classifications sharing the same features leads to interference issues. We address these two problems through visual-language interaction. In our visual-language interaction processing, we integrate the extracted visual and textual features, denoted as $\mathcal{X} = [X_v, X_s]$, and input them into the Vision-Semantic Fusion (VSF) Mamba module. This module, consisting of a stack of Mamba blocks, is designed to model the relationship between attributes and visual data across various layers. Ultimately, we employ the textual features $\mathcal{X}_F$ that are aggregated with visual information for classification.

PAR Classification Head and Loss Function. By using the Mamba networks to obtain effective feature representations, we input the features to the pedestrian attribute recognition prediction head for multi-label prediction. Specifically, our prediction head is composed of multiple feed-forward networks (FFN). It inputs features of pedestrian attributes to FFNs, outputs the respective category of each input, and completes the final prediction. Specifically,

$$P = Sigmoid(FFN(\mathcal{X})) \tag{5}$$

where $P = \{p_1, p_2, \ldots, p_L\}$, and L is the number of attributes. We use the weighted cross-entropy loss as the loss function for our prediction which can be formulated as:

$$\mathcal{L} = -\sum_{j=1}^{L} w_j (y_j \log(p_j) + (1 - y_j) \log(1 - p_j)) \tag{6}$$

where w is positively correlated with the attribute positive ratio in the training set.

3.3. Hybrid Mamba-Transformer for PAR

In our experiments, we find that the pure VMamba [33] based PAR framework performs better than the image-text fusion based one. Therefore, in this section, we exploit new network architectures to aggregate the VMamba [33] with Transformer networks for better pedestrian attribute recognition. Specifically speaking, we design eight variations of hybrid Mamba-Transformer networks to validate which one performs better for the PAR task. Due to the limited space in this paper, we briefly introduced the computation procedures of these hybrid Mamba-Transformer networks for the PAR task. For more details on these models, please check our source code.

• **Parallel Fusion (PaFusion)** is shown in Fig. 3 (a), where we connect the corresponding layers in the two models ViT-B and Vim-S in parallel, then add them together and feed into the next layer uniformly. It should be noted that the feature dimension of Vim-S is 384-D, while the feature dimension of ViT-B is 768-D. Thus, linear mapping is performed on the Vim features before feature addition each time. The integrated features are also subjected to reverse linear mapping for input to the next layer of Vim. Since Vim-S contains 24 Mamba layers and ViT-B contains 12 Transformer layers, we connect two Vim layers and one ViT layer in parallel in our practical implementation.

• **Non-alternating Serial Fusion (N-ASF)** As shown in Fig. 3 (b), we directly connect the Vim-S and ViT-B to build the N-ASF PAR model. After a couple of Vim-S layers, we perform a dimension transformation and feed it into four ViT-B blocks for the feature processing in the next step.

• **Alternating Serial Fusion (ASF)** is shown in Fig. 3 (c), where we alternately connect two models of ViT-B and Vim-S layers in serial. There is a linear mapping between the two-layer connections due to the inconsistent number of layers. We connect one layer of ViT to two layers of Vim in series.

• **Mamba-enhanced Transformer (MaFormer)** is shown in Fig. 3 (d), where we use Vim-S as a complementary information extractor to integrate the features extracted by Vim and the current layer input before feeding into ViT.

• **Mamba for Hierarchical Dense Fusion of Transformer (MaHDFT)** is shown in Fig. 3 (e), where we consider Vim-S as a processor that integrates the outputs of all layers of ViT. We use the ViT-B model trained in advance on the PA100k dataset and then freeze it. Following this, we integrate the outputs of each layer of ViT into four layers of Vim for integration, and then feed the outputs of Vim into the classification head for label prediction. Due to the limited memory of our used RTX3090 GPUs, it is not possible to integrate all tokens directly (2352 total tokens), so each output layer is first passed through a convolutional layer to reduce the number of tokens to 49 (588 total tokens). It should be mentioned here that we retain the classification head of ViT-B on the PA100K dataset for generating the labels corresponding to the final output features of ViT. The classification head used for Vim is recreated, and after obtaining the two sets of label distributions, we average them to obtain the final label.

• **Adapter for Mamba-Transformer Fusion (AdaMTF)** is shown in Fig. 3 (f), where we connect Vim and ViT backbone as a whole in parallel, and introduce adapter in the corresponding layer to perform feature interaction, where the adapter is an MLP in our implementation.

• **Knowledge Distill (KDTM)** is shown in Fig. 3 (g), where we focus on Vim and use ViT as the teacher network to perform *feature-* or *logit-level* distillation on Vim to enhance

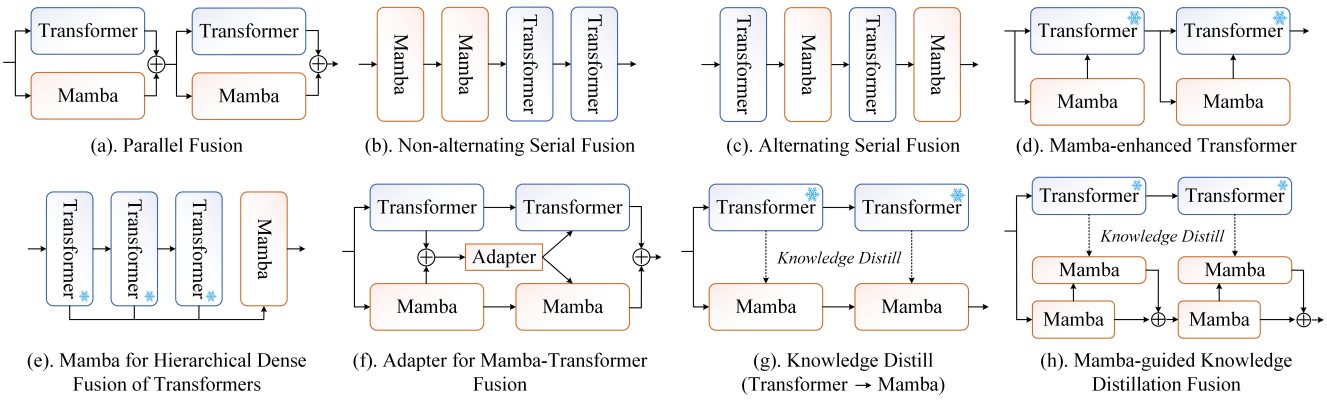


Figure 3. An illustration of various hybrid Mamba-Transformer frameworks for PAR.

the feature representations.

• **Mamba-guided Knowledge Distillation Fusion (MaKDF)** is shown in Fig. 3 (h), where we introduce a Vim module on top of the direct distillation to perform a distillation transition from the ViT teacher network and then integrate this feature back into the original Vim feature. This avoids the difference between ViT and Vim feature extraction, which brings performance damage to Vim.

4. Experiments

4.1. Datasets and Evaluation Metric

In our experiments, seven widely used PAR benchmark datasets are evaluated, including PA100K [31], PETA [5], RAP-V1 [26], RAP-V2 [25], WIDER [29], PETA-ZS [20], and RAP-ZS [20]. A brief introduction to these datasets is given below:

• **PA100K [31] dataset** is the largest pedestrian attribute recognition dataset, which contains 100,000 pedestrian images and 26 binary attributes. In our experiments, we split them into a training and validation set of 90,000 images, and a test subset of the remaining 10,000 images.

• **PETA [5] dataset** contains 19,000 outdoor or indoor pedestrian images and 61 binary attributes. These images are divided into 9500 as the training subset, 1900 as the validation subset, and 7600 as the test subset. In the experiment, we selected 35 pedestrian attributes according to the method of [5].

• **RAP-V1 [26] dataset** contains 41585 pedestrian images and 69 binary attributes, of which 33,268 images are used for training. In the current work, 51 attributes are usually selected for training and evaluation.

• **RAP-V2 [25] dataset** has 84,928 pedestrian images and 69 binary attributes, of which 67,943 are used for training. We selected 54 attributes to train and evaluate our model.

• **WIDER [29] dataset** is divided into a training and validation set of 28,345 images and a test set of 29,179 images with a total of 14 attribute labels. Following the default setup, the train-val set is used for training and the performance on the test set is evaluated.

• **PETA-ZS [20] dataset** is proposed by Jia et al., based on the PETA dataset, following a zero-shot protocol. The training, validation, and test sets consist of 11,241 | 3,826 | 3,933 samples respectively. For our experiments, we selected 35 common attributes according to Jia et al. [20].

• **RAP-ZS [20] dataset** is developed based on the RAPv2 dataset, where the training, validation, and testing sets contain 17062, 4628, and 4928 pedestrian images, respectively. There is no shared personal identity between training and inference data. In the experiment, we selected 53 attributes for evaluation following Jia et al. [20].

To evaluate our proposed Mamba-based pedestrian attribute recognition algorithm and other SOTA PAR models, we adopt the following evaluation metrics, including mA, Accuracy, Precision, Recall, and F1-measure. For the mA can be expressed as:

$$mA = \frac{\sum_{i=1}^{C} AP_i}{C}$$

where AP_i is the area under the precision-recall curve for the $attribute_i$, and C is the total number of attributes. The Accuracy can be expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP is a positive sample predicted correctly, TN is a negative sample predicted correctly, FP is a negative sample predicted incorrectly, and FN is a positive sample predicted incorrectly. The Precision, Recall, and F1-measure can be expressed as:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN},$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

4.2. Implementation Details

In our experiments, both VMamba-B and Vim-S versions are used as visual feature extractors. VMamba-B is a vision Mamba network in four stages with 2, 2, 27, and 2 layers in each stage, and the feature dimension is 768-D. Vim-S is a 24-layer vision Mamba model with 384 feature dimensions. For the text modality, the original Mamba model Mamba-130M, with a feature size of 768 is used. Note that, random cropping and flipping operations are employed for data augmentation.

For the detailed parameters, we train the model for 100 epochs using the Adam [37] optimizer. When using VMamba-B, we set the learning rate to 8e-5 and the batch size to 16. When using Vim-S, we set the learning to 2.5e-5 and the batch size to 64. Our source code is implemented using Python and the deep learning framework PyTorch [35]. The experiments were conducted on a server with GPU RTX3090s. More details can be found in our source code.

4.3. Comparison on Public PAR Benchmarks

In this section, we report the recognition results on seven datasets and compare them with existing state-of-the-art pedestrian attribute recognition algorithms. Note that the results for VTB* were obtained by replacing VTB's backbone network with ViT-L/14. Visualization of attribute predictions of some person images is given in Fig. 5.

Results on PETA dataset. As shown in Table 1, the evaluated MambaPAR model performs well on most evaluation metrics, with different branches and backbone networks showing varying results. Specifically, the VMamba-B visual branch model achieved mA, Acc, Prec, Recall, and F1 metrics of 86.28, 80.54, 87.45, 87.95, and 87.45, respectively; the VMamba-B visual plus language branch model achieved 85.01, 78.47, 85.12, 87.49, and 86.00; the Vim visual plus language branch model achieved 81.45, 74.92, 82.68, 84.27, and 83.15; and the Vim visual plus text branch model achieved 84.25, 76.07, 82.73, 87.20, and 84.56. Interestingly, we can find that the attribute labels are useful for the Vim-S based PAR, but not for the VMamba-B based version. Compared to the Transformer-based pedestrian attribute recognition models, our VMamba visual branch model surpassed VTB (ViT-B/16, 85.31/79.60/86.76/87.17/86.71) and VTB* (ViT-L/14, 86.34/79.59/86.66/87.82/86.97) in the mA, Acc, Prec, Recall, and F1 metrics. Therefore, we can conclude that our model achieves competitive results on the PETA dataset, verifying the feasibility of the Mamba model for pedestrian attribute recognition.

Results on PA100K dataset. As shown in Table 1, our MambaPAR model also performs well on the PA100K dataset. The VMamba-B visual branch model achieved mA, Acc, Prec, Recall, and F1 metrics of 83.63, 81.11, 87.59, 89.9, and 88.40, respectively; the VMamba-B visual plus language branch model achieved 81.50, 79.55, 87.54, 88.03, and 87.35; the Vim visual plus language branch model achieved 79.28, 76.77, 85.24, 86.58, and 85.44; and the Vim visual plus language model achieved 80.87, 79.33, 86.48, 88.79, and 87.21. The VMamba-B visual branch model's performance is close to that of VTB* (85.3/81.76/87.87/90.67/88.86). These experimental results indicate that the exploited Mamba-based PAR frameworks we explored have achieved preliminary success.

From the experimental results reported in the PETA and PA100K datasets, we can find that the VMamba-B achieves the best results over other Mamba-based PAR models. Therefore, we only report the results of pure VMamba-B based PAR framework for the rest of the datasets.

Results on RAP-V1 dataset. As shown in Table 2, our VMamba-B visual branch model achieved mA, Acc, Prec, Recall, and F1 metrics of 83.12, 69.47, 78.03, 84.81, and 80.88, respectively. In comparison, the strong baseline VTB* achieved 83.69, 69.78, 78.09, 85.21, and 81.10 on these metrics, and our results are very close to theirs. Meanwhile, our model's performance is also close to the DRFormer [42], which is also developed based on Transformer. This fully demonstrates that the performance of our Mamba-based pedestrian attribute recognition model is comparable to or even surpasses Transformer-based models.

Results on RAP-V2 dataset. As shown in Table 2, our VMamba-B visual branch model achieved mA, Acc, Prec, Recall, and F1 metrics of 81.91, 68.08, 76.41, 84.38, and 79.83, respectively, outperforming the Transformer-based VTB* (81.36/67.58/76.19/84.00/79.52). This further proves the effectiveness of our proposed pedestrian attribute recognition model.

Results on WIDER dataset. As shown in Table 3, our model achieved mA of 89.38, surpassing VTB (ViT-B/16, 88.2), which further validates the feasibility of our model. Compared with the pre-trained vision-language CLIP [36] model based PAR, e.g., the VTB* and PromptPAR, the VMamba-based model is still inferior to these models. In our future works, we will consider knowledge distillation strategies to further augment the performance of VMamba-based PAR.

Results on PETA-ZS dataset. In addition to standard evaluations, we also evaluated the model's performance on datasets using a zero-shot setting. According to the results reported in Table 4, our model achieved mA, Acc, Prec, Recall, and F1 metrics of 74.43, 60.56, 73.12, 74.51, and 73.37, respectively, which are close to or even better than VTB (ViT-B/16, 75.13/60.50/73.29/74.40/73.38) on certain metrics.

Results on RAP-ZS dataset. As shown in Table 4, the experimental results on the RAP-ZS dataset show that VTB

Table 1. Comparison with state-of-the-art methods on PETA and PA100K datasets. “V” means Vision branch, “L” means Language branch. “-” means this indicator is not available.

Methods	Backbone	PETA					PA100K				
		mA	Acc	Prec	Recall	F1	mA	Acc	Prec	Recall	F1
DeepMAR (ACPR 2015) [24]	CaffeNet	82.89	75.07	83.68	83.14	83.41	72.70	70.39	82.24	80.42	81.32
HPNet (ICCV 2017) [31]	Inception	81.77	76.13	84.92	83.24	84.07	74.21	72.19	82.97	82.09	82.53
JRL (ICCV 2017) [44]	AlexNet	82.13	-	82.55	82.12	82.02	-	-	-	-	-
GRL (IJCAI 2018) [56]	Inception-V3	86.70	-	84.34	88.82	86.51	-	-	-	-	-
MsVAA (ECCV 2018) [38]	ResNet101	84.59	78.56	86.79	86.12	86.46	-	-	-	-	-
RA (AAAI 2019) [57]	Inception-V3	86.11	-	84.69	88.51	86.56	-	-	-	-	-
VRKD (IJCAI 2019) [28]	ResNet50	84.90	80.95	88.37	87.47	87.91	77.87	78.49	88.42	86.08	87.24
AAP (IJCAI 2019) [16]	ResNet50	86.97	79.95	87.58	87.73	87.65	80.56	78.30	89.49	84.36	86.85
VAC (CVPR 2019) [15]	ResNet50	-	-	-	-	-	79.16	79.44	88.97	86.26	87.59
ALM (ICCV 2019) [41]	BN-Inception	86.30	79.52	85.65	88.09	86.85	80.68	77.08	84.24	88.84	86.46
JLAC (AAAI 2020) [40]	ResNet50	86.96	80.38	87.81	87.09	87.50	82.31	79.47	87.45	87.77	87.61
SCRL (TCSVT 2020) [51]	ResNet50	87.2	-	89.20	87.5	88.3	80.6	-	88.7	84.9	82.1
SSCsoft (ICCV 2021) [18]	ResNet50	86.52	78.95	86.02	87.12	86.99	81.87	78.89	85.98	89.10	86.87
IAA-Caps (PR 2022) [50]	OSNet	85.27	78.04	86.08	85.80	85.64	81.94	80.31	88.36	88.01	87.80
MCFL (NCA 2022) [2]	ResNet-50	86.83	78.89	84.57	88.84	86.65	81.53	77.80	85.11	88.20	86.62
DRFormer (NC 2022) [42]	ViT-B/16	89.96	81.30	85.68	91.08	88.30	82.47	80.27	87.60	88.49	88.04
VAC-Combine (IJCV 2022) [14]	ResNet50	-	-	-	-	-	82.19	80.66	88.72	88.10	88.41
DAFL (AAAI 2022) [19]	ResNet50	87.07	78.88	85.78	87.03	86.40	83.54	80.13	87.01	89.19	88.09
CGCN (TMM 2022) [6]	ResNet	87.08	79.30	83.97	89.38	86.59	-	-	-	-	-
CAS-SAL-FR (IJCV 2022) [53]	ResNet50	86.40	79.93	87.03	87.33	87.18	82.86	79.64	86.81	87.79	85.18
PARformer (TCSVT 2023) [7]	Swin-L	89.32	82.86	88.06	91.98	89.06	84.46	81.13	88.09	91.67	88.52
VTB (TCSVT 2022) [3]	ViT-B/16	85.31	79.60	86.76	87.17	86.71	83.72	80.89	87.88	89.30	88.21
VTB* (TCSVT 2022) [3]	ViT-L/14	86.34	79.59	86.66	87.82	86.97	85.30	81.76	87.87	90.67	88.86
SequencePAR (arXiv-2023) [21]	ViT-L/14	-	84.92	90.44	90.73	90.46	-	83.94	90.38	90.23	90.10
MambaPAR (V)	Vim-S	81.45	74.92	82.68	84.27	83.15	79.28	76.77	85.24	86.58	85.44
MambaPAR (V+L)	Vim-S	84.25	76.07	82.73	87.20	84.56	80.87	79.33	86.48	88.79	87.21
MambaPAR (V)	VMamba-B	86.28	80.54	87.45	87.95	87.45	83.63	81.11	87.59	89.98	88.40
MambaPAR (V+L)	VMamba-B	85.01	78.47	85.12	87.49	86.00	81.50	79.55	87.54	88.03	87.35

Table 2. Comparison with state-of-the-art methods on RAPv1 and RAPv2 datasets. “V” means Vision branch. “-” means this indicator is not available.

Methods	Backbone	RAPv1					RAPv2				
		mA	Acc	Prec	Recall	F1	mA	Acc	Prec	Recall	F1
DeepMAR (ACPR 2015) [24]	CaffeNet	73.79	62.02	74.92	76.21	75.56	-	-	-	-	-
HPNet (ICCV 2017) [31]	Inception	76.12	65.39	77.33	78.79	78.05	-	-	-	-	-
JRL (ICCV 2017) [44]	AlexNet	74.74	-	75.08	74.96	74.62	-	-	-	-	-
GRL (IJCAI 2018) [56]	Inception-V3	81.20	-	77.70	80.90	79.29	-	-	-	-	-
MsVAA (ECCV 2018) [38]	ResNet101	-	-	-	-	-	78.34	65.57	77.37	79.17	78.26
RA (AAAI 2019) [57]	Inception-V3	81.16	-	79.45	79.23	79.34	-	-	-	-	-
VRKD (IJCAI 2019) [28]	ResNet50	78.30	69.79	82.13	80.35	81.23	-	-	-	-	-
AAP (IJCAI 2019) [16]	ResNet50	81.42	68.37	81.04	80.27	80.65	-	-	-	-	-
VAC (CVPR 2019) [15]	ResNet50	-	-	-	-	-	79.23	64.51	75.77	79.43	77.10
ALM (ICCV 2019) [41]	BN-Inception	81.87	68.17	74.71	86.48	80.16	79.79	64.79	73.93	82.03	77.77
JLAC (AAAI 2020) [40]	ResNet50	83.69	69.15	79.31	82.40	80.82	79.23	64.42	75.69	79.18	77.40
SSCsoft (ICCV 2021) [18]	ResNet50	82.77	68.37	75.05	87.49	80.43	-	-	-	-	-
IAA-Caps (PR 2022) [50]	OSNet	81.72	68.47	79.56	82.06	80.37	-	-	-	-	-
MCFL (NCA 2022) [2]	ResNet50	84.04	67.28	73.44	87.75	79.96	-	-	-	-	-
DRFormer (NC 2022) [42]	ViT-B/16	81.81	70.60	80.12	82.77	81.42	-	-	-	-	-
VAC-Combine (IJCV 2022) [14]	ResNet50	81.30	70.12	81.56	81.51	81.54	-	-	-	-	-
DAFL (AAAI 2022) [19]	ResNet50	83.72	68.18	77.41	83.39	80.29	81.04	66.70	76.39	82.07	79.13
CGCN (TMM 2022) [6]	ResNet50	84.70	54.40	60.03	83.68	70.49	-	-	-	-	-
CAS-SAL-FR (IJCV 2022) [53]	ResNet50	84.18	68.59	77.56	83.81	80.56	-	-	-	-	-
VTB (TCSVT 2022) [3]	ViT-B/16	82.67	69.44	78.28	84.39	80.84	81.34	67.48	76.41	83.32	79.35
PARformer (TCSVT 2023) [7]	Swin-L	84.13	69.94	79.63	88.19	81.35	-	-	-	-	-
VTB* (TCSVT 2022) [3]	ViT-L/14	83.69	69.78	78.09	85.21	81.10	81.36	67.58	76.19	84.00	79.52
PromptPAR (arXiv-2023) [46]	ViT-L/14	85.45	71.61	79.64	86.05	82.38	83.14	69.62	77.42	85.73	81.00
SequencePAR (arXiv-2023) [21]	ViT-L/14	-	71.47	82.40	82.09	82.05	-	70.14	81.37	81.22	81.10
MambaPAR (V)	VMamba-B	83.12	69.47	78.03	84.81	80.88	81.91	68.08	76.41	84.38	79.83

Table 3. Comparison with state-of-the-art methods on WIDER datasets. "V" means Vision branch.

Methods	Backbone	mA
R*CNN (ICCV 2015) [8]	VGG16	80.5
DHC (ECCV 2016) [30]	VGG16	81.3
SRN (CVPR 2017) [58]	ResNet101	86.2
DIAA (ECCV 2018) [39]	ResNet101	86.4
Da-HAR (AAAI 2020) [52]	ResNet101	87.3
VAC-Combine (IJCV 2022) [14]	ResNet50	88.4
VTB (TCSVT 2022) [3]	ViT-B/16	88.2
VTB* (TCSVT 2022) [3]	ViT-L/14	91.2
PromptPAR (arXiv-2023) [46]	ViT-L/14	92.0
MambaPAR (V)	VMamba-B	89.38

based on ViT-B/16 achieved mA, Acc, Prec, Recall, and F1 metrics of 75.76, 64.73, 74.93, 80.85, and 77.35, respectively, while our experimental results were more excellent, reaching 77.34, 66.64, 75.37, 83.57, and 78.86, respectively. These results indicate that our model also performs excellently in zero-shot settings, and we believe the key reason is the high adaptability of the Mamba model to pedestrian attribute recognition.

4.4. Ablation Study

Effects of Vision-Semantic Fusion for PAR. As shown in Table 1, we conducted ablation experiments on the PA100K and PETA datasets for the visual semantic fusion (VSF) module. The effectiveness of this module has been validated by integrating with the Transformer networks, but not with the Mamba networks. Thus, we exploit two vision Mamba models, i.e., the VMamba-B and Vim-S, to check whether VSF still works for the Mamba-based PAR framework. It is easy to find that the pure Vim-S based model achieves (PA100K) 81.45, 74.92, 82.68, 84.27, 83.15 and (PETA) 79.28, 76.77, 85.24, 86.58, 85.44 in the five indicators of mA, Acc, Prec, Recall, and F1, respectively. After introducing the VSF module highlighted in the red dashed rectangle, the results reach 80.87, 79.33, 86.48, 88.79, 87.21 on the PA100K dataset and 84.25, 76.07, 82.73, 87.20, 84.56 on the PETA dataset, respectively, which proves the effectiveness of the semantic labels for the Vim-based PAR framework. However, we also find that similar conclusions do not hold for the VMamba-B based PAR framework, that is to say, the performance drops when adopting the semantic labels of all attributes.

Transformer *vs* Mamba for PAR. As shown in Table 5, we compare the performance of ViT-S, Vim-S, ViT-B, and VMamba-B on the PA100K dataset using only the visual branch. One can find that the vision Mamba-based PAR framework achieves comparable or even better performance than vision Transformer based models at the same size. These comparisons demonstrate that it is still a promising research direction to adopt the Mamba for pedestrian attribute recognition.

4.5. Efficiency Analysis

Analysis of Testing Time. As shown in Fig. 4(a), we test the running time of Vim-S, VMamba-B, ViT-S, and ViT-B to check their efficiency. Obviously, the Mamba-based models are slower than the ViT-based models. Also, the VMamba-B needs more time than the Vim-S due to larger parameters. Thus, research on further improving the efficiency of Vision Mamba is needed.

Analysis on GPU Memory Cost. As shown in Fig. 4(c), we compared the memory usage of different PAR models on various datasets. It is easy to find that the VMamba significantly costs more GPU memories (about $\times 2$ times) than the Vim-S, ViT-S, and ViT-B. Interestingly, we can also see that the ViT-B performs the best with batch size 24 or 64. These experimental results indicate that Mamba does not gain an advantage regarding memory usage on pedestrian attribute recognition tasks with limited image resolutions.

Analysis of Parameters and FLOPs. As illustrated in Fig. 4(b, d), the Vim-S model exhibits a notable reduction in parameters compared to the ViT-S and ViT-B models, with no appreciable decline in performance. Additionally, the VMamba-B model displays a more compact parameter count than the ViT-B model, while maintaining comparable performance. It has been demonstrated that the Mamba-based pedestrian attribute recognition method has a reduced number of parameters and facilitates the process of fine-tuning. Furthermore, the computational costs of these methods under different batch sizes were compared. It was observed that as the batch size increased, the Transformer-based PAR method exhibited a notable reduction in computational cost. However, the computational cost of VMamba-B was significantly lower than that of ViT-B, while that of Vim-S remained largely unchanged. This demonstrates that the Mamba-based pedestrian attribute recognition method continues to exhibit reduced computation at high batch sizes and is more hardware-friendly when deployed.

4.6. Comparison of Hybrid Mamba-Transformer

In this paper, we also exploit different versions of hybrid Mamba-Transformer networks for pedestrian attribute recognition, as discussed in sub-section 3.3. We will report and discuss the experimental results of these variations in this sub-section.

As shown in Table 6, among the six hybrid methods we proposed, method (e) achieves the best results, even outperforming the largest-scale ViT models. Others, such as (b) and (d), do not outperform ViT-B in performance, but also outperform ViT-S and Vim-S. Other than these, none of the other approaches resulted in any performance improvement, either because of design issues or because of the hybrid ap-

Table 4. Comparison with state-of-the-art methods on PETA-ZS and RAP-ZS datasets. "V" means Vision branch.

Methods	Backbone	PETA-ZS					RAP-ZS				
		mA	Acc	Prec	Recall	F1	mA	Acc	Prec	Recall	F1
MsVAA (ECCV 2018) [38]	ResNet101	71.53	58.67	74.65	69.42	71.94	72.04	62.13	75.67	75.81	75.74
VAC (CVPR 2019) [15]	ResNet50	71.91	57.72	72.05	70.64	70.90	73.70	63.25	76.23	76.97	76.12
ALM (ICCV 2019) [41]	BN-Inception	73.01	57.78	69.50	73.69	71.53	74.28	63.22	72.96	80.73	76.65
JLAC (AAAI 2020) [40]	ResNet50	73.60	58.66	71.70	72.41	72.05	76.38	62.58	73.14	79.20	76.05
Jia et al. (Arxiv 2021) [20]	ResNet50	71.62	58.19	73.09	70.33	71.68	72.32	63.61	76.88	76.62	76.75
MCFL (NCA 2022) [2]	ResNet50	72.91	57.04	68.47	74.35	71.29	74.37	63.37	71.21	83.86	77.02
VTB (TCSVT 2022) [3]	ViT-B/16	75.13	60.50	73.29	74.40	73.38	75.76	64.73	74.93	80.85	77.35
VTB* (TCSVT 2022) [3]	ViT-L/14	77.18	63.12	74.77	77.24	75.50	79.17	68.34	76.81	84.51	80.07
PromptPAR (arXiv-2023) [46]	ViT-L/14	80.08	66.02	76.53	80.49	77.77	80.43	70.39	78.48	85.57	81.52
SequencePAR (arXiv-2023) [21]	ViT-L/14	-	66.70	78.75	78.52	78.40	-	70.28	82.13	80.55	81.14
MambaPAR (V)	VMamba-B	74.43	60.56	73.12	74.51	73.37	77.34	66.64	75.37	83.57	78.86

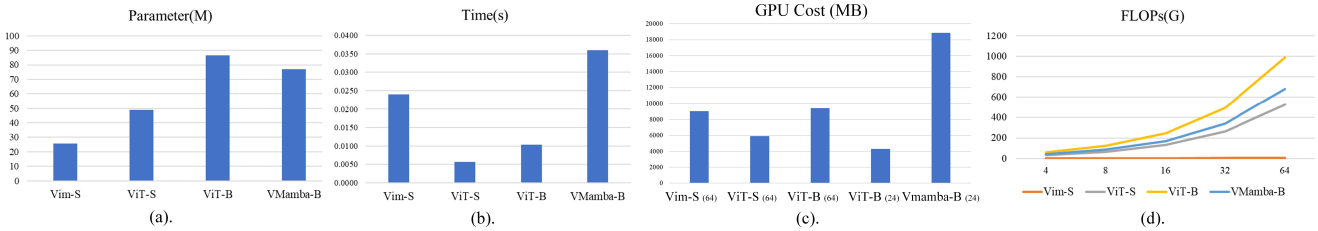


Figure 4. Comparison on (a) model parameters, (b) Time cost, (c) GPU cost, and (d) FLOPs.

Table 5. Results of different backbones on PA100k dataset. (Only Vision branch)

Methods	mA	Acc	Prec	Recall	F1
ViT-S	78.86	77.05	85.83	86.53	85.67
Vim-S	79.28	76.77	85.24	86.58	85.44
ViT-B	83.10	80.61	87.75	89.12	88.06
VMamba-B	83.63	81.11	87.59	89.98	88.40

Table 6. Experimental results of the proposed hybrid Mamba-Transformer networks for PAR on PETA dataset. The best and second results are highlighted in **Red** and **Green**, respectively.

Methods	mA	Acc	Prec	Recall	F1
ViT-B	**83.10**	**80.61**	**87.75**	**89.12**	**88.06**
ViT-S	78.86	77.05	85.83	86.53	85.67
Vim-S	79.28	76.77	85.24	86.58	85.44
(a) PaFusion	75.78	71.74	81.33	83.57	81.91
(b) N-ASF	**81.91**	**79.54**	**86.96**	**88.44**	**87.30**
(c) ASF	70.42	67.28	78.82	79.52	78.56
(d) MaFormer	**81.93**	**78.96**	**86.90**	**87.85**	**86.95**
(e) MaHDFT	**83.64**	**81.11**	**87.90**	**89.66**	**88.40**
(f) AdaMTF	76.64	73.68	83.70	84.00	83.30
(g) KDTM	**81.16**	**79.41**	**87.36**	**87.83**	**87.20**
(h) MaKDF	**80.72**	**78.57**	**86.55**	**87.61**	**86.66**

proach.

From the experimental results reported in Table 6, one can also find that in the method (g) KDTM, we experimented with feature distillation and predictive label distribution distillation. From the results, it seems that feature distillation has a negative effect, which may be caused by the differences between the different frameworks. Therefore, directly distilling from the predicted label distribution level can avoid the difficulty of matching the features of each layer caused by the differences between models. It can be deemed that response-level (i.e., the logits) distillation will give Vim a positive performance boost.

Based on the experimental results of KDTM, we improve the way of feature distillation and propose the method (h) MaKDF, which adds a lightweight transition module (single-layer Vim) between ViT distillation and Vim, and then adds the output of this transition module as additional information to the Vim backbone. As shown in Table 6, this assisted distillation with the Mamba module, which is independent of the backbone, brings some performance improvements, but it is still not better than the effect of direct label distribution distillation.

4.7. Limitation Analysis

Although we have explored some hybrid architectures purely visually and obtained some improvements, we have not explored a suitable scheme for how to use Mamba for multi-modal fusion. From the experimental results, directly applying Mamba to modal fusion cannot obtain a consistent improvement. Therefore, how to apply Mamba to the

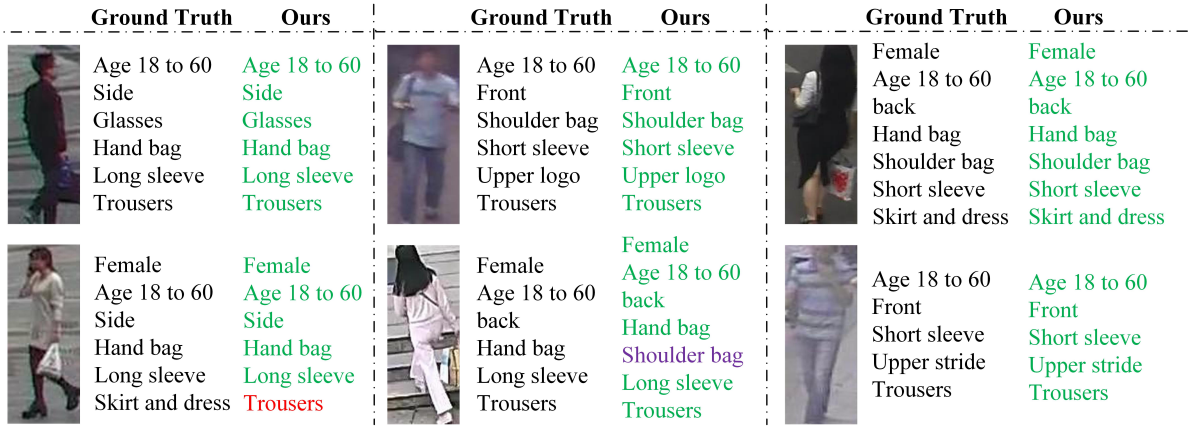


Figure 5. Visualization of the predicted human attributes using our Mamba-based PAR framework.

field of multi-modal fusion still needs to be explored, and whether the hybrid architecture can also reflect the effectiveness of modal fusion is a problem that needs to be explored. We leave these issues as our future works.

5. Conclusion

In this paper, our study has explored the potential of the Mamba architecture in the context of pedestrian attribute recognition (PAR) tasks. While Mamba, with its linear complexity, has shown promise in balancing accuracy and computational cost across various visual tasks, our findings suggest that its advantage in terms of memory usage is not pronounced when dealing with limited resolutions of input images in PAR tasks. The adaptation of Mamba into two typical PAR frameworks yielded mixed results; the text-image fusion approach with Vim showed enhancement potential, whereas the VMamba framework did not benefit from the interaction with attribute tags. Furthermore, our hybrid Mamba-Transformer variants demonstrated that the combination of Mamba with Transformer networks does not universally improve performance but can lead to better outcomes under specific conditions. This work serves as an empirical investigation that we hope will inspire further research into optimizing Mamba for PAR and its potential extension into multi-label recognition domains. The design and comprehensive experimentation with various network structures presented in this paper contribute valuable insights into the development of more efficient and effective models for pedestrian attribute recognition.

References

[1] Abrar H. Abdulnabi, Gang Wang, Jiwen Lu, and Kui Jia. Multi-task cnn model for attribute prediction. *IEEE Transactions on Multimedia*, 17(11):1949–1959, 2015.

[2] Lin Chen andJingkuan Song andXuerui Zhang andMingsheng Shang. Mcfl: multi-label contrastive focal loss for deep imbalanced pedestrian attribute recognition. *Neural Computing and Applications*, 2022.

[3] Xinhua Cheng, Mengxi Jia, Qian Wang, and Jian Zhang. A simple visual-textual baseline for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[4] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

[5] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792, 2014.

[6] Haonan Fan, Hai-Miao Hu, Shuailing Liu, Weiqing Lu, and Shiliang Pu. Correlation graph convolutional network for pedestrian attribute recognition. *IEEE Transactions on Multimedia*, 24:49–60, 2022.

[7] Xinwen Fan, Yukang Zhang, Yang Lu, and Hanzi Wang. Parformer: Transformer-based multi-task network for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[8] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r*cnn. In *International Conference on Computer Vision*, 2015.

[9] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2023.

[10] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33:1474–1487, 2020.

[11] Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35:35971–35983, 2022.

[12] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

[13] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent,

convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021.

[14] Hao Guo, Xiaochuan Fan, and Song Wang. Visual attention consistency for human attribute recognition. *International Journal of Computer Vision*, 130(4):1088–1106, 2022.

[15] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 729–739, 2019.

[16] Kai Han, Yunhe Wang, Han Shu, Chuanjian Liu, Chunjing Xu, and Chang Xu. Attribute aware pooling for pedestrian attribute recognition. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2456–2462. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[17] Ju Huang, Shiao Wang, Shuai Wang, Zhe Wu, Xiao Wang, and Bo Jiang. Mamba-fetrack: Frame-event tracking via state space model. *arXiv preprint arXiv:2404.18174*, 2024.

[18] Jian Jia, Xiaotang Chen, and Kaiqi Huang. Spatial and Semantic Consistency Regularizations for Pedestrian Attribute Recognition. *arXiv e-prints*, page arXiv:2109.05686, Sept. 2021.

[19] Jian Jia, Naiyu Gao, Fei He, Xiaotang Chen, and Kaiqi Huang. Learning disentangled attribute representations for robust pedestrian attribute recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):1069–1077, Jun. 2022.

[20] Jian Jia, Houjing Huang, Xiaotang Chen, and Kaiqi Huang. Rethinking of Pedestrian Attribute Recognition: A Reliable Evaluation under Zero-Shot Pedestrian Identity Setting. *arXiv e-prints*, page arXiv:2107.03576, July 2021.

[21] Jiandong Jin, Xiao Wang, Chenglong Li, Lili Huang, and Jin Tang. Sequencepar: Understanding pedestrian attributes via a sequence generation paradigm. *arXiv preprint arXiv:2312.01640*, 2023.

[22] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, 1960.

[23] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[24] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 111–115, 2015.

[25] D. Li, Z. Zhang, X. Chen, and K. Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE Transactions on Image Processing*, 28(4):1575–1590, 2019.

[26] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A Richly Annotated Dataset for Pedestrian Attribute Recognition. *arXiv e-prints*, page arXiv:1603.07054, Mar. 2016.

[27] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*, 2024.

[28] Qiaozhe Li, Xin Zhao, Ran He, and Kaiqi Huang. Pedestrian attribute recognition by joint visual-semantic reasoning and knowledge distillation. In *Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19*, 2019.

[29] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 684–700. Springer, 2016.

[30] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *European Conference on Computer Vision*, 2016.

[31] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 350–359, 2017.

[32] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[33] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model, 2024.

[34] Eric Nguyen, Karan Goel, Albert Gu, Gordon Downs, Preey Shah, Tri Dao, Stephen Baccus, and Christopher Ré. S4nd: Modeling images and videos as multidimensional signals with state spaces. *Advances in neural information processing systems*, 35:2846–2861, 2022.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[37] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. 2019.

[38] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. Deep Imbalanced Attribute Classification using Visual Attention Aggregation. *arXiv e-prints*, page arXiv:1807.03903, July 2018.

[39] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. Deep Imbalanced Attribute Classification using Visual Attention Aggregation. *arXiv e-prints*, page arXiv:1807.03903, July 2018.

[40] Zichang Tan, Yang Yang, Jun Wan, Guodong Guo, and Stan Z. Li. Relation-aware pedestrian attribute recognition with graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7):12055–12062, 2020.

[41] Chufeng Tang, Lu Sheng, Zhao-Xiang Zhang, and Xiaolin Hu. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4996–5005, 2019.

[42] Zengming Tang and Jun Huang. Drformer: Learning dual relations using transformer for pedestrian attribute recognition. *Neurocomputing*, 497:159–169, 2022.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[44] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Attribute recognition by joint recurrent learning of context and correlation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 531–540, 2017.

[45] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4):447–482, 2023.

[46] Xiao Wang, Jiandong Jin, Chenglong Li, Jin Tang, Cheng Zhang, and Wei Wang. Pedestrian attribute recognition via clip based prompt vision-language fusion, 2023.

[47] Xiao Wang, Shiao Wang, Yuhe Ding, Yuehang Li, Wentao Wu, Yao Rong, Weizhe Kong, Ju Huang, Shihao Li, Haoxiang Yang, et al. State space model for new-generation network alternative to transformers: A survey. *arXiv preprint arXiv:2404.09516*, 2024.

[48] Xiao Wang, Shaofei Zheng, Rui Yang, Aihua Zheng, Zhe Chen, Jin Tang, and Bin Luo. Pedestrian attribute recognition: A survey. *Pattern Recognition*, 121:108220, 2022.

[49] Zihan Wang, Fanheng Kong, Shi Feng, Ming Wang, Han Zhao, Daling Wang, and Yifei Zhang. Is mamba effective for time series forecasting? *arXiv preprint arXiv:2403.11144*, 2024.

[50] Junyi Wu, Yan Huang, Zhipeng Gao, Yating Hong, Jianqiang Zhao, and Xinsheng Du. Inter-attribute awareness for pedestrian attribute recognition. *Pattern Recognition*, 131:108865, 2022.

[51] Jingjing Wu, Hao Liu, Jianguo Jiang, Meibin Qi, Bo Ren, Xiaohong Li, and Yashen Wang. Person attribute recognition by sequence contextual relation learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3398–3412, 2020.

[52] Mingda Wu, Di Huang, Yuanfang Guo, and Yunhong Wang. Distraction-Aware Feature Learning for Human Attribute Recognition via Coarse-to-Fine Attention Mechanism. *arXiv e-prints*, page arXiv:1911.11351, Nov. 2019.

[53] Yang Yang, Zichang Tan, Prayag Tiwari, Hari Mohan Pandey, Jun Wan, Zhen Lei, Guodong Guo, and Stan Z Li. Cascaded split-and-aggregate learning with feature recombination for pedestrian attribute recognition. *International Journal of Computer Vision*, 129(10):2731–2744, 2021.

[54] Yajing Zhai, Yawen Zeng, Zhiyong Huang, Zheng Qin, Xin Jin, and Da Cao. Multi-prompts learning with cross-modal alignment for attribute-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6979–6987, 2024.

[55] Jialiang Zhang, Lixiang Lin, Jianke Zhu, Yang Li, Yun-chen Chen, Yao Hu, and Steven C. H. Hoi. Attribute-aware pedestrian detection in a crowd. *IEEE Transactions on Multimedia*, 23:3085–3097, 2021.

[56] Xin Zhao, Liufang Sang, Guiguang Ding, Yuchen Guo, and Xiaoming Jin. Grouping attribute recognition for pedestrian with joint recurrent learning. In *IJCAI*, volume 2018, page 27th, 2018.

[57] Xin Zhao, Liufang Sang, Guiguang Ding, Jungong Han, Na Di, and Chenggang Yan. Recurrent attention model for pedestrian attribute recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9275–9282, 2019.

[58] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning Spatial Regularization with Image-level Supervisions for Multi-label Image Classification. *arXiv e-prints*, page arXiv:1702.05891, Feb. 2017.

[59] Jianqing Zhu, Liu Liu, Yibing Zhan, Xiaobin Zhu, Huanqiang Zeng, and Dacheng Tao. Attribute-image person re-identification via modal-consistent metric learning. *International Journal of Computer Vision*, 131(11):2959–2976, 2023.

[60] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model, 2024.