RGB-Event based Pedestrian Attribute Recognition: A Benchmark Dataset and An Asymmetric RWKV Fusion Framework — Supplementary Material —

Xiao Wang[1], Haiyang Wang[1], Shiao Wang[1], Qiang Chen[1], Jiandong Jin[2], Haoyu Song[1], Bo Jiang[1]*, Chenglong Li[2]*
[1]School of Computer Science and Technology, Anhui University, Hefei, China
[2]School of Artificial Intelligence, Anhui University, Hefei, China
why2434961256@163.com, {*xiaowang, jiangbo*}*@ahu.edu.cn*, *wsa1943230570@126.com*, {*e23301220, e22214005*}*@stu.ahu.edu.cn*, {*jdjinahu, lcl1314*}*@foxmail.com*

1. Experiments

1.1. Dataset and Evaluation Metric

• **MARS-Attribute Dataset** is an annotated dataset for pedestrian attribute recognition, containing multiple multi-label and binary attributes such as pedestrian action, orientation, clothing color, gender, and others. These attributes are decomposed into 43 binary attributes to enhance the performance of pedestrian attribute recognition models for training and testing. The dataset is divided into a training subset and a testing subset, with 8,298 and 8,062 tracklets, respectively. Each tracklet contains approximately 60 frames on average. The dataset includes 625 and 626 distinct pedestrian identities for training and testing.

• **DukeMTMC-VID-Attribute Dataset** is an extension of the DukeMTMC-VID dataset, specifically designed for pedestrian attribute recognition. This dataset includes various annotated pedestrian attributes aimed at improving pedestrian re-identification performance. It features 2,032 identities and 16,522 video sequences captured across multiple scenarios, with a focus on real-world and challenging environments. In addition, the dataset provides rich attribute annotations. By splitting the multi-label attributes into binary attributes, the dataset includes a total of 36 binary attributes for training and testing. The training subset contains 702 distinct pedestrian identities and 16,522 images, while the testing subset includes 17,661 images corresponding to 702 pedestrians. The DukeMTMC-VID-Attribute dataset is widely used for evaluating attribute-based pedestrian re-identification models, offering valuable insights into how attributes influence recognition accuracy under varying conditions.

The evaluation metrics consist of **mean Accuracy** (mA), **Accuracy** (Acc), **Precision** (Prec), **Recall**, and **F1-score** (F1), which are formulated as follows:

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \tag{1}$$

$$Precision = \frac{TP}{FP + TP}, \quad Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{3}$$

where TP denotes the number of correctly predicted positive samples, TN is the number of correctly predicted negative samples, FP and FN denote the number of false positive and false negative samples, respectively.

1.2. Implementation Details

In our experiments, we use the VRWKV6-B version of the VRWKL [2] base model pre-trained on the ImageNet-1K [1] dataset as the visual encoder. This version consists of 12 block layers. We select SGD as the optimizer, We leverage the warm-up strategy and increase the learning rate from 0 to the initial learning rate 0.008 linearly in the first 10 epochs, and decrease the learning rate by a factor of 0.1 when the number of iterations increases, we set the batch size to 16 and train for 60 epochs, the filtering threshold is set to 0.75. During both the training and inference stages, we first pad and resize the images to 256×128. Training images are augmented with random horizontal flipping with a probability of 0.5 and random cropping with a padding size of 10. Feature interaction is performed using the last layer of the transformer. Our model is implemented based on the PyTorch deep learning framework, and the experiments are conducted on a server equipped with an NVIDIA RTX 3090 GPU. For more details about our framework, please refer to our source code.

*Corresponding Author: Bo Jiang, Chenglong Li

Table 1. Comparison of different threshold setting in the feature filtering module.

Threshold	mA	Acc	Prec	Recall	F1
0.60	87.70	84.14	88.56	89.07	88.68
0.65	87.55	84.16	88.62	89.09	88.70
0.75	87.70	84.94	89.15	89.48	89.18
0.80	87.62	84.17	88.77	88.94	88.71
0.95	86.53	82.46	87.37	87.49	87.12

1.3. Ablation Study

• **Analysis of different threshold setting in the similarity aggregation strategies** To further validate the impact of threshold settings in the similarity-based filtering strategy, we conducted experiments with varying similarity thresholds under identical conditions. The threshold controls the degree of fusion between event frames, preserving only those with sufficient semantic similarity to reduce redundant or noisy features. As shown in Table 1, a well-chosen threshold better preserves discriminative features and enhances performance, while extreme thresholds may impair fusion quality and degrade results.

• **Analysis of comparisons with standard fusion baselines.** As shown in Table 2, we further introduce several fusion comparison settings, including cross-attention, late fusion, token pruning baselines, and modality dropout. The experimental results further validate the effectiveness and design rationale of the proposed asymmetric architecture.

Table 2. Comparison of standard feature fusion methods.

Method	mA	Acc	Prec	Recall	F1
cross-attention	87.16	83.41	88.32	88.92	88.69
late fusion	86.62	82.93	87.53	88.72	88.06
token pruning	86.49	82.47	87.65	88.18	87.76
modality dropout	86.55	82.43	87.72	88.34	87.89
OTN-RWKV(ours)	87.70	84.94	89.15	89.48	89.18

• **Efficiency Analysis of RWKV.** As shown in Table 3, we further conduct ablation experiments to analyze and compare RWKV with other backbones in terms of FLOPs, latency, memory usage, as well as mA and F1. The results demonstrate that RWKV achieves a favorable balance between computational efficiency and performance in event-based sensing tasks.

• **Analysis of the benefit of the event modality.** As shown in Fig. 1, the upper part of the figure presents a comparison of 17 SOTA methods on the EventPAR dataset under RGB-only and RGB+Event settings. The results demonstrate that incorporating event data significantly improves overall performance and provides effective complementary support for attribute recognition. In addition, the lower part illustrates that when RGB data is degraded, the event modality can provide clear motion cues, thereby improving the recognition of certain attributes.

Table 3. Comparison of efficiency across different backbones.

Backbone	FLOPs	Latency	Memory	mA	F1
ViT	64.4G	31.47ms	444M	84.05	87.60
ResNet50	56.2G	23.11ms	550M	83.35	87.64
RWKV	59.6G	28.93ms	537M	87.70	89.18

Methods	mA	Acc	Prec	Recall	F1
DeepMAR	71.15 / 66.57	72.39 / 69.53	86.42 / 74.90	78.93 / 88.54	82.51 / 81.57
ALM	71.31 / 57.18	68.80 / 64.17	78.82 / 75.59	75.97 / 73.20	77.37 / 74.38
Strong Baseline	73.75 / 73.75	70.63 / 61.86	81.34 / 67.23	78.26 / 80.78	79.32 / 75.43
RethinkingPAR	71.23 / 81.37	69.00 / 80.84	80.80 / 86.31	77.16 / 87.57	78.94 / 86.93
SSCNet	72.12 / 63.10	69.95 / 66.07	78.36 / 72.72	80.10 / 83.22	79.22 / 77.62
VTB	76.96 / 88.41	73.81 / 83.83	80.74 / 87.89	83.15 / 89.31	81.67 / 88.53
Label2Label	72.47 / 72.49	69.54 / 74.01	78.75 / 86.60	79.14 / 79.02	78.55 / 82.19
DFDT	70.14 / 61.71	70.95 / 63.14	78.71 / 79.17	81.82 / 70.63	80.24 / 74.66
Zhou et al.	70.16 / 56.46	67.41 / 60.89	79.76 / 73.37	76.40 / 73.62	78.04 / 73.50
PARFormer	75.33 / 83.12	74.21 / 80.48	79.33 / 85.14	85.82 / 88.41	82.09 / 86.53
VTB-PLIP	73.72 / 67.25	71.50 / 68.37	79.97 / 77.75	80.57 / 79.72	79.99 / 78.37
Rethink-PLIP	57.91 / 68.75	62.78 / 70.03	78.27 / 81.82	70.81 / 78.04	74.35 / 79.89
PromptPAR	74.92 / 86.51	71.79 / 82.27	78.33 / 86.35	83.01 / 89.36	80.31 / 87.64
SSPNet	71.71 / 66.92	69.41 / 67.49	80.46 / 78.73	77.46 / 76.90	78.93 / 77.80
MambaPAR	72.28 / 50.01	70.57 / 42.32	77.99 / 54.81	81.70 / 57.31	79.50 / 55.63
MaHDFT	75.54 / 50.43	72.27 / 44.98	80.28 / 59.10	82.19 / 59.70	80.59 / 58.57
SequencePAR	73.96 / 86.27	73.40 / 84.42	81.47 / 88.81	81.17 / 89.12	81.11 / 88.83
OTN-RWKV (Ours)	**79.32 / 87.70**	**76.00 / 84.94**	**82.37 / 89.15**	**84.55 / 89.48**	**83.22 / 89.18**

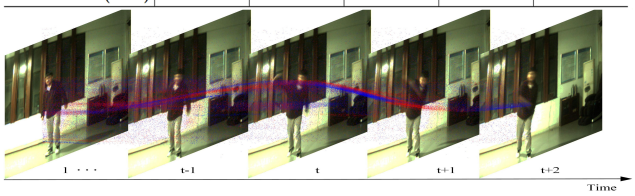

Figure 1. The upper part presents a comparison between RGB-only (left) and RGB+Event (right) results, while the lower part visualizes the spatiotemporal alignment and information interaction between the RGB and event streams. after filtering

1.4. Visualization

• **Visualization of Aggregation Strategy.** To provide an intuitive understanding of our proposed similarity-based aggregation strategy, we present a visual illustration in Fig. 2. In this figure, the colored regions represent event tokens that are retained and contribute to the final aggregation, while the black areas indicate tokens filtered out by the similarity-based selection process. As shown, the strategy effectively suppresses redundant or irrelevant information by adaptively retaining only semantically relevant tokens. This helps the model focus on discriminative patterns, thereby improving the overall recognition performance.

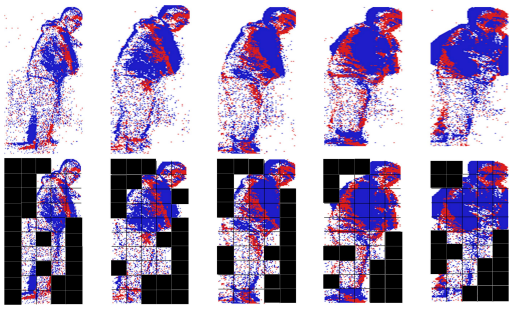

Figure 2. Illustration of our proposed similarity-based aggregation strategy for event token selection, where the black parts indicate the filtered-out tokens.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[2] Yuchen Duan, Weiyun Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Hongsheng Li, Jifeng Dai, and Wenhai Wang. Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. *arXiv preprint arXiv:2403.02308*, 2024. 1