Learning CLIP Guided Visual-Text Fusion Transformer for Video-based Pedestrian Attribute Recognition

Jun Zhu[1†], Jiandong Jin[2†], Zihan Yang[1], Xiaohao Wu[1], Xiao Wang[1*]
[1] School of Computer Science and Technology, Anhui University, Hefei 230601, China
[2] School of Artificial Intelligence, Anhui University, Hefei 230601, China

Abstract

Existing pedestrian attribute recognition (PAR) algorithms are mainly developed based on a static image. However, the performance is not reliable for images with challenging factors, such as heavy occlusion, motion blur, etc. In this work, we propose to understand human attributes using video frames that can make full use of temporal information. Specifically, we formulate the video-based PAR as a vision-language fusion problem and adopt pre-trained big models CLIP to extract the feature embeddings of given video frames. To better utilize the semantic information, we take the attribute list as another input and transform the attribute words/phrase into the corresponding sentence via split, expand, and prompt. Then, the text encoder of CLIP is utilized for language embedding. The averaged visual tokens and text tokens are concatenated and fed into a fusion Transformer for multi-modal interactive learning. The enhanced tokens will be fed into a classification head for pedestrian attribute prediction. Extensive experiments on a large-scale video-based PAR dataset fully validated the effectiveness of our proposed framework. Both the source code and pre-trained models will be released at https://github.com/Event-AHU/VTF_PAR.

1. Introduction

Pedestrian Attribute Recognition (PAR) [2, 17] is a very important research topic in computer vision and gets boosted greatly with the help of deep learning. Many representative PAR models are proposed in recent years based on convolutional neural networks (CNN) [5], and recurrent neural networks (RNN) [3]. Wang et al. [14] propose the JRL which learns the attribute context and correlation in a joint recurrent learning manner using LSTM [6]. The self-attention based Transformer networks are first proposed to handle the natural language processing tasks and then are borrowed into the computer vision community [4, 13, 15, 16, 19] due to their great performance. Some researchers also exploit the Transformer for the PAR problem to model the global context information [2, 12]. DRFormer [12] is proposed to capture the long-range relations of regions and relations of attributes. VTB [2] is also developed to fuse the image and language information for more accurate attribute recognition. In addition to understanding the pedestrian images using the attributes, this task also serves other computer vision problems, such as object detection [18], person re-identification [20], etc. Despite the great success of PAR, these works are developed based on a single RGB frame only which ignores the temporal information and maybe obtains sub-optimal results in practical scenarios.

As mentioned in work [1], the video frames can provide more comprehensive visual information for the specific attribute, but the static image fails to. The authors propose to understand human attributes using video clips and propose large-scale datasets for video-based PAR. They also build a baseline by proposing the multi-task video-based PAR framework based on CNN and temporal attention. Better performance can be obtained on their benchmark datasets, however, we think the following issues still limit their overall results. **Firstly**, they adopt CNN as the backbone network to extract the feature representation of input images which learns the local features well. As is known to all, global relation in the pixel-level space is also very important for fine-grained attribute recognition. Several researchers resort to the Transformer network to capture such global information [4, 13], however, their models can work for image-based attribute recognition only. **Secondly**, the authors formulate the video-based PAR as a multi-task classification problem and try to learn a mapping from a given video to attributes. The attribute labels are transformed into binary vectors for network optimization. However, the high-level semantic information is greatly lost which is very important for pedestrian attribute recognition.

To address the aforementioned two issues, in this paper, we propose a novel CLIP-guided Visual-Text Fusion

*† denotes equal contribution. Corresponding author: Xiao Wang, email: xiaowang@ahu.edu.cn

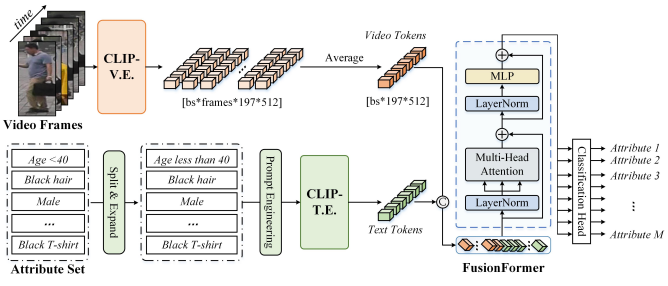

Figure 1. An overview of our proposed CLIP-guided Visual-Text Fusion Transformer for video-based PAR.

Transformer for Video-based Pedestrian Attribute Recognition. As shown in Fig. 1, we take the video frames and attribute set as the input and formulate the video-based PAR as a multi-modal fusion problem. To be specific, the video frames are transformed into video tokens using a pre-trained CLIP [11] which is a multimodal big model. The attribute set is transformed into corresponding language descriptions using split, expand, and prompt engineering. Then, the text encoder of CLIP is used for the language embedding. After that, we concatenate the video and text tokens and feed them into a fusion Transformer for multi-modal information interaction which mainly contains layer normalization, multi-head attention, and MLP (Multi-Layer Perceptron). The output will be fed into a classification head for pedestrian attribute recognition.

To sum up, the main contributions of this paper can be concluded as following two aspects:

• We propose a novel CLIP-guided Visual-Text Fusion Transformer for Video-based Pedestrian Attribute Recognition, which is the first work to address the video-based PAR from the perspective of visual-text fusion.

• We introduce the pre-trained big model CLIP as our backbone network, which makes our model robust to the aforementioned challenging factors. Extensive experiments validated the effectiveness of our proposed model.

2. Methodology

In this section, we will introduce our proposed framework from the following aspects, i.e., Input Processing and Embedding, Multimodal Fusion Transformer, and Optimization.

2.1. Input Processing and Embedding

Given the video frames $V = \{v_1, v_2, ..., v_T\}$ and attribute list $A = \{a_1, a_2, ..., a_M\}$, we first process these inputs to better utilize the pre-trained CLIP model. More in detail, the input frames are padded with zero pixels into a resolution 224×224, as the initial frames are slim but the pre-trained CLIP takes the fixed square resolution as the input. The ViT-B/16-based CLIP model is selected as the backbone considering its efficiency and accuracy. Therefore, the input frames are embedded into a set of visual tokens $T \times 197 \times 512$, here, 197 is the number of tokens and 512 is the dimension of each token. Then, we average this feature into a tensor $F_v \in \mathbb{R}^{197\times512} = \{f^1, f^2, ..., f^T\}$ along the temporal channel.

For the attribute set A, we first need to process them into corresponding language descriptions to make full of CLIP text encoder. Specifically, we first split and expand each attribute to get the natural phrases. For example, "*Age* $\leq$ *40*" is processed into "*age less than 40*". Then, prompt engineering is adopted to further process the phrases into natural language descriptions using carefully designed prompt templates. For example, *age less than 40* is transformed into "*the pedestrian has an attribute age less than 40*". After all the attributes are processed, we adopt the text encoder of CLIP to get the text tokens $F_t = \{t^1, t^2, ..., t^M\}$. Then, the video and text tokens are concatenated together $[F_v, F_t]$ as the input of fusion Transformer.

2.2. Multimodal Fusion Transformer

The Transformer network has been widely validated in its effectiveness in many research domains, including computer vision, natural language processing, and multimodal fusion [15]. In this work, we adopt the Transformer to fuse the vision and language information to enhance the multimodal feature representation learning. Given the multimodal input $[F_v, F_t]$, we first adopt a layer normalization layer to process the input. Then, the input is transformed into three branches, i.e., Q, K, and V. We adopt the multi-head self-attention to learn the long-range global relations, and the basic operation of each self-attention layer can be described as $MLP(\mathrm{SAttn}(Q, K)V)$, where the SAttn is $\mathrm{SAttn}(Q, K) = Softmax(\frac{QK^{Tr}}{\sqrt{c}})$, where c denotes feature dimension, Tr is the transpose operation. The output tokens will be fed into a classification head for attribute prediction. Because the MARS dataset has 43 attributes, in our practical implementation, 43 fully connected layers are used as the classification head.

2.3. Optimization

In this work, the video-based PAR task is formulated as a video-text fusion problem. Given the annotated attribute and raw video, we can train our framework in an end-to-end manner using supervised learning. The binary cross-entropy loss function is adopted for the optimization.

3. Experiments

3.1. Dataset, Metric, and Implementation Details

In our experiments, the MARS dataset [1] proposed by Chen et al. is used for both training and testing. The training

Table 1. Results on MARS video-based PAR dataset. w/o denotes without the following module.

Methods	Backbone	MARS		
		Prec	Recall	F1
3DCNN [7]	-	-	-	61.87
CNN-RNN [9]	-	-	-	70.42
VideoPAR (Image) [1]	ResNet50	-	-	67.28
VideoPAR (Video) [1]	ResNet50	-	-	72.04
VTB [2]	ViT-B/16	**78.96**	**78.42**	**78.32**
Ours	ViT-B/16	**81.76**	**82.95**	**81.94**
Improvements	-	**+2.80**	**+4.53**	**+3.62**
Ours (w/o FusionFormer)	ViT-B/16	77.60	81.32	78.69

subset of MARS contains 8,298 tracklets from 625 people, and the testing subset contains 8,062 tracklets corresponding to 626 pedestrians. For each tracklet, there are 60 frames on average. For the evaluation of our and the compared PAR models, we adopt the widely used Precision, Recall, and F1-score as the evaluation metric. Note that, the results reported in our experiments are obtained by averaging these metrics for multiple attribute groups.

The ViT-B/16 version of pre-trained CLIP is used in our experiments. In the training phase, the parameters of CLIP encoder are fixed. The learning rate of our model is 0.001, weight decay is 1e-4. Our model is trained for a total of 20 epochs. The Adam [8] is adopted as our optimizer. Our model is implemented using Python and PyTorch [10] framework and trained on a server with RTX3090s.

3.2. Compare with Other SOTA Models

In the experiments, we compare our model with multiple strong baseline methods on the MARS dataset, including 3DCNN [7], CNN-RNN [9], VideoPAR [1], and VTB [2]. As shown in Table 1, we can find that our model beats all these compared methods by a large margin. Specifically, the VTB [2] achieves $78.96, 78.42, 78.32$ on the Precision, Recall, and F1-score on this dataset, meanwhile, ours are $81.76, 82.95, 81.94$, the improvements are $+2.80, +4.53, +3.62$ on these metrics. Our results are also better than the VideoPAR proposed by Chen et al. (the video-based version, F1 score 72.04) by exceeding $+9.9$. For the fine-grained attribute results, we report them in Table 2. These experiments fully validated the effectiveness and advantages of our model.

3.3. Ablation Study

Component Analysis. In our proposed framework, the *fusion Transformer* and *pre-trained CLIP backbone* are our key components. In this section, we analyze the two components and report the recognition results in Table 1. The VTB [2] is our baseline which adopts the standard ViT-B/16 model as the backbone, and it achieves 78.96/78.42/78.32 on Precision, Recall, and F1-score. When the CLIP model is used, the results can be improved to 81.76/82.95/81.94,

Table 2. Results on MARS video-based PAR dataset. F1-score are reported for all the assessed attributes.

Attribute	VideoPAR (Image)	3DCNN	CNN-RNN	VideoPAR (Video)	Ours
top length	58.72	56.37	65.18	**71.61**	**97.26**
bottom length	92.29	89.35	93.33	**93.90**	**93.69**
shoulder bag	72.57	61.30	**75.89**	**76.08**	65.61
backpack	85.95	76.58	**87.17**	**87.62**	82.08
hat	57.57	57.69	**77.74**	**77.84**	72.76
hand bag	62.82	59.90	**71.68**	**73.55**	59.08
hair	**86.91**	82.77	87.11	**88.17**	86.37
gender	90.89	85.75	92.44	**92.50**	**92.88**
bottom type	81 .69	72.86	84.16	**86.62**	**97.21**
pose	56.91	47.69	58.36	**61.36**	**74.84**
motion	39.39	33.64	**43.92**	43.69	**93.50**
top color	**72.72**	65.63	69.28	71.44	**74.97**
bottom color	**44.63**	40.39	39.68	43.98	**69.76**
age	38.87	36.22	39.93	**40.21**	**87.07**
Average-F1	67.28	61.87	70.42	**72.04**	**81.94**

Figure 2. Visualization of our predictions and Ground Truth (GT).

Table 3. Results with different input frames on MARS dataset.

# Frames	6	4	2	1
Precision	81.76	81.23	79.56	76.43
Recall	82.95	82.32	81.20	78.88
F1-score	81.94	81.39	79.97	77.27

which validated the effectiveness of the pre-trained big model for video-based PAR. When replacing the FusionFormer using regular fully connected layers, the results are dropped from $81.76, 82.95, 81.94$ to $77.60, 81.32, 78.69$, which demonstrates that this fusion module also contributes to our final performance.

Analysis on Input Video Frames. The number of video frames plays an important role in video-based pedestrian attribute recognition. In this part, we train and test different numbers of input frames on the MARS dataset and report the experimental results in Table 3. We can find that the performance can be gradually improved with the increase of video frames, i.e., the F1-score is ranging from 77.27 to 81.94.

Visualization. In addition to the aforementioned quantitative analysis, we also give a qualitative analysis in this subsection. As shown in Fig. 2, we can find that our model predicts human attributes accurately.

4. Conclusion

Different from the mainstream image-based PAR, in this paper, we propose a novel CLIP-guided Visual-Text Fusion Transformer for video-based pedestrian attribute recognition, which makes full use of temporal information. More in detail, we formulate the video-based PAR as a vision-language fusion problem and adopt pre-trained big models CLIP to extract the feature embeddings of given video frames. To better utilize the semantic information, we take the attribute list as another input and fuse it with video tokens using a fusion Transformer. The enhanced tokens will be fed into a classification head for pedestrian attribute prediction. We conduct extensive experiments on a large-scale video-based PAR dataset and demonstrate that our model obtains superior recognition performance. In our future works, we will consider designing fine-grained partial region mining modules to realize a higher performance attribute recognition. Also, new prompt learning/tuning techniques are worthy to be exploited for pre-trained multi-modal big model guided video-based pedestrian attribute recognition.

Acknowledgement This paper is supported by the National Natural Science Foundation of China NO. 62102205.

References

[1] Zhiyuan Chen, Annan Li, and Yunhong Wang. A temporal attentive approach for video-based pedestrian attribute recognition. In *Pattern Recognition and Computer Vision: Second Chinese Conference, PRCV 2019, Xi'an, China, November 8–11, 2019, Proceedings, Part II 2*, pages 209–220. Springer, 2019.

[2] 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*, 32(10):6994–7004, 2022.

[3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[9] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2016.

[10] 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.

[11] 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.

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

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[14] 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.

[15] 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. *arXiv preprint arXiv:2302.10035*, 2023.

[16] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13763–13773, 2021.

[17] 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.

[18] Jialiang Zhang, Lixiang Lin, Jianke Zhu, Yang Li, Yun-chen Chen, Yao Hu, and Steven CH Hoi. Attribute-aware pedestrian detection in a crowd. *IEEE Transactions on Multimedia*, 23:3085–3097, 2020.

[19] Haojie Zhao, Xiao Wang, Dong Wang, Huchuan Lu, and Xiang Ruan. Transformer vision-language tracking via proxy token guided cross-modal fusion. *Pattern Recognition Letters*, 168:10–16, 2023.

[20] Aihua Zheng, Peng Pan, Hongchao Li, Chenglong Li, Bin Luo, Chang Tan, and Ruoran Jia. Progressive attribute embedding for accurate cross-modality person re-id. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4309–4317, 2022.