Learning CLIP for Video-based Pedestrian Attribute Recognition

Jun Zhu 23020241154480[1] Hai Jin 23020241154405[1] Kaiyun Wang 23020241154445[1] Lichao Chen 23020241154374[1] Junnan Ren 31520241154523[1]

[1]School of Informatics

Abstract

Existing pedestrian attribute recognition (PAR) algorithms are primarily developed based on static images, but their performance is not reliable when dealing with challenging factors such as severe occlusion and motion blur. This study proposes a method to understand human attributes using video frames to make full use of temporal information. Given that the large model CLIP performs well in aligning visual and language modalities, we formulate the video-based PAR problem as a vision-language fusion problem and utilize the pre-trained CLIP model to extract feature embeddings from the provided video frames.

Introduction

Pedestrian Attribute Recognition (PAR) (Wang et al. 2022; Cheng et al. 2022) 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) (He et al. 2016a), and recurrent neural networks (RNN) (Chung et al. 2014). Wang et al. (Wang et al. 2017) propose the JRL which learns the attribute context and correlation in a joint recurrent learning manner using LSTM (Hochreiter and Schmidhuber 1997). 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 (Vaswani et al. 2017; Dosovitskiy et al. 2020; Wang et al. 2023; Zhao et al. 2023; Wang et al. 2021) due to their great performance. Some researchers also exploit the Transformer for the PAR problem to model the global context information (Tang and Huang 2022; Cheng et al. 2022). DR-Former (Tang and Huang 2022) is proposed to capture the long-range relations of regions and relations of attributes. VTB (Cheng et al. 2022) 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 (Zhang et al. 2020), person re-identification (Zheng et al. 2022), 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 (Chen, Li, and Wang 2019), 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 (Dosovitskiy et al. 2020; Vaswani et al. 2017), 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 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 (Radford et al. 2021) 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

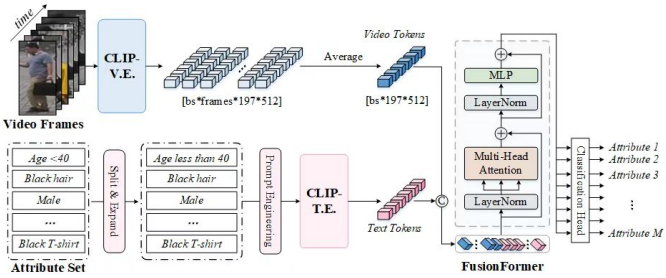

Figure 1: An overview of our proposed CLIP-guided Visual-Text Fusion Transformer for video-based PAR.

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.

Related Work

Pedestrian Attributr Recognition. Current pedestrian attribute recognition (PAR) can be divided into two main approaches: RGB frame-based and video-based methods. In RGB frame-based PAR, early research primarily focused on analyzing pedestrian attributes through multi-label classification using convolutional neural networks (CNNs).Recently, Transformers, with self-attention as their core operation, have gained increasing attention in the artificial intelligence community. Many PAR methods have also been developed based on Transformer networks. For instance, one approach introduces a PARformer to extract features by integrating global and local perspectives, replacing traditional CNNs. Another method, VTB, presents a novel baseline by incorporating an additional text encoder to enable interaction between different types of information.Video-based pedestrian attribute recognition (PAR) leverages temporal information, making it more effective than static image-based methods in handling challenging scenarios such as occlusion and motion blur. Early approaches primarily used convolutional neural networks (CNNs) to extract features from multiple frames and applied temporal aggregation methods like average pooling or RNNs, but they struggled to capture dynamic patterns between frames. With the growing adoption of Transformers, models utilizing self-attention mechanisms have been introduced, enabling the effective modeling of both global and local temporal dependencies in video PAR. Additionally, the emergence of large pre-trained models (e.g., CLIP) has advanced the development of vision-language fusion methods, aligning video frames with attribute descriptions and enabling multi-modal learning through Transformers, significantly improving recognition performance.

Pretrained Vision-Language Models. Integrating pre-trained vision-language models into pedestrian attribute recognition (PAR) has opened new opportunities for enhancing the performance of traditional methods. While early PAR approaches primarily focused on extracting visual features using convolutional neural networks (CNNs) and performing multi-label classification, these methods faced limitations when dealing with complex scenarios such as occlusion, motion blur, or ambiguous visual cues. The recently adopted vision-language models provide an innovative solution by combining semantic information from visual inputs and attribute descriptions.

Pretrained models like CLIP and ALIGN demonstrate significant potential in PAR by aligning visual and textual features in a shared latent space. These models are trained on large-scale image-text pairs, enabling the system to recognize attributes through text prompts, even achieving zero-shot learning without additional task-specific fine-tuning. This capability addresses a key challenge in PAR—handling unseen or rare attributes—by transforming attribute prediction into a multi-modal alignment task.

Method

In this section, we elaborate on our proposed framework, describing its Input Processing and Embedding, Multimodal Fusion Transformer, and Optimization.

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 preprocess the inputs to better utilize the pre-trained CLIP model.

The initial video frames V are zero-padded to the resolution of 224×224 which is required by the pre-trained CLIP model. The padded frames are then sliced into patches and processed by the CLIP visual encoder to obtain embedded tokens. Consequently, the input frames are embedded into a set of visual tokens $T \times N \times d$ where N is the number of tokens in a frame and d is the dimension of each token. We average these features into a tensor $F_v \in R^{N \times d} = \{f^1, f^2, ..., f^T\}$ along the temporal channel. In our case, $N = 197$ and $d = 512$ since we select ViT-B/16-based CLIP as the backbone.

The attribute set A is processed into corresponding natural language descriptions to make full use of CLIP's text encoder. Specifically, we split and expand each attribute to obtain the corresponding natural phrases. For example, "*Age* $\leq$ *40*" is processed into "*age less than 40*". Then, prompt engineering is adopted to further transform 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 $F_t = \{t^1, t^2, ..., t^M\}$.

Finally, the video tokens F_v and text tokens F_t are concatenated into $[F_v, F_t]$ as the input of fusion Transformer.

Multimodal Fusion Transformer

We adopt the multimodal Transformer to fuse and enhance vision and language features. As shown in Fig. , the input vision and language tokens are first normalized by LayerNorm(Xiong et al. 2020; Ba 2016), and then, these tokens are linearly transformed into three branches: query,

key and value, namely Q, K and V. Multi-Head Attention (MHA)(Vaswani et al. 2017) fuses the features of these tokens. Formally, the MHA layer can be expressed as $\mathrm{MHA}(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d}})V$, where K^T denotes the transposed K and d denotes the feature dimensions. The enchanced tokens are further transformed by another LayerNorm and an MLP. Two residual connections(He et al. 2016b) are used following the original Transformer achitecture.

Because the MARS(Zheng et al. 2016) dataset has 43 attributes, in our practical implementation, 43 fully connected layers are used as the classification head.

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.

Experiments

Dataset, Metric, and Implementation Details

In our experiments, the MARS dataset (Chen, Li, and Wang 2019) and DukeMTMC-VID-Attribute dataset (Ristani et al. 2016) proposed by Chen et al. are used for both training and testing. The training 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. The training subset of DukeMTMC-VID contains 702 different ID pedestrians and 16522 images and the testing subset contains 17661 images corresponding to 702 pedestrians. For each sequence, there are 169 frames on average, from which we randomly selected 6 frames for training and testing.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 (Kingma and Ba 2014) is adopted as our optimizer. Our model is implemented using Python and PyTorch (Paszke et al. 2019) framework and trained on a server with RTX3090s.

Compare with Other SOTA Models

In the experiments, we compare our model with multiple strong baseline methods on the MARS dataset, including 3DCNN (Ji et al. 2012), CNN-RNN (McLaughlin, Del Rincon, and Miller 2016), VideoPAR (Chen, Li, and Wang 2019), and VTB (Cheng et al. 2022). As shown in Table 1, we can find that our model beats all these compared methods by a large margin. Specifically, the VTB (Cheng et al. 2022) 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.

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

Methods	Backbone	MARS			DukeMTMC-VID		
		Prec	Recall	F1	Prec	Recall	F1
3DCNN	-	-	-	61.87	-	-	62.93
CNN-RNN	-	-	-	70.42	-	-	71.63
VideoPAR(Image)	ResNet50	-	-	67.28	-	-	69.66
VideoPAR(Video)	ResNet50	-	-	72.04	-	-	68.71
VTB	ViT-B/16	**78.96**	**78.42**	**78.32**	**77.23**	**81.44**	**78.83**
Ours	ViT-B/16	**81.76**	**82.95**	**81.94**	**78.19**	**83.18**	**80.45**
Improvements	-	**+2.80**	**+4.53**	**+3.62**	**+0.60**	**+0.13**	**+0.53**

Table 2: Results on MARS and DUKE 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**

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 (Cheng et al. 2022) 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, 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.

Visualization. In addition to the aforementioned quantitative analysis, we also give a qualitative analysis in this sub-

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

section. As shown in Fig. 2, we can find that our model predicts human attributes accurately.

Conclusion

In this paper, we formulate the video-based PAR task as a vision-language fusion problem, and resolve it with pre-trained CLIP encoder and Visual-Text Fusion Transformer. More in detail, we extract features embeddings of given video frames with CLIP, and fuse them with attribute list to better utilize the semantic informaction. The fusion is achieved by our adopted Visual-Text Transformer. The enhanced tokens are fed to a classification head for PAR. We conduct extensive experiments on MARS, a large-scale video-based PAR dataset, and demonstrate that our model reaches superior recognition performance.

In our future work, we will design finer-grained partial region mining modules to realize higher performance in PAR. Besides, it can be worthwhile to introduce advanced prompt learning and tuning techniques to large model guided PAR.

References

Ba, J. L. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Chen, Z.; Li, A.; and Wang, Y. 2019. 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*, 209–220. Springer.

Cheng, X.; Jia, M.; Wang, Q.; and Zhang, J. 2022. A Simple Visual-Textual Baseline for Pedestrian Attribute Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10): 6994–7004.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2012. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1): 221–231.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

McLaughlin, N.; Del Rincon, J. M.; and Miller, P. 2016. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1325–1334.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, 17–35. Springer.

Tang, Z.; and Huang, J. 2022. DRFormer: Learning dual relations using Transformer for pedestrian attribute recognition. *Neurocomputing*, 497: 159–169.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Zhu, X.; Gong, S.; and Li, W. 2017. Attribute recognition by joint recurrent learning of context and correlation. In *Proceedings of the IEEE International Conference on Computer Vision*, 531–540.

Wang, X.; Chen, G.; Qian, G.; Gao, P.; Wei, X.-Y.; Wang, Y.; Tian, Y.; and Gao, W. 2023. Large-scale multi-modal pre-trained models: A comprehensive survey. *arXiv preprint arXiv:2302.10035*.

Wang, X.; Shu, X.; Zhang, Z.; Jiang, B.; Wang, Y.; Tian, Y.; and Wu, F. 2021. 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*, 13763–13773.

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

Xiong, R.; Yang, Y.; He, D.; Zheng, K.; Zheng, S.; Xing, C.; Zhang, H.; Lan, Y.; Wang, L.; and Liu, T. 2020. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, 10524–10533. PMLR.

Zhang, J.; Lin, L.; Zhu, J.; Li, Y.; Chen, Y.-c.; Hu, Y.; and Hoi, S. C. 2020. Attribute-aware pedestrian detection in a crowd. *IEEE Transactions on Multimedia*, 23: 3085–3097.

Zhao, H.; Wang, X.; Wang, D.; Lu, H.; and Ruan, X. 2023. Transformer vision-language tracking via proxy token guided cross-modal fusion. *Pattern Recognition Letters*, 168: 10–16.

Zheng, A.; Pan, P.; Li, H.; Li, C.; Luo, B.; Tan, C.; and Jia, R. 2022. Progressive Attribute Embedding for Accurate Cross-modality Person Re-ID. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4309–4317.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. Mars: A video benchmark for large-scale person re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, 868–884. Springer.