arXiv:2304.10091v1 [cs.CV] 20 Apr 2023

用于基于视频的行人属性识别的 CLIP 引导视觉-文本融合 Transformer

Jun Zhu[1†], Jiandong Jin[2†], Zihan Yang[1], Xiaohao Wu[1], Xiao Wang[1*‡] 计算机科学与技术学院，安徽大学，中国合肥 230601[2] 人工智能学院，安徽大学，中国合肥 230601

摘要

现有的行人属性识别（*PAR*）算法主要基于静态图像开发。然而，对于存在严重遮挡、运动模糊等挑战因素的图像，其性能并不可靠。在本研究中，我们提出利用视频帧理解人体属性，从而充分利用时序信息。具体而言，我们将基于视频的 *PAR* 形式化为视觉-语言融合问题，并采用预训练大模型 *CLIP* 提取给定视频帧的特征嵌入。为了更好地利用语义信息，我们将属性列表作为另一项输入，并通过拆分、扩展和提示，将属性词语/短语转换为相应的句子。随后，利用 *CLIP* 的文本编码器进行语言嵌入。平均后的视觉词元与文本词元拼接后，输入融合 *Transformer* 进行多模态交互学习。增强后的词元将输入分类头，以进行行人属性预测。在大规模基于视频的行人属性识别数据集上开展的大量实验充分验证了所提框架的有效性。源代码和预训练模型将在 *https://github.com/Event-AHU/VTF_PAR*。

1. 引言

行人属性识别（PAR） [2, 17] 是计算机视觉领域一项非常重要的研究课题，并在深度学习的推动下取得了长足进展。近年来，基于卷积神经网络（卷积神经网络） [5], 和循环神经网络（循环神经网络） [3]的许多代表性行人属性识别模型相继提出。Wang 等人 [14]提出了 JRL，通过长短期记忆网络以联合循环学习的方式学习属性上下文和相关性 [6]。基于自注意力的 Transformer网络最初是为处理自然语言处理任务而提出的随后凭借出色的性能被引入计算机视觉领域 [4, 13, 15, 16, 19]。一些研究人员也将变换器用于行人属性识别问题，以建模全局上下文信息 [2, 12]。DR-Former [12] 旨在捕获区域之间的长程关系以及属性之间的关系。VTB [2]也被开发出来，用于融合图像与语言信息，从而实现更准确的属性识别。除了利用属性理解行人图像外，这项任务还服务于其他计算机视觉问题，例如目标检测 [18], 、行人重识别 [20], 等。尽管行人属性识别取得了巨大成功，但这些方法都仅基于单个 RGB帧开发，忽略了时间信息，因此在实际场景中可能无法获得最优结果。

如文献 [1], 所述，视频帧能够为特定属性提供更全面的视觉信息，而静态图像则无法做到这一点。作者提出利用视频片段理解人体属性，并提出了用于基于视频的行人属性识别的大规模数据集。他们还通过提出一种基于卷积神经网络和时间注意力的多任务视频行人属性识别框架，建立了一个基线模型。尽管在其基准数据集上取得了更好的性能，但我们认为以下问题仍限制了其整体结果。**首先**，他们采用卷积神经网络作为骨干网络来提取输入图像的特征表示，这有助于有效学习局部特征。众所周知，像素级空间中的全局关系对于细粒度属性识别也非常重要。一些研究者采用变换器网络来捕获此类全局信息 [4, 13], ，然而，他们的模型只能用于基于图像的属性识别。**其次**，作者将基于视频的行人属性识别建模为多任务分类问题，并尝试学习从给定视频到各项属性的映射。属性标签被转换为二值向量以进行网络优化。然而，这会严重丢失高层语义信息，而这对于行人属性识别至关重要。

为解决上述两个问题，本文提出一种新颖的 CLIP 引导视觉-文本融合

*† 表示贡献相同。通讯作者：Xiao Wang，邮箱：xiaowang@ahu.edu.cn

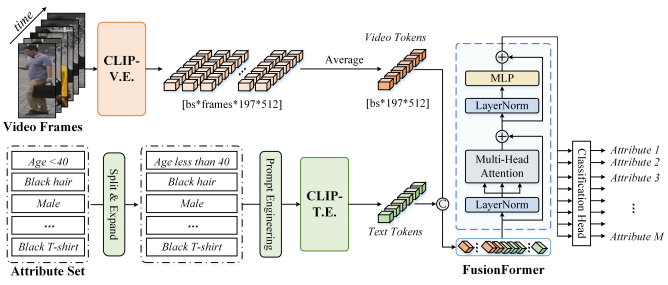

图 1. 我们提出的面向视频行人属性识别的 CLIP 引导的视觉-文本融合 Transformer 概览。

用于基于视频的行人属性识别的变换器。如图 1所示，我们将视频帧和属性集作为输入，并将基于视频的行人属性识别建模为一个多模态融合问题。具体而言，使用预训练的 CLIP 将视频帧转换为视频令牌 [11]，CLIP 是一个多模态大模型。通过拆分、扩展和提示工程，将属性集转换为相应的语言描述。然后，使用 CLIP 的文本编码器进行语言嵌入。之后，我们拼接视频令牌和文本词元，并将其输入融合 Transformer，以进行多模态信息交互，该模块主要包含层归一化、多头注意力和 MLP（多层感知器）。输出将输入分类头，用于行人属性识别。

总而言之，本文的主要贡献可归纳为以下两个方面：

• 我们提出了一种新颖的 CLIP 引导的视觉-文本融合 Transformer，用于视频行人属性识别。据我们所知，这是首项从视觉-文本融合角度研究视频行人属性识别的工作。

• 我们引入预训练大模型 CLIP 作为骨干网络，使模型能够应对上述挑战因素，具备较强的鲁棒性。大量实验验证了所提模型的有效性。

2. 方法

本节将从以下几个方面介绍我们提出的框架，即输入处理与嵌入、多模态融合 Transformer 以及优化。

2.1. 输入处理与嵌入

给定视频帧 $V = \{v_1, v_2, ..., v_T\}$ 和属性列表 $A = \{a_1, a_2, ..., a_M\}$，我们首先对这些输入进行处理，以便更好地利用预训练的 CLIP 模型。具体而言，我们将输入帧补零至分辨率 224×224，因为初始帧较窄，而预训练的 CLIP 要求输入采用固定的正方形分辨率作为输入。我们选择基于 ViT-B/16 的 CLIP 模型作为骨干网络，以兼顾效率和准确性。因此，输入帧被嵌入为一组视觉词元 $T \times 197 \times 512$，其中 197 是词元数量，512 是每个词元的维度。随后，我们沿时间维度对这些特征取平均，得到张量 $F_v \in \mathbb{R}^{197\times512} = \{f^1, f^2, ..., f^T\}$。

对于属性集合 A，我们首先需要将其处理为相应的语言描述，以充分利用 CLIP 文本编码器。具体而言，我们首先拆分并扩展每个属性，以获得自然短语。例如，“年龄 ≤ 40”会被处理为“年龄小于 *40* 岁”。然后，采用提示工程，使用精心设计的提示模板进一步将这些短语处理为自然语言描述。例如，年龄小于 *40* 岁 会被转换为“该行人的属性为年龄小于 *40* 岁”。处理完所有属性后，我们采用 CLIP 的文本编码器获得文本词元 $F_t = \{t^1, t^2, ..., t^M\}$。随后，将视频词元和文本词元拼接在一起 $[F_v, F_t]$，作为融合 Transformer 的输入。

2.2. 多模态融合变换器

变换器网络在许多研究领域中的有效性已得到广泛验证，包括计算机视觉、自然语言处理和多模态融合 [15]。在本研究中，我们采用变换器融合视觉与语言信息，以增强多模态特征表示学习。给定多模态输入 $[F_v, F_t]$，我们首先采用层归一化层对输入进行处理。随后，将输入转换为三个分支，即 Q、K 和 V。我们采用多头自注意力来学习长程全局关系，每个自注意力层的基本操作可描述为 $MLP(\text{SAttn}(Q、K)V)$，其中 SAttn 为 $\text{SAttn}(Q、K) = Softmax(\frac{QK^{Tr}}{\sqrt{c}})$，其中 c表示特征维度，Tr 为转置操作。输出词元将被输入分类头，用于属性预测。由于 MARS数据集包含 43 个属性，在我们的实际实现中，使用 43 个全连接层作为分类头。

2.3. 优化

在这项工作中，我们将基于视频的行人属性识别任务表述为视频文本融合问题。给定标注属性和原始视频，我们可以采用监督学习端到端地训练该框架。优化时采用二元交叉熵损失函数。

3. 实验

3.1. 数据集、指标与实现细节

在我们的实验中，采用了 Chen等人 [1] 提出的 MARS数据集用于训练和测试。训练

表 1. 在 MARS 基于视频的行人属性识别数据集上的结果。w/o 表示不包含以下模块。

方法	骨干网络	MARS		
		Prec	召回率	F1
3D 卷积神经网络 [7]	-	-	-	61.87
卷积神经网络-循环神经网络 [9]	-	-	-	70.42
VideoPAR（图像） [1]	ResNet50	-	-	67.28
VideoPAR（视频） [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**
改进	-	+2.80	+4.53	+3.62
我们的方法（不含 FusionFormer）	ViT-B/16	77.60	81.32	78.69

子集包含来自625人的8,298个轨迹片段，测试子集包含对应于626名行人的8,062个轨迹片段。每个轨迹片段平均包含60帧。为了评估我们的方法和对比行人属性识别（PAR）模型，我们采用广泛使用的精确率、召回率和F1分数作为评估指标。请注意，我们实验中报告的结果是对多个属性组的这些指标取平均后得到的。

在我们的实验中，使用预训练 CLIP 的 ViT-B/16 版本。在训练阶段，CLIP 编码器的参数保持固定。模型的学习率为 0.001，权重衰减为 1e-4。模型共训练 20 个周期。Adam [8]被用作优化器。模型使用 Python 和 PyTorch [10]框架实现，并在配备 RTX3090s 的服务器上训练。

3.2. 与其他 SOTA 模型的比较

在实验中，我们将模型与 MARS数据集上的多种强基线方法进行了比较，包括 3DCNN [7], 卷积神经网络-循环神经网络 [9], VideoPAR [1], 和 VTB [2]。如表 1 所示，我们的模型以较大优势胜过所有对比方法。具体而言，VTB [2] 在该数据集上的精确率、召回率和 F1 分数分别为 78.96、78.42、78.32，而我们的模型分别达到 81.76、82.95、81.94，在这些指标上的提升为+2.80、 +4.53、 +3.62 。我们的结果也优于 Chen等人提出的 VideoPAR（视频版本，F1 分数为 72.04），高出 +9.9。细粒度属性识别结果见表 2。这些实验充分验证了我们模型的有效性和优势。

3.3. 消融实验

组件分析。 在我们提出的框架中，融合 *Transformer* 和 预训练的 *CLIP* 骨干网络 是我们的关键组件。本节中，我们分析这两个组件，并在表 1中报告识别结果。VTB [2] 是我们的基线模型，它采用标准的 ViT-B/16 模型作为骨干网络，在精确率、召回率和 F1 分数上分别达到 78.96/78.42/78.32。使用 CLIP 模型后，结果可提升至 81.76/82.95/81.94,

表 2. MARS 基于视频的行人属性识别数据集上的结果。报告了所有评估属性的 F1 分数。

属性	VideoPAR（图像）	3D 卷积神经网络	卷积神经网络-循环神经网络	VideoPAR（视频）	Ours
上衣长度	58.72	56.37	65.18	**71.61**	**97.26**
下装长度	92.29	89.35	93.33	**93.90**	**93.69**
单肩包	72.57	61.30	**75.89**	**76.08**	65.61
背包	85.95	76.58	**87.17**	**87.62**	82.08
hat	57.57	57.69	**77.74**	**77.84**	72.76
手提包	62.82	59.90	**71.68**	**73.55**	59.08
hair	**86.91**	82.77	87.11	**88.17**	86.37
性别	90.89	85.75	92.44	**92.50**	**92.88**
下装类型	81 .69	72.86	84.16	**86.62**	**97.21**
pose	56.91	47.69	58.36	**61.36**	**74.84**
动作	39.39	33.64	**43.92**	43.69	**93.50**
上衣颜色	**72.72**	65.63	69.28	71.44	**74.97**
下装颜色	**44.63**	40.39	39.68	43.98	**69.76**
age	38.87	36.22	39.93	**40.21**	**87.07**
平均 F1	67.28	61.87	70.42	**72.04**	**81.94**



图 2. 我们的预测结果与真实标注（GT）的可视化。

表 3. MARS数据集上采用不同输入帧数的结果。

帧数	6	4	2	1
精确率	81.76	81.23	79.56	76.43
召回率	82.95	82.32	81.20	78.88
F1 分数	81.94	81.39	79.97	77.27

这验证了预训练大模型对基于视频的行人属性识别的有效性。将融合 Transformer 替换为常规全连接层后，结果从 81.76,82.95,81.94 降至 77.60,81.32,78.69，这表明该融合模块也有助于提升我们的最终性能。

输入视频帧分析。视频帧的数量对基于视频的行人属性识别具有重要作用。在本部分中，我们在 MARS数据集上使用不同数量的输入帧进行训练和测试，并在表3中报告实验结果。我们可以发现，随着视频帧数量的增加，性能逐步提升，即 F1 分数从 77.27 提升至 81.94。

可视化。除了上述定量分析外，本小节还进行了定性分析。如图2所示，我们可以发现，模型能够准确预测行人属性。

4. 结论

与主流的基于图像的行人属性识别不同，本文提出了一种新颖的、由 CLIP 引导的视觉-文本融合 Transformer，用于基于视频的行人属性识别，并充分利用时序信息。具体而言，我们将基于视频的行人属性识别建模为视觉-语言融合问题，并采用预训练的大型模型 CLIP 提取给定视频帧的特征嵌入。为更好地利用语义信息，我们将属性列表作为另一项输入，并使用融合 Transformer 将其与视频 token 融合。随后，将增强后的 token 输入分类头，以预测行人属性。我们在一个大规模的基于视频的行人属性识别数据集上进行了广泛实验，结果表明，模型取得了卓越的识别性能。在未来的工作中，我们将考虑设计细粒度的局部区域挖掘模块，以实现更高性能的属性识别。此外，值得探索新的提示学习/调优技术，以利用预训练多模态大型模型引导基于视频的行人属性识别。

致谢 本文得到中国国家自然科学基金项目（编号：62102205）的资助。

参考文献

[1] Zhiyuan Chen、Annan Li 和 Yunhong Wang。《一种用于基于视频的行人属性识别的时序注意力方法》。载于《模式识别与计算机视觉：第二届中国会议，*PRCV 2019*，中国西安，*2019*年*11*月*8*日至*11*日，论文集，第二部分 *2*》，第209–220页。Springer，2019年。[2] Xinhua Cheng、Mengxi Jia、Qian Wang 和 Jian Zhang。《一种用于行人属性识别的简易视觉-文本基线》。*IEEE* 视频技术电路与系统汇刊，32(10)：6994–7004，2022年。[3] Junyoung Chung、Caglar Gulcehre、KyungHyun Cho 和 Yoshua Bengio。《门控循环神经网络在序列建模中的实证评估》。*arXiv* 预印本 *arXiv:1412.3555*，2014年。[4] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。《一张图像胜过 16×16 个词：用于大规模图像识别的变换器》。载于国际学习表征会议。[5] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。《用于图像识别的深度残差学习》。载于*IEEE* 计算机视觉与模式识别会议论文集，第770–778页，2016年。[6] Sepp Hochreiter 和 J¨urgen Schmidhuber。《长短期记忆》。《神经计算》，9(8)：1735–1780，1997年。[7] Shuiwang Ji、Wei Xu、Ming Yang 和 Kai Yu。《用于人体动作识别的三维卷积神经网络》。*IEEE* 模式分析与机器智能汇刊，35(1)：221–231，2012年。[8] Diederik P Kingma 和 Jimmy Ba。《Adam：一种随机优化方法》。*arXiv* 预印本 *arXiv:1412 6980*，2014年。[9] Niall McLaughlin、Jesus Martinez Del Rincon 和 Paul Miller。《用于基于视频的行人重识别的循环卷积网络》。载于*IEEE* 计算机视觉与模式识别会议论文集，第1325–1334页，2016年。[10] Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga 等。《PyTorch：一种命令式风格、高性能的深度学习库》。神经信息处理进展，第32卷，2019年。[11]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。《通过自然语言监督学习可迁移视觉模型》。载于国际机器学习会议，第8748–8763页。PMLR，2021年。[12] Zengming Tang 和 Jun Huang。《Drformer：使用变换器学习双重关系以进行行人属性识别》。*Neurocomputing*，497：159–169，2022年。[13] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。《注意力就是一切》。神经信息处理进展，第30卷，2017年。[14] Jingya Wang、Xiatian Zhu、Shaogang Gong 和 Wei Li。《通过上下文与相关性的联合循环学习进行属性识别》。载于*IEEE* 国际计算机视觉会议论文集，第531–540页，2017年。[15] Xiao Wang、Guangyao Chen、Guangwu Qian、Pengcheng Gao、Xiao-Yong Wei、Yaowei Wang、Yonghong Tian 和 Wen Gao。《大规模多模态预训练模型：综合综述》。*arXiv* 预印本 *arXiv:2302 10035*，2023年。[16] Xiao Wang、Xiujun Shu、Zhipeng Zhang、Bo Jiang、Yaowei Wang、Yonghong Tian 和 Feng Wu。《借助自然语言实现更灵活、更准确的目标跟踪：算法与基准》。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第13763–13773页，2021年。[17] Xiao Wang、Shaofei Zheng、Rui Yang、Aihua Zheng、Zhe Chen、Jin Tang 和 Bin Luo。《行人属性识别：综述》。*PatternRecognition*，121：108220，2022年。[18] Jialiang Zhang、Lixiang Lin、Jianke Zhu、Yang Li、Yun-chen Chen、Yao Hu 和 Steven CH Hoi。《人群中的属性感知行人检测》。*IEEE* 多媒体汇刊，23：3085–3097，2020年。[19] Haojie Zhao、Xiao Wang、Dong Wang、Huchuan Lu 和 Xiang Ruan。《通过代理令牌引导的跨模态融合实现变换器视觉-语言跟踪》。模式识别快报，168：10–16，2023年。[20] Aihua Zheng、Peng Pan、Hongchao Li、Chenglong Li、Bin Luo、Chang Tan 和 Ruoran Jia。《通过渐进式属性嵌入实现准确的跨模态行人重识别》。载于第*30*届 *ACM* 国际多媒体会议论文集，第4309–4317页，2022年。