·图形图像处理·　　文章编号：1000-3428(2023)08-0215-08　　文献标志码：A　　中图分类号：TP391.41

结合语义与图像信息的行人属性识别算法

杨祖赫[1]，黎智辉[2]，唐云祁[1]，晏于文[2]，宋华青[2]

（1.中国人民公安大学 侦查学院，北京 100038；2.公安部物证鉴定中心，北京 100038）

摘　要：为提升行人属性的识别精度，充分利用行人属性间自然语义关联并解决不同属性相关图像信息的提取差问题，提出结合语义与图像信息的行人属性识别算法。通过自注意力机制的关系建模能力挖掘行人属性间的内在联系，利用交叉注意力机制建立属性间语义信息与图像特征信息的关系。在此基础上，依靠卷积融合图像的高阶与低阶特征并为模块增加局部特征信息，提升模型的泛化能力，通过设计属性预测模块，使模型可与任意骨干网络相拼接，进一步提升识别性能。实验结果显示，该算法的平均精度、准确率、F1值在PA-100K和PETA数据集上分别为84.04%、79.71%、88.03%和89.04%、82.39%、89.06%，与ALM、JLAC等算法相比，能够充分利用属性语义与图像特征信息，在多项评价指标上有明显提升。

关键词：行人属性识别；自注意力；卷积；特征融合；多标签分类

开放科学（资源服务）标志码（OSID）：

中文引用格式：杨祖赫，黎智辉，唐云祁，等. 结合语义与图像信息的行人属性识别算法[J]. 计算机工程，2023，49(8)：215-222，231.

英文引用格式：YANG Z H，LI Z H，TANG Y Q，et al. Pedestrian attribute recognition algorithm combining semantic and image information[J]. Computer Engineering，2023，49(8)：215-222，231.

Pedestrian Attribute Recognition Algorithm Combining Semantic and Image Information

YANG Zuhe[1]，LI Zhihui[2]，TANG Yunqi[1]，YAN Yuwen[2]，SONG Huaqing[2]

（1.School of Investigation，People's Public Security University of China，Beijing 100038，China；
2.Institute of Forensic Science of China，Beijing 100038，China）

【Abstract】To improve the recognition precision of pedestrian attributes and solve the problems of lack of use of natural semantic associations between pedestrian attributes and poor extraction of image information related to different attributes，this study proposes a pedestrian attribute recognition algorithm that combines semantic and image information. First，the relationship modeling ability of self-attention mechanism is utilized to explore the intrinsic relationship between pedestrian attributes，and cross-attention is utilized to establish the relationship between the semantic information between attributes and image feature information. Second，based on convolutional fusing high and low-order features，and adding local feature information into the module，the generalization ability of the model is improved. Owing to the design of the attribute prediction module，the model can be spliced with any backbone network and exhibits good performance. The experimental results show that the mean precision，accuracy，and F1 value of the proposed algorithm on the PA-100K and PETA datasets are 84.04%，79.71%，88.03%，and 89.04%，82.39%，89.06%，respectively. Compared with existing algorithms such as ALM and JLAC，this algorithm can exploit attribute semantics and image feature information and has a significant improvement in multiple evaluation indicators.

【Key words】pedestrian attribute recognition；self-attention；convolution；feature fusion；multi-label classification

DOI：10. 19678/j. issn. 1000-3428. 0064971

0　概述

行人属性识别旨在对输入的行人视频或图片进行多种属性预测，如性别、头发长度、年龄、穿着衣物等，具有将抽象图像信息转化为语义信息的能力。由于行人属性识别所需要的数据大多都来自与公安

基金项目：国家重点研发计划（2021YFF0602102）；公安部技术研究计划（2019JSYJA06）；公安部物证鉴定中心基本科研专项（2022JB024）。

作者简介：杨祖赫（1998—），男，硕士研究生，主研方向为图像识别、深度学习；黎智辉（通信作者），研究员、博士；唐云祁，副教授、博士；晏于文，助理研究员、硕士；宋华青，助理研究员、博士。

收稿日期：2022-06-13　　**修回日期：**2022-08-14　　**E-mail：**yangyangyang@stu.ppsuc.edu.cn

实战关联紧密的视频监控，因此其在公安工作中有着广阔的应用场景，如利用行人属性对监控视频进行结构化[1]、给出侦查分析意见[2]、作为庭审的相关证据[3]等。

随着深度学习的飞速发展，许多研究者也通过深度学习大幅提升了行人属性识别的计算速度和性能。最早的基于深度学习的行人属性识别算法通过提取全局特征进行行人属性预测，如DeepMAR[4]、ACN[5]等算法。这类方法虽然简单直接，但提取的特征过于粗糙，不利于行人属性的识别。针对提取的全局特征过于粗糙的问题，出现了提取局部特征的方法，如PGDM[6]、LGNet[7]、ALM[8]等算法。这类方法大多利用预训练的目标检测模型或注意力机制提取属性相关区域，虽然引入属性位置信息能够大幅提升属性识别效果，但都缺乏对于属性间内在语义联系的关注，并且需要额外训练目标检测器。针对该问题，研究者提出了利用属性间语义关系的方法，如JLAC[9]算法。

无论是早期提取全局特征还是利用属性间关系信息，都会将图像信息与属性间语义信息割裂，并缺少对于2种信息的联合利用。针对如何挖掘利用行人属性间自然语义关联，以及不同属性相关图像信息的提取问题，本文提出一种结合语义与图像信息的行人属性识别算法。该算法利用卷积和自注意力机制确定属性相关图像信息并挖掘属性间语义信息，同时结合属性语义与图像特征信息，实现对行人属性信息的充分利用，有效提升行人属性识别性能。

1 相关工作

近年来，深度学习的发展方兴未艾，许多基于深度学习的行人属性识别方法取得了显著的成果。本文以是否使用额外信息作为分界，将近期的方法分为2个类别[10]：

1）使用额外信息和属性标签作为网络训练的依据。这类方法通常使用人体关键点或是利用先验知识先定位属性相关区域来达到提升行人属性识别效果的目的。在相关研究中：LI等[6]提出将人体姿势的知识应用于行人属性识别，他们使用预训练好的人体关键点检测模型将属性相关区域提取出来，而后将其与整张图片输入神经网络进行行人属性识别；LIU等[7]提出使用预先提取的与属性相关的区域来辅助属性的识别。使用额外信息在解决行人属性问题上是一种可行的思路，但缺点在于过分依赖于额外信息的准确度和丰富程度，同时所需要的计算量也是庞大的。

2）不使用额外信息，只将属性标签作为网络训练的依据。在这类方法中，典型代表有以下3类方法：

（1）基于全局特征的方法。在相关研究中：LI等[4]将多属性识别与卷积神经网络结合，同时利用了图片的全局特征进行识别。

（2）基于注意力机制的方法。这类方法通过注意力机制自适应地定位图像上与属性相关的区域，从而提升识别效果。在相关研究中：LIU等[11]引入多向注意模块来提取像素级特征和语义级特征，用以定位细粒度属性；TANG等[8]将目标检测的特征金字塔结构和空间变换网络[12]结合注意力机制，用以提取多尺度特征图中的属性特征。

（3）通过挖掘属性间关系的方法。在相关研究中：WANG等[13]通过使用长短期注意力机制将行人属性识别问题转化为一个序列预测问题，挖掘出了不同属性间的关系；TAN等[9]使用图卷积神经网络模型[14]将属性间的关系转化为图结构，进而实现关系挖掘。

除了图卷积神经网络可以进行关系建模之外，在自然语言处理领域，Transformer[15]也可以实现关系建模，其独特的自注意力机制在许多自然语言处理问题上有着广泛的应用。最近在计算机视觉领域中，也出现了一些运用Transformer结构替换单纯以卷积神经网络为主体的方法。在目标识别领域：ZHENG等[16]提出了DETR方法，该方法将卷积神经网络强大的局部特征提取能力与Transformer独特的全局关系建模能力相结合，取得了性能提升。在图像分类领域：DOSOVITSKIY等[17]提出了ViT（Vision Transformer）方法，该方法完全采用Transformer结构构建网络，虽然在中等数据集上性能稍逊于传统卷积神经网络，但是在大数据集上取得了较好的表现；LIU等[18]提出一种将Transformer应用在计算机视觉领域的网络模型Swin Transformer，他们针对图像本身性质重新设计了Transformer结构，取得了优异的结果。鉴于Transformer针对关系建模的独特能力以及在计算机视觉领域的良好适应性，本文将以该结构为基础，针对行人属性识别存在的问题构建模型。

2 本文算法

本文所提出的网络模型主体结构如图1所示，其由2个部分组成：图中左侧为第1个部分，其依靠骨干网络的特征提取能力，将输入的行人图片转化为一组特征图；图中右侧为第2部分，其通过对不同属性间语义信息和属性相关图像信息的利用，实现行人属性的预测。得益于属性预测模块的设计，本文构建的模型可以选择任意骨干网络组合。为了验证其适应性，将属性预测模块分别与传统卷积神经网络ResNet[19]、轻量级卷积神经网络EfficientNet[20]、单纯将Transformer应用在机器视觉领域的ViT[17]以及针对视觉领域优化Transformer结构的Swin Transformer[18]相结合。

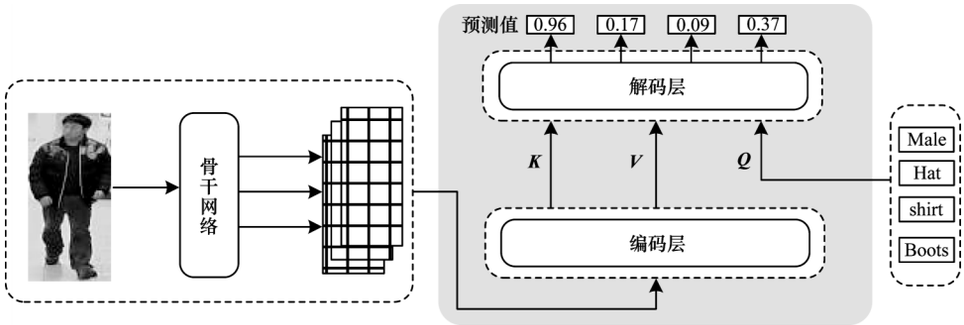

图1　网络模型结构

Fig.1　Network model structure

2.1　属性预测模块

在行人属性识别过程中，属性间的语义信息与图像特征所包含的信息十分重要，而若模型能够同时利用这2种信息，往往能够使识别的准确率得到提升。目前的研究大多只单独利用行人属性间的图像特征[8]或语义信息[9,13]来辅助识别，这将不能充分利用两者所包含的所有信息，而同时利用以上2种信息的方法大多只是通过堆叠不同模块来实现，并伴随着对不同模块输出的分别处理，割裂了这2种信息的天然联系，并极大地增加了模型的复杂度与参数量。本文对Query2Label[21]进行改进，设计一种结构简洁且即插即用的属性预测模块，其结构如图2所示，图中左侧结构为属性预测模块的编码层，右侧结构为解码层。该属性预测模块可以同时利用行人属性间的差异性和关联性，有效提升模型性能，并保证模型只需要处理单一模块输出，实现对于不同骨干网络的适配，同时降低模型的复杂度。

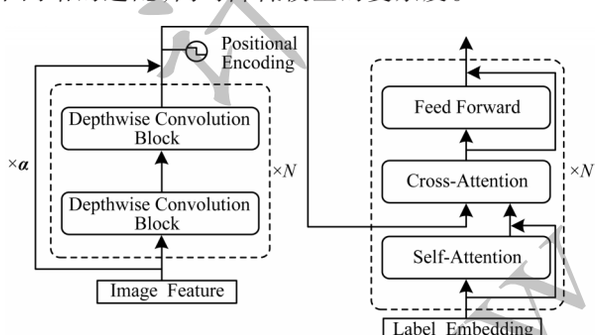

图2　属性预测模块结构

Fig.2　Attribute prediction module structure

2.1.1　编码层

在Query2Label原设计中，模块沿用了标准Transformer编码层的设计，依靠自注意力机制实现对于输入信息的关系建模，计算过程如下：

$$[\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}]=[\boldsymbol{W}_Q,\boldsymbol{W}_K,\boldsymbol{W}_V]\cdot\mathcal{F}_0 \tag{1}$$

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(\text{head}_1,\text{head}_2,\cdots,\text{head}_n)\boldsymbol{W}^{\circ}$$

$$\text{head}=\text{Attention}(\boldsymbol{Q}\boldsymbol{W}_{Q,i},\boldsymbol{K}\boldsymbol{W}_{K,i},\boldsymbol{V}\boldsymbol{W}_{V,i}) \tag{3}$$

对于输入的特征$\mathcal{F}_0\in\mathbb{R}^{H\times W\times d}$，其中，$H$和$W$代表特征的水平像素量和垂直像素量，$d$为维度，每一个元素都要与3个可学习的权重矩阵($\boldsymbol{W}_Q,\boldsymbol{W}_K,\boldsymbol{W}_V$)相乘，如式(1)所示，得到query($\boldsymbol{Q}$)、key($\boldsymbol{K}$)和value($\boldsymbol{V}$)，然后通过矩阵$\boldsymbol{Q}$与矩阵$\boldsymbol{K}^{\mathrm{T}}$相乘得到元素与其他每个元素之间的注意力权重，接着除以向量维度d_k防止内积过大，随后通过Softmax函数将得到的注意力权重平滑到0~1之间，便于反向传播的计算，如式(2)所示。为了考虑不同层面的信息，将每个层面的自注意力头拼接后乘以权重$\boldsymbol{W}^{\circ}$，就得到了多头自注意力特征表示，如式(3)所示。

由于自注意力机制计算复杂度高、计算量大，本文在编码层使用2个全新设计的逐通道卷积块代替原始结构中的基于多头自注意力机制的编码层，其结构如图3所示，其中，Dwconv为逐通道卷积，Conv为普通卷积，紧跟内容为卷积核大小，d为此层输出的通道，$\boldsymbol{\alpha}$为可学习的对角矩阵。使用逐通道卷积块替代原结构，首先可以使编码层计算时不考虑位置编码，其次结构中大量使用逐通道卷积以及1×1卷积，这样的设计既可以利用卷积计算能够天然获取局部信息的优势，同时又大幅减少了编码层的计算量。

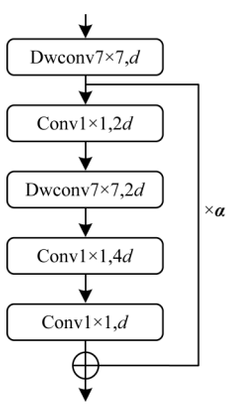

图3　逐通道卷积块结构

Fig.3　Depthwise convolution block structure

网络在不断加深的同时，也将低阶特征不断转化为高阶特征。高阶特征固然抽象程度更高，但过分依赖高阶特征会造成网络识别效果的退化。因此，本文在逐通道卷积块内以及编码层将高、低阶特

征进行融合，计算过程如下：

$$x'_{i+1}=x_{i+1}+f_{\text{conv}}(x_i)\times\text{diag}(\alpha_1,\alpha_2,\cdots,\alpha_d) \tag{4}$$

其中：x_i代表第i层的特征；f_{conv}代表对特征进行卷积计算；$\alpha_1,\alpha_2,\cdots,\alpha_d$是一组可学习的参数，下标$d$为相乘特征的通道数。本文使用1×1卷积将低阶特征与高阶特征通道对齐后，将低阶特征与可学习的对角矩阵$\boldsymbol{\alpha}$相乘[22]，对低阶特征的每个通道进行自适应调整，后与高阶特征相融合，使得编码层的高、低阶特征都能够在后续网络中产生影响，进而使得网络拥有更好的泛化性，同时融合的特征也可以减少网络对于高阶特征的依赖。

2.1.2 解码层

解码层由自注意力机制、交叉注意力机制和前馈神经网络组成，其输入由可学习的属性embedding和编码输出的图像特征组成。

首先，行人属性通过随机映射后变为一组可学习的矩阵，将此矩阵输入多头注意力机制，通过自注意力计算得到每一个属性之间的语义信息并输出信息矩阵$\boldsymbol{Q}_1$，如式(5)所示；然后，编码部分输出$\boldsymbol{S}$与上一个多头注意力机制的输出$\boldsymbol{Q}_1$共同组成下一个交叉注意力机制的输入，如式(6)所示，其中，query来自经过多头自注意力机制的属性语义信息，而key和value来自编码层的输出特征图；最后，通过一个前馈神经网络实现编码层的输出。

$$\boldsymbol{Q}_1\leftarrow\text{MultiHead}(\boldsymbol{Q},\boldsymbol{Q}_K,\boldsymbol{Q}_V) \tag{5}$$

$$\boldsymbol{Q}_2\leftarrow\text{MultiHead}(\boldsymbol{Q}_1,\boldsymbol{S}_K,\boldsymbol{S}_V) \tag{6}$$

在自然语言处理中，自注意力机制可以挖掘词与词之间的关系。参考自然语言对词组的处理，本文将属性转化成一个可学习的矩阵，使其能够进行自注意力计算，且该矩阵会随着模型训练而不断更新，这样挖掘到的属性间的语义信息可以规避虚假联系而不需要先验知识。交叉注意力机制则是将属性间的语义信息与图像特征信息进行自注意力计算：通过属性间的语义信息可以利用图像信息对自身进行修正，通过图像特征信息可以利用语义信息对特征图进行筛选，最终通过对2种信息自适应地建立相关联系，提升模型识别性能。属性预测模块针对关系建模的结构特点，能够挖掘行人属性间语义信息和行人属性相关联区域关系这2种信息并加以利用。

2.2 损失函数

现有的行人属性数据集大多都存在数据分布不平衡的问题，如PA-100K数据集中帽子(Hat)属性正样本占总样本不到10%。这种数据不平衡分布会使得模型在学习时过分关注数据量大的属性而学习不到数据量小的属性的特征，使属性识别效果变差。之前的研究普遍采用加权交叉熵损失函数[4]来解决长尾分布对于预测的影响。加权交叉熵损失虽然能够在一定程度上抑制分布不平衡的影响，但是需要预先知道数据分布。针对这个问题，本文引入一种非对称损失(Asymmetric Loss，ASL)函数[23]。ASL可以在不使用先验知识的前提下，对损失进行非对称调整来达到优化损失的目的。ASL损失函数如下所示：

$$p_m=\max(p_i-m) \tag{7}$$

$$L=\begin{cases}\dfrac{1}{M}\displaystyle\sum_{i=1}^{M}(1-p_i)^{\gamma^+}\lg p_i, y_i=1\\ \dfrac{1}{M}\displaystyle\sum_{i=1}^{M}(p_m)^{\gamma^-}\lg(1-p_m), y_i=0\end{cases} \tag{8}$$

在式(7)中：p_i是模型对第i个样本为正样本的概率预测；m是一个超参数。在式(8)中：M是总样本数；p_m的定义同式(7)；y_i是真实标签；γ^+和γ^-是2个超参数。可以通过调节m、γ^+、γ^-这3个超参数来解决数据不平衡的问题，其中，γ^+和γ^-可以进行软阈值优化，因为其底数是恒小于1大于0的实数，因此，如果更强调正样本对于损失的贡献，就可以设置$\gamma^-<\gamma^+$使负样本被抑制。相反地，也可以实现对于正样本的抑制，从而提升负样本对于损失的贡献。但如果数据不平衡相当严重，这时候就需要使用硬阈值m对损失进行优化。m可以将负样本中概率小于m的样本损失置为0，这样大量的负样本就无法对损失作出贡献，从而使模型优化时更多地考虑数量少或难以分辨的正样本。在本文的实验中，将超参数设置为m=0.2，γ^-=2，γ^+=0。

3 实验对比与分析

本节首先介绍数据集与评价方法和实验设置，然后与目前算法进行对比，并针对属性预测模块进行一系列实验，最后对模型进行消融实验。

3.1 数据集与评价方法

3.1.1 数据集

现有表现优异的算法普遍选择在PETA[24]和PA-100K[7]这2个公开行人属性数据集上进行训练测试。PETA数据集由10个小型行人重识别数据集重新标注构建而成，该数据集包含19 000张图片，标注了61个二分类属性和4个多分类属性。在本文中，PETA数据集被随机地分割成3个部分：9 500张图片用于训练，1 900张图片用于验证，7 600张图片用于测试。由于部分二分类属性中有些属性的正样本数量低于本属性总样本数量的5%，因此只选择其中35个正样本数量高于5%的属性作为测试属性。PA-100K数据集是目前最大的行人属性数据集，其中包含100 000张来自户外监控摄像头的图片，每张图片都标注了26个二分类属性，并按照8∶1∶1的比例划分训练集、验证集和测试集。

3.1.2 评估方法

本文采用4个实例级和1个属性级共5个评价指标对所提方法的性能表现进行衡量，其中4个实例级的评价指标分别为准确率、精确率、召回率和F1值，属性级的评价标准采用平均精度。5个评价指标的计算公式如下：

$$A_{\text{accuracy}}=\frac{1}{N_{\text{total}}}\sum_{i=1}^{N_{\text{total}}}\frac{T_{\text{P}i}}{T_{\text{P}i}+F_{\text{P}i}+F_{\text{N}i}} \tag{9}$$

$$P_{\text{precision}}=\frac{1}{N_{\text{total}}}\sum_{i=1}^{N_{\text{total}}}\frac{T_{\text{P}i}}{T_{\text{P}i}+F_{\text{P}i}} \tag{10}$$

$$R_{\text{recall}}=\frac{1}{N_{\text{total}}}\sum_{i=1}^{N_{\text{total}}}\frac{T_{\text{P}i}}{T_{\text{P}i}+F_{\text{N}i}} \tag{11}$$

$$F_1=\frac{2P_{\text{precision}}R_{\text{recall}}}{P_{\text{precision}}+R_{\text{recall}}} \tag{12}$$

$$m_{\text{mA}}=\frac{1}{2L}\sum_{i=1}^{L}\left(\frac{T_{\text{P}i}}{T_{\text{P}i}+F_{\text{N}i}}+\frac{T_{\text{N}i}}{T_{\text{N}i}+F_{\text{P}i}}\right) \tag{13}$$

其中：A_{accuracy}代表准确率；$P_{\text{precision}}$代表精确率；R_{recall}代表召回率；F_1代表F1值；m_{mA}代表平均精度；$T_{\text{P}i}$、$F_{\text{P}i}$、$F_{\text{N}i}$、$T_{\text{N}i}$分别代表对于第i个样本来说被正确预测的正样本、被错误预测的负样本、被错误预测的负样本、被正确预测的负样本的数量；N_{total}代表总样本数；L代表属性的总数。

3.2　实验设置

本文在操作系统Windows Server 2019、显卡NVIDIA TESLA V100 32 GB等实验条件下，使用CUDA加速，在PyTorch深度学习架构上进行实验。本文对实验数据进行预处理：输入图像大小为224×224像素，并使用RandAugment[25]进行数据增强。训练采用AdamW优化策略优化网络，设置权重衰减为0.000 5，初始学习率为0.001，batchsize设置为32，共训练50个epoch。同时，使用在ImageNet22K上预训练的参数作为骨干网络的初始参数。在实验中，将属性预测模块设置为1层编码层以及1层解码层的组合。

3.3　属性预测模块结构实验

对属性预测模块中结构改动以及编解码层组成进行测试实验，其中优化轨迹如图4所示，该图展示了在PETA数据集上从骨干网络开始对属性预测模块进行各项改动时平均精度的变化轨迹。

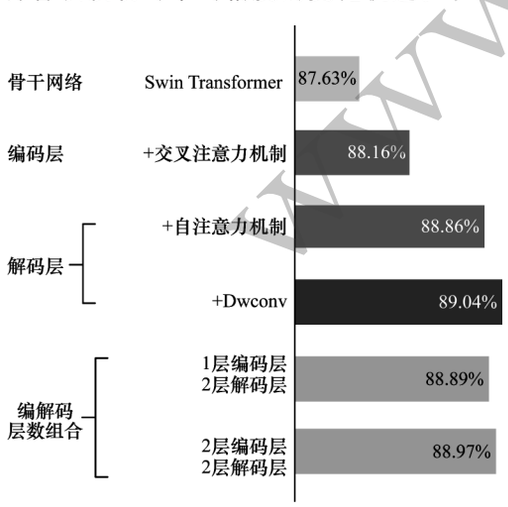

图4　属性预测模块优化轨迹

Fig.4　The attribute prediction module optimized trajectory

3.3.1　不同编解码层数组合对比

为了探究属性预测模块内的结构组成对于识别效果的影响，在保证其余条件一致的前提下，在PETA、PA-100K数据集上进行对比实验。采用平均精度作为识别的评价标准，并分别对比[编码层数，解码层数]设置为[1,1]、[1,2]、[1,3]、[2,1]、[2,2]、[2,3]这6种情况的实验结果，如图5所示。

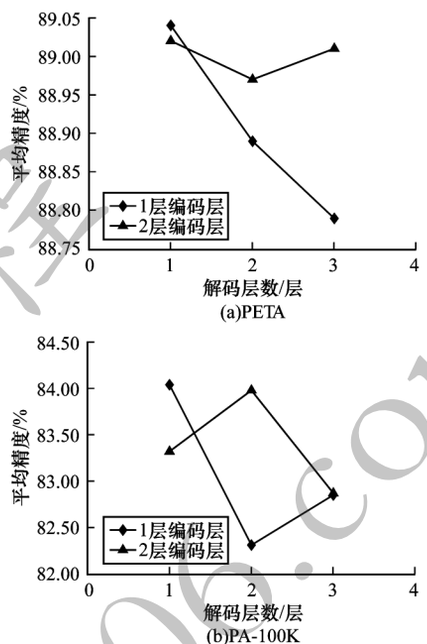

图5　属性预测模块内不同结构的实验结果对比

Fig.5　Experimental results comparison of different structures in attribute prediction module

在PETA数据集上，组合为[1,1]的效果最好，对于编码层数为1的实验组，添加更多的解码层不会使得平均精度提升，反而会一定程度下降，对于编码层数为2的实验组，平均精度随着解码层数的增加上下波动。在PA-100K数据集上，依然是组合为[1,1]的效果最好，对于所有组合，添加解码层都会使平均精度产生一定的波动。由图5可以看出，对于2个数据集来说，[编码层数，解码层数]组合为[1,1]都取得了最好的结果，虽然其他组合也可以取得相近的结果，但是添加更多层就意味着模型需要更多的计算量以及更大的参数量。因此，后续实验将采用1层编码层和1层解码层的组合作为属性预测模块的结构。

3.3.2　编码结构对比

本文使用2个逐通道卷积块代替原本基于自注意力机制的编码层，为了对比2种结构，在保证输入特征图大小及维度相同的前提下，对2种结构分别进行实验，实验结果如表1所示。其中：SA代表自注意力机制；Conv代表逐通道卷积块；参数量及计算量均为编码层数据；最后一行All代表使用逐通道卷积后的整体算法，参数量以及计算量都为整体算法。由表1可以看出：使用逐通道卷积不仅减少了参数

量，而且还减少了约75%的计算量；在2个数据集上，逐通道卷积块的表现也优于原始结构；对于整体算法而言，逐通道卷积结构不仅带来了计算量以及参数量的减少，使得算法整体的计算效率得到提高，同时也提升了算法识别的准确性。

表1 2种编码层结构的实验结果对比

Table 1 Experimental result comparison of two encoder layers structures

编码层结构	参数量/10^6	计算量/10^9	PETA		PA-100K	
			平均精度/%	F1值/%	平均精度/%	F1值/%
SA	33.58	2.98	88.86	89.01	83.76	87.69
Conv	29.89	0.73	89.04	89.06	84.04	88.03
All	135.72	16.81	89.04	89.06	84.04	88.03

为测试属性预测模块中编码层改进对于模型泛化能力的影响，参考迁移学习的训练模式，将模型在数据集A上训练后，冻结其除最后全连接层以外的所有参数，在新数据集B上进行训练。按照此实验方案，分别在PETA和PA-100K数据集上进行迁移学习。除数据集划分以及实验设置保持一致之外，考虑到2个数据集较为相近，故在新数据集上对所有训练样本只进行一次学习，实验结果如表2所示。其中：Before代表改进结构前的结果；After代表改进结构后的结果。由表2可以看出：改进结构后，在2个数据集相互迁移的情况下，各项评价指标都有较为明显的提升，说明依靠卷积融合高、低阶特征并增加局部信息的利用，能够使属性预测模块更好地提取泛化性更强的特征，并增强了模型在新数据集上的识别能力。

表2 2种编码层结构的泛化能力对比

Table 2 Generalization ability comparison of two encoder layers structures %

实验方案	平均精度	准确率	精确率	召回率	F1值
PETA→PA-100K(Before)	59.48	54.38	68.66	70.86	69.74
PA-100K→PETA(Before)	55.96	48.24	61.41	66.63	63.91
PETA→PA-100K(After)	65.57	62.31	74.44	77.53	75.98
PA-100K→PETA(After)	58.36	49.44	60.57	70.16	65.01

3.3.3 属性可视化分析

属性预测模块能够提升相关联属性的识别性能，主要依靠解码层中自注意力机制来挖掘不同属性间的语义信息。属性矩阵的生成采用高维映射，其在生成时均匀地分布于高维空间。经过训练的属性矩阵由于处于高维映射中，其最终结果并不可以直接观察，因此本文采用TSNE[26]降维方法，将原本高维的矩阵降至2维，进而实现对输入属性经过训练之后的结果可视化，如图6和图7所示（彩色效果见《计算机工程》官网HTML版）。图6是PA-100K数据集中26个属性标签的可视化结果，图7是PETA数据集中35个属性标签的可视化结果。由图6和图7可以看出：属性被分成了几个部分，且部分与部分之间的分隔较为明显。通过观察发现，一些正样本较少的属性往往处于边缘的位置，如PETA数据集中的Hat属性，这与其出现过于稀少且难以与其他有着大量正样本的数据产生强关联的客观事实相符。虽然在语义上不能解释为何有些属性标签关联度高，但是考虑到降维带来的信息损失和数据集数据分布不均等问题，通过属性预测模块确实在一定程度上得到了属性间的关联信息。

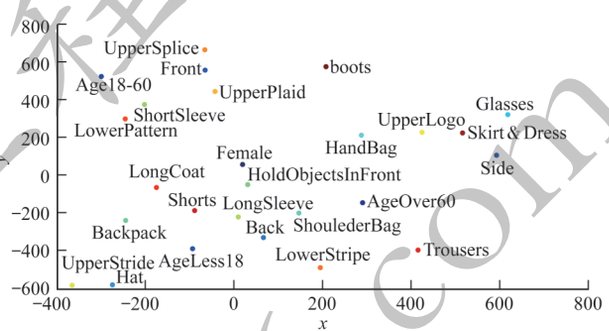

图6 PA-100K数据集上属性标签可视化结果

Fig.6 Attribute label visualization result on PA-100K dataset

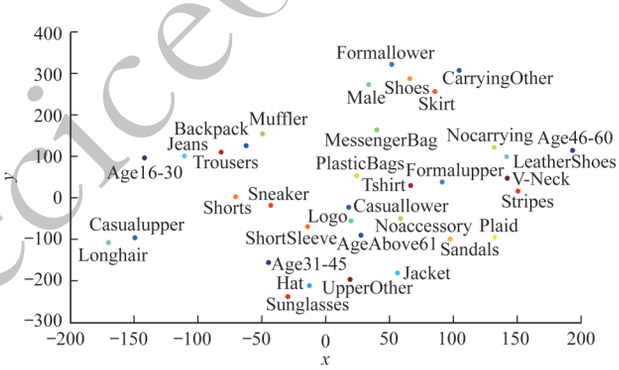

图7 PETA数据集上属性标签可视化结果

Fig.7 Attribute label visualization result on PETA dataset

3.4 属性预测模块适配性实验

为进一步验证属性预测模块对于不同骨干网络的适配性，将其与ResNet、EfficientNet、ViT、Swin Transformer这4种设计思路不同的骨干网络拼接进行实验。为保证实验公平性，4种骨干网络都使用ImageNet22K上预训练的参数。其中，ResNet使用ResNet50配置，EfficientNet使用EfficientNet_b3配置，ViT使用ViT-B/16配置，Swin Transformer使用Swin-B配置。2个数据集上的实验结果如表3和表4所示，其中，上部分baseline代表与属性预测模块组合前的实验结果，下部分是组合后的实验结果。表3

与表4实验结果显示，在2个数据集上，4种设计思路不同的骨干网络在与属性预测模块拼接后，平均精度、精确率和F1值都有不同程度的提升，同时，也可以观察到在PETA数据集上，ViT和Swin Transformer拼接属性预测模块后准确率都有小幅下降，表明属性预测模块牺牲了对于部分易识别样本的识别能力，转而更加关注难样本的识别，进而实现了平均精度的提升。总体上，属性预测模块为骨干网络提供的联合语义与图像信息能够帮助模型更好地预测行人属性。

表3　PA-100K数据集上骨干网络实验结果

Table 3　Experimental results of backbone network on PA-100K dataset　%

骨干网络	平均精度	准确率	精确率	召回率	F1值
ResNet(baseline)	79.84	73.04	76.89	86.26	81.31
EfficientNet(baseline)	78.45	74.74	81.06	86.50	83.69
ViT(baseline)	80.32	71.10	77.70	85.46	81.39
Swin Transformer(baseline)	82.58	78.54	85.27	89.23	87.21
ResNet	82.31	76.78	81.35	91.80	86.26
EfficientNet	81.52	76.29	84.42	86.40	85.40
ViT	83.21	79.33	87.10	88.21	87.65
Swin Transformer	84.04	79.71	84.74	91.59	88.03

表4　PETA数据集上骨干网络实验结果

Table 4　Experimental results of backbone network on PETA dataset　%

骨干网络	平均精度	准确率	精确率	召回率	F1值
ResNet(baseline)	83.76	75.33	81.28	88.39	84.69
EfficientNet(baseline)	81.29	71.07	77.32	86.91	81.83
ViT(baseline)	85.84	80.72	85.14	89.57	87.30
Swin Transformer(baseline)	87.63	82.72	86.36	88.91	87.62
ResNet	85.72	78.48	85.51	86.71	86.11
EfficientNet	83.74	76.23	83.51	85.66	84.57
ViT	87.82	79.78	88.11	87.31	87.71
Swin Transformer	89.04	82.39	87.18	91.01	89.06

3.5　算法对比

为验证本文算法的有效性以及各种相关优化对最终结果的影响，采用上文实验设置，在PETA数据集和PA-100K数据集上，将本文算法与DeepMar[4]、HP-Net[11]、LG-Net[7]、ALM[8]、JLAC[9]算法进行比较，如表5和表6所示，其中最优结果以粗体标出。由表5和表6可以看出：在2个数据集上，本文算法都在平均精度、准确率、召回率和F1值这4个评价指标上有着优异表现，说明将语义信息与图像信息联合能够实现对于行人属性的有效识别。

表5　PA-100K数据集上不同算法性能对比

Table 5　Performance comparison of different algorithms on PA-100K dataset　%

算法	平均精度	准确率	精确率	召回率	F1值
DeepMar	72.70	70.39	82.24	80.42	81.32
HP-Net	74.21	72.19	82.97	82.09	82.53
LG-Net	76.96	75.55	86.99	83.17	85.04
ALM	80.68	77.08	84.21	88.84	86.46
JLAC	82.31	79.47	**87.45**	87.77	87.61
本文算法	**84.04**	**79.71**	84.74	**91.59**	**88.03**

3.6　消融实验

本文算法使用了属性预测模块和ASL损失函数来提升模型识别性能，为了验证其有效性，在PETA和PA-100K数据集上进行消融实验，分别探讨添加内容是否对于结果产生了积极的作用，实验结果如表7所示。可以看出：ASL损失函数在2个数据集上都体现了其有效性，添加ASL损失函数后模型能够更好地针对难样本进行学习；属性预测模块也在2个数据集上展现了其结构的有效性。以上结果验证了添加组件的出色性能以及整体结构的有效性。

表6　PETA数据集上不同算法性能对比

Table 6　Performance comparison of different algorithms on PETA dataset　%

算法	平均精度	准确率	精确率	召回率	F1值
DeepMar	82.89	75.07	83.68	83.14	83.41
HP-Net	81.77	76.13	84.92	83.24	84.07
LG-Net	82.97	78.08	86.79	86.12	85.76
ALM	86.30	79.52	85.65	88.09	86.85
JLAC	86.96	80.38	**87.81**	87.09	87.45
本文算法	**89.04**	**82.39**	87.18	**91.01**	**89.06**

表7 消融实验结果

Table 7 Ablation experiment results %

数据集	骨干网络	属性预测模块	ASL损失函数	平均精度
PETA	√	×	×	87.63
	√	×	√	88.06
	√	√	×	88.22
	√	√	√	89.04
PA-100K	√	×	×	82.58
	√	×	√	83.20
	√	√	×	82.90
	√	√	√	84.04

4 结束语

本文提出一种结合语义与图像信息的行人属性识别算法，该算法通过骨干网络得到可靠的特征图，而后通过属性预测模块来强化相关联属性的关系以及自适应建立图像与语义信息关系。在PETA和PA-100K数据集上的实验结果表明，本文算法能够充分利用属性语义与图像特征信息，提升识别性能。后续将参考目标检测领域针对解码部分的改进[27]，结合其设计思路对属性预测模块做进一步优化。

参考文献

[1] 王治，韩祥. 视频结构化解析技术在公安警务实战中的建设与应用[J]. 警察技术，2018(5)：63-66.
WANG Z，HAN X. The construction and application of video structured analysis technology in public security police actual combat[J]. Police Technology，2018(5)：63-66. (in Chinese)

[2] 许磊，李志刚，黎智辉，等. 人像检验鉴定探讨[J]. 刑事技术，2020，45(2)：111-116.
XU L，LI Z G，LI Z H，et al. Cogitation into human image identification[J]. Forensic Science and Technology，2020，45(2)：111-116. (in Chinese)

[3] 黎智辉，谢兰迟，吕游，等. 视频侦查中多摄像头下嫌疑目标同一的概率研究[J]. 刑事技术，2022，47(1)：24-34.
LI Z H，XIE L C，LÜ Y，et al. Probabilistic approach to identifying same suspected target from multiple cameras in video investigation[J]. Forensic Science and Technology，2022，47(1)：24-34. (in Chinese)

[4] LI D W，CHEN X T，HUANG K Q. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios [C]//Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition. Washington D. C.，USA：IEEE Press，2016：111-115.

[5] SUDOWE P，SPITZER H，LEIBE B. Person attribute recognition with a jointly-trained holistic CNN model[C]// Proceedings of IEEE International Conference on Computer Vision. Washington D. C.，USA：IEEE Press，2016：329-337.

[6] LI D W，CHEN X T，ZHANG Z，et al. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios [C]//Proceedings of IEEE International Conference on Multimedia and Expo. Washington D. C.，USA：IEEE Press，2018：1-6.

[7] LIU P Z，LIU X H，YAN J J，et al. Localization guided learning for pedestrian attribute recognition [EB/OL]. [2021-11-22]. https://arxiv. org/abs/1808. 09102.

[8] TANG C F，SHENG L，ZHANG Z X，et al. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C.，USA：IEEE Press，2020：4996-5005.

[9] TAN Z C，YANG Y，WAN J，et al. Relation-aware pedestrian attribute recognition with graph convolutional networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence，2020，34(7)：12055-12062.

[10] JIA J，HUANG H，CHEN X，et al. Rethinking of pedestrian attribute recognition：a reliable evaluation under zero-shot pedestrian identity setting[EB/OL]. [2021-11-22]. https://arxiv. org/abs/2107. 03576.

[11] LIU X H，ZHAO H Y，TIAN M Q，et al. HydraPlus-Net：attentive deep features for pedestrian analysis [C]// Proceedings of IEEE International Conference on Computer Vision. Washington D. C.，USA：IEEE Press，2017：350-359.

[12] JADERBERG M，SIMONYAN K，ZISSERMAN A，et al. Spatial Transformer networks[EB/OL]. [2021-11-22]. https://arxiv. org/abs/1506. 02025.

[13] WANG J Y，ZHU X T，GONG S G，et al. Attribute recognition by joint recurrent learning of context and correlation [C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C.，USA：IEEE Press，2017：531-540.

[14] KIPF T N，WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. [2021-11-22]. https://arxiv. org/abs/1609. 02907.

[15] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York，USA：ACM Press，2017：6000-6010.

[16] ZHENG M H，GAO P，ZHANG R R，et al. End-to-end object detection with adaptive clustering Transformer [EB/OL]. [2021-11-22]. https://arxiv. org/abs/2011. 09315.

[17] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An image is worth 16×16 words：Transformers for image recognition at scale[EB/OL]. [2021-11-22]. https://arxiv. org/abs/2010. 11929.

[18] LIU Z，LIN Y T，CAO Y，et al. Swin Transformer：hierarchical Vision Transformer using shifted windows[C]// Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C.，USA：IEEE Press，2022：9992-10002.

[19] HE K M，ZHANG X Y，REN S Q，et al. Deep residual learning for image recognition[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2016：770-778.

（下转第231页）

（上接第222页）

[20] TAN M X，LE Q V. EfficientNet：rethinking model scaling for convolutional neural networks[C]//Proceedings of International Conference on Machine Learning. [S. l.]：PMLR，2019：6105-6114.

[21] LIU S L，ZHANG L，YANG X，et al. Query2Label：a simple Transformer way to multi-label classification[EB/OL]. [2021-11-22]. https：//arxiv. org/abs/2107. 10834v1.

[22] TOUVRON H，CORD M，SABLAYROLLES A，et al. Going deeper with image Transformers[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C. ，USA：IEEE Press，2022：32-42.

[23] BEN-BARUCH E，RIDNIK T，ZAMIR N，et al. Asymmetric loss for multi-label classification[EB/OL]. [2021-11-22]. https：//arxiv. org/abs/2009. 14119.

[24] DENG Y B，LUO P，LOY C C，et al. Pedestrian attribute recognition at far distance[C]//Proceedings of the 22nd ACM International Conference on Multimedia. New York，USA：ACM Press，2014：789-792.

[25] CUBUK E D，ZOPH B，SHLENS J，et al. RandAugment：practical automated data augmentation with a reduced search space[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C. ，USA：IEEE Press，2020：3008-3017.

[26] VAN DER MAATEN L，HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research，2008，9(11)：72-84.

[27] DAI X Y，CHEN Y P，YANG J W，et al. Dynamic DETR：end-to-end object detection with dynamic attention[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C. ，USA：IEEE Press，2022：2968-2977.

编辑　金胡考