DOI: 10.13382/j.jemi.B2407384

基于 ViT 的细粒度特征增强无监督行人重识别方法*

程思雨 陈 莹
(江南大学轻工过程先进控制教育部重点实验室 无锡 214122)

摘 要:行人重识别任务可以看做是细粒度视觉分类任务的一种。现有的无监督行人重识别方法通常只关注人体全局特征,不能获取准确的细粒度局部特征,进而影响模型的识别精度。为解决这一问题,提出了一种基于 ViT 的细粒度特征增强网络,该网络利用视觉-语言模型生成图像中人体局部区域的掩码,根据自注意力机制中可学习标记与图像块之间交互策略的不同,使类标记与引入的可学习变量局部标记分别学习全局与局部细粒度特征表示。此外,为进一步提升特征表示能力,设计了一个空间信息增强模块,该模块通过挖掘人体局部区域内代表性图像块之间的空间上下文关系来增强特征学习。最后,利用提取到的全局和局部细粒度特征,分别计算在线和离线相机感知对比损失,以增强模型在无监督环境下对于行人身份的鲁棒性。在 Market-1501、MSMT17 和 PersonX 数据集上的实验结果验证了所提方法的有效性,mAP/Rank-1 分别达到了 90.3%/95.9%、59.2%/83.5%、91.3%/96.1%。

关键词: 行人重识别;无监督;细粒度特征;vision transformer;自注意力
中图分类号: TP391.4; TN911.7 **文献标识码**: A **国家标准学科分类代码**: 510.40

Fine-grained feature enhancement unsupervised person re-identification method based on ViT

Cheng Siyu Chen Ying
(Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China)

Abstract: Person re-identification can be regarded as a form of fine-grained visual classification task. Existing unsupervised person Re-ID methods typically focus solely on global features of human bodies, failing to capture accurate fine-grained local features, thereby hindering the recognition accuracy of the models. To address this issue, we propose a ViT-based fine-grained feature enhancement network. This network leverages a vision-language model to generate masks for local regions of human bodies in images. Subsequently, based on the distinct interaction strategies between learnable tokens and image patches within the self-attention mechanism, the class token and introduced learnable local tokens are utilized to learn global and local fine-grained feature representations, respectively. Furthermore, to further enhance feature representation capabilities, a spatial information enhancement module is designed. This module augments feature learning by mining spatial contextual relationships among representative image patches within local regions of human bodies. Finally, utilizing the extracted global and local fine-grained features, online and offline camera-aware contrastive losses are computed separately to bolster the model's robustness to person identities in an unsupervised environment. Experimental results on the Market-1501, MSMT17, and PersonX datasets validate the effectiveness of the proposed method, achieving mAP/Rank-1 accuracies of 90.3%/95.9%, 59.2%/83.5%, and 91.3%/96.1%, respectively.

Keywords: person re-identification; unsupervised; fine-grained feature; vision transformer; self-attention

收稿日期:2024-03-29 Received Date: 2024-03-29
*基金项目:国家自然科学基金(62173160)项目资助

0　引　言

行人重识别(re-identification, Re-ID)的核心在于,给定一张行人图像,在跨摄像头的图像库中检索出该行人的图像。它能够弥补固定摄像头的视觉局限,与行人检测、行人跟踪等计算机视觉技术相结合,广泛应用于智能安防、智慧商业等领域[1]。随着深度学习技术的蓬勃发展,有监督的行人重识别方法获得了优异的性能。然而,由于数据标注成本高、跨域迁移性能降低等问题的存在,限制了行人重识别技术在实际应用中的推广。因此,无监督的行人重识别方法逐渐受到越来越多的关注。

无监督行人重识别方法主要分为无监督域自适应(unsupervised domain adaptation, UDA)方法和完全无监督学习(unsupervised learning, USL)方法两类。相较于 UDA 方法,USL 方法所需的无标签图像很容易通过目标检测技术从视频监控系统中获得。近几年来,USL 行人重识别方法得到了快速发展。大多数方法基于聚类技术生成伪标签,并在伪标签的监督下迭代训练模型。这些方法往往在各种对比损失的设计[2-4]、噪声伪标签的细化[4-5]等方面进行改进,却鲜少关注特征提取网络的改进。

行人往往具有相似的形状,再加上视角、光照、姿态和背景等因素的影响,行人图像之间存在许多相似之处,这使得行人重识别任务比普通的图像分类任务更加困难。在这种情况下,细粒度特征的提取就变得尤为重要,可以增强模型对不同个体之间细节信息的区分,从而提高行人重识别模型的准确性[6-7]。

基于卷积神经网络(convolutional neural network, CNN)的行人重识别方法通过刚性条纹[8-9]、注意力[10-11]等技术提取细粒度特征,取得了良好的效果。然而,基于 CNN 的特征提取方法存在两个局限性:1)感受野有限:由于卷积核的限制,CNN 在学习长距离的空间结构特征时有限,无法有效捕捉远程依赖关系。2)细节信息丢失:CNN 的下采样操作会降低输出特征图的空间分辨率,极大地影响识别具有相似外观的物体的能力。近年来,基于 Transformer[12] 的网络,例如 ViT (vision transformer)[13],逐渐兴起,并在行人重识别领域产生了一些新的解决方案。具体来说,基于 Transformer 的网络在行人重识别中的优势包括:1)多头自注意力机制能够有效建模图像的远程依赖关系,驱动模型关注不同的人体部位之间的关系。2)不依赖下采样操作,可以保留输入图像的高空间分辨率,从而更清晰地保留细节信息,增强模型对个体间细微差异的辨别能力。

ViT 网络将图像编码为块标记(patch token, PT)序列,输入到一系列标准 Transformer 编码器层中,能够在多个图像识别数据集上取得与最先进 CNN 模型比肩甚至更好的结果,同时训练所需的计算资源更少[13]。这些优点使得 ViT 在行人重识别任务中得到广泛应用。一些基于 ViT 的无监督行人重识别方法也探索了细粒度特征的提取。Li 等[14]提出一种双分支 Transformer 无监督行人重识别网络架构,通过对每个分支的图像块进行重构,并将其均匀划分为多个条纹,以提取细粒度特征。Sharma 等[15]提出在 ViT 网络框架中,采用类似 PCB 方法[16]的策略,将全局类标记(class token, CT)和图像块标记经加权组合得到的局部特征送入到分类器中。但是上述方法仍然存在一些问题,例如不能准确地定位人体部位,或聚合了不重要的背景区域中的信息,或忽略了包含可识别信息的非人体部位,如背包等。

最近,视觉-语言模型在多个领域取得了成功应用,为解决上述问题提供了新思路。CLIP[17]是该领域的代表性方法,它通过同时对图像和文本进行对比学习,使模型能够理解图像和文本之间的语义关系。CLIP[17]的设计使其在多个任务上表现出色,包括图像分类、物体检测、文本检索等,这使得 CLIP[17]成为一个通用的多模态模型,可以适用于多种不同的应用场景。GLIP 方法[18]是 CLIP[17]拓展版本,其架构与 CLIP[17]非常相似。只不过 CLIP[17]是图像-语言层面的,而 GLIP[18]是区域-短语层面的,更适合细粒度的任务。

为更好地使用 Transformer 网络提取行人图像中的细粒度特征,提出一种基于 ViT 的细粒度特征增强无监督行人重识别方法。首先,采用预训练的视觉-图像模型 GLIP[18]获取行人图像中人体局部区域的定位框并生成掩码。然后,利用该掩码引导 ViT 网络中的自注意力层分别计算全局和局部自注意力,以提取全局和局部特征。具体来说,在输入序列中引入可学习的局部标记(local token, LT),并使其与对应的图像块子集交互来学习局部特征表示。

此外,在网络的后期阶段,还添加了一个空间信息增强模块。该模块通过挖掘人体局部区域内代表性图像块之间的空间上下文关系来突显人体区域。借助 Transformer 的自注意特性,类标记或局部标记与图像块标记之间的注意力权重与图像块中是否包含目标信息高度相关,因此可以据此选择具有代表性的块标记。计算图像块之间的相对位置关系和语义关系可用于构造描述目标空间结构信息的图结构,并通过图卷积提取特征并注入到主干网络中,以增强网络对空间特征的表示能力。

最终,采用提取到的全局和局部特征,分别计算在线和离线对比损失,使模型能够学习如何在这些细粒度特征上进行有效比较,从而更准确地判断样本之间的相似性。

1 网络结构

网络采用与基线网络[14]相同的主干网络，即 TransReID-SSL[19]中使用的基础 Transformer 架构。该网络构建在 ViT 网络基础之上，并引入了两个策略，使其更加适应于行人重识别任务。首先，受到基于 CNN 的行人重识别方法中广泛使用的 IBN-Net 网络[20]的启发，该网络引入了基于实例-批次归一化（instance-batch normalization，IBN）的卷积流，替代 ViT 中传统的卷积结构，以提高训练的稳定性和泛化能力。其次，除了在图像块标记序列中添加类标记和位置编码外，该网络还引入了相机编码。通过对相机信息进行可学习的向量编码，有效地减轻了特征学习中的相机偏差。

在该网络基础上进行了两点创新。首先，在网络输入序列中添加可学习的局部标记，在自注意力层中使其与对应的局部区域内的图像标记进行交互，从而提取细粒度特征表示。其次，在网络的后期阶段引入空间信息增强模块，通过挖掘代表性图像块之间的空间上下文关系，增强网络对空间信息的提取能力。

1.1 网络整体框架

网络的整体框架如图 1 所示。

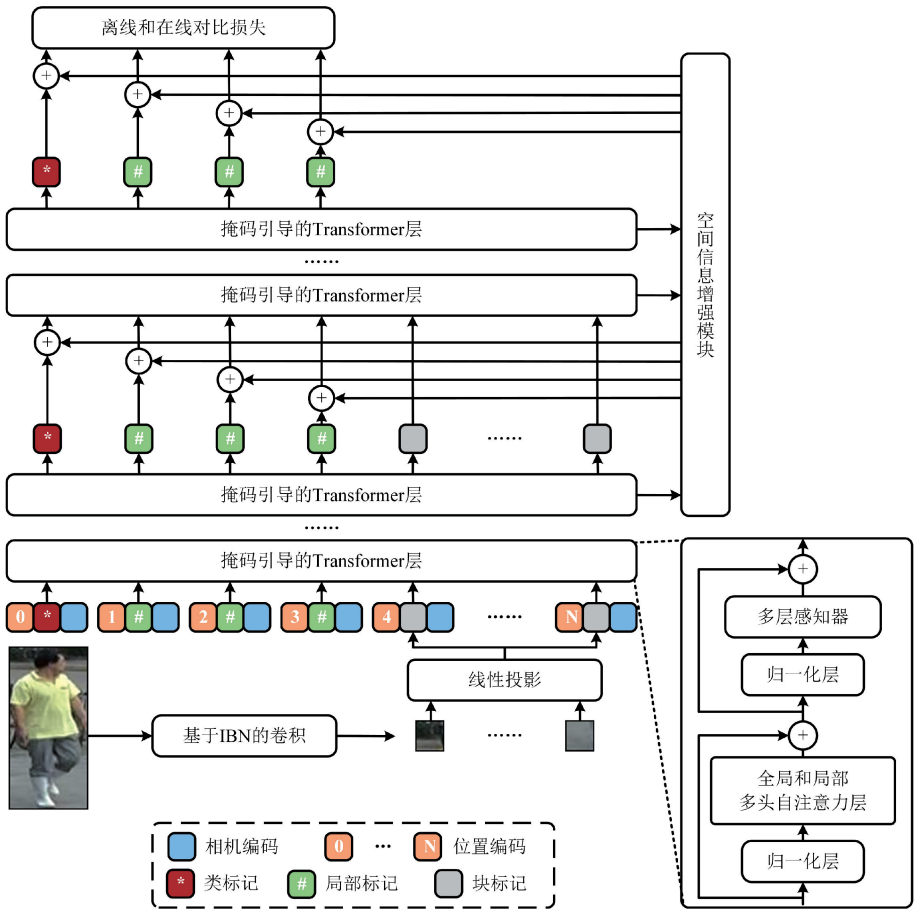


图 1 网络整体框架图

Fig. 1 The overall framework of our method

对于输入图像 $\boldsymbol{x} \in \boldsymbol{R}^{H\times W\times 3}$，经过 IBN 卷积流处理后生成特征图 $\boldsymbol{x}' \in \boldsymbol{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$，其中 H、W、C 分别表示图像的高度、宽度、通道数。将特征图划分为 $N = N_H \times N_W$ 个不重叠的图像块，其中 $N_H = H/S$、$N_W = W/S$，每个块的大小为 $S \times S$，以适应 Transformer 架构的输入。通过可训练的线性投影层将每个图像块映射到 D 维向量，将该映射的输出特征 $f \in \boldsymbol{R}^D$ 称为块标记 PT。在块标记序列的开头添加一个可学习的类标记 CT 用来提取全局特征。在类标记的后面添加 K 个可学习的局部标记 LT 用来提取局部特征。此外，将可学习的位置嵌入（position embedding，PE）和相机嵌入（camera embedding，CE）也添加到序列中，以保留位置信息和相机视角信息。因此，初始输入 z^0 的定义如式（1）所示。

$$\begin{aligned} &\boldsymbol{f}_i = \psi_i(\mathrm{ICS}(\boldsymbol{x})), i = 1,\cdots,N \\ &\boldsymbol{z}^0 = [CT;p_1;\cdots;p_K;\boldsymbol{f}_1;\cdots;\boldsymbol{f}_N] + \boldsymbol{PE} + \lambda_c\boldsymbol{CE} \end{aligned} \tag{1}$$

其中，ICS（·）表示 IBN 卷积流，ψ_i 表示划分和映射

操作。$CT \in \boldsymbol{R}^{1\times D}$ 表示类标记 CT，$\{p_i\}_{i=1}^{3} \in \boldsymbol{R}^{1\times D}$ 表示局部标记 LT，$\boldsymbol{PE} \in \boldsymbol{R}^{(N+4)\times D}$ 表示位置嵌入，$\boldsymbol{CE} \in \boldsymbol{R}^{(N+4)\times D}$ 表示相机嵌入，λ_c 表示加权参数。

接下来，z^0 将被输入到由 L 个掩码引导的 Transformer 层组成的网络中，每个层由全局和局部多头自注意力（global and local multi-head self-attention，GLMSA）和多层感知器（multi-layer perceptron，MLP）以及应用在 GLMSA 和 MLP 之前的两个归一化层（layer norm，LN）组成。其中 MLP 模块由两个全连接层和一个激活函数组成，用于对经过 GLMSA 模块处理后的特征进行非线性映射和变换。GLMSA 模块的详细结构将在 1.2 小节进行介绍。在网络的后期，采用空间信息增强模块（spatial information enhancement module，SIEM）来凸显行人区域。第 l 层（其中 $l \in \{1,2,\cdots,L\}$）的定义如式(2)所示。

$$\hat{z}^{l-1} = \mathrm{GLMSA}(\mathrm{LN}(z^{l-1})) + z^{l-1}$$
$$z^{l} = \begin{cases} \mathrm{MLP}(\mathrm{LN}(\hat{z}^{l-1})) + \hat{z}^{l-1}, l \leqslant 6 \\ \mathrm{SIME}(\mathrm{MLP}(\mathrm{LN}(\hat{z}^{l-1})) + \hat{z}^{l-1}), \text{其他} \end{cases} \tag{2}$$

网络最终的输出表示如式(3)所示。

$$z^{L} = [CT^{L}; p_1^{L}; \cdots; p_K^{L}; f_1^{L}; \cdots; f_N^{L}] \tag{3}$$

对于一个无标签数据集，在初始化阶段，首先通过网络提取特征，得到全局特征 $\boldsymbol{g} = CT^{L}$ 和局部特征 $\{p_k^{L}\}_{k=1}^{K}$。然后，采用 DBSCAN 聚类算法根据全局特征对所有图像进行聚类并分配伪标签。将带有伪标签的数据集表示为 $D = \{(x_i, \tilde{y}_i)\}_{i=1}^{N_I}$，其中 $\tilde{y}_i = \{1,\cdots,N_l\}$ 表示生成的伪标签，N_l 表示图像总数。在训练阶段，分别对全局特征和局部特征计算离线和在线对比损失。

1.2　全局和局部多头自注意力模块

在初始化阶段，使用视觉-语言模型定位图像中的行人区域，选择目前较为先进的 GLIP 模型[18]，该模型更适合细粒度的任务。如图 2 所示，输入检测图像以及用句号字符连接的所有待检测类别，即可得到图像中包含这些类别区域的定位框。通过对定位框进行去重、合并、补充等后处理操作，获得行人各局部区域的准确定位，主要包括头部、上半身、下半身三部分。

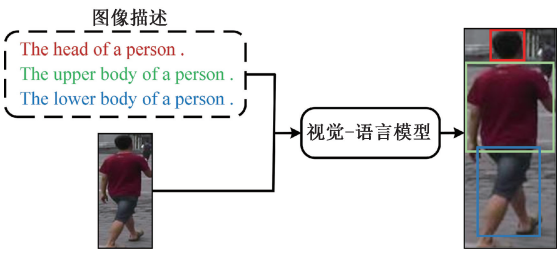


图 2　视觉-语言模型定位局部区域模块

Fig. 2　Visual language model localization local region module

下一步，定位框被转换为二进制掩码 $\{Mask_{p_k}\}_{k=1}^{K} \in \boldsymbol{R}^{H\times W}$。GLIP 模型[18]通过融合图像和文本信息，能够理解图像中的语义关系，从而更好地捕捉行人局部区域的细节信息，为后续整个网络的训练和优化提供可靠的初始信息。

每个 Transformer 层的自注意力模块都由全局注意力和局部注意力组成，如图 3 所示。类标记 CT 和块标记 PT 通过全局自注意力模块提取全局信息，而局部标记 LT 和块标记 PT 通过局部自注意力提取局部特征。全局注意力采用 Transformer 中的标准多头自注意力机制，定义如式(4)所示。

$$\mathrm{Attn}(z^{l-1}, h) = softmax\left(\frac{\boldsymbol{Q}_h^{l}(\boldsymbol{K}_h^{l})^{\top}}{\sqrt{d}}\right)\boldsymbol{V}_h^{l} \in \boldsymbol{R}^{(N+1)(N+1)}$$
$$\boldsymbol{Q}_h^{l} = \boldsymbol{W}_{q,h}^{l} z^{l-1}, \boldsymbol{K}_h^{l} = \boldsymbol{W}_{k,h}^{l} z^{l-1}, \boldsymbol{V}_h^{l} = \boldsymbol{W}_{v,h}^{l} z^{l-1} \tag{4}$$

其中，$l \in \{1,\cdots,L\}$，$h \in \{1,\cdots,H\}$，$d = D/H$ 表示注意力头的嵌入尺寸，$\boldsymbol{W}_{q,h}^{l}$、$\boldsymbol{W}_{k,h}^{l}$、$\boldsymbol{W}_{v,h}^{l}$ 表示线性投影矩阵，H 表示注意力头的个数。

整个多头自注意力机制的表达式如式(5)所示：

$$\mathrm{MSA}(z^{l-1}) = \mathrm{Concat}([\mathrm{Attn}(z^{l-1}, h)]_{h=1}^{H})\boldsymbol{W}^{l} \tag{5}$$

其中，Concat(·)表示拼接操作，$\boldsymbol{W}^{l}$ 表示输出的线性投影矩阵。

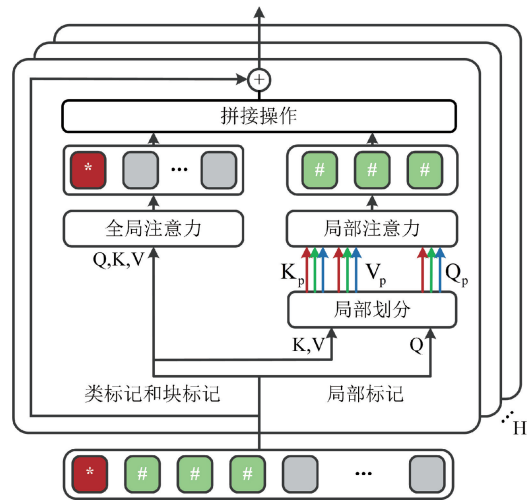


图 3　全局和局部自注意力模块

Fig. 3　Global and local self-attention module

局部自注意力与全局注意力采用计算公式相同，但局部标记 LT 只与部分块标记 PT 交互，而不是所有块标记。根据二进制掩码 $\{Mask_{p_k}\}_{k=1}^{K} \in \boldsymbol{R}^{H\times W}$ 可以得到图像块层面的掩码 $\{P\text{-}Mask_{p_k}\}_{k=1}^{K} \in \boldsymbol{R}^{H\times W}$，根据掩码即可获得局部标记 LT 对应的区域子集 $\{\Theta_k\}_{k=1}^{K}$，如图 4 所示。

局部标记 LT 只与对应子集内的块标记 PT 进行交互，从而只学习该局部的特征表示。具体来说，部分标记只用作查询向量 $\{(\boldsymbol{Q}_h^{l})_k^{LT}\}_{k=1}^{K}$，而且彼此之间不相互交

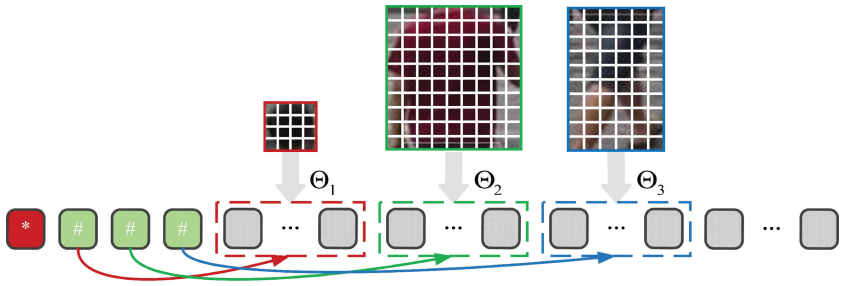


图 4 局部区域图像块划分

Fig. 4 Local area image token division

互,保证每个局部标记的独立性。各个局部的注意力表达式如式(6)所示。

$$\mathrm{Attn}(\boldsymbol{z}_{\Theta_k}^{l-1}) = softmax\left(\frac{(\boldsymbol{Q}_h^l)_k^{LT}(\boldsymbol{K}_h^l)_{\Theta_k}^{\top}}{\sqrt{d}}\right)(\boldsymbol{V}_h^l)_{\Theta_k} \tag{6}$$

其中,$(\boldsymbol{K}_h^l)_{\Theta_k}^{\top}$、$(\boldsymbol{V}_h^l)_{\Theta_k}$ 分别表示属于子集 Θ_k 的键向量和值向量。全局注意力和局部注意力计算完成后,将类标记 CT、局部标记 LT、经全局注意力计算后的块标记 PT 重新拼合,送入 MLP 层中。

1.3 空间信息增强模块

为了在不同图像块之间建立空间连接,建模行人不同区域之间的关系,学习空间上下文信息,并加强主干网络的重要信息提取能力,提出一个空间信息增强模块,如图 5 所示。在 Transformer 中,块标记和类标记之间的注意力权重表示每个块标记对分类结果的重要性,同时也反映该块标记内是否包含目标信息。因此,可以根据注意力权重来定位目标。根据式(4)所示,可以提取各个注意力头中块标记和类标记之间的注意力权重 $\boldsymbol{A}_h^{CT} \in \boldsymbol{R}^{N+1}$,总注意力权重为 $\boldsymbol{A} = \sum_{h=1}^{H}\boldsymbol{A}_h^{CT}$。假设在 $N_W \times N_H$ 的平面中,坐标为 (x,y) 的块标记与类标记之间的注意力权重表示为 $A_{(x,y)}$。为了过滤掉不重要的背景区域,使用平均注意力值作为阈值,重新定义注意力权重,如式(7)所示。

$$\boldsymbol{A}_{(x,y)}^{new} = \begin{cases} A_{(x,y)}, A_{(x,y)} > \bar{A} \\ 0, \text{其他} \end{cases} \tag{7}$$

采用极坐标来表示不同块之间的空间关系。给定一个参考块 $P_0 = P_{(x_0,y_0)}$,其中 (x_0,y_0) 表示在平面 $N_H \times N_W$ 中的坐标,以及一个参考水平方向。块 $P_{(x,y)}$ 的极坐标可以定义为 $(\rho_{(x,y)},\theta_{(x,y)})$,如式(8)所示。

$$\begin{aligned} \rho_{(x,y)} &= \sqrt{\left(\frac{x-x_0}{N_W}\right)^2 + \left(\frac{y-y_0}{N_H}\right)^2} \in (0,1] \\ \theta_{(x,y)} &= \frac{\arctan2(y-y_0, x-x_0) + \pi}{2\pi} \end{aligned} \tag{8}$$

其中,$\rho_{(x,y)}$ 表示 $P_{(x,y)}$ 和参考块 $P_0 = P_{(x_0,y_0)}$ 之间的相对距离,$\theta_{(x,y)}$ 表示相对于水平方向的归一化极角。函数 arctan2(·) 返回从笛卡尔坐标转换为极坐标的角度,范围属于[-π,π]。为了保证 θ 值分布范围较广,理想情况下,应该选择目标范围内的块作为参考块,因此选择注意力值最大的块作为参考块 $P_0 = P_{(x_0,y_0)}$。

为了将空间信息引入网络,采用图卷积神经网络来获得空间特征。首先构造图结构:1)使用图像块的极坐标 $(\rho_{(x,y)},\theta_{(x,y)})$ 作为节点;2)基于 Transformer 中类标记和块标记之间的注意力权重,可以得到节点间边的权重值,即 $\boldsymbol{Adj} = A^{new} \times (A^{new})^{\top}$,其中与不重要的块相连的边权值设为 0,以减轻这些块对网络的影响。然后使用两层图卷积提取空间信息并合并到 Transformer 中,如式(9)所示。

$$\boldsymbol{Spa} = \sigma(\boldsymbol{Adj} \times \sigma(\boldsymbol{Adj} \times \boldsymbol{x} \times W^1) \times W^2) \tag{9}$$

其中,W^1 和 W^2 为可学习参数,$\sigma(\cdot)$ 为激活函数。将得到的空间特征 $\boldsymbol{Spa}$ 添加到类标记和局部标记中,从而将空间信息引入 Transformer 网络主干。通过端到端训练,可以逐渐对目标的空间信息进行准确的建模,并突出了重要的图像块,从而提高模型的分类能力。

1.4 损失函数

在网络初始化阶段,首先构建一个代理级别的存储库 $\boldsymbol{M} \in \boldsymbol{R}^{N_P \times D}$,其中每个条目存储代理的全局特征中心。在反向传播过程中,当输入图像 $\boldsymbol{x}_i$ 时,与伪类 $\tilde{y}_i$ 相对应的条目会被更新,更新方式如式(10)所示。

$$\boldsymbol{M}[\tilde{y}_i] \leftarrow \mu\boldsymbol{M}[\tilde{y}_i] + (1-\mu)\boldsymbol{g}_i \tag{10}$$

其中,$\boldsymbol{M}[y_i]$ 表示存储库 $\boldsymbol{M}$ 中的第 $\tilde{y}_i$ 个条目,$\mu \in [0,1]$ 表示更新速率。

借助存储库 $\boldsymbol{M}$,分别计算离线和在线对比学习损失。对于图像 $\boldsymbol{x}_i$,离线对比损失根据离线聚类结果,生成一个正类别代理集 P_1 和一个困难负类别代理集 Q_1。P_1 和 Q_1 分别存储正类别和负类别代理的索引。离线对比损失计算公式如式(11)所示。

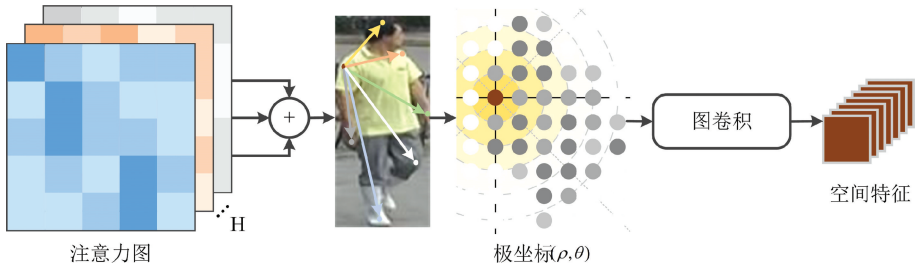


图 5　空间信息增强模块

Fig. 5　Spatial information enhancement module

$$L_{off}^{g}=-\sum_{i=1}^{B}\left(\frac{1}{|P_1|}\sum_{u\in P_1}\log\frac{S(u,\boldsymbol{g}_i)}{\sum_{p\in P_1}S(p,\boldsymbol{g}_i)+\sum_{q\in Q_1}S(q,\boldsymbol{g}_i)}\right)\tag{11}$$

其中，$S(u,\boldsymbol{g}_i)=\exp(\boldsymbol{M}[u]^{\mathrm{T}}g_i/\tau)$，$\tau$ 表示温度系数，$|\cdot|$表示集合的基数，B 表示批处理大小。

由于聚类结果可能存在噪声，进一步采用在线关联策略。通过实例代理平衡相似度和相机感知的最近邻方案，为每个图像 $\boldsymbol{x}_i$ 动态地选择正代理集 P_2 和负代理集 Q_2。在线对比损失计算公式如式(12)所示。

$$L_{on}^{g}=-\sum_{i=1}^{B}\left(\frac{1}{|P_2|}\sum_{u\in P_2}\log\frac{S(u,\boldsymbol{g}_i)}{\sum_{p\in P_2}S(p,\boldsymbol{g}_i)+\sum_{q\in Q_2}S(q,\boldsymbol{g}_i)}\right)\tag{12}$$

上述损失是根据全局特征进行定义的，为了采用局部特征，另外构建一个存储库，并为每个局部特征定义两种类型的损失，同时保持聚类步骤不变，即只使用全局特征进行聚类。训练时的整体损失如式(13)所示。

$$L=L_{off}^{g}+L_{on}^{g}+\lambda_p\frac{1}{P}\sum_{p=1}^{P}(L_{off}^{p}+L_{on}^{p})\tag{13}$$

其中，λ_p 是平衡全局和部分损失的权重参数。

2　实验结果与分析

2.1　数据集与评价指标

实验基于 3 个广泛使用的、大规模的数据集开展，分别是 Market-1501[21]、MSMT17[22] 和 PersonX[23]。具体而言，Market-1501 数据集源自于清华大学的真实场景，由 6 台摄像机捕获，收录了 1 501 个行人的 32 668 张图像；MSMT17 数据集则在北京大学校园内采集，规模更为庞大，覆盖了室内与室外多种复杂环境，通过 15 台摄像机共收集到 4 101 个行人的 126 411 张图像；PersonX 数据集采用 Unity 技术进行虚拟合成，模拟了 6 台摄像机视角视角下的 45 792 张图像，涉及 1 266 个虚拟行人。

选用的两个评估指标分别为：累计匹配特性(cumulative matching characteristics，CMC)曲线和平均准确率均值(mean average precision，mAP)。CMC 聚焦于前 K 幅图像匹配成功的概率，该方法主要分析 Rank-1、Rank-5、Rank-10 3 种情况，即分别为前 1 幅、5 幅、10 幅图像匹配成功的概率。此外，还计算了每个查询图像的平均准确率，并据此汇总得出所有查询图像平均准确率的均值，即 mAP 值。

2.2　实验参数设置

主干网络建立在 ViT-Small/16 模型[8] 的基础上，共包含 12 层，每个 GLMSA 中有 6 个注意力头，特征维度为 384。采用的预训练模型是在大规模未标记数据集 LUPerson[24] 上训练得到。每幅图像的大小被调整为 384×128，并进行随机水平翻转、裁剪、擦除操作。图像块大小 S 设置为 16，这意味着经过基于 IBN 卷积流后产生的特征图被分割成 192 个大小为 16×16 的图像块。

相机嵌入的权重参数 λ_c 设置为 3。存储库 M 的更新速率 μ 设置为 0.2，温度参数 τ 设置为 0.07，批处理大小 B 设置为 32。局部特征的对比损失权重参数 λ_p 设置为 0.1。

采用动量为 0.9 的 SGD 优化器，权重衰减设置为 0.000 5，共进行 50 个轮次的训练。学习率设置为 0.000 35，并且采用一个预热速度调度器进行调节，该调度器在初始学习率上乘以 0.01，并线性地扩大到前 10 个轮次，在第 20 和 40 个轮次，都除以 10。为了加速训练，训练时使用半精度浮点数(FP16)计算，测试时使用全精度(FP32)计算。所有实验都是在 Ubuntu18.04 系统中进行的，使用 1 张 24 GB 显存的 RTX3090Ti 显卡。

2.3　与最新方法的比较

本节在 Market-1501、MSMT17 和 PersonX 3 个数据集上，将所提方法的实验结果与近几年最新的方法进行比较，包括 UDA 方法和 USL 方法，比较结果如表 1 所示。其中，最优结果加粗表示，次优结果加下划线表示，“-”表示原文中没有该项结果。

表 1 所提方法与最新方法的比较

Table 1 The comparison between our method and the latest methods

方法		Market-1501		PersonX		MSMT17	
		mAP/%	Rank-1/%	mAP/%	Rank-1/%	mAP/%	Rank-1/%
Unsupervised domain adaptation(UDA)							
SpCL[25]	NeurIPS'20	76.7	90.3	78.5	91.1	26.8	53.7
Isobe et al[26]	ICCV'21	83.4	94.2	–	–	36.3	66.6
DARC[27]	AAAI'22	85.1	94.1	–	–	35.2	64.5
CDRL[28]	ICCV'23	84.7	93.8	–	–	40.3	70.0
PAT[29]	TST'23	86.3	94.2	–	–	45.3	70.1
Unsupervised learning(USL)							
HCT[30]	CVPR'20	56.4	80.0	–	–	–	–
SpCL[25]	NeurIPS'20	73.1	88.1	72.3	88.1	42.3	72.3
CAP[31]	AAAI'21	79.2	91.4	–	–	36.9	67.4
CC[2]	Arxiv'21	82.1	92.3	84.7	94.4	27.6	56.0
ICE[3]	ICCV'21	82.3	93.8	–	–	38.9	70.2
TransReID-SSL[19]	Arxiv'21	89.6	95.3	–	–	50.6	75.0
MCRN[32]	AAAI'22	80.8	92.5	–	–	31.2	63.6
O2CAP[4]	Arxiv'22	82.7	92.5	–	–	42.4	72.0
TransCL[33]	IJCNN'22	82.9	93.0	89.2	95.4	41.3	68.6
GRACL[34]	TCSVT'22	83.7	93.2	87.9	95.3	34.6	64.0
PPLR[5]	CVPR'22	84.4	94.3	–	–	42.2	73.3
SPLT[35]	ICSP'22	89.8	95.1	–	–	42.4	67.1
STDA[36]	TITS'23	82.7	93.1	–	–	31.8	62.6
DCCT[37]	TCSVT''23	86.3	94.4	87.6	95.0	41.8	68.7
TMGF[14]	WACVW'23	89.5	95.5	–	–	58.2	83.3
IIDCL[38]	TOMM'24	89.9	95.4	90.3	95.8	55.0	80.4
本文	–	**90.3**	**95.9**	**91.3**	**96.1**	**59.2**	**83.5**

表 1 中除 PAT[29]、TransReID-SSL[19]、TransCL[33]、TMGF[14]、SPLT[35]、IIDCL[38]方法之外,其他方法的主干网络采用的都是在 ImageNet 上预训练的 CNN 网络。可以看出,采用 Transformer 作为特征提取网络大大提高了识别性能。所提方法建立在 TMGF[14]方法的网络框架之上,但是改进了细粒度特征提取模块,并添加了空间信息增强模块。在 Market-1501 数据集上,所提方法的 mAP 和 Rank-1 分别超过 TMGF[14]方法 0.8% 和 0.4%。在 MSMT17 数据集上,所提方法的 mAP 和 Rank-1 分别超过 TMGF[14]方法 1.0% 和 0.2%。与使用 CC[1]中定义的聚类对比损失的 TransReID-SSL[19]方法相比,所提采用掩码引导的 Transformer 编码器提取多粒度特征,并采用 O2CAP[4]中的离线和在线对比学习方式计算全局和局部特征的损失。在 Market-1501 数据集上,所提方法的 mAP 和 Rank-1 分别超过 TransReID-SSL[19]方法 0.7% 和 0.6%。在最具挑战性的数据集 MSMT17 上,所提方法的 mAP 和 Rank-1 分别超过 TransReID-SSL[19]方法 8.6% 和 8.5%。与最新方法 IIDCL[35]相比,所提方法在 Market-1501 数据集上 mAP 和 Rank-1 分别提升了 0.4% 和 0.5%,在 PersonX 数据集上 mAP 和 Rank-1 分别提升了 1.0% 和 0.3%,在 MSMT17 数据集上 mAP 和 Rank-1 分别提升了 4.2% 和 3.1%。这些结果表明,所提方法在提取细粒度特征和空间特征方面具有更强的能力,从而显著提升了行人重识别的性能。

2.4 消融实验

在 Market-1501 与 MSMT17 两个数据集上实施了一系列详尽的消融实验,旨在验证所设计网络架构中各个关键模块的实际效能。实验成果汇总于表 2,其中,“M1”表示 TMGF[14]方法的原始结果,“M2”表示对 TMGF[14]方法的复现结果,“M3”表示去除 TMGF[14]中原有细粒度模块后的结果,“M4”和“M5”分别表示在“M3”基础上添加 GLMSA 和 SIME 模块后的实验结果,“M6”表示在“M3”基础上同时添加 GLMSA 和 SIME 模块后的实验结果。

对比表 2 中模型“M4”和“M2”的结果,可以观察到,提出的掩码引导的 Transformer 网络相较于 TMGF[14]方法的复现结果在细粒度处理方面展现出了显著的性能提升。在 Market-1501 数据集上测试,mAP 和 Rank-1 指标分别实现了 1.3% 和 0.4% 的提升。在 MSMT17 数据集上测试,mAP 和 Rank-1 指标分别实现了 1.6% 和 0.3% 的提升。进一步将表 2 中第 4 行的“M4”和第 1 行的“M1”的实验结果进行对比,可以看到所提方法在精度方面也

实现了对 TMGF[14] 原文结果的超越。在 Market-1501 数据集上测试，mAP 和 Rank-1 指标分别实现了 0.3% 和 0.4% 的提升。在 MSMT17 数据集上测试，mAP 指标实现了 0.4% 的提升。由此，可以看出提出的细粒度特征提取模块，相较于 TMGF[14] 方法，展现出了更为卓越的性能，从而验证了该模块设计的有效性和优越性。

表 2　Market-1501 和 MSMT17 数据集上消融实验结果

Table 2　Results of ablation studies on Market-1501 and MSMT17

方法	Market-1501				MSMT17			
	mAP/%	Rank-1/%	Rank-5/%	Rank-10/%	mAP/%	Rank-1/%	Rank-5/%	Rank-10/%
M1(TMGF[14])	89.5	95.5	98.0	98.7	58.2	83.3	90.2	92.1
M2	88.5	95.5	98.1	98.8	57.0	82.9	89.7	91.7
M3	89.1	95.7	97.8	98.7	56.7	82.4	89.6	91.6
M4	89.8	95.9	98.2	98.8	58.6	83.2	90.3	92.2
M5	89.3	95.9	98.0	98.5	58.0	83.5	90.0	92.1
M6(本文)	90.3	95.9	98.4	99.0	59.2	83.5	90.3	92.3

为了更好地理解局部标记 LT，本节使用所提方法的模型“M4”和基线模型 TMGF[14] 在 Market1501 数据集上进行了 Attention Rollout[39] 可视化实验，如图 6 所示。首先，分别使用基线模型和所提方法的模型中的类标记 CT 绘制了行人的整体热图，可以看出模型关注了行人的大致整体轮廓。这说明模型主要关注行人的全局信息，但是缺乏对行人局部细节信息的关注程度。然后，采用所提方法的模型中头部、上半身、和下半身对应的局部标记分别绘制了模型对局部的关注热图。与整体热图相比，局部热图展现了模型对于行人不同部位的更为细致的关注情况。头部标记显示了模型对行人头部特征的强烈关注，这可能涉及头部的表情等微小细节。上半身和下半身标记显示了模型对行人其他部位的关注，例如服装纹理、姿态等。

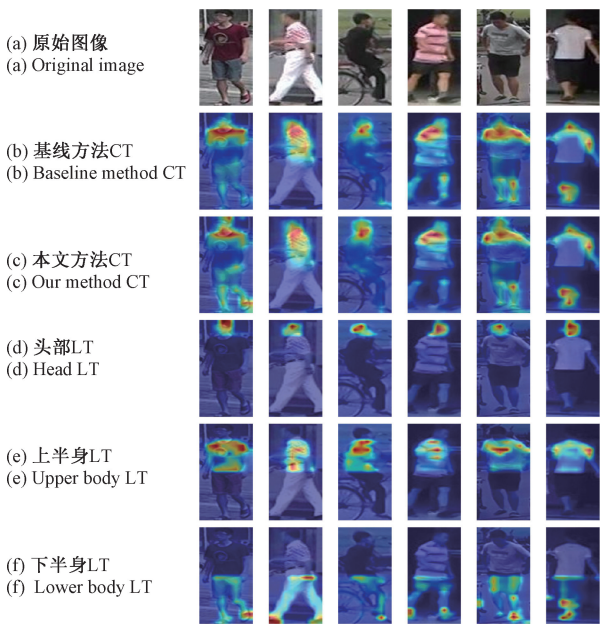


图 6　全局和局部注意力模块在 Market-1501 数据集上注意力可视化

Fig. 6　Global and local attention modules for attention visualization on the Market-1501 dataset

对比表 2 中“M5”和“M3”模型的测试结果，设计的空间信息增强模块在 Market1501 与 MSMT17 数据集上均展现出性能提升。Market-1501 上，mAP 与 Rank-1 指标均提升 0.2%；MSMT17 上，mAP 提升 0.6%，Rank-1 提升 0.2%，验证了该模块的有效性。所提方法的模型“M5”与基线模型 TMGF[14] 在 Market1501 数据集上部分图像的 Attention Rollout[39] 可视化如图 7 所示，从图中可以观察到，所提方法的模型对行人手里拿的矿泉水瓶、衣服上的花纹以及胳膊打弯处的特定区域有更强的关注。这表明空间信息增强模块能够帮助模型加强对这些特定区域的注意力，使模型能够更好地区分不同行人之间的差异，从而提高重识别的准确性。

2.5　超参数分析

随着 Transformer 网络层数的加深，网络对特征的抽象能力逐渐变强，从而能够更好地捕获图像中的语义和

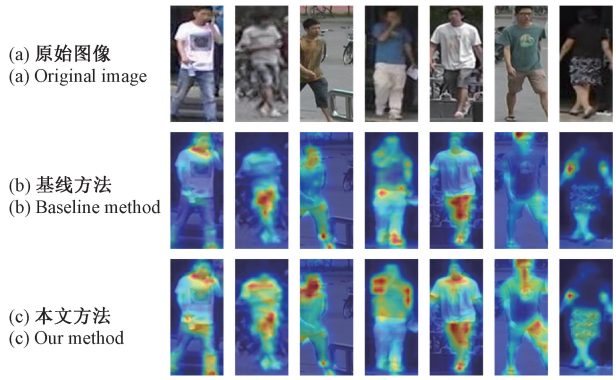


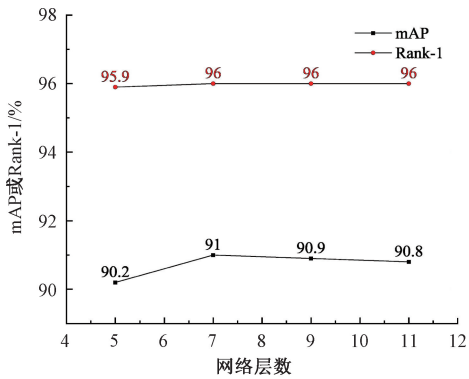
图 7 空间信息增强模块在 Market-1501 数据集上注意力可视化

Fig. 7 Attention visualization of spatial information enhancement module on Market-1501 dataset

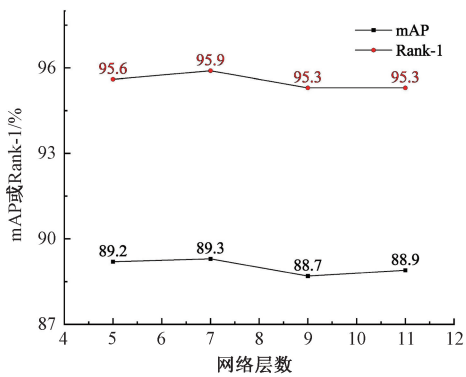
结构信息。因此,在引入空间信息增强模块时,并没有从网络的第一层加入,而是在网络的后半部分才引入。图 8 展示了在不同数据集、不同网络层添加空间信息增强模块的消融实验结果,可以看出,在更深的网络层引入空间信息增强模块,能够取得更好的结果。这说明随着网络的深入,网络更有效地学习了图像中的空间关系,并且更准确地对目标进行建模。

2.6 可视化分析

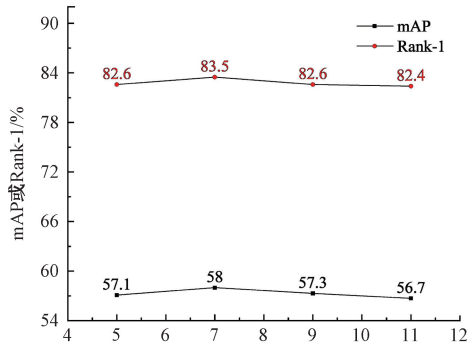
为直观评估模型检索性能,在 Market-1501 数据集上进行了 Rank-10 可视化排序测试。图 9 展示了随机选取的 8 张查询图像在基线模型 TMGF[14] 和所提方法的模型



(a) Market-1501数据集
(a) Market- 1501 dataset



(b) PersonX数据集
(b) PersonX dataset



(c) MSMT17数据集
(c) MSMT17 dataset

图 8 在不同数据集上不同网络层添加空间信息增强模块的实验结果

Fig. 8 Experimental results of adding spatial information enhancement module at different network layers on different datasets

上的对比结果,其中,无边框的图像为查询图像,绿色边框标识的为正确匹配,红色则为错误。从图中可以注意到,当面对外观相似的非同一身份行人时,基线模型表现出较高的误检率,尤其在第 2、4、6、7 个查询实例中,其检索结果几乎全为错误。相比之下,所提方法的模型则显著地改善了这一情况,有效降低了误匹配率,展现出了更强的识别能力。

3 结 论

针对细粒度特征提取不准确、特征表示能力不足的问题,提出了一种基于 ViT 的细粒度特征增强无监督行人重识别方法。首先,利用预训练的视觉-图像模型获取行人图像中人体局部区域的掩码。然后采用 Transformer 网络将图像编码成块标记序列,并利用掩码引导计算全局和局部自注意力,以提取全局和局部特征。随后,提出一个细粒度特征增强模块,通过挖掘人体局部区域内代

(a) 基线方法
(a) Baseline method

(b) 本文方法
(b) Our method

图 9　不同方法在 Market-1501 数据集上 Top-10 排序列表的比较

Fig. 9　Comparison of top-10 sorting lists using different methods on the market-1501 dataset

表性图像块之间的空间上下文关系来突显人体区域。通过计算块之间的相对位置关系和语义关系,构造描述目标空间结构信息的图结构,并通过图卷积进一步提取特征并注入到主干网络中,以增强特征。最终,采用提取到的全局和局部特征,分别计算在线和离线对比损失,使模型能够学习如何在这些细粒度特征上进行有效比较,从而更准确地判断样本之间的相似性。

参考文献

[1] 张晓艳, 张宝华, 吕晓琪,等. 深度双重注意力的生成与判别联合学习的行人重识别[J]. 光电工程, 2021, 48(5): 200388.
ZHANG X Y, ZHANG B H, LYU X Q, et al. The joint discriminative and generative learning for person re-identification of deep dual attention[J]. Opto-Electron Engineering, 2021, 48(5): 200388.

[2] DAI Z, WANG G, YUAN W, et al. Cluster contrast for unsupervised person re-identification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022: 1142-1160.

[3] CHEN H, LAGADEC B, BR′EMOND F. ICE: Inter-instance contrastive encoding for un-supervised person re-identification[C]. Proceedings of the IEEE International Conference on Computer Vision, 2021: 14960-14969.

[4] WANG M, LI J, LAI B, et al. Offline-online associated camera-aware proxies for unsupervised person re-identification [J]. IEEE Transactions on Image Proceeding, 2022, 31: 6548-6561.

[5] CHO Y, KIM W J, HONG S, et al. Part-based pseudo label refinement for unsupervised person re-identification [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022: 7308-7318.

[6] 钱亚萍, 王凤随, 熊磊. 基于局部细化多分支与全局特征共享的无监督行人重识别方法[J]. 电子测量与仪器学报, 2023, 37(1): 106-115.
QIAN Y P, WANG F S, XIONG L. Unsupervised person re-identification method based on local refinement multi-branch and global feature sharing [J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(1): 106-115.

[7] 张勃兴, 马敬奇, 张寿明,等. 利用全局与局部关联特征的行人重识别方法[J]. 电子测量与仪器学报, 2022, 36(6): 205-212.
ZHANG B X, MA J Q, ZHANG SH M, et al. Person re-identification method based on global and local relation features [J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(6): 205-212.

[8] SUN Y, ZHENG L, YANG Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline) [C]. European Conference on Computer Vision, 2018: 480-496.

[9] CHENG D, GONG Y, ZHOU S, et al. Person re-identification by multi-channel parts-based CNN with improved triplet loss function [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1335-1344.

[10] LI W, ZHU X, GONG S. Harmonious attention network for person re-identification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2285-2294.

[11] SI J, ZHANG H, LI C G, et al. Dual attention matching network for context-aware feature sequence based person re-identification [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 5363-5372.

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]. Advances in Neural Information Processing Systems, 2017: 5998-6008.

[13] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale [J]. ArXiv Preprint arXiv: 2010. 11929, 2020.

[14] LI J, WANG M, GONG X. Transformer based multi-grained features for unsupervised person re-identification [C]. Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops, 2023: 42-50.

[15] SHARMA C, KAPIL S R, CHAPMAN D. Person re-

identification with a locally aware transformer[J]. ArXiv Preprint arXiv: 2106.03720, 2021.

[16] SUN Y, ZHENG L, YANG Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]. Proceedings of the European Conference on Computer Vision, 2018: 480-496.

[17] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [J]. ArXiv Preprint arXiv: 2103.00020, 2021.

[18] ZHANG H, ZHANG P, HU X, et al. GLIPv2: Unifying localization and vision-language understanding [C]. Advances in Neural Information Processing Systems, 2022: 36067-36080.

[19] LUO H, WANG P, XU Y, et al. Self-supervised pre-training for transformer-based person re-identification[J]. ArXiv Preprint arXiv: 2021.12084, 2021.

[20] ZHENG L, SHEN L, TIAN L, et al. Scalable person re-identification: A benchmark[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1116-1124.

[21] WEI L, ZHANG S, GAO W, et al. Person transfer gan to bridge domain gap for person re-identifification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 79-88.

[22] SUN X, ZHENG L. Dissecting person re-identification from the viewpoint of viewpoint[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 608-617.

[23] PAN X, LUO P, SHI J, et al. Two at once: Enhancing learning and generalization capacities via ibn-net[C]. European Conference on Computer Vision, 2018: 464-479.

[24] FU D, CHEN D, BAO J, et al. Unsupervised pre-training for person re-identification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021: 14750-14759.

[25] GE Y, ZHU F, CHEN D, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id[C]. Advances in Neural Information Processing Systems, 2020: 11309-11321.

[26] ISOBE T, LI D, TIAN L, et al. Towards discriminative representation learning for unsupervised person re-identification[C]. Proceedings of the IEEE International Conference on Computer Vision, 2021: 8526-8536.

[27] HU Z, SUN Y, YANG Y, et al. Divide-and-regroup clustering for domain adaptive person re-identification[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022: 980-988.

[28] LEE G, LEE S, KIM D, et al. Camera-driven representation learning for unsupervised domain adaptive person re-identification[C]. Proceedings of the IEEE International Conference on Computer Vision, 2023: 11453-11462.

[29] YU S, DOU Z, WANG S. Prompting and tuning: A two-stage unsupervised domain adaptive person re-identification method on vision transformer backbone[J]. Tsinghua Science and Technology, 2023, 28(4): 799-810.

[30] ZENG K, NING M, WANG Y, et al. Hierarchical clustering with hard-batch triplet loss for person re-identification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 13657-13665.

[31] WANG M, LAI B, HUANG J, et al. Camera-aware proxies for unsupervised person re-identification[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 2764-2772.

[32] WU Y, HUANG T, YAO H, et al. Multi-centroid representation network for domain adaptive person re-id[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022: 2750-2758.

[33] TAO Y, ZHANG J, CHEN T, et al. Transformer-based contrastive learning for unsupervised person re-identification[C]. 2022 International Joint Conference on Neural Networks, 2022: 1-9.

[34] ZHANG H, ZHANG G, CHEN Y, et al. Global relation-aware contrast learning for unsupervised person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(12): 8599-8601.

[35] YANG E, LI C, LIU S, et al. Self-supervised pre-training with learnable tokenizers for person re-identification in Railway Stations[C]. 2022 16th IEEE International Conference on Signal Processing, 2022, 1: 325-330.

[36] HE Q, WANG Z, ZHENG Z, et al. Spatial and temporal dual-attention for unsupervised person re-identification[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(2): 1953-1965.

[37] CHEN Z, CUI Z, ZHANG C, et al. Dual clustering co-teaching with consistent sample mining for unsupervised person re-identification [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(10): 5908-5920.

[38] XIONG M, HU K, LYU Z, et al. Inter-camera identity discrimination for unsupervised person re-identification[J].

ACM Transactions on Multimedia Computing, Communications and Applications, 2024. https://doi.org/10.1145/3652858.

[39] ABNAR S, ZUIDEMA W. Quantifying attention flow in transformers[J]. arXiv, 2020, DOI: 10.18653/v1/2020.acl-main.385.

作者简介

程思雨，2024 年于江南大学获得硕士学位，主要研究方向为深度学习、行人重识别。

E-mail: 2446297319@qq.com

Cheng Siyu received her M. Sc. degree from Jiangnan University. Her main research interests include deep learning and person re-identification.

陈莹（通信作者），2005 年于西安交通大学获得博士学位，现为江南大学教授、博士生导师，主要研究方向为机器视觉、信息融合、模式识别。

E-mail: chenying@jiangnan.edu.cn

Chen Ying (Corresponding author) received her Ph. D. degree from Xi′an Jiaotong University in 2005. Now she is professor and Ph. D. supervisor in Jiangnan University. Her main research interests include machine vision, information fusion and pattern recognition.