基于CNN和TransFormer多尺度学习行人重识别方法

陈 莹* 匡 澄
(江南大学物联网工程学院 无锡 214122)

摘 要：行人重识别(ReID)旨在跨监控摄像头下检索出特定的行人目标。为聚合行人图像的多粒度特征并进一步解决深层特征映射相关性的问题，该文提出基于CNN和TransFormer多尺度学习行人重识别方法(CTM)进行端对端的学习。CTM网络由全局分支、深度聚合分支和特征金字塔分支组成，其中全局分支提取行人图像全局特征，提取具有不同尺度的层次特征；深度聚合分支循环聚合CNN的层次特征，提取多尺度特征；特征金字塔分支是一个双向的金字塔结构，在注意力模块和正交正则化操作下，能够显著提高网络的性能。大量实验结果表明了该文方法的有效性，在Market1501, DukeMTMC-reID和MSMT17数据集上，mAP/Rank-1分别达到了90.2%/96.0%, 82.3%/91.6%和63.2%/83.7%，优于其他现有方法。

关键词：行人重识别；TransFormer；CNN；金字塔结构
中图分类号：TN911.73; TP273 **文献标识码**：A **文章编号**：1009-5896(2023)06-2256-08
DOI: 10.11999/JEIT220601

Pedestrian Re-Identification Based on CNN and TransFormer Multi-scale Learning

CHEN Ying KUANG Cheng
(*School of Internet of Things Engineering*, *Jiangnan University*, *Wuxi* 214122, *China*)

Abstract: Person Re-IDentification (ReID) aims to retrieve specific pedestrian targets across surveillance cameras. For the purpose of aggregating the multi-granularity features of pedestrian images and further solving the problem of deep feature mapping correlation, Person Re-Identification based on CNN and TransFormer Multi-scale learning (CTM) is proposed. The CTM network is composed of a global branch, a deep aggregation branch and a feature pyramid branch. Global branch extracts global features of pedestrian images, and extracts hierarchical features with different scales. The deep aggregation branch aggregates recursively the hierarchical features of CNN and extracts multi-scale features. The feature pyramid branch is a two-way pyramid structure, under the attention module and orthogonal regularization operation, it can significantly improve the performance of the network. Experiments on three large scale datasets show the effectiveness of CTM. On the Market1501, DukeMTMC-reID and MSMT17 datasets, mAP/Rank-1 reached 90.2%/96.0%, 82.3%/91.6% and 63.2%/83.7%, which is superior to other existing methods.

Key words: Pedestrian Re-IDentification (ReID); TransFormer; CNN; Pyramid structure

1 引言

行人重识别[1](Person Re-Identification, ReID)也称为行人再识别，旨在跨监控摄像头下检索同一个行人，即给定一个摄像机拍摄的行人图片，从图片库中重新识别该行人的过程，属于图片检索子问题。近年来，随着深度学习的兴起和人们安全意识的提高，行人重识别技术在智能城市、社区安防、罪犯追踪等领域有着广阔的应用前景，已成为计算机视觉领域的热点问题，取得了快速的发展。但是由于现实场景的复杂性，行人重识别存在着诸如背景噪声、人体姿态变化、光照、遮挡、相似性人干扰等问题，且监控图片的分辨率较低，导致行人重识别仍然很具有挑战性。

行人重识别的常见步骤为特征提取和相似性度量，特征提取的目的是提取具有辨别力且鲁棒性强的特征表达，例如颜色、HOG、SIFT、行人步态特征[2]等。相似度度量的目的是设计度量函数，类内距离越小的同时使类间距离越大，例如LMNN，

收稿日期：2022-05-12；改回日期：2022-11-11；网络出版：2022-11-19
*通信作者： 陈莹 chenying@ jiangnan.edu.cn
基金项目：国家自然科学基金(62173160)
Foundation Item: The National Natural Science Foundation of China (62173160)

XQDA等。

基于卷积神经网络(CNN)的模型如ResNet[3], InceptionNet[4], OsNet[5]，能够作为骨干网络从图片中提取具有辨别力的特征表示。CTL-model[6]在检索过程中使用类质心作为特征表示，同时提出了新的损失函数(Centroid Triplet Loss, CTL)；MFN[7]提出了由全局分支、特征擦除分支和局部切块分支组成的网络体系结构，提取更丰富的多粒度特征。虽然CNN模型在行人重识别领域能够取得不错的结果，但会存在两个问题：(1)CNN模型更多地关注局部特征，缺少建立远程依赖关系。虽然注意力模块[8,9]的引入能够在一定程度上缓解这种问题，但其仍然倾向于关注大的连接区域，并不能解决CNN模型的原理问题。(2)下采样操作会降低输出特征图的空间分辨率，影响识别能力。

2017年Vaswani等人[10]率先提出TransFormer网络用于机器翻译，近年来，使用TransFormer来完成视觉任务成为一个新的研究方向[11]。ViT[12]首次证明了当数据足够大时，TransFormer结构模型可以达到最先进的图片分类精度。TransReID[13]首次将TransFormer网络结构引入行人重识别领域，提出拼图块模块(Jigsaw Patch Module, JPM)和侧信息嵌入(Side Information Embeddings, SIE)，以此来产生更鲁棒的特征，同时减少相机/视图变化带来的特征偏差。但是，与CNN模型相比，基于TransFormer的模型往往忽视了局部特征，并且缺乏诸如移位、尺度变化、分层结构等特性，这使得如今CNN模型仍是计算机视觉任务的主流方法。

对此，有学者提出可以把CNN与TransFormer这两种基本结构结合起来，融合这两种结构的优点。Conformer[14]以交互式的方式融合不同分辨率下的局部特征和全局特征；PVT[15]使用一个渐进的金字塔结构来实现高分辨率的输出，同时减少特征图的计算；HAT[16]将CNN+TransFormer结构引入行人重识别领域，提出深度监督聚合模块(Deeply Supervised Aggregation, DSA)来递归聚合来自CNN网络的层次特征，然后引入一种基于TransFormer的特征校准模块(Transformer-based Feature Calibration, TFC)，将TFC插入到每个层次特征中，极大提升模型的性能。然而HAT中TFC模块重复堆叠并不能有效降低深层特征映射的相关性，对图像多粒度特征的学习未做到充分的监督。

对此，本文在文献[16,17]的启发下，提出基于CNN和TransFormer多尺度学习行人重识别方法(person re-Identification based on CNN and TransFormer Multi-scale learning, CTM)。CTM模型是一种3分支网络结构，利用了CNN网络的优势(局部接受域、共享权重、空间自采样)，同时保持了TransFormer网络的优势(动态注意力、全局上下文融合、更好的泛化能力)。CTM模型由全局分支、深度聚合分支、特征金字塔分支组成，3条分支相互作用，相互促进，提取更加丰富的多粒度特征。在全局分支中，使用Resnet提取具有不同尺度和语义信息的层次特征；在深度聚合分支中，引入基于TransFormer的特征校准模块(Transformer-based Feature Calibration, TFC)[16]，通过从全局视图探索信息和促进局部信息来合并多尺度特征；在特征金字塔分支中，提出一个跨尺度连接的金字塔结构(Feature Prymaid Branch, FPB)，提取并整合不同尺度上的不同特征，此外使用正交正则化操作，能够有效地降低特征深层特征映射相关性，从而能够有效提高提取特征的多样性。最后在Market1501[18], DukeMTMC-reID[19], MSMT17[20]数据集上实验结果验证了本文方法的有效性，同时在mAP, Rank1两项指标上大多超越现有方法。

2　本文网络

本节分3个部分介绍CTM，首先介绍本文网络的整体架构；然后具体介绍网络的深度聚合分支、特征金字塔分支、正交正则化操作；最后介绍网络中使用的损失函数。

2.1　网络结构

图1是本文的网络结构图，该网络是一个3分支网络结构，由全局分支(global branch)、深度聚合分支(Deeply Supervised Aggregation Branch, DSAB)、特征金字塔分支(Feature Pyramid Branch, FPB)组成。全局分支，以行人图片作为输入，使用ResNet50作为主干网络提取行人特征。

深度聚合分支基于TFC模块提取多尺度特征，周期性地监督多尺度特征从低级到高级的聚合。首先从Res2, Res3, Res4和Res5中提取不同尺度的层次特征。然后，将TFC模块插入到每个层次特征中，TFC模块用于整合当前尺度特征的语义和细节信息，为下一尺度特征生成全局先验。

基于ResNet50骨干网络，特征金字塔分支首先把Res3, Res4中的浅层特征输入到金字塔结构中，这些浅层特征可以从图像中保存更多的局部特征。然后在该双层金字塔特征结构中，在注意力模块(self-attention module)[21]的基础上，通过正交正则化操作(orthogonal regularizer)降低深层特征相关性的同时，进一步加强了特征的多样性。同时具有不同空间分辨率的特征图之间存在两种不同尺度

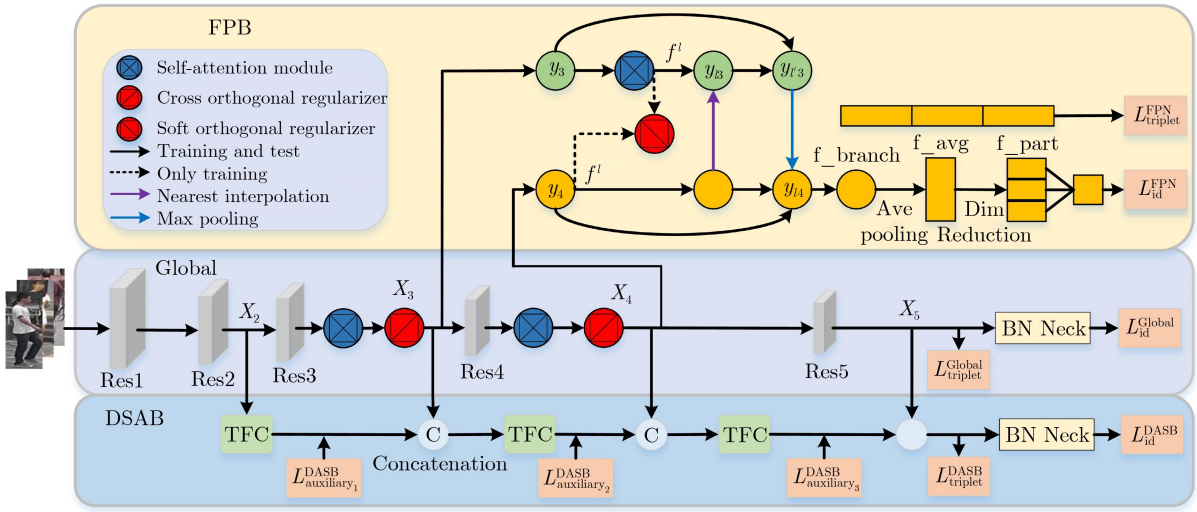


图 1 本文的网络结构

的连接，从下到上通过最近邻插值来增大特征图的尺寸，从上到下通过2×2最大池化来减小特征图的尺寸。最后，通过平均池化、降维操作将1024维度转换为256维度向量，优化分类损失。

2.1.1 深度聚合分支DSAB

在行人重识别任务中，单纯将低级、中级、高级特征进行简单的连接会导致更糟糕的性能，究其原因在于低级特征的语义信息较少。对此基于HAT，引入深度聚合分支(DSAB)，通过多粒度监督聚合过程，缓解低级特征中语义信息较少的问题。TFC模块将层次特征$\boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4$输入TFC模块中，作用在于从全局视图集成先前尺度的特征，然后聚合先前尺度和当前尺度的特征，为下一尺度特征生成全局先验。

如图2所示为TFC模块的结构图，对于第s层的层次特征$\boldsymbol{X}_s \in \boldsymbol{R}^{C_s \times H \times W}$，其中$C_s, H, W$分别表示当前特征图的通道数、高度和宽度。Bottleneck由一系列残差块组成，通过bottleneck将特征$\boldsymbol{X}_s$转换成特征嵌入。尺度变化由最大池化或双线性插值上采样组成，目的是将层次特征调整到相同的分辨率，进行集成。令$\boldsymbol{Z}_s = [\boldsymbol{X}_s; \boldsymbol{Z}_{s-1}] \in \boldsymbol{R}^{(C_s+C_{s-1}) \times H \times W}$，其中$\boldsymbol{X}_s$表示当前尺度的层次特征，$\boldsymbol{Z}_{s-1}$表示上一尺度TFC模块的输出特征。首先，先将$\boldsymbol{Z}_s$展开成2D特征块$\boldsymbol{Z}_s^p \in \boldsymbol{R}^{N \times C_p}$，其中$C_p$表示当前2D块的通道数，$N$表示切块的数量，在实现的过程中具体表现为$N = H \times W / P^2$, $C_p = C_s + C_{s-1} \times P^2$，其中$P$表示块的大小，本文中令$P = 1$。然后为每个块添加可学习的位置嵌入结合空间信息，进而输入到TransFormer中。

TransFormer包含多头注意力层(Multi-head Self-attention Layer)、前馈网络(Feed-forward Network)、归一化层和残差连接。值得注意的是，

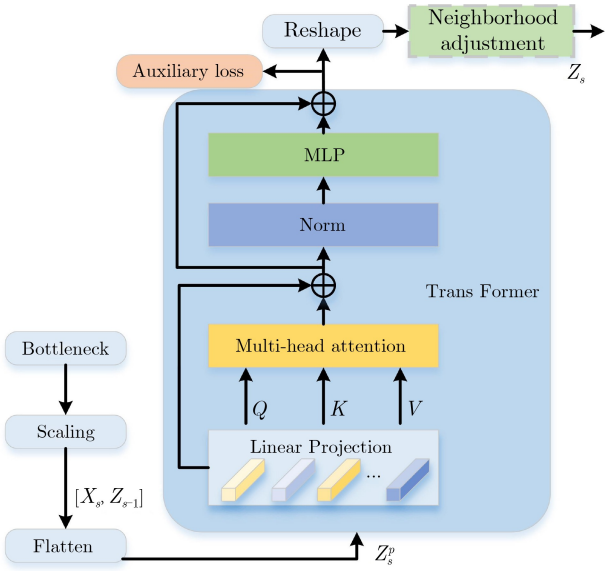


图 2 TFC模块

在自注意力层，输入向量首先会转换成3种截然不同的向量：查询向量 $\boldsymbol{q}$ 、键向量 $\boldsymbol{k}$ 和值向量 $\boldsymbol{v}$，三者的维度均为512。在TransFormer中，近邻域特征在局部特征区域更为明显。因此，除了CLS token，将TransFromer的输出特征大小重塑成与输入端相同的大小，然后输入到一个领域调整模块(neighborhood adjustment module)，其由一组具有批量归一化的卷积层组成。因此TFC模块的最终输出是

$$\boldsymbol{Z}_s = \mathrm{Conv}(\mathrm{Reshape}(\mathrm{TransFormer}(\boldsymbol{Z}_s^p))) \quad (1)$$

此外，引入辅助损失来监督分层聚合，以此来增强TFC交互中语义信息的提取，这个辅助损失由交叉熵损失和3元组损失组成。深度聚合分支的输出可以表示为

$$\mathrm{DSAB}(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n)=\begin{cases}\boldsymbol{X}_1,\ n=1\\ \mathrm{TFC}(\mathrm{CA}(\\ \mathrm{DSAB}(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_{n-1}),\boldsymbol{X}_n)),\ \text{其他}\end{cases}\tag{2}$$

其中，CA表示级联。

2.1.2 特征金字塔分支

随着模型训练加深，DSAB中TFC模块重复堆叠并不能有效降低深层特征映射的相关性。对此，本文引入所设计的特征金字塔分支(FPB)，与现有结构不同，本文特征金字塔存在一种双向的跨尺度链接，在注意力模块和正交正则化的协同下，有效降低特征映射相关性同时显著提升网络性能。

如图1所示，特征金字塔分支从主干网络Resnet50的第3层和第4层获取较浅的特征作为分支的输入，由于获取的较浅层特征具有不同分辨率，因此在两层FPN中存在这两层跨尺度连接，自上而下采用最近邻插值(Nearest Interpolation)，增加特征图大小，自下而上采用2×2的最大池化(Max Pooling)，以此来减小特征图尺寸。同时，如图1中弯曲的箭头所示，类似于Resnet中残差结构的降采样，每层FPN中存在着下采样操作，为输出提供额外具有鉴别力的信息。

特征f_branch的生成过程为

(1) 转换为统一的维度：$y_i=\mathrm{Lateral\_Filter}(X_i),\ i=3,4$；

(2) 最近邻插值计算：$y_{l_3}=\mathrm{Conv}_1(\mathrm{attention}(y_3)+\mathrm{Neareast\_Interpolate}(y_4))$;

(3) 最大池化计算：$y_{l'_3}=\mathrm{Conv}_2(y_{l_3})+\mathrm{Downsample}_3(y_3)$；

$$y_{l_4}=\mathrm{Conv}_4(\mathrm{Conv}_0(y_4)+\mathrm{Max\_Pooling}(y_{l'_3}))+\mathrm{Downsample}_4(y_4);$$

(4) 最终：$\mathrm{f\_branch}=\mathrm{Conv}_5(y_{l_4})$。

其中，$\mathrm{Conv}_i,\ i=1,2,\cdots,5$表示不同的卷积滤波器，attention表示注意力机制，受文献[21]启发，本文在FPB中引入空间信息感知全局注意力(Spatial Relation-aware Global Attention, RGA-S)和通道信息感知全局注意力(Channel Relation-aware Global Attention, RGA-C)，从空间(通道)位置探索节点间的亲和性，更好地捕获全局结构信息。

特征f_branch经过平均池化操作将通道数恢复到1024，得到f_avg。f_avg一方面经过2元自适应均值汇聚层(AdaptiveAvgPool2d)将特征图降采样成3×1，从而进行3元组损失计算；一方面经过降维操作后将特征图平均分割成3等份，级联切块后的局部特征得到f_part，进行ID损失的计算，旨在减少因分割带来的信息损失，提高行人重识别准确率。

2.1.3 正交正则化

本文在注意力模块后使用正交正则化，主要包括传统的软正交正则化[22](Soft Orthogonality Regularizers, SOR)和交叉正交正则化(Cross Orthogonality Regularizers, COR)。需要注意的是SOR只能作用于单张特征图，而本文提出的COR同时监督两张特征图生成，两种正交正则化操作相互作用，共同加强特征多样性。

传统硬正交正则化(Hard Orthogonality Regularizers)对权重的正交性施加严重的约束，其计算依赖于奇异值分解(SVD)。但是对于高维特征，SVD的计算是非常昂贵的，对此，软正交正则化不失为一种可行的替代方案。在此基础上，通过直接正则化$\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}$的条件数来加强正交正则化的效果，该SOR定义为

$$L_{\mathrm{SOR}}=\beta||\lambda_{\mathrm{la}}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}})-\lambda_{\mathrm{sm}}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}})||_2^2\tag{3}$$

其中，β是系数，$\lambda_{\mathrm{la}}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}})$和$\lambda_{\mathrm{sm}}(\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}})$分别表示矩阵$\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}$的最大特征值和最小特征值。

交叉正交正则化(COR)的工作原理与SOR类似，不同的是COR同时监督两张特征图$\boldsymbol{f}^l$和$\boldsymbol{f}^h$的生成，影响多个分支达到互补的效果，能够加强不同位置、不同通道上的正交性，如图1所示。通过最大池化操作，将两张分辨率不同的特征图统一起来，连接成一个更高维的特征图，再应用正交正则化操作进一步增加提取特征的多样性，能够显著提升模型的性能。

2.2 损失函数

在行人重识别研究中，大多通过联合度量损失函数和分类损失函数一起训练，共同约束特征。本文使用ID损失L_{id}(Softmax Loss)、困难3元组损失L_{tri}(Hard Triplet Loss)[23]两个损失函数共同训练网络。同时，对身份标签进行平滑操作(Label Smoothing, LS)[24]。ID损失函数L_{id}为

$$L_{\mathrm{id}}=-\sum_{i=1}^{N}q_i\ln(p_i)\tag{4}$$

$$q_i=\begin{cases}1-\dfrac{N-1}{N}\varepsilon,\ i=y\\ \varepsilon/N,\ \text{其他}\end{cases}\tag{5}$$

其中，N为训练集中行人的个数，p_i为输出的行人身份的预测概率，y表示行人身份的真实标签信息。式(5)表示对身份标签y进行LS操作，ε是一个数值较小的超参数，本文中令$\varepsilon=0.1$。

在三元组损失函数L_{tri}中，图片α和图片p为一对正样本对，图片α和图片n是一对负样本对，表达式如式(6)所示

$$L_{\text{tri}} = (d_{a,p} - d_{a,n} + \alpha)_+ \tag{6}$$

其中，$d_{a,n}$, $d_{a,p}$分别表示正、负样本对的相对距离，$(z)_+ = \max(z, 0)_+$，α是训练阈值，令$\alpha = 0.3$。

综上所述，损失函数为

$$L_{\text{total}} = \sum_i L_{\text{id}}^i + \sum_i L_{\text{tri}}^i + \lambda \sum_{j=1}^{n_{al}} L_{\text{aux}_j}^{\text{DSAB}},$$
$$i \in \{\text{Global}, \text{DSAB}, \text{FPB}\} \tag{7}$$

其中，λ为权重因子，本文令$\lambda = 0.5$。辅助损失L_{aux}由交叉熵损失和3元组损失组成。

3 实验及分析

3.1 数据集

实验在3个主流公开数据集进行有效性验证。

Market1501数据集是在清华大学校园中采集的，由6个摄像头拍摄的1 501个行人和32 668张图片组成。训练集包含751个行人的12 936张图片，测试集包含750个行人的19 732张图片。

DukeMTMC-reID数据集是DukeMTMC的行人重识别子集，包括8个摄像头拍摄的1 404个行人和36 411张图片。其中，训练集包含来自702个行人的16 522张图片，其他的702个行人的17 661张图片作为测试集。

MTMT17数据集是目前最大的单帧数据集。该数据集采集至15个摄像头包含4 101个行人，总计126 441张图片。训练集有3 060个行人，包含32 621张图片；测试集有3 060个行人，包含93 820张图片，其中 11 659张query和82 161张gallery。

3.2 实现细节

实验操作系统为Ubuntu 18.04.6，使用4张NVIDIA RTX 3090 GPU的工作站，显存为24 GByte。基于Pytorch框架，使用ResNet50作为骨干网络，在第4层取消最后一层下采样卷积操作，同时输入图像的大小调整为256×128，使用数据规范化、水平随机旋转和随机擦除作为数据增强的方法。训练时，batch_size设为64(16×4)，使用Adam优化策略优化网络，初始学习率为4e-06，并设置权重衰减为0.000 5，总共训练180个epoch，在前10个epoch中，用warm_up learning，使学习率增长至4e-05，在[50, 70, 90, 110, 130, 150]个epoch后，学习率分别是[1.64e-04, 6.40e-05, 2.56e-05, 1.02e-05, 4.10e-06, 1.64e-06]。

3.3 与最新方法的比较

在上述3个数据集上，与近3年最新的尤其是顶会(CVPR, ICCV, ECCV)所提方法进行比较，比较结果如表1所示。使用平均精度均值(mAP)和首位命中率(Rank-1)指标来评价模型的性能。

正如表1所示，在Market1501数据集上，CTM在mAP, Rank-1指标分别达到90.2%和96.0%，mAP指标以0.1%略低于最好的结果，但是CACE-Net[28]引入了额外的特征对齐模块。在DukeMTMC-ReID数据集上，mAP和Rank-1分别达到82.3%和91.6%，大幅领先其他方法。在MSMT17数据集中，相比于目前精度表现最好的CACE-Net, CTM在mAP和Rank-1上分别提高了1.2%, 0.2%。在3个数据集中，本文模型在各种最新的方法中均接近或取得了最好的性能，显著地提高行人重识别的准确率。

3.4 可视化分析

图3使用Grad-CAM得到可视化结果，图3(a)

表 1 不同方法在公开数据集上的性能比较(%)

方法	出处	Market-1501		DukeMTMC-reID		MSMT17	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
MHN[8]	CVPR2019	85.0	95.1	77.2	89.1	–	–
SONA[25]	ICCV2019	88.6	95.6	78.1	89.3	–	–
OSNet[5]	ICCV2019	84.9	94.8	73.5	88.6	52.9	78.7
HOReID[26]	CVPR2020	84.9	94.2	75.6	86.9	–	–
SNR[27]	CVPR2020	84.7	94.4	72.9	84.4	–	–
CACE-Net[28]	CVPR2020	**90.3**	**96.0**	81.3	90.1	**62.0**	**83.5**
ISP[29]	ECCV2020	88.6	95.3	80.0	89.6	–	–
CDNet[30]	CVPR2021	86.0	95.1	76.8	88.6	54.7	78.9
HAT[16]	MM2021	89.8	95.8	**81.4**	**90.4**	61.2	82.3
L3DS[31]	CVPR2021	87.3	95.0	76.1	88.2	–	–
PAT[32]	CVPR2021	88.0	85.4	78.2	88.8	–	–
本文		**90.2**	**96.0**	**82.3**	**91.6**	**63.2**	**83.7**

是输入图片，图3(b)是使用Torchreid[33]复现出Baseline网络热力图，图3(c)—图3(e)分别是CTM网络全局分支、深度聚合分支和特征金字塔分支输出特征的热力图，图中区域颜色越高亮，表示训练中网络关注越多。相比于Baseline网络关注行人图片的部分区域且时常关注到背景等无用信息，CTM关注区域基本覆盖整个行人，能够勾画出行人的轮廓，对网络有个很好的补充。

3.5 消融实验

3.5.1 不同分支对实验结果的影响

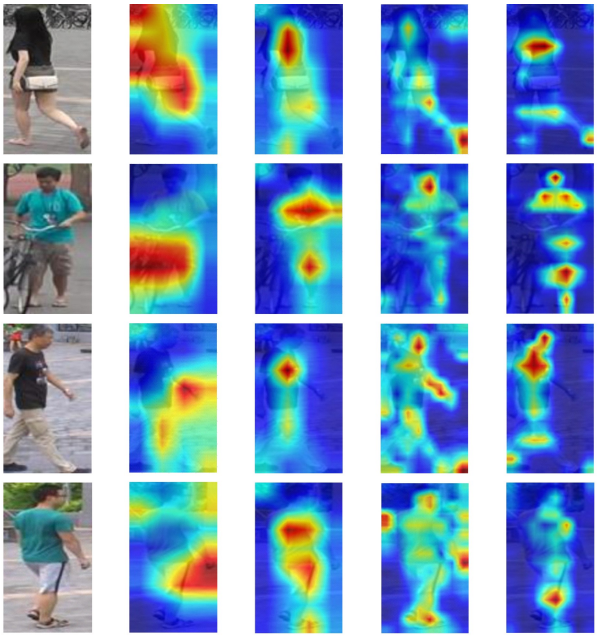
表2展示了不同分支对实验结果的影响，结果显示DSAB在HAT中加入正交正则化操作，同时引入通道(空间)注意力模块RGA-C(RGA-S)，在mAP和Rank-1上提升0.7%和0.6%。仅仅使用FPB，mAP和Rank-1能达到82.0%和91.2%。与基准网络HAT相比，CTM网络的识别率大大提升。

3.5.2 分块策略对实验结果的影响

表3展示了分块策略对实验结果的影响，结果显示把特征图分为3等份，而后将等分后的特征级联合成进行单一的损失计算，这时的网络性能最好，其关键指标mAP和Rank-1在DukeMTMC-reID数据集上达到82.3%和91.6%。

3.5.3 正交正则化对实验结果的影响

表4更加直观地展示了不同的正交正则化操作对实验结果的影响，可以看到在CTM模型中不使用SOR, COR操作，mAP和Rank-1指标较基线分别提升0.6%, 0.8%，同时在DukeMTMC-reID数据集上仅使用COR/SOR, mAP和Rank-1分别提升0.1%, 0.2%/0.3%, 0.3%。在SOR和COR相互作用下，CTM网络能够达到最佳的性能。



(a) 原图　(b) Baseline　(c) Global　(d) DSAB　(e) FPB

图 3 Market1501数据集可视化结果

3.5.4 注意力模块对实验结果的影响

表5展示了注意力模块对实验结果的影响，结果表明仅使用RGA-C(RGA-S)在mAP和Rank-1上提升0.7%, 0.8%(0.6%, 1.0%)，亦能够很好地提升网络性能。同时，注意力机制要放置在合适的位置上才能对整个模型产生积极的影响，本文在RseNet上第2, 3, 4层和FPB较浅层同时部署注意力模块，此时网络能够获得最佳的性能。

4 结束语

本文提出CTM，通过搭建全局分支、深度聚

表 2 不同分支对实验结果的影响(DukeMTMC-reID)(%)

Branch	mAP	Rank-1	Rank-5	Rank-10
HAT(Baseline)	81.4	90.4	95.6	97.1
+DSAB	82.1	91.0	95.7	96.8
+FPB	82.0	91.2	95.7	97.2
CTM	**82.3**	**91.6**	**95.9**	**97.5**

注：+表示网络中仅使用该分支

表 3 分块策略对比实验(DukeMTMC-reID)(%)

Part-Level	mAP	Rank-1	Rank-5	Rank-10
+2 part-level	81.9	90.8	95.3	96.8
+3 part-level	**82.3**	**91.6**	**95.9**	**97.5**
+4 part-level	**82.3**	91.2	95.4	97.0
+5 part-level	81.8	90.2	95.6	97.2

表 4 正交正则化对实验的影响(DukeMTMC-reID)(%)

Method	mAP	Rank-1	Rank-5	Rank-10
HAT(Baseline)	81.4	90.4	95.6	97.1
–SOR, COR	82.0	91.2	95.8	96.9
+COR	82.1	91.4	95.6	97.2
+SOR	**82.3**	91.5	95.6	97.1
CTM	**82.3**	**91.6**	**95.9**	**97.5**

(注：+表示网络中仅使用该操作，-表示均不使用操作)

表 5 注意力模块对实验结果的影响(DukeMTMC-reID)(%)

Attention	mAP	Rank-1	Rank-5	Rank-10
HAT(Baseline)	81.4	90.4	95.6	97.1
+RGA-C	82.1	91.2	95.8	96.8
+RGA-S	82.0	91.4	95.6	97.4
+Attention on backone	81.9	91.4	95.4	96.9
+Attention on FPB	82.2	91.3	**95.9**	97.2
CTM	**82.3**	**91.6**	**95.9**	**97.5**

合分支、特征金字塔3分支网络，最小化ID损失和3元组损失函数，完成表征提取。通过实验和特征可视化表明，本文的3分支模型中全局分支提取行人图像特征并生成层次特征作为另两个分支的输入，深度聚合分支能够通过从全局视图探索信息和促进局部信息来合并多尺度特征，特征金字塔分支能够整合不同尺度上的不同特征同时有效地降低特征深层特征映射相关性。后续工作将研究如何简化网络模型的复杂度，同时寻求更高识别率的方法。

参 考 文 献

[1] 邹国锋, 傅桂霞, 高明亮, 等. 行人重识别中度量学习方法研究进展[J]. 控制与决策, 2021, 36(7): 1547–1557. doi: 10.13195/j.kzyjc.2020.0801.
ZOU Guofeng, FU Guixia, GAO Mingliang, *et al*. A survey on metric learning in person re-identification[J]. *Control and Decision*, 2021, 36(7): 1547–1557. doi: 10.13195/j.kzyjc.2020.0801.

[2] 贲晛烨, 徐森, 王科俊. 行人步态的特征表达及识别综述[J]. 模式识别与人工智能, 2012, 25(1): 71–81. doi: 10.16451/j.cnki.issn1003-6059.2012.01.010.
BEN Xianye, XU Sen, and WANG Kejun. Review on pedestrian gait feature expression and recognition[J]. *Pattern Recognition and Artificial Intelligence*, 2012, 25(1): 71–81. doi: 10.16451/j.cnki.issn1003-6059.2012.01.010.

[3] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778. doi: 10.1109/CVPR.2016.90.

[4] SZEGEDY C, LIU Wei, JIA Yangqing, *et al*. Going deeper with convolutions[C]. The 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 1–9. doi: 10.1109/CVPR.2015.7298594.

[5] ZHOU Kaiyang, YANG Yongxin, CAVALLARO A, *et al*. Omni-scale feature learning for person re-identification[C]. The 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 2019: 3701–3711. doi: 10.1109/ICCV.2019.00380.

[6] WIECZOREK M, RYCHALSKA B, and DABROWSKI J. On the unreasonable effectiveness of centroids in image retrieval[C]. The 28th International Conference on Neural Information Processing, Sanur, Indonesia, 2021: 212–223. doi: 10.1007/978-3-030-92273-3_18.

[7] 匡澄, 陈莹. 基于多粒度特征融合网络的行人重识别[J]. 电子学报, 2021, 49(8): 1541–1550. doi: 10.12263/DZXB.20200974.
KUANG Cheng and CHEN Ying. Multi-granularity feature fusion network for person re-identification[J]. *Acta Electronica Sinica*, 2021, 49(8): 1541–1550. doi: 10.12263/DZXB.20200974.

[8] CHEN Binghui, DENG Weihong, and HU Jiani. Mixed high-order attention network for person re-identification[C]. 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 2019: 371–381. doi: 10.1109/ICCV.2019.00046.

[9] CHEN Xuesong, FU Canmiao, ZHAO Yong, *et al*. Salience-guided cascaded suppression network for person re-identification[C]. The 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 3297–3307. doi: 10.1109/CVPR42600.2020.00336.

[10] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 6000–6010.

[11] HAN Kai, WANG Yunhe, CHEN Hanting, *et al*. A survey on visual transformer[EB/OL]. https://doi.org/10.48550/arXiv.2012.12556, 2012.

[12] 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.

[13] HE Shuting, LUO Hao, WANG Pichao, *et al*. TransReID: Transformer-based object re-identification[C]. 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 14993–15002. doi: 10.1109/ICCV48922.2021.01474.

[14] PENG Zhiliang, HUANG Wei, GU Shanzhi, *et al*. Conformer: Local features coupling global representations for visual recognition[C]. 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 357–366. doi: 10.1109/ICCV48922.2021.00042.

[15] WANG Wenhai, XIE Enze, LI Xiang, *et al*. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]. 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021. doi: 10.1109/ICCV48922.2021.00061.

[16] ZHANG Guowen, ZHANG Pingping, QI Jinqing, *et al*. HAT: Hierarchical aggregation transformers for person re-identification[C]. The 29th ACM International Conference on Multimedia, Chengdu, China, 2021: 516–525. doi: 10.1145/3474085.3475202.

[17] ZHANG Suofei, YIN Zirui, WU X, *et al*. FPB: Feature pyramid branch for person re-identification[EB/OL]. https://doi.org/10.48550/arXiv.2108.01901, 2021.

[18] ZHENG Liang, SHEN Liyue, TIAN Lu, *et al*. Scalable person re-identification: A benchmark[C]. The 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 1116–1124. doi: 10.1109/ICCV.2015.133.

[19] RISTANI E, SOLERA F, ZOU R, *et al*. Performance measures and a data set for multi-target, multi-camera tracking[C]. The European Conference on Computer Vision, Amsterdam, The Netherlands, 2016: 17–35. doi: 10.1007/978-3-319-48881-3_2.

[20] WEI Longhui, ZHANG Shiliang, GAO Wen, *et al*. Person transfer GAN to bridge domain gap for person re-identification[C]. The 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 79–88. doi: 10.1109/CVPR.2018.00016.

[21] ZHANG Zhizheng, LAN Cuiling, ZENG Wenjun, *et al*. Relation-aware global attention for person re-identification[C]. The 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 3183–3192. doi: 10.1109/CVPR42600.2020.00325.

[22] CHEN Tianlong, DING Shaojin, XIE Jingyi, *et al*. ABD-net: Attentive but diverse person re-identification[C]. The 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 2019: 8350–8360. doi: 10.1109/ICCV.2019.00844.

[23] HERMANS A, BEYER L, and LEIBE B. In defense of the triplet loss for person re-identification[EB/OL]. https://doi.org/10.48550/arXiv.1809.05864, 2017.

[24] SZEGEDY C, VANHOUCKE V, IOFFE S, *et al*. Rethinking the inception architecture for computer vision[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2818–2826. doi: 10.1109/CVPR.2016.308.

[25] BRYAN B, GONG Yuan, ZHANG Yizhe, *et al*. Second-order non-local attention networks for person re-identification[C]. The 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 2019: 3759–3768. doi: 10.1109/ICCV.2019.00386.

[26] WANG Guan'an, YANG Shuo, LIU Huanyu, *et al*. High-order information matters: Learning relation and topology for occluded person re-identification[C]. The 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 6448–6457. doi: 10.1109/CVPR42600.2020.00648.

[27] JIN Xin, LAN Cuiling, ZENG Wenjun, *et al*. Style normalization and restitution for generalizable person re-identification[C]. The 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 3140–3149. doi: 10.1109/CVPR42600.2020.00321.

[28] YU Fufu, JIANG Xinyang, GONG Yifei, *et al*. Devil's in the details: Aligning visual clues for conditional embedding in person re-identification[EB/OL]. https://doi.org/10.48550/arXiv.2009.05250, 2020. doi: 10.48550/arXiv.2009.05250,2020.

[29] ZHU Kuan, GUO Haiyun, LIU Zhiwei, *et al*. Identity-guided human semantic parsing for person re-identification[C]. The 16th European Conference on Computer Vision, Glasgow, UK, 2020: 346–363. doi: 10.1007/978-3-030-58580-8_21.

[30] LI Hanjun, WU Gaojie, and ZHENG Weishi. Combined depth space based architecture search for person re-identification[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 6725–6734. doi: 10.1109/CVPR46437.2021.00666.

[31] CHEN Jiaxing, JIANG Xinyang, WANG Fudong, *et al*. Learning 3D shape feature for texture-insensitive person re-identification[C]. The 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 8142–8151. doi: 10.1109/CVPR46437.2021.00805.

[32] LI Yulin, HE Jianfeng, ZHANG Tianzhu, *et al*. Diverse part discovery: Occluded person re-identification with part-aware transformer[C]. The 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 2897–2906. doi: 10.1109/CVPR46437.2021.00292.

[33] ZHOU Kaiyang and XIANG Tao. Torchreid: A library for deep learning person re-identification in pytorch[EB/OL]. https://doi.org/10.48550/arXiv.1910.10093, 2019.

陈　莹：女，教授，博士生导师，研究方向为图像处理、信息融合、模式识别.

匡　澄：男，硕士，研究方向为行人重识别.

责任编辑：马秀强