用于车辆重识别的部件耦合 Transformer 网络

孙伟[1,2)], 胡亚华[1)], 代广昭[1)], 张小瑞[2,3,4)], 徐凡[1)], 赵宇煌[1)]

[1)] (南京信息工程大学自动化学院 南京 210044)
[2)] (南京信息工程大学江苏省大气环境与装备技术协同创新中心 南京 210044)
[3)] (南京信息工程大学电子与信息工程学院 南京 210044)
[4)] (南京信息工程大学无锡研究院 无锡 214100)
(sunw0125@163.com)

摘 要: 基于卷积神经网络的车辆重识别模型在执行卷积和池化操作时, 不可避免地会出现全局感受野狭小和局部信息丢失的情况, 当光照、视角和分辨率等发生剧烈变化时, 导致车辆重识别的鲁棒性和精确性急剧下降. 为此, 提出了部件耦合 Transformer 的车辆重识别网络, 通过堆叠部件耦合 Transformer 块来搭建重识别模型, 每一个部件耦合 Transformer 块利用部件自适应嵌入模块提取区分性的局部特征和 Transformer 层提取鲁棒性的全局特征. 首先, 部件自适应嵌入模块按照位置和伸缩量动态划分和调整特征图, 增强模型对局部部件信息的感知能力; 其次, Transformer 层中利用自注意力机制增强网络模型对全局特征的表示能力; 最后, 部件自适应嵌入模块和 Transformer 层之间的耦合关系促进全局和局部特征协同合作. 在 VeRi-776 和 VehicleID 数据集上的实验结果表明, CMC@1/CMC@5 分别达到 0.970/0.988 和 0.865/0.985, 优于对比模型.

关键词: 车辆重识别; Transformer; 部件自适应嵌入
中图法分类号: TP391 **DOI:** 10.3724/SP.J.1089.2023.19641

Part Coupled Transformer Network for Vehicle Re-Identification

Sun Wei[1,2)], Hu Yahua[1)], Dai Guangzhao[1)], Zhang Xiaorui[2,3,4)], Xu Fan[1)], and Zhao Yuhuang[1)]

[1)] (*School of Automation, Nanjing University of Information Science & Technology, Nanjing* 210044)
[2)] (*Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science & Technology, Nanjing* 210044)
[3)] (*School of Electronics & Information Engineering, Nanjing University of Information Science & Technology, Nanjing* 210044)
[4)] (*Wuxi Research Institute, Nanjing University of Information Science & Technology, Wuxi* 214100)

Abstract: When vehicle re-identification models based on convolutional neural network perform convolutional and pooling operation, it is inevitable that the global sensitivity field will be narrow and local information will be lost. When illumination, perspective and resolution change dramatically, the robustness and accuracy of vehicle re-identification will decline sharply. Therefore. the part coupled Transformer (PCT) network for vehicle re-identification is proposed, which stacks multiple PCT blocks to build a re-identification model. Each PCT block uses a part adaptation embedding (PAE) module to extract discriminative local features and employs a Transformer layer to extract robust global features. Firstly, the PAE that splits and adjusts dynamically feature map according to positions and scales, enhances the ability to

收稿日期: 2022-03-24; 修回日期: 2022-06-06. **基金项目**: 国家自然科学基金(62376128, 62272236); 江苏省自然科学基金(BK20201136, BK20191401); 江苏省研究生科研与实践创新计划(SJCX21_0363). **孙伟**(1980—), 男, 博士, 教授, 硕士生导师, 主要研究方向为车辆重识别、车辆分类、机器学习; **胡亚华**(1996—), 男, 硕士研究生, 主要研究方向为计算机视觉、车辆重识别; **代广昭**(1995—), 男, 硕士研究生, 主要研究方向为模式识别、车辆重识别; **张小瑞**(1980—), 女, 博士, 教授, 硕士生导师, 主要研究方向为虚拟现实、图像处理; **徐凡**(1996—), 男, 硕士研究生, 主要研究方向为车辆重识别; **赵宇煌**(1999—), 男, 硕士研究生, 主要研究方向为车辆重识别.

capture local parts. Secondly, the self-attention mechanism of Transformer layers improves the representation of global features. Finally, the coupling between PAE module and Transformer layer results in an effective cooperation between global and local features. Experimental results on VeRi-776 datasets and VehicleID datasets show that CMC@1/CMC@5 of PCT reaches 0.970/0.988 and 0.865/0.985 respectively, which is superior to the comparison models

Key words: vehicle re-identification; Transformer; part adaptation embedding

给定一张待查询的车辆图片, 车辆重识别旨在从多个摄像机采集的大规模图库集中找到所有相同的车辆图片. 近年来, 车辆重识别(vehicle re-identification, Vehicle Re-ID)在智能交通[1-2]和智慧城市[3-4](如车辆轨迹分析、车辆交通流分析和智能高速等方面)发挥越来越重要的作用, 受到学者的广泛关注.

然而, 真实交通场景下采集的车辆图片不可避免地会遭受光照、视角和分辨率的影响, 这给 Vehicle Re-ID 带来了极大的挑战. 相同车辆在不同视角下外观差异巨大, 仅从整体外观上难以对其进行区分; 相同车型和颜色的车辆拥有相似的外观, 如果不从局部细节上辨别, 则极易将其误识别为同一车辆. 因此, 如何提取和利用鲁棒性的全局特征和区分性的局部特征成为提高 Vehicle Re-ID 性能的关键.

近年来, 一些大规模车辆数据集的出现, 促进了基于深度卷积神经网络(deep convolutional neural networks, DCNNs)的 Vehicle Re-ID 的发展, 显著地提高了 Vehicle Re-ID 的精度和可靠性. 在这些网络模型中, 其核心是通过堆叠卷积和池化层来学习获得车辆的特征, 包括全局特征和局部特征等. Liu 等[2]通过 VGG(visual geometry group)[5]提取车辆图片特征, 将整张车辆图片输入到欧几里得空间来计算车辆之间的距离, 使用车辆 ID 和车辆型号对神经网络的训练进行监督, 以提升模型对同一车型的不同车辆进行重识别的鲁棒性. 与全局特征相比, 局部特征能够被用于提供稳定而有效的线索. Wang 等[6]对车身表面 20 个关键点区域标注, 并分为前、后、左、右 4 个视角, 通过对关键点的检测生成视角不变特征; He 等[7]利用预训练的 YOLO 检测网络[8]提取重要的局部区域(如车窗、车灯和车牌等), 以增强神经网络对局部细微差异的感知. 然而, 这些基于 DCNNs(如 VGG[5], ResNet[9]等)的 Vehicle Re-ID 网络模型每次在执行卷积和池化操作时, 其输出只能映射到车辆图片的一小部分像素点, 使得网络模型每一层对车辆图片的感受野范围十分狭小, 导致局部细节信息丢失. 特别是在遇到遮挡和模糊等情况时, 狭小的感受野会使得全局特征无法完整地表征车辆图片的宏观信息, 而局部信息的丢失也将导致局部特征缺乏区分性.

针对 DCNNs 本质上的局限性, TransReID[10]引入 Transformer[11]作为主干网络提取车辆图片的全局特征. 基于 Transformer 的网络模型通过引入自注意力模块, 并去除卷积和池化操作, 可有效地解决基于 DCNNs 模型的本质缺陷、增大模型的全局感受野, 显著提升模型全局的宏观特征表征能力. 然而, 在 TransReID 中运用到的视角和摄像机身份信息均属于额外的辅助信息, 其收集和标注工作需要耗费大量的时间和人力成本. 此外, Transformer 片段划分中固定大小和位置的设置, 使得车辆的部件(如车灯、车标等)不可避免地被分割到不同片段中, 破坏了车辆局部部件完整的语义信息.

此外, 如何有效地利用全局特征和局部特征同样是重识别性能提升的关键. 三分支网络 TBE-Net[12]中全局分支和局部分支分别用于提取局部特征和全局特征, 并将全局特征损失函数和局部特征损失函数加权求和作为总损失对网络进行约束. 文献[13]中通过图网络在每个阶段输出的特征图上学习局部相关性, 并将局部相关性补充到全局特征. 然而, 这种关系是单方面的, 局部和全局之间缺乏有效的协同合作.

不同于上述网络模型, 本文提出了部件耦合 Transformer 网络, 以提取鲁棒性和区分性的车辆特征. 该网络通过堆叠部件耦合 Transformer (part coupled Transformer, PCT)块构建重识别模型, PCT 块由部件自适应模块和 Transformer 层 2 部分组成. 每一个 PCT 块通过 Transformer 层中自注意力机制计算图片每一个像素点之间的相似性, 提取车辆鲁棒性的全局特征. 受可变形卷积的启发, PCT 块

中的部件自适应嵌入模块在特征图上对划分的片段进行偏移、伸缩, 并通过特征图坐标的更新来表示这种变化, 既能保留车辆部件完整的语义信息, 又能增加局部特征的区分性. 此外, 通过每个 PCT 块中部件自适应嵌入模块和 Transformer 层之间构建耦合的关系, 使得全局特征和局部特征能够有效的协同工作, 与上述 TBE-Net, TransReID 等网络模型相比, 本文模型能够更好地提取和利用全局和局部特征.

基于 VeRi-776[14], VehicleID[2]这 2 个大规模数据集, 对本文所提出的 Vehicle Re-ID 网络模型 PCT 进行测试, 以验证其鲁棒性和精确性.

1　相关工作

1.1　基于 DCNNs 的 Vehicle Re-ID

近年来, 由于 Vehicle Re-ID 在智能交通和智慧城市领域上的广泛应用, 引起了越来越多的学者的注意. 目前, 基于 DCNNs 的 Vehicle Re-ID 网络模型可以分为以下 2 类.

(1) 基于全局特征. 车辆拥有较为丰富的全局外观信息, 如车型和颜色等. 不同车型和颜色的车辆差异明显, 仅从全局外观便可识别. Li 等[15]提出了三分支网络来分别学习车辆的车型、颜色和视角信息, 并将其作为输入通过交叉熵损失函数进行优化; Huynh 等[16]利用 ResNeXt_ibn_a101[17]提取特征, 并通过训练技巧提升网络的性能. 这类网络模型利用全局外观特征进行重识别, 但是在视角、光照等外界条件发生变化时, 车辆外观会发生改变. 因此, 全局外观特征的鲁棒性是提升重识别性能的关键之一.

(2) 基于局部特征. 在面对全局外观相似的车辆时, 如车型和颜色相同, 仅利用全局外观特征无法正确地识别相同车辆. 有些学者开始将车辆局部区域的差异作为区分性的线索, 如年检标志、车标、车灯等. Meng 等[18]通过 3D 视角对齐网络将车辆图片转换为前后 2 个固定视角, 并从这 2 个视角选取 9 个局部感兴趣区域作为比较线索. 然而, 这类网络模型需要其他神经网络对车辆图片进行预处理, 时间和人工成本耗费较大.

1.2　基于 Transformer 的 Vehicle Re-ID

鉴于 Transformer 在自然语言处理的优异表现, 部分学者[19-20]开始调研其在计算机视觉领域的应用. 随着 ViT[11]的提出, 基于 Transformer 的 Vehicle Re-ID 研究成果陆续发表. TranReID[10]是第 1 个使用纯 Transformer 进行重识别的工作, 将 Transformer 作为主干网络来提取车辆的全局特征. 虽然 He 等[10]利用视角和摄像机信息进一步提升了模型的表现, 但是其依旧缺少对局部特征的利用. Li 等[21]利用基于 Transformer 的框架, 从多视角的成对图片中学习车辆表征. 该工作使用了卷积层提取结构化的片段表征和 Transformer 层学习不同视角片段之间的相关性, 是 Transformer 层和卷积层联合使用的一种尝试.

2　本文网络

Vehicle Re-ID 旨在从由不重叠的摄像机捕获的图库集中找到所有与待查询车辆相同的图片. 一个合理的 Vehicle Re-ID 网络模型应该同时考虑到鲁棒性的全局特征和区分性的局部特征. 本文针对 Vehicle Re-ID 提出了一个 PCT 网络, 如图 1 所示.

2.1　PAE 模块

PAE 模块被用于建模车辆部件完整的语义信息, 如车灯、车窗和车标等, 从而提取图片的局部特征. PAE 模块通过对特征图进行片段坐标表示、中心点偏移和边长伸缩来实现. PAE 的模块结构如图 2 所示, 假设 $F \in \mathbb{R}^{C\times H\times W}$ 表示一个特征图, 其中, C 表示通道数, H 和 W 分别表示特征图的长和宽. 与 ViT[11]中的片段划分类似, 将特征图划分成大小一致的矩形图 $O_i \in \mathbb{R}^{h\times w}$, 其中 $h\times w$ 表示矩形图的大小, i 表示第 i 个矩形图并且 $i \in \{1,2,\cdots, HW/hw\}$. 为了使得每个矩形图能包含完整的部件语义信息, 本文将原本位置固定的矩形图进行偏移和伸缩, 并通过坐标来表示这种转化. 将初始矩形图 O_i 的中心点设置为 $\left(x_{\text{cent}}^{(i)}, y_{\text{cent}}^{(i)}\right)$, 初始矩形图的左下角和右上角则分别表示为 $\left(x_{\text{cent}}^{(i)} - h^{(i)}/2,\ y_{\text{cent}}^{(i)} - w^{(i)}/2\right)$ 和 $\left(x_{\text{cent}}^{(i)} + h^{(i)}/2, y_{\text{cent}}^{(i)} + w^{(i)}/2\right)$. 通过中心点、左下角和右上角的坐标完成了矩形图坐标表示的初始化.

在矩形图初始化坐标的基础上, 本文引入中心点偏移量 φ 和边长伸缩量 τ, 来描述形变后的矩形图的坐标. 形变后第 i 个矩形图左下角和右上角的坐标 $\left(x_{\text{lb}}^{(i)}, y_{\text{lb}}^{(i)}\right)$, $\left(x_{\text{rt}}^{(i)}, y_{\text{rt}}^{(i)}\right)$ 分别表示为

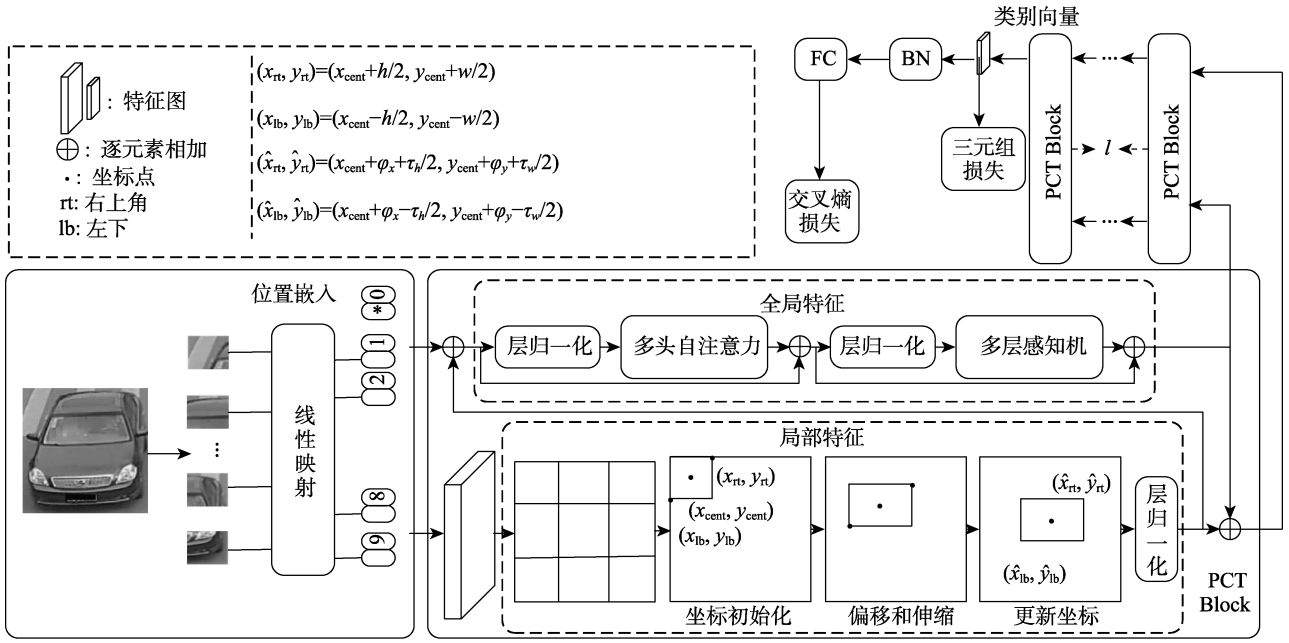


图 1 PCT 网络结构

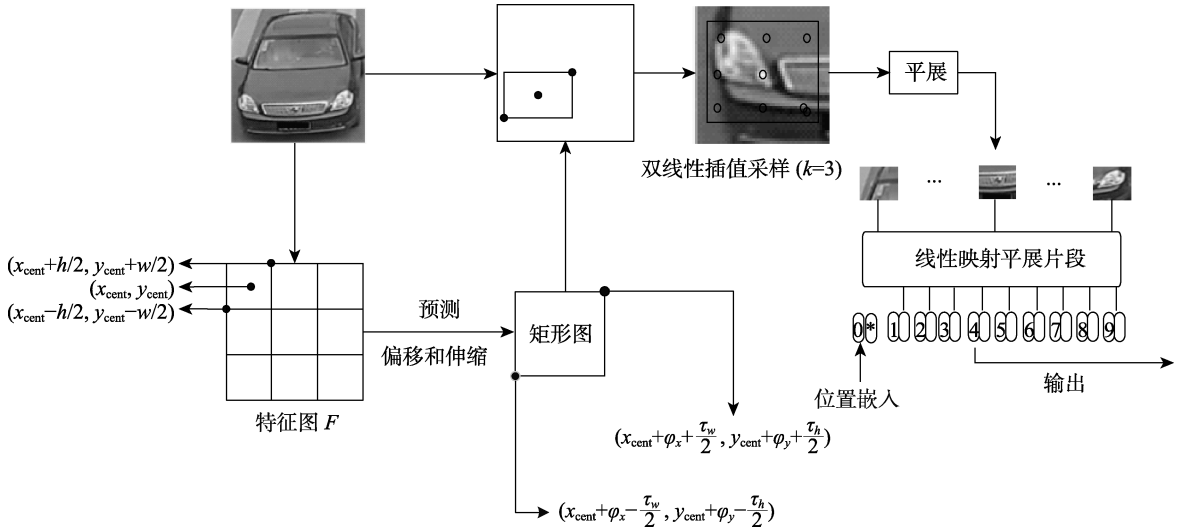


图 2 PAE 模块结构

$$\begin{cases} x_{\mathrm{lb}}^{(i)} = x_{\mathrm{cent}}^{(i)} + \varphi_x^{(i)} - \dfrac{\tau_h^{(i)}}{2} \\ y_{\mathrm{lb}}^{(i)} = y_{\mathrm{cent}}^{(i)} + \varphi_y^{(i)} - \dfrac{\tau_w^{(i)}}{2} \\ x_{\mathrm{rt}}^{(i)} = x_{\mathrm{cent}}^{(i)} + \varphi_x^{(i)} + \dfrac{\tau_h^{(i)}}{2} \\ y_{\mathrm{rt}}^{(i)} = y_{\mathrm{cent}}^{(i)} + \varphi_y^{(i)} + \dfrac{\tau_w^{(i)}}{2} \end{cases} \tag{1}$$

其中，(φ_x, φ_y) 和 (τ_h, τ_w) 分别表示矩形图水平、垂直方向上中心点的偏移量和边长的伸缩量. 上标 (i) 表示第 i 个矩形图.

从式(1)可以看出，矩形图的形变由参数 $\varphi_x, \varphi_y, \tau_h, \tau_w$ 决定，通过对参数前向和后向传播进行更新来预测矩形图. 偏移量和伸缩量计算公式分别为

$$\varphi_x, \varphi_y = \mathrm{Mish}\left(\omega_{\mathrm{shift}} \cdot g(F)\right) \tag{2}$$

$$\tau_x, \tau_y = \mathrm{LeakyReLU}\left(\mathrm{Mish}\left(\omega_{\mathrm{scale}} \cdot g(F) + b_{\mathrm{scale}}\right)\right) \tag{3}$$

其中，$F \in \mathbb{R}^{C\times H\times W}$ 表示输入的特征图；$g(\cdot)$ 表示一线性特征提取器；Mish 和 LeakyReLU 分别表示不同的激活函数；ω_{shift} 和 ω_{scale} 分别表示偏移权重和伸缩权重.

值得注意的是，形变后矩形图只是一个范围，

并不能体现图片特征, 网络需要在该范围内提取局部特征. 由于形变后的矩形图大小不一致且坐标可能是浮点型, 因此本文采用采样插值的方法进行特征提取. 在每个矩形图内部统一选取 $m\times m$ 个采样点, m 为超参数, 本文设置 m=3. 每一个采样点的坐标被标记为 $V^{(j)}=\left(v_x^{(j)},v_y^{(j)}\right)$, $j\in\{1,\cdots,k\times k\}$. 每个采样点的特征 $\left\{f^{(j)}\right\}_{1\leqslant j\leqslant k\times k}$ 通过平展和线性映射等处理转化成嵌入序列, 即

$$\boldsymbol{Z}^{(i)}=W[f^{(1)},f^{(2)},\cdots,f^{(m\times m)}]+b \tag{4}$$

对于采样点 $\left(v_x,v_y\right)$, 本文通过双线性插值采样获得该点的相关特征, 即

$$\begin{aligned}F\left(v_x,v_y\right)=&F\left(D_{11}\right)\left(x_2-v_x\right)\left(y_2-v_y\right)+\\&F\left(D_{21}\right)\left(v_x-x_1\right)\left(y_2-v_y\right)+\\&F\left(D_{12}\right)\left(x_2-v_x\right)\left(v_y-y_1\right)+\\&F\left(D_{22}\right)\left(v_x-x_1\right)\left(v_y-y_1\right)\end{aligned} \tag{5}$$

其中, $\left(v_x,v_y\right)$ 表示采样点坐标; $D_{11},D_{12},D_{21},D_{22}$ 表示靠近 $\left(v_x,v_y\right)$ 的 4 个坐标点, 即

$$\begin{cases}D_{11}=\left(x_1,y_1\right),D_{12}=\left(x_1,y_2\right)\\D_{21}=\left(x_2,y_1\right),D_{22}=\left(x_2,y_2\right)\end{cases} \tag{6}$$

2.2 Transformer 层

Transformer 层被用于建模不同片段间的全局特征, 其由一个多头自注意力(multi-head self-attention, MHSA), 一个多层前向感知机(multi-layer perceptron, MLP)和 2 个层归一化(layer normalization, LN)组成. 输入车辆图片被划分为若干片段, 通过映射和平展将其转化为向量序列. 假设 $\boldsymbol{X}\in\mathbb{R}^{N\times M}$ 是 N 个片段的序列, 其中 $M=P^2\times C$. 每一张图片的大小和通道数分别为 $P\times P$ 和 C. 为了保留片段的空间信息, 本文采用可学习的一维位置嵌入与片段嵌入直接相加, 使得同一数据在不同位置出现差异化. 输入的序列可以表示为

$$\boldsymbol{Z}_0=\left[x_{\text{cls}};F\left(x^{(1)}\right);F\left(x^{(2)}\right);\cdots;F\left(x^{(N)}\right)\right]+P \tag{7}$$

其中, $P\in\mathbb{R}^{(N+1)\times D}$ 表示位置嵌入; x_{cls} 表示全局特征表征; F 表示将片段映射到 D 维空间的线性函数. 输入的序列 $\boldsymbol{Z}$ 通过线性映射转换为查询向量 $\boldsymbol{Q}$, 键值向量 $\boldsymbol{K}$ 和数值向量 $\boldsymbol{V}$. $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 通过放缩点积注意力计算相似度, 公式为

$$\text{Att}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_v}}\right)\boldsymbol{V} \tag{8}$$

MHSA 将 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 分离成 h 份, 并且平行运行 h 次自注意力函数后, 将每部分自注意力函数的输出进行拼接和映射, 形成最终的输出. MLP 进行前向传播, 并调整输出的特征维度; 残差结构用于防止梯度爆炸. 因此, Transformer 层的最终输出 $Y\in\mathbb{R}^{N\times M}$ 的计算过程为

$$Y_l=\text{MLP}\left(\text{LN}\left(X_l'\right)\right)+X_l' \tag{9}$$

$$X_l'=\text{MHSA}\left(\text{LN}\left(X_{l-1}\right)\right)+X_{l-1} \tag{10}$$

其中, LN 表示层归一化.

2.3 PCT

由图 1 可知, PCT 网络模型有 l 个 PCT 块. 每一个 PCT 块包含一个 PAE 模块和一个 Transformer 层. 值得注意的是, 本文提出的 PCT 能够在 Transformer 层提取的全局特征和 PAE 模块提取的局部特征之间构建一种耦合的关系. 即在整个训练过程中, 全局特征和局部特征之间可以双向信息交互. 第 l 个 PCT 块中的 PAE 模块 $P^{(l)}$ 和 Transformer 层 $T^{(l)}$ 的输入/输出定义分别为

$$P^{(l)}=\text{PAE}\left(P^{(l-1)},T^{(l-1)}\right) \tag{11}$$

$$T^{(l)}=\text{Trans}\left(T^{(l-1)},P^{(l)}\right) \tag{12}$$

由式(10)(11)可以看出, 在第 l 个 PCT 块的 PAE 模块, 从第 $(l-1)$ 个 PCT 块中 PAE 模块和 Transformer 层提取出局部和全局特征, 在此基础上学习局部特征. 类似地, Transformer 层的输入由第 l 层 PAE 模块的输出和第 $l-1$ 层 Transformer 的输出构成. 因此, 网络在 PAE 模块和 Transformer 层之间构建了耦合的关系, 使得全局特征和局部特征能够有效协同合作.

除此之外, PCT 网络模型有 3 个阶段, 每个阶段有 4 个 PCT 块来处理特征图; 每个 PCT 块具有相同结构. 每张图片和片段的大小与 ViT[11]中一致, 分别设置为 224×224 和 16×16.

2.4 损失函数

大多数工作将 ReID 视为多任务学习, 主流工作是将其分为 2 个任务: 一个是分类任务, 将同一车辆的不同图片当成一个类别, 通过分类损失进行优化训练, 如交叉熵损失函数; 另一个是图像检索问题, 通过三元组损失约束拉近相同车辆图片的距离和增大不同车辆图片间的距离. 为了使预测车辆 ID 靠近真实 ID, 并且拉近正样本距离和增大负样本距离、提高 Vehicle Re-ID 的准确率, 本文采用交叉熵损失和三元组损失来设计 PCT 的损失函数. 由图 1 可知, 与 ViT 一致, 将最后一个 PCT

块输出的类别向量作为车辆图片的特征表示. 受 BNNeck[16]在 DCNNs 中优异表现的启发, 本文在类别向量后引入 BNNeck. 类别向量表示的特征通过三元组损失函数[21]进行约束, 使得相似特征之间的距离越来越小; 不同特征之间的距离越来越大. 通过交叉熵损失函数优化类别向量经过批量归一化处理后表示的特征. 因此, 本文提出的 PCT 网络模型总的损失为

$$L_{\text{total}} = \alpha L_{\text{CE}} + \beta L_{\text{Tri}} \tag{13}$$

其中, L_{CE} 和 L_{Tri} 分别表示交叉熵损失函数和三元组损失函数; α 和 β 表示人工设置常量, 用于保持 2 种损失之间的平衡.

交叉熵损失函数计算公式为

$$L_{\text{CE}}(X,y) = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{K}(y_i,j)\ln\left(p_{ij}\right) \tag{14}$$

其中, X, y 分别表示训练集和类别标签信息; M 和 K 分别表示训练样本数和类别数量; x_i 表示第 i 个训练样本; $j \in \{1,2,\cdots,K\}$ 表示车辆图片的类别标签. 如果车辆图片的样本标签为 j, 那么 y_i=1; 否则, y_i=0. p_{ij} 表示输入图片 x_i 的类别标签为 j 的预测概率, 其由 Softmax 函数计算而得, 即

$$p_{ij} = \frac{\exp(\boldsymbol{W}_j^{\mathrm{T}} x_i)}{\sum_{k=1}^{K}\exp(\boldsymbol{W}_k^{\mathrm{T}} x_i)} \tag{15}$$

其中, $\boldsymbol{W} = [\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3, \cdots, \boldsymbol{W}_K]$ 表示可学习的参数矩阵. 此外, 三元组损失函数可以表示为

$$L_{\text{Tri}}\left(x^a, x^n, x^p\right) = -\frac{1}{M}\sum_{i=1}^{M}\max\left(\left\|x_i^a - x_i^n\right\|_2 - \left\|x_i^a - x_i^p\right\|_2, 0\right) \tag{16}$$

其中, $\left(x^a, x^n, x^p\right)$ 表示训练三元组的集合; M 表示三元组的数量. 对于第 i 个训练三元组, $\left(x^a, x^n\right)$ 表示一负样本对, 即类别标签不同; $\left(x^a, x^p\right)$ 表示一正样本对, 即拥有相同的样本标签; $\|\cdot\|$ 表示欧几里得距离. 此外, 在每个小批量训练中, 硬样本挖掘策略[22]被用于寻找最困难的正样本和负样本图像对, 以改善三元组损失.

3 实验与分析

本节将介绍实验中使用的多个数据集和实施的细节, 通过消融实验以及与先进网络模型的对比来证明本文提出的 PCT 的鲁棒性和高准确性.

3.1 数据集和评估指标

3.1.1 数据集

本文在 2 个公开的基准数据集, 即 VeRi-776 数据集和 VehicleID 数据集进行, 以评估 Vehicle Re-ID 网络模型.

VeRi-776[14]是一个大规模车辆数据集, 由大约 776 种车辆的 50000 张图片构成. 车辆图片由 20 个摄像机从不同视角捕获而得, 并且被标注了多种属性, 如颜色、车型和车标. 其训练集包含了 576 种车辆, 测试集包含了其他 200 种车辆.

VehicleID[2]也是一个经典的车辆数据集, 由大约 26267 种车辆的 221736 张图片构成. 车辆图片由 12 个摄像头从前或后 2 个视角采集而得. 其训练集包含了 13134 种车辆, 训练集进一步细分为 6 个子集, 其中包含 800, 1600, 2400 种车辆的 3 个子集通常被用于评估实验.

上述 2 个基准数据集中车辆的车型和数量十分丰富. 既有全局外观差异大的相同车辆, 又有局部差异细微的不同车辆, 能够较好地模拟真实场景的情况.

3.1.2 评估指标

基于上述2个数据集, 本文采用2个评价标准, 即平均精度均值(mean average precision, mAP)和累计匹配曲线(cumulative matching characteristic, CMC)来评测 Vehicle Re-ID 模型的表现. CMC@k 表示在前 k 个查询结果中找到正确结果的概率. 使用 mAP 和 CMC 在 VeRi-776 数据集上进行评估. 对于 VehicleID, 在随机的不同规模的测试集合中, CMC 度量取得了更加令人信服的结果.

3.2 实验细节

本文在 PyTorch[23]框架下进行所有实验, 并取得实验结果. 在训练时, 所有的输入图片大小均被调整为 224×224, 并通过数据增强, 如随机翻转、填充和随机擦除[24]等, 增加样本的多样性. 设置学习率的初始值为 5×10^{-4}, 并在第 60 和 120 周期分别下降至 5×10^{-5} 和 5×10^{-6}. 在训练时, 批量大小和周期总数分别为 64 和 150. 本文采用批量梯度下降算法[25]来优化网络模型, 设置动量参数为 0.9, 权重衰减为 10^{-4}. 在推理阶段, 批量大小更改为 128, 其他设置不变.

3.3 消融实验

为了验证本文网络模型各个模块的有效性, 本节在 VeRi-776 数据集上执行了多组不同设置的消融实验. 消融实验的评价指标分别 mAP, CMC@1 和 CMC@5. 消融实验包含 Transformer

骨干网络的有效性、PCT 中各模块的有效性、PAE 中不同插值采样方法的影响以及损失函数中平衡参数的影响.

3.3.1　Transformer 骨干网络的有效性

为了证明 Transformer 编码器作为主干网络的有效性, 本文比较了基于 DCNNs 和基于 Transformer 主干网络在 Vehicle Re-ID 上的表现, 实验结果如表 1 所示. 其中, DenseNet, ResNet50, ResNet101 和 ResNet152 属于基于 DCNNs 系列的不同类别、不同深度主干网络; Baseline 则使用 Transformer 作为主干网络. 主干网络的对比实验去除了本文提出的各模块, 仅利用各主干网络学习车辆特征.

表 1　不同主干网络比较

主干网络	mAP	CMC@1	CMC@5
DenseNet[26]	0.752	0.950	0.958
ResNet50[9]	0.763	0.952	0.960
ResNet101[9]	0.767	0.952	0.961
ResNet152[9]	0.772	0.959	0.965
Baseline	0.783	0.965	0.972

从表 1 可以看出, 基于 Transformer 主干网络的性能表现要远好于基于 CNNs 不同类别、不同深度的主干网络, 其原因是去除了卷积和池化操作而完全使用自注意力提取特征. 实验结果证明, Transformer 中自注意力机制增大的全局感受野和保留细微的局部特征, 能够解决 DCNNs 中卷积和池化操作的局限性.

3.3.2　PCT 中各模块的有效性

为了验证网络模型 PCT 中各模块的有效性, 本文执行了多组不同设置的消融实验, 实验结果如表 2 所示. 其中, Baseline 表示 PCT 去除了所有其他模块, 仅利用 Transformer 学习车辆的全局特征; Baseline+PAE 表示在学习全局特征的基础上线性增加对局部部件特征的学习, 即将全局损失和局部损失加权求和对网络进行约束; PCT 表示在学习全局和局部特征的基础上构建一种耦合的合作关系.

表 2　PCT 各模块作用比较

方法	mAP	CMC@1	CMC@5
Baseline	0.783	0.965	0.972
Baseline+PAE	0.788	0.966	0.982
PCT	0.811	0.970	0.988

由表 2 可知, PCT 在所有设置的实验中取得最佳的重识别表现, 这是因为本文网络模型不仅能够学习鲁棒性的全局特征和区分性的局部部件特征, 还能使两者有效协同合作. 此外, PCT 和 Baseline+PAE 分别与 Baseline 进行对比, PCT 重识别精度提升的幅度更大, 其在 mAP, CMC@1 和 CMC@分别增加了 2.8%, 0.5%和 1.6%. 这说明全局和局部特征有效协同合作有助于提高重识别的表现能力.

为了进一步证明网络模型的有效性, 本文将 VeRi-776 数据集上的部分实现结果可视化, 如图 3 所示. 图 3 中最左边的一列表示待查询图片; 右边 10 列表示 Baseline 和 PCT 按相似度得分排序的前 10 个检索结果. 红色框的车辆图片表示正确的实验结果, 即与待查询车辆身份相同的图片; 蓝色框的车辆图片则表示错误的实验结果, 即与待查询车辆身份不一致的图片. 由图 3 观察可知, 本文所提出的 PCT 网络模型在前 10 个检索结果中正确的概率远大于 Baseline, 使得 Vehicle Re-ID 的鲁棒性和准确性得到了大幅提升.

3.3.3　PAE 中不同插值采样方法的影响

本文对几种比较常见的采样插值方法进行了对比实验, 结果如表 3 所示. 其中, 双线性插值采样和双三次插值方法的精度均优于最近邻方法. 虽然在采样后的效果方面双线性插值和双三次插

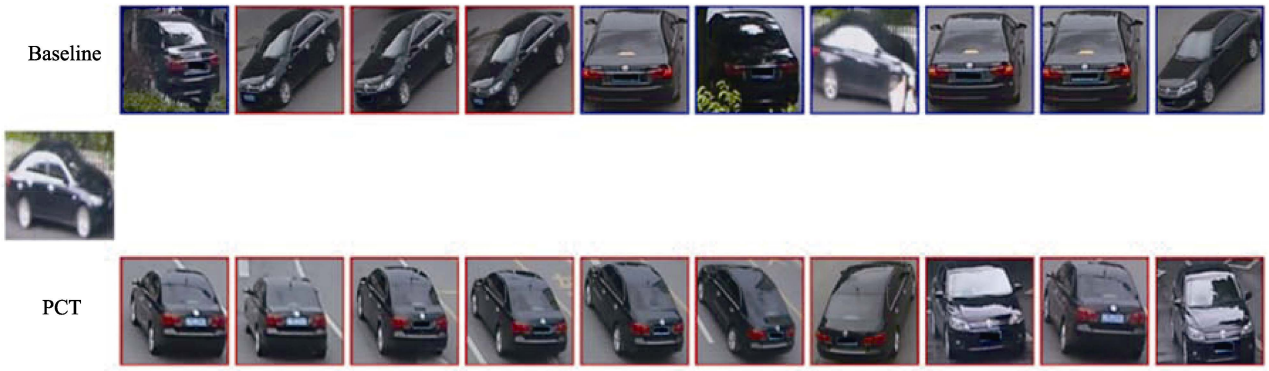


图 3　TOP10 可视化结果

表 3 PAE 不同插值方法比较

方法	mAP	CMC@1	CMC@5
双线性差值	0.778	0.957	0.974
双三次差值	0.811	0.970	0.988
最近邻	0.812	0.969	0.989

值的差距不大, 但是在计算量方面, 使用 16 个周围坐标点的双三次插值的计算量要远大于双线性插值. 故本文选取双线性插值作为采样方法, 以提高模型的计算效率.

3.3.4 损失函数中平衡参数的影响

本节对损失函数中平衡参数的选取进行了对比实验, 结果如表 4 所示. 其中, 设置 α=0.0, 表示只是用三元组损失; 设置 β=0.0, 表示只是用了交叉熵损失.

表 4 平衡参数 α 和 β 对实验结果的影响

α	β	mAP	CMC@1	CMC@5
0.0	1.0	0.726	0.949	0.973
0.2	1.0	0.755	0.953	0.979
0.4	1.0	0.771	0.962	0.981
0.6	1.0	0.785	0.967	0.985
0.8	1.0	0.796	0.969	0.987
1.0	0.0	0.712	0.947	0.970
1.0	0.2	0.753	0.951	0.974
1.0	0.4	0.778	0.959	0.977
1.0	0.6	0.787	0.966	0.982
1.0	0.8	0.792	0.968	0.986
1.0	1.0	0.811	0.970	0.988

由表4可以看出, 当只是用三元组损失或交叉熵损失时, 重识别模型性能急剧下降. 这说明重识别模型的高性能表现得益于三元组损失和交叉熵损失的联合约束, 又或者说这2种损失函数均有助于模型的训练, 两者的存在是必要的. 此外, 当 α=1.0 并以间隔 0.2 多次更改 β 的值时, 模型的性能随着 β 的增大逐步提升. 类似地, 当 β=1.0 时, 以相同设置更改 α, 模型性能依旧可以逐步提升. 当 α=β=1.0 时, 模型的性能达到最佳. 因此, 在其他对比和消融实验中本文选取 α=β=1.0.

3.4 与其他网络模型比较

将本文提出的 PCT 在 2 个车辆公开数据集 VeRi-776 和 VehicleID 上与其他 Vehicle Re-ID 网络模型进行对比, 比较结果分别如表 5 和表 6 所示.

3.4.1 在 VeRi-776 数据集上的表现

不同网络模型在 VeRi-776 数据集上性能比较如表 5 所示. 在 VeRi-776 数据集上, 首先, 与其他的重识别网络模型相比, 本文的 PCT 网络模型重识别精度均远超其他 Vehicle Re-ID 网络模型, PCT 取得了最高的 Map, CMC@1, CMC@5. 这得益于本文的 PCT 不仅拥有映射全局的感受野, 还能够保留详细的局部特征. 此外, 部件自适应嵌入模块和 Transformer 层之间的耦合关系可以学习全局特征和局部特征之间的相互关系, 使得全局特征和局部特征能够有效的协同合作.

表 5 不同网络模型在 VeRi-776 数据集上性能比较

网络模型	mAP	CMC@1	CMC@5
OIFE[6]	0.480	0.894	
VAMI[27]	0.613	0.859	0.918
RAM[28]	0.615	0.886	0.940
AAVER[29]	0.612	0.890	0.947
PRN[7]	0.743	0.943	0.989
PVEN[30]	0.795	0.965	0.984
TBE-Net[12]	0.795	0.960	0.985
PCT	0.811	0.970	0.988

3.4.2 在 VehicleID 数据集上的表现

为证明所提出模型的泛化性, 本文也将 PCT 在 VehicleID 上进行了测试. 表 6 所示为 VehicleID 数据集上的网络模型与其他基于 DCNNs 的 Vehicle Re-ID 网络模型的 CMC@1 和 CMC@5 比较结果. 由表 6 观察可知, 本文网络模型在 VehicleID 800, 1 600, 2 400 大小的 3 个测试子集上均取得了最优的结果. 特别是对于 VehicleID 中规

表 6 不同网络模型在 VehicleID 数据集上性能比较

网络模型	800		1 600		2 400	
	CMC@1	CMC @5	CMC @1	CMC @5	CMC @1	CMC @5
OIFE[6]					0.670	0.829
VAMI[27]	0.631	0.833	0.529	0.751	0.473	0.703
RAM[28]	0.752	0.915	0.723	0.870	0.677	0.845
AAVER[29]	0.747	0.938	0.686	0.900	0.635	0.856
PRN[7]	0.784	0.923	0.750	0.883	0.742	0.864
PVEN[30]	0.847	0.970	0.806	0.945	0.778	0.920
TBE-Net[12]	0.860	0.984	0.823	0.966	0.807	0.949
PCT	0.865	0.985	0.826	0.970	0.825	0.963

模最大的子集, 由于其不仅包含了更多车辆图片, 而且光照、视角和分辨率等变化更加剧烈, 因而其比小和中 2 个测试子集时更难对车辆进行重识别. 尽管如此, 在此数据集的 2400 测试子集上, 本文的网络模型的 CMC@1 值比现有最好的 TBE-Net 网络模型高出了 1.8%, CMC@5 值高出了 1.4%. 因此, 相对于 PVEN, TBE-Net 等基于 DCNNs 的网络模型, 本文的网络模型能够更好地提取鲁棒性和鉴别性的特征, 从而提高重识别的精度.

4 结　语

本文提出了一种 PCT 网络模型对车辆进行重识别, 其构建基于 Transformer 主干网络的 Vehicle Re-ID 模型, 鲁棒地实现了对宏观的全局特征的学习. 提出 PAE 模块, 通过自适应划分矩形图, 使得模型能够提取到车辆区分性的局部部件特征. 提出的耦合机制将 PAE 模块和 Transformer 层进行耦合, 能够学习到全局和局部特征之间的互补表达关系, 使得全局和局部特征能够更有效地协同工作. 在 2 个大型标准的数据集 VeRi-776 和 VehicleID 上进行的拓展实验, 证明了本文网络模型 PCT 的鲁棒性和精确性.

受 cross-attention 在跨模态[31-32]和跨域[33]任务中优异表现的启发, 未来考虑利用 cross-attention 的特征对齐能力, 将目标域和源域的数据映射到统一的特征空间进行学习, 以减少 Vehicle Re-ID 模型在新环境部署过程中数据标注所花费的人力和时间成本, 提高基于 Transformer 的 Vehicle Re-ID 模型的跨域自适应能力, 促进 Vehicle Re-ID 在实际交通场景的广泛应用.

参考文献(References):

[1] Guo H Y, Zhao C Y, Liu Z W, *et al*. Learning coarse-to-fine structured feature embedding for vehicle re-identification[C] //Proceedings of the 32nd the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 6853-6860

[2] Liu H Y, Tian Y H, Yang Y W, *et al*. Deep relative distance learning: tell the difference between similar vehicles[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2167-2175

[3] Lou Y H, Bai Y, Liu J, *et al*. VERI-wild: a large dataset and a new method for vehicle re-identification in the wild[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 3230-3238

[4] Hu Yu, Chen Xiaobo, Liang Jun, *et al*. Vehicle re-identification method based on part features and multi-attention fusion[J]. Journal of Computer Research and Development, 2022, 59(11): 2497-2506(in Chinese)
(胡煜, 陈小波, 梁军, 等. 基于组件特征与多注意力融合的车辆重识别方法[J]. 计算机研究与发展, 2022, 59(11): 2497-2506)

[5] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. [2022-03-24]. https://arxiv.org/abs/1409.1556

[6] Wang Z D, Tang L M, Liu X H, *et al*. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 379-387

[7] He B, Li J, Zhao Y F, *et al*. Part-regularized near-duplicate vehicle re-identification[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 3992-4000

[8] Redmon J, Divvala S, Girshick R, *et al*. You only look once: unified, real-time object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 779-788

[9] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778

[10] He S, Luo H, Wang P C, *et al*. TransReID: transformer-based object re-identification[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 14993-15002

[11] Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: transformers for image recognition at scale[OL]. [2022-03-24]. https://arxiv.org/abs/2010.11929

[12] Sun W, Dai G Z, Zhang X R, *et al*. TBE-net: a three-branch embedding network with part-aware ability and feature complementary learning for vehicle re-identification[OL]. [2022-03-24]. https://doi.org/10.1109/TITS.2021.3130403

[13] Zhao J J, Zhao Y F, Li J, *et al*. Heterogeneous relational complement for vehicle re-identification[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 205-214

[14] Liu X C, Liu W, Mei T, *et al*. A deep learning-based approach to progressive vehicle re-identification for urban surveillance[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2016: 869-884

[15] Li H C, Lin X M, Zheng A H, *et al*. Attributes guided feature learning for vehicle re-identification[OL]. [2022-03-24]. https://doi.org/10.1109/TETCI.2021.3127906

[16] Huynh S V, Nguyen N H, Nguyen N T, *et al*. A strong baseline for vehicle re-identification[C] //Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2021: 4142-4149

[17] Pan X G, Luo P, Shi J P, *et al*. Two at once: enhancing learning and generalization capacities via IBN-net[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 484-500

[18] Meng D C, Li L, Liu X J, *et al*. Viewpoint alignment and discriminative parts enhancement in 3D space for vehicle ReID[OL]. [2022-03-24]. https://doi.org/10.1109/TMM.2022.3154102

[19] Han K, Wang Y H, Chen H T, *et al*. A survey on visual transformer[OL]. [2022-03-24]. https://doi.org/10.1109/TPAMI.2022.3152247

[20] Khan S, Naseer M, Hayat M, *et al*. Transformers in vision: a survey[J]. ACM Computing Surveys, 2022, 54(10s): Article No.200

[21] Li M, Liu J, Zheng C, *et al*. Exploiting multi-view part-wise correlation via an efficient transformer for vehicle re-identification[OL]. [2022-03-24]. https://doi.org/10.1109/TMM.2021.3134839

[22] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[OL]. [2022-03-24]. https://arxiv.org/abs/1703.07737

[23] Paszke A, Gross S, Massa F, *et al*. PyTorch: an imperative style, high-performance deep learning library[J]. Advances in Neural Information Processing Systems, 2019, 32: 8026-8037

[24] Zhong Z, Zheng L, Kang G L, *et al*. Random erasing data augmentation[C] //Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 13001-13008

[25] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012: 1097-1105

[26] Huang G, Liu Z, van der Maaten L, *et al*. Densely connected convolutional networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2261-2269

[27] Zhou Y, Shao L. Viewpoint-aware attentive multi-view inference for vehicle re-identification[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 6489-6498

[28] Khorramshahi P, Kumar A, Peri N, *et al*. A dual-path model with adaptive attention for vehicle re-identification[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 6131-6140

[29] Lou Y H, Bai Y, Liu J, *et al*. Embedding adversarial learning for vehicle re-identification[J]. IEEE Transactions on Image Processing, 2019, 28(8): 3794-3807

[30] Meng D C, Li L, Liu X J, *et al*. Parsing-based view-aware embedding network for vehicle re-identification[C] //Processing of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 7101-7110

[31] Tsai Y H H, Bai S J, Liang P P, *et al*. Multimodal transformer for unaligned multimodal language sequences[C] //Proceedings of the Conference Association for Computational Linguistics Meeting. Stroudsburg: ACL Press, 2019: 6558-6569

[32] Li X Y, Hou Y H, Wang P C, *et al*. Trear: transformer-based RGB-D egocentric action recognition[J]. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14(1): 246-252

[33] Xu T K, Chen W H, Wang P C, *et al*. CDTrans: cross-domain transformer for unsupervised domain adaptation[OL]. [2022-03-24]. https://arxiv.org/abs/2109.06165