基于自适应局部划分Vision Transformer的车辆重识别方法*

徐萌兮[1] 陈海鑫[2] 徐焕宇[3]

（1.无锡学院交通与车辆工程学院，无锡 214105；2.南京信息工程大学计算机学院，南京 210044；3.无锡学院物联网工程学院，无锡 214105）

【摘要】针对复杂场景下因摄像头视角变化、遮挡及车辆外观相似性导致模型的特征表达不充分和鲁棒性不足等问题，提出一种基于自适应局部划分Vision Transformer的车辆重识别方法。结合卷积和注意力设计自适应局部划分（ALP）模块；通过层次化注意力融合（HAF）模块整合低层视觉细节与高层全局语义，为自适应区域划分提供特征引导；引入多重特征嵌入（MFE）模块，基于相机和视角的动态加权机制提升多视角、多相机环境下的特征区分能力。试验结果表明：在车辆重识别任务中，所提出方法在VeRi-776数据集上的mAP和Rank-1分别达到81.0%、97.1%，在VehicleID数据集上的Rank-1达到80.2%，显著提升了模型的识别精度和鲁棒性。

主题词：车辆重识别 Vision Transformer 局部划分 注意力融合

中图分类号：TP391.41 **文献标志码：**A **DOI:** 10.19620/j.cnki.1000-3703.20250449

Vehicle Re-Identification Method Based on Adaptive Local Partitioning Vision Transformer

Xu Mengxi[1], Chen Haixin[2], Xu Huanyu[3]

（1. School of Transportation and Vehicle Engineering, Wuxi University, Wuxi 214105; 2. School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044; 3. School of Internet of Things Engineering, Wuxi University, Wuxi 214105）

【Abstract】To address insufficient feature representation and poor robustness caused by viewpoint changes, occlusions, and visual similarity, this paper proposes an adaptive local partitioning Vision Transformer for vehicle re-identification. Firstly, an Adaptive Local Partitioning (ALP) module is designed to perform dynamic regional division. Secondly, by integrating low-level visual details with high-level global semantics through a Hierarchical Attention Fusion (HAF) module, it provides feature guidance for adaptive region segmentation. Finally, a Multi-Feature Embedding (MFE) module is introduced to enhance discriminability using a dynamic weighting mechanism for camera and viewpoint information. The experimental results show that in the vehicle re-identification task, the proposed method achieved an mAP of 81.0% and a Rank-1 of 97.1% on the VeRi-776 dataset, and a Rank-1 of 80.2% on the VehicleID dataset, significantly improving the model's identification accuracy and robustness.

Key words: Vehicle re-identification, Vision Transformer, Local partition, Attention fusion

【引用格式】徐萌兮，陈海鑫，徐焕宇. 基于自适应局部划分Vision Transformer的车辆重识别方法[J]. 汽车技术，2026(4): 17-25.

XU M X, CHEN H X, XU H Y. Vehicle Re-Identification Method Based on Adaptive Local Partitioning Vision Transformer[J]. Automobile Technology, 2026(4): 17-25.

1 前言

在智慧城市建设中，车辆重识别（Vehicle Re-Identification，Vehicle ReID）作为智能交通与公共安全监控的核心技术，旨在跨摄像头视角精准匹配同一车辆。然而，实际交通场景中的车辆重识别极具挑战性：一方面，同款车型具有高度的类间相似性，要求模型具备极强的细粒度特征捕捉能力；另一方面，光照变化、

*基金项目：江苏省高校哲学社会科学研究课题（2022SJYB0987，2022SJYB0986）；陕西省自然科学基础研究计划（2025JC-YBMS-713）。

低分辨率及视角变化等因素导致巨大的类内差异，要求算法具备极高的鲁棒性。因此，在复杂环境下构建判别性强、泛化能力稳健的特征表示至关重要。

随着大规模车辆数据集（VeRi-776[1]、VehicleID[2]等）的推广，卷积神经网络（Convolutional Neural Network，CNN）在车辆重识别领域具有显著优势。在细粒度特征挖掘方面，He等[3]聚焦于外观高度相似的车辆，通过部件正则化机制增强了模型对车辆局部细节的感知能力，但在遮挡或目标结构不完整条件下极易发生特征错位与信息丢失。在应对视角多变与环境复杂性方面，常采用特征转换与空间对齐机制约束特征表示。Zhou等[4]提出视角感知的注意力多视角推断模型VAMI，利用注意力机制自动选取输入图像与目标视角间重叠的核心视觉区域，使模型能够在同一视角的特征空间优化距离度量。Meng等[5]通过空间视角对齐与判别性部件增强机制，使模型在对齐后的图像上精准提取并增强判别性部件的特征。

为了克服外部环境变化带来的特征混淆，Zhu等[6]联合建模车辆身份、视角与摄像头信息的分支网络框架，通过特征维度的解耦与多维信息的相互补充，缓解了由跨摄像头复杂背景干扰和同款车型形状相似性引发的误判问题，但数据训练依赖先验标注，难以有效捕获图像各区域间的长距离全局依赖关系。TransReID利用Vision Transformer（ViT）的全局特征提取能力[7-8]，引入结构信息增强融合视角与摄像头标签，有效缓解了传统CNN在处理大尺度视角变化时的局限性，但局部细粒度特征提取和多层次特征融合方面仍存在不足，特别是外观高度相似的车辆中包含决定性的个体差异信息。

针对ViT对局部细节感知不足的问题，现有研究主要从特征加权、空间划分和引入语义掩码等维度进行局部建模。Gong等[9]提出VRcvt模型，以类别标记CLS token（Classification token）查询向量，其他图像块标记（Patch token）对键向量和值向量进行加权融合，但该模型仅对最后一层进行特征融合，忽视了ViT多层结构的不同抽象层次信息（如边缘、纹理、轮廓），导致细粒度语义捕捉不充分；Qian等[10]构建了双分支的条带与属性感知网络SAN，通过结构感知的条带划分策略学习局部判别特征，提高模型的局部特征提取能力，但固定条带划分极易受拍摄角度变化影响，易致局部区域失配；Meng等[11]提出了一种基于解析的视角感知嵌入网络PVEN，通过语义解析掩码引导注意力聚焦高判别性区域，但显式掩码方法过度依赖外部车辆部件的解析精度，外部解析器在复杂遮挡、光照突变或低分辨率场景，严重影响后续重识别特征的判别力。

因此，本文提出一种基于ViT的自适应局部划分（Adaptive Local Partitioning-Vision Transformer，ALP-ViT）网络，通过增强局部区域的上下文关联，引入动态侧信息调节机制，提升复杂交通场景下模型的鲁棒性。模型主要包含3个关键模块：设计自适应局部划分模块，结合注意力引导与卷积操作，实现对车辆图像关键区域的动态划分；引入层次化注意力融合（Hierarchical Attention Fusion，HAF）模块，整合Transformer不同层级的注意力信息；采用多重特征嵌入（Multiple Feature Embedding，MFE）模块，对摄像头和视角信息进行解耦建模与动态加权融合，显著提升模型在多视角、多相机环境下的判别与泛化能力。

2 本文方法

2.1 基本架构

本文提出的ALP-ViT架构如图1所示，模型整体流程包括全局建模与局部细粒度特征增强模块。模型主干由多层Transformer编码器构成，每层包含多头自注意力（Multi-Head Self-Attention，MHSA）机制、前馈网络等，通过深层堆叠建模图像的语义关系，使用CLS token进行提取图像的全局表示。将图像划分为多个图像块（Patch），编码为一系列词元（Token）并输入ViT主干网络。

采用自适应局部划分模块，对图像特征图进行动态区域划分。该模块利用ViT中CLS token的注意力分布，引导对关键区域进行分割，并通过区域加权池化提取局部结构特征。网络最终输出由全局特征与经ALP提取的多个局部区域特征共同组成，后续进行特征融合以提升模型识别精度。在训练阶段引入局部灰度补丁替换[12]（Grayscale Patch Replacement，GPR）策略，以一定概率随机选取图像中的矩形区域，使用对应位置的灰度图像素模拟遮挡或信息缺失，从而提升模型对局部遮挡与风格变化的鲁棒性。

2.2 全局特征提取

在Transformer层，输入图像经过特征嵌入模块被映射为一系列Patch token，随后被送入多层Transformer编码器中进行处理。各层通过MHSA建立token间的全局关系，利用层归一化（Layer Normalization，LN）和残差连接保证特征稳定传播。在所有token中，第一个CLS token作为专门用于分类任务的代理token，承担全局信息的汇聚作用，即提供车辆整体轮廓、颜色等信息。

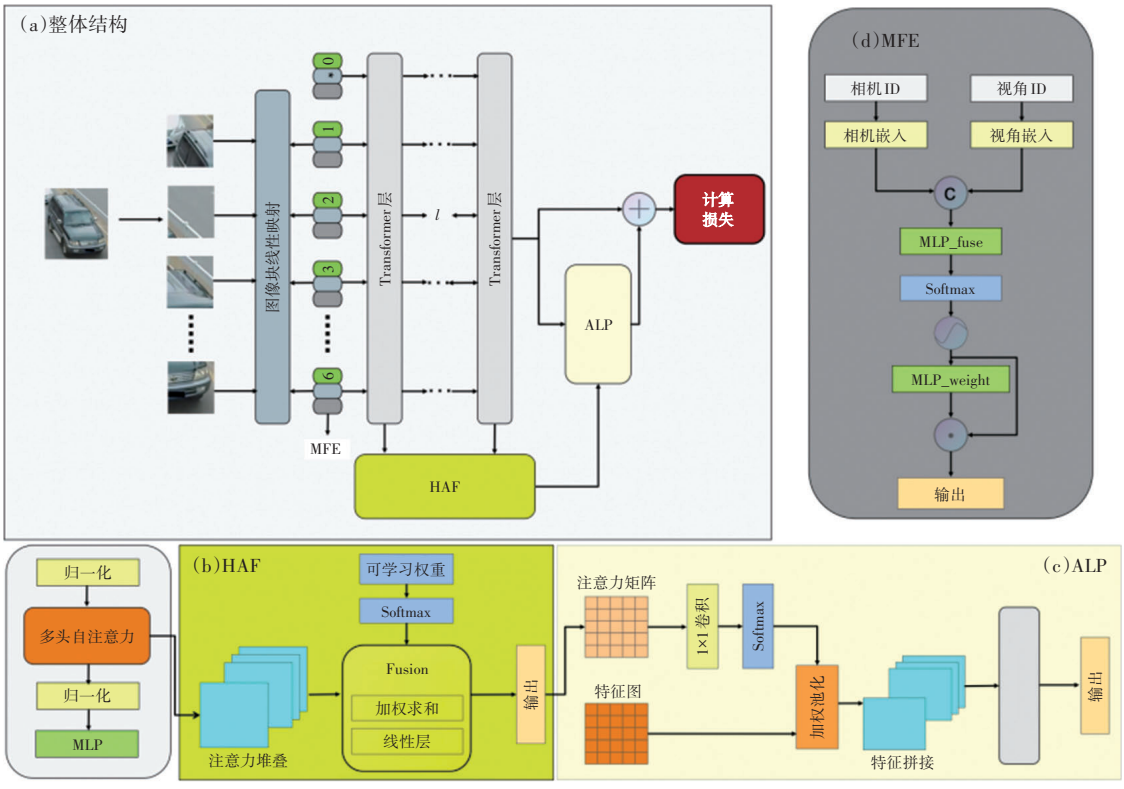


图1　ALP-ViT结构

其输出的特征为当前图像的全局表示，后续用于分类或与局部特征融合。

2.3　局部特征提取与增强

2.3.1　自适应局部划分模块

传统ViT架构通过全局池化获得图像的整体特征，极易丢失车窗、车牌等局部细节信息。为此，本文设计自适应局部划分模块，该模块结合注意力机制和卷积网络，动态关注图像的感兴趣区域，根据局部区域的重要性进行加权池化，从而提升模型对细节的捕捉能力。

ALP模块以HAF模块输入的注意力矩阵为引导信号，从而进行图像的区域划分。HAF模块融合Transformer不同层的注意力信息，得到注意力矩阵。利用矩阵中CLS token对patch token的注意力分布，指引局部区域划分，对显著性区域赋予较高的注意力权重。其中，重塑后二维空间维度的注意力矩阵为：

$$\boldsymbol{A}_{\text{cls}} = reshape(\boldsymbol{A}_{\text{fused}}),\ \boldsymbol{A}_{\text{cls}} \in \mathbf{R}^{B\times C\times H\times W} \tag{1}$$

式中：$\boldsymbol{A}_{\text{fused}}$为CLS token对patch token的注意力，H、W、$C=1$、B分别为特征图的高、宽、通道数和批大小，$reshape()$为重塑操作。

在软注意力掩膜生成阶段，输入为Transformer层提取的空间特征图和HAF的注意力矩阵。首先，从注意力矩阵中提取CLS token对剩余patch token的注意力向量。假设每张图片被划分为N个patch，则注意力向量维度为$[B,N]$。将其重塑为与patch排列相对应的二维注意力，其维度为$[B,1,H,W]$，其中$H\times W=N$。各位置的像素值表示CLS token对该局部区域的关注程度。随后，通过1个1×1卷积层将单通道的注意力图映射为K个通道，生成K个初步的区域掩膜$\boldsymbol{M}_{\text{mask}}$：

$$\boldsymbol{M}_{\text{mask}} = \text{Softmax}(Conv(\boldsymbol{A}_{\text{cls}})),\ \boldsymbol{M}_{\text{mask}} \in \mathbf{R}^{B\times K\times H\times W} \tag{2}$$

式中：K为划分区域的数量。

模型可依据注意力分布划分区域，每个输出通道对应一个潜在的划分区域。最后，为了实现图像局部的软划分，在通道维度上应用Softmax归一化处理。该操作将各空间位置（像素）对应K个区域的响应值映射为联合概率分布，确保同一位置上的K个掩膜权重值均位于区间(0,1)且总和为1。

在区域特征池化阶段，自适应局部划分模块对各区域分别提取特征。将i个区域掩膜$\boldsymbol{M}_{\text{mask},i}$作为权重矩阵，对输入特征图进行加权平均池化，得到该区域的特征表示$\boldsymbol{F}_i$，K个区域的特征向量堆叠即为图像的局部特征集合：

$$\boldsymbol{F}_i = \frac{\sum_{h=1}^{H}\sum_{w=1}^{W}\boldsymbol{M}_{\text{mask},i}(h,w)\boldsymbol{F}(h,w)}{\sum_{h=1}^{H}\sum_{w=1}^{W}\boldsymbol{M}_{\text{mask},i}(h,w)+\varepsilon} \tag{3}$$

式中：$\boldsymbol{F}(h,w)$为特征图在位置(h,w)处的特征向量，ε为极小值常数。

因此，自适应局部划分模块基于注意力的动态区域

划分和特征提取，使模型学习边界划分，针对不同车辆图像关注最具判别力的部位，如车头造型、尾灯形状、车牌位置等细节差异，解决了ViT模型对局部信息利用不足的问题。

2.3.2　跨层次注意力融合模块

原始ViT的各层注意力具有明显的层次特征：低层关注局部细节和边缘模式，高层则倾向于建模全局语义和高级表示。若模型仅使用最后一层注意力矩阵，容易忽略浅层的细粒度信息，难以全面识别图像的关键区域。受ViT-Adapter[13]中跨层特征融合思想的启发，HAF模块充分挖掘ViT不同层次的注意力信息，以提升模型对上下文语义的理解能力。

对每层注意力图进行标准化处理，使其分布一致，从而缓解梯度爆炸风险并增强训练稳定性。随后，为每层Transformer的注意力矩阵分配可训练的融合权重α，并通过Softmax函数进行归一化，保证各层间的贡献尺度一致。将所有层的注意力矩阵按权重进行加权求和，得到融合后的注意力图。最后，通过一组线性投影操作对融合结果进行多头整合，使生成的注意力表示$\boldsymbol{A}_{\text{fused}}$同时包含跨层次与多头注意力的结构信息，为ALP模块提供更加稳健的上下文语义引导：

$$\boldsymbol{A}_{\text{fused}} = proj\left(\sum_{l=1}^{L}\text{Softmax}(\boldsymbol{w}_l)\boldsymbol{A}_l\right) \tag{4}$$

式中：$\boldsymbol{A}_l$为第l层注意力，$\boldsymbol{w}_l$为学习权重，$proj()$为线性投影操作。

2.4　多重特征嵌入

MFE模块通过对相机和视角信息解耦，将摄像头嵌入$\boldsymbol{s}_1$和视角嵌入$\boldsymbol{s}_2$视为2类独立且互补的领域，分别学习后再进行拼接，构成联合表示$(\boldsymbol{s}_1,\boldsymbol{s}_2)$。通过轻量注意力网络为不同嵌入分配通道级别的动态权重，从而实现细粒度的融合，避免信息冗余或语义冲突：

$$\begin{cases}\boldsymbol{\alpha} = \text{Softmax}(f(\boldsymbol{s}_1,\boldsymbol{s}_2)) \\ \boldsymbol{e}_{\text{fused}} = \sum_{i=1}^{2}\boldsymbol{\alpha}_i \odot \boldsymbol{S}_i\end{cases} \tag{5}$$

式中：$f()$为带有非线性变换的可学习网络，$\boldsymbol{\alpha}$为网络输出的通道级权重，$\boldsymbol{S}_i$为摄像头嵌入和视角嵌入，$\boldsymbol{e}_{\text{fused}}$为融合后的辅助特征。

为了提升辅助信息的表征能力，对融合后的嵌入$\boldsymbol{e}_{\text{fused}}$应用可学习的动态缩放机制，使其对主干特征的影响程度可控，最终输出的动态缩放特征为：

$$\boldsymbol{e}_{\text{MFE}} = (1 + 4\boldsymbol{w})\boldsymbol{e}_{\text{fused}} \tag{6}$$

式中：$\boldsymbol{w}$为$\boldsymbol{e}_{\text{fused}}$通过多层感知机学习的动态权重。

模型根据输入图像自适应调整辅助信息的贡献，基于试验得到：当权重过小时，辅助信息的影响会被削弱，导致模型难以利用；当权重过大时，辅助信息对原始图像补丁信息产生干扰。因此，$(1 + 4\boldsymbol{w})$能够在保证辅助信息有效贡献的同时，避免过度干扰，有效提升模型的整体表现。

2.5　损失函数

本文损失函数使用交叉熵损失函数和三元损失函数。其中，交叉熵损失函数用于监督模型对车辆身份的准确分类[14]，其定义为：

$$L_{\text{ID}}(\theta; x) = -\frac{1}{P \times M}\sum_{i=1}^{P \times M}\log(p(y_i|x_i)) \tag{7}$$

式中：P为类别数量，M为每类样本数量，$p(y_i|x_i)$为模型预测输入样本x_i属于其真实类别y_i的概率。

三元组损失函数[15]通过减少同一车辆图像特征间的距离，同时增加不同车辆图像特征间的距离，进一步优化特征空间的结构。本文采用软间隔版本的三元组损失，其定义为：

$$L_{\text{tri}}(\theta; x) = -\frac{1}{P \times M} - \frac{1}{P \times M}\sum_{i=1}^{P}\sum_{a=1}^{M} \log\left(1 + \frac{\exp\left(\max\limits_{p=1,2,\cdots,K} D(f_\theta(x_i^a), f_\theta(x_i^p))\right)}{\exp\left(\min\limits_{j \neq i,\ p=1,2,\cdots,K} D(f_\theta(x_i^a), f_\theta(x_j^p))\right)}\right) \tag{8}$$

式中：$D()$为两个特征向量间欧氏距离的平方；$f_\theta(x_i)$为样本x_i通过特征提取网络后的表征向量；x_i^a为当前选定的基准图，即第i个车辆的第a张图；x_i^p为与基准图片属于同一辆车（下标同为i）的其他图片，即第i个车辆的第p张图；x_j^p为与基准图片属于不同车辆（下标为j）的图，即第j个车辆的第p张图。

因此，整体损失函数由交叉熵损失函数和三元损失函数相加构成：

$$L_{\text{total}}(\theta; x) = L_{\text{ID}}(\theta; x) + L_{\text{tri}}(\theta; x) \tag{9}$$

3　试验与分析

3.1　数据集与评价指标

3.1.1　数据集

选择车辆重识别领域大规模基准数据集VeRi-776和VehicleID进行试验。其中，VeRi-776数据集涵盖多视角变化（前、后、侧面）、复杂光照条件、遮挡干扰及相似车型的外观混淆数据，同时提供车辆身份（Identification，ID）、摄像头ID、品牌和颜色等标注信息。该数据集共计49 357张图像，随机选择37 778张作为训练图像（576个ID）、11 579张作为测试图像（200个ID）。

VehicleID数据集由清华大学发布，所有图像均来

自真实监控场景，具有多视角、不同光照和复杂背景等特点，主要用于大规模车辆识别和跨摄像头检索。该数据集共计221 763张图像，涵盖26 267个车辆ID。为了全面评估模型在不同规模VehicleID数据集的检索性能，从测试集中提取包含800个、1 600个和2 400个车辆ID的3个子集进行测试评估，依次命名为测试子集1~3。

3.1.2 评价指标

为了全面评估模型在车辆重识别任务的性能，将平均精度均值（mean Average Precision，mAP）和累积匹配特征（Cumulative Match Characteristic，CMC）曲线作为评价指标：

$$\begin{cases} A_{\mathrm{AP}} = \dfrac{\sum_{k=1}^{n} P(k) \times rel(k)}{N_{\mathrm{gt}}} \\ A_{\mathrm{mAP}} = \dfrac{1}{Q} \sum_{q=1}^{Q} A_{\mathrm{AP},q} \end{cases} \tag{10}$$

式中：n为单次查询检索出的图库样本总数，$P(k)$为检索结果列表中前k个样本的准确率，$rel(k)$为指示函数，N_{gt}为图库中与该查询图像真实匹配的目标样本总数，Q为测试集中包含的查询图像的总数，$A_{\mathrm{AP},q}$为第q张查询图像的平均精度。

其中，mAP衡量模型对所有查询图像在整个检索结果中的排序质量，反映了模型的整体检索能力。CMC曲线衡量不同横坐标Rank-K下，模型成功检索目标的准确率，反映了模型在实际检索系统中的效果。本文使用Rank-1和Rank-5的识别准确率，分别代表查询图像的正匹配样本出现在检索序列前1位和前5位的概率。

3.2 试验环境

本文试验在PyTorch框架下运行，图像处理器为英伟达RTX 4090。所有输入图像分辨率为256像素×256像素，使用数据增强提升模型的泛化能力，包括随机水平翻转（概率为0.5）、图像边缘填充（10像素填充）和随机擦除（擦除概率为0.5）等。

模型采用基于ImageNet预训练的ViT-B/16主干网络进行迁移学习。在训练阶段，初始学习率设置为8×10^{-3}，采用线性预热策略。总训练周期设置为120轮，批大小为64。优化器选用带动量的随机梯度下降（Stochastic Gradient Descent，SGD）算法，动量系数为0.9，权重衰减系数为1×10^{-4}，并对偏置项进行正则化约束。在推理阶段，批大小为256，其余参数与训练阶段一致。

3.3 消融试验

为了分析本文模型中各模块的有效性，使用VeRi-776数据集进行消融试验。以ViT提取车辆的全局特征作为基线（Baseline）方法，各模块的性能表现如表1所示。

表1 消融试验结果 %

模块	mAP	Rank-1	Rank-5
Baseline	78.2	96.5	97.2
Baseline+ALP+HAF(10)	79.6	96.8	98.4
Baseline+GPR	78.4	96.3	97.2
Baseline+MFE	79.6	96.9	98.4
ALP-ViT	81.0	97.1	98.8

由表1可知，本文模型在车辆重识别任务中具有有效性。由于基线方法仅依赖ViT提取全局特征，虽然具备一定的表达能力，但对局部细节的刻画能力有限，在复杂场景下的表现并不理想。ALP和HAF模块通过引入以层次融合注意力引导的自适应局部划分机制，能够根据Transformer注意力对图像区域进行动态划分，从而提取更加精细的局部特征，使mAP和CMC曲线指标显著提升；GPR策略提高了模型在遮挡、颜色变化下的鲁棒性和泛化能力；MFE模块将视角与摄像头信息融入训练过程，增强了模型在跨视角、跨摄像头场景下的泛化能力。

因此，ALP-ViT通过对全局与局部、多尺度与跨层信息的联合建模，模型的mAP达到81.0%，Rank-1与Rank-5分别达到97.1%和98.8%，实现了更丰富的特征表达能力。

为了直观展示模块效果，分析Baseline和Baseline+ALP+HAF的关注区域，结果如图2所示。同时，比较ALP-ViT与Baseline方法，对前10个（TOP-10）检索结果进行可视化分析，结果如图3所示。

由图2和图3可知，Baseline的注意力较为分散，而Baseline+ALP+HAF聚焦于车辆的关键区域，如车灯、格栅等。这得益于ALP模块对局部特征的定位和HAF模块的注意力引导，模型能够将注意力更精确地集中在车辆最具判别力的部位，有效抑制了背景区域的干扰。在ALP-ViT各模块的协同作用下，模型在复杂环境中具备更强的车辆重识别能力，对于提升模型整体性能具有显著优势。

3.4 HAF模块提取层数试验

由于Transformer中各层对特征提取具有不同的语义层次和表达能力，需选择合适的层数以达到最佳特征

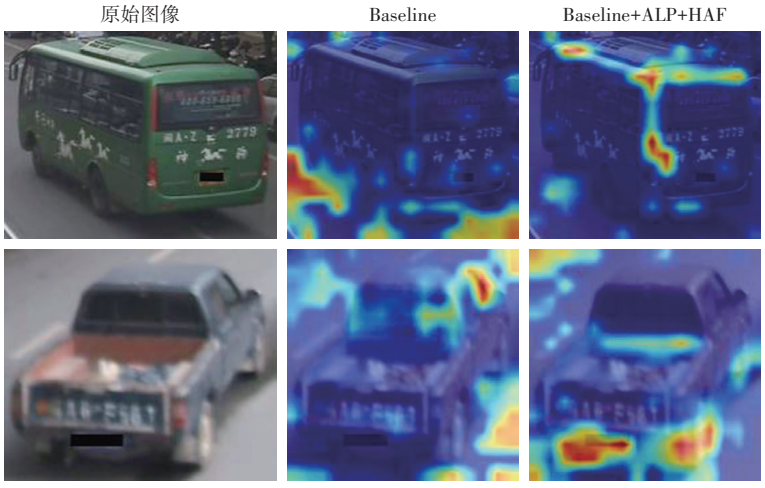


图2　模型关注区域可视化对比



（a）样本1



（b）样本2

图3　TOP-10可视化结果对比

融合效果。本文通过 VeRi-776 数据集，对 HAF 模块提取注意力矩阵的层数进行检验，结果如表2所示。

由表2可知，当仅使用最后一层的注意力引导 ALP 时，模型的 mAP 和 Rank-1 分别达到 80.7%、96.9%；当使用前 10 层时，mAP 上升至 81.0%，达到试验的最佳效果；当层数增加至 11 层时，模型训练出现边际效应，试验效果无明显提升且计算量激增；当层数增加至 12 层时，模型训练出现梯度爆炸问题，尽管试验已采用标准化措施，但在训练至 24 轮后，模型准确率仍明显下降，后续训练仍处于较低水平。造成这一现象的主要原因为：融合 12 层的注意力图意味着模型基于原始 ViT 增加了多个非线性梯度通道，虽然每一层的权重受控，但整体仍

呈现指数级增长。标准化无法从根本上约束梯度在高维空间中的非线性传播，从而引发梯度爆炸。

表2　HAF模块特征提取层数试验结果　%

方法	mAP	Rank-1
仅最后1层	80.7	96.9
前6层	80.3	96.4
前8层	80.8	97.1
前10层	81.0	97.1
前11层	81.0	97.1

因此，HAF模块在融合跨层注意力信息时，存在“最佳融合深度”：过少的层数可能无法充分建模不同语义层次的信息，且全局特征拼接后会破坏原有特征；过多则可能出现边际效应甚至造成梯度爆炸，计算消耗过多。本文选取10层作为HAF模块的提取层数，不仅覆盖了从浅层到深层的多尺度注意力语义信息，也避免了训练不稳定，进一步验证了跨层融合机制的有效性和敏感性，为后续模型的模块设计提供了参考依据。

3.5　不同车辆类型与光照条件下的模型性能评估

为了探究本文模型的泛化性能和细粒度识别能力，评估不同车辆类型与光照条件下的模型性能。考虑到VeRi-776数据集具有完备的车辆类型标签（Sedan、SUV等），使用该数据集进行细粒度试验，并且按照人工标注和平均像素亮度方法设定阈值（90）。将图像依据光照强度分为充足和不足2个类别。试验中，选取查询集中样本量较充足的主要车型，在数据集中进行检索，从而验证模型在特定细粒度属性约束下的检索效能，结果如表3所示。

由表3可知，ALP-ViT在识别车辆类型和应对光照变化方面表现出色。该模型在多种常见车辆类型识别中，均取得了较高的总体mAP，明显优于基线方法，尤其对Van、Bus等外形相对独特的类型识别，显示出良好的泛化能力。在不同光照条件下，本文模型具有较强的鲁棒性：虽然在光线不足场景下，模型的识别精度普遍受到挑战，但ALP-ViT在多数车型依然保持了较高的mAP，甚至在Truck、Estate等类型上表现出优于光线充足场景。

该模型通过结合自适应局部特征提取、层次化注意力融合等关键模块，增强了局部特征提取和多层次语义特征的建模能力，不同车辆类型和光线不足条件下保持较高的识别精度。

3.6　对比试验

分别使用VeRi-776和VehicleID数据集，对比现有先进模型的整体性能，验证本文模型的有效性，结果如表4和表5所示。

表3　不同车辆类型与光照条件下的性能对比

标签	光照条件	样本/张	mAP/%	
			Baseline	ALP-ViT
Sedan	充足	545	78.2	81.5
	不足	231	73.0	76.5
	总体	776	76.7	79.9
SUV	充足	197	79.9	81.0
	不足	121	74.0	75.5
	总体	318	77.8	79.0
Van	充足	37	92.3	94.3
	不足	5	89.5	96.0
	总体	42	92.0	94.5
Hatchback	充足	94	77.2	97.0
	不足	20	81.5	84.5
	总体	114	78.0	80.0
Pickup	充足	79	78.5	83.0
	不足	27	72.5	80.5
	总体	106	77.0	82.3
Bus	充足	43	98.0	99.7
	不足	13	91.0	99.6
	总体	56	96.5	99.7
Truck	充足	212	77.8	80.0
	不足	38	80.5	82.5
	总体	250	78.2	80.4
Estate	充足	14	75.5	79.0
	不足	2	96.5	99.0
	总体	16	78.0	81.5
所有车型	总体	1678	78.2	81.0

表4　VeRi-776数据集上各模型性能结果　%

模型	mAP	Rank-1	Rank-5
VAMI	61.3	85.9	91.8
AAVER[16]	61.2	89.0	94.7
VANet[17]	66.3	89.8	96.0
SAN	72.5	93.3	97.1
UMT[18]	75.9	95.8	98.1
PVEN	79.5	95.6	98.4
SAVER[19]	79.6	96.4	98.6
TransReID	80.6	96.9	98.7
VRcvt	80.9	97.1	98.7
A^2PM-TransReID[20]	81.4	97.0	98.5
ALP-ViT	81.0	97.1	98.8

表5　VehicleID数据集上各模型性能结果　%

模型	测试子集1		测试子集2		测试子集3	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
VAMI	63.1	83.3	52.9	75.1	47.3	70.3
QD-DLF[21]	72.3	92.5	70.7	88.9	64.1	83.8
AAVER	74.7	93.8	68.6	90.0	63.5	85.6
SAN	79.7	94.3	78.4	91.3	75.6	88.3
UMT	80.9	87.0	78.8	84.2	76.1	82.8
PVEN	84.7	97.0	80.6	94.5	77.8	92.0
SAVER	79.9	95.2	77.6	91.1	75.3	88.3
VRcvt	86.2	97.0	80.5	94.7	78.2	93.6
ALP-ViT	85.7	97.2	81.4	95.8	80.2	94.9

由表4可知，本文模型在VeRi-776数据集上取得了优异的性能，在Rank-1和Rank-5指标上实现了当前最优结果，分别达到97.1%和98.8%。同时，mAP在所有方法中处于领先水平。与主流基于ViT的方法TransReID相比，ALP-ViT的各项指标均有提升，表明在整体检索准确性和排名效果上均有明显优势。与融合方法A^2PM-TransReID相比，虽然ALP-ViT的mAP略低，但Rank-1和Rank-5均有提升，显示出更强的最终识别能力，说明模型在保持高整体匹配质量的同时，在检索精度方面更加出色。

由表5可知，本文模型在VehicleID数据集上的提升相对有限，可能的原因是VehicleID数据集仅提供视角ID，未提供摄像头ID，无法调用模型的MFE模块，但是ALP模块的投入依然使模型的重识别效果更加出色。ALP-ViT在测试子集1（800个ID）中虽略低于VRcvt，但在其他测试子集中表现优异，表明其在样本复杂、角度多变的场景中更具优势，也证明了本文模型在样本更复杂、场景更丰富的条件下依然保持较高的性能。

4　结束语

本文提出基于ViT的自适应局部划分网络，通过创新设计自适应局部划分、层次化注意力融合以及多重特征嵌入模块，增强了模型对局部细节与多层次语义特征的建模能力，显著提升了在复杂场景下（如视角变化、遮挡、外观相似车辆）的判别力与鲁棒性，展现出良好的特征表达能力。

试验结果充分验证了各模块协同工作的有效性，在复杂场景和跨摄像头匹配中表现突出。未来，可进一步探索模块优化以及扩展其他细粒度视觉识别任务，以推动模型的广泛应用。

参　考　文　献

[1] LIU X C, L W, MA H D, et al. Large-Scale Vehicle Re-Identification in Urban Surveillance Videos[C]// 2016 IEEE International Conference on Multimedia and Expo (ICME). Seattle, WA, USA: IEEE, 2016.

[2] LIU H Y, TIAN Y H, WANG Y W, et al. Deep Relative Distance Learning: Tell the Difference between Similar Vehicles[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 2167-2175.

[3] HE B, LI J, ZHAO Y F, et al. Part-Regularized Near-Duplicate Vehicle Re-Identification[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 3997-4005.

[4] ZHOU Y, SHAO L. Viewpoint-Aware Attentive Multi-View Inference for Vehicle Re-Identification. in 2018 IEEE[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018.

[5] MENG D C, LI L, LIU X J, et al. Viewpoint Alignment and Discriminative Parts Enhancement in 3D Space for Vehicle ReID[J]. IEEE Transactions on Multimedia, 2022, 25: 2954-2965.

[6] ZHU X Y, LUO Z B, FU P, et al. VOC-ReID: Vehicle Re-Identification Based on Vehicle-Orientation-Camera[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, WA, USA: IEEE, 2020.

[7] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale[EB/OL]. (2021-06-03) [2025-11-24]. https://arxiv.org/abs/2010.11929.

[8] HE S T, LUO H, WANG P C, et al. TransrReID: Transformer-Based Object Re-Identification[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 15013-15022.

[9] GONG R, ZHANG X, PAN J N, et al. Vehicle Re-Identification Based on Convolution and Vision Transformer Feature Fusion[J]. IEEE MultiMedia, 2024, 31(2): 61-68.

[10] QIAN J J, JIANG W, LUO H, et al. Stripe-Based and Attribute-Aware Network: A Two-Branch Deep Model for Vehicle Re-Identification[J]. Measurement Science and Technology, 2020, 31(9).

[11] MENG D C, LI L, HUANG Q M, et al. Parsing-Based View-Aware Embedding Network for Vehicle Re-Identification[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, Washington, USA: IEEE, 2020: 7103-7112.

[12] PHAN N, HUY T D, DUONG S T M, et al. LogoViT: Local-Global Vision Transformer for Object Re-Identifi-

cation[C]// ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023.

[13] CHEN Z, DUAN Y C, WANG W H, et al. Vision Transformer Adapter for Dense Predictions[EB/OL]. (2023-02-13) [2025-11-24]. https://arxiv.org/abs/2205.08534.

[14] ZHAI Y, GUO X, LU Y, et al. In Defense of the Classification Loss for Person Re-Identification[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, CA, USA: IEEE, 2019.

[15] LI M, LIU J, ZHENG C, et al. Exploiting Multi-View Part-Wise Correlation via an Efficient Transformer for Vehicle Re-Identification[J]. IEEE Transactions on Multimedia, 2021, 25: 919-929.

[16] 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.

[17] CHU R H, SUN Y F, LI Y D, et al. Vehicle Re-Identification with Viewpoint-Aware Metric Learning[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 8282-8291.

[18] JIN X, LAN C L, ZENG W J, et al. Uncertainty-Aware Multi-Shot Knowledge Distillation for Image-Based Object Re-Identification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press, 2020.

[19] KHORRAMSHAHI P, PERI N, CHEN J C, et al. The Devil is in the Details: Self-Supervised Attention for Vehicle Re-Identification[C]// 2020 European Conference on Computer Vision (ECCV). Glasgow, United Kingdom: Springer, 2020: 369-386.

[20] QIU M, CHRISTOPHER L A, CHIEN S, et al. Adaptive Aspect Ratios with Patch-Mixup-ViT-Based Vehicle ReID [EB/OL]. (2024-11-09) [2025-11-24]. https://arxiv. org/abs/2411.06297.

[21] ZHU J Q, ZENG H Q, HUANG J C, et al. Vehicle Re-Identification Using Quadruple Directional Deep Learning Features[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 21(1): 410-420.

（责任编辑　瑞　秋）

修改稿收到日期为2025年11月24日。