文章编号：1007-2780(2026)08-1200-13

基于自适应侧信息与残差层级注意力的车辆重识别网络

徐萌兮[1]，陈海鑫[2]，徐焕宇[3]，胡立发[4*]
(1. 无锡学院 交通与车辆工程学院，江苏 无锡 214105；
2. 南京信息工程大学 计算机学院，江苏 南京 210044；
3. 无锡学院 物联网工程学院，江苏 无锡 214105；
4. 江南大学 光电信息与物理科学学院，江苏 无锡 214122)

摘要：车辆重识别旨在跨摄像头场景下准确匹配同一车辆，是智能交通与公共安全监控的重要任务。针对现有方法局部细粒度特征利用不足、侧信息融合静态以及多层注意力融合不稳定等问题，提出一种基于自适应侧信息与残差层级注意力的车辆重识别网络。该网络以 Vision Transformer 为基础架构，引入自适应侧信息整合模块，对 Patch token 进行动态调制与上下文增强，以缓解跨摄像头、跨视角条件下的特征分布偏移；设计稳定残差跨层注意力聚合模块，在多层可学习注意力融合中引入残差结构，提高多尺度语义融合稳定性和全局引导可靠性；构建多头层级自适应划分模块，自适应提取具有一定物理可解释性的局部表征。实验结果表明，该方法在 VeRi-776 数据集上取得 83.8% 的 mAP、97.6% 的 Rank-1 和 98.8% 的 Rank-5；在 VehicleID 数据集 Test-1600 和 Test-2400 子集上分别取得 80.6% 和 78.6% 的 Rank-1。本文方法在复杂车辆重识别场景中的准确性上具有一定优势，在大规模智能交通系统中具有潜在应用价值。

关　键　词：车辆重识别；自适应侧信息；层级注意力融合；局部特征学习
中图分类号：TP391.41　　**文献标识码**：A　　**doi**：10.3724/CJLCD.2026-0118　　**CSTR**：32172.14.CJLCD.2026-0118

Vehicle re-identification network based on adaptive side information and residual hierarchical attention

XU Mengxi[1], CHEN Haixin[2], XU Huanyu[3], HU Lifa[4*]

(1. *School of Transportation and Vehicle Engineering, Wuxi University, Wuxi* 214105, *China*;
2. *School of Computer Science, Nanjing University of Information Science and Technology, Nanjing* 210044, *China*;
3. *School of Internet of Things Engineering, Wuxi University, Wuxi* 214105, *China*;
4. *School of Optoelectronic Information and Physical Science, Jiangnan University, Wuxi* 214122, *China*)

Abstract: Vehicle re-identification aims to match the same vehicle across different camera views and is a critical task for intelligent transportation and public security surveillance. Aiming at the drawbacks of

收稿日期：2026-06-29；**修订日期**：2026-07-17.
基金项目：国家自然科学基金(No. 11704377, No. 61475152)
Supported by National Natural Science Foundation of China (No. 11704377, No. 61475152)
*通信联系人，E-mail：hulifa@jiangnan.edu.cn

existing Vision Transformer-based vehicle re-identification methods, including insufficient mining of fine-grained local features, static side-information fusion strategies, and poor stability during multi-layer attention aggregation, this paper proposes a Vision Transformer vehicle re-identification network with adaptive side-information and residual hierarchical attention. Built upon the Vision Transformer backbone, the network embeds an Adaptive Side-information Integration module to dynamically modulate patch tokens and strengthen contextual cues, mitigating feature distribution discrepancies caused by varying cameras and viewing angles. Meanwhile, a Stable Residual Cross-layer Attention Aggregation module is devised, which embeds residual connections into multi-layer learnable attention fusion to stabilize multi-scale semantic aggregation and strengthen global feature guidance. In addition, a Multi-Head Hierarchical Adaptive Partitioning module is constructed to automatically extract local feature representations with a certain degree of spatial interpretability. Experimental results demonstrate that the proposed method reaches 83.8% mAP, 97.6% Rank-1 and 98.8% Rank-5 on the VeRi-776 dataset, and achieves Rank-1 accuracies of 80.6% and 78.6% on the Test-1600 and Test-2400 subsets of VehicleID respectively. The presented Vision Transformer model achieves remarkable accuracy in complicated vehicle re-identification scenarios, and possesses great practical application potential for large-scale intelligent transportation systems.

Key words: vehicle re-identification; adaptive side-information; residual hierarchical attention; local feature learning

1　引　　言

车辆重识别(Vehicle Re-Identification,ReID)旨在从不同摄像头图像中识别并匹配同一车辆，是智能交通管理、城市安防监控和公共安全分析的重要研究任务。相比一般图像检索，车辆重识别面临更复杂的应用环境：不同车辆可能具有相近品牌、型号、颜色和外观结构，类间差异较小；同一车辆也会因视角、光照、分辨率、遮挡和背景变化产生较大类内差异。因此，在复杂交通场景中学习兼具判别性与鲁棒性的车辆表征一直是该领域的关键问题。

车辆重识别的发展与大规模公开数据集密切相关。Liu等[1]构建的VeRi-776提供了多摄像头、多视角的跨区域评测基准；随后，VehicleID[2]和VERI-Wild[3]等更大规模数据集进一步推动了细粒度检索和跨场景泛化研究。早期车辆重识别方法多以卷积神经网络(CNN)为基础，围绕视角建模、局部区域学习和车辆属性融合展开。Zhou等[4]通过VAMI从单视图图像推断多视图判别特征；Liu等[5]将全局特征、局部区域特征和车辆属性联合建模；Wan等[6]结合方向不变特征与时空约束改善跨视角匹配。在细粒度对齐方面，Meng等[7]利用语义解析掩码和可见性掩码关注可见结构区域；Qian等[8]则通过结构感知条带划分增强局部特征表达。行人与车辆重识别均需学习跨摄像头鲁棒身份表征，行人重识别中的特征融合和关系建模方法可为车辆重识别提供借鉴。涂明君等[9]提出了一种多机制融合的跨模态行人重识别方法；寇旗旗等[10]则提出一种基于关系建模和频谱变换的跨模态行人重识别方法。

尽管基于CNN的ReID方法[11-12]已取得较好效果，但卷积操作受局部感受野限制，难以充分建模远距离区域之间的全局依赖。近年来，深度学习、注意力机制与Transformer架构已广泛应用于相位恢复、散射成像、光学神经形态计算、图像超分辨率及视觉识别等任务中[13-20]。Vision Transformer(ViT)凭借自注意力机制的全局建模能力，逐渐成为图像重识别任务的重要骨干网络[21]。He等[22]提出的TransReID将Transformer架构用于重识别任务，提升了全局建模能力。此后，研究者进一步探索了多尺度特征、局部注意力和辅助先验融合等策略。Gong等[23]提出的VRcvt以CLS Token为查询聚合Patch token；Qiu等[24]则结合自适应长宽比和Patch Mixup策

略提升车辆ReID的准确性与鲁棒性。

然而，现有基于ViT的车辆重识别方法仍存在不足。首先，标准ViT更侧重全局上下文建模，对车灯、车牌、轮毂、年检标志等细粒度区域的显式刻画不足，而这些细节往往是区分同型号、同颜色车辆的关键线索。其次，车辆图像常伴随复杂背景、固定摄像头视角和特定光照条件，模型可能学习到与身份无关的场景偏差，形成“背景-身份”的虚假关联，进而削弱遮挡、视角剧变和背景变化下的泛化能力。此外，不同车辆在轮胎、行李架、车灯和保险杠等局部部件上常具有相似性，这些区域在不同视角下提供了稳定视觉提示。若模型不能充分挖掘局部结构信息，容易将注意力转向无关背景，难以获得鲁棒的车辆身份表征。

针对上述问题，本文提出基于自适应侧信息与残差层级注意力的车辆重识别网络（Residual Hierarchical Attention and Adaptive Side Information Network，RHA-SI-Net）。本文关注的核心目的是在复杂成像条件下，从摄像头、视角和背景等非身份因素中逐步提纯车辆身份信息。为此，构建“条件校正-全局引导-局部解耦”的表征学习路径：ASI利用摄像头和视角信息调制Patch token，以缓解成像条件引起的特征偏移；SRA-CAA以残差方式融合不同层级的注意力响应，获得稳定的车辆主体空间先验；MH-HAP在该先验约束下竞争式分配局部区域并加权聚合，提取互补的局部身份表征。三者依次完成条件校正、全局引导和局部表征，共同提升复杂场景下身份表征的鲁棒性。本文主要贡献如下：

（1）自适应侧信息整合（Adaptive Side-information Integration，ASI）模块：利用摄像头与视角信息对Patch token进行动态调制和上下文增强，提高跨摄像头、跨视角特征对齐能力。

（2）稳定残差跨层注意力聚合（Stable Residual Cross-Layer Attention Aggregation，SRA-CAA）模块：融合ViT多层注意力，并引入残差结构提升多尺度语义融合稳定性和全局引导可靠性。

（3）多头层级自适应划分（Multi-Head Hierarchical Adaptive Partitioning，MH-HAP）模块：将层级注意力引导转化为具有物理分工意义的局部表征，增强细粒度判别能力。

2 自适应侧信息与残差层级注意力的车辆重识别网络

2.1 总体架构

针对车辆重识别中侧信息和局部细粒度特征利用不足、多层注意力融合不稳定的问题，本文提出RHA-SI-Net，其整体结构如图1所示。模型以ViT-B/16为主干，输入图像经分块和线性嵌入后，首先通过ASI模块对Patch token进行自适应调节，再送入多层ViT提取特征。各层包含多头自注意力（MHSA）和前馈网络（FFN），并采用残差连接与层归一化。随后，SRA-CAA模块对多层注意力进行稳定自适应融合，MH-HAP模块在其引导下生成具有物理分工意义的局部表征。

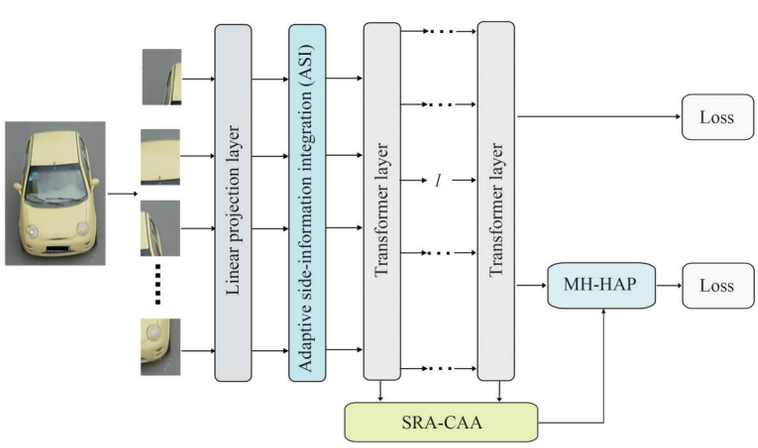


图1 RHA-SI-Net整体结构

Fig. 1 Overall architecture of RHA-SI-Net

2.2　自适应侧信息整合模块

为利用摄像头与视角辅助信息，本文设计自适应侧信息整合模块（ASI）。给定 Camera ID 和 View ID，模块通过可学习嵌入层获得摄像头与视角嵌入，拼接后输入轻量级 MLP，形成联合侧信息表示。该表示分别进入 FiLM 和 Prompt 分支：FiLM 分支生成缩放因子 γ 和偏移因子 β，对 LayerNorm 后的 Patch token 进行通道调制；Prompt 分支生成 camera-view prompt token，并通过交叉注意力与视觉 token 交互，为局部区域提供条件相关的上下文。增强 token 与原始 token 经残差融合后输入 Transformer 编码器。相比静态相加式嵌入，ASI 能更灵活地校正摄像头与视角引起的特征偏移，其结构如图 2 所示。ASI 可视为一种条件相关的特征校正机制。摄像头和视角信息反映了车辆图像的成像条件，FiLM 分支根据该条件生成缩放因子和偏移量，对 Patch token 进行自适应调制，从而减小不同摄像头和视角下的特征分布差异。Prompt 分支进一步提供条件相关的上下文信息，使不同区域能够选择性利用成像条件先验。

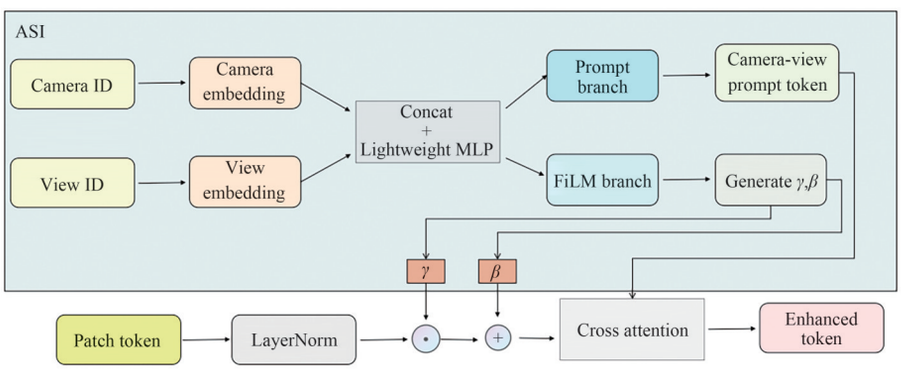


图 2　ASI 模块结构

Fig. 2　Structure of the ASI module

2.3　稳定残差跨层注意力聚合模块

ViT 的特征层级分布具有显著的语义差异性：浅层注意力多聚焦于高频纹理细节（如车灯格栅），而深层注意力则高度抽象，侧重于全局轮廓响应。若仅依赖单一特征层引导局部划分，极易造成细粒度判别线索的丢失。因此，本文设计层级注意力融合（CAA）模块，旨在实现跨层级多尺度信息的有机聚合。针对 ViT 编码器输出的前 L 层（本文设定 $L=10$），提取对应的原始注意力亲和矩阵 $\boldsymbol{A}_l$。为自适应评估各层级信息的贡献度，引入可学习的层级权重 $\boldsymbol{w}_l$，并经 Softmax 函数归一化后执行加权求和操作，公式如下：

$$T_{\text{out}}=\text{proj}\left(\sum_{l=1}^{L}\text{softmax}\left(\boldsymbol{w}_l\right)\cdot\boldsymbol{A}_l\right),\tag{1}$$

式中，proj 表示多头特征维度的线性投影算子。融合生成的注意力分布 T_{out} 有效整合了浅层细节与深层语义，为后续模块提供了具备强鲁棒性与尺度兼容性的全局上下文引导。

为进一步提升跨层注意力融合的稳定性，在 CAA 模块基础上提出稳定残差跨层注意力聚合（SRA-CAA）模块，其结构如图 3 所示。与原始 CAA 直接采用 T_{out} 作为后续局部划分引导不同，SRA-CAA 选取单层实验中性能最佳的第 10 层注意力作为稳定基准注意力 T_{base}，并通过残差形式生成最终融合注意力：

$$T_{\text{fused}}=T_{\text{base}}+\lambda T_{\text{out}},\tag{2}$$

其中，λ 为可学习缩放系数，初始化为较小值。该残差结构使模型在训练初期主要依赖稳定的高层注意力分布，同时逐步引入多层聚合注意力中的细粒度补充信息，从而缓解浅层背景噪声和深层语义冲突对局部划分的干扰。最后，从 T_{fused} 中提取 CLS token 对 Patch token 的注意力响应，并重塑为全局空间引导图 $\boldsymbol{A}_{\text{cls}}$，输入 MHHAP 模块以约束车辆局部区域的自适应划分。该设计仅增加一个轻量级可学习缩放参数，在几乎不增加模型复杂度的情况下，提高了 CAA 模块的训练稳定性和局部引导可靠性。公式（2）中的残差结构以基准层注意力为主要引导，并通过较小的可学习系数逐步引入跨层注意力信息。

这样可以限制多层注意力在训练初期对原始注意力分布的扰动，减少浅层背景响应和深层语义冲突造成的训练不稳定，并在训练后期补充多尺度细节信息。

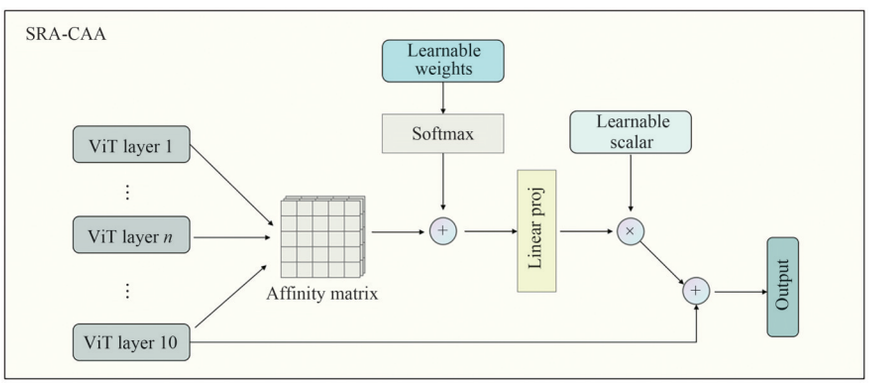


图 3 SRA-CAA 模块结构

Fig. 3 Structure of the SRA-CAA module

2.4 多头层级自适应划分模块

为弥补 ViT 局部特征建模能力的不足，本文设计多头层级自适应划分(MH-HAP)模块，结构如图 4 所示。该模块以 SRA-CAA 输出的层级融合注意力为全局先验，结合空间多头注意力和头间竞争，将 ViT 特征转化为具有明确物理划分意义的局部表征。

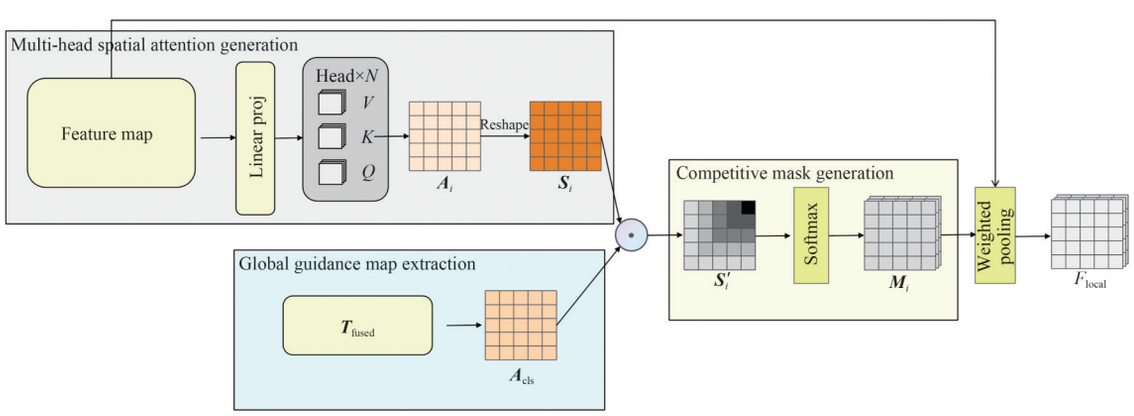


图 4 MH-HAP 模块结构

Fig. 4 Structure of the MH-HAP module

2.4.1 基于层级引导的多头竞争掩码

SRA-CAA 模块输出的全局空间引导图 $\boldsymbol{A}_{\mathrm{cls}}$ 融合了多层语义，天然具备抑制背景噪声和高亮车辆主体的全局引导能力。为了实现部件级的特征解耦，本模块对 ViT 编码器输出的 C 维空间特征图引入多头自注意力分发机制。具体而言，通过 G 个并行的注意力头，空间特征被线性投影为 G 组查询向量 $\boldsymbol{Q}_i$ 和键向量 $\boldsymbol{K}_i$。通过计算 $\boldsymbol{Q}_i$ 和 $\boldsymbol{K}_i$ 的点积并进行归一化，得到每个头内部的 patch 间亲和力矩阵 $\boldsymbol{A}_i$：

$$\boldsymbol{A}_i=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d_k}}\right), \tag{3}$$

式中，d_k 为缩放因子。为了鼓励不同的头关注不同的车辆部件，模块对每个亲和力矩阵 $\boldsymbol{A}_i$ 沿查询(Query)维度进行聚合，并将其重塑为二维空间注意力图 $\boldsymbol{S}_i$。该操作的数学表达如式(4)所示：

$$\boldsymbol{S}_i=\mathrm{Reshape}_{H\times W}\left(\sum_{q=1}^{N}\boldsymbol{A}_i(q,\cdot)\right), \tag{4}$$

式中，$A_i(q,\cdot)$ 表示第 q 个查询对所有键的注意力分布。通过沿 q 维度聚合，生成的 $\boldsymbol{S}_i$ 刻画了

第 i 个注意力头自发关注的局部模式。随后，为了确保模型重点关注车辆主体而非背景噪声，将层级引导图 $\boldsymbol{A}_{\text{cls}}$ 与每个 $\boldsymbol{S}_i$ 进行逐元素相乘，得到融合空间注意力图 $\boldsymbol{S}_i'$。最终，为了迫使 G 个头关注互不重叠的物理部件（如车灯、车牌、侧窗），在头维度（Head dimension）执行 Softmax 归一化产生空间竞争，从而生成第 i 个部件的多样性掩码 $\boldsymbol{M}_i$：

$$\boldsymbol{M}_i = \text{Softmax}_{\text{dim}=\text{head}}\left(\boldsymbol{A}_{\text{cls}} \odot \boldsymbol{S}_i\right),\ i \in \{1, \cdots, G\}. \tag{5}$$

MH-HAP 首先利用全局引导图抑制车辆主体之外的背景响应，再通过头维度 Softmax 使不同注意力头在车辆主体内部进行竞争式分配，从而减少多个注意力头重复关注同一区域的现象，并获得互补的局部表征。

其中，多头 Q/K 投影用于捕获不同局部纹理，头维度 Softmax 在车辆主体内部进行竞争分配，全局引导图负责抑制背景，三者协同实现细粒度部件的自动解耦与分配。

2.4.2　自适应区域特征聚合

在获得 G 个部件多样性掩码后，模块采用加权池化技术，将离散的 Patch 序列聚合成具备鲁棒性的局部特征向量 $\boldsymbol{f}_{l,i}$：

$$\boldsymbol{f}_{l,i} = \frac{\sum_{h=1}^{H}\sum_{w=1}^{W} \boldsymbol{M}_i(h,w) \cdot \boldsymbol{F}(h,w)}{\sum_{h=1}^{H}\sum_{w=1}^{W} \boldsymbol{M}_i(h,w) + \epsilon}, \tag{6}$$

式中：$\boldsymbol{F}(h,w)$ 为编码器输出的空间特征图，ϵ 为极小值常数以保证数值稳定。该聚合策略根据掩码的置信度，自适应调整聚合重心，使模型在视角偏转或局部遮挡时仍能从可见部件中提取判别性线索。

2.5　损失函数

为监督 RHA-SI-Net 的全局与局部分支，本文以分类损失提升身份判别力，并通过度量学习优化特征空间。对 MH-HAP 提取的局部表征和 ViT 主干输出的全局表征分别实施监督，使不同语义维度的特征得到有效优化。

2.5.1　身份判别损失

身份判别损失用于监督模型在已知类别空间内的分类精度。对于全局特征 $\boldsymbol{f}_{\text{g}}$ 和由 G 个局部头拼接而成的增强特征 $\boldsymbol{f}_{\text{l}}$，本研究分别应用独立的线性分类器，并采用交叉熵损失函数，以增强全局特征和局部特征的身份判别能力。损失函数定义如式（7）所示：

$$\mathcal{L}_{\text{ID}} = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i|x_i), \tag{7}$$

式中：N 为批次大小，$p(y_i|x_i)$ 为模型预测样本 x_i 属于真实类别 y_i 的后验概率。

2.5.2　特征度量损失

为了增强特征对外观极相似车辆（如同型号、同颜色车辆）的细粒度区分度，本研究引入了软间隔三元组损失（Soft-margin Triplet Loss）[25]。该损失通过约束样本间的欧氏距离，强制拉近同类样本（正样本对），同时推远异类样本（负样本对）。公式定义为：

$$\mathcal{L}_{\text{tri}} = \frac{1}{P \times M}\sum_{i=1}^{P}\sum_{a=1}^{M} \ln\left(1 + \exp\left(d_{\text{pos}} - d_{\text{neg}}\right)\right), \tag{8}$$

式中：P 为类别数量，M 为每类样本数。d_{pos} 和 d_{neg} 分别表示锚点与正、负样本在特征空间内的平方欧氏距离。

2.5.3　总损失函数

最终，模型的总损失函数 L_{total} 是作用于两个分支的各个损失项的总和，其构成如式（9）所示：

$$\mathcal{L}_{\text{total}} = (\mathcal{L}_{\text{ID}}^{\text{global}} + \mathcal{L}_{\text{tri}}^{\text{global}}) + (\mathcal{L}_{\text{ID}}^{\text{local}} + \mathcal{L}_{\text{tri}}^{\text{local}}), \tag{9}$$

式中，$\mathcal{L}_{\text{ID}}^{\text{global}}$ 与 $\mathcal{L}_{\text{tri}}^{\text{global}}$ 分别为单独全局分支计算的 ID 损失和三元组损失。$\mathcal{L}_{\text{ID}}^{\text{local}}$ 与 $\mathcal{L}_{\text{tri}}^{\text{local}}$ 分别为单独局部分支计算的 ID 损失和三元组损失。本模型为双分支模型，两个分支分别进行损失函数计算，相加得到最终结果。

3　实验与分析

3.1　数据集与评价指标

3.1.1　数据集

为验证所提出的 RHA-SI-Net 在车辆重识别任务中的有效性与泛化能力，本文选取 VeRi-776[1] 和 VehicleID[2] 两个公开大规模基准数据集进行实验评估。VeRi-776 是车辆重识别领域常用的真实城市交通数据集，由 20 个监控摄像头采集，包含 776 辆车的 49 357 张图像，人工为每一张图像中的运动车辆标注矩形边界框。该数据集覆盖多摄像头、多视角和复杂交通环境，可反映实际监控场景下的视角变化、光照差异与背景干扰。VehicleID 是另一个具有代表性的大规模车

辆重识别数据集，包含221 763张车辆图像和26 267个车辆ID，以预裁剪的车辆矩形图像形式发布，边界框裁剪方式未具体说明。相比VeRi-776，其ID规模更大，更适合评估模型在大规模检索场景下的区分能力和泛化性能。

3.1.2 评价指标

采用平均精度均值(mean Average Precision，mAP)和累积匹配特性(Cumulative Matching Characteristic，CMC)曲线评价各方法的检索性能。mAP衡量整体排序精度，CMC反映查询图像在前若干检索结果中成功匹配目标车辆的概率，通常使用Rank-1、Rank-5等指标。两者结合可从整体检索质量和前列命中率两个角度评估模型性能。

3.2 实验配置与参数设置

本文实验基于PyTorch框架，使用NVIDIA RTX 4090 GPU。训练阶段，输入图像统一调整为256×256，并采用随机水平翻转、边缘填充和随机擦除等增强策略。模型以ImageNet预训练的ViT-B/16为主干进行迁移学习。

训练初始学习率为8×10^{-3}，采用线性预热策略；总训练周期为120，批量大小为64。优化器为带动量的SGD，动量系数为0.9，权重衰减为1×10^{-4}。推理阶段输入尺寸保持256×256，批量大小设为128，并开启特征归一化以保证尺度一致性。

3.3 实验结果与分析

VeRi-776与VehicleID在数据采集、标注信息和评测协议方面存在差异，因此本文未强行采用完全相同的对比方法集合。VeRi-776包含摄像头信息，通常采用mAP、Rank-1和Rank-5评测；VehicleID不提供摄像头ID，主要采用Test-800、Test-1600和Test-2400子集上的Rank-1、Rank-5评测。本文依据两个数据集的标准划分和主流协议分别选择代表性方法，并尽可能保留在两者上均有公开结果的方法。对比结果主要引自相关文献，各方法均采用原文报告的标准评测结果。部分方法仅报告其中一个数据集的结果，因而无法直接用于另一数据集的比较。由于不同方法在骨干网络、输入尺寸、预训练方式和训练策略上可能存在差异，结果主要用于相对性能参考。

3.3.1 VeRi-776数据集的实验结果与分析

为评估综合性能，本文在VeRi-776数据集上与多种先进车辆重识别方法进行对比。实验结果见表1，Top10可视化结果见图5。

表1 VeRi-776数据集上的实验结果

Tab. 1 Experimental results on the VeRi-776 dataset

方法	mAP/%	Rank-1/%	Rank-5/%
AAVER[26]	61.2	89.0	94.7
VAMI	61.3	85.9	91.8
SAN	72.5	93.3	97.1
UMTS[27]	75.9	95.8	98.1
PVEN	79.5	95.6	98.4
SAVER[28]	79.6	96.4	98.6
GiT[29]	80.3	96.8	98.6
TransReID	80.6	96.9	98.7
VRcvt	80.9	97.1	98.7
A^2PM-TransReID	81.4	97.0	98.5
CLIP-ReID[30]	83.3	97.4	98.8
MATNet[31]	82.3	96.9	97.3
EvoPrompt-ReID[32]	84.2	97.5	—
Ours	83.8	97.6	98.8

实验结果表明，本文方法在VeRi-776上取得83.8% mAP、97.6% Rank-1和98.8% Rank-5。与PVEN和SAVER相比，本文方法的mAP分别提高4.3和4.2个百分点；相比于TransReID方法，mAP提升3.2个百分点；相比于CLIP-ReID，mAP和Rank-1分别提高0.5和0.2个百分点，Rank-5保持一致。EvoPrompt-ReID虽然在mAP上高0.4个百分点，但本文方法在Rank-1指标上高0.1个百分点。Top10可视化结果中，仅有第5测试样例Rank-9与Rank-10出现与极为相似的同类型出租车误判，说明在不依赖额外多模态知识的情况下，层级引导与局部部件解耦即可有效提升车辆重识别性能。

3.3.2 VehicleID数据集的实验结果与分析

VehicleID包含Test-800、Test-1600和Test-2400三个规模递增的测试子集。由于该数据集不提供摄像头ID，仅提供前视和后视两类视角，ASI模块无法充分利用全部侧信息。由表2可见，VRcvt在Test-800的Rank-1略优于本文方法；在Test-1600和Test-2400上，本文方法的

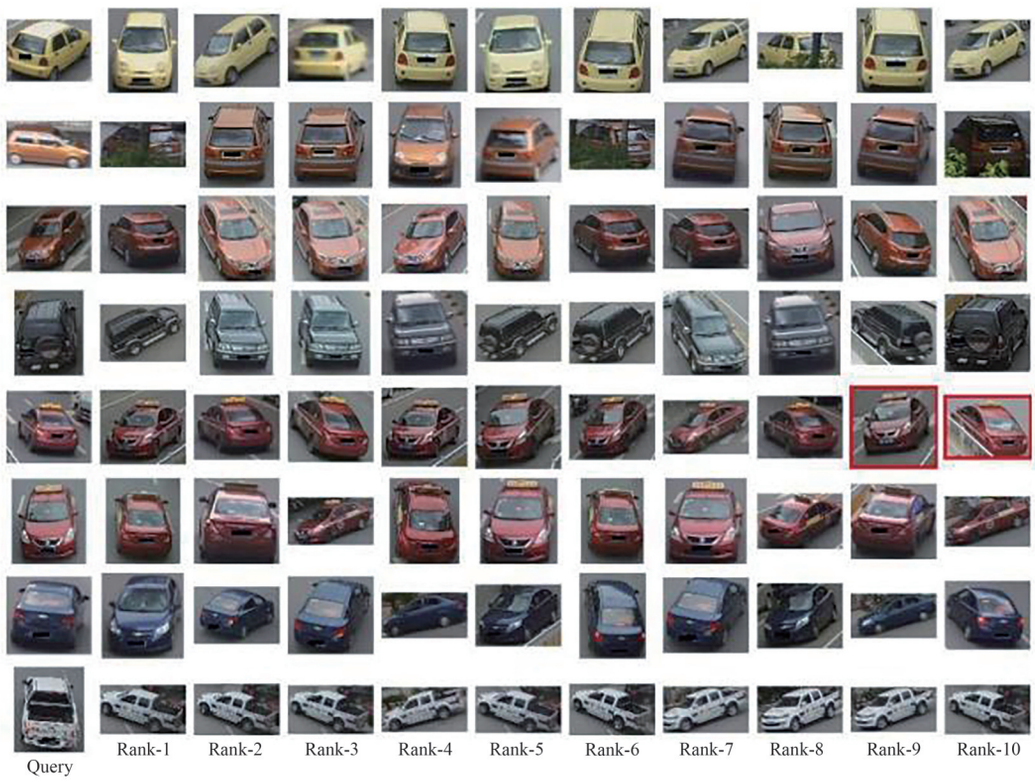


图 5　VeRi-776 数据集上的 Top10 可视化结果

Fig. 5　Top10 visualization results on VeRi-776 dataset

表 2　VehicleID 数据集上的实验结果

Tab. 2　Experimental results on the VehicleID dataset

方法	Test-800		Test-1600		Test-2400	
	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[33]	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
UMTS	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
HCI-Net[34]	83.8	96.5	79.4	92.7	76.4	91.2
VRcvt	86.2	97.0	80.5	94.7	78.2	93.6
Ours	85.6	97.9	80.6	94.8	78.6	94.5

Rank-1 分别为 80.6% 和 78.6%，略高于 VRcvt，表现出较好的大规模检索能力。结果表明，局部部件解耦与细粒度特征挖掘有助于缓解大量相似车辆造成的语义混叠。

3.3.3　跨域车辆重识别对比实验

为检验模型的泛化能力，我们进行了跨域对比实验：在一个数据集上进行训练，在另一个未参与训练的数据集上直接进行测试，并与 CycleGAN、GLCAM-Net 及 DAVR 等方法进行比较，测试条件相同，结果如表 3 所示。

在 VeRi-776→VehicleID 方向上，本文方法取得 46.7% mAP 和 40.9% Rank-1，mAP 明显高于 CycleGAN 和 GLCAM-Net，并略高于 DAVR，表明该方法具有较好的跨数据集迁移能力。

表3 跨域车辆重识别对比实验结果

Tab. 3 Comparison results of cross-domain vehicle re-identification

方法	VeRi-776→VehicleID		VehicleID→VeRi-776	
	mAP/%	Rank-1/%	mAP/%	Rank-1/%
CycleGAN[35]	31.9	27.1	21.8	55.4
GLCAM-NET[36]	43.1	34.2	27.8	63.6
DAVR[37]	45.1	40.7	26.3	63.6
Ours	46.7	40.9	27.8	63.8

在VehicleID→VeRi-776方向上，本文方法的mAP为27.8%，与GLCAM-Net持平；Rank-1为63.8%，略高于GLCAM-Net和DAVR，并显著高于CycleGAN。结果表明，本文方法具有一定的跨域泛化能力。

3.4 消融实验

为验证RHA-SI-Net各模块的有效性，本文在VeRi-776数据集上以ViT-B/16为基线模型(Baseline)进行消融实验。

如表4所示，基线模型的mAP和Rank-1分别为78.2%和96.5%。引入CAA后mAP提升至79.8%，引入SRA-CAA后mAP提升至81.2%，说明添加稳定残差结构后，CAA模块性能有所提升。由于MH-HAP依赖SRA-CAA输出的全局引导图，无法脱离SRA-CAA独立运行，引入SRA-CAA与MH-HAP后，mAP提升至81.9%，较基线模型提高3.7个百分点，说明稳定层级注意力引导与多头局部竞争能够有效增强细粒度特征表达。单独引入ASI时，mAP达到80.9%，表明摄像头和视角信息对跨场景特征对齐具有积极作用。当ASI、SRA-CAA与MH-HAP联合使用时，mAP提升至83.8%，较基线模型提高5.6个百分点，说明了本文方法所引入的动态侧信息与局部特征解耦之间的互补性，能够共同提升模型性能。

表4 消融实验结果

Tab. 4 Ablation study results %

方法	mAP	Rank-1	Rank-5
Baseline	78.2	96.5	97.2
Baseline+ASI	80.9	96.8	98.4
Baseline+CAA	79.8	96.5	98.2
Baseline+SRA-CAA	81.2	96.9	98.4
Baseline+MH-HAP+SRA-CAA	81.9	96.9	98.4
Baseline+ASI+MH-HAP+SRA-CAA	83.8	97.6	98.8

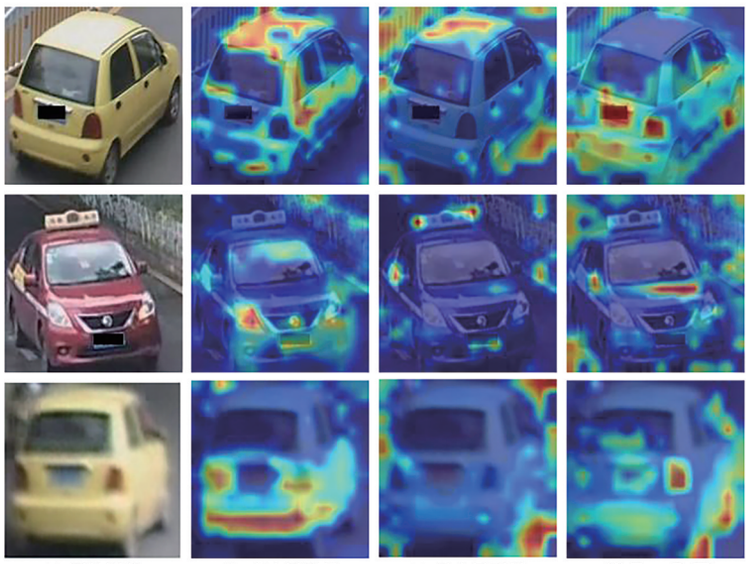
3.5 全局空间引导图可视化分析

为了更直观地体现SRA-CAA模块输出的全局空间引导图有效性，本文对空间注意力引导图进行了一组可视化实验。分别为有全局空间引导图、无全局空间引导图，以及与TransReID注意力引导图的对比。在有全局引导组中，将SRA-CAA生成的引导图与MH-HAP局部注意力逐元素相乘后进行区域划分；在无全局引导组中移除该引导图，其余网络结构和实验设置保持不变。

如图6所示，“有全局引导”的SRA-CAA模块，热力图较为精准地聚焦在车辆本身的关键部件上，如前大灯、尾灯和车牌区域，几乎没受到背景干扰。但在“无全局特征引导”的对照实验中，注意力极为发散，错误地激活了大量背景区域。这说明SRA-CAA多层次融合注意力在引导关注车辆主体、抑制背景噪声方面起到了重要作用。相较于TransReID方法，本文的注意力引导精确地聚焦在车灯等高区分度的关键局部部件上，具有更好的细粒度特征提取能力，受背景噪声影响更小。

3.6 SRA-CAA模块提取层数实验与分析

为分析层级跨度对注意力融合的影响，本文对SRA-CAA的融合层数L进行参数敏感性实验，除融合层数外，ASI、MH-HAP及训练参数均保持不变。如表5所示，仅使用单一层性能最优的第10层注意力时，获得82.8%的mAP，说明单一深层语义仍可能忽略局部细节。融合前6层时，mAP降至81.5%，表明过浅层容易引入背景噪声。随着融合深度增加，性能逐步提升，前10层取得最佳结果，mAP和Rank-1分别为83.8%和97.6%。融合前11层时出现性能下降，融合



(a) 原始图像 (a) Original image　(b) 有全局引导 (b) With global map　(c) 无全局引导 (c) Without global map　(d) TransReID (d) TransReID

图 6　全局引导图的可视化结果

Fig. 6　Visualization of global guidance map

前12层会导致明显的梯度波动，评价指标明显下降，原因在于更深层注意力过度耦合预训练语义，易与局部竞争机制冲突。实验结果表明，模型存在最佳层数，网络层数设置偏少时，模型难以充分挖掘多层级语义信息；而层数过度堆叠不仅会产生性能边际递减效应，易引发梯度波动问题，大幅提升模型计算开销。因此，本文将最大融合层数设为$L=10$，并将第10层作为残差基准T_{base}，以兼顾判别性、稳定性和优化安全性。

表 5　SRA-CAA 模块提取不同层数实验结果

Tab. 5　Results of SRA-CAA with different extraction layers

方法	mAP/%	Rank-1/%
第10层	82.8	96.9
前6层	81.5	96.5
前8层	82.2	97.3
前10层	83.8	97.6
前11层	83.2	97.5
前12层	80.1	95.8

3.7　MH-HAP模块多头数量实验与分析

为确定MH-HAP模块中多头的数量，本文对其进行对比实验，不启用ASI。表6显示，MH-HAP在$G=3$时取得最佳性能（mAP 81.9%，Rank-1 96.9%）。该划分粒度与车辆常见3大拓扑区域相匹配，即上部车顶/车窗区域、中部车灯/车身区域和下部底盘/车牌区域。当头数继续增加时，性能逐步下降，说明过度竞争会造成语义碎片化，破坏车辆部件的空间连续性并引入冗余局部噪声。因此，本文设定$G=3$，以平衡局部解耦粒度与车辆结构先验。

表 6　在VeRi-776上的MH-HAP模块多头数量实验结果

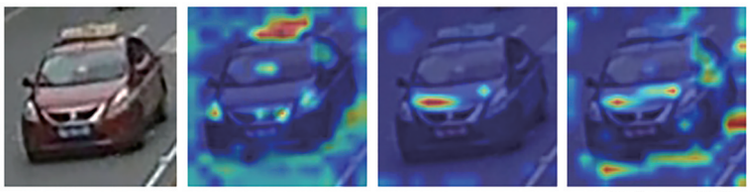
Tab. 6　Experimental results on the number of heads in the MH-HAP module on VeRi-776　%

Head数量	mAP	Rank-1
Baseline+MH-HAP(2)+SRA-CAA	80.9	96.5
Baseline+MH-HAP(3)+SRA-CAA	81.9	96.9
Baseline+MH-HAP(4)+SRA-CAA	81.7	96.8
Baseline+MH-HAP(5)+SRA-CAA	81.2	96.7
Baseline+MH-HAP(6)+SRA-CAA	80.6	96.5

为进一步分析MH-HAP的机制，本文将$G=3$时各注意力头生成的热力图可视化，如图7所示。3个头分别关注车顶、车辆前侧和车辆底部区域，表明MH-HAP能够学习互补且具有物理意义的局部表征。

3.8　模型复杂度与推理效率分析

在相同硬件与测试配置下，本文比较TransReID与所提模型的参数量和推理速度，结果见表7。本文模型包含88M参数，仅比TransReID增加1M；



(a) 原始图像 (a) Original image　(b) 注意力头1图像 (b) Head 1 image　(c) 注意力头2图像 (c) Head 2 image　(d) 注意力头3图像 (d) Head 3 image

图 7　MH-HAP 模块内部注意力头可视化结果

Fig. 7　Visualization of internal attention heads in MH-HAP

表 7　TransReID 与本模型参数、推理效率对比

Tab. 7　Comparison of parameters and inference efficiency between TransReID and our model

方法	参数量/M	FPS
TransReID	87	208
Ours	88	192

推理速度为 192 FPS，比 TransReID 低 16 FPS，仍保持较高的推理效率。

4　结　　论

针对车辆重识别中局部细粒度特征利用不足、跨摄像头视角差异明显以及多层注意力融合稳定性不足等问题，本文提出基于自适应侧信息与残差层级注意力的车辆重识别网络（RHA-SI-Net）。ASI 自适应建模摄像头和视角侧信息，缓解拍摄条件引起的特征偏移；SRA-CAA 稳定融合 ViT 多层注意力，为局部划分提供全局引导；MH-HAP 在车辆主体内进行多头竞争式划分，提取具有物理意义的局部表征。实验结果表明，RHA-SI-Net 在 VeRi-776 数据集上取得 83.8% mAP、97.6% Rank-1 和 98.8% Rank-5；在 VehicleID 数据集 Test-1600 和 Test-2400 子集上分别取得 80.6% 和 78.6% Rank-1。与其他方法相比，本文方法表现出较强的竞争力，并具有一定的跨数据集泛化能力。后续研究将进一步探索轻量化设计、跨域泛化以及时空信息融合等方向，以提升模型在真实大规模交通监控系统中的应用能力。

参　考　文　献：

[1] LIU X C, LIU W, MA H D, *et al*. Large-scale vehicle re-identification in urban surveillance videos [C]// *Proceedings of the* 2016 *IEEE International Conference on Multimedia and Expo* (*ICME*). Seattle: IEEE, 2016: 1-6.

[2] LIU H Y, TIAN Y H, WANG Y W, *et al*. Deep relative distance learning: tell the difference between similar vehicles [C]//*Proceedings of the* 2016 *IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas: IEEE, 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* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 3230-3238.

[4] ZHOUY Y, SHAO L. Viewpoint-aware attentive multi-view inference for vehicle re-identification [C]// *Proceedings of the* 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018: 6489-6498.

[5] LIU X B, ZHANG S L, HUANG Q M, *et al*. RAM: a region-aware deep model for vehicle re-identification [C]// *Proceedings of the* 2018 *IEEE International Conference on Multimedia and Expo* (*ICME*). San Diego: IEEE, 2018: 1-6.

[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* 2017 *IEEE International Conference on Computer Vision*. Venice: IEEE, 2017: 379-387.

[7] MENG D H, LI L, LIU X J, *et al*. Parsing-based view-aware embedding network for vehicle re-identification [C]// *Proceedings of the* 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2020: 7101-7110.

[8] 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): 095401.

[9] 涂明君,林国军,冯凯,等. 基于多机制融合增强的跨模态行人重识别方法[J]. 液晶与显示,2026,41(5):675-686.
TU M J, LIN G J, FENG K, *et al*. Cross-modal pedestrian re-identification method enhanced by multi-mechanism fusion [J]. *Chinese Journal of Liquid Crystals and Displays*, 2026, 41(5): 675-686. (in Chinese)

[10] 寇旗旗,乔鑫,牛凯凯,等. 关系建模和频谱变换的跨模态行人重识别方法[J]. 光学 精密工程,2026,34(6):973-989.
KOU Q Q, QIAO X, NIU K K, *et al*. Cross-modal person re-identification method relation modeling and spectral transformation [J]. *Optics and Precision Engineering*, 2026, 34(6): 973-989. (in Chinese)

[11] 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* 14*th European Conference on Computer Vision—ECCV* 2016. Amsterdam: Springer, 2016: 869-884.

[12] 姬晓飞,赵帅,宋京浩,等. 基于姿势估计和特征融合的行人重识别算法[J]. 电子测量与仪器学报,2024,38(4):187-194.
JI X F, ZHAO S, SONG J H, *et al*. Person re-identification algorithm based on pose estimation and feature fusion [J]. *Journal of Electronic Measurement and Instrumentation*, 2024, 38(4): 187-194. (in Chinese)

[13] WANG K Q, SONG L, WANG C T, *et al*. On the use of deep learning for phase recovery [J]. *Light*: *Science & Applications*, 2024, 13(1): 4.

[14] LIU H S, WANG F, JIN Y, *et al*. Learning-based real-time imaging through dynamic scattering media [J]. *Light*: *Science & Applications*, 2024, 13(10): 194.

[15] DONG Y B, BAI Y C, ZHANG Q M, *et al*. High-throughput optical neuromorphic graphic processing at millions of images per second [J]. *eLight*, 2025, 5(1): 29.

[16] DUN X, ZHANG J, CHEN F S, *et al*. Neural array meta-imaging [J]. *eLight*, 2025, 5(1): 33.

[17] JEONG H, YOON C, LIM H, *et al*. MT-former: multi-task hybrid transformer and deep support vector data description to detect novel anomalies during semiconductor manufacturing [J]. *Light*: *Advanced Manufacturing*, 2025, 6(2): 32.

[18] 周颖,裴盛虎,陈海永,等. 基于多尺度自适应注意力的图像超分辨率网络[J]. 光学 精密工程,2024,32(6):843-856.
ZHOU Y, PEI S H, CHEN H Y, *et al*. Image super-resolution network based on multi-scale adaptive attention [J]. *Optics and Precision Engineering*, 2024, 32(6): 843-856. (in Chinese)

[19] 滕嘉玮,田蕾,江山,等. 卫星关键部件智能识别与三维重建方法[J]. 中国光学(中英文),2025,18(6):1376-1387.
TENG J W, TIAN L, JIANG S, *et al*. Intelligent recognition and 3D reconstruction method for satellite key components [J]. *Chinese Optics*, 2025, 18(6): 1376-1387. (in Chinese)

[20] 黄庆东,苏宇辉,刘依华,等. 基于PATN和自注意力机制的多分辨率人体姿态迁移方法[J]. 液晶与显示,2025,40(9):1381-1389.
HUANG Q D, SU Y H, LIU Y H, *et al*. Multi-resolution human posture transfer algorithm based on PATN and self-attention mechanism [J]. *Chinese Journal of Liquid Crystals and Displays*, 2025, 40(9): 1381-1389. (in Chinese)

[21] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An image is worth 16×16 words: transformers for image recognition at scale [C]. International Conference on Learning Representations. Online: OpenReview.net, 2021.

[22] HE S T, LUO H, WANG P C, *et al*. TransReID: transformer-based object re-identification [C]//*Proceedings of the* 2021 *IEEE/CVF International Conference on Computer Vision*. Montreal: IEEE, 2021: 14993-15002.

[23] GONG R, ZHANG X, PAN J A, *et al*. Vehicle reidentification based on convolution and vision transformer feature fusion [J]. *IEEE MultiMedia*, 2024, 31(2): 61-68.

[24] QIU M, CHRISTOPHER L A, CHIEN S, *et al*. Adaptive aspect ratios with patch-mixup-ViT-based vehicle

ReID [C]//*Proceedings of the* 2025 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). Hyderabad: IEEE, 2025: 1-5.

[25] ZHAI Y, GUO X, LU Y, *et al*. In defense of the classification loss for person re-identification [C]//*Proceedings of the* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. Long Beach: IEEE, 2019: 1526-1535.

[26] KHORRAMSHAHI P, KUMAR A, PERI N, *et al*. A dual-path model with adaptive attention for vehicle re-identification [C]//*Proceedings of the* 2019 *IEEE/CVF International Conference on Computer Vision*. Seoul: IEEE, 2019: 6132-6140.

[27] 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* 34*th AAAI Conference on Artificial Intelligence*. New York: AAAI Press, 2020: 11165-11172.

[28] KHORRAMSHAHI P, PERI N, CHEN J C, *et al*. The devil is in the details: self-supervised attention for vehicle re-identification [C]//*Proceedings of the* 16*th European Conference on Computer Vision—ECCV* 2020. Glasgow: Springer, 2020: 369-386.

[29] SHEN F, XIE Y, ZHU J Q, *et al*. GiT: graph interactive transformer for vehicle re-identification [J]. *IEEE Transactions on Image Processing*, 2023, 32: 1039-1051.

[30] LI S Y, SUN L, LI Q L. CLIP-ReID: exploiting vision-language model for image re-identification without concrete text labels [C]//*Proceedings of the* 37*th AAAI Conference on Artificial Intelligence*. Washington: AAAI Press, 2023: 1405-1413.

[31] ZHANG X, LING Y N, LI K G, *et al*. Multimodality adaptive transformer and mutual learning for unsupervised domain adaptation vehicle re-identification [J]. *IEEE Transactions on Intelligent Transportation Systems*, 2024, 25(12): 20215-20226.

[32] HE Y L, WANG Z C, CHEN J, *et al*. EvoPrompt-ReID: a bilevel optimization framework for prompt-encoder co-evolution in image re-identification [C]//*Proceedings of the* 2026 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Denver: IEEE, 2026: 6758-6767.

[33] 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*, 2020, 21(1): 410-420.

[34] SUN K, PANG X Y, ZHENG M F, *et al*. Heterogeneous context interaction network for vehicle re-identification [J]. *Neural Networks*, 2024, 169: 293-306.

[35] ZHU J Y, PARK T, ISOLA P, *et al*. Unpaired image-to-image translation using cycle-consistent adversarial networks [C]//*Proceedings of the* 2017 *IEEE International Conference on Computer Vision*. Venice: IEEE, 2017: 2242-2251.

[36] XU Y, RONG L L, ZHOU X L, *et al*. Cross-domain evaluation for vehicle re-identification [C]//*Proceedings of the* 2021 *IEEE 6th International Conference on Cloud Computing and Big Data Analytics* (*ICCCBDA*). Chengdu: IEEE, 2021: 474-477.

[37] PENG J J, WANG H B, XU F Q, *et al*. Cross domain knowledge learning with dual-branch adversarial network for vehicle re-identification [J]. *Neurocomputing*, 2020, 401: 133-144.

作者简介：



徐萌兮，女，硕士，讲师，2014年于南京理工大学获得硕士学位，主要从事计算机视觉、图像处理方面的研究。E-mail: 880186@cwxu.edu.cn



胡立发，男，博士，研究员，2003年于东北大学获得博士学位，主要从事液晶自适应光学的研究。E-mail: hulifa@jiangnan.edu.cn