

文章编号:1007-2780(2026)05-0675-12

基于多机制融合增强的跨模态行人重识别方法

涂明君¹, 林国军^{1,2*}, 冯 凯¹, 马 润^{1,2}

(1. 四川轻化工大学 自动化与信息工程学院, 四川 宜宾 644000;

2. 智能感知与控制四川省重点实验室, 四川 宜宾 644000)

摘要:针对现有基于 ViT 的跨模态行人重识别方法存在的注意力噪声干扰、缺乏精细化空间关系建模及计算冗余等问题,本文提出了一种多机制融合增强的跨模态行人重识别方法。该方法以 ViT 为基准模型,采用多维特征融合策略提升表达能力。首先,引入差分注意力(DAM)模块,通过计算两组平行注意力图的差异,有效抑制背景中高频出现的共性噪声,显著提升模型对行人主体区域的关注度;其次,设计层级渐进的关系感知模块(RAM),自适应进行由局部到全局的空间关系建模,有效克服了 ViT 在处理遮挡和姿态变化时易丢失空间结构信息的局限;最后,融合交叉窗口自注意力(CSWA)机制,通过将特征图划分为水平和垂直条带进行并行自注意力计算,在将计算复杂度降至线性级的同时,高效捕捉长距离全局依赖。实验结果表明,在 MSMT17 数据集上, mAP 和 Rank-1 分别达到 68.1% 和 86.3%;在 Occluded-Duke 数据集上, mAP 和 Rank-1 分别提升了 5.5% 和 6.3%;在 Market-1501 与 DukeMTMC-reID 数据集上也取得了优于传统 ViT 的检索性能。综合结果表明,本方法有效解决了传统 ViT 模型在行人重识别任务中面临的特征发散和空间信息丢失问题,显著增强了模型在遮挡和复杂背景下的鲁棒性与泛化能力。

关键词:行人重识别;跨模态;ViT;DAM;CSWA;RAM

中图分类号:TP391.4 文献标识码:A doi:10.37188/CJLCD.2026-0043 CSTR:32172.14.CJLCD.2026-0043

Cross-modal pedestrian re-identification method enhanced by multi-mechanism fusion

TU Mingjun¹, LIN Guojun^{1,2*}, FENG Kai¹, MA Run^{1,2}

(1. College of Automation and Information Engineering, Sichuan University of Light Industry,
Yibin 644000, China;

2. Sichuan Provincial Key Laboratory of Intelligent Perception and Control,
Yibin 644000, China)

Abstract: To tackle attention noise, insufficient fine-grained spatial modeling, and computational redundancy in current Vision Transformer (ViT)-based cross-modal person re-identification (ReID) methods, a multi-mechanism fusion-enhanced framework is proposed. Building upon the ViT backbone, the proposed method

收稿日期:2026-03-17;修订日期:2026-04-10.

基金项目:国家自然科学基金(No.42405145);四川省科技厅项目(No. 2024YFHZ0026);四川省自然科学基金(No. 2025ZNSFSC0477, No. 2024NSFSC2042)

Supported by National Natural Science Foundation of China(No. 42405145);Project of Science and Technology Department of Sichuan Province (No. 2024YFHZ0026); Natural Science Foundation of Sichuan Province (No. 2025ZNSFSC0477, No. 2024NSFSC2042)

*通信联系人, E-mail:tiemujian123400@163.com

leverages a multi-dimensional feature fusion strategy to optimize feature representation. Firstly, a Difference Attention Module(DAM) is introduced to compute the discrepancy between parallel attention maps, effectively filtering out high-frequency background noise and significantly enhancing the model's focus on pedestrian targets. Secondly, a hierarchical Relation-Aware Module(RAM) is designed to adaptively conduct progressive local-to-global spatial relationship modeling. This design effectively overcomes ViT's inherent limitation of losing spatial structural information when facing occlusions and severe pose variations. Finally, a Cross-Strip Window Attention(CSWA) mechanism is integrated. By dividing the feature maps into horizontal and vertical strips for parallel self-attention computation, the method reduces computational complexity to linear time while efficiently capturing long-range global dependencies. The superiority of the proposed method is demonstrated through extensive experiments. On the MSMT17 dataset, an mAP of 68.1% and a Rank-1 accuracy of 86.3% are achieved. On the challenging Occluded-Duke dataset, mAP and Rank-1 are improved by 5.5% and 6.3%, respectively. Furthermore, the standard ViT baseline is outperformed on both Market-1501 and DukeMTMC-reID datasets. These results indicate that the proposed framework successfully overcomes feature divergence and spatial information loss, thereby boosting robustness and generalization under occlusions and complex backgrounds.

Key words: person re-identification; cross-modal; ViT; DAM; CSWA; RAM

1 引言

行人重识别(ReID)旨在实现无重叠视域下的跨摄像头身份精准匹配。在实际工程部署中,即便搭配全向视觉定位等先进硬件系统^[1],行人目标仍易受光照剧变、遮挡及背景杂波等复杂视觉退化因素干扰^[2-3],导致传统方法难以提取高鲁棒性特征。

早期基于部件切分(如PCB)的模型^[4]或常规卷积神经网络(CNN)^[5],受限于局部卷积特性,难以捕获身体部位间的长程语义依赖^[6],也无法应对复杂场景中分辨率偏低等内在瓶颈^[7]。针对上述缺陷,研究者引入端到端目标检测 Transformer(如 DETR)^[8]以获取全局信息,或融合多尺度特征改进 SSD 算法,提升拥挤场景下小目标的定位精度^[9]。

随着研究不断发展,视觉 Transformer(ViT)^[10]凭借全局自注意力机制,构建了特征提取新范式。该全局架构虽具优势,但夜间等低照度场景仍需强化局部边缘与结构细节特征。为此,陆燕等^[11]改进 RT-DETR 算法,引入双阶段注意力与多尺度特征融合,增强弱光模糊目标的特征表征能力。ReID 领域中,TransReID^[12]率先采用纯 ViT 架构,显著提升识别性能。同时,王栋等人^[13]设计的 CORM-YOLO 算法,借助链式思考网络,丰富弱光行人图像的语义与纹理信息。针对跨域泛化缺陷,Dai

等人提出相关性感知混合专家模型^[14],有效缓解域偏移引发的性能衰退。此外,融合多头注意力的密集检测模型^[15]解决了目标重叠漏检问题;可变形 DETR^[16]进一步提升了复杂拓扑场景的特征聚合效率。

近年来,融合高层语义的跨模态预训练模型(如 CLIP-ReID^[17])有效提升了跨模态匹配精度。然而,标准 ViT 在该任务中仍存在背景噪声干扰、计算复杂度呈平方级增长的双重缺陷,尽管 Swin Transformer 通过局部窗口注意力机制^[18]控制了计算开销,却不可避免地限制了模型感受野范围。

针对上述局限,本文提出多机制融合增强的跨模态行人重识别方法。其核心创新为:首先,引入差分注意力(DAM)自适应抑制背景噪声并增强前景显著性^[19];其次,设计关系感知模块(RAM)显式建模空间逻辑,填补遮挡区域的语义缺失^[20];最后,融合十字交叉窗口机制(CSWA),在降维至线性计算量的同时扩展感受野,高效捕获人体细长的拓扑结构特征^[21],以实现复杂环境下的精准行人重识别。

2 多机制融合增强方法

本文选用标准的 ViT-B/16 作为骨干网络,其架构如图 1 所示。模型首先将输入图像切分为

N 个固定大小为 16×16 的图像块(Patch),通过线性投影映射为一维向量序列,并嵌入可学习的位置编码以补偿切片过程中丢失的空间位置信息。随后,该特征序列被输入至 Transformer 编码器中,通过多层多头自注意力(MHSA)与前馈神经网络(MLP Block)的交替作用实现深层特征交互。

编码器的内部结构如图 2 所示,通过层归一化(LayerNorm)与残差连接(Residual Connection)确保了深层网络中梯度的稳定传播与特征的高效聚合。在行人重识别流程中,通常提取编码器输出的分类令牌([CLS] Token)作为行人的高层语义特征表示,进而实现精准的身份度量。

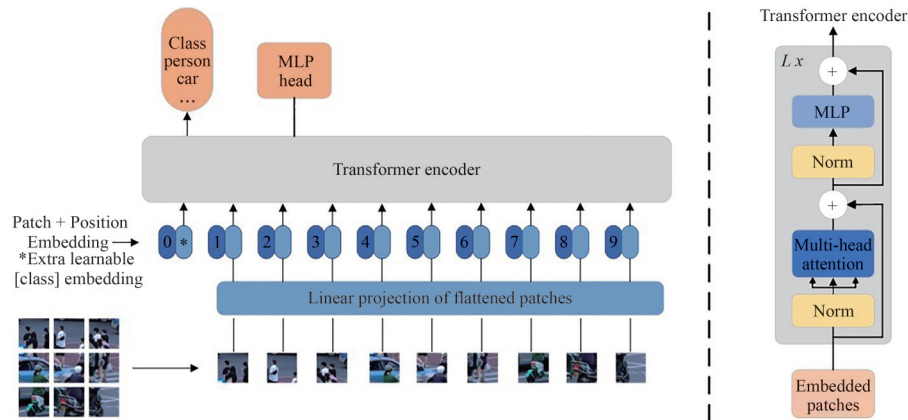


图 1 ViT-B/16 网络结构图

Fig. 1 ViT-B/16 architecture diagram

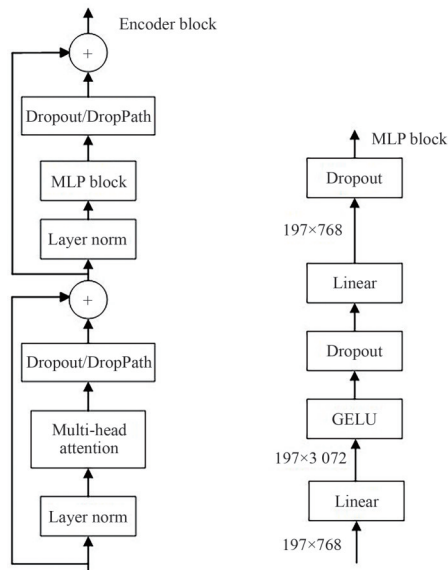


图 2 Transformer Encoder 网络结构图

Fig. 2 Structure diagram of Transformer Encoder network

然而,标准 ViT 在跨模态识别与复杂遮挡场景中存在显著缺陷:全局注意力易受背景噪声干扰,造成特征判别力衰减;同时,其固定空间建模策略难以精准捕获行人狭长形态的局部细粒度特征。针对上述问题,本文提出多机制融合增强方法(图 3):采用差分注意力(DAM)抑制背景噪

声,利用关系感知模块(RAM)自适应重构空间拓扑,嵌入十字交叉窗口自注意力(CSWA),在降低计算复杂度的同时高效捕捉长距离全局依赖。

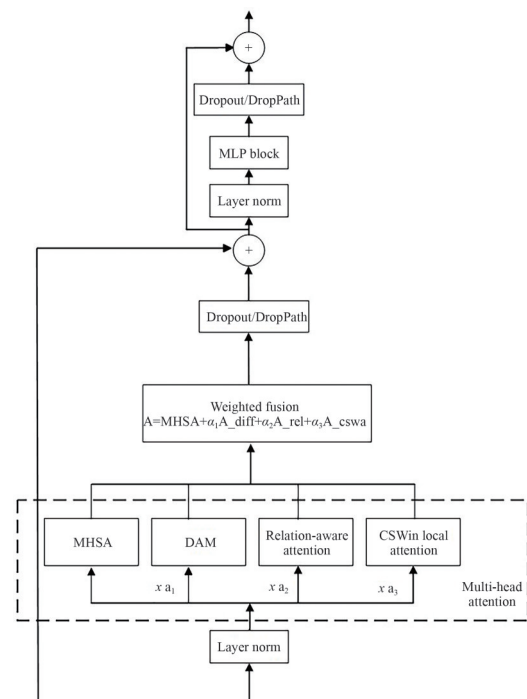


图 3 改良网络结构图

Fig. 3 Structure diagram of the improved netwo

2.1 DAM 模块

标准 ViT 的多头自注意力机制在复杂场景中, 易受 Softmax 平滑权重分配策略影响, 被建筑物、相似纹理等背景杂波干扰, 产生“注意力噪声”, 削弱遮挡条件下的特征提取鲁棒性。密集人群计数等视觉任务研究表明, 嵌入双通道注意力机制可有效抑制背景干扰、增强目标特征^[22]。

基于上述跨任务共性, 本文在 ViT 骨干网络中引入差分注意力 (DAM) 模块。有别于传统仅关注特征相似性的计算方式, DAM 采用差分去噪策略: 通过计算两组互补注意力图 (一组捕获含噪的宽泛上下文, 另一组聚焦关键信号) 的差异, 自动抵消背景噪声。该设计在不显著增加计算开销的前提下, 生成更稀疏、更精准的注意力分布, 显著提升模型对行人核心特征的感知能力。

该机制的具体执行流程如下: 假设输入特征序列为 $X \in \mathbb{R}^{N \times d}$, 其中 N 表示图像块 (Patch) 的数量, d 为特征维度。首先, 通过线性投影生成查询 (Query, Q)、键 (Key, K) 和值 (Value, V) 向量。为了构建差分机制, DAM 将查询向量 Q 和键向量 K 在通道维度上均分为两组, 分别记为 $\{Q_1, Q_2\}$ 和 $\{K_1, K_2\}$ 。具体公式如下:

$$[Q_1; Q_2] = XW^Q, \quad (1)$$

$$[K_1; K_2] = XW^K, \quad (2)$$

$$V = XW^V, \quad (3)$$

其中: W^Q, W^K, W^V 为可学习的投影矩阵, $Q_1, Q_2, K_1, K_2 \in \mathbb{R}^{N \times \frac{d}{2}}$ 。

利用切分后的两组向量, 分别计算独立的注意力图 A_1 和 A_2 。为适应通道减半, 缩放点积的缩放因子相应调整 $\sqrt{\frac{d}{2}}$ 以适应通道减半的变化:

$$A_1 = \text{softmax} \left(\frac{Q_1 K_1^T}{\sqrt{\frac{d}{2}}} \right), \quad (4)$$

$$A_2 = \text{softmax} \left(\frac{Q_2 K_2^T}{\sqrt{\frac{d}{2}}} \right). \quad (5)$$

为消除冗余背景干扰, DAM 引入可学习标量 λ , 通过差分操作生成最终的注意力权重。输出特征 $\text{DiffAttn}(X)$ 的聚合过程表示为:

$$\text{DiffAttn}(X) = (A_1 - \lambda A_2)V, \quad (6)$$

式中, 参数 λ 起到了至关重要的平衡作用, 它自适

应地控制第二组注意力分布 (即噪声估计) 的扣除强度:

$$\lambda = \exp(\lambda_{q_1} \lambda_{k_1}) - \exp(\lambda_{q_2} \lambda_{k_2}) + \lambda_{\text{init}}, \quad (7)$$

其中: $\lambda_{q_1}, \lambda_{k_1}, \lambda_{q_2}, \lambda_{k_2}$ 为可学习的权重参数, λ_{init} 为初始超参数。通过从主信号 A_1 中剥离冗余的背景关联 A_2 , 该机制显著抑制了环境噪声响应, 生成稀疏且聚焦于前景行人的注意力权重。

上述机制可扩展为多头差分注意力 (Diff MHA)。对于第 i 个注意力头, 其内部独立执行如下差分操作:

$$\text{DiffAttn}_i(X) = \left(\text{softmax} \left(\frac{Q_{1,i} K_{1,i}^T}{\sqrt{d_h/2}} \right) - \lambda \text{softmax} \left(\frac{Q_{2,i} K_{2,i}^T}{\sqrt{d_h/2}} \right) \right) V_i, \quad (8)$$

其中, $Q_{1,i}, Q_{2,i}, K_{1,i}, K_{2,i} \in \mathbb{R}^{N \times (d_h/2)}$ 。最终将所有头 h 的输出拼接, 并通过线性层 W^O 进行融合:

$$\text{DAM}(X) =$$

$$[\text{DiffAttn}_1(X) \parallel \dots \parallel \text{DiffAttn}_h(X)] W^O. \quad (9)$$

DAM 模块通过引入差分机制, 从根本上解决了传统 ViT 在复杂背景下注意力分散的问题。它不仅保留了 Transformer 全局建模的优势, 更赋予了模型对细粒度特征的敏锐挖掘能力, 在处理行人重识别中的遮挡、姿态变化及相似背景干扰时, 展现出优越的鲁棒性。

2.2 RAM 模块

空间拓扑结构的衰减是传统 ViT-B/16 在行人重识别任务中的核心缺陷。标准自注意力机制高度依赖纹理相似度, 且深层网络中位置编码信息趋于平滑, 致使模型弱化对行人刚性体态结构的感知能力。空间语义缺失易导致外观相似但语义不匹配区域产生错误关联, 造成识别精度下降。虽有研究证实, 联合局部显著特征与全局信息可增强遮挡场景下的抗干扰性能^[23], 但传统 ViT 仍未建立针对拓扑关系的显式建模机制。

为突破模型在结构感知层面的局限, 本文受 LP-DETR 启发引入 RAM 模块。该模块旨在将显式几何先验 (Geometric Prior) 融入注意力计算, 通过“内容-几何”双流机制对特征交互进行重标定 (Re-calibration)。

与 DAM 模块的去噪侧重点不同, RAM 模块着重挖掘 Patch 间的相对位置关系以调制注意力

分布,其网络结构如图4所示。模块以 Patch 特征序列为输入生成查询(Q)、键(K)和值(V)向量。区别于标准自注意力, RAM在 Softmax 归一化阶段显式注入了外部几何先验:由公式(11)生成的几何权重矩阵(Geometric Weight)直接参与注意力计算路径,在“Weighted Softmax”单元内与缩放点积内容分数(Scaled Dot-Product)实现深度融合。

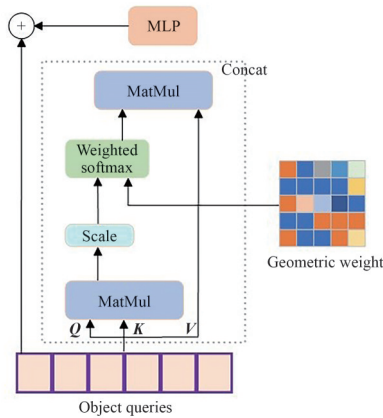


图4 RAM网络结构图

Fig. 4 Structure diagram of RAM encoder network

在运算逻辑方面, Weighted Softmax 单元执行了公式(9)中的运算, 利用几何权重对内容注意力进行逐项调制。调制后的注意力图随即与值向量(V)相乘, 经过多头拼接(Concat)与线性映射(MLP)层处理后, 通过残差连接(Residual Connection)与原始特征相加, 输出融合了空间结构信息的增强特征。

将输入的特征序列标记为 $X \in \mathbb{R}^{N \times C}$, 其中 N 为 Token 数量, C 为通道维度。不同于仅依赖查询(Q)和键(K)点积的传统注意力, 而是通过引入几何权重 w^p 来调制内容权重 w^c 。对于第 i 个 Token 和第 j 个 Token, 其融合后的关系感知注意力权重 w_{ij}^r 计算公式如下:

$$w_{ij}^r = \frac{w_{ij}^p \cdot \exp(w_{ij}^c)}{\sum_{k=1}^N w_{ik}^p \cdot \exp(w_{ik}^c)}, \quad (10)$$

其中, 内容权重 w_{ij}^c 为缩放点积注意力的标准形式:

$$w_{ij}^c = \frac{(W^Q x_i)^T (W^K x_j)}{\sqrt{d_k}}, \quad (11)$$

式中, W^Q 和 W^K 为可学习的投影矩阵, d_k 为缩放因子。

针对 ViT-B/16 的网格化特征结构, 对 LP-

DETR 中基于预测框的几何度量进行了适配。定义 Patch_i 和 Patch_j 之间的相对空间坐标为 $R(i, j) = (\Delta x_{ij}, \Delta y_{ij})$, 其中 Δx_{ij} 和 Δy_{ij} 分别表示 Token_i 和 Token_j 在网格中的归一化坐标差值。该相对距离能够精确描述行人身体部位间的空间邻接性与方向性。几何权重的生成过程如下:

$$w_{ij}^p = W_G \cdot \text{ReLU} \left(\text{MLP} \left(\text{PosEnc} \left(R(i, j) \right) \right) \right), \quad (12)$$

式中: $\text{PosEnc}(\cdot)$ 表示正弦位置编码(Sinusoidal Encoding), 用于将低维坐标差映射到高维空间; MLP 和 W_G 则是用于提取空间几何特征的可学习变换层。

2.3 CSWA 模块

传统全局自注意力机制存在平方级计算复杂度, 而现有局部窗口注意力虽能削减计算开销, 却会显著限制模型感受野范围。为实现计算效率与全局感受野的最优权衡, 本文引入十字交叉窗口自注意力(CSWA)模块, 其网络结构与工作机制如图5所示。

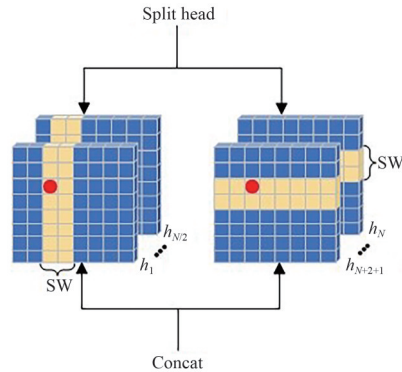


图5 CSWA网络结构图

Fig. 5 Structure diagram of CSWA encoder network

该模块沿用了标准 Transformer 的残差架构(依次经由层归一化、CSWA 模块与多层感知机处理), 以缓解梯度消失并保障特征表征的稳定性。其核心创新在于特征分组与并行条带处理机制: 在多头自注意力计算中, CSWA 摒弃了传统方形窗口, 将总数为 N 的注意力头均分为两组。其中, 前 $N/2$ 个注意力头专注于水平条带(Horizontal stripes)的局部自注意力计算, 后 $N/2$ 个注意力头则专注于垂直条带(Vertical stripes)的计算。该正交并行架构可在控制计算开销的前提下, 有效

捕获融合全局感知信息的十字形上下文特征。

在具体的数学实现上,设输入特征图为 $X \in \mathbb{R}^{H \times W \times C}$,条带宽度为 SW 。对于水平条带自注意力计算,输入特征 X 被均匀划分为 M 个互不重叠的水平条带 $X = [X_h^1, X_h^2, \dots, X_h^M]$ 。以第 n 个注意力头为例(其中 $1 \leq n \leq N/2$),其在第 i 个水平条带内的查询(Q)、键(K)和值(V)矩阵的映射及注意力输出 Y_n^i 计算过程如下:

$$Q_n^i = W_n^Q X_h^i, \quad (13)$$

$$K_n^i = W_n^K X_h^i, \quad (14)$$

$$V_n^i = W_n^V X_h^i, \quad (15)$$

$$Y_n^i = \text{Softmax}\left(\frac{Q_n^i (K_n^i)^T}{\sqrt{d_n}}\right) V_n^i, \quad (16)$$

其中: $X_h^i \in \mathbb{R}^{sw \times W \times C}$ 表示第 i 个水平条带的特征图; $W_n^Q \in \mathbb{R}^{C \times d_n}$ 、 $W_n^K \in \mathbb{R}^{C \times d_n}$ 和 $W_n^V \in \mathbb{R}^{C \times d_n}$ 分别表示第 n 个头的权重投影矩阵; $d_n = \frac{C}{N}$ 为通

道维度的缩放因子,用于防止 Softmax 梯度消失。将所有 M 个水平条带的输出进行拼接,即可得到第 n 个头的完整水平注意力输出 $\text{H-Attention}_n(X)$:

$$\text{H-Attention}_n(X) = [Y_n^1, Y_n^2, \dots, Y_n^M]. \quad (17)$$

同理,对于后 $N/2$ 个注意力头,输入特征被划分为 S 个垂直条带 $X = [X_v^1, X_v^2, \dots, X_v^S]$,其垂直注意力输出 $\text{V-Attention}_n(X)$ 的计算方式与水平方向完全对称:

$$Q_n^i = W_n^Q X_v^i, \quad K_n^i = W_n^K X_v^i, \quad V_n^i = W_n^V X_v^i, \quad (18)$$

$$Y_n^i = \text{Softmax}\left(\frac{Q_n^i (K_n^i)^T}{\sqrt{d_n}}\right) V_n^i, \quad (19)$$

$$\text{V-Attention}_n(X) = [Y_n^1, Y_n^2, \dots, Y_n^S]. \quad (20)$$

在分别完成水平与垂直条带的自注意力计算后,CSWA 模块将上述两组特征沿通道维度进行拼接,并通过线性变换矩阵 $W^o \in \mathbb{R}^{C \times C}$ 融合不同头的注意力信息:

$$h_n = \begin{cases} \text{H-Attention}_n(X), & n = 1, 2, \dots, \frac{N}{2} \\ \text{V-Attention}_n(X), & n = \frac{N}{2} + 1, \frac{N}{2} + 2, \dots, N \end{cases}, \quad (21)$$

$$\text{CSWin-Attention}(X) = \text{concat}(h_1, h_2, \dots, h_N) W^o. \quad (22)$$

结合上述核心注意力机制,整个 CSWin Transformer Block 在第 l 层的完整前向传播过程可定义为:

$$\hat{X}^l = \text{CSWin-Attention}(\text{LN}(X^{l-1})) + X^{l-1}, \quad (23)$$

$$X^l = \text{MLP}(\text{LN}(\hat{X}^l)) + \hat{X}^l. \quad (24)$$

CSWA 模块通过划分水平与垂直条带,有效打破了传统局部窗口的感受野局限。拼接后的特征深度融合了横纵双向的上下文信息,在不增加额外计算开销的前提下,显著拓宽了 ViT-B/16 的感受野,为后续的高精度特征匹配与检索提供了更为全面、鲁棒的特征表征。

3 实验结果与分析

3.1 数据集与实验环境

本实验的环境配置如表 1 所示,训练时采用统

一的参数,实验参数配置如表 2 所示。实验在 4 个行人重新识别数据集上对此方法进行评估,包括 MSMT17 (Wei *et al.*, 2018)、Market-1501 (Zheng *et al.*, 2015)、DukeMTMC-reID (Ristani *et al.*, 2016)、Occluded-Duke (Miao *et al.*, 2019)。4 个数据集的详细信息如表 3 所示,本实验遵循常规方法,采用 Rank-1(R1)处的累积匹配特征(CMC)和平均平均精度(MAP)评估性能。

表 1 实验环境配置

Tab. 1 Experimental environment configuration

参数	配置
系统环境	Windows11
显卡	NVIDIA RTX 5070ti
深度学习框架	Pytorch2.8
加速环境	CUDA12.9
语言	Python3.10.19

Occluded-Duke 是专门针对遮挡行人重识别构建的数据集,其查询集包含大量受行人、车辆或静态障碍物严重遮挡的样本。该数据集涵盖局

部区域缺失、几何畸变与姿态偏移等复杂挑战,是评估模型遮挡鲁棒性与局部特征解析能力的核心基准。

表2 实验参数配置

Tab. 2 Experimental parameter configuration

参数	配置
图片大小(Image size)	256×128
训练迭代次数(Epochs)	60
每次更新的批次大小(Batch size)	64
线程数(Workers)	8
优化器(Optimizer)	AdamW
基础学习率(Lr)	0.000 005
部分权重衰减系数(Decay)	0.000 1

表3 实验数据集

Tab. 3 Experimental dataset

Dataset	Image	ID	Cam+View
MSMT17	126 441	4 101	15
Market-1501	32 668	1 501	6
DukeMTMC-reID	36 411	1 404	8
Occluded-Duke	35 489	1 404	8

MSMT17是当前规模最大、场景最复杂的重识别基准之一,涵盖多时段、多视角的室内外密集监控网络。该数据集蕴含由昼夜光照突变、复杂天气及大范围视角更迭引发的严重环境域偏移,能够严苛检验模型在超大规模跨域场景下的泛化能力、特征判别力与系统稳定性。对比Market-1501与DukeMTMC-reID等早期基准,MSMT17在独立身份数量、摄像头网络拓扑规模及图像视觉复杂度等维度均实现了大幅跃升,故被公认为现阶段行人重识别领域极具挑战性的评估基准之一。该数据集能够全方位、严苛地检验模型在复杂开放场景下的跨域泛化能力、特征判别力及系统稳定性。

3.2 评价指标

为客观评估所提方法在行人重识别任务中的检索性能,本文采用累积匹配特征曲线(Cumulative Matching Characteristic, CMC)和平均精度均值(mean Average Precision, mAP)作为主要评价指标。其中,CMC用于衡量查询图像在检索结果中的匹配排名情况,而mAP用于综合评价模型在多匹配目标条件下的整体检索性能。

在行人重识别任务中,给定一张查询图像

(Query),模型需要在测试集图库(Gallery)中检索与其身份相同的图像。CMC曲线描述了正确匹配样本出现在检索结果前 k 位中的概率,通常报告Rank-1、Rank-5和Rank-10等指标。其中,Rank-1表示正确匹配样本位于检索结果第一位的比例,是衡量模型识别能力的重要指标,其计算公式定义如下:

$$\text{Rank-1} = \frac{1}{Q} \sum_{q=1}^Q \delta(r_q = 1), \quad (25)$$

式中: Q 为查询集中的样本总数; r_q 表示针对第 q 个查询样本,系统检索到的首个正确匹配图像的排名; $\delta(\cdot)$ 为指示函数,当条件成立($r_q = 1$)时取值为1,否则取值为0。

对于单个查询样本,其平均精度(Average Precision, AP)定义为:

$$\text{AP} = \sum_{k=1}^N P(k) \Delta R(k), \quad (26)$$

其中: $P(k)$ 表示在第 k 个检索位置处的精度, $\Delta R(k)$ 表示对应位置召回率的变化量, N 为图库中样本总数。AP能够反映模型在不同召回水平下的综合排序性能。

在整个测试集中,设共有 Q 个查询样本,则平均精度均值(mAP)定义为:

$$\text{mAP} = \frac{1}{Q} \sum_{q=1}^Q \text{AP}_q, \quad (27)$$

其中, AP_q 表示第 q 个查询样本对应的平均精度。

相比仅采用Rank-1指标,mAP能够更全面地衡量模型在存在多个正确匹配样本情况下的整体检索性能。因此,本文在所有实验中同时采用Rank-1和mAP作为主要评价指标,以保证实验结果的全面性与可靠性。

3.3 对比实验

为验证所提算法的有效性,本节在4个主流数据集上将其与10种前沿方法(涵盖CNN与Transformer架构)进行对比,定量评估结果详见表4。

在常规数据集中,本算法综合性能优势显著。相较于基线模型TransReID,本算法在MSMT17上的mAP与Rank-1分别提升2.0%与1.9%(峰值达68.1%与86.3%);在Market-1501数据集上,上述两项指标分别增至87.8%与95.0%,较基准模型分别取得1.4%与1.7%的性能增益;而在DukeMTMC-reID数据集上,本算法的mAP与

Rank-1 分别达到 81.7% 与 90.6%,较 TransReID 分别提升 1.7% 与 1.8%。此外,对比表现优异的 AAformer 与 DCAL 方法,本算法在复杂挑战

性极高的 MSMT17 数据集上,mAP 分别高出 4.9% 与 4.1%,展现出卓越的特征提取与身份匹配效能。

表 4 各方法在 ReID 数据集上的对比实验结果
Tab. 4 Comparison experimental results of different methods on ReID dataset

Methods	MSMT17		Market-1501		DukeMTMC-reID		Occluded-Duke	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
MGN	—	—	86.9	95.7	78.4	88.7	—	—
OSNeT	52.9	78.7	84.9	94.8	73.5	88.6	—	—
SAN	55.7	79.2	88.0	96.1	75.5	87.9	—	—
ALDER	59.1	82.5	88.9	95.6	78.9	89.9	—	—
LTReID	58.6	81.0	89.0	95.9	80.4	90.5	—	—
DRL-Net	55.3	78.4	86.9	94.7	76.6	88.1	50.8	65.0
CNN	60.7	82.1	88.1	94.7	79.3	88.6	47.4	54.2
AAformer	63.2	83.6	87.7	95.4	80.0	90.1	58.2	67.0
DCAL	64.0	83.1	87.5	94.7	80.1	89.0	—	—
TransReID(baseline)	66.1	84.4	86.4	93.3	80.0	88.8	53.5	60.8
Ours	68.1	86.3	87.8	95.0	81.7	90.6	59.0	67.1

注:粗体数字表示最优值

在评估遮挡鲁棒性的 Occluded-Duke 数据集上,所提算法的优势更为突出(mAP 达 59.0%)。该结果较专为遮挡设计的 DRL-Net 大幅提升 8.2%,且 mAP 与 Rank-1 较同类注意力模型 AAformer 分别提升 0.8% 与 0.1%。尤为显著的是,相较于基线模型 TransReID,本算法实现了 5.5% 与 6.3% 的大幅增长(提升幅度远超常规数据集)。上述实验结果表明,所提方法在大面积遮挡、强背景干扰等极端退化场景中,具有优异的检索鲁棒性与泛化性能。

3.4 消融实验

为验证差分注意力(DAM)、十字交叉窗口自注意力(CSWA)及关系感知(RAM)模块的有

效性,本节以标准 ViT 为基线在 4 个数据集上进行消融实验,定量评估结果详见表 5。

由表 5 可知,单模块引入均带来稳定的性能增益:DAM 使 Occluded-Duke 的 mAP 提升 2.8%,证明其有效抑制了背景噪声干扰;CSWA 使 MSMT17 的 mAP 提升 0.8%,印证其成功捕获了长距离空间依赖;RAM 则显著改善了 Duke 系列数据集的 Rank-1 指标,验证了层级关系建模在弥补空间结构信息缺失方面的作用。

双模块级联展现出显著的协同增强效应:DAM 与 CSWA 组合使 MSMT17 的 mAP 升至 67.7%,表明前端去噪优化了全局特征的提取质量;DAM 与 RAM 联合使 Occluded-Duke 的 Rank-1

表 5 消融实验结果
Tab. 5 Ablation experiment results

CSWA	DAM	RAM	MSMT17		Market-1501		DukeMTMC-reID		Occluded-Duke	
			mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
—	—	—	66.1	84.4	86.4	93.3	80.0	88.8	53.5	60.8
✓			66.3	84.6	86.4	93.7	80.2	89.5	56.3	65.9
	✓		66.9	85.1	86.2	93.6	80.1	89.3	54.4	62.0
		✓	67.3	84.9	87.3	93.9	80.8	90.0	56.7	64.2
✓	✓		67.7	85.7	87.5	94.3	81.5	90.2	58.8	66.1
✓		✓	67.5	85.4	87.4	94.2	81.3	90.1	58.2	66.5
	✓	✓	67.6	85.6	87.6	94.3	81.4	90.3	57.8	65.5
✓	✓	✓	68.1	86.3	87.8	95.0	81.7	90.6	59.0	67.1

达66.5%,体现了应对遮挡场景的高度互补性;CSWA与RAM结合则在Market-1501上取得87.6%的mAP,验证了全局依赖捕获与空间重建的双重增益。

在三模块协同驱动下,最终构建的融合模型在最具挑战性的Occluded-Duke数据集上,mAP与Rank-1分别达到59.0%与67.1%,较基准模型分别实现了5.5%与6.3%的显著跨越;同时,在MSMT17上的mAP增至68.1%,在Market-1501上的Rank-1则高达95.0%。综上,各核心组件针对遮挡干扰、复杂背景及姿态畸变均提供了行之有效的优化策略,其深度融合构建了涵盖“特征净化-全局建模-空间补全”的有机多机制增强体系,从根本上提升了模型在复杂开放环境下的鲁棒性与泛化能力。

3.5 改进算法可视化效果

为了更直观地验证本文提出的多机制融合增强策略在应对复杂遮挡与背景干扰时的有效性,本节采用GradCAM++技术对网络深层特征图进行类激活映射(Class Activation Mapping)可视化分析,以清晰展现模型对行人关键区域的聚焦能力。

图6的可视化结果揭示了模块内部特征传输与差分去噪机制的作用过程。图中生成两组互

补的中间层注意力分布矩阵:第一组(图6(b)“宽泛注意力”)用于捕获全局上下文特征,激活区域同时包含行人目标有效信号与复杂背景噪声;第二组(图6(c)“噪声注意力”)自适应聚焦于背景中的高频共性冗余噪声。通过对两组注意力图实施差分运算,模型可从全局上下文特征中分离背景噪声分量,实现背景干扰的有效抑制。最终融合输出的增强特征(图6(d)“最终特征图”)验证了该机制的合理性,证实其可显著降低环境干扰的不利影响、重新校准特征权重,促使模型精准聚焦于行人核心目标区域。

图7可视化结果揭示了RAM模块自适应建模局部与全局空间关系的特征演化机理。由图7(b)可知,原始特征因缺少显式空间约束,在遮挡条件下易出现特征响应发散问题,无法有效聚焦目标主体结构。RAM模块对多头注意力交互后的特征实施通道维度拼接,通过线性映射层(MLP)完成降维与特征重组,融合多维上下文信息并嵌入显式几何先验,生成图7(c)的残差特征。该特征精准刻画遮挡区域的空间语义补偿关系,其与原始特征融合后得到增强特征(图7(d))。对比图7(b)与图7(d)可证实,RAM模块实现了局部-全局空间关联的自适应重构,有效填补了遮挡导致的语义空缺,显著提升了模型对行人空间拓扑结构的

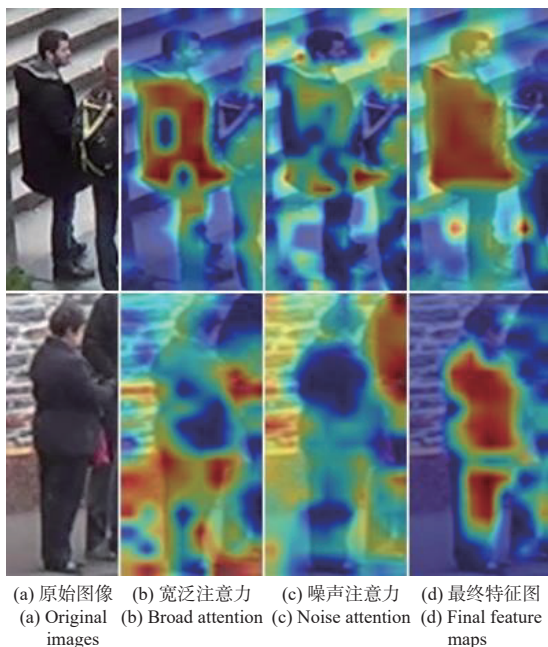


图6 DAM模块差异去噪可视化

Fig. 6 Visualization of differential denoising in DAM

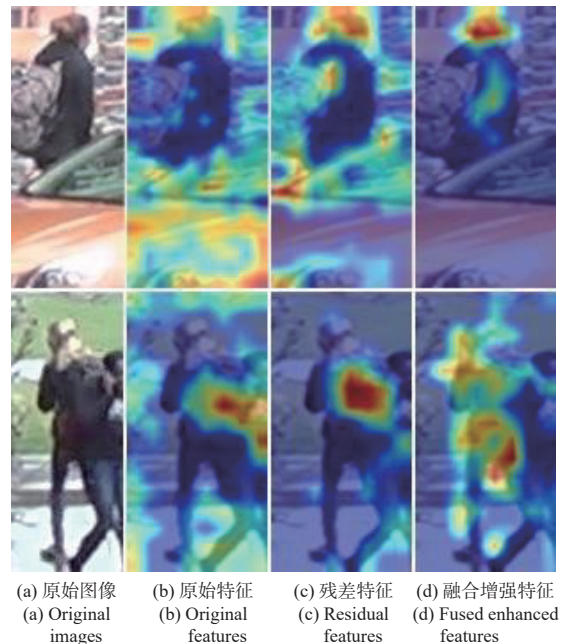


图7 RAM模块残差融合可视化

Fig. 7 Visualization of residual fusion in RAM

完整表征性能。

图 8 可视化对比了水平与垂直条带特征通道维度多头拼接前后的激活响应特征。该模块对输入特征实施正交分条,生成水平、垂直两组独立条带并并行执行自注意力计算。由图 8(b)(水平注意力)与图 8(c)(垂直注意力)可见,两组分支分别建模横向、纵向单向上下文依赖,受条带固有几何约束,二者感受野均存在明显的单视角感知缺陷。相较于拼接前的单向特征响应,图 8(d)(拼接融合特征)直观反映了通道维度多头拼接后的特征增强效果。对比结果表明,水平与垂直并行特征的深度融合可有效消除单维度空间感知盲区,构建全局十字交叉感受野,显著增强模型对行人长程结构依赖及精细化拓扑特征的完整表征能力。

图 9 可视化结果对比了原始图像、基线 ViT 模型与本文算法的注意力热力图差异。基线模型易受背景杂波干扰,在目标遮挡或姿态形变条件下难以有效聚焦行人目标主体,特征输出存在冗余度高、完整性不足的问题。本文算法则表现出显著优越性:DAM 模块可有效滤除高频背景噪声,驱动注意力权重精准聚焦行人主体区域;RAM 模块强化空间结构信息传递,于全局视野下精准捕

捉衣物、背包等高辨识度局部特征;CSWA 机制构建长程全局依赖关系,促使高响应区域连续均匀覆盖行人整体拓扑结构,从根本上突破传统局部窗口感受野碎片化的固有缺陷。

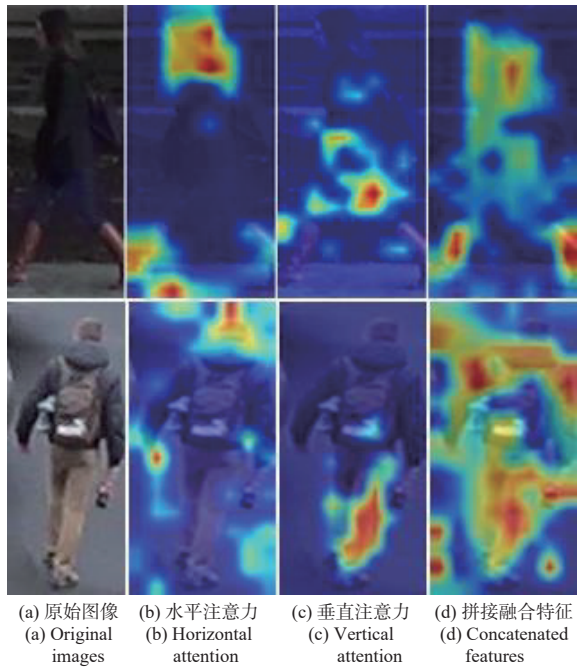


图 8 CSWA 模块特征拼接对比

Fig. 8 Comparison of feature concatenation in CSWA

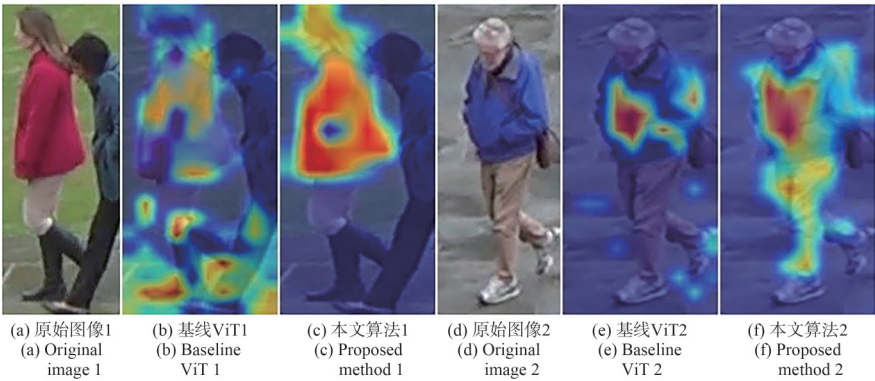


图 9 热力图检索效果对比

Fig. 9 Comparison of heat map retrieval effects

4 结 论

本文提出一种基于多机制融合增强的跨模态行人重识别方法。通过集成差分注意力(DAM)、关系感知(RAM)与十字交叉窗口自注意力(CSWA)模块,本模型有效克服了传统 ViT

在复杂环境下的注意力噪声干扰与空间信息丢失问题。实验表明,该方法在最具挑战性的 Occluded-Duke 数据集上,mAP 和 Rank-1 分别达 59.0% 和 67.1%(较基线提升 5.5% 和 6.3%);在 MSMT17 等主流基准上的检索性能亦显著优于传统 ViT,大幅增强了模型在遮挡和复杂背景

下的鲁棒性。

面向大规模动态监控等真实场景,如何进一步突破模型的计算开销与推理实时性瓶颈是未来的重要方向。近年来,基于光子计算的架构在兼顾高算力与低能耗方面展现出巨大潜力。未来工作拟探索光电融合的计算范式,尝试引入轨道角动量(OAM)全光神经网络^[24],以期在特征编码与降维过程中,既能保证高并发的计算效率,又能

维持超低的能耗水平;同时,最新研究提出的可重构多功能集成光子计算芯片,在单一物理架构内实现了全连接、卷积和循环神经网络的统一构建。该芯片有效支持静态与动态时序任务的协同处理,其面积计算效率达到2.45 TOPS/mm²^[25]。以期在维持特征提取高鲁棒性的同时实现超高速推理,进而满足安防场景严苛的实时部署需求。

参 考 文 献:

- [1] 吴军,王豪爽,单腾飞,等. 基于双自由度转台的全向空间单目视觉室内定位测量方法[J]. 中国光学(中英文), 2024,17(3):605-616.
WU J, WANG H S, SHAN T F, *et al.* Omnidirectional spatial monocular vision indoor localization measurement based on a two-degree-of-freedom rotary platform [J]. *Chinese Optics*, 2024, 17(3): 605-616. (in Chinese)
- [2] 冯霞,杜佳浩,段仪浓,等. 基于深度学习的行人重识别研究综述[J]. 计算机应用研究,2020,37(11):3220-3226,3240.
FENG X, DU J H, DUAN Y N, *et al.* Research on person re-identification based on deep learning [J]. *Application Research of Computers*, 2020, 37(11): 3220-3226, 3240. (in Chinese)
- [3] 段嗣钊,金忠. 一种基于逐类别点到集合距离损失的行人重识别方法[J]. 计算机与数字工程,2025,53(9):2609-2613,2633.
DUAN S Z, JIN Z. A person re-identification method based on class-by-class point-to-set distance loss [J]. *Computer and Digital Engineering*, 2025, 53(9): 2609-2613, 2633. (in Chinese)
- [4] SUN Y F, ZHENG L, LI Y L, *et al.* Learning part-based convolutional features for person re-identification [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43(3): 902-917.
- [5] 周飞燕,金林鹏,董军. 卷积神经网络研究综述[J]. 计算机学报,2017,40(6):1229-1251.
ZHOU F Y, JIN L P, DONG J. Review of convolutional neural network [J]. *Chinese Journal of Computers*, 2017, 40(6): 1229-1251. (in Chinese)
- [6] RAGHU M, UNTERTHINER T, KORNBLITH S, *et al.* Do vision Transformers see like convolutional neural networks? [C]//*Proceedings of the 35th International Conference on Neural Information Processing Systems*. Online: Curran Associates Inc., 2021: 927.
- [7] 童康,吴一全. 基于深度学习的小目标检测基准研究进展[J]. 电子学报,2024,52(3):1016-1040.
TONG K, WU Y Q. Research advances on deep learning based small object detection benchmarks [J]. *Acta Electronica Sinica*, 2024, 52(3): 1016-1040. (in Chinese)
- [8] CARION N, MASSA F, SYNNAEVE G, *et al.* End-to-end object detection with transformers [C]//*Proceedings of the 16th European Conference on Computer Vision (ECCV)*. Glasgow: Springer, 2020: 213-229.
- [9] 张秀再,邱野,沈涛. 基于改进SSD算法的地铁场景小行人目标检测[J]. 计算机研究与发展,2025,62(2): 397-407.
ZHANG X Z, QIU Y, SHEN T. Small pedestrian target detection in subway scene based on improved SSD algorithm [J]. *Journal of Computer Research and Development*, 2025, 62(2): 397-407. (in Chinese)
- [10] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al.* An image is worth 16×16 words: Transformers for image recognition at scale [C]//*Proceedings of the 9th International Conference on Learning Representations*. OpenReview.net, 2021.
- [11] 陆燕,李富,祁铭瑞,等. 基于改进RT-DETR的夜间低光照行人检测算法[J]. 液晶与显示,2025,40(12):1881-1893.
LU Y, LI F, QI M R, *et al.* Improved RT-DETR-based pedestrian detection algorithm for low-light conditions at night [J]. *Chinese Journal of Liquid Crystals and Displays*, 2025, 40(12): 1881-1893. (in Chinese)
- [12] HE S T, LUO H, WANG P C, *et al.* TransReID: Transformer-based object re-identification [C]//*Proceedings*

- of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021: 14993-15002.
- [13] 王栋, 罗亚波, 张峰, 等. CORM-YOLO: 低照度环境下交通行人检测算法[J]. 液晶与显示, 2025, 40(9): 1369-1380.
WANG D, LUO Y B, ZHANG F, *et al.* CORM-YOLO: pedestrian detection algorithm in low-light conditions for traffic environments [J]. *Chinese Journal of Liquid Crystals and Displays*, 2025, 40(9): 1369-1380. (in Chinese)
- [14] DAI Y X, LI X T, LIU J, *et al.* Generalizable person re-identification with relevance-aware mixture of experts [C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville: IEEE, 2021: 16140-16149.
- [15] 王泽宇, 徐慧英, 朱信忠, 等. 基于YOLOv8改进的密集行人检测算法: MER-YOLO[J]. 计算机工程与科学, 2024, 46(6): 1050-1062.
WANG Z Y, XU H Y, ZHU X Z, *et al.* An improved dense pedestrian detection algorithm based on YOLOv8: MER-YOLO [J]. *Computer Engineering & Science*, 2024, 46(6): 1050-1062. (in Chinese)
- [16] ZHU X Z, SU W J, LU L W, *et al.* Deformable DETR: deformable Transformers for end-to-end object detection [C]// *Proceedings of the 9th International Conference on Learning Representations*. Online: OpenReview.net, 2021.
- [17] 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 37th AAAI Conference on Artificial Intelligence*. Washington: AAAI, 2023: 1405-1413.
- [18] LIU Z, LIN Y T, CAO Y, *et al.* Swin Transformer: hierarchical vision transformer using shifted windows [C]// *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal: IEEE, 2021: 9992-10002.
- [19] HAMMOUD H A A K, GHANEM B. DiffCLIP: differential attention meets CLIP [J/OL]. *arXiv*, 2025: 2503.06626.
- [20] KANG Z J, ZHANG Y, DENG X Y, *et al.* LP-DETR: layer-wise progressive relation for object detection [C]// *Proceedings of the 21st International Conference on Advanced Intelligent Computing Technology and Applications*. Ningbo: Springer, 2025: 144-156.
- [21] LIU X, GAO P, YU T, *et al.* CSWin-UNet: Transformer UNet with cross-shaped windows for medical image segmentation [J]. *Information Fusion*, 2025, 113: 102634.
- [22] 陈永, 董珂, 安卓奥博, 等. 密集连接注意力与尺度感知重组增强的人群计数[J]. 光学精密工程, 2024, 32(22): 3395-3408.
CHEN Y, DONG K, AN Z A B, *et al.* Crowd counting method based on dense connection attention and scale perception recombination enhancement [J]. *Optics and Precision Engineering*, 2024, 32(22): 3395-3408. (in Chinese)
- [23] 马晞茗, 李宁, 吴迪, 等. 基于头部与整体信息联合的行人检测算法[J]. 光学精密工程, 2025, 33(14): 2278-2290.
MA X M, LI N, WU D, *et al.* Pedestrian detection algorithm based on joint head and overall information [J]. *Optics and Precision Engineering*, 2025, 33(14): 2278-2290. (in Chinese)
- [24] FANG X Y, HU X N, LI B L, *et al.* Orbital angular momentum-mediated machine learning for high-accuracy mode-feature encoding [J]. *Light: Science & Applications*, 2024, 13(1): 49.
- [25] WANG Y F, LIAO K, ZHANG K, *et al.* Reconfigurable versatile integrated photonic computing chip [J]. *eLight*, 2025, 5(1): 20.

作者简介:



涂明君, 男, 硕士研究生, 2024 年于四川轻化工大学获得学士学位, 主要从事行人重识别技术和图像处理方面的研究。E-mail: tumingjun7769@163.com



林国军, 男, 博士, 副教授, 2014 年于电子科技大学获得博士学位, 主要从事人脸识别、图像处理和计算机视觉方向的研究。E-mail: tiemujian123400@163.com