引用格式：虞智, 刘佳文, 黄智勇, 等. 基于层级化 Token 选择 Transformer 的多模态目标重识别方法 [J]. 电子科技大学学报, 2026, 55(4): 600-612.
Yu Zhi, Liu Jiawen, Huang Zhiyong, et al. Multimodal object re-identification based on hierarchical token selection Transformer[J]. Journal of University of Electronic Science and Technology of China, 2026, 55(4): 600-612. (in Chinese)

基于层级化 Token 选择 Transformer 的多模态目标重识别方法

虞 智，刘佳文，黄智勇*，仲元红

（重庆大学微电子与通信工程学院，重庆 400044）

摘要：多模态目标重识别旨在融合不同模态的互补信息，在视域不重叠的监控场景中实现对同一目标的检索，主要挑战在于缓解跨模态差异。为此，提出一种层级化 token 选择 Transformer（HTSTrans）。HTSTrans 由多模态共享特征学习模块（MSFLM）、层级化 token 选择模块（HTSM）和跨模态对齐约束模块（CACM）构成。MSFLM 使用权重共享的视觉 Transformer（ViT）学习多模态共享特征，在不同模态间初步建立特征一致性，为后续更精细化的特征学习提供统一的表示空间。HTSM 使用层级感知的递进式窗口选择关键 token，基于 token 融合实现对模态特定关键细节特征的提取，再通过多模态特征融合降低跨模态差异。CACM 使用模态间传输约束（ITC）和循环一致性约束（CCC）降低同一目标身份在不同模态空间中的特征差异，并提升模态不变性特征的判别性。实验结果表明，HTSTrans 可有效提升多模态目标重识别性能。

关键词：多模态目标重识别；token 选择；跨模态对齐；Transformer
中图分类号：TP391 **文献标志码：**A DOI: 10.12178/1001-0548.2025230

Multimodal object re-identification based on hierarchical token selection Transformer

YU Zhi, LIU Jiawen, HUANG Zhiyong*, and ZHONG Yuanhong

(School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China)

Abstract: Multimodal object re-identification aims to integrate complementary information from different modals to achieve the retrieval of the same target across non-overlapping surveillances. The primary challenge in this process lies in mitigating the cross-modal discrepancies. To address this issue, a hierarchical token selection transformer (HTSTrans) is proposed. HTSTrans consists of a multi-modal shared feature learning module (MSFLM), a hierarchical token selection module (HTSM), and a cross-modal alignment constraint module (CACM). MSFLM employs a weight-shared vision transformer (ViT) to learn shared multi-modal features, establishing preliminary feature consistency across different modals and providing a unified representation space for subsequent fine-grained feature learning. HTSM utilizes a hierarchically-aware progressive window mechanism to select crucial tokens. Through token fusion, it extracts modal-specific crucial detailed features, followed by multi-modal feature fusion to reduce cross-modal differences. CACM applies an inter-modal transmission constraint (ITC) and a cycle consistency constraint (CCC) to reduce feature discrepancies of the same identity in different modal spaces and enhance the discriminability of modal-invariant features. Experiments demonstrate that HTSTrans effectively improves the performance of multi-modal object re-identification.

Key words: multi-modal object re-identification; token selection; cross-modal alignment; Transformer

传统单模态目标重识别主要针对可见光（RGB）图像，通过学习不同摄像头下同一目标特征实现跨视域匹配[1-2]。然而，光照变化、昼夜转换、天气条件等因素会影响可见光图像质量，导致模型在复杂场景下识别性能大幅下降。为克服这一局限，研究者将目标重识别扩展至多模态场景，利用近红外（near infrared, NIR）和热红外（thermal infrared, TIR）等模态的互补特性，在低照度、夜

收稿日期：2025−11−15
基金项目：国家自然科学基金（62372074）；重庆市自然科学基金（CSTB2023NSCQ-MSX0274）；中央高校基本科研基金（2023CDJKYJH050）
作者简介：虞智，博士生，主要从事多模态目标重识别方面的研究。
*通信作者 E-mail: zyhuang@cqu.edu.cn

间或极端环境下保持目标重识别稳定性。尽管多模态设备能同步采集同一场景下的目标，但由于各模态的成像机理不同，不同模态间差异显著。同时，各模态在目标结构呈现与语义表达上也存在偏差。这种成像差异性和语义不一致性使多模态特征难以直接统一，提升了多模态目标重识别的挑战性[3-5]。

随着深度学习的发展，多模态目标重识别研究聚焦于使用神经网络融合不同模态信息。文献 [6] 提出生成式注意力融合网络以解决多模态目标重识别中的模态融合问题。在输入层面，通过生成过渡模态图像来丰富数据单元。这些图像有效缩小了模态间差异并增强数据单元内部一致性。在特征层面，采用注意力融合对原始数据与过渡模态特征进行了多维度融合以实现特征对齐。文献 [7] 提出跨向一致性网络以解决多模态目标重识别中的双重差异问题。该方法通过跨向中心损失函数将同一目标在不同模态下的特征中心相互拉近以减小模态差异，同时拉近同一样本在不同视图下的特征中心以缓解样本差异。此外，引入自适应层归一化单元降低模态内特征分布差异。文献 [8] 提出解耦子空间架构以解决多模态目标重识别中的特征判别性不足问题。该方法将特征分解为公共成分和独立成分，有效解决了高维特征处理和模态变化带来的挑战，显著提升了多模态目标识别的鲁棒性。文献 [9] 提出基于原型的多样性和完整性学习网络以解决多模态目标重识别在真实场景中面临的光照变化和模态缺失问题。该方法首先通过光照增强和模态特定原型交互构建多样性表征以提升特征稳定性，然后利用原型与实例特征的完整性重建机制处理模态缺失问题。考虑到 Transformer 在特征建模方面的高效性，更多研究将其引入多模态重识别中。文献 [10] 提出一种用于多模态目标重识别的混合视觉 Transformer 以解决单一可见光模态在不良光照条件下的性能受限问题。该架构包括模态特定控制器和模态信息嵌入结构两个模块。模态特定控制器通过配置模态专用层并控制位置嵌入的共享属性，有效缓解了异构模态差异带来的学习困难。模态信息嵌入结构则通过联合学习模态间与模态内信息，显著减少了特征偏差。文献 [11] 提出一种渐进式混合 Transformer 以解决多模态目标重识别中的模态融合问题。该架构包括随机混合增强和特征混合两个模块。随机混合增强通过图像随机裁剪和局部区域混合技术，在输入层面生成具有多模态结构特征的混合区域。特征混合则借助模态特定控制器与模态信息嵌入，在特征层面实现有效的多模态信息融合。文献 [12] 提出一种循环令牌置换 Transformer 以解决多模态目标重识别中存在的分布差异和局部特征利用不足问题。该方法采用多流网络来保留不同模态信息，通过令牌置换实现跨模态空间特征对齐，并利用互补重建模块在令牌级别施加重建约束以减少模态间分布差异。文献 [13] 提出一种光照感知跨模态增强 Transformer 以解决强光干扰下的特征退化问题。该方法通过自监督方式联合预测可见光与近红外模态的光照掩码，利用热红外光谱的光照免疫特性指导受损特征的恢复，并设计多模态一致性损失函数以增强跨模态语义一致性。

以上多模态特征融合方法可有效缓解模态间特征差异，但在融合过程中主要侧重于整合各模态整体特征，而无法充分聚焦模态特定的关键细节特征。为解决此问题，本文提出一种层级化 token 选择 Transformer（hierarchical token selection Transformer，HTSTrans）用于多模态目标重识别。HTSTrans 由多模态共享特征学习模块（multi-modal shared feature learning module，MSFLM）、层级化 token 选择模块（hierarchical token selection module，HTSM）和跨模态对齐约束模块（cross-modal alignment constraint module，CACM）构成。首先，MSFLM 使用权重共享视觉 Transformer（vision Transformer，ViT）对多模态图像进行特征学习以提取原始多模态特征。这可以在不同模态间初步建立特征一致性，为后续更精细化的特征学习提供统一表示空间。之后，HTSM 选择模态特定的关键特征并融合多模态特征以聚焦模态特定的关键细节并缓解模态间差异。具体地，基于 ViT 中不同编码器块对特征的建模能力差异，HTSM 将 ViT 编码器块划分为不同层级。随着层级深入，HTSM 部署递进式窗口以选择关键 token。通过融合不同层的 token 选择图生成单模态 token 选择图，通过融合不同模态的 token 选择图生成最终的多模态 token 选择图。将其与原始多模态特征融合即可获得多模态融合特征。该特征在聚焦模态特定关键细节特征的同时，还通过融合学习降低了模态间特征差异。最后，CACM 可对多模态特征进行模态间传输约束（inter-modal transmission constraint, ITC）和循环一致性约束（cycle consistency constraint, CCC），以对齐模态间的特征分布并保持多模态映

射的一致性。两种约束相互协作，进一步缓解了多模态特征间的差异，提升了模态不变性特征的判别性。

1 方法

1.1 整体架构

HTSTrans 架构如图 1 所示，包含 MSFLM、HTSM 和 CACM 这 3 个模块。首先，多模态图像被输入至 MSFLM 以提取原始特征。之后，HTSM 选择模态特定的关键特征并融合多模态特征以生成多模态特征描述符。最后，CACM 对多模态特征进行约束以缓解多模态特征间差异并提升模态不变性特征的判别性。

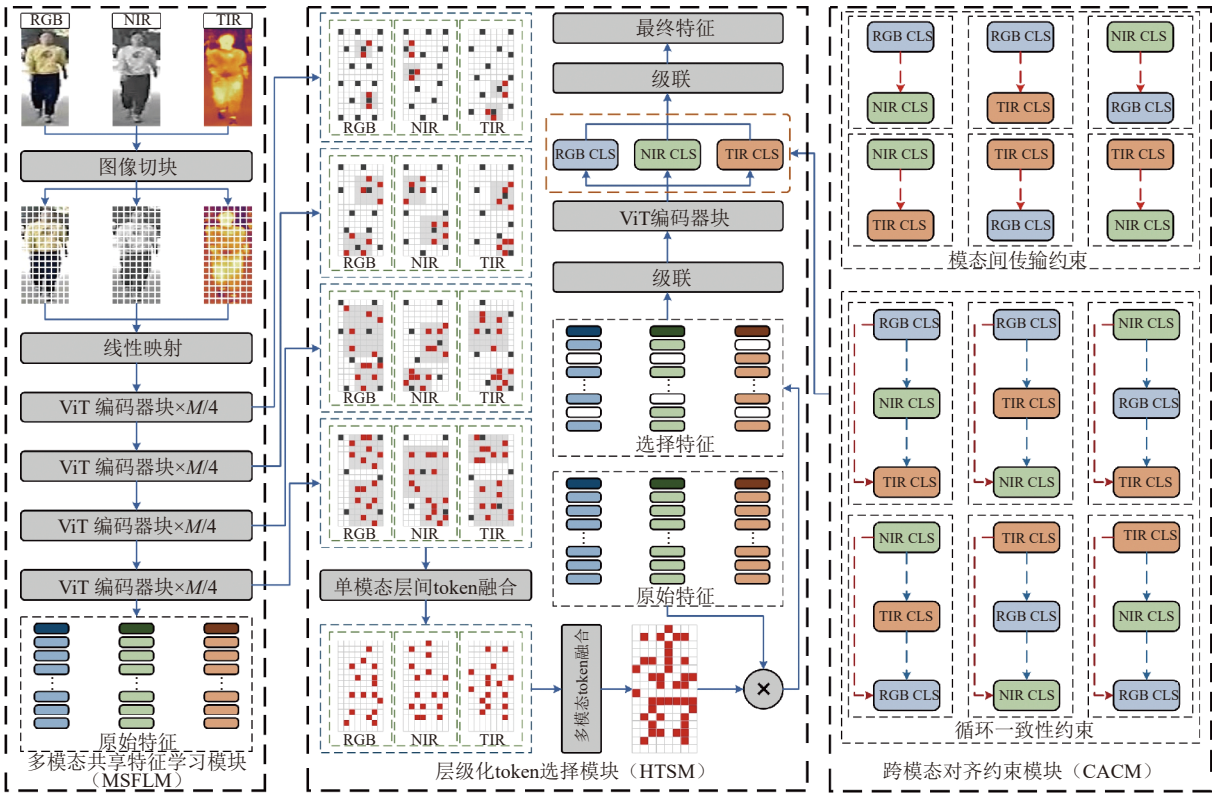


图 1 HTSTrans 整体框架图

1.2 多模态共享特征学习模块

多模态目标重识别的核心在于有效整合不同模态（RGB、NIR、TIR）的互补特征，使模型能在复杂环境下保持稳定识别。然而，由于不同模态的成像机理存在差异，各模态图像在亮度分布、纹理细节、边缘结构等方面呈现强烈异质性。MSFLM 作为 HTSTrans 的基础特征提取主干，主要目的是在保留各模态信息的同时，初步建立跨模态特征一致性表示。该模块采用权重共享 ViT 作为主干网络，对多模态图像进行统一特征编码。

给定一组 RGB-NIR-TIR 多模态图像 X_R，X_N，$X_T \in \mathbb{R}^{H\times W\times C}$，图像被分割为 N 个图像块 $\{x_p^1, x_p^2, \cdots, x_p^N\}$。经线性映射后，加入位置编码，形成初始序列：

$$Z_0 = \left[x_{\text{cls}}; f_{\text{LP}}(x_p^1, x_p^2, \cdots, x_p^N)\right] + E_{\text{pos}} \tag{1}$$

式中，$Z_0 \in \mathbb{R}^{(N+1)\times d}$ 为初始序列，表示输入 ViT 编码器的初始 token 序列；N 为图像块数量；d 为特征维度，一般设置为 768；x_{cls} 为分类 token，通过与图像块 token 的注意力交互聚合全图语义信息，代表图像全局表征；f_{LP} 为线性映射；x_p^i 为图像块；$E_{\text{pos}} \in \mathbb{R}^{(N+1)\times d}$ 为可学习位置编码，通过可学习参数矩阵实现，在训练过程中与网络参数一同更新，用于保留图像块之间的空间位置信息，使模型在特征建模过程中同时感知局部空间布局与全局结构关系。

构建初始序列 Z_0 之后，将其输入 MSFLM 的编码器块中进行特征提取。设 MSFLM 包含 M 个 ViT 编码器块，每个编码器块由多头自注意力和前馈网络组成，则第 m 层（$m = 1,2,\cdots,M$）编码器块的特征计算过程为：

$$Z_m' = Z_{m-1} + f_{\text{MHSA}}(f_{LN}(Z_{m-1})) \tag{2}$$

$$Z_m = Z_m' + f_{\mathrm{FFN}}(f_{LN}(Z_m')) \tag{3}$$

式中，f_{MHSA}为多头自注意力；f_{FFN}为前馈网络；f_{LN}为层归一化。

通过权重共享 ViT，MSFLM 能从不同模态中初步学习对齐的特征表示，有效捕获跨模态的共享语义信息。通过 MSFLM 提取的多模态原始特征为：

$$[f_R^o, f_N^o, f_T^o] = f_{\mathrm{MSFLM}}[X_R, X_N, X_T] \tag{4}$$

式中，f_R^o、f_N^o、f_T^o为多模态原始特征；X_R、X_N、X_T为输入的 RGB、NIR、TIR 多模态图像；f_{MSFLM}为 MSFLM 模块的整体处理流程。

多模态原始特征将在后续模块中被进一步处理，为精细化特征选择提供统一的表示空间。

1.3　层级化 token 选择模块

现有多数方法直接融合模态整体特征以降低模态差异，但无法聚焦模态特定关键细节。为此，提出层级化 token 选择模块（HTSM），其由 ViT 编码器块分层、种子 token 选择、邻域窗口选择、窗内 token 选择、窗间 token 融合、单模态层间 token 融合、多模态 token 融合、多模态特征融合构成，使用层级感知的递进式窗口选择并融合关键 token，再通过特征融合降低模态差异。

图像的不同区域信息各有差异，选择关键区域可反映关键细节。基于 ViT 特征机制，选择注意力响应较高的切块 token 定位关键区域。然而，不同深度编码器块的特征建模能力不同。若对不同深度的编码层均基于全局注意力图进行高响应 token 选择，则会忽略关键细节信息。为此，在选择关键 token 之前，需对不同深度编码器块进行差异化处理。根据深度将 ViT 中的M个编码器块划分为$H=4$个层级，每个层级包含$M/4$个编码器块。

HTSM 筛选可表征模态特定关键细节信息的 token，种子 token 选择（图 1 中黑色方块）是整个过程的起点。以 RGB 模态为例，第m层编码器块中，第h个注意力头对应的注意力响应图为A_m^h，将N_h个注意力响应相乘，生成该层的聚合注意力响应图为：

$$A_m = \prod_{h=1}^{N_h} A_m^h \tag{5}$$

基于A_m选取注意力响应最大的 token 作为第 1 个种子 token：

$$s_m^1 = \arg\max A_m \tag{6}$$

之后，从剩余 token 中选择响应值较高的 token。然而，高响应 token 易出现局部聚集现象，从而忽略其他关键区域。为此，引入距离约束机制，使种子 token 呈现较为均匀的全局分布。设第k个候选种子 token 为s_m^k，已选择的种子 token 集合为S_m，选择s_m^k成为种子 token，需满足以下条件：

$$\min_{s_m^i \in S_m} D(s_m^k, s_m^i) > D_m \tag{7}$$

式中，D为欧式距离计算函数，计算方式为$D(s_m^k, s_m^i) = \left\| s_m^k - s_m^i \right\|_2$；$D_m$为距离阈值，鉴于 ViT 不同层级的特征建模能力差异，引入差异化距离阈值机制，对于划分的 4 个层级，设置$D_m = \{2,3,4,5\}$。浅层编码器块更集中于局部细节，因此设定较小距离阈值，以保证密集采样和细粒度覆盖。深层编码器块更关注高层语义关联，因此增大距离阈值，使种子 token 更加稀疏并具有更强的全局代表性。

生成每个层级上的种子 token 后，以种子 token 为中心构建邻域窗口，将特征分析限制在以种子 token 为核心的邻域内，从而更精细地捕捉模态特定的关键细节特征。浅层特征包含丰富细节信息，采用小窗口以聚焦关键细节区域。深层特征的语义信息更抽象且感受野增大，采用大窗口以覆盖更广泛的语义区域。因此，为每个种子 token 设计层级感知的自适应递进式窗口机制。窗口半径不仅与当前所在层级有关，还与种子 token 自身的注意力响应强度相关联。对于注意力响应越高的种子 token，其邻域可能包含越多的相关上下文信息，需要更大窗口来囊括信息。对于种子 token s_m^k，其自适应窗口半径为：

$$r_m^k = \left\lceil R_h \left(0.5 + 0.5 \frac{A_m(s_m^k) - \min(A_m)}{\max(A_m) - \min(A_m)} \right) \right\rceil \tag{8}$$

式中，R_h为 4 个层级对应的基础半径。设s_m^k的空间坐标为(x_m^k, y_m^k)，对应邻域窗口为：

$$w_m^k = \left\{ (x,y) \middle| |x - x_m^k| \leqslant r_m^k, |y - y_m^k| \leqslant r_m^k \right\} \tag{9}$$

构建以种子 token 为中心的邻域窗口后，进行窗内 token 选择。每个邻域窗口内包含以种子 token 为中心的多个相邻的 token，直接融合窗内 token 特征会稀释核心特征表达强度。因此，需在窗口内部实施二次选择，聚焦与中心种子 token 语义高度相关且自身具有较高全局注意力响应的 token。具体地，使用窗内注意力重评估机制进行 token 选择。对于窗口w_m^k，以中心种子 token 的特征$f_{s_m^k}$作为查询，窗口内所有 token 的特征f_i作为

键，计算窗内 token 的注意力权重：

$$a_{\text{win}}(i)=\frac{\exp(\text{sim}(f_{s_m^k},f_i))}{\sum_{j\in w_m^k}\exp(\text{sim}(f_{s_m^k},f_j))} \tag{10}$$

式中，sim 为余弦相似度计算函数，计算方式为 $\text{sim}(f_{s_m^k},f_i)=(f_{s_m^k}f_i)/(\|f_{s_m^k}\|_2\|f_i\|_2)$。

窗内 token 选择结合全局注意力和窗内注意力：

$$A_i^{'}=A_m(i)+a_{\text{win}}(i) \tag{11}$$

选择固定数量的 token 作为该窗口的关键 token：

$$t_m^k=f_{\text{top}-\left\lceil\frac{N_T}{|S_m|}\right\rceil}(A_i^{'}) \tag{12}$$

式中，N_T 为选择的总 token 数，设置 $N_T=32$。

完成各邻域窗口内部 token 选择后，窗间 token 融合将同层不同窗口中的关键 token 进行整合。首先，基于每个窗口生成掩码：

$$M_m^k(i,j)=\begin{cases}1 & \text{该位置token}\in t_m^k\\ 0 & \text{该位置token}\notin t_m^k\end{cases} \tag{13}$$

然后，对所有 K 个窗口的掩码求并集，得到的全局掩码标识了第 m 层中所有关键 token 的空间分布：

$$M_m=\bigcup_{k=1}^{K}M_m^k \tag{14}$$

完成各层窗间 token 融合后，单模态层间 token 融合模块整合 M 个编码器块的掩码信息，生成单模态关键 token 选择图。首先，计算层间累积掩码图：

$$M_M=\sum_{m=1}^{M}M_m \tag{15}$$

每个位置的值表示该 token 被选为关键 token 的次数。根据累积掩码选择出现次数最多的 N_G 个 token，获得 RGB 模态的 token 选择图：

$$M_R=f_{\text{mask}}(f_{\text{top}-N_G}(M_M)) \tag{16}$$

式中，f_{mask} 为掩码处理。通过以上步骤，NIR 和 TIR 模态的 token 选择图 M_N 和 M_T 也可生成，这些 token 选择图反映了模态特定的关键细节信息分布。

生成不同模态特定的 token 选择图后，多模态 token 融合整合不同模态的关键 token 信息，利用多模态数据互补特性形成更具判别力的跨模态特征表示。具体地，采用掩码并集策略整合不同模态的关键 token 选择图：

$$M_{\text{fuse}}=M_R\cup M_N\cup M_T \tag{17}$$

图 2 展示了单模态层间 token 融合和多模态 token 融合的过程。

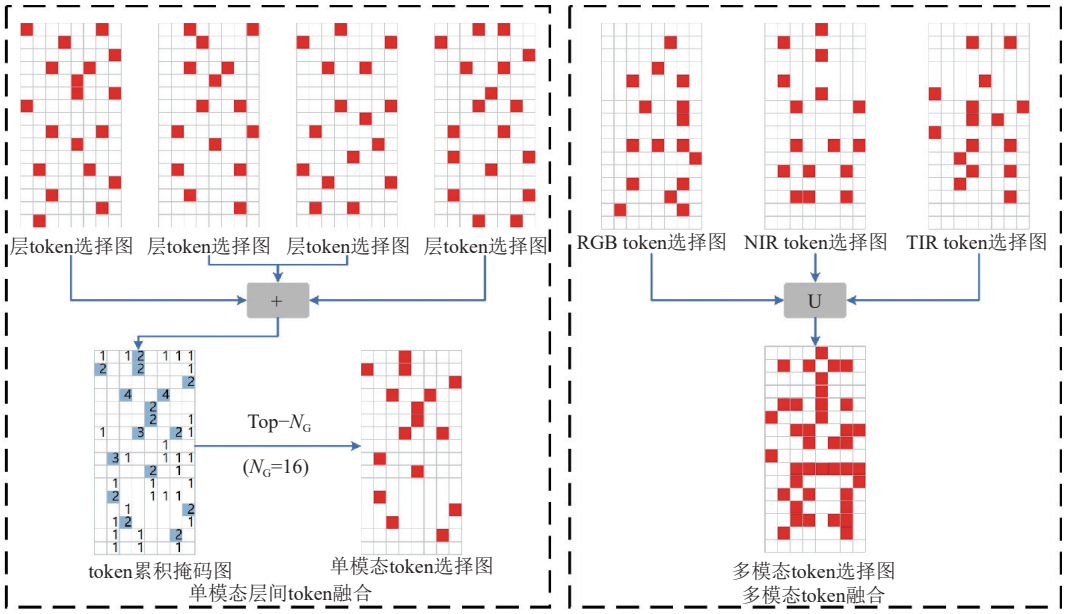


图 2 单模态层间 token 融合和多模态 token 融合

完成模态间 token 融合后，进行最终的多模态特征融合。此阶段基于融合掩码 M_{fuse}，将原始多模态特征进行有效整合，生成具有强判别性的多模态特征描述符。具体地，将 MSFLM 提取的原始特征 $\{f_R^o,f_N^o,f_T^o\}$ 与 M_{fuse} 对应相乘即可获得选择特征：

$$\left[f_R^s,f_N^s,f_T^s\right]=\left[f_R^o,f_N^o,f_T^o\right]\odot[1;M_{\text{fuse}}] \tag{18}$$

式中，1表示将分类 token 的掩码置为 1。

为降低跨模态差异，将多模态选择特征进行级联并通过后续编码器块提取融合特征，最后抽取并级联多模态分类 token 作为最终多模态特征描述符：

$$f_{\text{fuse}} = f_{\text{cat}}\left(f_{\text{cls}}\left(f_e\left(f_{\text{cat}}\left(f_R^s, f_N^s, f_T^s\right)\right)\right)\right) \tag{19}$$

式中，f_{cat}为级联处理；f_e为后续编码处理；f_{cls}为抽取每个模态对应的分类 token 后的$\{f_R, f_N, f_T\}$。

基于 HTSM 原理，对其进行复杂度分析。设编码器块数量为M，切块数量为N，特征维度为d，注意力头数量为N_h，窗口数量为K。注意力响应聚合的复杂度为$O(MNN_h)$，种子 token 选择的复杂度为$O(MN\log N + MNK)$，邻域窗口构建的复杂度为$O(MNK + MN)$，窗内 token 选择的复杂度为$O(MNd + MN\log(N/K))$，窗间 token 融合的复杂度为$O(MNK)$，单模态层间 token 融合的复杂度为$O(MN + N\log N)$，多模态 token 融合的复杂度为$O(3N)$，多模态特征融合的复杂度为$O(3Nd + 9HN^2d + 24Nd^2)$。

1.4　跨模态对齐约束模块

由于多模态图像的成像机理差异，来源于同一身份的多模态特征在特征空间中不能有效对齐。为此，提出跨模态对齐约束模块（CACM）实现跨模态特征约束以减小同一身份的模态间特征差异性，从而提升模态不变性特征的判别性。该模块包含 ITC 和 CCC。

ITC 用于构建模态间特征传输约束，使同一身份的多模态特征具备分布一致性。设每个训练批包含P个身份，每个身份各包含Z个三模态样本。对于每个身份$y \in \{1,2,\cdots,P\}$，提取多模态分类 token 特征：

$$F_R^y = \left\{f_{R,1}^y, f_{R,2}^y, \cdots, f_{R,Z}^y\right\} \tag{20}$$

$$F_N^y = \left\{f_{N,1}^y, f_{N,2}^y, \cdots, f_{N,Z}^y\right\} \tag{21}$$

$$F_T^y = \left\{f_{T,1}^y, f_{T,2}^y, \cdots, f_{T,Z}^y\right\} \tag{22}$$

基于此，首先构建从 RGB 模态到 NIR 模态的传输路径代价。定义传输代价矩阵$\boldsymbol{C}_{R\to N}^y \in \mathbb{R}^{Z\times Z}$，每个元素表征样本从源模态传输到目标模态所需代价：

$$C_{R\to N}^y(i,j) = 1 - \frac{f_{R,i}^y f_{N,j}^y}{\left\|f_{R,i}^y\right\|_2 \left\|f_{N,j}^y\right\|_2} \tag{23}$$

从 RGB 模态到 NIR 模态的特征传输损失定义为：

$$L_{R\to N}^y = \frac{1}{Z^2}\sum_{i=1}^{Z}\sum_{j=1}^{Z} C_{R\to N}^y(i,j) \tag{24}$$

为构建完整传输体系，建立其他方向的传输路径，形成全方位的特征传输网络。身份特定的完整传输约束损失通过集成所有传输路径的代价得到：

$$L_{\text{ITC}}^y = L_{R\to N}^y + L_{R\to T}^y + L_{N\to R}^y + L_{N\to T}^y + L_{T\to R}^y + L_{T\to N}^y \tag{25}$$

虽然模态间传输约束能对齐模态特征分布，但仅考虑两两模态间的成对匹配。为保证多模态对齐满足全局传递一致性，引入 CCC。其在多模态嵌入空间中建立传递一致的映射链路，确保任意两模态间的映射关系不仅相互独立对齐，同时在跨模态循环路径上保持稳定一致性。根据模态间的代价矩阵，首先定义 3 组匹配概率矩阵：

$$P_{R\to N}^y(i,j) = \frac{\exp(-C_{R\to N}^y(i,j)/\tau)}{\sum_{j'}\exp(-C_{R\to N}^y(i,j')/\tau)} \tag{26}$$

$$P_{N\to T}^y(j,k) = \frac{\exp(-C_{N\to T}^y(j,k)/\tau)}{\sum_{k'}\exp(-C_{N\to T}^y(j,k')/\tau)} \tag{27}$$

$$P_{R\to T}^y(i,k) = \frac{\exp(-C_{R\to T}^y(i,k)/\tau)}{\sum_{k'}\exp(-C_{R\to T}^y(i,k')/\tau)} \tag{28}$$

式中，τ为温度参数，设置τ为 0.5，用于控制 softmax 函数输出分布的平滑程度。

每个元素反映了跨模态样本间的匹配强度。理想情况下，若模态映射一致，应满足以下传递关系：

$$P_{R\to T}^y(i,k) \approx P_{R\to N}^y(i,j)\,P_{N\to T}^y(j,k) \tag{29}$$

上式表示从 RGB 模态到 TIR 模态的直接映射与通过模态 NIR 的间接映射结果一致。基于此，RGB-NIR-TIR 的循环一致性损失定义为：

$$L_{R\to N\to T}^y = \frac{1}{Z^2}\left\|P_{R\to T}^y(i,k) - P_{R\to N}^y(i,j)\,P_{N\to T}^y(j,k)\right\|_{\text{F}} \tag{30}$$

同样建立其他方向的一致性约束，形成全方位的一致性约束网络。身份特定的完整循环一致性约束损失通过集成所有循环路径的代价得到：

$$\begin{aligned} L_{\text{CCC}}^y =& L_{R\to N\to T}^y + L_{R\to T\to N}^y + L_{N\to R\to T}^y + \\ & L_{N\to T\to R}^y + L_{T\to R\to N}^y + L_{T\to N\to R}^y \end{aligned} \tag{31}$$

将模态间传输约束损失和循环一致性约束损失结合，构成最终的跨模态对齐约束损失：

$$L_{\mathrm{CACM}}=\frac{1}{P}\sum_{y=1}^{P}\left(L_{\mathrm{ITC}}^{y}+L_{\mathrm{CCC}}^{y}\right) \tag{32}$$

2 实验验证

2.1 实验设置

多模态目标重识别中，多模态行人/车辆重识别是最为重要的任务。现有方法的性能评估主要在 RGBNT201[14]、RGBNT100[15]、MSVR310[7] 这3 个数据集上进行。RGBNT201 包含 201 个行人，共 4 787 组多模态图像对，141 个行人作为训练集，30 个行人作为验证集，30 个行人作为测试集。RGBNT100 包含 100 个车辆，共 17 250 组多模态图像对，50 个车辆作为训练集，50 个车辆作为测试集。MSVR310 包含 310 个车辆，共 2 087 组多模态图像对，155 个车辆作为训练集，155 个车辆作为测试集。

特征提取主干包含 $M=12$ 个 ViT 编码器块，被划分为 $H=4$ 个层级。行人图像和车辆图像分别被缩放至 256×128 和 128×256。SGD 优化器优化模型，初始学习率为 0.001，动量为 0.9，衰减因子为 0.000 1。式 (16) 中的参数设定为 $N_G=16$。GPU 为 NVIDIA GeForce RTX 3090Ti，PyTorch 版本为 1.7.1。

平均精度均值（mean average precision, mAP）和累计匹配特性（cumulative matching characteristic, CMC）为评估指标，使用 Rank-k 来反映 CMC 指标。

2.2 不同方法对比

为验证 HTSTrans 用于多模态目标重识别的有效性，将其与其他主流方法进行对比，结果如表 1 所示。所对比的其他方法包括 IEEE [16]、CCNet [7]、DENet [17]、LRMM [18]、TIENet [19]、UniCat [20]、EDITOR [21]、RSCNet [4]、FACENet [13] 和 WTSF-ReID [22]。前 5 种方法是基于 CNN 构建的，后 5 种方法则是基于 ViT 构建的。IEEE [16] 通过跨模态特征交互提取模态互补信息，并将全局信息嵌入局部特征以增强特征表达能力，最后利用多模态间隔损失引导网络学习更具判别性的模态特征。CCNet [7] 通过跨方向一致性损失同时约束模态级与样本级分布，并结合自适应层归一化缓解不同模态和复杂环境带来的分布偏移，从而在保持模态判别性的同时提升跨模态一致性与鲁棒性。DENet [17] 通过跨模态特征变换来恢复缺失模态的表示，并根据不同的模态缺失状态动态调整模态间的特征增强方式，从而在模态不完整和缺失状态变化的情况下，仍能保持多模态特征的判别性与鲁棒性。LRMM [18] 通过多尺度多模态交互在不同尺度上建模模态间关系，引导各模态充分交换互补信息，同时引入低秩多模态融合以降低融合冗余、提升融合效率，并结合多模态原型损失与分类损失共同约束网络，使模型能增强跨模态一致性。TIENet [19] 通过在频域中进行幅度引导的相位学习来捕获跨模态共享的全局结构信息，同时在空间域中保留局部细节，并结合模态内交互、模态间交互与模态内特征融合 3 种交互机制，系统性地建模多模态间及单模态内的关系。UniCat [20] 通过在训练阶段彻底解耦各模态特征，并且仅在推理阶段进行特征拼接的后融合策略，有效地避免了多模态联合优化过程带来的模态惰性问题，使每个模态都能充分学习其自身的判别性特征信息。EDITOR [21] 通过空间特征域–频率特征域联合的特征分析机制从不同模态特征中自适应筛选关键区域，并通过自适应信息聚合在模态内与模态间建模特征交互，同时引入背景一致性约束和对象中心特征细化损失以稳定特征分析并压缩类内分布。RSCNet [4] 通过注意力域与傅里叶域联合的特征稀疏化机制保留各模态特征中的有效判别性信息，并在此基础上进行跨模态多表征耦合，在缩小模态差异的同时避免无关信息干扰，同时引入信息统一约束策略，从特征层面和分布层面同时约束多模态表示的一致性。FACENet [13] 通过建模不同光照对多模态特征的影响，引入光照感知与抑制机制，削弱光照相关成分，同时保留与身份相关的结构与语义信息，并结合跨模态一致性约束，引导模型学习更加稳健、光照不变的多模态特征表示。WTSF-ReID [22] 通过局部区域扩展机制在不同尺度的区域内对各模态特征进行建模与筛选，使模型能够从局部到全局逐步聚焦于更具判别力的关键特征信息。在此基础上，通过信息驱动的特征分析与一致性约束，实现对重要特征的有效融合与跨模态对齐。

从表 1 中的实验结果可以看出，由于 ViT 比 CNN 具备更优秀的特征建模能力，整体而言，ViT-based 方法比 CNN-based 方法的重识别准确率更高。本文的 HTSTrans 架构在 3 个数据集上均获得了最佳性能指标。具体地，在 RGBNT201 上的 mAP、Rank-1、Rank-5、Rank-10 分别为 70.1%、74.0%、84.3%、89.4%，在 RGBNT100 上的 mAP、

Rank-1 分别为 83.6%、97.4%，在 MSVR310 上的 mAP、Rank-1 分别为 40.9%、56.6%。相较于其他方法，在 RGBNT201 上的 mAP、Rank-1、Rank-5、Rank-10 性能指标提升分别为 1.9%～27.7%、1.5%～31.8%、1.0%～29.0% 和 1.2%～25.2%，在 RGBNT100 上的 mAP 和 Rank-1 性能提升分别为 1.3%～22.3% 和 0.5%～9.6%，在 MSVR310 上的 mAP 和 Rank-1 性能提升分别为 1.4%～19.9% 和 1.4%～15.6%。对比表现较佳的 CNN-based 方法 LRMM [18]，HTSTrans 在 3 个数据集上的性能提升分别为 17.8%、20.6%、19.7%、16.2%、5.0%、0.7%、4.2%、6.9%。对比表现较佳的 ViT-based 方法 RSCNet [4]，HTSTrans 对应的性能提升分别为 1.9%、1.5%、1.5%、1.6%、1.3%、0.8%、1.4%、7.0%。以上实验验证了 HTSTrans 用于多模态目标重识别的可行性，可实现具备竞争力的性能。HTSTrans 取得上述优势性能在于其针对多模态目标重识别任务特性所设计的分层特征建模与关键判别信息选择机制。一方面，HTSTrans 基于 ViT 骨干网络构建统一的跨模态表示空间，通过权重共享的多模态共享特征学习模块有效缓解了不同模态间成像差异带来的特征分布偏移问题，为后续跨模态匹配提供了稳定的语义基础。另一方面，本文的层级化 token 选择模块从不同语义层次对 ViT 编码器输出的 token 进行筛选与融合，使模型能够自适应聚焦于更具身份判别力的关键区域信息，避免大量背景或模态冗余信息对跨模态匹配造成干扰，从而显著提升特征表达的判别性与鲁棒性。此外，引入的跨模态对齐约束模块进一步从跨模态对齐角度对特征空间进行约束，通过联合优化模态间传输和循环一致性，有效缩小同一目标在不同模态下的特征距离，同时拉大不同目标之间的分布间隔。上述模块在整体框架中的协同作用，使 HTSTrans 能够在充分挖掘多模态互补信息的同时保持良好的模态不变性，最终在多个数据集上均表现出稳定且显著的性能优势。

表 1　HTSTrans 与不同方法对比　　%

Method	Backbone	Reference	RGBNT201				RGBNT100		MSVR310	
			mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	mAP	Rank-1
IEEE [16]	CNN	AAAI-2022	46.4	47.1	58.5	64.2	61.3	87.8	21.0	41.0
CCNet [7]	CNN	InFusion-2023	—	—	—	—	77.2	96.3	36.4	55.2
DENet [17]	CNN	arXiv-2023	42.4	42.2	55.3	64.5	68.1	89.2	—	—
LRMM [18]	CNN	ESWA-2025	52.3	53.4	64.6	73.2	78.6	96.7	36.7	49.7
TIENet [19]	CNN	TNNLS-2025	54.4	54.4	66.3	71.1	—	—	—	—
UniCat [20]	ViT	arXiv-2023	57.0	55.7	—	—	79.4	96.2	—	—
EDITOR [21]	ViT	CVPR-2024	66.5	68.3	81.1	88.2	82.1	96.4	39.0	49.3
RSCNet [4]	ViT	TCSVT-2025	68.2	72.5	82.8	87.8	82.3	96.6	39.5	49.6
FACENet [13]	ViT	InFusion-2025	—	—	—	—	81.5	96.9	36.2	54.1
WTSF-ReID [22]	ViT	ESWA-2025	67.9	72.2	83.3	88.0	82.2	96.5	39.2	49.1
HTSTrans	ViT	Ours	70.1	74.0	84.3	89.4	83.6	97.4	40.9	56.6

2.3　消融实验

HTSTrans 主要由 MSFLM、HTSM 和 CACM 这 3 个模块构成，为验证各模块有效性，进行消融实验，结果如表 2 所示。在仅使用 MSFLM 时，模型性能最低，在 RGBNT201 上的 mAP、Rank-1、Rank-5、Rank-10 为 55.2%、54.3%、67.9%、76.3%，在 RGBNT100 上的性能指标为 76.4%、94.8%、96.4%、97.5%，在 MSVR310 上的性能指标为 28.1%、44.2%、60.4%、68.9%。在引入 HTSM 之后，mAP、Rank-1、Rank-5、Rank-10 提升分别为 12.7%、18.5%、7.6%、7.0%、4.5%、1.1%、0.8%、0.4% 和 6.3%、7.9%、4.9%、5.9%。说明 HTSM 通过选择模态特定的关键特征并融合多模态特征可聚焦模态特定的关键细节信息并缓解模态间差异，从而提升特征表达的判别性与鲁棒性。在继续引入 CACM 之后，性能进一步提升。说明 CACM 通过对多模态特征进行模态间传输约束和循环一致性约束，缓解了多模态特征差异并提升模态不变性特征的判别性。以上实验验证了 MSFLM、HTSM 和 CACM 设计的有效性。

表 2 MSFLM、HTSM 和 CACM 模块消融分析

%

HTSTrans			RGBNT201				RGBNT100				MSVR310			
MSFLM	HTSM	CACM	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
√	×	×	55.2	54.3	67.9	76.3	76.4	94.8	96.4	97.5	28.1	44.2	60.4	68.9
√	√	×	67.9	72.8	75.5	83.3	80.9	95.9	97.2	97.9	34.4	52.1	65.3	74.8
√	√	√	70.1	74.0	84.3	89.4	83.6	97.4	97.8	98.5	40.9	56.6	69.8	79.2

进一步对 HTSTrans 的计算效率进行分析，主要通过模型复杂度进行评估，结果如表 3 所示，复杂度指标包括模型参数量（Parameters）、浮点运算量（FLOPs）以及单次推理时间（Inference Time）。在仅使用 MSFLM 时，模型参数量为 124.6×10^6，浮点运算量为 66.4×10^9，单次推理时间为 5.5 ms。引入 HTSM 之后，模型参数量和浮点运算量分别增加至 135.7×10^6 和 70.6×10^9，单次推理时间提升至 6.3 ms。结合表 2 可知，HTSM 在显著提升模型性能的同时，仅带来了较小幅度的额外复杂度开销。这是由于 HTSM 主要基于 token 筛选实现关键特征选择，其计算过程集中于局部 token 集合，并未引入大规模额外的网络参数或高复杂度运算。在进一步引入 CACM 之后，模型在测试阶段的参数量、浮点运算量以及单次推理时间均保持不变。这是因为 CACM 仅在训练阶段通过损失函数形式对多模态特征施加约束，并不引入额外的网络结构或推理操作。因此，CACM 不会增加模型在推理阶段的计算复杂度。上述结果表明，HTSTrans 在获得性能显著提升的同时，整体复杂度增长可控，在性能与效率之间实现了较为合理的平衡。

表 3 HTSTrans 复杂度分析

HTSTrans			Complexity		
MSFLM	HTSM	CACM	Parameters	FLOPs	Inference Time/ms
√	×	×	124.6×10^6	66.4×10^9	5.5
√	√	×	135.7×10^6	70.6×10^9	6.3
√	√	√	135.7×10^6	70.6×10^9	6.3

HTSTrans 使用权重共享 ViT 提取多模态原始特征。为验证该策略的有效性，对比采用权重共享 ViT 和权重独立 ViT 提取多模态原始特征对应的性能，结果如表 4 所示。使用权重独立 ViT 时的性能较低，在 RGBNT201 上的 mAP、Rank-1、Rank-5、Rank-10 为68.4%、72.8%、83.7%、88.6%，在 RGBNT100 上的性能指标为 81.5%、96.6%、97.0%、97.4%，在 MSVR310 上的性能指标为 38.4%、52.9%、66.7%、78.0%。在 3 个数据集上，采用独立权重设置的性能均低于共享权重方式。这表明权重共享 ViT 能够在特征提取阶段引导不同模态学习统一的特征建模方式，构建统一的跨模态表示空间，使模型更加关注模态间的共享语义，从而在早期阶段缓解由成像机理差异引起的特征分布偏移。此外，权重共享还在一定程度上减少了模型过拟合某一特定模态的风险，为后续层级化 token 选择与跨模态对齐约束模块提供了更加一致、稳定的特征基础。上述结果验证了在 HTSTrans 中采用权重共享 ViT 进行多模态特征学习的合理性与有效性。

表 4 HTSTrans 中权重共享 ViT 和权重独立 ViT 对应性能对比

%

ViT Weight	RGBNT201				RGBNT100				MSVR310			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Independent Weight	68.4	72.8	83.7	88.6	81.5	96.6	97.0	97.4	38.4	52.9	66.7	78.0
Shared Weight	70.1	74.0	84.3	89.4	83.6	97.4	97.8	98.5	40.9	56.6	69.8	79.2

HTSTrans 所使用的 ViT 共包含 12 个编码器块，被划分为$H=4$个层级。为验证不同分层数对性能的影响，进行对比实验，结果如表 5 所示。当 $H=1$时，性能最低，在 RGBNT201 上的 mAP、Rank-1、Rank-5、Rank-10 为66.1%、71.7%、76.5%、84.2%，在 RGBNT100 上的性能指标为 81.4%、

96.3%、96.5%、97.3%，在 MSVR310 上的性能指标为 36.9%、52.1%、65.9%、77.6%。当层级数较少时，不同深度编码器块被粗粒度地归为同一层级，浅层以局部纹理和边缘信息为主、深层以高层语义和结构关系为主的建模差异未能得到充分利用，导致 HTSM 在执行 token 选择时难以同时兼顾细粒度细节建模与高层语义聚合，从而限制了性能提升。随着层级数逐渐增加至$H=4$，不同深度编码器块的特征建模特性得以更清晰地分层表达，使 HTSM 能够针对不同层级特征分别采用差异化的种子 token 分布与邻域窗口策略，在保证关键细节充分覆盖的同时，又能有效引入具有全局判别性的高层语义信息，从而显著提升多模态特征选择与融合的效果。然而，随着层级数继续增大（$H>4$），相邻层级之间的特征语义差异逐渐减小，性能提升幅度较小。且过细的层级划分会引入冗余的 token 选择结果，使不同层级选取的关键 token 高度重叠，难以带来新的有效判别信息增益。因此，在 12 个编码器块的 ViT 结构下，划分层级$H=4$能够在充分挖掘不同深度特征建模差异的同时，避免过度细分带来的冗余问题，是一种兼顾性能与稳定性的最优选择。

表 5　不同层级数下的性能对比

%

H	RGBNT201				RGBNT100				MSVR310			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
1	66.1	71.7	76.5	84.2	81.4	96.3	96.5	97.3	36.9	52.1	65.9	77.6
2	68.3	72.9	79.2	86.9	82.4	96.6	97.4	97.9	38.7	54.7	68.3	78.1
3	69.3	73.5	83.1	88.7	82.9	97.0	97.7	98.2	39.5	56.2	68.8	78.5
4	70.1	74.0	84.3	89.4	83.6	97.4	97.8	98.5	40.9	56.6	69.8	79.2
6	70.2	74.0	84.4	89.4	83.8	97.5	97.9	98.4	41.1	56.8	69.9	79.3
12	70.3	74.2	84.5	89.4	83.9	97.5	97.9	98.6	41.2	56.7	69.8	79.4

HTSTrans 在构建邻域窗口时，基于层级信息和注意力信息自适应设定窗口半径。为验证该策略的有效性，对比使用自适应窗口半径（adaptive window radius, AWR）和固定窗口半径（fixed window radius, FWR）时的性能，结果如图 3 所示。使用固定窗口半径机制时的性能较低，在RGBNT201 上的mAP、Rank-1、Rank-5、Rank-10 为 68.8%、72.5%、83.1%、88.6%，在 RGBNT100 上的性能指标为 82.1%、96.4%、97.2%、97.7%，在 MSVR310 上的性能指标为 38.8%、54.7%、68.4%、77.9%。在使用自适应窗口半径机制之后，mAP、Rank-1、Rank-5、Rank-10 均获得相应提升。说明在固定窗口半径机制下，统一窗口尺寸难以适应不同层级特征差异性。而自适应半径机制根据层级深度感知和注意力强度动态调整邻域窗口半径，既能在浅层保留局部细节，又能在深层覆盖上下文，从而有效挖掘关键细节。

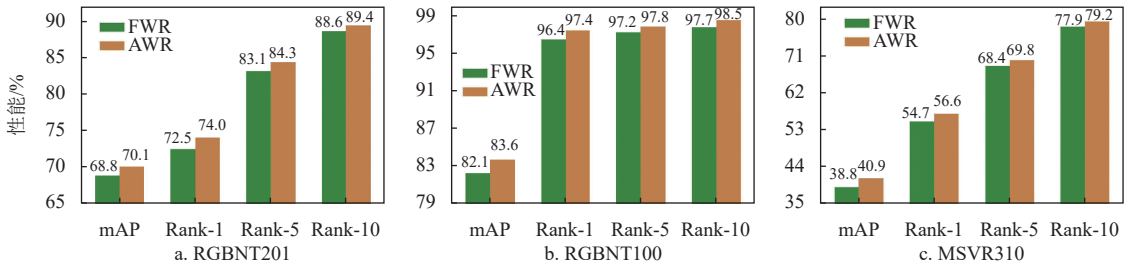


图 3　固定窗口半径和自适应窗口半径下的性能对比

HTSTrans 在进行窗内 token 选择时，同时结合了全局注意力（GLOBAL attention, GA）和窗内注意力（window attention, WA）。为验证该策略的有效性，进行对比实验，结果如图 4 所示。仅使用全局注意力进行窗内 token 选择所对应的性能最低，在RGBNT201 上的mAP、Rank-1、Rank-5、Rank-10 为69.0%、72.8%、83.9%、88.8%，在RGBNT100 上的性能指标为 82.4%、96.7%、97.0%、97.6%，在 MSVR310 上的性能指标为 39.7%、54.9%、68.5%、77.9%。在将全局注意力和窗内注意力进行融合之后

再进行窗内 token 选择可获得最高性能。说明全局注意力和窗内注意力具有互补作用，全局注意力能识别图像中具有高判别性的整体区域，窗内注意力能捕捉细粒度特征，将两者结合可实现更精准的 token 选择。

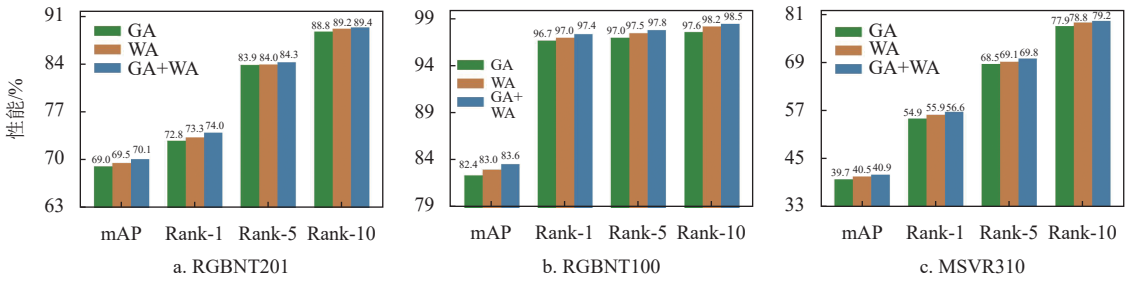


图 4 基于不同注意力进行 token 选择的性能对比

HTSTrans 在进行单模态层间 token 融合时，根据累积掩码选择出现次数最多的$N_G = 16$个 token。为验证N_G对性能的影响，进行对比实验，结果如图 5 所示。当$N_G = 4$增大至$N_G = 16$时，性能均逐渐上升，当$N_G = 16$时，性能达到峰值。之后，随着N_G进一步增大，性能会逐渐下降。说明当N_G值过小时，无法充分覆盖关键细节区域，导致关键特征缺失。适当增大N_G可增强对关键特征的捕捉。而当N_G过大时，会引入过多与身份无关的背景或噪声 token，稀释关键特征的表达强度，从而降低特征判别性。

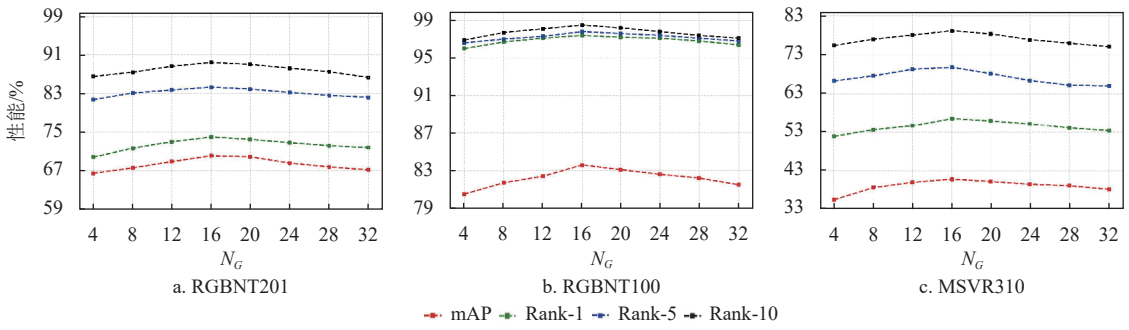


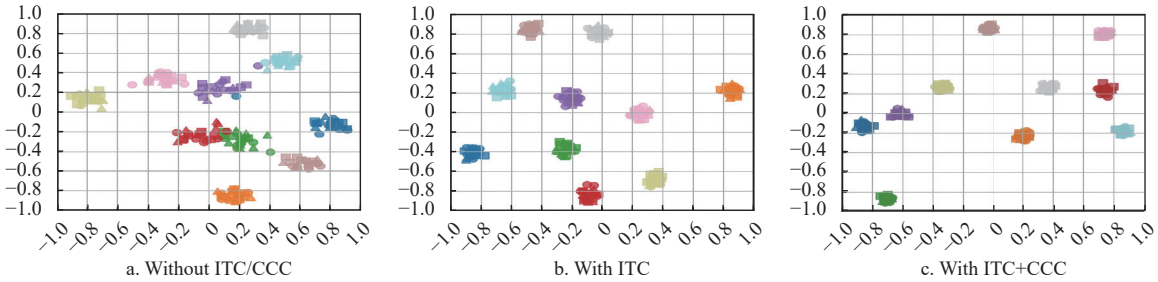
图 5 不同参数N_G下的性能对比

HTSTrans 使用 CACM 降低不同模态空间中的特征差异并提升模态不变性特征的判别性。CACM 由模态间传输约束（ITC）和循环一致性约束（CCC）构成，为验证两种约束的互补性，进行对比实验，结果如表 6 所示。在单独引入模态间传输约束或循环一致性约束之后，mAP、Rank-1、Rank-5、Rank-10 均可以获得提升。将两种约束同时引入，性能提升最为显著。说明两种约束机制在特征对齐中具有互补作用，模态间传输约束实现了模态间的显式对齐，能有效拉近不同模态的特征分布，循环一致性约束通过维护多模态循环路径上的映射一致性，确保了跨模态特征转换的传递稳定性和全局一致性。为直观体现两种约束的作用，对不同约束下的特征分布进行可视化，结果如图 6 所示，图中不同颜色表示不同身份，圆形、三角形、正方形分别表示 RGB、NIR、TIR 样本。在引入模态间传输约束或循环一致性约束之后，样本类间距离增大，类内距离缩小。将两种约束结合之后，性能进一步改善。

表 6 CACM 中 ITC 和 CCC 消融分析

%

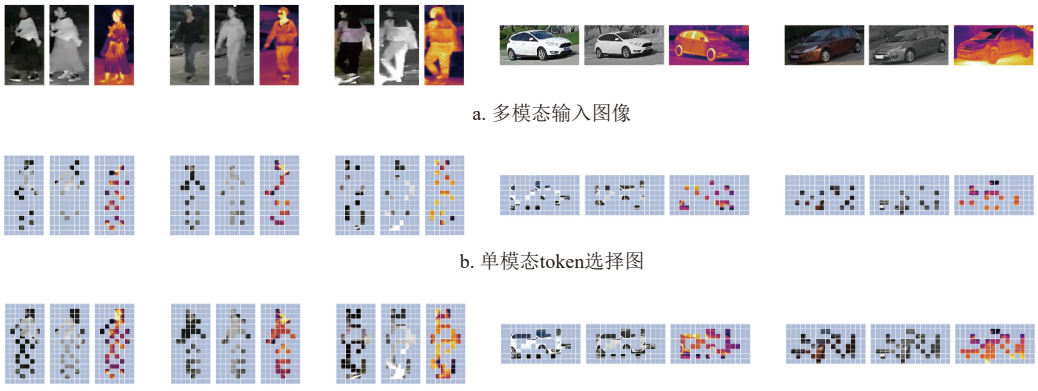
CACM		RGBNT201				RGBNT100				MSVR310			
ITC	CCC	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
×	×	67.9	72.8	75.5	83.3	80.9	95.9	97.2	97.9	34.4	52.1	65.3	74.8
√	×	69.3	73.7	82.7	87.9	82.6	96.8	97.6	98.3	39.2	55.1	68.3	77.9
×	√	69.0	73.3	82.1	86.9	81.9	96.6	97.3	98.1	38.6	53.9	68.0	77.0
√	√	70.1	74.0	84.3	89.4	83.6	97.4	97.8	98.5	40.9	56.6	69.8	79.2



a. Without ITC/CCC　　b. With ITC　　c. With ITC+CCC

图 6　不同约束下的多模态特征分布可视化

HTSTrans 的核心在于通过层级化 token 选择聚焦模态特定的关键细节特征，为更加直观地分析 HTSTrans 对关键细节的聚焦能力，对模型所生成的 token 选择结果进行可视化分析，实验结果如图 7 所示。第 1 行为原始的多模态输入图像，第 2 行为单模态 token 选择图，第 3 行为融合的多模态 token 选择图。可见，HTSTrans 能够在不同模态下稳定地聚焦与目标身份高度相关的关键细节区域，如行人的头肩区域、躯干轮廓或车辆的车身结构等，而对背景及噪声区域的响应较弱。这表明层级化 token 选择机制通过有效结合不同深度特征的建模特性，可以实现对模态特定关键细节的精准定位。该可视化结果从直观角度验证了 HTSTrans 在特征选择阶段的聚焦能力和对更具判别性特征信息的挖掘能力，也进一步解释了其在多模态目标重识别任务中取得性能优势的原因。



a. 多模态输入图像

b. 单模态token选择图

c. 多模态token选择图

图 7　Token 选择结果可视化

3　结束语

本文提出 HTSTrans 用于多模态目标重识别，其由多模态共享特征学习模块、层级化 token 选择模块和跨模态对齐约束模块 3 个主要部件组成。多模态共享特征学习模块基于权重共享 ViT 对多模态图像进行特征学习以提取原始多模态特征，为后续特征学习提供统一表示空间。层级化 token 选择模块基于层级感知的递进式窗口选择并融合关键的 token，从而聚焦模态特定的关键细节特征，后续再通过多模态特征融合降低跨模态差异并提升多模态特征的表达能力。跨模态对齐约束模块基于模态间传输约束和循环一致性约束对齐跨模态特征，两种约束相互协作以进一步缓解多模态特征间差异并提升模态不变性特征的判别性。在多个数据集上的实验结果验证了本文方法的有效性。

尽管所提出的 HTSTrans 在多模态目标重识别任务中取得了具备竞争力的性能，但仍存在一些有待提升的方面。首先，HTSTrans 在结构设计中涉及多个需手动设定的结构参数，如层级划分方式、token 选择数量及相关模块的配置策略，这些参数在不同应用场景下可能需要进行针对性调整，增加了模型部署的复杂度。如何设计自适应或与参数无关的结构，以降低对人工经验的依赖，是未来工作中值得重点关注的方面。其次，尽管 HTSTrans 在现有公开数据集上表现稳定，但在更加复杂和多变的实际应用环境中，如遮挡严重、背景干扰增强、极端视角变化或模态信息质量显著退化等情况下，

模型的鲁棒性与稳定性仍有进一步提升空间。未来工作可结合更复杂场景下的数据建模与鲁棒优化策略，以增强模型在真实环境中的泛化性与可靠性。

参 考 文 献

[1] Qian Yan, Barthelemy J, Karuppiah E, et al. Identifying re-identification challenges: Past, current and future trends[J]. SN Computer Science, 2024, 5(7): 937.

[2] Ye Mang, Chen Shuoyi, Li Chenyue, et al. Transformer for object re-identification: A survey[J]. International Journal of Computer Vision, 2025, 133(5): 2410-2440.

[3] Zhang Shizhou, Luo Wenlong, Cheng De, et al. Prompt-based modality alignment for effective multi-modal object re-identification[J]. IEEE Transactions on Image Processing, 2025, 34: 2450-2462.

[4] Yu Zhi, Huang Zhiyong, Hou Mingyang, et al. Representation selective coupling via token sparsification for multi-spectral object re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(4): 3633-3648.

[5] Wang Yuhao, Lv Yongfeng, Zhang Pingping, et al. IDEA: Inverted text with cooperative deformable aggregation for multi-modal object re-identification[C]//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2025: 29701-29710.

[6] Guo Jinbo, Zhang Xiaojing, Liu Zhengyi, et al. Generative and attentive fusion for multi-spectral vehicle re-identification[C]//Proceedings of the 7th International Conference on Intelligent Computing and Signal Processing. Xi'an: IEEE, 2022: 1565-1572.

[7] Zheng Aihua, Zhu Xianpeng, Ma Zhiqi, et al. Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark[J]. Information Fusion, 2023, 100: 101901.

[8] Yan Tianying, Ma Huixin, Wang Changhai, et al. Generalizable multi-spectral vehicle re-identification via decoupled subspaces[C]//Proceedings of the Second International Conference on Applied Intelligence. Zhengzhou: Springer, 2024: 13-24.

[9] Wang Zi, Li Chenglong, Li Pengyu, et al. Prototype-based diversity and integrity learning for all-day multi-modal person re-identification[J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 11385-11400.

[10] Pan Wenjie, Wu Hanxiao, Zhu Jianqing, et al. H-ViT: Hybrid vision transformer for multi-modal vehicle re-identification[C]//Proceedings of the 2nd CAAI International Conference on Artificial Intelligence. Beijing: Springer, 2022: 255-267.

[11] Pan Wenjie, Huang Linhan, Liang Jianbao, et al. Progressively hybrid transformer for multi-modal vehicle re-identification[J]. Sensors, 2023, 23(9): 4206.

[12] Wang Yuhao, Liu Xuehu, Zhang Pingping, et al. TOP-ReID: Multi-spectral object re-identification with token permutation[C]//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2024: 5758-5766.

[13] Zheng Aihua, Ma Zhiqi, Sun Yongqi, et al. Flare-aware cross-modal enhancement network for multi-spectral vehicle Re-identification[J]. Information Fusion, 2025, 116: 102800.

[14] Zheng Aihua, Wang Zi, Chen Zihan, et al. Robust multi-modality person re-identification[C]//Proceedings of 35th AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2021: 3529-3537.

[15] Li Hongchao, Li Chenglong, Zhu Xianpeng, et al. Multi-spectral vehicle re-identification: A challenge[C]// Proceedings of 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 11345-11353.

[16] Wang Zi, Li Chenglong, Zheng Aihua, et al. Interact, embed, and EnlargE: Boosting modality-specific representations for multi-modal person re-identification[C]//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2022: 2633-2641.

[17] Zheng Aihua, He Ziling, Wang Zi, et al. Dynamic enhancement network for partial multi-modality person re-identification[PP/OL]. V1. arXiv (2023-05-25)[2025-05-25]. https://arxiv.org/abs/2305.15762.

[18] Wu Di, Liu Zhihui, Chen Zihan, et al. LRMM: Low rank multi-scale multi-modal fusion for person re-identification based on RGB-NI-TI[J]. Expert Systems with Applications, 2025, 263: 125716.

[19] Yang Xi, Dong Wenjiao, Cheng De, et al. TIENet: A tri-interaction enhancement network for multimodal person reidentification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(6): 9852-9863.

[20] Crawford J, Yin Haoli, Mcdermott L, et al. UniCat: Crafting a stronger fusion baseline for multimodal re-identification[PP/OL]. V1. arXiv (2023-10-28)[2025-06-28]. https://arxiv.org/abs/2310.18812.

[21] Zhang Pingping, Wang Yuhao, Liu Yang, et al. Magic tokens: Select diverse tokens for multi-modal object re-identification[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 17117-17126.

[22] Yu Zhi, Huang Zhiyong, Hou Mingyang, et al. WTSF-ReID: Depth-driven window-oriented token selection and fusion for multi-modality vehicle re-identification with knowledge consistency constraint[J]. Expert Systems with Applications, 2025, 274: 126921.

责任编辑 税 红