智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

用于可见－红外行人重识别的模态共享门控双流Transformer

刘智, 李广登, 李诗雨, 王伟, 张小川

引用本文:
刘智, 李广登, 李诗雨, 等. 用于可见－红外行人重识别的模态共享门控双流Transformer[J]. 智能系统学报, 2026, 21(4): 979-987.
LIU Zhi, LI Guangdeng, LI Shiyu, et al. Modality-shared gated two-stream Transformer for visibleinfrared person re-identification[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(4): 979-987.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202511021

DOI: 10.11992/tis.202511021
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20260614.2004.002

用于可见–红外行人重识别的模态共享门控双流 Transformer

刘智[1], 李广登[1], 李诗雨[1], 王伟[2], 张小川[1,3]
（1. 重庆理工大学 两江人工智能学院, 重庆 401135; 2. 军事科学院 战略评估咨询中心, 北京 100091; 3. 重庆工程学院 软件学院, 重庆 400056）

摘　要: 可见光–红外行人重识别旨在匹配不同模态下的同一身份行人图像, 是计算机视觉领域一项极具挑战性的任务。在现实场景中, 光照不足往往导致识别困难, 严重限制了行人重识别系统的实用性。尽管基于卷积神经网络的传统方法已得到广泛研究, 但受限于局部感受野及下采样操作, 这类方法容易造成模态信息的丢失。针对上述问题, 本文提出了一种基于 vision Transformer(ViT) 的可见光–红外行人重识框架——模态共享门控双流 Transformer。针对红外与 RGB 图像间的模态差异, 采用灰度图像增强策略, 将 RGB 图像转换为灰度图以最小化模态间隙。设计模态特有嵌入模块, 引导 ViT 更有效地提取模态共享特征。随后, 利用模态共享模块充分挖掘 RGB、灰度及红外 3 类图像中的模态无关特征。最后, 引入模态共享门控层对特征进行精细筛选, 保留高鉴别力特征以用于后续重识别任务。本文框架能够全面提取多模态信息, 并有效保留模态无关特征。在 SYSU-MM01 和 RegDB 数据集上的大量实验表明, 该方法优于当前主流技术。

关键词: 可见–红外行人重识别; 多模态学习; 视觉 Transformer; 对比语言–图像预训练; 模态共享特征提取; 模态共享门控
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2026)04−0979−09

中文引用格式: 刘智，李广登，李诗雨，等. 用于可见–红外行人重识别的模态共享门控双流 Transformer[J]. 智能系统学报, 2026, 21(4): 979–987.
英文引用格式: LIU Zhi, LI Guangdeng, LI Shiyu, et al. Modality-shared gated two-stream Transformer for visibleinfrared person re-identification[J]. CAAI transactions on intelligent systems, 2026, 21(4): 979–987.

Modality-shared gated two-stream Transformer for visibleinfrared person re-identification

LIU Zhi[1], LI Guangdeng[1], LI Shiyu[1], WANG Wei[2], ZHANG Xiaochuan[1,3]
(1. Liangjiang School of Artificial Intelligence, Chongqing University of Technology, Chongqing 401135, China; 2. Strategic Assessment and Consulting Center, Academy of Military Sciences, Beijing 100091, China; 3. School of Software, Chongqing Institute of Engineering, Chongqing 400056, China)

Abstract: Visible-infrared person re-identification (VI-ReID) aims to match person images of the same identity across different modalities, representing a highly challenging task in the field of computer vision. In real-world scenarios, poor illumination often leads to recognition difficulties, severely limiting the practicality of person re-identification systems. Although traditional methods based on convolutional neural networks have been widely investigated, they are limited by local receptive fields and down-sampling operations, which tend to result in the loss of key modality information. To address these issues, this paper proposes a vision Transformer-based VI-ReID framework, termed modality-shared gating two-stream Transformer (MgtFormer). First, aiming at the modality discrepancy between infrared and RGB images, a grayscale image augmentation strategy is adopted to convert RGB images into grayscale ones to minimize the modality gap. Second, a modality-specific embedding module is designed to guide the ViT to extract modality-shared features more effectively. Subsequently, a modality-shared module is utilized to fully mine modality-invariant features from RGB, grayscale, and infrared images. Finally, a modality-shared gating layer is introduced to finely screen the features, retaining highly discriminative ones for subsequent re-identification tasks. MgtFormer is capable of comprehensively extracting multi-modal information and effectively preserving modality-invariant features. Extensive experiments on the SYSU-MM01 and RegDB datasets demonstrate that the proposed method outperforms current state-of-the-art techniques.

Keywords: visible-infrared person re-identification; multi-modal learning; vision Transformer; contrastive language–image pre-training; modality-shared features extraction; modality-shared gated

行人重识别(person re-identification, ReID)旨在根据给定的查询图像(Query), 在跨摄像头的图

收稿日期: 2025−11−15.　**网络出版日期:** 2026−06−15.
通信作者: 张小川. E-mail: cqpczxc@qq.com.

像库(Gallery)中检索具有相同身份的目标行人[1]。尽管针对单模态可见光图像的 ReID 研究已取得了显著进展[2-11],但传统方法在光照条件较差的监控场景下难以提取有效的判别特征。随着现代监控设备普遍具备在低光环境下自动切换至红外模式的功能,可见光−红外行人重识别(visible-infrared ReID, VI-ReID)逐渐成为研究热点。然而,由于成像波段不同,红外(IR)与可见光(RGB)图像之间存在巨大的模态差异。此外,VI-ReID 还面临色彩缺失和模态未对齐等特有难题,这些挑战与传统 ReID 中存在的视角变化、姿态多样性及遮挡等问题相互叠加,极大增加了任务的难度。

现有的跨模态行人重识别(VI-ReID)方法大多基于卷积神经网络(convolutional neural networks, CNN)[12-15],通常利用 CNN 作为特征提取器,并已得到广泛研究。文献 [16-18] 等提出了以 CNN 为主干的双流架构,利用不共享权值的浅层网络分别提取两种模态的特征,随后通过共享层提取共性特征。然而,随着 Transformer 技术的兴起,其自注意力机制带来的全局感受野相较于 CNN 具有显著优势,并在传统[7,19]及遮挡[20-26]行人重识别领域逐渐取得了更优的性能。

尽管基于 ViT 的 ReID 方法在效果上普遍优于以 ResNet50 为主干的方法,但将其直接应用于 VI-ReID 任务时,与当前最先进的方法相比仍存在差距。因此,设计一种专用于 VI-ReID 任务的 ViT 架构显得尤为必要。众所周知,可见光(RGB)与红外(IR)图像存在显著差异:前者为三通道且色彩丰富,后者为单通道且缺乏色彩,这给 VI-ReID 任务带来了巨大挑战。为缓解模态差异,现有方法普遍采用将红外图像进行通道复制以扩展为三通道图像[27],从而适应网络对颜色信息的处理需求。然而,这种做法往往忽略了对轮廓、纹理等模态间共有特征的有效挖掘。针对上述问题,本文提出模态共享门控双流 Transformer(modality-shared gating two-stream Transformer, MgtFormer)网络结构。首先,为了消除颜色差异的影响,本文将可见光图像转换为灰度图像作为输入,引导网络专注于学习模态间的共性特征。其次,受文献 [28-34] 启发,MgtFormer 采用双流架构:先对输入图像进行模态特有嵌入 (modality-specific embedding),再利用权重共享的 ViT 网络进行特征提取。模态特有嵌入模块能有效保留各模态的独有信息,进而辅助共享网络捕捉模态间的共性。最后,本文设计了模态共享门控层,融合门控线性单元 (gated linear unit, GLU) 与注意力机制,对特征图进行像素级筛选。该机制通过过滤背景杂波,进一步增强了特征的鉴别力。本文的研究为挖掘 ViT 在 VI-ReID 任务中的潜力以及普适性的跨 RGB-IR 数据增强提供了新的见解。研究成果将为非重叠视域下的智能监控以及低光照环境下的 RGB-IR 跨模态匹配提供技术支撑,有助于实现全天候监控系统。本文的主要贡献总结如下:

1) 针对 VI-ReID 任务,提出了 MgtFormer 框架,该框架通过对 RGB 进行灰度增强,融合模态特有嵌入、模态共享、模态共享门控模块提取不同模态间的共性特征。在两个流行 VI-Reid 任务数据集 SYSU-MM01 与 RegDB 上的实验,证明了 MgtFormer 的有效性;

2) 通过对 RGB 图像进行灰度增强并设计模态特有嵌入双流结构,有效保留原有 RGB、IR 图像信息的同时,充分挖掘 RGB、IR 图像间的共性特征,以缓解 VI-ReID 中模态差异问题;

3) 融合门控线性单元与注意力机制,设计模态共享门控模块,对输出的特征图进行像素级特征筛选,驱使 MgtFormer 学习跨模态共有的模态特征,使提取的特征更具有鉴别性。

1 对比语言−图像预训练与图文自监督

1.1 整体网络

MgtFormer 包括灰度增强、模态特有嵌入、模态共享特征提取、模态共享门控 4 个部分,如图 1 所示。灰度增强模块将 RGB 图像转为灰度图像。模态特有嵌入模块包括两个分支,分支一将 RGB 和灰度图像进行 RGB 模态嵌入,分支二将 IR 图像进行 IR 模态嵌入。模态共享特征提取网络,基于传统的 ViT 结构对两个分支的模态嵌入进行融合并提取模态不相关的特征。模态共享门控层关注共性特征、过滤背景信息,突出模态不相关的特征。

对于模型输入,假定 RGB、Gray 和 IR 分别代表可见光、灰度以及红外模态,则可见光、灰度和红外模态数据分别表示为 $\boldsymbol{X}^{\mathrm{RGB}} \in \mathbb{R}^{H\times W\times C}$, $\boldsymbol{X}^{\mathrm{GRY}} \in \mathbb{R}^{H\times W\times C}$, $\boldsymbol{X}^{\mathrm{IR}} \in \mathbb{R}^{H\times W\times C}$。其中 $H\times W$ 为图像分辨率,C 为图像通道数。灰度增强模块将可见光图像生成灰度图像,模态特有嵌入模块对输入的 RGB、灰度、IR 图像进行嵌入特征表示。模块使用 ViT 作为特征提取器,其中的 Transformer 编码层共有 L 层,每一层由多头自注意力机制 (multi-head self-attention, MSA) 模块和多层感知机 (multi-layer perceptron, MLP) 模块组成,在每个模块之前加入层正则化 (layer normalization, LN) 操作。因

此每个编码层的输出可以表示为

$$z'_l = \mathrm{MSA}(\mathrm{LN}(z_{l-1})) + z_{l-1} \quad l \in 1,2,\cdots,L$$

$$z_l = \mathrm{MLP}\left(\mathrm{LN}\left(z'_{l-1}\right)\right) + z'_{l-1} \quad l \in 1,2,\cdots,L$$

最后，模态共享门控模块结合了多头自注意力机制，通过对 Transformer 编码层的输出进行自注意力，并且经过激活层门控，过滤掉背景杂波等噪声信息。

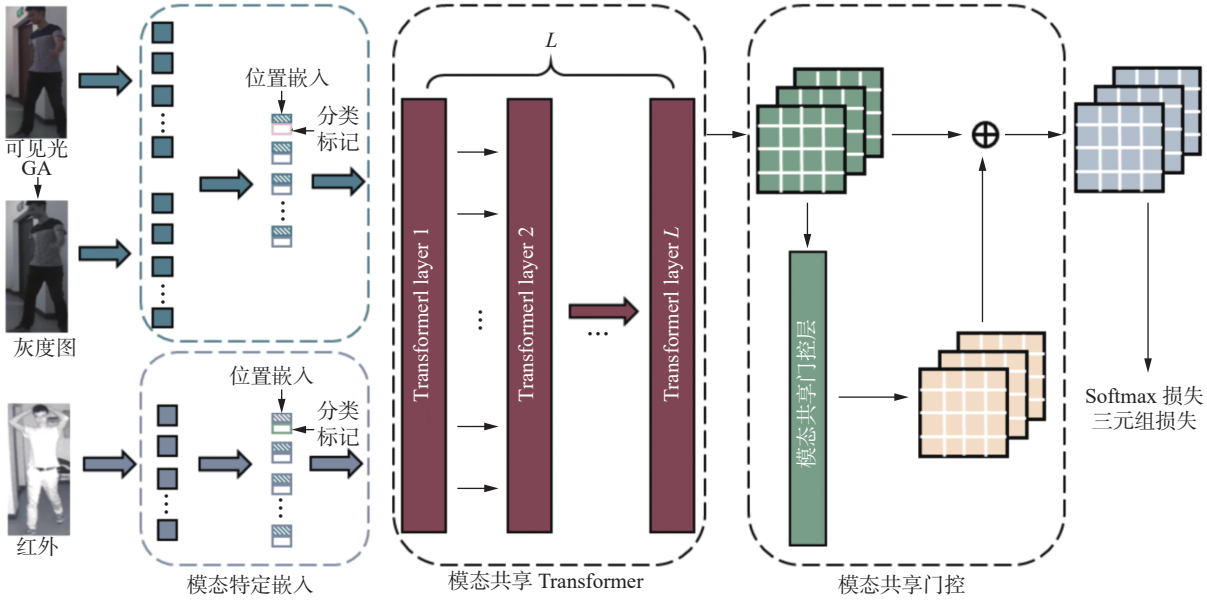

图 1　模态共享门控双流 Transformer

Fig. 1　Modality-shared gating dual-stream Transformer

1.2　灰度图像增强与模态特有嵌入

根据图像捕获的特性，三通道的可见光图像与单通道的红外图像存在着明显的差异，两种模态的图像共享了许多非颜色信息，也称模态无关信息，例如对同一个行人的不同模态图像，其身高、发型、衣着等信息是极其相似的。因此需要一种机制推动和约束模型更多地去关注各种模态无关的特征。灰度增强模块将每个可见光图像转换为灰度图像，通过近似于红外图像的风格，进一步推动网络更多地关注轮廓、纹理方面的信息，削弱其对 RGB 图像中色彩的关注程度。对于可见光图像 $\boldsymbol{X}^{\mathrm{RGB}}$，可通过计算得到灰度图像 $\boldsymbol{X}^{\mathrm{GRY}}$。

$$G_{\mathrm{ray}} = f(R,G,B)$$

式中：R、G、B 分别指的是可见光图像 $\boldsymbol{X}^{\mathrm{RGB}}$ 中某像素在红、绿、蓝 3 个通道的色彩值；$f(\cdot)$是一个灰度变换函数，用于对可见光图像的每个像素中红、绿、蓝 3 个通道的值做一组特定的叠加；G_{ray} 指的是该像素经过红、绿、蓝 3 个通道转换后的灰度值。计算全部像素的灰度值即可得到灰度图像 $\boldsymbol{X}^{\mathrm{GRY}}$。

由于单通道的 IR 与灰度图像均缺乏颜色信息，具有高度的视觉相似性，而灰度图像含有轮廓线条特征。利用灰度图像强化 MgtFormer 网络对模态无关特征提取的敏感性，使模态共享模块最终输出中可以包含更多的模态无关特征。

为学习和捕获每个模态固有的信息和特征，MgtFormer 网络中引入了模态特有嵌入。模态特有嵌入模块将输入图像 $\boldsymbol{X}^{\mathrm{RGB}}$、$\boldsymbol{X}^{\mathrm{GRY}}$、$\boldsymbol{X}^{\mathrm{IR}}$ 分别被划分为 patch 序列 $\boldsymbol{x}_i \in \mathbb{R}^{N\times(P^2\cdot C)}$，$i \in \{\mathrm{RGB},\mathrm{GRY},\mathrm{IR}\}$，每张 patch 分辨率为 $P\times P$。滑动窗口的步长为 S，本文采用 S=P 来对图像划分，N 为 patch 数量，可以表示为

$$N = \left(\frac{H+S-P}{S}\right)\times\left(\frac{W+S-P}{S}\right)$$

然后，使用可学习的线性投影将 patch 序列进行展平映射到 D 维，并将一个额外的可学习分类标记 ([CLS]) 添加到 patch 嵌入。为方便表示，用 $\boldsymbol{E}^{\mathrm{RGB}}_{\mathrm{pose}}$ 表示可见光图像、灰度图像输入中加入可见光模态位置后的嵌入，$\boldsymbol{E}^{\mathrm{IR}}_{\mathrm{pose}}$ 表示红外图像输入中加入 IR 模态位置后的嵌入。灰度图像位置嵌入与 RGB 模态采用相同的位置嵌入。在模态特有嵌入中，RGB 图像分支和 IR 图像分支采用不共享的参数以提取各自的特征。模态特有嵌入的输出可以表示为

$$\boldsymbol{Z}_{\mathrm{RGB}} = \left[\boldsymbol{x}^{\mathrm{RGB}}_{\mathrm{CLS}}\ \ \boldsymbol{x}^{\mathrm{RGB}}_{1}\boldsymbol{E}\ \ \boldsymbol{x}^{\mathrm{RGB}}_{2}\boldsymbol{E}\ \ \cdots\ \ \boldsymbol{x}^{\mathrm{RGB}}_{N}\boldsymbol{E}\right] + \boldsymbol{E}^{\mathrm{RGB}}_{\mathrm{pose}}$$

$$\boldsymbol{Z}_{\mathrm{GRY}} = \left[\boldsymbol{x}^{\mathrm{GRY}}_{\mathrm{cls}}\ \ \boldsymbol{x}^{\mathrm{GRY}}_{p}\boldsymbol{E}\ \ \boldsymbol{x}^{\mathrm{GRY}}_{p}\boldsymbol{E}\ \ \cdots\ \ \boldsymbol{x}^{\mathrm{GRY}}_{p}\boldsymbol{E}\right] + \boldsymbol{E}^{\mathrm{RGB}}_{\mathrm{pose}}$$

$$\boldsymbol{Z}_{\mathrm{RG}} = \{\boldsymbol{Z}_{\mathrm{RGB}},\ \boldsymbol{Z}_{\mathrm{GRY}}\}$$

$$\boldsymbol{Z}_{\mathrm{IR}} = \left[\boldsymbol{x}^{\mathrm{IR}}_{\mathrm{cls}}\ \ \boldsymbol{x}^{\mathrm{IR}}_{p}\boldsymbol{E}\ \ \boldsymbol{x}^{\mathrm{IR}}_{p}\boldsymbol{E}\ \ \cdots\ \ \boldsymbol{x}^{\mathrm{IR}}_{p}\boldsymbol{E}\right] + \boldsymbol{E}^{\mathrm{IR}}_{\mathrm{pose}}$$

$$\boldsymbol{Z} = \{\boldsymbol{Z}_{\mathrm{RG}}\ \ \boldsymbol{Z}_{\mathrm{IR}}\}$$

式中：$\boldsymbol{E}\in\mathbb{R}^{(P^2\cdot C)\times D}$为 patch 嵌入映射投影；$\boldsymbol{E}_{\text{pos}}\in\mathbb{R}^{N\times D}$为位置嵌入；$\boldsymbol{Z}$即为模态特有嵌入模块的输出，送入模态共享特征提取模块提取特征后，进入模态共享门控模块。

1.3　灰度图像增强与模态特有嵌入

为了能够进一步筛选跨模态共有的特征，结合门控线性单元和多头自注意力机制，本文设计了如图 2 所示的模态共享门控模块，旨在提取更具有鉴别性的特征。

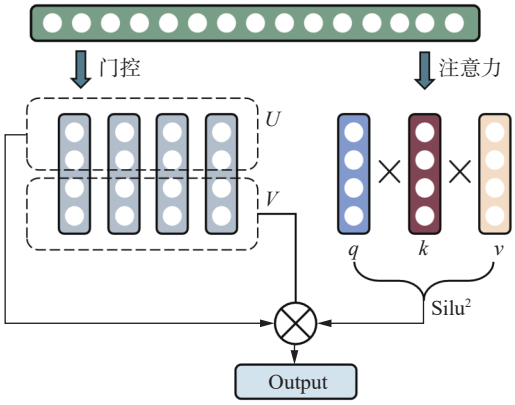

图 2　模态共享门控

Fig. 2　Modality-shared gating

对 Encoder Block 的输出 $\boldsymbol{X}\in\mathbb{R}^{T\times D}$，其中 T 表示词元数量，D 为线性投影层维度。模态共享门控对最终的特征向量先进行自我特征筛选，之后与传统的自注意力矩阵相乘进行门控筛选，门控筛选由门控激活层的激活函数对特征向量进行权重再调整，放大模态共有的特征权重，从而得到最终用来分类的特征向量。可以表示为

$$\boldsymbol{O}=\boldsymbol{A}(\boldsymbol{U}\odot\boldsymbol{V})\boldsymbol{W}_0$$

式中：$\boldsymbol{O}\in\mathbb{R}^{T\times D}$；$\odot$为点乘操作；$\boldsymbol{A}\in\mathbb{R}^{T\times T}$为注意力权重矩阵，融合了词元之间的信息，表示为

$$\boldsymbol{A}=\text{silu}^2\left(\boldsymbol{Q}(Z)K(\boldsymbol{Z})^{\mathrm{T}}+\boldsymbol{b}\right)$$

$$\boldsymbol{Z}=\phi_Z(\boldsymbol{X}\boldsymbol{W}_Z)$$

式中：$\boldsymbol{X}$为输入；$\boldsymbol{Z}\in\mathbb{R}^{T\times s}$；$Q(\cdot)$、$K(\cdot)$为简单的仿射变换，类似于 LayerNorm；$\boldsymbol{b}$ 为偏置项，采用 silu 激活函数。

对于$\boldsymbol{U},\boldsymbol{V}\in\mathbb{R}^{T\times e}$可以表示为

$$\boldsymbol{U}=\phi_U(\boldsymbol{X}\boldsymbol{W}_U)$$

$$\boldsymbol{V}=\phi_V(\boldsymbol{X}\boldsymbol{W}_V)$$

式中：$\boldsymbol{W}_U\in\mathbb{R}^{D\times e}$，$\boldsymbol{W}_V\in\mathbb{R}^{e\times D}$，$e$ 为扩展的中间层维度，ϕ为激活函数。

经由模态共享门控层的筛选作用，可以实现类似门控的效果，通过对特征向量进行自我关注与门控过滤，模态共有的特征的权重将得到放大，而背景杂波等信息则会被过滤掉，进一步提升了网络识别能力。

2　实验结果与分析

2.1　数据集和评价指标

为验证所提出方法的有效性，在两个公共数据集 SYSU-MM01[28] 和 RegDB[35] 上评估了所提出的方法。

SYSU-MM01[28] 中使用了 2 个可见光和 4 个红外摄像机，其中含有室内和室外场景。训练集包含 395 个身份的 22 258 张可见光图像和 11 909 张红外图像。测试集包含了 96 个身份，3 803 张红外图像供查询集，301 张随机选择的可见光图像作为图库集。该数据集有全搜索和室内搜索两种查询模式。全搜索模式下，使用所有场景的可见光图像作为图库集，红外图像作为查询集。室内搜索模型下，使用室内场景的可见光图像作为图库集。

RegDB[35] 由一对可见光和红外相机构建。它包含了 412 个身份的 8 240 张图像，每个身份都有 10 张可见光图像和 10 张红外图像。数据集被随机分成 206 个身份的训练集和 206 个身份的测试集。该数据集包含两种测试模式，对于 Vis-to-Therm 模式，使用可见光图像作为图库集，红外图像作为查询集。Therm-to-vis 模式与之相反。

与绝大多数其他工作一致[28, 36-46]，本文遵循了 Re-ID 社区的惯例。采用累积匹配特征（CMC）[47] 曲线和平均精度（mAP）[48] 来评估不同 Re-ID 模型的性能。

2.2　实验设置

将所有行人图像的尺寸统一调整为 288 像素×144 像素。在训练阶段，采用填充、随机裁剪、水平翻转及随机擦除策略对图像进行数据增强。所有实验均在单张 NVIDIA RTX 3090 GPU 上完成。ViT 模型在 ImageNet-21K 数据集上进行预训练，随后在 ImageNet-1K 上进行了微调。为了获得最佳模型性能，本文参考主流方法设置并经由验证集上的经验调试，确定了以下超参数配置。在模态特有特征嵌入中采用步距为 16 的滑动窗口，批处理大小为 32。对于每个批次，随机抽取 8 个身份和每个身份的 4 张图像。采用 SGD 优化器，动量为 0.9 和 5×10^{-4} 的重量衰减。初始学习速率设置为 0.1，在 20、50 个 epochs 时衰减 0.1 和 0.01。在前 10 个 epochs 时采用了预热策略。

2.3　与最先进模型的比较

表 1 给出了 MgtFormer 方法在 SYSU-MM01 数据集上与现有主流方法 DZP[28](deep zero-padding)、eBDTR[36](bi-directional center-constrained

top-ranking)、Hi-CMD[37](hierarchical cross-modality disentanglement)、JSIA[46]、MPMN[38](deep multi-patch matching network)、LbA[39](learning by aligning)、NFS[40](neural feature search)、DFLN-ViT[44](discriminative feature learning network based on a visual Transformer)、cm-SSFT[18](cross-modality person re-identification with shared-specific feature transfer)、USLVI-ReID[49](unsupervised learning visible-infrared person re-identification、ZS-IDA[50](zero-shot infrared domain adaptation) 的对比结果。其中，DZP[28] 是 VI-ReID 任务中的经典基准方法；JSIA[46] 利用生成对抗网络 (generative adversarial network, GAN) 生成跨模态配对图像；DFLN-ViT[44] 引入 Transformer 架构以增强特征学习；cm-SSFT[18] 则通过模态互补策略提取模态共享与特有特征。

表 1　在 SYSU-MM01 数据集上与最先进模型的性能比较

Table 1　Performance comparison with state-of-the-art models on SYSU-MM01 datasets　%

查询模式	全场景搜索				室内场景搜索			
方法	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
DZP[28]	14.80	54.12	71.33	15.95	20.58	68.38	85.79	26.92
eBDTR[36]	27.80	67.34	81.34	28.42	32.46	77.42	89.62	42.46
Hi-CMD[37]	34.90	77.60	—	35.90	—	—	—	—
JSIA[46]	38.10	80.70	89.90	36.90	43.80	86.20	94.20	52.90
MPMN[38]	48.98	90.33	97.13	62.41	64.89	**96.85**	**99.22**	76.47
LbA[39]	55.41			54.14	61.02			67.98
NFS[40]	56.91	91.34	96.52	55.45	62.79	96.53	99.07	69.79
DFLN-ViT[44]	59.84	92.49	97.20	57.70	62.13	94.83	98.24	69.03
cm-SSFT[18]	61.60	89.20	93.90	**63.20**	70.50	94.90	97.70	72.60
USLVI-ReID[49]	50.36	89.02	95.92	47.36	53.47	92.24	97.84	61.73
ZS-IDA[50]	37.97	51.26	52.73	37.25	34.96	51.39	53.23	34.14
本文模型	**64.50**	**94.11**	**97.92**	62.51	**71.10**	96.20	99.20	**76.90**

注：加粗表示该结果在本列效果最好。

与上述方法相比，MgtFormer 在最具挑战性的全场景搜索模式下取得了显著提升，Rank-1 准确率达到 64.5%，mAP 达到 62.51%。在室内场景搜索模式下，Rank-1 准确率和 mAP 分别达到 71.1% 和 76.9%。尽管 MgtFormer 在室内模式的主要指标上保持领先，但在 Rank-10 等次要指标上略逊于 MPMN[38]。这主要是因为室内场景相对简单，背景干扰较小，基于局部特征对齐的方法（如 MPMN）在解决此类简单样本时已趋于性能饱和；而 MgtFormer 基于 Transformer 的全局注意力机制更侧重于解决遮挡和复杂背景下的模态差异问题，因此在全场景模式下的优势更为明显。

表 2 为 MgtFormer 方法在 RegDB 数据集上与当前已有方法的比较。在该数据集上，本方法优于以前最先进的方法。然而，在红外到可见光模式下，本方法的 Rank-1（75.58%）和 mAP（70.1%）略低于 DTRM[43] 和 cm-SSFT[18]。造成这一现象的原因可能是：首先，RegDB 数据集规模较小，而 Transformer 架构通常需要比 CNN更大规模的数据来打破归纳偏置（Inductive Bias）以学习更优的特征表示；其次，当红外图像作为查询图像时，缺失了色彩纹理信息，使得模型难以充分利用预训练阶段学到的颜色语义，导致在该特定检索方向上的性能受到一定限制。尽管如此，MgtFormer 依然在该数据集上取得了具有竞争力的结果。

表 2　在 RegDB 数据集上与最先进模型的性能比较

Table 2　Performance comparison with state-of-the-art models on RegDB datasets　%

查询模式	可见光到红外				红外到可见光			
方法	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
DZP[28]	17.75	34.21	44.35	18.90	16.63	34.68	44.25	17.82
eBDTR[36]	34.62	58.96	69.72	33.46	34.21	58.74	68.64	32.49
JSIA[46]	48.50			49.30	48.10			48.90

续表 2

查询模式	可见光到红外				红外到可见光			
方法	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Hi-CMD[37]	70.93	86.39		66.04				
LbA[39]	74.17			67.64	72.43			65.46
DTRM[43]	79.09	92.25	95.66	70.09	78.02	91.75	95.19	69.56
NFS[40]	80.54	91.96	95.07	72.10	77.95	90.45	93.62	69.79
cm-SSFT[18]	72.30			72.90	71.00			71.70
VT-GE[51]	73.00	88.10	94.40	67.70	72.80	87.10	90.10	66.20
本文模型	79.80	**93.00**	**95.70**	**73.50**	75.58	91.20	94.90	70.10

注：加粗表示该结果在本列效果最好。

2.4　消融实验

为说明 MgtFormer 框架中各模块的有效性，本文在 SYSU-MM01 数据集上进行了消融实验。结果如表 3 所示。其中 AGW 是以 ResNet-50 为骨干的双流主干网络，ViT 使用 Vision Transformer 作为骨干网络，但未加入模态特有嵌入模块。

表 3　在 SYSU-MM01 数据集上的消融研究

Table 3　Ablation study over SYSU-MM01

模型	灰度增强	模态门控	全场景搜索		室内场景搜索	
			Rank-1	mAP	Rank-1	mAP
AGW[41]（ResNet-50）			47.50	47.65	54.17	62.97
ViT			53.90	53.60	58.60	67.00
	√		63.40	61.00	70.20	74.60
	√	√	63.70	61.60	70.30	75.30
MgtFormer			55.10	55.00	63.10	69.90
	√		64.00	61.80	70.70	75.30
	√	√	64.50	62.51	71.10	76.90

从表 3 可以看到，MgtFormer 与 ViT 相比，在室内场景搜索模式下 mAP 提升了 2.9 百分点。模态特有嵌入与模态共享特征提取网络可以取得比单独使用 ViT 或者 ResNet-50 更好的性能，这说明模态特有嵌入与模态共享特征提取网络能够有效保留原有 RGB、IR 图像信息并且提取到更全面的不同模态的信息。

从表 3 可以看到在 ViT 和 MgtFormer 使用灰度增强均能有较大的提升。在全场景搜索模式下，ViT 的 mAP 提升了 7.4 百分点，cm-ViT 的 mAP 提升了 6.8 百分点，这说明灰度增强能够有效地减轻数据集模态间的差异。MgtFormer 使用模态共享门控层后在全场景搜索模式下 mAP 提升了 0.71 百分点，这说明设计的模态共享门控层能够有效驱使网络学习到跨模态共有的模态特征。结果表明，本文模型可以达到最佳的性能，在 SYSU-MM01 上的全场景搜索模式比 ResNet-50 方法的 Rank-1 提升了 17 百分点、mAP 提升 14.86 百分点，在室内场景搜索模式下 Rank-1 提升了 16.93 百分点、mAP 提升 13.93 百分点，这表明了我们的模型的有效性。

3　结束语

本文提出了一种基于 Transformer 的行人重识别网络框架 MgtFormer。利用 RGB 图像生成灰度图像，将 RGB、灰度、IR 图像同时作为模型的输入，并采用双流结构的模态特有嵌入模块分别对 RGB(包括灰度) 和 IR 进行特征嵌入，不仅有效保留了原始图像信息，也缓解了不同模态之间的视觉差异。基于传统 ViT 架构的模态共享特征提取模块能更好提取来自 RGB 模态与 IR 模态的信息。模态共享门控层进一步筛选具有鉴别性的特征，驱使网络学习跨模态共有的模态特征。SYSU-MM01 和 RegDB 数据集上的实验结果表明，MgtFormer 方法优于当前的最先进的方法。在 SYSU-MM01 数据集全场景搜索模式，实现了

64.5% 的 Rank-1(至少提升 2.9 百分点) 和 62.51% 的 mAP，在室内场景搜索模式，实现了 71.1% 的 Rank-1(至少提升 0.6 百分点) 和 76.9% 的 mAP；在 RegDB 数据集可见光到红外模式下取得了 73.5% 的 mAP，红外到可见光模式下取得了 70.1% 的 mAP。

参考文献：

[1] ZHENG Liang, YANG Yi, HAUPTMANN A G. Person re-identification: past, present and future[EB/OL]. (2016−10−10)[2025−11−15]. https://arxiv.org/abs/1610.02984.

[2] SUN Yifan, ZHENG Liang, YANG Yi, et al. Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline)[C]//Computer Vision–ECCV 2018. Cham: Springer International Publishing, 2018: 501−518.

[3] ZHENG Feng, DENG Cheng, SUN Xing, et al. Pyramidal person re-IDentification via multi-loss dynamic training[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2020: 8506−8514.

[4] JIN Xin, LAN Cuiling, ZENG Wenjun, et al. Style normalization and restitution for generalizable person re-identification[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 3140−3149.

[5] SUN Yifan, CHENG Changmao, ZHANG Yuhan, et al. Circle loss: a unified perspective of pair similarity optimization[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 6397−6406.

[6] NGUYEN B X, NGUYEN B D, DO T, et al. Graph-based person signature for person re-identifications[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville: IEEE, 2021: 3487−3496.

[7] HE Shuting, LUO Hao, WANG Pichao, et al. TransReID: transformer-based object re-identification[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2022: 14993−15002.

[8] REN Min, HE Lingxiao, LIAO Xingyu, et al. Learning instance-level spatial-temporal patterns for person re-identification[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2022: 14910−14919.

[9] LI Dengjie, CHEN Siyu, ZHONG Yujie, et al. DiP: learning discriminative implicit parts for person re-identification[EB/OL]. (2022−12−24)[2025−11−15]. https://arxiv.org/abs/2212.13906.

[10] LI Siyuan, SUN Li, LI Qingli. CLIP-ReID: exploiting vision-language model for image re-identification without concrete text labels[J]. Proceedings of the AAAI conference on artificial intelligence, 2023, 37(1): 1405−1413.

[11] SOMERS V, DE VLEESCHOUWER C, ALAHI A. Body part-based representation learning for occluded person re-identification[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2023: 1613−1623.

[12] TIAN Xudong, ZHANG Zhizhong, WANG Cong, et al. Variational distillation for multi-view learning[EB/OL]. (2022−06−20)[2025−11−15]. https://arxiv.org/abs/2206.09548.

[13] JOSI A, ALEHDAGHI M, CRUZ R M O, et al. Multimodal data augmentation for visual-infrared person ReID with corrupted data[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. Waikoloa: IEEE, 2023: 1−10.

[14] ALEHDAGHI M, JOSI A, CRUZ R M O, et al. Visible-infrared person re-identification using privileged intermediate information[C]//Computer Vision–ECCV 2022 Workshops. Cham: Springer Nature Switzerland, 2023: 720−737.

[15] FENG Jiawei, WU Ancong, ZHENG Weishi. Shape-erased feature learning for visible-infrared person re-identification[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 22752−22761.

[16] YE Mang, WANG Zheng, LAN Xiangyuan, et al. Visible thermal person re-identification *via* dual-constrained top-ranking[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence Main track. Stockholm: IJCAI, 2018: 1092−1099.

[17] HAO Yi, WANG Nannan, LI Jie, et al. HSME: hypersphere manifold embedding for visible thermal person re-identification[C]//Proceedings of the AAAI conference on artificial intelligence, Honolulu: ACM, 2019: 8385−8392.

[18] LU Yan, WU Yue, LIU Bin, et al. Cross-modality person re-identification with shared-specific feature transfer [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 13376−13386.

[19] LUO Hao, WANG Pichao, XU Yi, et al. Self-supervised pre-training for Transformer-based person re-identification[EB/OL]. (2021−11−23)[2025−11−15]. https://arxiv.org/abs/2111.12084.

[20] JIA Mengxi, CHENG Xinhua, LU Shijian, et al. Learning disentangled representation implicitly via Transformer for occluded person re-identification[J]. IEEE transactions on multimedia, 2023, 25: 1294−1305.

[21] DOU Shuguang, ZHAO Cairong, JIANG Xinyang, et al. Human co-parsing guided alignment for occluded person re-identification[J]. IEEE transactions on image processing, 2023, 32: 458−470.

[22] TAN Lei, XIA Jiaer, LIU Wenfeng, et al. Occluded person re-identification via saliency-guided patch transfer [C]//Proceedings of the 2024 AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2024: 5070−5078.

[23] XIA Jiaer, TAN Lei, DAI Pingyang, et al. Attention disturbance and dual-path constraint network for occluded person re-identification[C]//Proceedings of the 2024 AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2024: 6198−6206.

[24] GAO Liying, JIAO Bingliang, LONG Yuzhou, et al. Contrastive pedestrian attentive and correlation learning network for occluded person re-identification[J]. IEEE transactions on circuits and systems for video technology, 2024, 34(9): 8862−8880.

[25] WANG Tao, LIU Mengyuan, LIU Hong, et al. Feature completion transformer for occluded person re-identification[J]. IEEE transactions on multimedia, 2024, 26: 8529−8542.

[26] LI Yanping, LIU Yizhang, ZHANG Hongyun, et al. Occlusion-aware transformer with second-order attention for person re-identification[J]. IEEE transactions on image processing, 2024, 33: 3200−3211.

[27] YE Mang, RUAN Weijian, DU Bo, et al. Channel augmented joint learning for visible-infrared recognition [C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2022: 13547−13556.

[28] WU Ancong, ZHENG Weishi, YU Hongxing, et al. RGB-infrared cross-modality person re-identification[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 5390−5399.

[29] 王晋溪, 鲁鸣鸣. 基于场景图知识的文本到图像行人重识别[J]. 模式识别与人工智能, 2024, 37(11): 947−959.
WANG Jinxi, LU Mingming. Scene graph knowledge based text-to-image person re-identification[J]. Pattern recognition and artificial intelligence, 2024, 37(11): 947−959.

[30] 石瑞鑫, 智敏, 殷雁君. 多模态行人重识别研究综述[J]. 计算机应用研究, 2025, 42(7): 1921−1929.
SHI Ruixin, ZHI Min, YIN Yanjun. Review of multimodal pedestrian re-identification[J]. Application research of computers, 2025, 42(7): 1921−1929.

[31] 李俊峰, 楼琼, 钱亚冠, 等. 基于像素对齐和特征对齐的跨模态行人重识别[J]. 浙江科技学院学报, 2022(3): 251−260.
LI Junfeng, LOU Qiong, QIAN Yaguan, et al. Cross-modality person re-identification based onpixel alignment and feature alignment[J]. Journal of Zhejiang University of Science and Technology, 2022(3): 251−260.

[32] 冯展祥, 赖剑煌, 袁藏, 等. 走向通用行人重识别: 预训练大模型技术在行人重识别的应用综述[J]. 中国图象图形学报, 2025, 30(6): 1638−1660.
FENG Zhanxiang, LAI Jianhuang, YUAN Zang, et al. Advancing universal person reidentification: a survey on the applications of large-scale, pretraining models for identifying individuals[J]. Journal of image and graphics, 2025, 30(6): 1638−1660.

[33] 孙锐, 杜云, 陈龙, 等. 隐式多尺度对齐与交互的文本-图像行人重识别方法[J]. 软件学报, 2025, 36(10): 4846−4863.
SUN Rui, DU Yun, CHEN Long, et al. Implicit multi-scale alignment and interaction for text-image person re-identification method[J]. Journal of software, 2025, 36(10): 4846−4863.

[34] 金昌胜, 王海瑞. 基于关系挖掘的跨模态行人重识别[J]. 空军工程大学学报, 2024, 25(1): 106−114.
JIN Changsheng, WANG Hairui. A cross-modal person re-identification based on relationship mining[J]. Journal of Air Force Engineering University, 2024, 25(1): 106−114.

[35] NGUYEN D T, HONG H G, KIM K W, et al. Person recognition system based on a combination of body images from visible light and thermal cameras[J]. Sensors, 2017, 17(3): 605.

[36] YE Mang, LAN Xiangyuan, WANG Zheng, et al. Bi-directional center-constrained top-ranking for visible thermal person re-identification[J]. IEEE transactions on information forensics and security, 2020, 15: 407−419.

[37] CHOI S, LEE S, KIM Y, et al. Hi-CMD: hierarchical cross-modality disentanglement for visible-infrared person re-identification[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10254−10263.

[38] WANG Pingyu, ZHAO Zhicheng, SU Fei, et al. Deep multi-patch matching network for visible thermal person re-identification[J]. IEEE transactions on multimedia, 2021, 23: 1474−1488.

[39] PARK H, LEE S, LEE J, et al. Learning by aligning: visible-infrared person re-identification using cross-modal

correspondences[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2022: 12026−12035.

[40] CHEN Y, WAN Lin, LI Zhihang, et al. Neural feature search for RGB-infrared person re-identification[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 587−597.

[41] YE Mang, SHEN Jianbing, LIN Gaojie, et al. Deep learning for person re-identification: a survey and outlook[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 44(6): 2872−2893.

[42] WANG Pingyu, SU Fei, ZHAO Zhicheng, et al. Deep hard modality alignment for visible thermal person re-identification[J]. Pattern recognition letters, 2020, 133: 195−201.

[43] YE Mang, CHEN Cuiqun, SHEN Jianbing, et al. Dynamic tri-level relation mining with attentive graph for visible infrared re-identification[J]. IEEE transactions on information forensics and security, 2022, 17: 386−398.

[44] ZHAO Jiaqi, WANG Hanzheng, ZHOU Yong, et al. Spatial-channel enhanced transformer for visible-infrared person re-identification[J]. IEEE transactions on multimedia, 2023, 25: 3668−3680.

[45] WANG Guanan, ZHANG Tianzhu, CHENG Jian, et al. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 3622−3631.

[46] WANG Guanan, ZHANG Tianzhu, YANG Yang, et al. Cross-modality paired-images generation for RGB-infrared person re-identification[C]//The AAAI 2020 proceedings include all papers presented at the 34th AAAI Conference. New York: AAAI, 2020: 12144−12151.

[47] WANG Xiaogang, DORETTO G, SEBASTIAN T, et al. Shape and appearance context modeling[C]//2007 IEEE 11th International Conference on Computer Vision. Rio de Janeiro: IEEE, 2007: 1−8.

[48] ZHENG Liang, SHEN Liyue, TIAN Lu, et al. Scalable person re-identification: a benchmark[C]//2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2016: 1116−1124.

[49] CHENG De, HUANG Xiaojian, WANG Nannan, et al. Unsupervised visible-infrared person ReID by collaborative learning with neighbor-guided label refinement [C]//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM, 2023: 7085−7093.

[50] ZHANG Xu, LIU Yinghui, GUO Liangchen, et al. Zero-shot infrared domain adaptation for pedestrian re-identification via deep learning[J]. Electronics, 2025, 14(14): 2784.

[51] 张红颖, 樊世钰, 罗谦, 等. 结合视觉文本匹配和图嵌入的可见光-红外行人重识别[J]. 电子与信息学报, 2024, 46(9): 3662−3671.
ZHANG Hongying, FAN Shiyu, LUO Qian, et al. Visible-infrared pedestrian recognition combined with visual text matching and graph embedding[J]. Journal of electronics & information technology, 2024, 46(9): 3662−3671.

作者简介：

刘智，副教授，主要研究方向为计算机视觉、机器学习、视频与信号分析。主持或参与国家自然科学基金、重庆市自然科学基金等纵向及企业横向项目 20 余项。获国家发明专利授权 4 项，以第一或通讯作者发表学术论文 20 余篇。E-mail: liuzhi@cqut.edu.cn。

李广登，硕士，主要研究方向为计算机视觉和行人重识别。E-mail: guangdeng.lee@gmail.com。

张小川，教授，CAAI 杰出会员，CAAI 机器博弈专委会主任委员，重庆工程学院智能系统工程中心主任。主要研究方向为软件工程、机器博弈、自然语言处理、机器学习和智能机器人。主持和参与纵向项目 38 项，获省部级自然科学奖等 2 项，出版专著和教材 5 部，发表论文 100 余篇。E-mail: cqpczxc@qq.com。