中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2025)10-3335-11

论文引用格式:Wei S and Yang W L. 2025. Visible-infrared person re-identification algorithm integrating structural and visual features. Journal of Image and Graphics, 30(10):3335-3345(魏思, 杨文璐. 2025. 融合结构与视觉特征的可见光—红外行人重识别. 中国图象图形学报, 30(10):3335-3345)[DOI:10.11834/jig.240600]

融合结构与视觉特征的可见光—红外行人重识别

魏思*,杨文璐

上海海事大学信息工程学院,上海　201306

摘　要:**目的**　可见光—红外行人重识别(visible-infrared person re-identification,VI-ReID)因可见光与红外图像间的模态差异而面临挑战,现有方法在特征分辨力方面存在不足。本研究旨在设计一种全新算法以获取高分辨力的行人特征,弥补跨模态识别任务中的不足。**方法**　提出一种融合结构与视觉特征的VI-ReID算法,通过双流分支进行处理。首先,借助姿态估计提取骨骼关键点生成结构特征图,通过图卷积网络(graph convolutional network,GCN)学习骨骼的结构化信息,以构建结构特征提取分支;同时,以ResNet50(residual network)作为视觉提取分支获取图像视觉特征。在此基础上,提出结构—视觉跨模态注意力机制(structural-visual interactive attention mechanism,SVIAM),融合骨骼和视觉特征,得到高分辨力的联合特征表示。此外,为增强骨骼特征的一致性,提出结构内聚损失(structural cohesion loss,SCLoss)函数,持续优化骨骼特征,有效减少模态内差异,保证算法的稳定性与准确性。**结果**　实验结果表明,所提算法在SYSU-MM01数据集上表现卓越,相较于基线DEEN(diverse embedding expansion network),在all search模式下,Rank-1准确率提高4.21%,mAP(mean average precision)提高3.52%;在indoor search模式下,Rank-1准确率提高7.39%,mAP提高2.56%。**结论**　本研究提出融合结构与视觉特征的VI-ReID算法,有效提升了跨模态行人重识别的识别精度,并在复杂场景中展现较高的鲁棒性和准确性。

关键词:可见光—红外行人重识别(VI-ReID);层次化特征提取;骨骼结构特征;结构—视觉跨模态注意力机制(SVIAM);结构内聚损失(SCLoss)

Visible-infrared person re-identification algorithm integrating structural and visual features

Wei Si*, Yang Wenlu

School of Information Engineering, Shanghai Maritime University, Shanghai　201306, China

Abstract: **Objective** Visible-infrared person re-identification (VI-ReID) has emerged as a challenging task primarily due to the pronounced modal discrepancies between visible and infrared images. In the visible light spectrum, images are repleted with vivid colors and intricate textures, but they are highly susceptible to perturbations caused by varying illumination conditions. For instance, during dawn or dusk, the subdued light can distort the visual appearance of pedestrians, which causes difficulty in accurately discerning their unique features. By contrast, infrared images, which predominantly capture thermal radiation, offer a distinct advantage in low-light or obscured scenarios. However, they lack the detailed visual cues present in their visible counterparts, such as clothing patterns or facial features. These fundamental differences

收稿日期:2024-10-16;**修回日期**:2025-01-04;**预印本日期**:2025-01-11

*通信作者:魏思　202230310007@stu.shmtu.edu.cn

基金项目:国家自然科学基金项目(62102242)

Supported by:National Natural Science Foundation of China (62102242)

have led to great difficulties in achieving reliable person re-identification across modalities. Compounding this issue, existing methods lack sufficient feature discrimination. In numerous real-world datasets and scenarios, they struggle to distinguish between pedestrians with similar postures or occluded body parts, which compromises the overall accuracy and reliability of the recognition process. **Method** To directly address these challenges, this research designed an innovative algorithm capable of extracting high-resolution pedestrian features, with the ultimate aim of bridging the existing gaps in cross-modal recognition tasks. The methodological framework of this study centered around a novel VI-ReID algorithm that incorporated structural and visual features, which operated through a meticulously designed dual-stream branch architecture. The first step in this process involved the extraction of skeletal key points using advanced pose estimation techniques. By leveraging state-of-the-art algorithms, such as OpenPose or custom-developed variants with enhanced capabilities in some cases, we precisely localized the key joints of the human body even in the presence of partial occlusions or extreme postural variations. This accuracy was achieved through various complex computational steps, which started with the initial detection of body regions, followed by the refinement of joint positions based on anatomical constraints and probabilistic models. The extracted skeletal key points were then used to generate detailed structural feature maps, which served as the foundation for further analysis. Subsequently, a graph convolutional network (GCN) was employed to delve deep into the structured information encapsulated within the skeletal framework. The GCN architecture was meticulously designed, which comprised multiple layers, with each having a carefully calibrated node connection pattern. The choice of activation functions was optimized to enhance the propagation of relevant information while suppressing noise. Weighted processing was implemented to account for the varying importance of different joints in characterizing a person's gait or posture, which ensured that the most discriminative features were emphasized. This comprehensive approach enabled the construction of a highly effective structural feature extraction branch. Simultaneously, ResNet50, which is a popular deep learning model renowned for its prowess in visual feature extraction, was adapted to serve as the visual extraction branch. In this context, several fine-tuning procedures were conducted to tailor the model to the unique characteristics of visible and infrared images. This optimization involved adjusting the parameters of the pretrained model based on the statistical properties of the target datasets, as well as devising innovative strategies for leveraging the outputs from different hierarchical levels of the network. For instance, by selectively combining features from the early and late layers, we captured low-level details and high-level semantic information. In certain cases, attention mechanisms were integrated to further focus on the most salient visual regions, which enhanced the overall discriminative power of the visual features. Based on the two parallel streams, a structure-visual inter-modal attention mechanism (SVIAM) was proposed to seamlessly fuse the skeletal and visual features. This mechanism was underpinned by a sophisticated computational process that involved the precise calculation of correlation weights between the two modalities. Through the use of detailed schematics and mathematical formulas, we illustrated the attention distribution, which highlighted the most relevant regions and features for recognition. Compared with simplistic concatenation or rudimentary fusion methods, SVIAM demonstrated a remarkable superiority in terms of feature integration, which led to a more cohesive and discriminative joint feature representation. Furthermore, to bolster the consistency of the skeletal features and mitigate intra-modal differences, a structure cohesion loss (SCLoss) function was devised. The mathematical formulation of SCLoss was derived with great care, with the geometric and topological properties of the skeletal data being considered. Each parameter within the function was meticulously calibrated to serve a specific purpose: whether to penalize deviations from the expected skeletal structure or encourage the alignment of related joints. Through extensive experimental validation and theoretical analysis, we demonstrated the mechanism by which SCLoss effectively optimized the skeletal features, which enhanced the overall stability and accuracy of the algorithm. **Result** Experimental results provided unequivocal evidence of the superiority of the algorithm. On the widely recognized SYSU-MM01 dataset, our proposed algorithm outperformed the baseline DEEN by significant margins. In the all-search mode, the Rank-1 accuracy rate remarkably improved by 4.21%, while the mean average precision (mAP) soared by 3.52%. Similarly, in the indoor search mode, the Rank-1 accuracy rate achieved an even more impressive increase of 7.39%, which was accompanied by a 2.56% elevation in mAP. These results not only validated the effectiveness of our approach in enhancing cross-modal person re-identification accuracy but also showcased its robustness and reliability in complex scenarios. **Conclusion** This research introduced a pioneering VI-ReID algorithm that integrated structural and

visual features, which effectively addressed the challenges posed by modal differences and substantially elevated the recognition precision in cross-modal person re-identification. The performance of the algorithm in complex and dynamic environments further attested to its high level of robustness and accuracy, which lays the foundation for future advancements in this critical area of research.

Key words: visible-infrared person re-identification (VI-ReID); hierarchical feature extraction; skeletal structure features; structural-visual interactive attention mechanism (SVIAM); structural cohesion loss (SCLoss)

0 引　言

在智能监控与公共安全等实际应用场景需求的驱动下，可见光—红外行人重识别(visible-infrared person re-identification，VI-ReID)迅速发展，逐渐成为行人重识别领域的关键研究方向。VI-ReID面临的核心难题，源于可见光图像与红外图像间显著的模态差异：可见光图像依赖环境光，能够清晰呈现物体的颜色和纹理；而红外图像通过捕捉物体的热辐射信息，在低光或夜间环境中发挥优势，与可见光图像形成互补，更好地服务于VI-ReID。因此，如何巧妙且有效地弥合两模态图像的差异，实现行人的精准识别与匹配，是VI-ReID领域迫切需要解决的关键技术问题。

深度学习技术在图像分类(Khandelwal等，2022；Liang等，2023；梁志军和刘栋，2021；Hu等，2023；He等，2023)、目标检测(Nayeem等，2022)等诸多领域取得了显著成果，并逐步拓展至VI-ReID领域，催生出不同的研究路径。在VI-ReID的研究进程中，基于深度学习的主流路线通常采用卷积神经网络(convolutional neural network，CNN)(Kim，2014；Kipf和Welling，2016；张勃兴 等，2022)提取图像特征，并借助度量学习或分类学习(Ye等，2021)实现行人识别。

起初，研究人员将目光聚焦于全局特征对齐，尝试构建共享特征空间来减少可见光和红外图像的模态差异(杨磊，2023)，然而，这类方法存在明显弊端，在捕捉各模态独特特征方面能力有限，在面对复杂场景中那些细微变化时，往往难以精准地捕捉关键信息。在意识到全局特征对齐的局限性后，部分研究者另辟蹊径，探索模态转换与多模态特征融合的方法，利用生成对抗网络(generative adversarial network，GAN)，将红外图像转换为中间模态，例如朱敏等人(2022)尝试转换为灰度图像，随后再结合风格迁移和注意力机制，力求提升识别性能，Xia等人(2021)的研究便是这一思路的典型代表。然而，GAN网络的对抗训练在跨模态任务中稳定性欠佳，容易导致生成图像的质量不稳定，这在很大程度上限制了该方法在复杂实际场景中的广泛应用。

面对前两类方法遭遇的困境，人体骨骼结构和姿态信息(李帅和蒋敏，2023)因其几何稳定性和模态非相关性，为研究者带来了新的曙光，结构信息建模逐渐走入人们的视野。在动作识别(何智敏和许佳云，2023；白忠玉 等，2023)、手势识别(杨和稳 等，2018)以及单模态行人重识别(Lyu等，2024)等领域，基于骨骼特征的方法(刘婧，2021；冯赛楠，2023)已经取得一定进展。如，Rao等人(2024)将骨骼信息建模应用于单模态行人重识别，成功提高识别的准确性，这无疑为VI-ReID研究提供了新的借鉴方向。

在VI-ReID中，骨骼特征作为一种模态无关的几何特征，具有结构稳定、可迁移的特性，使用骨骼特征可以捕捉行人独特的几何结构，但要完成行人识别与匹配，需要实现可见光和红外模态间的有效特征对齐。

鉴于此，本研究将骨骼信息建模引入VI-ReID任务，提出一种融合结构与视觉特征的全新算法。具体而言，在结构层次上运用姿态估计技术提取骨骼关键点，生成关节和肢体的热图表示，结合图卷积网络(graph convolutional network，GCN)学习骨骼的结构化特征表示；在视觉层次上，采用ResNet50(residual network)(He等，2016)提取图像的视觉特征。与此同时，提出结构—视觉跨模态注意力机制(structural-visual interactive attention mechanism，SVIAM)，打破骨骼结构与图像视觉特征之间的信息壁垒，深度融合骨骼结构特征和图像视觉特征，生成联合特征表示。此外，为进一步强化骨骼特征的一致性，引入结构内聚损失(structural cohesion loss，SCLoss)，通过联合监督训练，有机结合损失函数，全

方位优化跨模态行人重识别的综合性能。

1 本文方法

本文方法的总体框架如图1所示，包括3个主要模块：全局特征提取模块、热图特征提取模块以及结构—视觉注意力机制融合模块，在baseline（Zhang和Wang，2023）模型基础上进行改进，以提升跨模态行人重识别性能。

1.1 全局特征提取模块

采用预训练的ResNet50网络（He等，2016），分别对可见光和红外模态的输入图像进行处理，移除ResNet50的最后两层网络，加入最大平均池化层，以获取紧凑的全局特征向量，得到两模态的全局特征图，表示为

$$\boldsymbol{F}_{\text{cnn-v}} \in \mathbf{R}^{C\times H\times W} \tag{1}$$

$$\boldsymbol{F}_{\text{cnn-t}} \in \mathbf{R}^{C\times H\times W} \tag{2}$$

式中，$\boldsymbol{F}_{\text{cnn-v}}$代表可见光模态的全局特征，$\boldsymbol{F}_{\text{cnn-t}}$代表红外模态的全局特征，$C$、$H$和$W$分别代表图像通道数、高度和宽度。

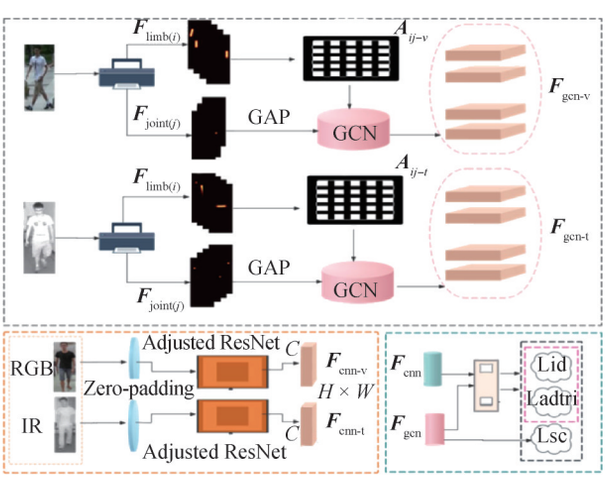


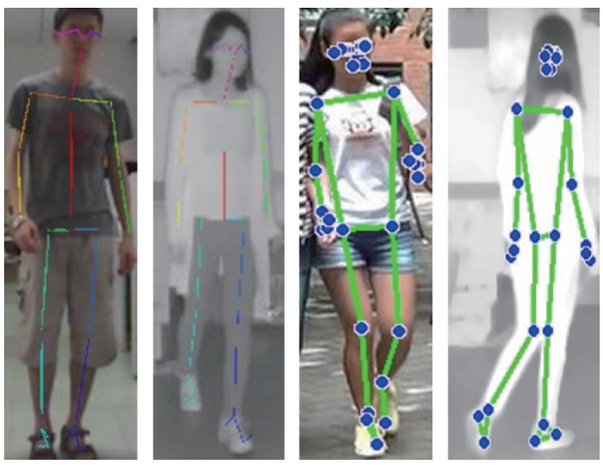
图1 总体框架

Fig. 1 The overall framework

1.2 热图特征提取模块

1.2.1 骨骼点数据获取

考虑到数据集中红外图像的分辨率较低，为匹配获取更多样本图像的骨骼信息，采用OpenPose和MediaPipe两种方法结合进行骨骼点提取，通过对原始图像进行噪声去除等图像增强（Lyu等，2024）后，补充获取更多的骨骼样本。部分示例如图2所示，具体样本量如表1所示。随后，对提取的骨骼点数据进行整合和对齐，以确保数据的一致性和准确性。



(a) OpenPose提取的可见光图像　(b) OpenPose提取的红外骨架图像　(c) MediaPipe提取的可见光图像　(d) MediaPipe提取的红外骨架图像

图2 部分骨骼提取图像

Fig. 2 Partial skeleton extraction images ((a) visible light image extracted by OpenPose; (b) infrared skeleton image extracted by OpenPose; (c) visible light image extracted by MediaPipe; (d) infrared skeleton image extracted by MediaPipe)

表1 图像增强前后样本量

Table 1 Sample sizes before and after image enhancement

数据集	模态	原始数量	图像增强前	图像增强后
SYSU-MM01	可见光	30 071	13 356	15 455
	红外	15 792	3 945	5 526
RegDB	可见光	4 120	1 778	2 095
	红外	4 120	989	1 434

由于人脸区域的骨骼点对行人重识别贡献有限，本研究采用大框架（头部以下、脚踝以上等12点）的骨骼数据作为有效骨骼点，通过获取的关节点坐标情况，计算出各肢体部分长度、对应比例以及实时的方向信息，用于生成关节和肢体热图。

1.2.2 特征提取

如图2所示，获取到的骨骼点数据通过一定的拓扑关系构成，可以视为以关节点为节点、肢体为连接的图结构。使用骨骼热图，结合GCN处理骨骼热图数据（Lyu等，2024）。

具体而言，将关节热图和肢体热图按照同一行人的12个关节点进行堆叠；接着对堆叠后的关节数据进行全局平均池化，得到节点集$\boldsymbol{V}$；同时，将肢体热图及其连接关系作为边集$\boldsymbol{E}$，形成输入图结构。

采用边缘预测（Lyu等，2024）更新邻接矩阵。

获取两个模态结构上的特征输出，表示为

$$\boldsymbol{f}_{\text{gcn-v}} = GCN(\boldsymbol{Rv}_{\text{joint}}, \boldsymbol{Av}_{\text{limb}}) \tag{3}$$

$$\boldsymbol{f}_{\text{gcn-t}} = GCN(\boldsymbol{Rt}_{\text{joint}}, \boldsymbol{At}_{\text{limb}}) \tag{4}$$

式中，$\boldsymbol{Rv}_{\text{joint}}$和$\boldsymbol{Rt}_{\text{joint}}$分别表示堆叠后可见光和红外的关节热图；$\boldsymbol{Av}_{\text{limb}}$和$\boldsymbol{At}_{\text{limb}}$分别表示两个模态的更新邻接矩阵。

1.3 结构—视觉注意力机制融合模块

由于结构特征与视觉特征获取原理存在差异，因此，在融合上，设计结构—视觉注意力机制，将从图卷积网络（GCN）和卷积神经网络（CNN）中提取的结构特征与视觉特征通过分特征进行融合，与简单拼接融合（Zhang和Wang，2023）相比，获取的特征能够协同增效、相互促进，具体处理如下：

首先，针对可见光—可见光和红外—红外的同模态特征，计算两种特征在模态内的相关性矩阵$\boldsymbol{C}$，用以衡量每个结构特征与每个视觉特征之间的相似度，计算为

$$C_{ab\text{-v}} = \frac{(\boldsymbol{f}^{a}_{\text{gcn-v}})^{\text{T}}(\boldsymbol{F}^{b}_{\text{cnn-v}})}{\left\| \boldsymbol{f}^{a}_{\text{gcn-v}} \right\|^{2} \left\| \boldsymbol{F}^{b}_{\text{cnn-v}} \right\|^{2}} \tag{5}$$

$$C_{ab\text{-t}} = \frac{(\boldsymbol{f}^{a}_{\text{gcn-t}})^{\text{T}}(\boldsymbol{F}^{b}_{\text{cnn-t}})}{\left\| \boldsymbol{f}^{a}_{\text{gcn-t}} \right\|^{2} \left\| \boldsymbol{F}^{b}_{\text{cnn-t}} \right\|^{2}} \tag{6}$$

式中，$\boldsymbol{f}^{a}$和$\boldsymbol{F}^{b}$分别对应代表可见光、红外的第a个结构特征和第b个图像特征。

接着，使用softmax 函数（黄光红 等，2022）对相关性矩阵中每一行进行归一化，生成注意力权重$\boldsymbol{A}$，计算为

$$\boldsymbol{A}_{ab\text{-v}} = \frac{\text{e}^{C_{ab\text{-v}}}}{\sum_{q=1}^{w} \text{e}^{C_{aq\text{-v}}}} \tag{7}$$

$$\boldsymbol{A}_{ab\text{-t}} = \frac{\text{e}^{C_{ab\text{-t}}}}{\sum_{q=1}^{w} \text{e}^{C_{aq\text{-t}}}} \tag{8}$$

式中，$\boldsymbol{A}_{ab\text{-v}}$和$\boldsymbol{A}_{ab\text{-t}}$分别表示第$a$个结构特征和第$b$个视觉特征之间的注意力权重，$w$代表图像特征的空间维度。

最后，依据注意力权重$\boldsymbol{A}$对特征进行加权融合，具体为

$$\boldsymbol{F}^{a}_{v} = \sum_{b=1}^{w} (\boldsymbol{A}_{ab\text{-v}} \times \boldsymbol{F}^{b}_{\text{cnn-v}}) \tag{9}$$

$$\boldsymbol{F}^{a}_{t} = \sum_{b=1}^{w} (\boldsymbol{A}_{ab\text{-t}} \times \boldsymbol{F}^{b}_{\text{cnn-t}}) \tag{10}$$

式中，等号左侧的$\boldsymbol{F}$特征分别是第a个结构特征在$\boldsymbol{A}$加权融合下的结果。

1.4 损失函数

1.4.1 结构内聚损失函数

在人体骨骼图中，通常相邻的关节（如肩关节和肘关节）之间具有高度的结构关联性。为捕捉这种结构关系，并确保相邻节点的特征更加一致，本研究提出结构内聚损失。该损失的目标是使得在图卷积网络（GCN）的输出中，相邻节点的特征在特征空间中的距离最小化，以避免大幅波动或不一致性。

为反映结构关联性，使用拉普拉斯矩阵$\boldsymbol{L}$结合GCN 的输出特征矩阵$\boldsymbol{X}$进行定义。具体为

$$L_{\text{sc}} = \frac{1}{2} \sum_{i,j} A_{ij} \left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|^{2} = \text{tr}(\boldsymbol{x}^{\text{T}} \boldsymbol{L} \boldsymbol{x}) \tag{11}$$

式中，A_{ij}是邻接矩阵的元素，有连接值为1，否则为0；$\boldsymbol{x}_i$和$\boldsymbol{x}_j$分别表示节点i和节点j的特征向量，这里指两个模态经过GCN后得到的特征向量，tr(·)表示矩阵的迹，即矩阵对角线元素的和。$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$是图的拉普拉斯矩阵，$\boldsymbol{D}$是度矩阵，表示每个节点的度数。

1.4.2 总损失函数

将结构内聚损失函数结合经典的损失函数，使用提出的平滑的标签损失L_{id}（Radenović 等，2019）和自适应权重的困难三元组损失L_{adtri}（黄驰涵和沈肖波，2024）进行联合监督训练，通过最小化总和来优化网络模型，总损失函数的表述为

$$L_{\text{total}} = L_{\text{id}} + L_{\text{adtri}} + L_{\text{sc}} \tag{12}$$

2 实验结果与分析

2.1 数据集和评价指标

本研究对所提出的方法在两个公开数据集中进行实验验证，其中包括SYSU-MM01（Wu等，2017）和RegDB（Nguyen等，2017）。

SYSU-MM01 数据集包含 491 名行人的多幅图像数据，其中彩色可见光图像有 287 628 幅，红外图像有 15 792 幅。该数据集采用 4 台可见光摄像机（cam1、cam2、cam4、cam5）和 2 台红外摄像机（cam3、cam6），图像采集涵盖室外和室内场景。SYSU-MM01还提供两种搜索模式：全搜索（all search）和室内搜索（indoor search）。全搜索模式使用所有可见光摄像机（cam1、cam2、cam4、cam5）进行匹配，而室内搜索模式则仅使用cam1和cam2进行匹配。这是可见光—红外行人重识别领域常用的大规模数

据集。

RegDB数据集是一个由模拟环境采集的图像数据集，该数据集采用两个摄像头，拍摄获取412个不同行人的图像序列，其中针对每个行人分别采集10幅可见光图像和10幅红外图像。RegDB数据集提供两种评估模式，分别是可见光到红外（visible to infrared, V2I）和红外到可见光（infrared to visible, I2V）。

为评估本研究模型的性能，实验采用平均精度均值（mean average precision, mAP）、正确排名准确率（Rank-k）和最小反向距离平均精度（mean inverse negative penalty, mINP）（Ye等，2021）等指标，全面分析和评估模型性能，确保模型系统的稳健性和准确性。

2.2 实验配置

本研究模型基于PyTorch深度学习框架进行实验，使用NVIDIA 3090 24 GB GPU训练。采用ResNet50作为视觉特征提取器，双层GCN作为骨骼特征提取器。每次输入8个行人ID（identification）作为一个批次，其中每个ID各加载4幅红外图像和4幅可见光图像。

在训练阶段，考虑到热图采集后各ID图像的样本量不均匀情况，采用SYSU-MM01中的387个ID作为训练集的行人ID，同步加载各ID下的图像对应热图。在预训练阶段使用SGD（stochastic gradient descent）作为优化器，动量参数为0.9，预热学习策略，初始学习率为0.1，分别在第20次迭代和第80次迭代时衰减到0.01和0.001，总的迭代次数是105次。

2.3 主流方法对比

在SYSU-MM01和RegDB数据集上，将本研究所提方法与当前多种最新方法进行详细的对比实验，以证明本研究方法的优越性，具体如表2和表3所示。为便于分析，以是否采用骨骼数据或姿态数据，将各方法分为骨骼姿态相关和骨骼姿态不相关两类。

在SYSU-MM01数据集中，本模型在all search和indoor search模式下进行评估；在all search模式下Rank-1高达78.91%，超过所有对比方法。在indoor search模式下，针对较难样本挖掘的mINP值达到

表2 在SYSU-MM01数据集上的方法对比

Table 2 Comparison of methods on the SYSU-MM01 dataset

/%

	方法	all search					indoor search				
		Rank-1	Rank-10	Rank-20	mAP	mINP	Rank-1	Rank-10	Rank-20	mAP	mINP
骨骼不相关	TSME（Liu等，2022）	64.23	95.19	98.73	61.21	–	64.80	96.92	99.31	71.53	–
	SFANet（Liu等，2023）	65.74	92.98	97.05	60.83	–	71.60	96.60	99.45	80.05	–
	PMT（Lu等，2023）	67.53	95.36	98.64	64.98	51.86	71.66	96.73	99.25	76.52	72.74
	SIDA（Gong等，2023）	68.36	95.91	98.56	64.19	–	73.28	97.35	99.52	77.49	–
	MFCS（Yang等，2024）	70.59	96.22	98.77	67.49	–	75.98	98.12	99.62	80.24	–
	MSALNet（Zhang等，2024a）	71.16	96.69	99.02	68.01	54.62	79.35	98.45	99.65	82.84	79.33
	DEEN（Zhang和Wang，2023）	74.70	97.60	99.20	71.80	–	80.30	99.00	99.80	83.30	–
	MUN（Yu等，2023）	76.24	–	97.84	73.81	–	79.42	–	98.09	82.06	–
	SGIEL（Feng等，2023）	77.12	97.03	99.08	72.33	–	82.07	97.42	98.87	82.95	–
	FDNM（Zhang等，2025）	77.80	**97.80**	**99.60**	75.10	–	87.30	**99.40**	**100.00**	**89.10**	–
骨骼相关	TGLFC-Net（Guo等，2024）	59.10	90.90	96.17	54.86	–	60.98	93.05	98.12	66.83	–
	PAGP（Qian和Tang，2024）	63.66	–	–	60.99	–	66.26	–	–	71.55	–
	PEAT（Miao等，2023）	71.21	95.35	98.81	67.15	–	72.55	97.15	98.60	77.05	–
	本文	**78.91**	95.68	98.53	**75.32**	**57.37**	**87.69**	97.73	99.32	85.86	**80.35**

注：加粗字体表示各列最优结果。“–”表示该数据在对应文献中未计算。

80.35%，领先于对比方法，表明模型在处理复杂室内场景时表现极为鲁棒。考虑到在复杂场景下，可能存在一些特征较为相似的行人，导致模型在区分这些样本时出现一定的混淆，相对于部分方法的提升幅度并不如Rank-1准确率那般显著，但在全局性能上可比保持很强的稳定性。

在RegDB数据集中，本模型在V2I和I2V模式下表现出色，尤其在处理难样本和全局跨模态检索上具有显著优势。具体而言，在Rank-1和Rank-20上分别达到92.08%和99.05%，与领先的对比方法相比，依然具备高度竞争力。在I2V模式下，接近领先方法FDNM，而在mINP这一评价困难样本处理能力上，达到77.92%，超过了其他所有对比方法，表明本研究的模型在应对困难样本时具有更强的鲁棒性和更高的识别精度，能够在不同模态转换的情况下识别行人。

表3　在RegDB数据集上的方法对比

Table 3　Comparison of methods on the RegDB dataset

/%

方法		V2I					I2V				
		Rank-1	Rank-10	Rank-20	mAP	mINP	Rank-1	Rank-10	Rank-20	mAP	mINP
骨骼不相关	TSME(Liu等,2022)	87.35	97.10	98.90	76.94	–	86.41	96.39	98.20	75.70	–
	SFANet(Liu等,2023)	76.31	91.02	94.27	68.00	–	70.15	85.24	89.27	63.77	–
	PMT(Lu等,2023)	84.83	–	–	76.55	–	84.16	–	–	75.13	–
	SIDA(Gong等,2023)	81.73	–	96.55	75.07	–	79.71	–	95.47	72.60	–
	MFCS(Yang等,2024)	85.34	–	–	76.39	–	83.88	–	–	75.16	–
	MSALNet(Zhang等,2024a)	91.14	–	–	85.19	72.66	90.23	–	–	84.26	70.49
	DEEN(Zhang和Wang,2023)	91.10	97.80	98.90	85.10	–	89.50	96.80	98.40	83.40	–
	MUN(Yu等,2023)	95.19	98.93	–	87.15	–	91.86	97.99	–	85.01	–
	SGIEL(Feng等,2023)	70.23	86.11	–	52.54	–	67.65	84.73	–	52.30	–
	FDNM(Zhang等,2025)	**95.50**	**99.00**	**99.60**	**90.00**	–	**94.00**	**98.50**	**99.30**	**88.70**	–
骨骼相关	TGLFC-Net(Guo等,2024)	92.36	97.76	98.80	80.32	–	88.82	96.18	97.93	77.47	–
	PAGP(Qian和Tang,2024)	88.35	–	–	83.18	–	86.46	–	–	80.08	–
	PEAT(Miao等,2023)	92.14	98.16	99.22	87.88	–	91.36	97.57	98.88	86.70	–
	本文	92.08	98.04	99.05	89.59	**87.92**	90.56	96.44	99.05	87.59	**77.92**

注：加粗字体表示各列最优结果。“–”表示该数据在对应文献中未计算。

2.4　消融实验与分析

为进一步验证本研究方法的有效性，评估每个组件对模型的贡献，本研究在SYSU-MM01数据集的all search模式开展消融实验，通过逐步移除模型中的模块来观察其对性能的影响，结果如表4所示。

1）骨骼热图模块。与基线模型相比，加入骨骼热图(heatmap)后，Rank-1从74.70%提升至75.61%，mAP也从71.80%提升到72.95%。尽管提升幅度相对较小，但骨骼热图提供额外的结构信息，尤其在跨模态任务中，能有效帮助模型理解姿态变化。

2）SCLoss函数。当引入SCLoss来平滑GCN的

表4　各组件消融效果

Table 4　Ablation results of each component

/%

实验设置				SYSU-MM01	
base	heatmap	L_{sc}	SVIAM	Rank-1	mAP
√				74.70	71.80
√	√			75.61	72.95
√	√	√		77.03	73.68
√	√		√	78.05	74.30
√	√	√	√	**78.91**	**75.32**

注：加粗字体表示各列最优结果。“√”表示加入。

输出后，模型的Rank-1进一步提高到77.03%，mAP达到73.68%。SCLoss有效地帮助模型平滑邻接节点的特征，使其更加一致。

3）结构—视觉跨模态注意力机制（SVIAM）模块。引入SVIAM模块来融合骨骼热图和原始图像的特征，模型性能再次提升，Rank-1达到78.05%，mAP提升至74.30%。这一结果说明，通过将骨骼热图与图像特征进行融合，模型能够在跨模态行人重识别任务中更好地理解不同模态下的特征差异。

综合来看，通过结合骨骼热图、SCLoss和SVIAM模块，模型在跨模态数据集上的适应性和识别能力显著提升，证明各模块在提高模型整体性能中发挥着不可或缺的重要作用。

2.5 实验结果的可视化

为直观展现本研究方法的优势，本研究在SYSU-MM01数据集上对输入的查询图和查询结果图进行可视化。实验主要包括t-SNE分布（t-distributed stochastic neighbor embedding）可视化和检索结果的可视化。

1）t-SNE分布可视化。从SYSU-MM01数据集中随机抽取24个不同身份的图像，采用t-SNE可视化分析其在嵌入空间中的分布情况。实验针对初始特征、基线模型和完整模型这3种状态进行可视化分析，其中不同颜色表示不同身份的图像，不同形状标记则区分可见光（visible，VIS）和红外（infrared，IR）图像。如图3所示，与初始特征分布和基线特征分布相比，本研究提出模块能够增大不同身份的特征之间的距离，缩短相同身份的特征距离，从而缩小模态差异。

2）检索结果的可视化。本研究在SYSU-MM01数据集上随机选择4个案例，以Rank-10的排序结果进行可视化展示。对于每个检索案例，检索正确用带绿色方框的图像表示，反之，检索错误用带红色方框表示。如图4所示，图4（a）中识别正确数量与识别错误数量之比为7∶3，差于图4（b）中的3∶1；与此同时，相对于baseline，本研究方法能够更加有效地提高排序结果，使匹配正确率较高的图像排在前几位，证明了本研究方法的有效性。

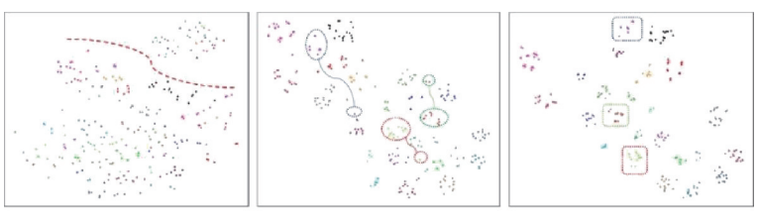


图3 SYSU-MM01数据集不同阶段的特征t-SNE图

Fig. 3 t-SNE plots of features at different stages on the SYSU-MM01 dataset

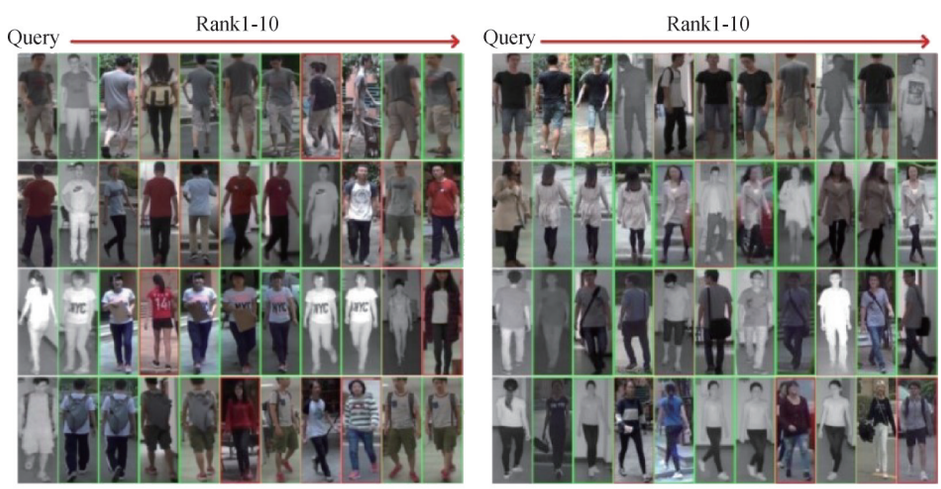


图4 在SYSU-MM01数据集上，通过基线与本文方法得到的检索结果

Fig. 4 Search results obtained from the baseline and the algorithm of this study on the SYSU-MM01 dataset ((a) baseline; (b) ours)

3 结 论

本研究围绕可见光—红外行人重识别展开深入探索，引入骨骼建模法应对模态差异挑战。提出融合结构与视觉特征的VI-ReID算法进行层次化特征提取。通过GCN学习骨骼结构特征，ResNet50提取图像特征，结合结构—视觉跨模态注意力机制（SVIAM）巧妙融合骨骼与视觉特征，塑造高分辨力联合特征表示，同时，结构内聚损失（SCLoss）函数的设计，增强了骨骼特征一致性，有力保障算法稳定准确，减少模态内差异。实验结果令人瞩目，在特征一致性和模型性能上表现优异，充分证明本文算法在提升跨模态行人重识别精度上的卓越成效。

在实际应用中，受限于当前数据集图像分辨率较低，导致精细视觉特征提取受阻、骨骼信息样本量不足等情况，模型的性能被制约。进一步探究发现，样本识别准确度高低与图像质量紧密相关，优质样本能驱动多模块协同作战，实现高效特征提取融合。

今后的研究可以朝着两个关键方向迈进：1）积极探寻高分辨力数据或借助前沿图像增强技术，可以通过自建数据集的方式丰富VI-ReID的样本数据库，为模型注入高质量“燃料”；2）引入深度图像等更多模态，构建更强大的多模态识别生态系统，全方位强化行人重识别效能，以从容应对现实世界的复杂多变，实现VI-ReID领域更多的突破和成就。

参考文献（References）

Bai Z Y, Ding Q C, Xu H L and Wu C D. 2023. Human similar action recognition by fusing saliency image semantic features. Journal of Image and Graphics, 28(9): 2872-2886（白忠玉，丁其川，徐红丽，吴成东. 2023. 融合显著性图像语义特征的人体相似动作识别. 中国图象图形学报，28(9)：2872-2886）[DOI: 10.11834/jig.220028]

Feng J W, Wu A C and Zheng W S. 2023. Shape-erased feature learning for visible-infrared person re-identification//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 22752-22761 [DOI: 10.1109/CVPR52729.2023.02179]

Feng S N. 2023. Research on Human Action Recognition Method Based on Skeletal Key Points. Beijing, China: North China University of Technology（冯赛楠. 2023. 基于骨骼关键点的人体动作识别方法研究. 北京：北方工业大学）[DOI: 10.26926/d.cnki.gbfgu.2023.000883]

Gong J H, Zhao S Y, Lam K M, Gao X and Shen J B. 2023. Spectrum-irrelevant fine-grained representation for visible-infrared person re-identification. Computer Vision and Image Understanding, 232: #103703 [DOI: 10.1016/j.cviu.2023.103703]

Guo J T, Ye Y F, Du H S and Hao X X. 2024. A triple-path global-local feature complementary network for visible-infrared person re-identification. Signal, Image and Video Processing, 18(1): 911-921 [DOI: 10.1007/s11760-023-02789-4]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE [DOI: 10.1109/CVPR.2016.90]

He Z J, Zhao H B, Wang J R and Feng W Q. 2023. Pose matters: pose guided graph attention network for person re-identification. Chinese Journal of Aeronautics, 36(5): 447-464 [DOI: 10.1016/j.cja.2022.11.017]

He Z M and Xu J Y. 2023. Research progress on person re-identification algorithms based on deep learning. Intelligent Manufacturing, (3): 80-83（何智敏，许佳云. 2023. 基于深度学习的行人重识别算法研究进展. 智能制造，(3)：80-83）[DOI: 10.3969/j.issn.1671-8186.2023.03.018]

Hu L, Zou X Z and Zhang P P. 2023. Learning progressive modality-shared transformers for effective visible-infrared person re-identification//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 1835-1843 [DOI: 10.1609/aaai.v37i2.25273]

Huang C H and Shen X B. 2024. Cross-modal person re-identification based on fused attention and feature enhancement. Journal of Nanjing University of Information Science and Technology, 16(4): 451-460（黄驰涵，沈肖波. 2024. 基于融合注意力和特征增强的跨模态行人重识别. 南京信息工程大学学报，16(4)：451-460）[DOI: 10.13878/j.cnki.jnuist.20240330001]

Huang G H, Lin G D, Wu E J, Zhao X D and Song L L. 2022. Design of fixed-point algorithm for softmax of DNN. China Integrated Circuit, 31(7): 60-64（黄光红，林广栋，吴尔杰，赵旭东，宋亮亮. 2022. 深度神经网络Softmax函数定点算法设计. 中国集成电路，31(7)：60-64）[DOI: 10.3969/j.issn.1681-5289.2022.07.012]

Khandelwal A, Chandra M G and Pramanik S. 2022. On classifying images using quantum image representation//2022 IEEE/ACM the 7th Symposium on Edge Computing (SEC). Seattle, USA: IEEE [DOI: 10.1109/SEC54971.2022.00067]

Kim Y. 2014. Convolutional neural networks for sentence classification//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics [DOI: 10.3115/v1/D14-1181]

Kipf T N and Welling M. 2016. Semi-supervised classification with graph convolutional networks [EB/OL]. [2024-10-01]. https://arxiv.org/pdf/1609.02907.pdf

Li R and Jiang M. 2023. Person re-identification based on pose estimation with feature similarity. Laser and Optoelectronics Progress, 60(6): #0610001 (李帆, 蒋敏. 2023. 基于姿态估计与特征相似度的行人重识别算法. 激光与光电子学进展, 60(6): #0610001) [DOI: 10.3788/lop212869]

Liang T F, Jin Y, Liu W and Li Y D. 2023. Cross-modality transformer with modality mining for visible-infrared person re-identification. IEEE Transactions on Multimedia, 25: 8432-8444 [DOI: 10.1109/TMM.2023.3237155]

Liang Z J and Liu D. 2021. Pose-based modular network for human-object interaction detection. Application Research of Computers, 38(8): 2299-2302 (梁志军, 刘栋. 2021. 基于姿态信息的人与物体交互检测模块网络. 计算机应用研究, 38(8): 2299-2302) [DOI: 10.19734/j.issn.1001-3695.2020.11.0429]

Liu H J, Ma S, Xia D X and Li S Z. 2023. SFANet: a spectrum-aware feature augmentation network for visible-infrared person reidentification. IEEE Transactions on Neural Networks and Learning Systems, 34(4): 1958-1971 [DOI: 10.1109/TNNLS.2021.3105702]

Liu J. 2021. Research on Smoking Action Recognition Based on Skeleton Information. Xuzhou, China: China University of Mining and Technology (刘婧. 2021. 基于骨骼信息的吸烟动作识别方法研究. 徐州: 中国矿业大学) [DOI: 10.27623/d.cnki.gzkyu.2021.000175]

Liu J N, Wang J L, Huang N C, Zhang Q and Han J G. 2022. Revisiting modality-specific feature compensation for visible-infrared person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 32(10): 7226-7240 [DOI: 10.1109/TCSVT.2022.3168999]

Lyu Y H, Wang G X, Zhao W Q, Zhao W and Guan Z Y. 2024. Edge-weight-embedding graph convolutional network for person reidentification. IEEE Intelligent Systems, 39(4): 74-82 [DOI: 10.1109/MIS.2024.3385381]

Lu H, Zou X Z and Zhang P P, 2023. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. Proceedings of the AAAI Conference on Artificial Intelligence, 37(2): 1835-1843 [DOI: 10.1609/aaai.v37i2.25273.]

Miao Y Q, Huang N C, Ma X and Han J. 2023. On exploring pose estimation as an auxiliary learning task for visible-infrared person re-identification. Neurocomputing, 556: #126652 [DOI: 10.1016/j.neucom.2023.126652]

Nayeem T A, Motaharuzzaman S M, Hoque A T and Rahman M H. 2022. Computer vision based object detection and recognition system for image searching//Proceedings of the 12th International Conference on Electrical and Computer Engineering (ICECE). Dhaka, Bangladesh: IEEE [DOI: 10.1109/ICECE57408.2022.10089019]

Nguyen D T, Hong H G, Kim K W and Park K R. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors, 17(3): #605 [DOI: 10.3390/s17030605]

Qian Y H and Tang S K. 2024. Pose attention-guided paired-images generation for visible-infrared person re-identification. IEEE Signal Processing Letters, 31: 346-350 [DOI: 10.1109/LSP.2024.3354190]

Radenović F, Tolias G and Chum O. 2019. Fine-tuning CNN image retrieval with no human annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(7): 1655-1668 [DOI: 10.1109/TPAMI.2018.2846566]

Rao H C, Leung C and Miao C Y. 2024. Hierarchical skeleton meta-prototype contrastive learning with hard skeleton mining for unsupervised person re-identification. International Journal of Computer Vision, 132(1): 238-260 [DOI: 10.1007/s11263-023-01864-0]

Wu A C, Zheng H X, Yu H X, Gong S G and Lai J H. 2017. RGB-infrared cross-modality person re-identification//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE [DOI: 10.1109/ICCV.2017.575]

Xia D X, Liu H J, Xu L L and Wang L N. 2021. Visible-infrared person re-identification with data augmentation via cycle-consistent adversarial network. Neurocomputing, 443: 35-46 [DOI: 10.1016/j.neucom.2021.02.088]

Yang H W, Yang P P and Guo H C. 2018. Gesture recognition based on skeleton information. Computer Applications and Software, 35(12): 228-232, 292 (杨和稳, 杨萍萍, 郭海晨. 2018. 基于骨骼信息下的手势识别研究. 计算机应用与软件, 35(12): 228-232), 292) [DOI: 10.3969/j.issn.1000-386x.2018.12.042]

Yang L. 2023. Overview of pedestrian re-identification based on deep learning. China Water Transport, 23(7): 57-59 (杨磊. 2023. 基于深度学习的行人重识别综述. 中国水运, 23(7): 57-59) [DOI: 10.3969/j.issn.1006-7973.2023.07.0057]

Yang X, Dong W J, Li M J, Wei Z Y, Wang N N and Gao X B. 2024. Cooperative separation of modality shared-specific features for visible-infrared person re-identification. IEEE Transactions on Multimedia, 26: 8172-8183 [DOI: 10.1109/TMM.2024.3377139]

Ye M, Shen J B, Lin G J, Xiang T and Hoi S C H. 2021. Deep learning for person re-identification: a survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6): 2872-2893 [DOI: 10.1109/TPAMI.2021.3054775]

Yu H, Cheng X, Peng W, Liu W H and Zhao G Y. 2023. Modality unifying network for visible-infrared person re-identification//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 11151-11161 [DOI: 10.1109/ICCV51070.2023.01027]

Zhang B X, Zhang S M and Zhong Z Y. 2022. Person re-identification based on multi-granularity feature fusion network. Journal of Optoelectronics·Laser, 33(9): 977-983 (张勃兴, 张寿明, 钟震宇. 2022. 基于多粒度特征融合网络的行人重识别. 光电子·激光, 33(9): 977-983) [DOI: 10.16136/j.joel.2022.09.0886]

Zhang H D, Cheng S L and Du A Y. 2024a. Multi-stage auxiliary learn-

ing for visible-infrared person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 34(11): 12032-12047[DOI: 10.1109/TCSVT.2024.3425536]

Zhang Y K, Wang H Z, Lu Y, Yan Y and Li X L. 2025. Frequency domain nuances mining for visible-infrared person re-identification. IEEE Transactions on Information Forensics and Security. 20: 5411-5424[DOI:10.1109/TIFS.2025.3569176]

Zhang Y K and Wang H Z. 2023. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 2153-2162[DOI: 10.1109/CVPR52729.2023.00214]

Zhu M, Ming Z Q, Yan J R, Yang Y and Zhu J M. 2022. A survey on generative adversarial network based person re-identification method. Journal of Computer-Aided Design and Computer Graphics, 34(2): 163-179 (朱敏, 明章强, 闫建荣, 杨勇, 朱佳旻. 2022. 基于生成对抗网络的行人重识别方法研究综述. 计算机辅助设计与图形学学报, 34(2): 163-179)[DOI: 10.3724/SP.J.1089.2022.18852]

作者简介

魏思，女，硕士研究生，主要研究方向为信息与通信工程。
E-mail: 202230310007@stu.shmtu.edu.cn
杨文璐，男，副教授，主要研究方向为图像和视频处理。
E-mail: wlyang@shmtu.edu.cn