E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2023)05-1422-12

论文引用格式:Ma X F and Cheng W G. 2023. Dual grained feature fusion network-relevant cross-modality pedestrian re-identification. Journal of Image and Graphics, 28(05):1422-1433(马潇峰, 程文刚. 2023. 双粒度特征融合网络的跨模态行人再识别. 中国图象图形学报, 28(05):1422-1433)[DOI:10.11834/jig.220387]

双粒度特征融合网络的跨模态行人再识别

马潇峰[1],程文刚[1,2*]

1. 华北电力大学控制与计算机工程学院,北京　102206; 2. 复杂能源系统智能计算教育部工程研究中心,保定　071003

摘　要:**目的**　可见光—红外跨模态行人再识别旨在匹配具有相同行人身份的可见光图像和红外图像。现有方法主要采用模态共享特征学习或模态转换来缩小模态间的差异,前者通常只关注全局或局部特征表示,后者则存在生成模态不可靠的问题。事实上,轮廓具有一定的跨模态不变性,同时也是一种相对可靠的行人识别线索。为了有效利用轮廓信息减少模态间差异,本文将轮廓作为辅助模态,提出了一种轮廓引导的双粒度特征融合网络,用于跨模态行人再识别。**方法**　在全局粒度上,通过行人图像到轮廓图像的融合,用于增强轮廓的全局特征表达,得到轮廓增广特征。在局部粒度上,通过轮廓增广特征和基于部件的局部特征的融合,用于联合全局特征和局部特征,得到融合后的图像表达。**结果**　在可见光—红外跨模态行人再识别的两个公开数据集对模型进行评估,结果优于一些代表性方法。在SYSU-MM01(Sun Yat-sen University multiple modality 01)数据集上,本文方法rank-1准确率和平均精度均值(mean average precision,mAP)分别为62.42%和58.14%。在RegDB(Dongguk body-based person recognition database)数据集上,本文方法rank-1和mAP分别为84.42%和77.82%。**结论**　本文将轮廓信息引入跨模态行人再识别,提出一种轮廓引导的双粒度特征融合网络,在全局粒度和局部粒度上进行特征融合,从而学习具有判别性的特征,性能超过了近年来一些具有代表性的方法,验证了轮廓线索及其使用方法的有效性。

关键词:行人再识别;跨模态;特征融合;轮廓;特征学习

Dual grained feature fusion network-relevant cross-modality pedestrian re-identification

Ma Xiaofeng[1], Cheng Wengang[1,2*]

1. School of Control and Computer Engineering, North China Electric Power University, Beijing　102206, China;

2. Engineering Research Center of Intelligent Computing for Complex Energy Systems, Ministry of Education, Baoding　071003, China

Abstract: **Objective** Visible-infrared cross-modality pedestrianre-identification (VI-ReID) is focused on same identity-related images between the visible and infrared modality. As a popular technique of intelligent surveillance, it is still challenged for cross-modality. To optimize intra-class variations in RGB image-based pedestrian re-identification (RGB-ReID) task, one crucial challenge in VI-ReID is to bridge the modality gap between the RGB and infrared(IR) images of the same identity because current methods are mainly derived from modality-incorporated feature learning or modality transformation approaches. Specifically, modality-incorporated feature learning methods are used to map the inputs of RGB and IR images into a common embedding space for cross-modality feature alignment. A two-stream convolutional neural network (two-stream CNN) architecture has been recognized and some discriminative constraints are developed as well. However, each filter is constrained for a small region, convolutions is challenged to interlinked to spatial-ranged concepts. Quantitative

收稿日期:2022-04-24;**修回日期**:2022-06-28;**预印本日期**:2022-07-05

*通信作者:程文刚　wgcheng@ncepu.edu.cn

tests illustrates that the CNNs are strongly biased toward textures rather than shapes. Moreover, existing methods in VI-ReID focus on the global or local feature representation only. Another path of modality transformation approaches can be used to generate cross-modality images-relevant pedestrain images or transform them into an intermediate modality. Generative adversarial network (GAN) and encoder-decoder structures are commonly used for these methods. However, due to the distorted IR-to-RGB translation and additional noises, image generation is incredible and GAN models are difficult to be converged. The RGB images consist of three-color channels, while IR images only contain a single channel reflecting the thermal radiation emitted from the human body and its contexts. Compared to the missing colors and textures in IR images, we review VI-ReID problem and the contour is realized in terms of a relative effective feature. Furthermore, contour is a modality-shared path as it keeps consistent for both of IR and RGB images, which is more accurate and reliable than generated intermediate modality. We implement VI-ReID-integrated contour strategy. To develop its optimization, we take the contour as an auxiliary modality to narrow the modality gap. Meanwhile, we tend to introduce local features into our model to collaborate with global ones further beyond part-based features. **Method** A contour-guided dual-grained feature fusion network (CGDGFN) is developed for VI-ReID. It consists of two types of fusion. The first type is concerned of the image to contour fusion, which is called global-grained fusion (G-Fusion) at image-level. G-Fusion can output the augmented contour features. The other type can realize fusion between augmented contour features and local features, which is oriented at image and part mixed level. As the local feature is involved in, it called as local-grained fusion (L-Fusion) for simplicity. The proposed CGDGFN consists of four branches, which are 1) RGB images, 2) IR images, 3) RGB contour and 4) IR contour. First, the input of the network is a pair of RGB and IR images, while they are fed into RGB branch and IR branch, and a contour detector generates their contour images. Then, RGB and IR-relevant contour images of two modalities are fed into RGB-contour branch and IR-contour branch. The ResNet50 is used as the backbone architecture for each branch. The first convolutional layer in each branch has independent parameters to capture modality-specific information, while the remaining blocks are used to learn weight-shared modality-invariant features. In addition, RGB branch and IR branch average pooling layer structure is optimzed for part-based features extraction. G-Fusion is used to fuse an image to its corresponding contour image. After G-Fusion, the augmented contour features will be produced by the global average pooling layer of RGB-contour branch and IR-contour branch. Meanwhile, RGB branch and IR branch can output corresponding local features. RGB local feature as well as IR local feature is an array of feature vectors and its length is clarified by the partition setting. Two of local feature extraction methods are involved in: 1) uniform partition and 2) soft partition. L-Fusion is responsible for fusing the augmented contour features and corresponding local ones. The implementation of our method is based on the Pytorch framework. We adopt the ResNet50 pre-trained on ImageNet as the backbone network, and the stride of the last convolutional layer is set to 1 to obtain feature maps with higher spatial size. The batch size is set to 64. For each batch, we select 4 identities in random and each identity includes 8 visible images and 8 infrared images. The input images are resized to 288×144 pixels, random cropping and random horizontal flip are used for data augmentation. The stochastic gradient descent (SGD) optimizer is used for optimization and the momentum is set to 0.9. We train the model for 60 epochs at first. The initial learning rate is set as 0.01 and warmup strategy is applied to enhance performance. To realize its soft partition, we fine-tune our model for additional 20 epochs as well. **Result** The proposed CGDGFN method is compared with state-of-the-art (SOTA) VI-ReID approaches on two databases Sun Yat-sen University multiple modality 01(SYSU-MM01) and Dongguk body-based person recognition database(RegDB), which is composed of the methods of global-feature, local-feature, and image generation. The standard cumulated matching characteristics (CMC) and mean average precision (mAP) are employed to evaluate the performance. Our proposed method has obtained 62.42% rank-1 identification rate and 58.14% mAP score on SYSU-MM01, and the values of rank-1 and mAP on RegDB are reached to 84.42% and 77.82% each in comparison with popular SOTA approaches on both datasets. **Conclusion** We introduce the contour clue to VI-ReID. To leverage contour information, we took the contour as an auxiliary modality, and a contour-guided dual-grained feature fusion network (CGDGFN) is developed for VI-ReID. Global-grained fusion (G-Fusion) can enhance the original contour representation and produce augmented contour features. Local-grained fusion (L-Fusion) can fuse the part-based local features and the augmented contour features to output its powerful image representation further.

Key words: pedestrian re-identification; cross-modality; feature fusion; contour; feature learning

0 引 言

跨模态行人再识别是指给定一种模态的行人图像作为查询，从另一种模态的候选集中检索具有相同身份图像的技术。本文针对可见光(RGB)模态和红外(infrared，IR)模态进行研究。跨模态行人再识别广泛应用于智能监控、安防和刑侦等领域，但由于存在较大的跨模态差异，准确匹配行人图像仍然很具有挑战性。因此，跨模态行人再识别受到了工业界和学术界的共同关注。

除了在单模态行人再识别中已经存在的模态内变化外，跨模态行人再识别的一个关键问题在于如何缩小相同身份的可见光图像和红外图像之间的模态差异。现有的工作主要采用模态共享特征学习或模态转换的方法。模态共享特征学习方法致力于将可见光和红外图像投影到特定的公共嵌入空间，以实现跨模态特征对齐，可细分为全局特征学习(Wu等，2017；Ye等，2020)和局部特征学习(Hao等，2019b；Zhu等，2020)。全局特征学习用一个特征向量表示行人图像整体，而局部特征学习用基于部件或区域的特征向量集合表示该行人图像。双路卷积神经网络(two-stream convolutional neural network，two-stream CNN)结构常应用于这类方法，并配合损失函数(如身份损失、三元组损失等)进行约束(Ye等，2022)。然而，现有的模态共享特征学习方法通常致力于发掘全局或局部特征表示，很少结合两种特征的优势。基于模态转换的方法旨在生成行人图像对应的跨模态图像(Wang等，2019a，b，2020)或中间模态图像(Li等，2020；Zhang等，2021)，将异构模态图像转换到统一的模态中，从而减小模态间差异。这类方法通常采用生成对抗网络(generative adversarial network，GAN)和编码器—解码器(encoder-decoder)结构。然而，红外图像到可见光图像的转换是不适定的，还可能引入附加噪声，无法生成准确、真实的可见光图像，并且基于GAN的模型存在难以收敛的问题。生成的中间模态试图在特征分布上拉近异构图像的距离，但两种模态仍存在着较大差异(Wei等，2021)。

不同的成像机制决定了可见光和红外两种图像本质上的差异。可见光图像由红、绿、蓝3个颜色通道构成，而红外图像只包含反映物体热辐射的单通道，这导致颜色这一关键特征无法应用于跨模态匹配。而轮廓是一种相对可靠的识别线索，事实上，人类通过视觉检验红外监控进行判断时，主要依靠的就是轮廓信息。红外图像丢失了颜色和纹理等特征，但轮廓、形状等信息则仍然明确，如图1伪彩色红外图像所示。由图1可见，轮廓在可见光和红外图像间具有一定的跨模态不变性。

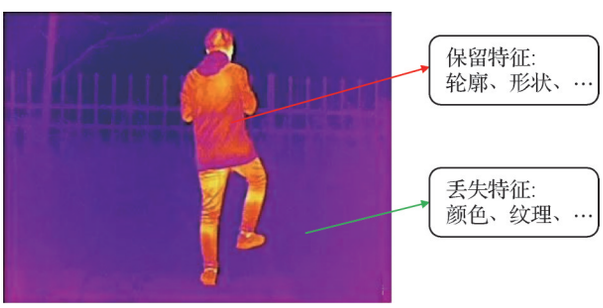

图1 伪彩色红外图像的示例

Fig. 1 An example of pseudo-color IR images

基于CNN的方法在行人再识别问题中取得了巨大成功，这归因于其具有强大的深层判别特征表达能力。然而，由于每个卷积核都限制在局部区域(感受野)上，使其在特征学习过程中并没有充分利用全局上下文信息(Wu等，2021)。因此，计算机视觉研究引入了Non-local(Wang等，2018)机制以建模长距离关系，如各种视觉Transformer(Han等，2022)通过自注意力机制捕捉全局信息。同时，Geirhos等人(2022)的研究表明，CNN更倾向于提取纹理信息而非形状信息。因此，轮廓这种图像级全局特征的引入亦有助于弥补现有CNN方法的上述缺陷。

基于以上考虑，本文提出将轮廓信息引入到跨模态行人再识别研究中。然而，轮廓也存在变形和遮挡等问题，如何恰当利用轮廓线索也非常具有挑战性的。为此，本文将轮廓作为一种辅助模态，希望借助深度网络强大的特征表达能力来缩小可见光和红外的模态间差异。轮廓是行人的一种整体性而非局部性的特征描述，因此对全局特征进行了轮廓增广。同时，受到局部特征具有良好判别能力的启发，期冀将轮廓与模态共享特征学习得到的局部特征进一步融合，增强特征表达能力。相应地，提出了一种轮廓引导下的双粒度特征融合网络，如图2所示。该网络包括两种类型的融合，一种是图像到轮廓的融合，在图像级进行，称为全局粒度融合，输出轮廓

增广特征；另一种是在轮廓增广特征和局部特征之间进行融合，由于涉及局部特征，称为局部粒度融合。

本文的主要贡献如下：1）将轮廓作为一种辅助模态引入到跨模态行人再识别模型中进行特征嵌入。这是在跨模态行人再识别问题中利用显式轮廓信息的首次尝试。2）提出了一种轮廓引导的双粒度特征融合网络，在统一的端到端网络中同时学习全局粒度和局部粒度特征。在两个公开数据集SYSU-MM01（Sun Yat-sen University multiple modality 01）和RegDB（Dongguk body-based person recognition database）上的实验结果验证了模型的有效性。

1　相关工作

1.1　跨模态行人再识别

跨模态行人再识别不仅要面对遮挡、不同视角和行人姿势造成的模态内差异（史维东 等，2020），还要解决由于异构图像而形成的跨模态差异。其中，减小跨模态差异至关重要，因为模态间差异也会加剧已经存在的模态内差异。现有方法主要可以分为模态共享特征学习和模态转换两类。

模态共享特征学习旨在从异构模态中学习具有判别力和鲁棒性的特征。Wu等人（2017）设计了一种深度零填充（zero-padding）结构，使单路网络的节点自动提取两种模态的特征。Ye等人（2018a）提出了一个结合特征学习和度量学习的两阶段框架，并通过后续工作逐步完善该框架，使双路卷积神经网络成为目前跨模态行人再识别领域一个常用的基线模型（Ye等，2022），其通常包括特定于模态的浅层结构和模态共享的深层结构，最终将行人图像映射到共享特征空间进行相似度学习。双路网络主要采用身份损失（identity loss）和三元组损失（triplet loss）进行约束。一些工作从优化损失函数的角度出发增强其学习能力。Zhu等人（2020）提出异质中心损失（hetero-center loss），以缩小不同模态下同一行人图像的中心距离。Liu等人（2021）设计了异质中心三元组损失（hetero-center triplet loss），试图结合三元组损失和异质中心损失的优点。尽管基于CNN的方法取得了巨大成功，但对长距离依赖关系建模能力有限，使网络偏向于识别纹理而非形状（Geirhos等，2022）。轮廓是一种图像级的特征，引入轮廓可以引导CNN学习基于形状的行人判别特征，并弥补其在长距离关系建模上的不足。

模态转换通常采用基于GAN的方法和编码器—解码器结构。Wang等人（2020）提出的JSIA-ReID（joint set-level and instance-level alignment Re-ID）执行集合级和实例级的对齐，以生成跨模态成对图像。Li等人（2020）通过一个轻量级网络引入了辅助X模态图像，并联合优化三种模态的特征。Zhang等人（2021）提出一种非线性中间模态生成器，采用编码器—解码器结构生成M模态图像，使模态间特征分布尽可能接近。由于红外模态到可见光模态的转换是不适定的，生成的图像可能包含额外的噪声。而轮廓在红外和可见光图像中保持不变，是一种良好的模态共享特征。从这点上看，固有的轮廓比生成的图像更加可靠，然而现有的跨模态行人再识别方法没有关注到轮廓信息。Chen等人（2019）在可见光单模态行人再识别中考虑了行人轮廓的影响，本文则深入探究轮廓在跨模态行人再识别的价值，并提出了一种双粒度特征融合策略以实现更有效的特征学习。

1.2　全局特征和局部特征

全局特征学习为每幅行人图像提取全局特征表示，跨模态行人再识别中的大多数方法都采用全局特征来描述行人。Ye等人（2022）设计了一个简单但广泛使用的基线模型，使用双路网络提取全局特征，由身份损失和三元组损失联合优化整个网络。因其易于实现且泛化能力强，大部分特征学习相关方法（Wu等，2017；Ye等，2018b，2020；Dai等，2018）和基于模态转换的方法（Wang等，2019a，b，2020；Li等，2020）都倾向于使用全局特征。局部特征学习能够获得部件或区域的特征，对行人图像错位具有鲁棒性。一些方法（Zhu等，2020；Hao等，2019b）侧重于利用局部细粒度特征，将可见光和红外图像分成几个水平部件，每个部件独立预测行人身份。但目前的跨模态行人再识别模型通常只关注全局或局部特征学习方法，本文则在轮廓信息引导下，融合全局特征和局部特征，使其具有更强的判别能力。

2　轮廓引导的双粒度特征融合网络

2.1　网络架构

在双路网络基础上，本文设计了两个特定的分

支用于学习可见光图像和红外图像所对应轮廓的特征，将轮廓图像作为辅助模态联合优化整个网络，从而缩小模态间差异。

提出的轮廓引导下的双粒度特征融合网络架构如图2所示，由4个分支组成，分别对应于可见光轮廓图像、可见光图像、红外图像和红外轮廓图像。为了便于叙述，从上到下将其依次命名为分支1、分支2、分支3和分支4。选取ResNet50（50-layer residual network）作为每个分支的主干网络。各分支的第1个卷积层使用独立的参数来捕获特定于模态的信息，而剩余的残差块则共享权重以学习模态不变特征，即分支2和分支3，分支1和分支4共享各自残差块Stage1—Stage4的参数。此外，将分支2与分支3中的最后一个全局平均池化（global average pooling，GAP）层替换为用于局部特征提取的结构。

网络的输入是一组可见光和红外图像，可见光图像送入分支2，红外图像送入分支3。根据给定的图像，轮廓检测器相应地生成其轮廓图像。然后，将可见光轮廓图像和红外轮廓图像（如图2所示）这两种模态的轮廓图像分别送入分支1和分支4。通过这种方式，轮廓图像作为辅助模态信息进入网络。

全局粒度融合是指行人图像到轮廓的融合，包括可见光—轮廓融合以及红外—轮廓融合。经过全局粒度融合后，由分支1和分支4的全局平均池化层分别生成可见光轮廓增广特征和红外轮廓增广特征。同时，分支2和分支3输出可见光局部特征和红外局部特征，局部特征是一组特征向量，具体数量由区域划分相关参数决定。局部粒度融合负责连接轮廓增广特征和相应的局部特征。例如，通过局部粒度融合将红外轮廓增广特征和红外局部特征拼接在一起，以获得红外图像的表示。

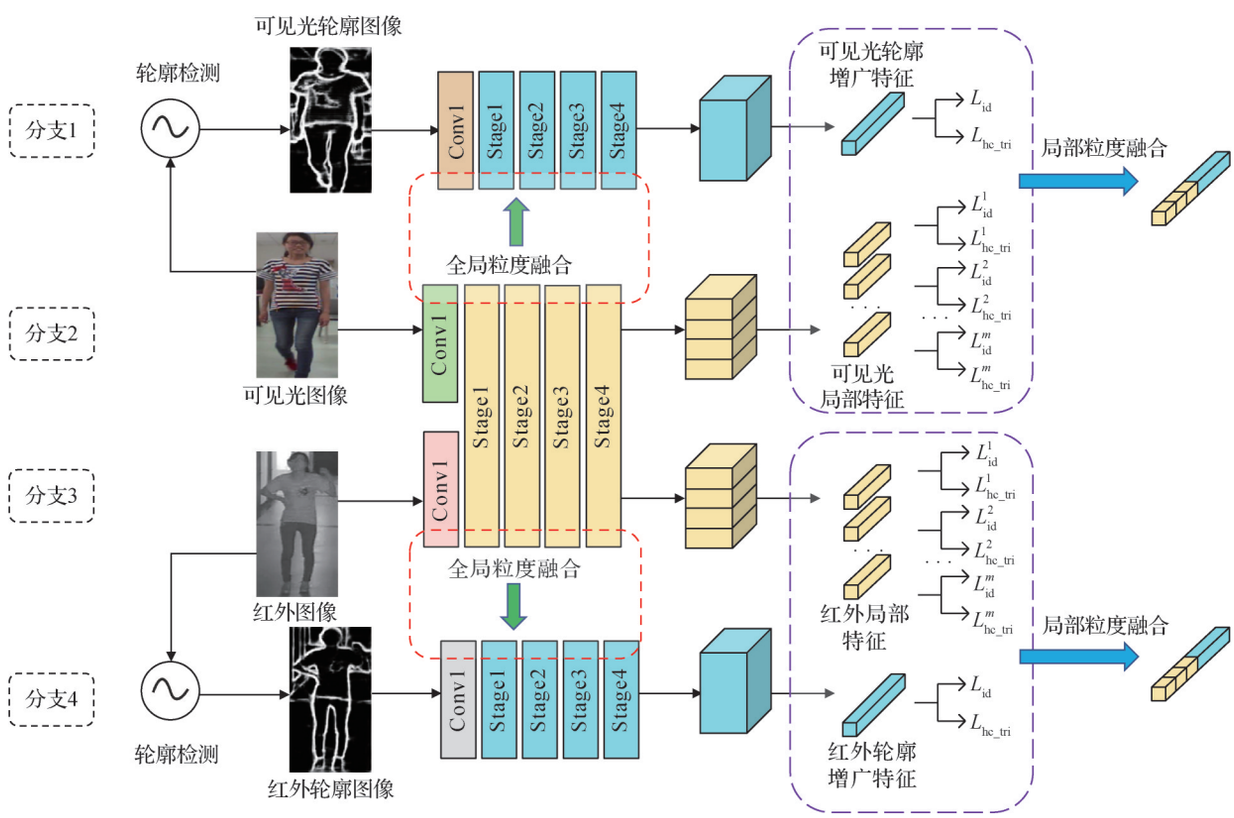

图2　轮廓引导的双粒度特征融合网络结构

Fig. 2　The structure of contour-guided dual-granularity feature fusion network

2.2　双粒度特征融合

2.2.1　全局粒度特征融合

全局粒度融合是指将行人图像特征融合到其对应的轮廓图像中，借助轮廓作为先验知识，增强轮廓的全局特征表达。以红外图像为例，其红外—轮廓融合过程如图3所示。特征融合可以在不同的层次上分别开展，如图3中的箭头所示，浅层网络融合低层细节相关特征，而深层网络则融合高层语义相关

特征。实验检验了各个不同融合位置的作用。

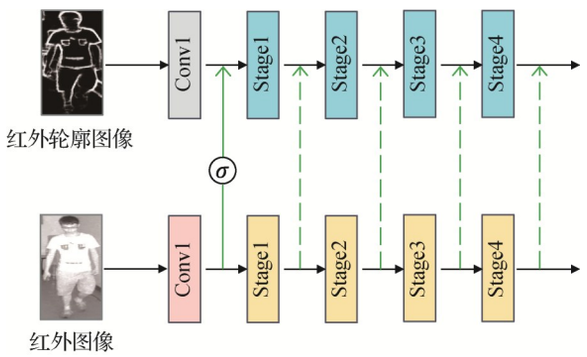

图 3　全局粒度融合的示意图

Fig. 3　Illustration of global-granularity fusion

采用 RCF（richer convolutional features）（Liu 等，2017）作为轮廓检测器，其主干架构是一个经过预训练的 VGG16（Visual Geometry Group 16-layer network）网络。轮廓特征提取的表达式为

$$\boldsymbol{x}_C^k = \phi(\boldsymbol{x}^k) \tag{1}$$

式中，$\boldsymbol{x}^k$ 和 $\boldsymbol{x}_C^k$ 表示原始图像和生成的轮廓图像，$k \in \{\boldsymbol{V}, \boldsymbol{I}\}$ 分别代表可见光模态或红外模态，$\phi(\cdot)$ 表示轮廓检测器。

此外，本文探讨了不同融合操作对实验的影响，包括按元素乘、按元素加和拼接。按元素乘旨在通过轮廓图像特征对行人图像特征进行筛选过滤，保留行人轮廓信息而忽略其他区域的信息；按元素加则着重为轮廓图像特征补充行人图像相关的语义信息；拼接是在特征维度上扩展，而不损失行人图像和轮廓图像各自的信息。本文模型在 Conv1 后采用按元素加的方式对特征进行融合。全局粒度融合的表达式为

$$\boldsymbol{F}_i^{\mathrm{AugVC}} = \sigma(\boldsymbol{F}_i^{\mathrm{V}}, \boldsymbol{F}_i^{\mathrm{VC}}) \tag{2}$$

$$\boldsymbol{F}_i^{\mathrm{AugIC}} = \sigma(\boldsymbol{F}_i^{\mathrm{I}}, \boldsymbol{F}_i^{\mathrm{IC}}) \tag{3}$$

式中，$\sigma(\boldsymbol{x}, \boldsymbol{y})$ 指特征融合操作，$\sigma \in \{\odot, \oplus, ⓒ\}$ 分别代表按元素乘、按元素加和拼接；$\boldsymbol{F}_i^{\mathrm{V}}$，$\boldsymbol{F}_i^{\mathrm{I}}$，$\boldsymbol{F}_i^{\mathrm{VC}}$，$\boldsymbol{F}_i^{\mathrm{IC}}$ 分别表示经过网络的第 i 个残差块后，可见光图像、红外图像、可见光轮廓和红外轮廓各自对应的特征图；$\boldsymbol{F}_i^{\mathrm{AugVC}}$ 和 $\boldsymbol{F}_i^{\mathrm{AugIC}}$ 分别表示可见光轮廓增广特征图和红外轮廓增广特征图。

2.2.2　局部粒度特征细化与融合

局部粒度融合是指将轮廓增广特征与基于部件的局部特征进行融合，从而联合全局特征和局部特征，得到具备更强判别能力的图像表达。由于局部特征通常与特定的身体部位有关，在不同的模态之间相对稳定，从而有助于异构模态下的对齐。

现有工作在提取局部特征时通常采用均匀分割法，首先将经过主干网络的特征图平均划分为几个水平部件，每个部件的特征图经过全局平均池化层生成特征向量，随后送入各自的分类器独立地预测行人身份。为了提高识别准确率，进一步采用了软分割方法（Sun 等，2018）细化局部粒度特征。具体而言，首先由区域分类器对原始特征图的各个列向量进行 m 分类，并得到区域划分掩膜，每个区域划分掩膜表示列向量属于该部件区域的概率。区域分类器由全连接层和 softmax 函数构成。最后，将 m 个区域划分掩膜分别与原始特征图相乘，通过平均池化操作得到 m 个特征向量。软分割法可以表达为

$$\boldsymbol{p}_j^k = \omega(\boldsymbol{u}) = \mathrm{softmax}(\boldsymbol{W}_j \boldsymbol{u}) \tag{4}$$

$$\boldsymbol{f}_j^k = g(\boldsymbol{F}^k \times \boldsymbol{p}_j^k) \tag{5}$$

式中，$\omega(\cdot)$ 指区域分类器，$g(\cdot)$ 指全局平均池化操作，$\mathrm{softmax}(\cdot)$ 指 softmax 激活函数，$\boldsymbol{W}_j$ 为全连接层的权重矩阵，$\boldsymbol{F}^k$、$\boldsymbol{u}$ 表示行人图像经过主干网络输出的特征图和其中的每个列向量；$\boldsymbol{p}_j^k$ 和 $\boldsymbol{f}_j^k$ 分别表示图像第 j 个区域的划分掩膜和特征向量，其中 $j \in \{1, \cdots, m\}$。

获得局部特征后，将轮廓增广特征向量和局部特征向量拼接，完成局部粒度融合。以可见光图像为例，针对均匀分割和软分割这两种局部特征提取方法，局部特征融合过程如图 4 所示，该图省略了全局特征融合的表示。局部粒度融合的表达为

$$\boldsymbol{f}^{\mathrm{V}} = Concat(\boldsymbol{f}^{\mathrm{AugVC}}, \boldsymbol{f}_1^{\mathrm{V}}, \cdots, \boldsymbol{f}_m^{\mathrm{V}}) \tag{6}$$

$$\boldsymbol{f}^{\mathrm{I}} = Concat(\boldsymbol{f}^{\mathrm{AugIC}}, \boldsymbol{f}_1^{\mathrm{I}}, \cdots, \boldsymbol{f}_{\mathrm{m}}^{\mathrm{I}}) \tag{7}$$

式中，$\boldsymbol{f}^{\mathrm{AugVC}}$ 和 $\boldsymbol{f}^{\mathrm{AugIC}}$ 分别表示经过全局平均池化层得到的可见光轮廓增广特征向量和红外轮廓增广特征向量，$\boldsymbol{f}^{\mathrm{V}}$ 和 $\boldsymbol{f}^{\mathrm{I}}$ 分别表示可见光行人图像和红外行人图像最终的特征表示，$Concat(\cdot)$ 代表向量拼接操作。

2.3　损失函数

为了优化提出的模型，采用身份损失和三元组损失。身份损失将训练过程视为一个分类问题，使每幅行人图像尽可能分类到正确的身份类别中，从而学习具有判别性的特征。三元组损失将训练视为一个检索排序问题（赵才荣 等，2021），在特征空间拉近相同行人身份的图像特征，推远不同行人身份的图像特征。身份损失一般由交叉熵损失函数实

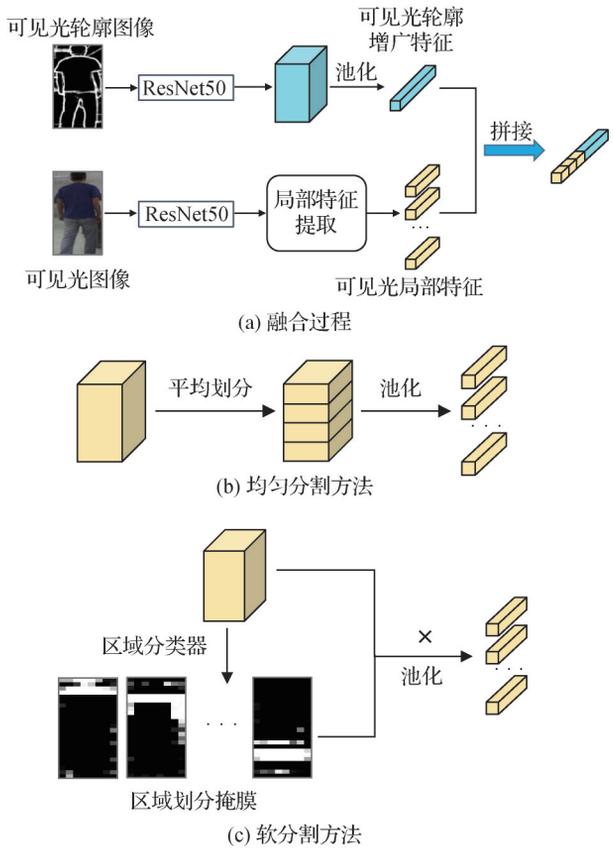

图4　局部粒度融合的示意图

Fig. 4　Illustration of local-granularity fusion ((a) fusion process; (b) uniform partition method; (c) soft partition method)

现，本文使用Liu等人(2021)提出的异质中心三元组损失替代传统三元组损失。异质中心三元组损失结合了传统三元组损失和异质中心损失(Zhu等，2020)的优点，同时考虑了类内的紧凑性和类间的可分离性。总体的损失函数为

$$L = L_{\mathrm{id}} + L_{\mathrm{hc\_tri}} + \lambda \sum_{j=1}^{m} \left(L_{\mathrm{id}}^{j} + L_{\mathrm{hc\_tri}}^{j} \right) \tag{8}$$

式中，L_{id}和$L_{\mathrm{hc_tri}}$分别表示全局特征向量对应的身份损失和异质中心三元组损失，L_{id}^{j}和$L_{\mathrm{hc_tri}}^{j}$分别表示第j个局部特征向量对应的身份损失和异质中心三元组损失。本文实验将权衡参数λ的值设置为1.0，异质中心三元组损失的边距值设置为0.3。

3　实　验

3.1　数据集和评估标准

在可见光—红外跨模态行人再识别的两个公开数据集SYSU-MM01(Wu等，2017)和RegDB(Nguyen等，2017)上对提出的方法进行实验评估。通过与基线模型和一些近年来的SOTA(state-of-the-art)方法进行性能比较，验证模型的有效性。

SYSU-MM01数据集由4个可见光摄像头和2个红外摄像头拍摄，包含491个行人的287 628幅可见光图像和15 792幅红外图像。其中，训练集有395个行人，测试集有96个行人。数据集有室内搜索(indoor-search)和全搜索(all-search)两种评估模式，前者不包括室外摄像头拍摄的图像，后者使用全部摄像头拍摄的图像。本文采用最具挑战性的单次全搜索(single-shot all-search)模式评估提出的方法。

RegDB数据集由可见光—红外双成像系统拍摄，包含412个行人的8 240幅图像，每个行人都有10幅不同的可见光图像和10幅不同的红外图像。其中，训练集和测试集各有206个行人。沿用Ye等人(2018a)提出的策略，本文通过10次实验的结果评估模型，以获得稳定的结果。

实验参照现有的跨模态行人再识别中的评估标准，采用累积匹配特征(cumulative matching characteristics，CMC)和平均精度均值(mean average precision，mAP)两项指标来评估方法的性能。其中，CMC-k(rank-k匹配准确率)表示在排名前k的检索结果中出现正确匹配的概率，而mAP则度量具有多个正确匹配时的平均检索性能。

3.2　实验设置

使用深度学习框架Pytorch来实现本文方法，硬件配置如下：GPU为NVIDIA RTX 3090 24 GB，CPU为Intel(R) Core(TM) i7-11700 @ 2.50 GHz，内存32 GB。

实验采用在ImageNet上预先训练的ResNet50作为主干网络，且最后一个卷积层的stride设置为1，以获得更大空间尺寸的特征图。参照Zhu等人(2020)的实验设置，训练的batch size设置为64，每个batch随机选取4个行人，每个行人包括8幅可见光图像和8幅红外图像。输入图像的大小统一调整为288×144像素，并采用随机裁剪和随机水平翻转进行数据增强。局部特征的分割区域数量设置为6。

实验使用随机梯度下降(stochastic gradient descent，SGD)优化器，其中动量设置为0.9。初始学习率设置为0.01，并采用warm up策略调整学习率。具体来说，在前10个epoch，学习率可以通过0.01×

(epoch+1)来计算;在第10～20个epoch之间时,学习率保持为0.01不变;在第20个epoch和第50个epoch时,学习率分别衰减为0.001和0.000 1。经过60个epoch后停止训练。此外,当采用软分割方法时,还需对模型进行另外20个epoch的微调。在这个过程中,首先固定其他组件,单独训练区域分类器,然后联合优化整个网络。

3.3　对比实验

为了验证方法的有效性,在SYSU-MM01和RegDB两个数据集上与经典和SOTA方法进行对比实验。包括基于全局特征的方法Zero-Padding(Wu等,2017)、TONE(two-stream CNN network)+HCML(hierarchical cross-modality metric learning)(Ye等,2018a)、HSME(hypersphere manifold embedding)(Hao等,2019a)、cmGAN(cross-modality generative adversarial network)(Dai等,2018)、BDTR(bi-directional dual-constrained top-ranking)(Ye等,2018b)、AGW(attention generalized mean pooling with weighted triplet loss)(Ye等,2022)、MACE(modality-aware collaborative ensemble)(Ye等,2020)、Hi-CMD(hierarchical cross-modality disentanglement)(Choi等,2020)、NFS(neural feature search)(Chen等,2021)、MSO(multi-feature space joint optimization)(Gao等,2021)、基于局部特征的方法DFE(dual-alignment feature embedding)(Hao等,2019b)、TSLFN(two-stream local feature network)(Zhu等,2020)、LBA(learning by aligning)(Park等,2021),以及基于模态转换的方法D^2RL(dual-level discrepancy reduction learning)(Wang等,2019b)、JSIA-ReID(joint set-level and instance-level alignment Re-ID)(Wang等,2020)、AlignGAN(alignment generative adversarial network)(Wang等,2019a)、X-Modality(Li等,2020)。

在SYSU-MM01数据集上的对比实验结果如表1所示,轮廓引导的双粒度特征融合网络在最具挑战性的单次全搜索模式下的rank-1和mAP分别为62.42%和58.14%。结果表明,双粒度特征融合有利于模型学习判别性特征,局部特征和全局特征相结合比单独使用其中一种粒度的特征具有更好的性能。此外,本文方法的性能超过了基于GAN的方法,模型更容易收敛并具有更快的训练速度,不会引入额外的噪声。在RegDB数据集上的对比实验结果如表2所示,本文方法的rank-1和mAP分别为84.42%和77.82%,相比于其他方法具有较高的识别准确率。在两个公开数据集SYSU-MM01和RegDB上的对比实验结果证明了本方法的优越性。

表1　不同方法在SYSU-MM01数据集上的比较结果

Table 1　Comparison results of different methods on SYSU-MM01 dataset

/%

方法	全搜索			室内搜索		
	rank-1	rank-10	mAP	rank-1	rank-10	mAP
Zero-Padding	14.80	54.12	15.95	20.58	68.38	26.92
TONE+HCML	14.32	53.16	16.16	24.52	73.25	30.08
HSME	20.68	62.74	23.12	–	–	–
cmGAN	26.97	67.51	27.80	31.63	77.23	42.19
BDTR	27.32	66.96	27.32	31.92	77.18	41.86
D^2RL	28.90	70.60	29.20	–	–	–
Hi-CMD	34.94	77.58	35.94	–	–	–
JSIA-ReID	38.10	80.70	36.90	43.80	86.20	52.90
AlignGAN	42.40	85.00	40.70	45.90	87.60	54.30
AGW	47.50	84.39	47.65	54.17	91.14	62.97
DFE	48.71	88.86	48.59	52.25	89.86	59.68
X-Modality	49.92	89.79	50.73	–	–	–
MACE	51.64	87.25	50.11	57.35	93.02	64.79
LBA	55.41	–	54.14	58.46	–	66.33
TSLFN	56.96	91.50	54.95	59.74	92.07	64.91
NFS	56.91	91.34	55.45	62.79	96.53	69.79
MSO	58.70	**92.06**	56.42	63.09	**96.61**	**70.31**
本文	**62.42**	92.01	**58.14**	**64.18**	93.21	69.28

注:加粗字体表示各列最优结果,"–"表示原文献未提供数据。

3.4　消融实验与分析

为了验证轮廓增广和模型各组成部分的有效性,并探究不同特征融合方法和权衡参数的影响,进行消融实验。相比于RegDB数据集,SYSU-MM01数据集的图像数量更多,拍摄场景和相机视角也更加复杂多变。各种方法在SYSU-MM01数据集上的性能远不如在RegDB数据集上的性能,对其做更深入的探究是很有必要的。因此,消融实验在SYSU-MM01

表2　不同方法在RegDB数据集上的比较结果

Table 2　Comparison results of different methods on RegDB dataset

/%

方法	rank-1	rank-10	mAP
Zero-Padding	17.75	34.21	18.90
TONE+HCML	24.44	47.53	20.08
HSME	50.85	73.36	47.00
BDTR	33.47	58.42	31.83
D^2RL	43.40	66.10	44.10
Hi-CMD	70.93	86.39	66.04
JSIA-ReID	48.50	–	48.90
AlignGAN	57.90	–	53.60
AGW	70.05	86.21	66.37
DFE	70.13	86.32	69.14
X-Modality	62.21	83.13	60.18
MACE	72.37	88.40	69.09
LBA	74.17	–	67.64
NFS	80.54	91.96	72.10
MSO	73.60	88.60	66.90
本文	**84.42**	**94.85**	**77.82**

注：加粗字体表示各列最优结果，“–”表示原文献未提供数据。

数据集上进行。

3.4.1　组成部分的有效性

为了评估各组成部分的有效性，实验在基线模型上添加不同的组件，并对性能指标进行定量分析。实验1使用双路网络作为基线，原始可见光图像和红外图像作为输入。实验2和实验3分别表示仅使用全局粒度轮廓特征或局部粒度部件特征作为行人的特征表示。实验4表示将实验3的均匀分割方法替换为软分割方法。实验5指融合两种粒度的特征，这里在Conv1后使用按元素加的方式完成全局融合操作。实验6表示将实验5的均匀分割方法替换为软分割方法。

在SYSU-MM01数据集上各组成部分的有效性如表3所示。与实验1相比，实验2的rank-1提升了7.76%，mAP提升了6.60%；而实验3对应的提升值分别为6.90%和4.81%。实验2的提升效果更显著，表明了在本文提出的模型中，全局粒度轮廓特征比局部粒度部件特征更有效，同时也体现了轮廓是一种具有较强判别性的模态共享特征。与实验2和实验3相比，实验5的结果证明了融合全局特征和局部特征的重要性。全局特征包含整体的语义信息，但可能会受到背景噪声的干扰；局部特征是细粒度的，通常与行人身体部位相关。因此，为了尽可能减少模态差异，有必要将两种粒度的特征结合起来。此外，与实验3和实验5相比，实验4和实验6表明，软分割方法可以进一步提高模型的识别准确率。然而，由于可见光模态和红外模态之间的巨大差异，其效果不如可见光单模态下的行人再识别(Sun等，2018)。

表3　各组成部分在SYSU-MM01数据集上的有效性

Table 3　Effectiveness of each component on SYSU-MM01 dataset

/%

实验设置		rank-1	rank-10	rank-20	mAP
实验1	基线	49.89	86.67	93.70	49.13
实验2	基线+全局	57.65	91.13	96.27	55.73
实验3	基线+局部	56.79	90.44	95.78	53.94
实验4	基线+局部+软分割	57.32	90.98	95.69	54.23
实验5	基线+全局+局部	61.66	91.43	96.00	58.02
实验6	基线+全局+局部+软分割	**62.42**	**92.01**	**96.32**	**58.14**

注：加粗字体表示各列最优结果。

3.4.2　融合方法的影响

为了研究全局粒度融合方法对性能的影响，实验尝试了在不同位置使用不同操作进行特征融合。在SYSU-MM01数据集上的实验结果如表4所示。结果表明，在较浅层融合的性能优于在较深层融合。因为CNN的浅层更倾向于提取图像的形状、边缘和纹理特征，而深层则更偏向于学习抽象特征，且浅层生成的特征图具有更大的空间尺寸。融合操作可以在浅层结合原始图像和轮廓图像各自的细节信息，以便于后续的网络进行学习，从而取得比在深层融合更好的效果。在各种特征融合方式中，拼接操作的性能整体上优于其他方法，因为与按元素的乘或加相比，拼接不会损失信息。但由于拼接操作增加了特征图维度，对计算资源的消耗大于其他两种方法。综合以上考虑，本文实验在Conv1后采用按元素加的方式对特征进行融合。

表 4　不同融合方法在 SYSU-MM01 数据集上的性能
Table 4　Performance of different fusion methods on SYSU-MM01 dataset

/%

位置	按元素乘		按元素加		拼接	
	rank-1	mAP	rank-1	mAP	rank-1	mAP
Conv1	56.56	54.34	57.65	55.73	57.76	55.29
Stage1	**56.68**	**54.92**	56.24	54.16	**59.50**	56.96
Stage2	55.32	53.62	56.56	55.01	59.15	**57.01**
Stage3	54.67	52.96	**57.98**	**55.88**	56.36	54.62
Stage4	50.55	47.97	49.35	49.43	54.69	52.25

注：加粗字体表示各列最优结果。

3.4.3　轮廓增广的有效性

为了验证轮廓增广的有效性，实验分别探究了在无轮廓增广、局部特征轮廓增广和全局特征轮廓增广下双粒度特征融合网络的性能，表 5 给出了在 SYSU-MM01 数据集上的实验结果。数据表明，对局部特征或全局特征进行轮廓增广的结果好于没有轮廓引导的结果，从而验证了轮廓增广的有效性。同时，可以发现采用全局特征轮廓增广带来的性能提升显著高于局部特征的增广。这是因为，轮廓是行人的一种整体性而非局部性的特征描述，对全局特征进行轮廓增广可以引导模型学习基于形状的行人判别特征，并弥补其在长距离关系建模上的不足。而在局部特征轮廓增广中，由于图像会被划分成不同的区域，整体性的轮廓将被分解为局部性的边缘，导致模型无法感知图像级的关联信息。因此，本文所提出的模型仅对全局特征进行了轮廓增广。

表 5　轮廓增广在 SYSU-MM01 数据集上的有效性
Table 5　Effectiveness of contour augmentation on SYSU-MM01 dataset

/%

实验设置	rank-1	rank-10	rank-20	mAP
无轮廓增广	57.04	90.30	95.82	54.11
局部特征轮廓增广	58.16	90.56	95.71	54.90
全局特征轮廓增广	**61.66**	**91.43**	**96.00**	**58.02**

注：加粗字体表示各列最优结果。

3.4.4　权衡参数的影响

为了探究全局特征损失和局部特征损失的比例系数对性能的影响，在 SYSU-MM01 数据集上采用不同的权衡参数 λ 进行实验，结果如表 6 所示。结果表明，当权衡参数 λ 介于 1.0～1.5 时，模型的性能较好。考虑到 λ = 1.0 时，rank-1 和 mAP 性能突出，且 rank-10 和 rank-20 的值亦接近最优，本文实验将权衡参数 λ 的值设置为 1.0。

表 6　不同权衡参数在 SYSU-MM01 数据集上的性能
Table 6　Performance of different trade-off parameters on SYSU-MM01 dataset

/%

权衡参数 λ	rank-1	rank-10	rank-20	mAP
0.2	53.69	90.01	95.24	52.67
0.5	58.61	90.85	95.45	56.00
1.0	**61.66**	91.43	96.00	**58.02**
1.5	60.74	**91.93**	**96.45**	56.97
2.0	59.72	90.90	96.13	56.37

注：加粗字体表示各列最优结果。

4　结　论

本文将显式轮廓信息引入红外—可见光跨模态行人再识别中，旨在减小模态间差异。为了充分利用轮廓特征，本文将轮廓作为辅助模态，提出了一种轮廓引导的双粒度特征融合网络，用于跨模态行人再识别。全局粒度融合增强了原始图像的轮廓特征表示，生成轮廓增广特征。局部粒度融合进一步融合基于行人部件的局部特征和轮廓增广特征，从而得到具备更强判别能力的图像表达。在两个公开数据集 SYSU-MM01 和 RegDB 上的实验结果验证了模型的有效性。

本文模型验证了轮廓引导和双粒度特征融合的有效性，然而模型的性能仍有待提高。后续工作将探索如何更有效地利用轮廓线索增强特征的表达能力。例如，尝试其他的轮廓特征融合方法或设计相应的损失函数，进一步提高识别准确率。此外，将考虑采用随机擦除、噪声添加等数据扩张技术提升模型的泛化能力，以适应更加复杂多变的真实行人再识别场景。

参考文献（References）

Chen J X, Yang Q Z, Meng J K, Zheng W S and Lai J H. 2019. Contour-guided person re-identification//Proceedings of the 2nd

Chinese Conference on Pattern Recognition and Computer Vision. Xi'an, China: Springer: 296-307 [DOI: 10.1007/978-3-030-31726-3_25]

Chen Y H S, Wan L, Li Z H, Jing Q Y and Sun Z Y. 2021. Neural feature search for RGB-infrared person re-identification//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 587-597 [DOI: 10.1109/CVPR46437.2021.00065]

Choi S, Lee S, Kim Y, Kim T and Kim C. 2020. Hi-CMD: hierarchical cross-modality disentanglement for visible-infrared person re-identification//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10254-10263 [DOI: 10.1109/CVPR42600.2020.01027]

Dai P Y, Ji R R, Wang H B, Wu Q and Huang Y Y. 2018. Cross-modality person re-identification with generative adversarial training//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: AAAI Press: 677-683 [DOI: 10.24963/ijcai.2018/94]

Gao Y J, Liang T F, Jin Y, Gu X Y, Liu W, Li Y D and Lang C Y. 2021. MSO: multi-feature space joint optimization network for RGB-infrared person re-identification//Proceedings of the 29th ACM International Conference on Multimedia. Virtual, China: Association for Computing Machinery: 5257-5265 [DOI: 10.1145/3474085.3475643]

Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann F A and Brendel W. 2022. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness [EB/OL]. [2022-04-02]. https://arxiv.org/pdf/1811.12231.pdf

Han K, Wang Y H, Chen H T, Chen X H, Guo J Y, Liu Z H, Tang Y H, Xiao A, Xu C J, Xu Y X, Yang Z H, Zhang Y M and Tao D C. 2022. A survey on vision transformer [EB/OL]. [2022-04-02]. https://arxiv.org/pdf/2012.12556.pdf

Hao Y, Wang N N, Li J and Gao X B. 2019a. HSME: hypersphere manifold embedding for visible thermal person re-identification. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1): 8385-8392 [DOI: 10.1609/aaai.v33i01.33018385]

Hao Y, Wang N N, Gao X B, Li J and Wang X Y. 2019b. Dual-alignment feature embedding for cross-modality person re-identification//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: Association for Computing Machinery: 57-65 [DOI: 10.1145/3343031.3351006]

Li D G, Wei X, Hong X P and Gong Y H. 2020. Infrared-visible cross-modal person re-identification with an X modality. Proceedings of the AAAI Conference on Artificial Intelligence, 34(4): 4610-4617 [DOI: 10.1609/aaai.v34i04.5891]

Liu H J, Tan X H and Zhou X C. 2021. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. IEEE Transactions on Multimedia, 23: 4414-4425 [DOI: 10.1109/TMM.2020.3042080]

Liu Y, Cheng M M, Hu X W, Wang K and Bai X. 2017. Richer convolutional features for edge detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5872-5881 [DOI: 10.1109/CVPR.2017.622]

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]

Park H, Lee S, Lee J and Ham B. 2021. Learning by aligning: visible-infrared person re-identification using cross-modal correspondences//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 12026-12035 [DOI: 10.1109/ICCV48922.2021.01183]

Shi W D, Zhang Y Z, Liu S W, Zhu S D and Bao J N. 2020. Person re-identification based on deformation and occlusion mechanisms. Journal of Image and Graphics, 25(12): 2530-2540 (史维东, 张云洲, 刘双伟, 朱尚栋, 暴吉宁. 2020. 针对形变与遮挡问题的行人再识别. 中国图象图形学报, 25(12): 2530-2540) [DOI: 10.11834/jig.200016]

Sun Y F, Zheng L, Yang Y, Tian Q and Wang S J. 2018. Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline)//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 501-518 [DOI: 10.1007/978-3-030-01225-0_30]

Wang G A, Zhang T Z, Cheng J, Liu S, Yang Y and Hou Z G. 2019a. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 3622-3631 [DOI: 10.1109/ICCV.2019.00372]

Wang G A, Zhang T Z, Yang Y, Cheng J, Chang J L, Liang X and Hou Z G. 2020. Cross-modality paired-images generation for RGB-infrared person re-identification. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7): 12144-12151 [DOI: 10.1609/aaai.v34i07.6894]

Wang X L, Girshick R, Gupta A and He K M. 2018. Non-local neural networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7794-7803 [DOI: 10.1109/CVPR.2018.00813]

Wang Z X, Wang Z, Zheng Y Q, Chuang Y Y and Satoh S. 2019b. Learning to reduce dual-level discrepancy for infrared-visible person re-identification//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 618-626 [DOI: 10.1109/CVPR.2019.00071]

Wei Z Y, Yang X, Wang N N and Gao X B. 2021. Syncretic modality collaborative learning for visible infrared person re-identification//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 225-234 [DOI: 10.1109/ICCV48922.2021.00029]

Wu A C, Zheng W S, 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. Venice, Italy: IEEE: 5390-5399 [DOI: 10.1109/ICCV.2017.575]

Wu B C, Xu C F, Dai X L, Wan A, Zhang P Z, Yan Z C, Tomizuka M, Gonzalez J, Keutzer K and Vajda P. 2021. Visual transformers: where do transformers really belong in vision models?//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 579-589 [DOI: 10.1109/ICCV48922.2021.00064]

Ye M, Lan X Y, Leng Q M and Shen J B. 2020. Cross-modality person re-identification via modality-aware collaborative ensemble learning. IEEE Transactions on Image Processing, 29: 9387-9399 [DOI: 10.1109/TIP.2020.2998275]

Ye M, Lan X Y, Li J W and Yuen P C. 2018a. Hierarchical discriminative learning for visible thermal person re-identification. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1): 7501-7508 [DOI: 10.1609/aaai.v32i1.12293]

Ye M, Shen J B, Lin G J, Xiang T, Shao L and Hoi S C H. 2022. 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]

Ye M, Wang Z, Lan X Y and Yuen P C. 2018b. Visible thermal person re-identification via dual-constrained top-ranking//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: AAAI Press: 1092-1099 [DOI: 10.24963/ijcai.2018/152]

Zhang Y K, Yan Y, Lu Y and Wang H Z. 2021. Towards a unified middle modality learning for visible-infrared person re-identification//Proceedings of the 29th ACM International Conference on Multimedia. Virtual, China: Association for Computing Machinery: 788-796 [DOI: 10.1145/3474085.3475250]

Zhao C R, Qi D, Dou S G, Tu Y P, Sun T L, Bai S, Jiang X Y, Bai X and Miao D Q. 2021. Key technology for intelligent video surveillance: a review of person re-identification. Scientia Sinica Informationis, 51(12): 1979-2015 (赵才荣, 齐鼎, 窦曙光, 涂远鹏, 孙添力, 柏松, 蒋忻洋, 白翔, 苗夺谦. 2021. 智能视频监控关键技术：行人再识别研究综述. 中国科学：信息科学, 51(12): 1979-2015) [DOI: 10.1360/SSI-2021-0211]

Zhu Y X, Yang Z, Wang L, Zhao S, Hu X and Tao D P. 2020. Hetero-center loss for cross-modality person re-identification. Neurocomputing, 386: 97-109 [DOI: 10.1016/j.neucom.2019.12.100]

作者简介

马潇峰，男，硕士研究生，主要研究方向为跨模态行人再识别。E-mail：maxf@ncepu.edu.cn

程文刚，通信作者，男，副教授，主要研究方向为多媒体信息处理。E-mail：wgcheng@ncepu.edu.cn