DOI: 10.13382/j.jemi.B2307051

细微特征增强的多级联合聚类跨模态行人重识别算法*

范馨月 张 阔 张 干 李嘉辉
(重庆邮电大学通信与信息工程学院 重庆 400065)

摘 要:目前跨模态行人重识别研究注重于通过身份标签在全局特征或局部特征上提取模态共享特征来减少模态差异,但却忽视了具有辨别力的细微特征。为此提出了一种基于特征增强的聚类学习网络,该网络通过全局和局部特征来挖掘并增强不同模态的细微特征,并结合多级联合聚类学习策略,最小化模态差异和类内变化。针对训练数据设计了随机颜色转换模块,在图像输入端增加模态之间的交互,以克服颜色偏差的影响。通过在公共数据集上进行实验,验证了所提方法的有效性,其中在 SYSU-MM01 数据集的全搜索模式下 Rank-1 和 mAP 分别达到了 70.52%和 64.02%;在 RegDB 数据集的 V2I 检索模式下 Rank-1 和 mAP 分别达到了 88.88%和 80.93%。

关键词: 行人重识别;跨模态;随机颜色转换;细微特征增强;多级联合聚类学习
中图分类号: TP391.4 **文献标识码**:A **国家标准学科分类代码**: 510.40

Cross-modal person re-identification algorithm based on multi-level joint clustering with subtle feature enhancement

Fan Xinyue Zhang Kuo Zhang Gan Li Jiahui
(School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: The current cross-modal person re-identification research focuses on extracting modality-shared features from global features or local features via identity labels to reduce modality differences, but ignores the Subtle features of discernment. This paper proposes a feature enhanced clustering learning (FECL) network. The network mines and enhances the subtle features of different modalities through global and local features, and combines a multilevel joint clustering learning strategy to minimize the modal differences and intraclass variation. In addition, this paper also designs a random color transition module for training data, which increases the interaction between modalities at the image input to overcome the influence of color deviation. The experiments on public datasets verify the effectiveness of the proposed methods. In the All-search mode of SYSU-MM01 dataset, the Rank-1 and mAP reach 70.52% and 64.02%. In the V2I retrieval mode of RegDB dataset, the Rank-1 and mAP reach 88.88% and 80.93%.

Keywords: person re-identification; cross-modality; random color transition; subtle feature enhancement; multilevel joint clustering learning

0 引 言

近年来,随着人们对于公共安全的日益重视,行人重识别技术[1](Re-IDentification, Re-ID)在智能监控领域中得到广泛应用,Re-ID 利用机器学习的方法在多个不重叠摄像头之间进行检索。在实际监控系统中,可见光相机无法在黑暗的条件下清晰成像,难以拍摄到有效的行人图像,行人的细节和特征无法准确提取。红外相机被应用到视频监控系统中,用于拍摄夜间行人图像。但由于在不同光照条件下,收集到的图像清晰度会存在较大差异,导致可见光图像和红外图像存在较大的模态差异,其次由于相机环境变化、行人姿态变化以及障碍物遮挡等因素的影响,可能会出现与另一个具有类似着装和

收稿日期:2023-11-15 Received Date: 2023-11-15
*基金项目:国家自然科学基金(62271096)项目资助

体型的行人更为相似的情况。

针对这一问题，Sun 等[2]将全局特征水平划分得到若干个局部特征，扩大了对局部细粒度特征的关注，但忽略了全局特征的粗粒度信息与更加细腻的细微特征。Wu 等[3]设计了一种零填充网络，通过将不同模态的特征填充到特定域节点中，对来自不同模态的特征进行匹配。Hao 等[4]提出了一种超球面嵌入网络（hypersphere manifold embedding，HSME），通过将特征嵌入到一个超球面流形空间中，约束不同模态特征间的距离。Dai 等[5]通过对抗性学习的方式提出跨模态生成对抗网络（cross-modality generative adversarial network，cmGAN），通过生成器和判别器的对抗性学习来实现跨模态特征的生成和匹配。生成器负责将其中一个模态的特征映射到另一个模态，以生成对应的合成特征。判别器则负责区分合成特征和真实的原始特征，以提供对生成特征质量的反馈。Zhang 等[6]提出了一种对偶相互学习方法（dual mutual learning，DML），该方法在两种模态之间进行相互学习，学习用于外观相似性度量的模态共享特征。Hao 等[7]设计了一种模态混淆学习机制（modality confusion learning network，MCLNet），将可见光模态与红外模态进行混淆，学习模态不变特征。这些方法通常侧重于从行人实例层面通过身份标签在全局特征或局部特征上提取模态共享特征来缓解模态差异，忽视了通道层面语义信息的联系和对行人细微判别特征的挖掘，从而影响模型的检索能力。

本文提出了一种基于特征增强的聚类学习（feature enhanced clustering learning，FECL）网络，在图像输入端增加模态间的交互，缓解颜色偏差的影响；利用双流网络提取行人特征，通过局部特征与全局特征来挖掘并增强细微特征；同时利用多级联合聚类学习策略来约束网络学习，缓解模态差异、提高类内特征相似性和扩大类间差异，提升网络判别能力。本文主要贡献如下：

1）本文提出了一种 FECL 跨模态行人重识别方法，该方法旨在增强细微特征并缓解模态差异。

2）由于拍摄角度和姿态变化，行人动作、鞋子、眼镜等细微特征不明显，因此提出细微特征增强（subtle feature enhancement，SFE）模块，学习具有辨别力的细微特征。

3）设计了一种多级联合聚类学习（multilevel joint clustering learning，MJCL）策略，在通道层面缓解模态差异，扩大类间距离和提高类内相似性。

4）针对输入数据设计了随机颜色转换（random color transition，RCT）模块，通过转换训练数据的部分颜色信息来平衡神经网络中颜色特征与颜色无关特征之间的权重，从而克服颜色偏差的影响。

1　本文方法

1.1　网络总体框架

由于两种模态图像之间的颜色差异和不断变化的相机环境，网络难以提取到模态间共有的辨别特征。对此，本文提出 FECL 网络，其中包括随机颜色转换模块、特征提取模块、细微特征增强模块和多级联合聚类学习策略，其网络结构如图 1 所示。FECL 网络在数据预处理阶段转换不同模态的颜色信息，随后构造全局特征与局部特征关系对，最后结合多级联合聚类学习策略来约束网络学习。

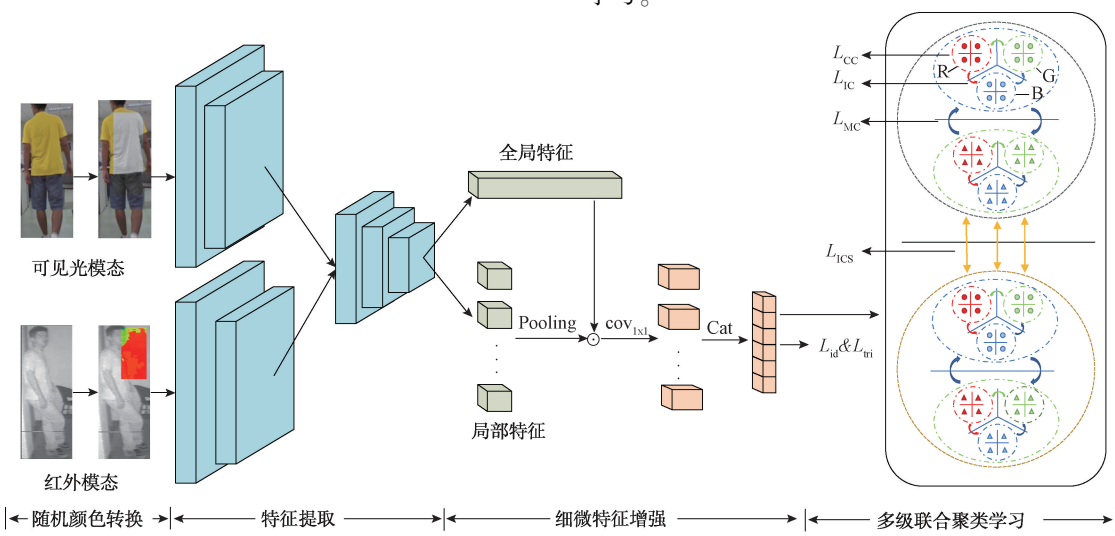

图 1　FECL 网络结构

Fig. 1　FECL network structure

1.2 随机颜色转换模块

在 Re-ID 领域,通过数据增强的方法对输入数据进行细微改动或生成新的数据集,可提高数据的丰富性,从而使网络学习到更多的信息。Zhong 等[8]提出了随机擦除策略,在图像中随机选取一个矩形框,并擦除矩形框内像素,模拟不同遮挡程度的真实图像。Ye 等[9]提出了一种通道增强联合学习策略,通过将通道交换和随机擦除相结合,提升模型对颜色变化的鲁棒性。受这些方法的启发,本文提出了一种随机颜色转换模块,在输入端转换两个模态图像的颜色信息,使图像互相包含另一模态的颜色信息,增强两个模态信息之间的交互,缓解模态差异。

1)可见光模态颜色信息转换

每个批次数据(batch)随机采样 K 个行人,每个行人提取 M 张图像,batch 大小为 $N = K \times M$。可见光图像表示为 $\boldsymbol{x}^{\mathrm{v}} = \{\boldsymbol{x}_i^{\mathrm{v}} \mid i = 1,2,\cdots,N/2\}$,其中 $\boldsymbol{x}_i^{\mathrm{v}} = \{\boldsymbol{x}_i^{\mathrm{v}} \mid \boldsymbol{y}_i\}$ 表示 batch 中第 i 张可见光图像,$\boldsymbol{y}_i$ 表示行人标签,红外图像 $\boldsymbol{x}^{\mathrm{r}} = \{\boldsymbol{x}_i^{\mathrm{r}} \mid i = 1,2,\cdots,N/2\}$,$\boldsymbol{x}_i^{\mathrm{r}} = \{\boldsymbol{x}_i^{\mathrm{r}} \mid \boldsymbol{y}_i\}$ 表示 batch 中第 i 张红外图像。

如图 2 所示,可见光模态颜色信息转换的主要思想是随机选择一个通道来替代其他通道,其中虚线矩形框内为转换后的红外颜色信息图像,生成新的训练图像,如式(1)~(3)所示:

$$\tilde{\boldsymbol{x}}_i^{\mathrm{v}} = t^{\mathrm{v-r}}(\mathrm{rect}(\boldsymbol{x}_i^{\mathrm{v}})) \tag{1}$$

$$\mathrm{rect} = \mathrm{Randrect}(\boldsymbol{x}) \tag{2}$$

$$t^{\mathrm{v-r}}(\boldsymbol{x}^{\mathrm{v}}) = \begin{cases} \boldsymbol{x}^{\mathrm{v,R}}, \boldsymbol{x}^{\mathrm{v,R}}, \boldsymbol{x}^{\mathrm{v,R}}, p = 0 \\ \boldsymbol{x}^{\mathrm{v,G}}, \boldsymbol{x}^{\mathrm{v,G}}, \boldsymbol{x}^{\mathrm{v,G}}, p = 1, p \in \mathrm{rand}[0,1,2] \\ \boldsymbol{x}^{\mathrm{v,G}}, \boldsymbol{x}^{\mathrm{v,G}}, \boldsymbol{x}^{\mathrm{v,G}}, p = 2 \end{cases} \tag{3}$$

其中 $\tilde{\boldsymbol{x}}_i^{\mathrm{v}}$ 表示转换后包含红外颜色信息的图像,$\mathrm{Randrect}(\cdot)$ 用于在图像中随机生成一个矩形框,$t^{\mathrm{v-r}}(\cdot)$ 表示红外模态信息转换函数,用于随机选择一个通道来替换可见光图片的其余两个通道。

2)红外模态颜色信息转换

如图 3 所示,与可见光模态同理,红外模态信息转换将其局部信息转换为伪彩色图像信息,其中虚线矩形框内为转换后的伪彩色颜色信息图像。由于红外图片是单通道,在输入前需要将红外图片转换为三通道,然后再进行模态信息转换,如式(4)、(5)所示:

$$\boldsymbol{x}_i^{\mathrm{r}} = (\boldsymbol{x}_i^{\mathrm{r}^*}, \boldsymbol{x}_i^{\mathrm{r}^*}, \boldsymbol{x}_i^{\mathrm{r}^*}) \tag{4}$$

$$\tilde{\boldsymbol{x}}_i^{\mathrm{r}} = t^{\mathrm{r-v}}(\mathrm{rect}(\boldsymbol{x}_i^{\mathrm{r}})) \tag{5}$$

其中,$\boldsymbol{x}_i^{\mathrm{r}^*}$ 表示单通道的红外图像,$\boldsymbol{x}_i^{\mathrm{r}}$ 表示三通道的红外图像,$\tilde{\boldsymbol{x}}_i^{\mathrm{r}}$ 表示转换后包含可见光模态信息的图像,

图 2 可见光模态颜色信息转换

Fig. 2 Visible modal color information conversion

$t^{\mathrm{r-v}}(\cdot)$ 表示可见光模态信息转换函数,用于生成伪彩色信息。

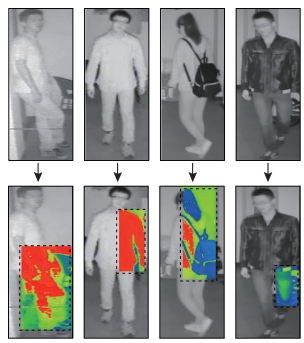

图 3 红外模态颜色信息转换

Fig. 3 Infrared modal color information conversion

1.3 特征提取模块

本文骨干网络采用双流网络,采用 ResNet50[10]作为骨干网络。对 ResNet50 网络的进行修改,如图 1 所示,Stage0 和 Stage1 参数独立,分别提取可见光和红外模态的特定特征,Stage2~Stage4 参数共享,提取两个模态的共享特征。

1.4 细微特征增强模块

通常神经网络的关注区域更多集中在图像的躯干、面部等显著特征,而忽视了对衣服、眼镜和鞋子等细微特征的关注,这些细微特征往往是容易被忽视但却具有辨别力的特征。现有基于全局特征[11]的方法大多直接从输入图像中提取特征,但光照、遮挡和人体姿势变化会对特征提取造成较大的干扰。相比全局特征,基于局部特征[2]的方法将图像划分为若干部分,可排除背景等不相关信息,受干扰和背景杂波的影响更小,但图像切分容易破坏图像结构,造成细微特征丢失。文献[12-13]通过全局特征与局部特征相互协同,虽然能形成更全面的特征描述符,但仍然忽视了对细微语义信息的关注。基于上述方法的改进,为挖掘行人的细微语义信息,本文通过构

建局部特征和全局特征的成对关系,挖掘并增强其中所包含的细微特征。

如图 4 所示,将骨干网络提取到的全局特征 $\boldsymbol{f}_{\text{global}}$ 划分为 l 个局部特征,如式(6)所示:

$$\boldsymbol{f}_{\text{global}} = [\boldsymbol{f}_{\text{local}}^{1}, \boldsymbol{f}_{\text{local}}^{2}, \cdots, \boldsymbol{f}_{\text{local}}^{l}] \in \mathbb{R}^{C\times H/l\times W} \tag{6}$$

其中, C 表示通道数目, H 和 W 表示图像的长和宽,每个局部特征 $\boldsymbol{f}_{\text{local}}^{k}$ 代表行人身体的不同部位,包含了不同的语义信息,经过维度变化后与全局特征 $\boldsymbol{f}_{\text{global}}$ 按元素相乘,如式(7)所示:

$$\boldsymbol{f}_k = \boldsymbol{f}_{\text{local}}^{k} \text{e} \boldsymbol{f}_{\text{global}}, k \in [1,2,\cdots,l] \tag{7}$$

其中,e 表示点积,新的特征向量 $\boldsymbol{f}_k$ 包含全局特征和增强后的细微特征。利用广义平均池化(generalized-mean, Gem)来提取第 k 个特征,Gem 表达式如式(8)所示:

$$\boldsymbol{x}_{\text{gem}} = \left(\frac{1}{|\boldsymbol{X}|}\sum_{\boldsymbol{x}_i\in\boldsymbol{X}}\boldsymbol{x}_i^{p}\right)^{\frac{1}{p}} \in \mathbb{R}^{C} \tag{8}$$

其中, $\boldsymbol{x}_{\text{gem}}$ 表示池化后的特征, p 是一个超参数,可以在网络中反向传播,当 $p \to \infty$ 时,Gem 近似于最大池化, $p \to 1$ 时,Gem 近似于平均池化。通过卷积层对特征进行降维,减少特征通道数,然后在通道维度上进行特征组合,如式(9)、(10)所示:

$$\boldsymbol{f}_k^{*} = \text{cov}_{1\times1}(\text{Gem}(\boldsymbol{f}_k)) \tag{9}$$

$$\boldsymbol{F} = [\boldsymbol{f}_1^{*}, \boldsymbol{f}_2^{*}, \cdots \boldsymbol{f}_k^{*}], k \in [1,2,\cdots,l] \tag{10}$$

其中, $\text{cov}_{1\times1}$ 是 1 × 1 的卷积层, $\boldsymbol{F}$ 表示将 k 个特征组合后的特征。

1.5　多级联合聚类学习策略

考虑到跨模态特征的语义和结构分布不同,本文提出多级联合聚类学习策略,在通道级逐步优化特征分布,由通道聚类、类内聚类、模态聚类、和类间分离 4 个部分组成,其关系如图 1 中联合聚类学习模块所示。

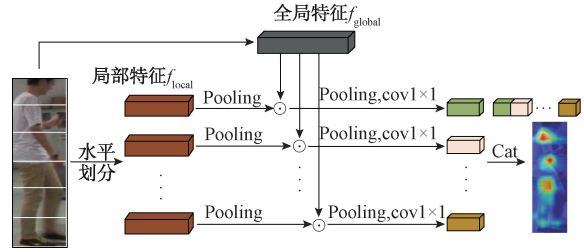

图 4　细微特征增强模块结构

Fig. 4　Structure of subtle feature enhancement module

1)通道聚类

在同一批次的特征空间中,不同行人图像的通道内信息存在差异和分布不集中的现象,因此首先进行通道内聚类。将细微特征增强模块得到特征 $\boldsymbol{F}$ 拆分为可见光模态特征 $\boldsymbol{f}^{\text{v}} = [\boldsymbol{f}_i^{\text{v,R}}, \boldsymbol{f}_i^{\text{v,G}}, \boldsymbol{f}_i^{\text{v,B}}]$ 和红外模态特征 $\boldsymbol{f}^{\text{r}} = [\boldsymbol{f}_i^{\text{r,R}}, \boldsymbol{f}_i^{\text{r,G}}, \boldsymbol{f}_i^{\text{r,B}}]$,通道聚类如式(11)~(13)所示:

$$L_{\text{CC}}^{\text{v}} = \frac{1}{K}\sum_{i=1}^{K}\left(\|\boldsymbol{f}_i^{\text{v,R}} - \psi(\boldsymbol{f}^{\text{v,R}})\|_2 + \|\boldsymbol{f}_i^{\text{v,G}} - \psi(\boldsymbol{f}^{\text{v,G}})\|_2 + \|\boldsymbol{f}_i^{\text{v,B}} - \psi(\boldsymbol{f}^{\text{v,B}})\|_2\right) \tag{11}$$

$$L_{\text{CC}}^{\text{r}} = \frac{1}{K}\sum_{i=1}^{K}\left(\|\boldsymbol{f}_i^{\text{r,R}} - \psi(\boldsymbol{f}^{\text{r,R}})\|_2 + \|\boldsymbol{f}_i^{\text{r,G}} - \psi(\boldsymbol{f}^{\text{r,G}})\|_2 + \|\boldsymbol{f}_i^{\text{r,B}} - \psi(\boldsymbol{f}^{\text{r,B}})\|_2\right) \tag{12}$$

$$L_{\text{CC}} = L_{\text{CC}}^{\text{v}} + L_{\text{CC}}^{\text{r}} \tag{13}$$

其中 $\psi(\cdot)$ 是计算通道内特征的平均值,将其视为某一通道的中心, $\|\cdot\|$ 表示欧式距离。如图 5(a)所示,通过优化 L_{CC} ,将通道内特征聚集到通道中心,解决通道内数据分布不集中的问题,以便更好地进行后续聚类。

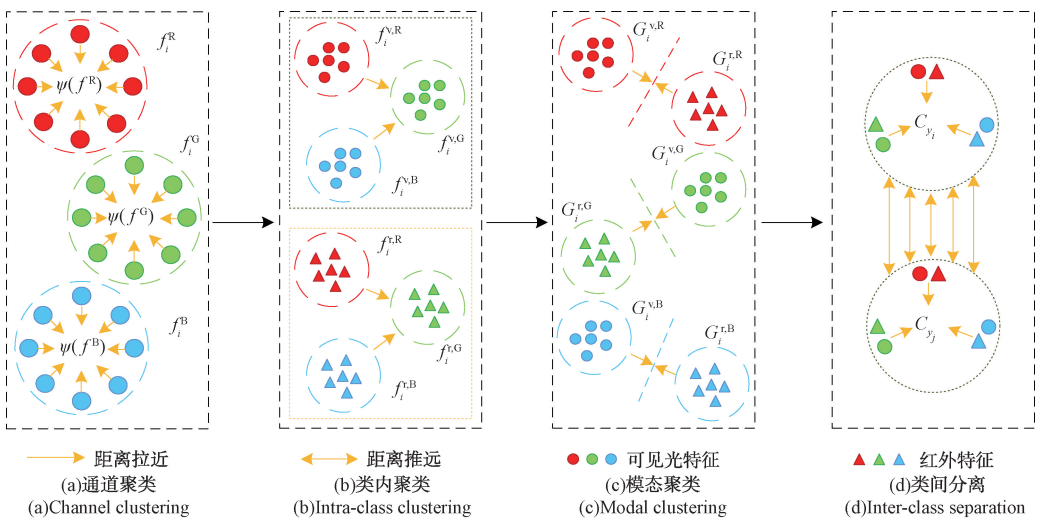

图 5　联合聚类学习策略示意图

Fig. 5　Schematic diagram of the multilevel joint clustering learning

2)类内聚类

与单模态的 Re-ID 相似,跨模态行人重识别中人的外貌也容易受到衣物和背景等因素的影响,这使得跨模态检索任务更加困难。为解决这一问题,现有的方法大多采用中心损失[14]来学习每类特征的类中心,并约束样本与相应类之间的距离,但却忽略了通道间语义特征的联系。文献[15]中指出在图像的 R、G 和 B 3 个通道里,G 通道的卷积核空间能更可靠的提取到通道语义信息,因此将 G 通道作为类中心进行类内聚类,如式(14)所示:

$$L_{\mathrm{IC}}=\frac{1}{K}\sum_{i=1}^{K}\left(\boldsymbol{f}_i^{\mathrm{v,R}}\cdot\log\frac{\boldsymbol{f}_i^{\mathrm{v,R}}}{\boldsymbol{f}_i^{\mathrm{v,G}}}+\boldsymbol{f}_i^{\mathrm{v,B}}\cdot\log\frac{\boldsymbol{f}_i^{\mathrm{v,B}}}{\boldsymbol{f}_i^{\mathrm{v,G}}}\right)+\frac{1}{K}\sum_{i=1}^{K}\left(\boldsymbol{f}_i^{\mathrm{r,R}}\cdot\log\frac{\boldsymbol{f}_i^{\mathrm{r,R}}}{\boldsymbol{f}_i^{\mathrm{r,G}}}+\boldsymbol{f}_i^{\mathrm{r,B}}\cdot\log\frac{\boldsymbol{f}_i^{\mathrm{r,B}}}{\boldsymbol{f}_i^{\mathrm{r,G}}}\right)\tag{14}$$

其中,L_{IC} 表示 R、G、B 通道的语义一致性,如图 5(b)所示,通道 R 和通道 B 向通道 G 靠拢,缩小了各通道之间的距离,使样本接近相应的类中心,特征分布更加紧簇。

3)模态聚类

为减小模态间差异,保持通道间的同一性,使 $\boldsymbol{f}_i^{\mathrm{v,R}}$ 与 $\boldsymbol{f}_i^{\mathrm{r,R}}$ 、$\boldsymbol{f}_i^{\mathrm{v,G}}$ 与 $\boldsymbol{f}_i^{\mathrm{r,G}}$ 以及 $\boldsymbol{f}_i^{\mathrm{v,B}}$ 与 $\boldsymbol{f}_i^{\mathrm{r,B}}$ 分布一致,利用互相关矩阵来约束不同模态的通道间距离,如式(15)、(16)所示:

$$\boldsymbol{G}_i^{\mathrm{m,c}}=\frac{\boldsymbol{f}_i^{\mathrm{m,c}}\cdot\boldsymbol{f}_i^{\mathrm{m,c}^{\mathrm{T}}}}{\|\boldsymbol{f}_i^{\mathrm{m,c}}\cdot\boldsymbol{f}_i^{\mathrm{m,c}^{\mathrm{T}}}\|_2}\tag{15}$$

$$L_{\mathrm{MC}}=\frac{1}{K}\sum_{i=1}^{K}(\|\boldsymbol{G}_i^{\mathrm{v,R}}-\boldsymbol{G}_i^{\mathrm{r,R}}\|_2+\|\boldsymbol{G}_i^{\mathrm{v,G}}-\boldsymbol{G}_i^{\mathrm{r,G}}\|_2+\|\boldsymbol{G}_i^{\mathrm{v,B}}-\boldsymbol{G}_i^{\mathrm{r,B}}\|_2)\tag{16}$$

其中,$\mathrm{m}\in[\mathrm{v},\mathrm{r}]$,$\mathrm{c}\in[\mathrm{R},\mathrm{G},\mathrm{B}]$,$\boldsymbol{G}$ 为互相关性矩阵;如图 5(c)所示,通过优化 L_{MC},约束不同模态间的相同通道的距离。

4)类间分离

在缩小类内距离,使类内相似度最大化的同时,类间分离通过拉远不同类之间的距离,最大化类间差异,以便更好地进行跨模态检索,如式(17)、(18)所示:

$$\mathbf{X}_{ij}=\|\mathbf{C}_{y_i}-\mathbf{C}_{y_j}\|_2^2\tag{17}$$

$$L_{\mathrm{ICS}}=\sum_{i}^{K}\boldsymbol{X}_{ii}+\sum_{i}^{K}\sum_{j,j\neq i}^{K}(\alpha-\boldsymbol{X}_{ij})\tag{18}$$

其中,X_{ij} 表示不同类中心之间的距离,K 为行人类别数,α 为限制不同类别之间距离的距离因子。通过优化 L_{ICS},使 X_{ii} 趋近于 0,X_{ij} 趋近于 α。$X_{ii}\to 0$ 表示拉近相同类别之间的距离,$X_{ij}\to\alpha$ 表示增加不同类之间的距离。如图 5(d)所示,类间距离不断变大,使模型能够更清晰地进行判别。最后多级联合聚类学习策略的损失函数 L_{MJC} 如式(19)所示:

$$L_{\mathrm{MJC}}=L_{\mathrm{CC}}+L_{\mathrm{IC}}+L_{\mathrm{MC}}+L_{\mathrm{ICS}}\tag{19}$$

1.6 目标函数

使用交叉熵损失和三元组损失[14]来约束网络学习,限制不同特征之间的距离。交叉熵损失用于衡量预测结果和真实类别标签的差异,约束网络将正确分类预测为 1,其他类别预测为 0,交叉熵损失如式(20)所示:

$$L_{\mathrm{id}}=\sum_{i=1}^{S}-q_i\log(p_i)$$
$$\text{s.t. }q_i=\begin{cases}1-\dfrac{S-1}{S}\xi, y=i\\ \dfrac{\xi}{S}, y\neq i\end{cases}\tag{20}$$

其中,S 表示标签数量,q_i 表示预测结果,若预测正确,即 q_i 置为 1,若预测错误,则 q_i 置为 0。ξ 为超参数,本文将其设置为 0.1,用于在训练时模拟标签可能存在的错误。

三元组损失的目的是限制特征之间的距离,优化网络学习,其表达式如式(21)所示:

$$L_{\mathrm{tri}}=\sum_{i=1}^{N}[\|\boldsymbol{f}_i^{\mathrm{a}}-\boldsymbol{f}_i^{\mathrm{pos}}\|_2-\|\boldsymbol{f}_i^{\mathrm{a}}-\boldsymbol{f}_i^{\mathrm{neg}}\|_2+\sigma]_+\tag{21}$$

其中,$\boldsymbol{f}_i^{\mathrm{a}}$ 表示锚样本,$\boldsymbol{f}_i^{\mathrm{pos}}$ 表示正样本,$\boldsymbol{f}_i^{\mathrm{neg}}$ 表示负样本。总体损失如式(22)所示:

$$L=L_{\mathrm{id}}+L_{\mathrm{tri}}+\lambda L_{\mathrm{MJC}}\tag{22}$$

其中,λ 用于调节相应损失函数的占比权重。

2 实验结果与分析

2.1 数据集与评估指标

SYSU-MM01[3]:SYSU-MM01 数据集包含 491 个行人的 30 671 张可见光图像和 15 792 张红外图像。SYSU-MM01 数据集具有两种搜索模式,在全搜索(All-search)模式下,摄像头 1、2、4、5 组成 Gallery 集,摄像头 3、6 组成 Query 集。在室内搜索(Indoor-search)模式下,摄像头 1、2 组成 Gallery 集,摄像头 3、6 组成 Query 集。

RegDB[16]:RegDB 数据集包含 412 个行人的 8 240 张图像,训练集包含 4 120 张图像,测试集包含 4 120 张图像。在测试阶段,本文采用 Visible-to-Infrared(V2I)模式进行实验,即利用可见光图像检索红外图像。

评估指标:采用累积匹配特征(cumulative matching characteristics, CMC)曲线中的 Rank-n 和均值平均精度(mean average precision, mAP)作为本文的评估指标。

2.2 实验设置

本文实验使用的软件平台为 64 位的 Ubuntu20.04 操作系统,采用 PyTorch 深度学习框架,硬件配置为 AMD EPYC 7452 处理器和 GeForce RTX 2080ti 显卡。在开始训练前,将行人数量 K 设置为 3,每个行人提取图片数量

M 设置为 12,批次大小 N 设置为 36, l 和 λ 分别设置为 6 和 1.0。在训练阶段,使用 SGD 算法进行优化。

2.3　消融实验

为验证网络模型和所提方法的有效性,在 SYSU-MM01 数据上进行消融实验,实验结果如表 1 所示。

表 1　在 SYSU-MM01 数据集全搜索模式下的消融实验结果

Table 1　Results of ablation experiments in All-search mode on SYSU-MM01 dataset

方法				全搜索模式	
Baseline	SFE	RCT	MJC	Rank-1/%	mAP/%
Π				55.34	53.76
Π	Π			62.79	58.95
Π		Π		63.21	59.81
Π			Π	63.11	58.10
Π	Π	Π		65.08	60.42
Π	Π		Π	64.71	61.19
Π	Π	Π	Π	70.52	64.02

细微特征增强模块:将细微特征增强模块加入到基线中进行试验,实验结果如表 1 所示,其 Rank-1 和 mAP 分别为 62.79%和 58.95%,在基线的基础上分别提高了 7.45%和 5.19%。同时,将特征增强模块与其他模块组合能很大程度提升模型性能,说明细微特征增强模块能够很好地挖掘到行人的细微语义信息,获得更有辨别性的细微特征。

随机颜色转换模块:将随机颜色转换模块加入到基线中进行试验,实验结果如表 1 所示,其 Rank-1 和 mAP 分别为 63.21%和 59.81%,在基线的基础上分别提高了 7.87%和 6.05%。基线、随机颜色转换模块和细微特征增强模块协同工作时,Rank-1 和 mAP 分别为 65.08%和 60.42%,较基线分别提高了 9.74%和 6.66%。这说明随机颜色转换模块能较好地增强模态之间的交互,缩小模态之间的差异,有效地提高模型的精度。

多级联合聚类学习策略:将多级联合聚类学习策略加入到基线中进行试验,实验结果如表 1 所示,其 Rank-1 和 mAP 分别为 63.11%和 58.10%,在基线的基础上分别提高 7.77%和 4.34%。基线、细微特征增强模块和联合聚类学习策略协同工作时,Rank-1 和 mAP 分别为 64.71%和 61.19%,在基线的基础上分别提高了 9.37%和 7.43%。这说明联合聚类学习策略能够有效地进行特征聚合,使模型能更好地辨别各个类别,证明了联合聚类学习的有效性。

综上所述,基线、随机颜色转换模块、细微特征增强模块和多级联合聚类学习策略联合使用时网络模型性能最优,实验结果 Rank-1 和 mAP 分别为 70.52%和 64.02%,较基线分别提升了 15.18%和 10.26%。这说明 FECL 网络够有效地挖掘具有辨别力的细微特征,缩小模态差异,提升模型性能。

2.4　参数分析

在 FECL 网络上进行一系列实验来研究参数 λ 的影响,参数 λ 影响 L_{MJC} 的权重。在 SYSU-MM01 数据集的全搜索模式下,实验结果如图 6 所示。其中 λ 的最佳值是 1.0,随着 λ 的上升和下降 FECL 网络性能逐渐下降。分析其原因是两个模态的共享信息较少, λ 太大会导致模态聚类过程无法将两个模态的距离拉得更近,从而导致过拟合。同时, λ 太小会导致类间分离过程无法将类间距离拉的更远,从而影响网络的判别。为验证细微特征增强模块中局部特征数量 l 的影响,分别将全局特征水平划分为不同 l 的局部特征进行实验,实验结果如图 7 所示。实验发现 l 的最佳值为 6,将全局特征划分为 6 个局部特征时,FECL 网络精度最高。分析原因是局部特征数过多会破坏图像结构,从而造成特征丢失,局部特征数过少导致提取到的特征不够细腻,不足以获得到具有辨别力的特征。

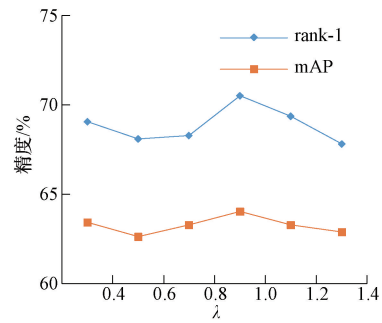

图 6　不同 λ 的实验结果

Fig. 6　Experimental results for different λ

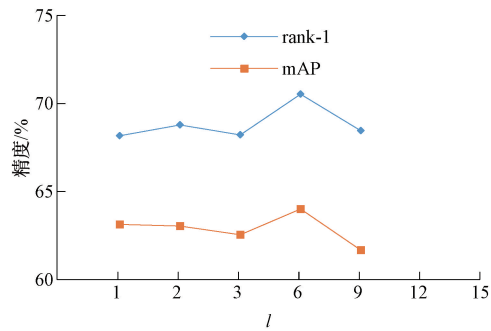

图 7　不同 l 的实验结果

Fig. 7　Experimental results for different l

2.5　可视化分析

通过 T-SNE 将特征进行可视化,进一步分析本文方法的有效性,如图 8 为基线的聚类可视化结果,图 9 为 FECL 网络的聚类可视化结果,其中虚线框内为不同标签的特征,同一虚线框内的三角形和圆形分别表示同一标

签下的可见光模态特征和红外模态特征。与基线相比，FECL网络能够将相同标签的特征更加清晰地区分和聚合在一起，不同类之间的界限更加明显，验证了多级联合聚类学习策略的有效性。由于颜色偏差是引起模态间差异的直接原因，不同模态特征之间的距离直接反映了网络受颜色偏差的影响程度。通过对比同一标签下的可见光模态特征和红外模态特征之间的距离，FECL网络的两种模态特征之间的距离比基线网络更小，可以说明FECL网络中的随机颜色转换模块可以有效缓解颜色偏差的影响，减小模态差异。

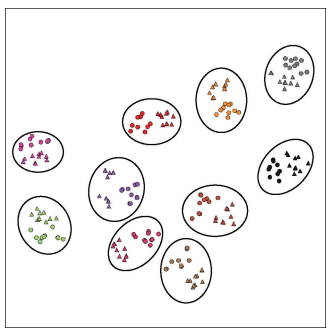

图8　baseline特征空间中的聚类可视化结果

Fig. 8　Clustering visualization results in the feature space of baseline

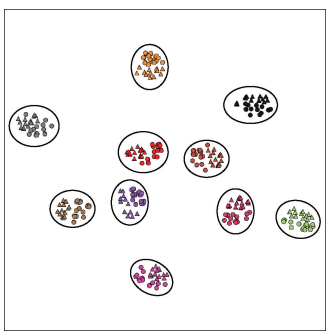

图9　FECL网络特征空间中的聚类可视化结果

Fig. 9　Clustering visualization results in the feature space of FECL networks

为验证FECL网络具有挖掘细腻特征的能力，通过Grad-Cam[17]获得特征热力图。如图10所示为基线的可视化结果，图11所示为FECL网络的可视化结果，基线网络仅有身体部分区域被绘制为高亮，而FECL网络中的行人眼、鞋子以及身体关节等细节部位被绘制为高亮，这些通常是容易被网络所忽视的判别性特征。通过与基线网络对比，可以说明FECL网络比基线所提取的特征更加细腻，能够较准确的挖掘行人具有判别性的细微特征。

在测试阶段，通过将SYSU-MM01数据集上的检索结果可视化，验证FECL的检索能力。在query集中取一张红外图像，在gallery集中的可见光图像中进行匹配。基线检索结果如图12所示，FECL网络检索结果如图13所示，其中实线框表示正确匹配，虚线框表示错误匹配。与基线相比，FECL网络具有较高的匹配精度，检索能力较强。

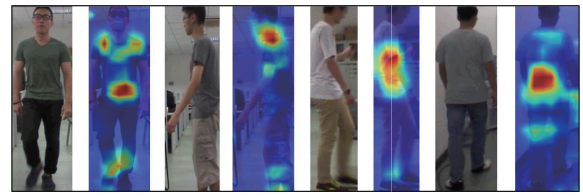

图10　baseline热力图可视化结果

Fig. 10　Visualization result of heat map of baseline

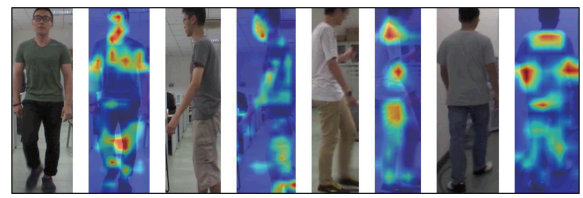

图11　FECL网络热力图可视化结果

Fig. 11　Visualization result of heat map of FECL network

图12　baseline的top-10检索结果

Fig. 12　Top-10 retrieval results of baseline

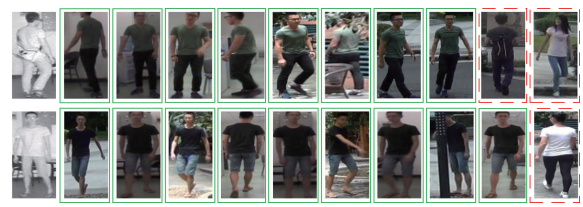

图13　FECL网络的top-10检索结果

Fig. 13　Top-10 retrieval results of FECL network

2.6　与其他方法的对比

为证明FECL网络的先进性，将FECL网络与多个较为先进的跨模态行人重识别方法在公共数据集上进行对比，其中包括Zero-Pad[3]、HCML[18]、cmGAN[5]、D2RL[19]、Hi-CMD[20]、JSIA[21]、MACE[22]、AGW[1]、CM-NAS[23]、MCLNet[7]、DML[6]、MAUMG[24]。在SYSU-

MM01 数据集上的实验结果如表 2 所示,在全搜索模式下的 Rank-1 为 70.52%、mAP 为 64.02%,在室内搜索模式下的 Rank-1 为 74.23%、mAP 为 76.61%。与 MCLNet 等主要研究在全局特征或局部特征上提取模态共享特征的方法相比,FECL 网络致力于在通道层面消除模态间差异。与 cmGAN 等生成对抗性学习的方法相比,FECL 网络不需要生成新数据,耗费资源较少。与 DDAG 相同,FECL 网络同样使用全局特征和局部特征相结合,与之不同的是 FECL 网络注重挖掘并增强局部特征中的细微特征,DDAG 中使用的模态内局部加权聚合方法使用了自注意力机制,其算法复杂度为 $O(n^2 \cdot d)$,FECL 网络中的细微特征增强模块的复杂度为 $O(n \cdot d)$,故 FECL 网络的计算复杂度更低,计算开销更少。在 RegDB 数据集上的实验结果如表 3 所示,FECL 网络在 Visible-to-Infrared 检索模式下有较好的表现,Rank-1 和 mAP 分别为 88.88%和 80.93%。

表 2　在 SYSU-MM01 数据集上的对比结果

Table 2　Comparison results on the SYSU-MM01 dataset

方法	期刊	全搜索模式				室内搜索模式			
		$r=1$/%	$r=10$/%	$r=20$/%	mAP/%	$r=1$/%	$r=10$/%	$r=20$/%	mAP/%
Zero-Pad[3]	ICCV17	14.8	54.12	71.33	15.95	20.58	68.38	85.79	26.92
HCML[18]	AAAI18	14.32	53.16	69.17	16.16	24.52	73.25	86.73	30.08
cmGAN[5]	IJCAI18	26.97	67.51	80.56	27.80	31.63	77.23	89.18	42.19
D2RL[19]	CVPR19	28.9	70.6	82.4	29.2	–	–	–	–
Hi-CMD[20]	CVPR20	34.94	77.58	–	35.94	–	–	–	–
JSIA[21]	AAAI20	38.10	80.70	89.90	36.90	43.80	86.20	94.20	52.90
MACE[22]	TIP20	51.64	87.25	94.44	50.11	57.35	93.02	97.47	64.79
AGW[1]	TPAMI21	47.50	84.39	92.14	47.65	54.17	91.14	95.98	62.97
CM-NAS[23]	CVPR21	61.99	92.87	97.25	60.02	67.01	97.02	99.32	72.95
MCLNet[7]	ICCV21	65.40	93.33	97.14	61.98	72.56	96.88	99.20	76.58
DML[6]	TCSVT22	58.40	91.20	95.80	56.10	62.40	95.20	98.70	69.50
MAUMG[24]	CVPR22	61.59	–	–	59.96	67.07	–	–	73.58
FECL	–	**70.52**	**94.58**	**97.29**	**64.02**	**74.23**	**95.61**	**98.41**	**76.61**

表 3　在 RegDB 数据集上的对比结果

Table 3　Comparison results on the RegDB dataset

方法	期刊	Visible to Infrared			
		$r=1$/%	$r=10$/%	$r=20$/%	mAP/%
Zero-Pad[3]	ICCV17	17.75	34.21	44.35	18.90
HCML[18]	AAAI18	24.44	47.53	56.78	20.08
D2RL[19]	CVPR19	43.4	66.1	76.3	44.1
Hi-CMD[20]	CVPR20	34.94	77.58	–	35.94
JSIA[21]	AAAI20	48.10	–	–	48.90
MACE[22]	TIP20	72.37	–	–	69.09
AGW[1]	TPAMI21	70.05	86.21	91.55	66.37
CM-NAS[23]	CVPR21	84.54	95.18	97.85	80.32
MCLNet[7]	ICCV21	80.31	92.70	96.03	73.07
DML[6]	TCSVT22	77.60	–	–	84.30
MAUMG[24]	CVPR22	83.39	–	–	78.75
FECL	–	**88.88**	**97.48**	**98.64**	**80.93**

3　结　论

本文针对目前跨模态行人重识别存在的问题,提出了 FECL 网络,FECL 网络致力于增强细微特征,消除模态差异。首先提出细微特征增强模块,增强行人眼镜、鞋子和衣服等细微特征,以便网络提取到更加细腻的判别特征;其次提出随机颜色转换模块,在输入端加强两个模

态之间的交互,缓解颜色偏差的影响;最后设计了一种多级联合聚类学习策略,在通道层面逐步优化特征分布,最小化通道间和模态间差异的同时,最大化类间距离。在SYSU-MM01和RegDB数据集上进行实验,证明了所提方法的先进性和有效性。下一步工作将加强对模态特定特征的提取,充分利用各模态内部的有用信息,进一步提升模型性能。

参考文献

[1] YE M, SHEN J, LIN G, et al. Deep learning for person re-identification: A survey and outlook [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(6): 2872-2893.

[2] SUN Y, ZHENG L, YANG Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline) [C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 480-496.

[3] WU A, ZHENG W S, YU H X, et al. RGB-infrared cross-modality person re-identification[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 5380-5389.

[4] HAO Y, WANG N, LI J, et al. HSME: Hypersphere manifold embedding for visible thermal person re-identification[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 8385-8392.

[5] DAI P, JI R, WANG H, et al. Cross-modality person re-identification with generative adversarial training[C]. International Joint Conference on Artificial Intelligence, 2018, 1(3): 677-683.

[6] ZHANG D, ZHANG Z, JU Y, et al. Dual mutual learning for cross-modality person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(8): 5361-5373.

[7] HAO X, ZHAO S, YE M, et al. Cross-modality person re-identification via modality confusion and center aggregation [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 16403-16412.

[8] ZHONG Z, ZHENG L, KANG G, et al. Random erasing data augmentation [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 13001-13008.

[9] YE M, RUAN W, DU B, et al. Channel augmented joint learning for visible-infrared recognition[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 13567-13576.

[10] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[11] 周航,黄春光,程海. 基于全局多粒度池化的可见光红外行人重识别[J]. 电子测量技术, 2022, 45(1): 122-128.
ZHOU H, HUANG CH G, CHENG H. Visible infrared pedestrian re-identification based on global multi-granularity pooling [J]. Electronic Measurement Technology, 2022, 45(1): 122-128.

[12] 张勃兴,马敬奇,张寿明,等. 利用全局与局部关联特征的行人重识别方法[J]. 电子测量与仪器学报, 2022,36(6):205-212.
ZHANG B X, MA J Q, ZHANG SH M, et al. Pedestrian re-identification method using global and local correlation features[J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(6): 205-212.

[13] 钱亚萍,王凤随,熊磊. 基于局部细化多分支与全局特征共享的无监督行人重识别方法[J]. 电子测量与仪器学报,2023,37(1):106-115.
QIAN Y P, WANG F S, XIONG L. Unsupervised pedestrian re-identification method based on local refinement of multiple branches and global feature sharing [J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(1): 106-115.

[14] LUO H, GU Y, LIAO X, et al. Bag of tricks and a strong baseline for deep person re-identification [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019: 1487-1495.

[15] ZHANG Y, KANG Y, ZHAO S, et al. Dual-semantic consistency learning for visible-infrared person re-identification [J]. IEEE Transactions on Information Forensics and Security, 2022, 18: 1554-1565.

[16] 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): 605-605.

[17] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 618-626.

[18] YE M, LAN X, LI J, et al. Hierarchical discriminative learning for visible thermal person re-identification[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 7501-7508.

[19] WANG Z, WANG Z, ZHENG Y, et al. Learning to

reduce dual-level discrepancy for infrared-visible person re-identification [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 618-626.

[20] CHOI S, LEE S, KIM Y, et al. Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10257-10266.

[21] WANG G A, ZHANG T, YANG Y, et al. Cross-modality paired-images generation for RGB-infrared person re-identification [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12144-12151.

[22] YE M, LAN X, LENG Q, et al. Cross-modality person re-identification via modality-aware collaborative ensemble learning [J]. IEEE Transactions on Image Processing, 2020, 29: 9387-9399.

[23] FU C, HU Y, WU X, et al. CM-NAS: Cross-modality neural architecture search for visible-infrared person re-identification [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 11823-11832.

[24] LIU J, SUN Y, ZHU F, et al. Learning memory-augmented unidirectional metrics for cross-modality person re-identification [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 19366-19375.

作者简介

范馨月，2002 年于四川师范大学获得学士学位，2005 年于电子科技大学获得硕士学位，现为重庆邮电大学副教授，主要研究方向为计算机视觉、网络信息安全。

E-mail: fanxy@ cqupt. edu. cn

Fan Xinyue received her B. Sc. degree from Sichuan Normal University in 2002, and her M. Sc. degree from the University of Electronic Science and Technology of China in 2005. Now she is an associate professor at Chongqing University of Posts and Telecommunications. Her main research interests include computer vision and network information security.

张阔（通信作者），2021 年于成都信息工程大学获得学士学位，现为重庆邮电大学硕士研究生，主要研究方向为计算机视觉与图像识别。

E-mail: s210101189@ stu. cqupt. edu. cn

Zhang Kuo (Corresponding author) received his B. Sc. degree from Chengdu University of Information Technology in 2021. Now he is a M. Sc. Candidate in Chongqing University of Posts and Telecommunications. His main research interests include computer vision and image recognition.