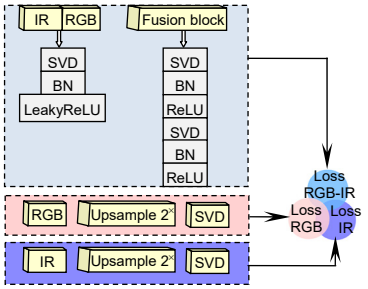
Opto-Electronic Engineering
光 电 工 程

Article

2020 年，第 47 卷，第 12 期

DOI: 10.12086/oee.2020.190669

# 基于红外和可见光模态的随机融合特征金子塔行人重识别




汪荣贵，王　静，杨　娟*，薛丽霞

合肥工业大学计算机与信息学院，安徽 合肥 230009

**摘要：**目前行人重识别的研究只关注了可见光下跨摄像头提取图像不变的特征表示，忽视了红外条件下的成像特点，并结合两种模态的研究成果很少。此外，当前行人重识别在判别两个图像时，通常是计算单个卷积层特征图的相似性，这会导致弱特征学习现象。为了解决上述问题，本文提出了基于特征金字塔的随机融合网络，它可以同时计算多个特征层级的相似性，匹配图像时是基于多个语义层的判别因子。该模型关注到红外图像的特性，并且缩小了可见光和红外模态内部负作用的偏差，平衡了模态间的异质差距，综合了局部特征和全局特征学习的优势，有效地解决了跨模态行人重识别问题。实验在 SYSU-MM01 数据集上对平均精确度和收敛速度进行验证。结果表明，所提的模型优于现有的先进算法，特征金字塔随机融合网络实现了快速收敛且平均精确度达到了 32.12%。

**关键词：**行人重识别；可见光；红外；特征金字塔

**中图分类号：**TP391.4　　**文献标志码：**A

**引用格式：**汪荣贵，王静，杨娟，等. 基于红外和可见光模态的随机融合特征金子塔行人重识别[J]. 光电工程，2020，**47**(12): 190669

# Feature pyramid random fusion network for visible-infrared modality person re-identification

Wang Ronggui, Wang Jing, Yang Juan*, Xue Lixia

School of Computer & Information, Hefei University of Technology, Hefei, Anhui 230009, China

**Abstract:** Existing works in person re-identification only considers extracting invariant feature representations from cross-view visible cameras, which ignores the imaging feature in infrared domain, such that there are few studies on visible-infrared relevant modality. Besides, most works distinguish two-views by often computing the similarity in feature maps from one single convolutional layer, which causes a weak performance of learning features. To handle the above problems, we design a feature pyramid random fusion network (FPRnet) that learns discriminative multiple semantic features by computing the similarities between multi-level convolutions when matching the person. FPRnet not only reduces the negative effect of bias in intra-modality, but also balances the heterogeneity gap between inter-modality, which focuses on an infrared image with very different visual properties. Meanwhile, our work integrates the advantages of learning local and global feature, which effectively solves the problems of visible-infrared person re-identification. Extensive experiments on the public SYSU-MM01 dataset from aspects of mAP

收稿日期：2019-11-02； 收到修改稿日期：2020-04-10

作者简介：汪荣贵(1966-)，男，博士，教授，博士生导师，主要从事智能视频处理与分析、视频大数据与云计算、智能视频监控与公共安全、嵌入式多媒体技术等领域的研究。E-mail：wangrgui@hfut.edu.cn

通信作者：杨娟(1983-)，女，博士，讲师，硕士生导师，主要从事视频信息处理、视频大数据处理技术、深度学习与二进神经网络理论与应用等的研究。E-mail：yangjuan6985@163.com

and convergence speed, demonstrate the superiorities in our approach to the state-of-the-art methods. Furthermore, FPRnet also achieves competitive results with 32.12% mAP recognition rate and much faster convergence.

**Keywords:** person re-identification; visible; infrared; feature pyramid

**Citation:** Wang R G, Wang J, Yang J, *et al.* Feature pyramid random fusion network for visible-infrared modality person re-identification[J]. *Opto-Electronic Engineering*, 2020, **47**(12): 190669

## 1 引 言

传统可见光的监控摄像头已经不适应于一般场合，例如道路交通、监狱等场景都需要 24 小时监控。目前很多监控系统和智能设备配置了可见光到红外模式自动切换功能[1]，基于可见光模式的行人重识别在 Market1501 数据集[2]下 rank-1 已经达到了 95%[3]，而基于红外(IR)模式的行人重识别，由于红外图像低分辨率和低对比度等问题，使得目前研究成果很少[4-6]。

跨模态行人重识别研究面临着来自 RGB 和 IR 模态内部以及 RGB-IR 模态间三方面挑战。模态内的图像存在着视觉外观差异，模态间图像存在着异质差距，比如在分辨率和对比度方面，两种模态的图像是未对齐的。而且，可见光图像包含三种通道的颜色信息，红外图像本质是单通道的灰度图。这种不一致的信息分布，会扰乱神经网络特征学习过程。因而，跨模态行人重识别研究重点是减缓三种差异的消极影响。最近，Wu 等人[4]提出了 SYSU-MM01 数据集，它囊括了可见光和红外两种模式。作者提出的单流网络，在融合模态间差距方面取得了初步成效。基于前人 Wu 的研究并利用生成对抗网络，Dai 等人[5]提出了跨模态融合 CmGAN 方法，它能显著提升跨模态行人重识别领域的实验效果。然而，单流网络准确率低，CmGAN 网络的高参数量导致收敛速度慢，同时两者均忽视了红外和可见光不同的成像特点，会造成网络学习的特征量不对称。

针对上述跨模态问题，本文提出了基于特征金字塔的随机融合网络。首先，引入超分辨重建方法，增强图像对比度，提升其分辨率，进而将行人 RGB 图拉至清晰水平，以此缩小模态内差异，用可见光模态在训练中的优势，去降低红外图像由于视觉糊化效应造成的训练困扰。其次，利用网络底层和高层分别学习图像的局部和全局的特性，来构建特征金子塔，可实现单模态不同网络层的特征学习。这种多级特征学习方法，结合了全局特征和局部特征学习的优势，既关注了整体表现，又保留了细节特性，设计的特征金字塔产生了跨域高层联合特征，这种跨模态的融合特征在实际训练时减缓了异质差距的负影响，使得网络学习的特征量对称且稳定。然后，通过随机机制讨论了不同网络层级特征与模态间的组合关系，得到了随机融合金字塔的优化组合方式，相应地产生了随机融合特征。接下来，将随机融合特征与高层联合特征相互博弈，以实现基于多级多语义判别因子的联合模态和单模态的特征相似性识别。最后，提出了跨模态混合损失函数，它通过显性化模态的差异并增加其权重，再利用联合模态与单模态间偏置将 RGB-RGB、IR-IR、RGB-IR 三种模态的损失函数融合，有效地修正了网络模型，使得跨模态的重识别学习具有综合性。

## 2 相关工作

神经网络的最大优势是学习图像的深层语义信息，在复杂场景下具有较好的识别效果。然而，这种高重识别率仅限于白天场景，现有的可见光行人重识别不能很好地解决夜晚情况下特定人的红外图像识别，对此本文将融合 RGB 与 IR 两种模态图像来锁定目标，接下来将从两方面进行分析。

### 2.1 行人重识别

行人重识别主要任务是提取合适特征，使得同一行人不同场景中提取到的特征是相似的，而不同人的图像应提取到不同特征。早期的特征提取，专注于图像的颜色纹理信息，包括局部特征集成[7]和形状外观建模[8]等手工特征。近年来，深度学习方法开始用于行人重识别领域，并取得了比传统方法更好的效果。Li[9]等人提出了一种过滤配对网，这是第一个利用深度学习方法处理重识别中特征提取问题。手动标记大量图像很昂贵，使得用于深度学习训练的数据(即真实数据)量不足。为了解决这个问题，生成性对抗网络(GAN)[10]被用于生成人工样本数据。但是 GAN 网络的超参数多，时间成本高，不适合少量 GPU 训练。针对设备性能不高，图像场景下要求较精准的识别时，可采用 ResNet[2]，其较小的网络尺寸，使得网络优化更为简单，可实现与 GAN 网络等同的识别性能。因此，本文将 ResNet 作为基本特征提取器。

现有的行人重识别方法主要基于单一尺度的外观信息，它忽略了其他不同尺度潜在的有用信息。针对上述问题，Liu 等人[11]提出了多尺度三元组神经网络 MSTriCNN，它通过组合子网络将不同分辨率以及不同尺度的特征融合。Qian 等人[12]提出了多尺度深度模型 MuDeep，它实现了 MSTriCNN 不具备的自动加权。Chen 等人[13]提出的深层特征金字塔网络，它抛弃了 MuDeep 不同尺度匹配不同行人的做法，而是将多尺度特征构造成金字塔，来解决跨尺度学习问题。本文将不同尺度特征融合表征多种语义的思想，运用到卷积层面，就是将 ResNet 网络中不同卷积层级的特征融合，以此实现图像多种语义信息的单一输出，这种方式融合了高层特征强语义表征能力与低层特征强几何细节信息表征能力，从而形成了高分辨强语义的特征。

考虑到图像清晰度会影响特征提取效果，可采用多尺度联合学习框架[14]和尺度距离函数[15]，去处理损失部分信息的低分辨图像。也可利用低分辨(LR)图像和高分辨(HR)图像字典对[16]，将 LR 转化成 HR。本文运用超分辨重建策略将 LR 图像提升至 HR 状态，在某种程度上能够弥补低分辨条件下不同类型的视觉特征损失。此外，为了将同一子空间相同标签的行人图像推得更近，将不同标签的行人图像拉开得更远，需要比较图像的特征距离，在跨模态行人重识别中，常见的度量特征间距离的方法，有欧氏距离[4]、SCM[17]和 CRAFT[18]，本文则采用欧氏度量法学习区分样本。

### 2.2 跨模态检索

跨模态检索的对象是不同模态数据，包括可见光图像、红外图像、文本信息和视频。传统跨模态学习常采用哈希法，Zhu 等人[19]提出了缩放索引哈希法，局部自适应哈希法也被提出[20]。再引入语义信息后，监督学习的哈希法[17]出现了。最近几年，深度学习被广泛应用在跨模态学习领域，基于玻尔兹曼机的深度学习[21]在多模态问题上取得了出色效果。

如今，多模态问题的研究在人体识别[22]和面部识别[23]以及行人重识别[24]方面均有成果。特别是跨模态行人重识别[25-26]，它在机场、车站等全天候人流密集的场所，具有广泛应用前景。有研究者通过深度相机将红外图像和可见光图像这种跨模态信息域[27]引入到行人重识别方向。但目前跨模态行人重识别的研究成果很少，导致这种现象的核心原因是模态间异质差距。针对这一原因，最小化模态内语义模糊的标签，最大化模态间语义相似标签的思想被提出，基于该思想的 CmGAN[5]方法被用来判别两种模式图像的匹配度。相对于 CmGAN 结构，端到端的 dual-path 网络[6]结构更为简单。为了融合不同模态的特征，dual-path 网络提出了 top-ranking 方法。该方法在 SYSU-MM01 数据集上与单流网络的 zero-padding[4]方法对比，数据显示该方法效果要更优。上述 dual-path 网络的实验效果虽然略高于单流网络，但明显低于 CmGAN 模型。然而，CmGAN 网络收敛速度慢，单流网络识别精度低。因此，本文希望在保证良好识别率的同时有不错的收敛性能，为此，注意到了跨模态的特性与图像成像特点。本文通过构造特征金子塔将底层细节特征与高层语义特征融合，不仅减小了模型尺寸保障收敛速度，而且降低了不同图像成像差异的消极影响，再由随机学习，提升了跨模态重识别的精度。

## 3 方 法

### 3.1 网络模型

本文提出的特征金字塔随机融合网络结构，可从红外场景和可见光场景中抽取和融合行人特征。前者特征抽取器，抽取了从高级到低级不同层级两种模态的图像特征。后者融合模块，将抽取到的不同层级的特征不区分模态地随机融合。

#### 3.1.1 特征抽取器

本文以 ImageNet 上预训练的 ResNet-50 为基本框架，学习 RGB 模态和 IR 模态下的图像特征，记 r 表示 RGB 模态，i 表示 IR 模态，生成的特征图用 $C$ 表示。标记 Resnet-50 卷积层生成的不同层级特征，第 $k$ 层为 $C(k)(C(k)\in C)$，且不同模态的特征图分别用 $C_r(k)$与 $C_i(k)$ 表示，即有 $C(k)=C_i(k)\bigcup C_r(k)$ 。输入的图像是预处理后的，通过预处理，平衡了两种模态间异质差距造成的不对称的特征学习过程。输入的图像尺寸为 288×144，它经过 2 步长和 7×7 大小的卷积核，生成二维特征图矩阵，其中行分量为 72，列分量为 36，且通道数是 64，具体的每层卷积参数设置如表 1 所示，5 层卷积后生成了 2048 通道数的特征图 $C(5)$，$C(5)=\{C_i(5),C_r(5)\}$ 。$C(5)$再经平均池化，形成包含两种模态大小均为 1×1 的特征，并分别用 RGB block 和 IR block 标记。然后，将特征 RGB block 和 IR block 串联的结果，作为高层联合特征 IR-RGB block。

特征 $C(5)$用 1×1 大小的卷积核$\omega(1)$来降维，以形成通道数为 256 的特征 $P'(5)$，如图 1 所示，其中，$P'(k)$是特征图 $C(k)$卷积系列的表示。上采样可使图像具备

表 1 基本网络结构

Table 1 Basic network structure

| Layer | Input size | Output size | Structure |
|---|---|---|---|
| *C*(1) | 288×144 | 72×36 | 7×7,64,stride=2 |
| *C*(2) | 72×36 | 72×36 | 3×3 max pool, stride=2<br>$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$ |
| *C*(3) | 72×36 | 36×18 | $\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$ |
| *C*(4) | 36×18 | 18×9 | $\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$ |
| *C*(5) | 18×9 | 9×5 | $\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$ |
| RGB block and IR block | 9×5 | 1×1 | Average pool |

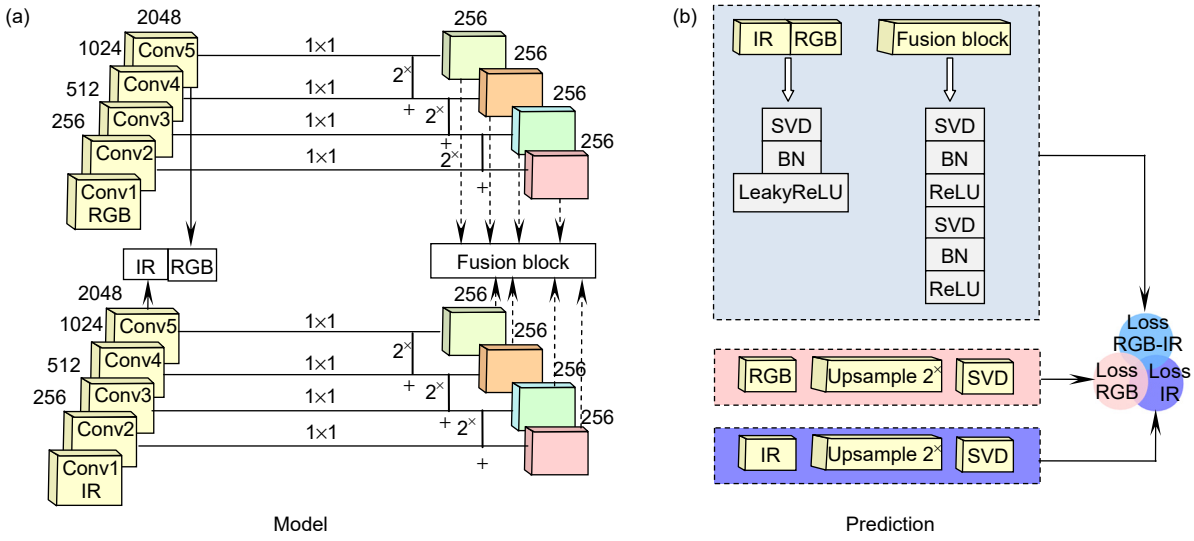



图 1 特征金字塔随机融合网络结构示意图。(a) 网络模型包括高层联合特征 IR-RGB block 与随机融合特征 fusion block，具体来说，IR-RGB block 由 RGB block 和 IR block 串联形成，fusion block 由不同层级不同模态的特征随机融合生成；(b) 跨模态预测是由蓝色的联合模态块、粉色的 RGB 单模态块以及紫色 IR 单模态块组成，并产生三种模态的分类损失，其中，联合模态分类损失由 IR-RGB block 与 fusion block 博弈产生

Fig. 1 An illustration of the framework of feature pyramid random fusion network. (a) The model generates a top-level joint feature (termed IR-RGB block) and a random fusion feature (termed fusion block). Specifically, the IR-RGB block is concatenated by RGB block with IR block; the fusion block is generated by randomly blend features from different levels and distinct modalities; (b) The prediction consists of a blue cross-domain, a pink RGB-domain, and a purple IR-domain, which generates three types of classification loss. The IR-RGB block and fusion block use a minimax game to beat each other for learning the joint-modal classification loss

高分辨率的特质，本文将双线性插值上采样后的 $P'(5)$ 与 1×1 卷积后的特征图 $C(4)$，两者按元素相加，之后再用 3×3 的卷积核 $\omega(2)$来消除上采样引起的混叠效应，形成了特征 $P'(4)$，如表 2 所示。

第一步：获取特征 $C(k)$通过 1×1 大小卷积核 $\omega(1)$ 的结果。本文注意到 $C(1)$过大的参数量，因此，只对特征金字塔 $C=\{C(5),C(4),C(3),C(2)\}$ 做上述操作，$C(1)$ 不参与下面步骤。

第二步：获取特征 $P'(k+1)$上采样的结果。

第三步：将前两步的两个结果进行侧向连接；由于低层特征图忽略了图像的全局特性，高层特征图细节信息不够丰富，借鉴侧向连接方法的思想，本文综合了低层网络的丰富细节和高层特征的全局特性，获取了不同尺度特征图的优势。

第四步：经过 1 步长，3×3 大小卷积核 $\omega(2)$，生成了通道数均为 256 的特征 $P'(2)$、$P'(3)$、$P'(4)$。

第五步：平均池化，以形成 1×1 大小的二维特征矩阵 $\boldsymbol{P}(k)$。上述步骤正如式(1)所示，⊙代表卷积操作，$2^{\times}$ 表示上采样过程。

$$\boldsymbol{P}(k)=P_{\text{avgpool}}(\omega(2)\odot(C(k)\odot\omega(1)+2^{\times}(P'(k+1)))) \quad ,(1)$$

其中 $\boldsymbol{P}(k)\subseteq\boldsymbol{P}_{\text{i}}(k)\bigcup\boldsymbol{P}_{\text{r}}(k)$。若用字母 $P$ 表示上述构建的特征金字塔，那么有：

$P=\{P_{\text{i}}(2),P_{\text{i}}(3),P_{\text{i}}(4),P_{\text{i}}(5)\}\bigcup\{P_{\text{r}}(2),P_{\text{r}}(3),P_{\text{r}}(4),P_{\text{r}}(5)\}$，再由特征融合模块对 $P$ 进行相应操作，生成随机融合特征 fusion block。接下来，本文将详细介绍特征融合模块。

### 3.1.2 特征融合

特征金子塔 $P$ 包含两种模态四种层级的特征图。考虑到网络负载成本，故而从 $P$ 中只抽取一组融合特征 fusion block，作为模型输出。抽取的融合特征直接影响分类器学习效果，所以需要讨论特征选取和融合方式。前者特征选取，需要考虑模态和层级间的关系，存在三种组合情况，如图 2 所示。第一种，抽取同一网络层级不同模态的特征，比如$(P_{\text{r}}(5)，P_{\text{i}}(5))$，$(P_{\text{r}}(4)，P_{\text{i}}(4))$组合。第二种，抽取不同尺度同一模态的特征，例如$(P_{\text{r}}(5)，P_{\text{r}}(2))$，$(P_{\text{i}}(2)，P_{\text{i}}(2))$。第三种，抽取不同层级跨模态的特征，如$(P_{\text{r}}(4)，P_{\text{i}}(4)，P_{\text{r}}(5)，P_{\text{i}}(5))$。后者特征融合，考虑的是特征间的串联形式，如图 3 所示，也存在三种组合情况。第一种，特征间彼此横向串联，如$(P_{\text{r}}(5)，P_{\text{i}}(5)，P_{\text{r}}(4)，P_{\text{i}}(4))$，第二种，特征间彼此纵向串联，比如$(P_{\text{r}}(5)，P_{\text{i}}(5)，P_{\text{r}}(4)，P_{\text{i}}(4))^{\text{T}}$，第三种，特征间以横向串联和纵向串联混合形式融合，例如

$$\begin{pmatrix}P_{\text{r}}(5) & P_{\text{r}}(4)\\ P_{\text{i}}(5) & P_{\text{i}}(4)\end{pmatrix}。$$

此时，特征图的尺度必须被考虑到。因此，特征间混合串联可细分为以下三种情况：组合 1，不同层级特征横向串联时，纵向串联的特征也是不同尺度的；组合 2，当同一层级特征横向串联时，不同层级特征采用纵向串联方式；组合 3，不同层级特征横向串联时，纵向串联的特征是同一尺度的。通过随机选取融合方式和随机抽取特征，再对比多组实验，可以发现不同尺度潜在的有用信息，有利于找到模态与网络层级间的关联，进而解决多模态跨尺度学习问题。

## 3.2 跨模态预测

由于数据集 SYSU-MM01 样本图像非均匀分布，使得全连接层的每一个权重向量都高度相关，直接影响了行人重识别的性能。为了解决这个问题，使用奇

表 2 构建特征金字塔

Table 2 Building feature pyramid

| Layer | $C(5)$ | $C(4)$ | $C(3)$ | $C(2)$ |
|---|---|---|---|---|
| Step1 | | 1×1,256, stride=1 | | |
| Step2 | | $P'(5)\uparrow$ | $P'(4)\uparrow$ | $P'(3)\uparrow$ |
| Step3 | | + | + | + |
| Step4 | | 3×3,256, stride=1 | | |
| Output size | 9×5 | 18×9 | 36×18 | 72×36 |
| Hidden layer | $P'(5)$ | $P'(4)$ | $P'(3)$ | $P'(2)$ |
| Step5 | | Average pool | | |
| Output size | 1×1 | 1×1 | 1×1 | 1×1 |
| Result layer | $P(5)$ | $P(4)$ | $P(3)$ | $P(2)$ |

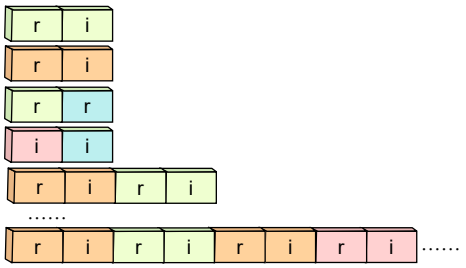



图 2 特征选取方式。
r、i 分别表示 RGB、IR 模态，特征 *P*(5), *P*(4), *P*(3), *P*(2)分别用绿色、橙色、蓝色、粉色表示

Fig. 2 Feature selection. r and i represent RGB and IR domain respectively. Features *P*(5), *P*(4), *P*(3), and *P*(2) are shown by green, orange, blue and pink respectively

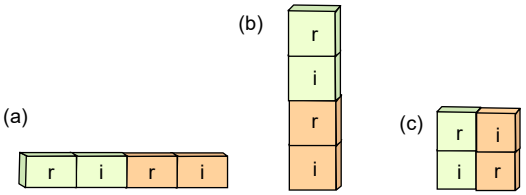



图 3 特征融合方式。(a) 横向串联; (b) 纵向串联; (c) 混合串联。r、i 分别表示 rgb、IR 模态

Fig. 3 The method of feature fusion. (a) Horizontal concatenation; (b) Vertical concatenation; (c) Hybrid concatenation. Let r be the RGB-modality, and i be the IR-modality

异矩阵分解[28]的全连接层，以降低特征向量间的相关性，并将该层记作 SVD。如图 1 所示，高层联合特征 IR-RGB block 和随机融合特征 fusion block 分别输入到 SVD 层，再将 leakyReLU[29]用于联合特征的训练，被 ReLU[30]激活的 fusion block 再次输入 SVD 层(见图 1(b)上方)。最后，将 IR-RGB block 与 fusion block 互相博弈，把相似性排名最高的作为联合模态特征 $U$，再经过交叉熵分类器输出联合模态的损失 $L_{\text{r-i}}$。

影响跨模态行人重识别性能的核心因素是：模态内视觉外观差异与模态间异质差距。本文认为将这种差距显性化，再增强其权重，作用于网络模型，是能够弥合数据间不同分布引起的差异变化，以修正原有模型偏差的。具体是通过双线性插值法将 RGB block 和 IR block 的分辨率放大 2 倍后，输入到 SVD 层，再经过交叉熵分类输出两种模态的损失 $L_{\text{r}}$ 与 $L_{\text{i}}$(见图 1(b)下方)，并结合联合模态损失，得到混合损失函数 $L$，如下：

$$L=\lambda_1 L_{\text{r-i}}+\lambda_2 L_{\text{r}}+\lambda_3 L_{\text{i}} \ , \tag{2}$$

其中：$\lambda_1$，$\lambda_2$ 和$\lambda_3$ 是预定义的权重参数，用于网络训练时微调，以提升重识别精度；$L_{\text{r-i}}$、$L_{\text{r}}$ 与 $L_{\text{i}}$ 具体形式如下：

$$L_{\text{r-i}}(U,U_{\text{t}})=-\log\frac{\mathrm{e}^{U[U_{\text{t}}]}}{\sum_{j=0}^{M}\mathrm{e}^{U[j]}} \ , \tag{3}$$

$$L_{\text{r}}(B_{\text{R}},U_{\text{r}})=-\log\frac{\mathrm{e}^{B_{\text{R}}[U_{\text{r}}]}}{\sum_{j=0}^{M}\mathrm{e}^{B_{\text{R}}[j]}} \ , \tag{4}$$

$$L_{\text{i}}(B_{\text{I}},U_{\text{i}})=-\log\frac{\mathrm{e}^{B_{\text{I}}[U_{\text{i}}]}}{\sum_{j=0}^{M}\mathrm{e}^{B_{\text{I}}[j]}} \ , \tag{5}$$

其中：$U_{\text{t}}$ 是联合模态特征 $U$ 的真实标签；$U_{\text{r}}$ 是特征 RGB block(简写为 $B_{\text{R}}$)的真实标签；$U_{\text{i}}$ 是特征 IR block(简写为 $B_{\text{I}}$)的真实标签；$M$ 是训练集真实标签的数目，其值为 395。

无法忽视的一个问题是，联合模态和单模态间也存在着异质差距，造成这种差距的原因是网络模型。通过调整联合模态与单模态间偏置，进一步修正网络模型。本文将提取到的偏置记作 $d$，如式(6)所示，并以对数形式来引导网络学习，见式(7)：

$$d=\frac{L_{\text{r}}+L_{\text{i}}}{2}-L_{\text{r-i}} \ , \tag{6}$$

$$L_{\text{c}}=\mathrm{e}^{|d|} \ , \tag{7}$$

其中：$L_{\text{c}}$ 表示修正损失。因此式(2)可改写成如下：

$$L'=\lambda_1 L_{\text{r-i}}+\lambda_2 L_{\text{r}}+\lambda_3 L_{\text{i}}+L_{\text{c}} \ 。 \tag{8}$$

交叉熵损失相对于目前重识别中流行的三元组或四元组损失，是属于小复杂度训练。因此，本文采用改进交叉熵的混合损失函数 $L'$训练模型，在收敛性能方面更具优势。

## 4 实 验

### 4.1 数据集

SYSU-MM01 是跨模态(RGB-IR)行人重识别领域第一个标准数据集，它由 6 个摄像头(4 个 RGB 和 2 个 IR)组成。该数据集包含 491 个人，总共 287628 张 RGB 图像和 15792 张 IR 图像。每个人至少被两个不同位置和视角的摄像头捕获，本文采用 single-shot all-search[5]评估协议。训练集由 395 人组成，图像包括 22258 张 RGB 图像和 11909 张 IR 图像。测试集包含 96 人，其中 3803 个 IR 图像作为 query set，301 个随

机选择的 RGB 图像作为 gallery set。

### 4.2 实验细节

累积匹配特性的精度表示 rank1、rank5、rank10 和 rank20，它们用来评估识别性能。对于 SYSU-MM01 数据集，平均精度(mAP)也作为评估性能的附加标准。本文模型在 Pytorch 上训练，基于 NVIDIA 1070Ti GPU 实现。随机梯度下降(SGD)[31]用于网络优化，初始学习率设为 0.01，且每 30 次迭代后学习率逐渐降低。迭代次数(epoch)置为 60，批处理大小置为 32。通过将图像随机裁剪到 288×144，然后随机水平翻转图像以实现数据增广。经过多次实验，确定了合适的权重参数值，具体为 $\lambda_1$=1、$\lambda_2$=1、$\lambda_3$=1，部分实验结果如表 3 所示。

在表 3 中，列 2~4 使得所有的权重参数值一致，列 5~7 保证 $\lambda_2$ 与 $\lambda_3$ 的数值相同，列 8~10 控制所有的权重参数值之和为 1。具体来说，列 2~4 反映了权重参数值设为大于 1 或小于 1 的情况，然而实验结果并不理想；列 5~7 控制了一个权重参数值与另外两个权重不同，但数据显示其识别精度仍低于列 1；列 8~10 是将 $\sum_{i=1}^{3}\lambda_i$ 维持在数值 1，然而也未找到期望的实验效果。

#### 4.2.1 模型分析

表 4 列出部分实验结果，其中，列表示特征随机融合方式，行代表随机抽取不同特征。列反映了特征间融合方法对实验结果的影响，行体现了模态和网络层级对行人重识别精度的影响。同时，特征间融合方法的对比，有助于发现最佳的特征融合方式。模态和网络层级对比，则能在模态与网络层级间建立某种联系，这种关联包括模态间组合规律和层级间构建技巧，它使得特征抽取更加明确。通过对行列的整体分析，

表 3 选择权重参数值

Table 3 Selecting value of weight parameter

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda_1$ | 1 | 1.5 | 1.2 | 0.8 | 1 | 1 | 1.5 | 0.5 | 0.8 | 0.4 |
| $\lambda_2$ | 1 | 1.5 | 1.2 | 0.8 | 0.5 | 0.3 | 1.2 | 0.2 | 0.15 | 0.35 |
| $\lambda_3$ | 1 | 1.5 | 1.2 | 0.8 | 0.5 | 0.3 | 1.2 | 0.3 | 0.05 | 0.25 |
| mAP | 31.43 | 28.78 | 29.97 | 30.05 | 26.53 | 27.46 | 29.29 | 30.11 | 30.14 | 28.17 |

表 4 分析特征抽取和特征融合方法的实验结果

Table 4 Analysis of feature extracting and fusion method

| | Random feature pyramid | Horizontal concatenation | | | Vertical concatenation | | | Hybrid concatenation | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | rank-1 | rank-5 | mAP | rank-1 | rank-5 | mAP | T | rank-1 | rank-5 | mAP |
| Different level same modality | $P_r(5)P_r(2)$<br>$P_r(5)P_r(2)$ | 27.79 | 53.98 | 30.05 | 26.37 | 52.27 | 29.05 | 1 | 27.61 | 53.23 | 29.90 |
| | | | | | | | | 3 | 27.02 | 52.26 | 29.06 |
| | $P_i(4)P_i(3)$<br>$P_i(4)P_i(3)$ | 27.89 | 53.58 | 30.10 | 27.13 | 51.87 | 28.96 | 1 | 28.18 | 53.21 | 30.23 |
| | | | | | | | | 2 | 27.40 | 51.79 | 29.42 |
| | $P_r(5)P_r(3)$ | 28.42 | 53.84 | 29.87 | 26.31 | 53.71 | 29.31 | ---------- | | | |
| Same level cross- modality | $P_i(3)P_r(3)$ | 27.96 | 53.79 | 30.15 | 26.10 | 52.22 | 28.64 | ---------- | | | |
| | $P_r(5)P_r(5)$<br>$P_i(5)P_i(5)$ | 26.82 | 51.71 | 29.06 | 25.29 | 50.69 | 27.32 | ---------- | | | |
| Different level cross-modality | $P_r(5)P_r(2)$<br>$P_i(5)P_i(2)$ | 28.03 | 54.31 | 30.28 | 26.80 | 53.24 | 29.62 | 1 | 27.18 | 53.35 | 29.72 |
| | | | | | | | | 3 | 26.83 | 52.59 | 29.30 |
| | $P_r(4)P_r(3)$<br>$P_i(4)P_i(3)$ | 28.05 | 54.48 | 30.39 | 26.81 | 52.83 | 29.93 | 1 | 27.97 | 53.40 | 29.84 |
| | | | | | | | | 2 | 24.89 | 50.92 | 26.79 |
| | $P_r(5)P_i(5)$<br>$P_i(4)P_r(4)$<br>$P_r(5)P_i(5)$<br>$P_i(2)P_r(2)$ | 29.28 | 55.35 | 31.43 | 26.84 | 53.12 | 29.65 | ---------- | | | |

有利于找到一组最适合跨模态行人重识别的特征组合。其中，T 表示混合串联的组合类型，组合 1 表示不同层级的行列融合，组合 2 表示不同尺度的列融合，组合 3 表示行融合的层级不同。

表 4 对比发现，横向串联明显比纵向串联更具融合优势；即使混合串联在多种组合情况下，横向串联也平均高于混合串联形式。因此，不同特征最佳融合方法是横向串联。同时注意到，在混合串联中组合 1 实验效果大于组合 2，说明横向融合时应采用不同尺度的特征。对比组合 1 和组合 3，发现相同层级纵向融合的效果低于不同层级特征的纵向融合效果。因此，横向与纵向上取不同层级特征是混合串联的最佳形式。佐证了设计的低层与高层相结合的多级特征学习方式。本文认为造成这种现象的原因是类内平均相似性会随着融合层级数目的增加而增加，多级映射后不同层级特征图中相同内容的耦合被增强。

相同模态与不同网络层级间对比发现，包含红外数据的行人重识别中跨模态融合能够提升效果。除混合串联情况外，跨模态的特征融合效果均高于单模态融合。对比相同层级与不同层级，发现跨模态融合时，相同层级特征的融合效果低于不同层级，表明不同层级跨模态融合的泛化能力更大，合成数据确实能弥合两种模态间差距。因此，特征抽取的标准可定为抽取不同模态不同层级的特征。综合特征的抽取标准和最佳融合方式，本文找到一组最适合跨模态行人重识别的特征集合 $P_r(5)P_i(5)P_i(4)P_r(4)P_r(5)P_i(5)P_i(2)P_r(2)$。

#### 4.2.2 深层分析结构组件

网络框架中的组成模块会引起行人重识别性能上的变化。表 5 则展示了对性能产生关键影响的结构块。它是基于网络整体框架(如图 1)，然后移除指定结构，即排除了其他影响因素。表 6 基于跨模态 prediction 结构，更改指定模块，形成自身对照，以阐明网络构建过程。

表 5 Model 结构对比

Table 5 Component comparison with network

| Remove out component | rank-1 | rank-5 | rank-20 | mAP |
|---|---|---|---|---|
| RGB-domain and IR-domain | 24.30 | 49.61 | 75.62 | 27.51 |
| IR-RGB block | 25.40 | 50.41 | 76.81 | 28.79 |
| Fusion block | 26.08 | 51.69 | 76.73 | 28.78 |
| Upsample $2^{\times}$ | 26.91 | 52.66 | 77.96 | 29.55 |

表 5 展开分析，当 RGB-RGB 与 IR-IR 结构不采用上采样方法时，识别精度明显下降。证实了放大模态内差异因子的权重能有效修正网络模型的观点。再移除 RGB-RGB 与 IR-IR 结构后，实验效果则 rank-1 降低了 4.98%。说明该结构对于整体网络是至关重要的，仅采用联合模态网络结构对于跨模态数据的利用是不够充分的，将单模态与联合模态结合能有效利用跨模态的特性。验证模态间偏置的利用对网络模型的正效应，是通过对比联合模态 RGB-IR 结构移除前后效果。研究发现，网络模型的设计要基于 RGB-RGB 和 IR-IR 以及 RGB-IR 三种模态。从另一角度证明了，提出的特征金字塔随机融合网络结构是有效的。

表 6 具体来说，研究了奇异矩阵的使用和激活函数的选择。无论是将 fusion block 后接单层 SVD，还是在 IR-RGB block 后连双层 SVD，其效果均下降了。因此，对 fusion block 做双层 SVD 处理以消除不必要的相关性，单层 SVD 则应用在 IR-RGB block 上。此外，本文尝试过多种激活函数，不同激活函数产生不同分类效果。通过多组实验得到了网络中表现最优的一组激活函数，分别是基于 ReLU 激活的 fusion block 和基于 LeakyReLU 分类的 IR-RGB block，而其他组合的激活函数效果均低于它们。如两者均采用 ReLU 或者 LeakyReLU 激活，同等条件下，rank-5 效果分别降低了 3.39%和 2.61%，这或许是部分权重无法更新。因此，如图 1 所示的网络模型的设计是合理的。

### 4.3 对比实验

#### 4.3.1 数据预处理

本文采用超分辨重建技术对数据预处理，在跨模态 SYSU-MM01 数据集上训练超分辨网络 SRCNN[32]，其模型原理决定了若原图片视觉模糊，经过超分重建后，图像会更模糊。而 SYSU-MM01 数据集中红外摄像头捕获的图片很模糊，故不能作为 SRCNN 网络的

表 6 Prediction 结构比较

Table 6 Module comparison with network

| Convert into component | rank-1 | rank-5 | rank-20 | mAP |
|---|---|---|---|---|
| Single SVD | 26.33 | 51.76 | 78.95 | 29.35 |
| Double SVD | 26.93 | 53.27 | 78.86 | 29.40 |
| LeakyReLU | 26.96 | 52.74 | 77.67 | 29.42 |
| ReLU | 26.34 | 51.96 | 78.70 | 28.94 |

输入，只能预处理可见光摄像头下 RGB 图像。为了提高重识别效果，只取四个 RGB 摄像头中一个进行重建，所取摄像头的编号没有影响。将 128×128 大小的三通道彩图做水平翻转，作为超分辨网络的输入，然后使用双三次插值将图像放大 2 倍，并取颜色空间 YCrCb 中 Y 通道；接着对 Y 通道做 9×9 卷积以提取特征，此时特征的通道数为 64；再经过 1×1 卷积拟合特征的非线性映射，最后反卷积重构出大小 512×512 的特征，具体内容如表 7 所示。网络微调后输出高分辨 256×256 大小的图像，并将该图像输入进 FPRnet，FPR 网络训练完成后，再用原始 SYSU-MM01 数据集测试。

本文 SRCNN 模型在 Pytorch 上训练，基于 NVIDIA 1070Ti GPU 实现，初始学习率设为 0.001，批处理大小定成 32，dropout 设置为 0.5，经过随机梯度下降(SGD)[31]800 次迭代实现了网络优化。

#### 4.3.2 实验结果

最近几年，ResNet-50 和 SVDnet 常被作为行人重识别的标准对比框架。本文将提出的 FPRnet 基于 SYSU-MM01 数据集与 One-stream、Two-stream、Zero-padding、BCTR、BDTR、CmGAN、ResNet-50 以及 SVDnet 模型进行了比较，如表 8 和图 4 所示。其中，BCTR 和 BDTR 是 dual-path 的改进版本，具体细节见文献[6]。本文从识别精度(如图 5 所示)和收敛速度(如图 6 所示)两方面对比。通过将 FPRnet 与 SRCNN 结合，来提升实验效果，作为最终训练框架。此外，将 reRanking 微调策略运用到 FPRnet 上，作为补充实验。

实验结果表明，本文方法优于上述其他方法。特别是所提的 FPR 方法 rank-1 和 rank-20 分别高于单流网络 zero-padding 方法 17.24% 和 18.75%。而且 FPR+SRCNN 的 mAP 高于 CmGAN 方法 4.32%，并且 FPR 网络在收敛速度上有了大幅提高，见图 6。

在表 8 的最后，显示了在跨模态行人重识别中加入超分辨的优势。使用行人重识别领域流行的微调策略来优化本文的模型，发现并没有提升 FPR 网络性能，甚至在某种程度上拉低识别效果，本文推测度量相似性时出现伪最近邻问题，导致训练过程中保存的模型不是最佳的。

表 7 超分辨结构

Table 7 Super-resolution structure

| Layer | Input size | Input channel | Output size | Output channel | Structure |
|---|---|---|---|---|---|
| Preprocessing | 128×128 | 3 | 256×256 | 1 | Bicubic interpolation (2) |
| Conv1 | 256×256 | 1 | 256×256 | 64 | 9×9, stride=1 |
| Conv2 | 256×256 | 64 | 256×256 | 32 | 1×1 |
| Conv3 | 256×256 | 32 | 512×512 | 1 | 5×5, stride=1, PixelShuffle (2) |

表 8 在 SYSU-MM01 数据集上对比最新方法

Table 8 Comparison with state of the art on SYSU-MM01

| Methods | rank-1 | rank-10 | rank-20 | mAP |
|---|---|---|---|---|
| One-stream | 12.04 | 49.68 | 66.74 | 13.67 |
| Two-stream | 11.65 | 47.99 | 65.50 | 12.85 |
| Zero-padding | 14.80 | 54.12 | 71.33 | 15.95 |
| CmGAN | 26.97 | 67.51 | 80.56 | 27.80 |
| BCTR | 16.12 | 54.90 | 71.47 | 19.15 |
| BDTR | 17.01 | 55.43 | 71.96 | 19.66 |
| ResNet-50 | 19.36 | 59.89 | 73.47 | 23.85 |
| SVDnet | 21.75 | 58.57 | 73.02 | 25.61 |
| FPRnet | 29.28 | 68.43 | 81.01 | 31.43 |
| FPRnet+SRCNN | 30.02 | 69.08 | 81.19 | 32.12 |
| FPRnet+reranking | 30.99 | 67.91 | 79.76 | 31.17 |

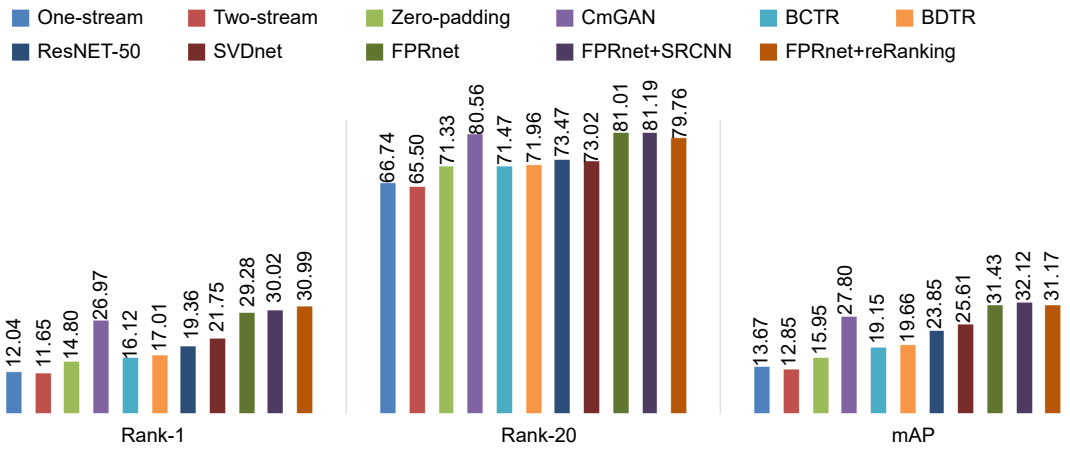



图 4 对比最新方法的实验结果

Fig. 4 Comparison with state of the art on SYSU-MM01

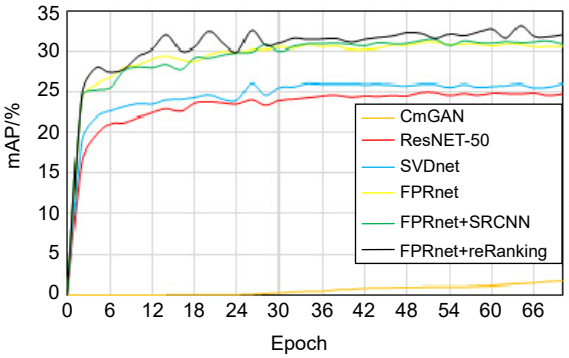



图 5 训练阶段 mAP 变化趋势

Fig. 5 The trend of mAP during training

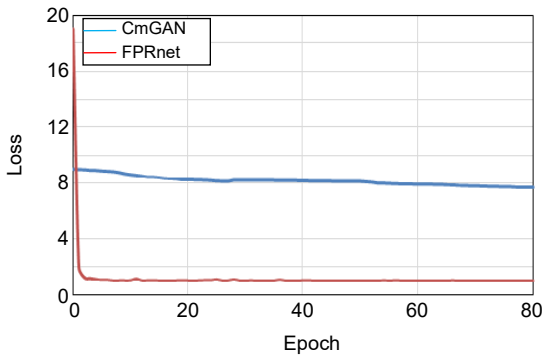



图 6 训练阶段 loss 变化趋势

Fig. 6 The trend of loss during training

## 5 结 论

针对红外与可见光图像的跨模态行人重识别，本文提出了基于混合损失函数的特征金子塔随机融合结构。首先，该网络通过融合高层特征的强语义表征能力与低层特征的强几何细节信息表征能力，实现了高分辨强语义特征作为相似性匹配的唯一输入。其次，结合了全局学习和局部学习的优势后，设计了多层级不同模态特征的随机融合机制。同时，本文引入了超分辨重建 SRCNN 方法来处理红外图像视觉模糊化对网络训练的干扰。最后，提出了跨模态混合损失函数，它利用 RGB-RGB、IR-IR、RGB-IR 三种模态间的偏置有效地修正了模型。实验在 SYSU-MM01 数据集上验证，结果表明，所提的模型优于现有的先进算法。程序开源 https://github.com/KyreneLaura/FPRnet。

## 参考文献

[1] Xu M, Yu X S, Chen D Y, *et al*. Pedestrian detection in complex thermal infrared surveillance scene[J]. *Journal of Image and Graphics*, 2018, **23**(12): 1829–1837.
许茗，于晓升，陈东岳，等. 复杂热红外监控场景下行人检测[J]. 中国图象图形学报, 2018, **23**(12): 1829–1837.

[2] Zheng L, Shen L Y, Tian L, *et al*. Scalable person re-identification: a benchmark[C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*, Santiago, 2015: 1116–1124.

[3] Dai Z Z, Chen M Q, Zhu S Y, *et al*. Batch feature erasing for person re-identification and beyond[Z]. arXiv: 1811.07130[cs:CV], 2018.

[4] Wu A C, Zheng W S, Yu H X, *et al*. RGB-infrared cross-modality person re-identification[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*, Venice, 2017: 2380–7504.

[5] Dai P Y, Ji R R, Wang H B, *et al*. Cross-modality person re-identification with generative adversarial training[C]// *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, Stockholm, 2018: 677–683.

[6] Ye M, Wang Z, Lan X Y, *et al*. Visible thermal person re-identification via dual-constrained top-ranking[C]// *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Palo Alto, 2018: 1092–1099.

[7] Gray D, Tao H. Viewpoint invariant pedestrian recognition with an ensemble of localized features[C]//*Proceedings of the 10th European Conference on Computer Vision*, Marseille, France, 2008: 262–275.

[8] Wang X G, Doretto G, Sebastian T, *et al*. Shape and appearance context modeling[C]//*Proceedings of the 11th International Conference on Computer Vision*, Rio de Janeiro, 2007: 1–8.

[9] Li W, Zhao R, Xiao T, *et al*. DeepReID: deep filter pairing neural network for person re-identification[C]//*Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014: 152–159.

[10] Huang Y, Xu J S, Wu Q, *et al*. Multi-pseudo regularized label for generated data in person re-identification[J]. *IEEE Transactions on Image Processing*, 2018, **28**(3): 1391–1403.

[11] Liu J W, Zha Z J, Tian Q, *et al*. Multi-scale triplet CNN for person re-identification[C]//*Proceedings of the 24th ACM International Conference on Multimedia*, Amsterdam, The Netherlands, 2016: 192–196.

[12] Qian X L, Fu Y W, Jiang Y G, *et al*. Multi-scale deep learning architectures for person re-identification[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*, Venice, 2017: 5399–5408.

[13] Chen Y B, Zhu X T, Gong S G. Person re-identification by deep learning multi-scale representations[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision Workshops*, Venice, 2017: 2590–2600.

[14] Li X, Zheng W S, Wang X J, *et al*, Gong S. Multi-scale learning for low-resolution person re-identification[C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*, Santiago, 2015: 3765–3773.

[15] Wang Z, Hu R M, Yu Y, *et al*. Scale-adaptive low-resolution person re-identification via learning a discriminating surface[C]//*Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, New York, 2016: 2669–2675.

[16] Jing X Y, Zhu X K, Wu F, *et al*. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning[J]. *IEEE Transactions on Image Processing*, 2017, **26**(3): 1363–1378.

[17] Zhang D Q, Li W J. Large-scale supervised multimodal hashing with semantic correlation maximization[C]//*Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, Quebec City, 2014: 2177–2183.

[18] Chen Y C, Zhu X T, Zheng W S, *et al*. Person re-identification by camera correlation aware feature augmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **40**(2): 392–408.

[19] Zhu X F, Huang Z, Shen H T, *et al*. Linear cross-modal hashing for efficient multimedia search[C]//*Proceedings of the 21st ACM International Conference on Multimedia*, Barcelona, 2013: 143–152.

[20] Zhai D M, Chang H, Zhen Y, *et al*. Parametric local multimodal hashing for cross-view similarity search[C]//*Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, Beijing, 2013: 2754–2760.

[21] Srivastava N, Salakhutdinov R. Multimodal learning with deep Boltzmann machines[J]. *Journal of Machine Learning Research*, 2014, **15**(84): 2949–2980.

[22] 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.

[23] Sarfraz M S, Stiefelhagen R. Deep perceptual mapping for cross-modal face recognition[J]. *International Journal of Computer Vision*, 2017, **122**(3): 426–438.

[24] Xiao T, Li H S, Ouyang W L, *et al*. Learning deep feature representations with domain guided dropout for person re-identification[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016: 1249–1258.

[25] Wang F Q, Zuo W M, Lin L, *et al*. Joint learning of single-image and cross-image representations for person re-identification[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016: 1288–1296.

[26] Jiang X Y, Wu F, Li X, *et al*. Deep compositional cross-modal learning to rank via local-global alignment[C]//*Proceedings of the 23rd ACM International Conference on Multimedia*, Brisbane, 2015: 69–78.

[27] Møgelmose A, Bahnsen C, Moeslund T B, *et al*. Tri-modal person re-identification with RGB, depth and thermal features[C]//*Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Portland, OR, 2013: 301–307.

[28] Sun Y F, Zheng L, Deng W J, *et al*. SVDNet for pedestrian retrieval[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*, Venice, 2017: 3800–3808.

[29] Maas A L, Hannun A Y, Ng A Y. Rectifier nonlinearities improve neural network acoustic models[C]//*Proceedings of 30th International Conference on Machine Learning*, Atlanta, Georgia, 2013: 18–23.

[30] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]//*Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, 2011: 315–323.

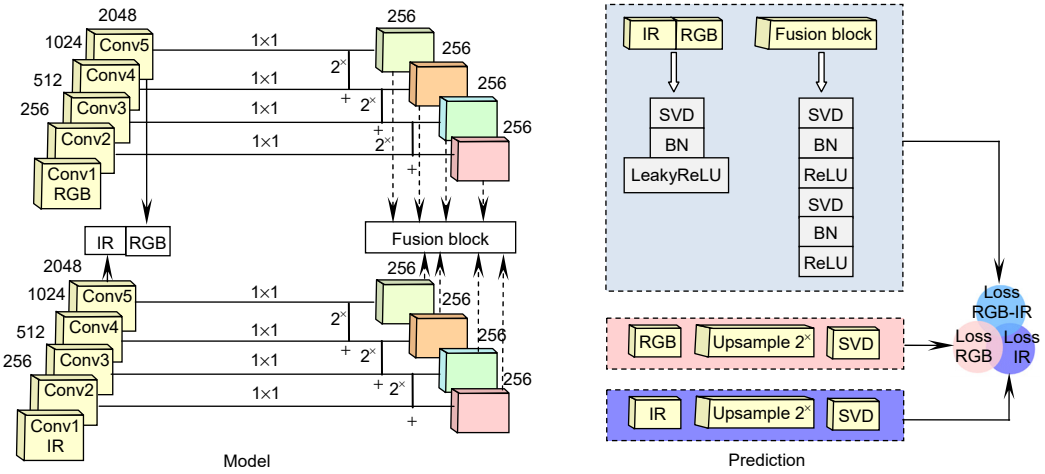
[31] Bottou L. Stochastic gradient descent tricks[M]//Montavon G, Orr G B, Müller K R. *Neural Networks*: *Tricks of the Trade*. Berlin, Heidelberg: Springer, 2012: 421–436.

[32] Dong C, Loy C C, He K M, *et al*. Image super-resolution using deep convolutional networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **38**(2): 295–307.

# Feature pyramid random fusion network for visible-infrared modality person re-identification

Wang Ronggui, Wang Jing, Yang Juan*, Xue Lixia

School of Computer & Information, Hefei University of Technology, Hefei, Anhui 230009, China




The framework of feature pyramid random fusion network

Overview: Existing works in person re-identification only considers extracting invariant feature representations from cross-view visible cameras, which ignores the imaging feature in infrared domain, such that there are few studies on visible-infrared relevant modality. Besides, most works distinguish two-views by often computing the similarity in feature maps from one single convolutional layer, which causes a weak performance of learning features. To handle the above problems, we design a feature pyramid random fusion network (FPRnet). Firstly, we introduce SRCNN of a super-resolution reconstruction method to preprocess, and the purpose is to alleviate the interference of IR-images blur and make feature learning more robust. Secondly, we take ResNet-50 pre-trained on ImageNet dataset as a baseline to learn feature representations of images in RGB-domain and IR-domain. The re-identification based on the residual network can only learn features with one resolution scale. However, tracking a specific person requires multi-directional learning, including the pedestrian's overall properties, local attributes and important characteristics to reduce the occurrence of misjudgment. For this reason, referring to the thought of the feature pyramid network, the features of different convolution layers in ResNet-50 network are constructed into a pyramid structure. It can calculate the similarity between multiple features at the same time, and abandons the approach of the original pyramid network using different scales to adapt to pedestrian bounding box images. Instead, it embeds the spirit of the pyramid structure into the depth residual network as a feature extraction module to extract the IR-RGB block. This learning method integrates the advantages of learning local and global feature, and represents the features with strong semantics and strong geometric details. Then, the random fusion mechanism is used as the basis of the feature fusion module to complete the end-to-end design of the double-branch, and the fusion block is obtained, which can avoid the problem of excessive parameters in the pyramid model. Thirdly, after the tasks of feature extraction and feature fusion are completed, the cross-modality prediction is carried out. It consists of a blue cross-domain, a pink RGB-domain, and a purple IR-domain. It generates three types of classification loss, and then uses a hybrid loss function to reduce the gaps between the intra-modality visual appearance and the inter-modality heterogeneity issue. The IR-RGB block and fusion block use a minimax game to beat each other for learning the joint-modal classification loss. Finally, the original dataset is utilized for FPRnet testing. Extensive experiments on the public SYSU-MM01 dataset from aspects of mAP and convergence speed, demonstrate the superiorities in our approach to the state-of-the-art methods. Furthermore, FPRnet also achieves competitive results with 32.12% mAP recognition rate and much faster convergence. The source code of the FPRnet can be available from https://github.com/KyreneLaura/FPRnet.

Citation: Wang R G, Wang J, Yang J, *et al*. Feature pyramid random fusion network for visible-infrared modality person re-identification[J]. *Opto-Electronic Engineering*, 2020, **47**(12): 190669

* E-mail: yangjuan6985@163.com