计算机科学
COMPUTER SCIENCE

基于双目叠加仿生的微换衣行人再识别

陈坤峰，潘志松，王家宝，施蕾，张锦

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210600140

基于双目叠加仿生的微换衣行人再识别

陈坤峰　潘志松　王家宝　施　蕾　张　锦
陆军工程大学指挥控制工程学院　南京 210007
(kfchenhn@163.com)

摘　要　微换衣行人再识别是以换衣幅度不大的情况为前提,从不同摄像头场景中查找某特定身份的行人的一项计算机视觉技术。现有行人再识别方法的实现通常是基于行人衣着不变的假设,因此它们依赖的是与衣着相关的特征。那么,当此假设不成立时,这些方法就难以实现理想的识别效果。考虑到行人换衣幅度不大时行人体态基本不发生改变这一重要特点,针对微换衣行人再识别展开研究。受生物视觉系统中双目叠加效应的启发,采取仿生思想提出一个自注意力孪生网络,类比生物双眼获取信息的过程。首先,该网络以同一行人不同衣着的两类图像作为双分支输入,并利用孪生架构实现叠加效应。随后对输出的多个特征进行对比学习和融合学习,进而得到具有身份辨别力的行人特征表示。最后,在微换衣行人再识别相关数据集上进行了充分实验,结果表明该方法可达到当前最好的识别性能。

关键词:行人再识别;微换衣;双目叠加效应;自注意力;孪生网络
中图法分类号　TP391

Moderate Clothes-Changing Person Re-identification Based on Bionics of Binocular Summation

CHEN Kun-feng,PAN Zhi-song,WANG Jia-bao,SHI Lei and ZHANG Jin
College of Command and Control Engineering,Army Engineering University of PLA,Nanjing 210007,China

Abstract　Moderate clothes-changing person re-identification is to find the same person from different camera scenes under the premise of considering the moderate change of clothes. The implementation of existing person re-identification methods is usually based on the assumption that the pedestrian's clothing is invariant,so they rely on clothing-related features. Then,when the above assumptions are not valid,these methods are difficult to achieve the ideal recognition performance. Considering the important characteristic that pedestrian's shape hardly change when the change of clothes is moderate,the moderate clothes-changing person re-identification is studied. Inspired by the binocular summation in biological vision system,a self-attention siamese network is proposed in this paper. Analogous to biological binocular information acquisition process,the network takes two types of images of the same pedestrian with different clothes as two branch inputs,and then achieves summation effect with siamese architecture. Subsequently,the contrastive learning and fusion learning of multiple features are carried out to obtain the pedestrian feature representation with identity discrimination. Finally,empirical studies show that the proposed method achieves best performance at present on clothes-changing person re-identification benchmark.

Keywords　Person re-identification,Moderate clothes-changing,Binocular summation,Self-attention,Siamese network

1　引言

行人再识别(Person Re-Identification,Person ReID)的目的是从不同摄像头场景中检索出特定身份的行人。当人脸识别等近距离识别方法难以奏效时,行人再识别可以根据身体特征对行人身份进行区分。作为计算机视觉研究中的一项重要技术,行人再识别被广泛应用于智能监控和视频分析等领域。2006 年,Gheissari 等[1]首次提出行人再识别的概念。早期关于行人再识别的研究常常使用颜色和纹理等信息手工设计行人特征[2-3]。近年来,得益于神经网络[4]和特征度量学习[5-6]等方法的蓬勃发展,我们可以利用深度学习提取出具有辨别力的行人特征[7]。因此,行人再识别技术取得越来越多的创新和突破,并且在相关的数据集上实现了较高的识别准确率[8-10]。这些行人再识别方法虽看似性能良好,但大多数都是基于行人衣着几乎不变这一假设。在不同时间、不同地点的环境下,该假设显然不满足现实需求。由于传统不考虑

到稿日期:2021-06-17　返修日期:2021-10-19
基金项目:国家自然科学基金(62076251);江苏省自然科学基金(BK20200581)
This work was supported by the National Natural Science Foundation of China(62076251) and Natural Science Foundation of Jiangsu Province (BK20200581).
通信作者:潘志松(hotpzs@hotmail.com)

换衣情况的行人再识别方法通常使用行人的 RGB 图像作为输入，其衣着颜色自然成为了特征提取过程中所依赖的重要信息。那么，一旦这些传统方法遇到换衣行人，则很容易错误地将某一行人识别为与其穿着类似的另一行人，从而降低识别率。

面对行人换衣这一技术需求，现有一些研究工作开展了突破性探索。这些方法大体可分为两类：基于视频的方法和基于图像的方法。一些研究工作提到基于视频的行人再识别技术[11]和步态识别技术[12]，通过视频序列来获取行人的动作特征。这些方法考虑了特定身份的行人在动态行进过程中的独特性。显然，这一独特性将有利于对行人身份特征的获取。但是很多时候某些行人出现在摄像头视野中的时间较短，难以体现完整的动作周期。因此，基于图像的行人再识别技术依然很有研究意义。在基于图像的方法中，一些研究采用深度图像作为行人再识别的算法输入[13-15]。深度图像可以得到 3D 的人体信息，因此可以利用丰富的特征很好地应对换衣的影响。但是深度图像的获取需要昂贵的深度图像摄像机，而这种摄像机在当前大部分监控场景中尚未普及。因此基于深度图像的方法的推广面临很大的现实阻力。另一些方法则采用强调面部的手段来学习与衣着无关的特征[16]。面部信息具有较高的区分性，因此加入人脸识别的约束可以更好地找到目标行人。但是，长距离的拍摄难以获得清晰的人脸，而且对于行进中的人体，被拍摄到的常常是侧脸，因而给人脸识别造成困难。此外，Yang 等[17]的研究发现，在行人换衣幅度不大时，其体型轮廓不会有较大变化，因而提出基于轮廓图的深度学习网络来应对微换衣行人再识别问题。同时，Li 等[18]在其研究中也采用体型信息来克服换衣挑战。但是这些基于人体形态的方法需要借助边缘检测、图像分割等辅助手段，难免会出现估计不准确的情况。

通过分析以上方法的优点和不足，我们认识到，即使存在行人换衣情况，依然可以根据不依赖衣着的图像特征来判别行人身份，但仍需探索更加灵活可靠的特征提取方法。为了避免添加时序信息、动作信息和深度信息等辅助信息造成的算法过于复杂，我们考虑仅使用图像本身的新奇性来解决微换衣行人再识别问题。本文充分利用微换衣情况下行人体型变化不大这一特点，提出了双目叠加仿生的微换衣行人再识别方法。如图 1 所示，观察生物视觉系统的结构可以发现，双眼获取到的信息进入脑部神经系统的过程中，视交叉的部分形成的双目叠加效应(Binocular Summation)[19-20]尤为重要。利用叠加效应可以对双眼获取到的信息进行对比和融合，从而更好地看清和识别各类物体。受此启发，本文采取仿生的思想提出了自注意力孪生行人再识别网络(Self-attention Siamese Person Re-Identification Network，SSN)，以解决微换衣挑战下的行人再识别问题。在 SSN 中，类比输入视觉信息的双眼。首先，将同身份不同着装的两类行人图像作为网络两个分支的输入；其次，为了模拟视觉系统中的视交叉结构，利用本文提出的孪生网络架构实现了叠加效应；然后对多个输出特征进行对比学习和融合学习，进而得到具有身份辨别力的行人特征表示。

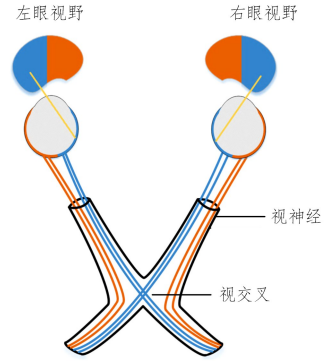

图 1 生物视觉系统中双目叠加效应

Fig. 1 Binocular summation in biological vision system

在 SSN 的宏观网络架构的基础上，本文方法还对网络细节模块做了创新性设计。受 Yang 等[17]对行人轮廓图像的运用的启发，我们考虑灰度图像应该也具有和轮廓图像相似的作用，亦可让模型更加关注与行人衣着无关的辨别性特征。此外，灰度图像比轮廓图像保留有更多的原有信息。因此，本文首次提出使用行人的原始 RGB 图像、灰度图像以及轮廓图像作为融合输入。特别地，利用文中提出的与衣着无关的两个自注意力模块(Cloth-independent Self-attention Module，CSM)来挖掘灰度图像(轮廓图像)和 RGB 图像中的身份不变特征，生成两种信息增强图像。因为不同的生成图像具有不同的特征属性，所以本文方法在特征提取器中设计了双属性融合(Dual-Attribute Fusion，DAF)策略。此外，受 SIFT[21]和 YOLOV3 等[22]研究工作中“金字塔思维”的启发，本文利用空洞卷积金字塔方法提出了一个轻量级特征增强自注意力模块(Lightweight Feature-enhanced Self-attention Module，FSM)。实验表明，在微换衣行人再识别任务中，FSM 可以在增加较小计算量的情况下获得较大的识别性能提升。综上所述，本文方法的创新性贡献可以总结为以下 4 点：

(1)受生物视觉系统中双目叠加感知的启发，本文采取仿生思想提出了自注意力孪生网络。以同一行人不同衣着的两类图像作为双分支输入，利用孪生网络架构实现叠加效应，然后对多个输出特征进行对比学习和融合学习。该创新性方法可有效解决着装可变环境下的行人再识别问题。

(2)首次采用 3 种异质图像融合输入的方式，利用与着装无关的自注意力模块获得了 2 种具有不同属性的生成图像，用以增强与衣着无关的行人身份特征。

(3)在特征提取器中，本文提出了双属性融合策略来充分融合使用两种不同特征属性的生成图像中蕴含的不同图像信息。此外，我们在特征提取器中设计了一个轻量化特征增强自注意力模块，在不增加较多计算量的情况下提升特征表示的鲁棒性。

(4)在相关的行人再识别数据集上的实验结果表明，本文方法可以取得当前最好的识别效果。

2 本文方法

本节首先对所提方法的网络架构和模块设计进行概述，然后对其中的两个创新性模块的实现和损失函数的计算方法展开介绍。

2.1 方法概述

为实现着装可变环境下同身份行人图像特征的对比学习和融合学习，受生物视觉系统中双目叠加感知的启发，本文采取仿生的思想提出了自注意力孪生行人再识别网络（SSN）。整体网络架构如图 2 所示。在网络的双分支中分别输入同身份行人两种不同穿着的图像，把 RGB 图像、轮廓图像和灰度图像作为各分支的融合输入，其中 RGB 图像直接来自监控视频的行人检测结果，轮廓图像是利用边缘检测技术[23]处理原 RGB 图像所得，灰度图像则来自对 RGB 图像的灰度化处理。因此，整个框架的输入输出可以被认为是端到端的。在孪生特征提取器的作用下得到两种输出特征，然后利用特定的损失函数对其做对比学习和融合学习，以获得具有身份判别力的行人特征表示。在特征提取器中，本文提出两种自注意力模块和一种融合操作，分别是着装无关自注意力模块（CSM）、轻量级特征增强自注意力模块（FSM）和双属性融合（DAF）。

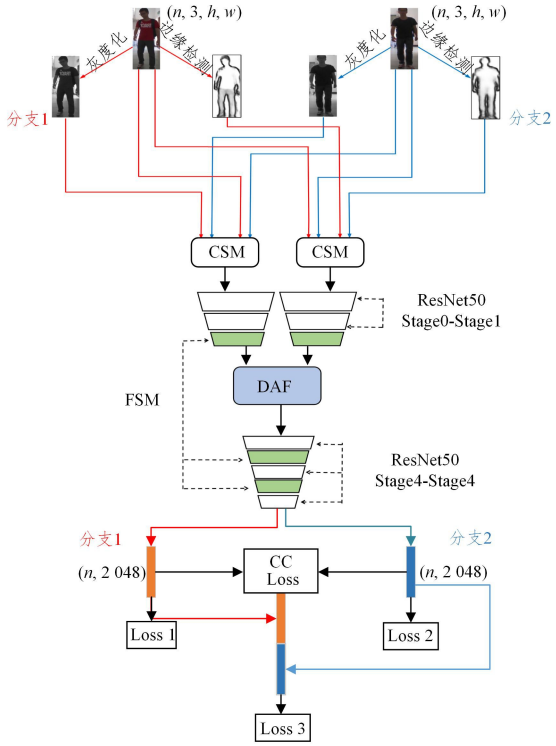

图 2 自注意力孪生行人再识别网络架构

Fig. 2 Architecture of self-attention siamese person re-identificatio network

如图 2 所示，与生物双眼间对称关系类似，SSN 可以分为对称的左右两部分。在网络训练时使用两个部分，在测试推理过程使用一个部分即可。因为两部分呈对称关系，所以下文仅介绍其中一个部分。首先，RGB 图像被输入到网络中并被转化为轮廓图像和灰度图像。然后，两对异质图像，也就是 RGB 图像和灰度图像，以及 RGB 图像和轮廓图像，被分别送进两个 CSM。这一操作的目的是将轮廓图像和灰度图像中蕴含的与衣着无关的身份特征迁移到 RGB 图像中。因此，我们可以在 CSM 输出时获得两种生成图像，然后将生成的两种三通道图像送入基于 ResNet50 的特征提取器。在特征提取器的前面若干个阶段，用不共享参数的两个网络分支分别提取两种生成图像的低级特征。为了实现双属性特征融合（DAF），以上得到的来自两个分支的低级特征在通道维度被级联在一起，紧接着利用 1×1 卷积将通道维度减半。随后把融合后的特征张量送进特征提取器中其余的卷积层中。按照常规，本文把 ResNet50 分为 5 个阶段：Stage0（浅层卷积层）、Stage1－Stage4（残差卷积层）。后续的实验结果表明，当特征张量在 Stage1 处融合时可以得到最佳识别性能。此外，为了提升模型性能，本文方法在特征提取器的 Stage1，Stage2 和 Stage3 之后插入 FSM。

最后，两种着装的行人图像被编码为两种特征向量。为了让特征体现出与衣着无关的特性，需要挖掘两种特征的共性特点。因此，我们在网络学习过程中计算了两种特征的换衣损失（Clothes-Changing Loss，CC Loss），对两分支输出特征以及两种特征的级联结果计算交叉熵损失（Cross Entropy Loss，CE Loss）和三元组损失（Triplet Loss），并运用 SGD 优化器实现网络训练。

2.2 与衣着无关的自注意力模块

现有方法表明，若不考虑行人换衣的情况，使用 RGB 图像作为输入是效果最好的方式。Yang 等[17]的研究证实行人的轮廓图像有助于实现受换衣因素影响的行人再识别，原因是轮廓图像摒弃了行人的着装信息，仅保留了行人的体型信息。而且，当衣着变化不大时，行人的体型轮廓具有独特性，能较好地体现行人身份。除此之外，我们认为着装不同的行人图像最大的区别应该是颜色。因此，如果将灰度图像利用起来，就可解耦获得与颜色无关的行人身份特征。当然，这也属于与衣着无关的特征。为了实现在有无换衣情况皆可达到较高识别正确率的行人再识别算法，我们考虑综合使用 RGB 图像、灰度图像和轮廓图像 3 种图像作为输入。那么，如何才能实现这 3 种异质图像的充分利用呢？

作为计算机视觉研究中的重要角色，注意力机制可以引导卷积神经网络学习到最具信息价值的图像特征。例如，Hu 等[24]考虑特征张量的不同通道间的关系及其对识别结果的影响力，提出了 SE（Squeeze-and-Excitation）注意力模块，用于给不同通道分配不同的权值。作为注意力机制的变体，自注意力减少了对外部信息的依赖，更擅长捕捉数据或特征的内部相关性。受此启发，本文希望能够利用注意力机制的这种“聚焦”思想，让模型更加关注与衣着无关的行人特征，因此提出了衣着无关自注意力模块（CSM）。CSM 的结构如图 3 所示。

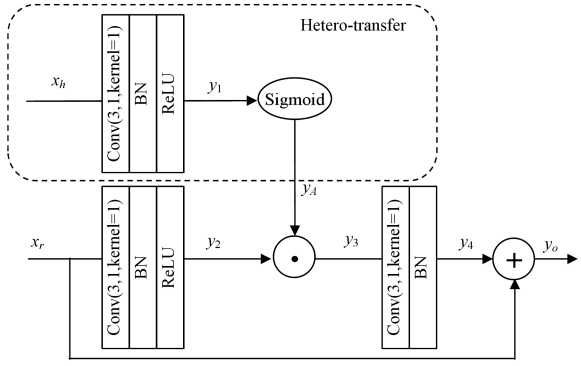

图 3 与衣着无关的自注意力模块结构

Fig. 3 Structure of clothes-independent self-attention module

本文提出的CSM具有两个输入和一个输出。x_h用来输入轮廓图像或灰度图像，x_r用来输入RGB图像。而输出的y_o表示合成的三通道图像。在CSM中，x_h和x_r首先被1×1卷积从三通道降维到单通道，并经过BatchNormalization层转化和ReLU层激活，得到y_1和y_2。对y_1做Sigmoid操作，即得到注意力掩码y_A。其次，通过计算y_A和y_2的点积得y_3，实现异质图像的信息迁移(Hetero-Transfer)。然后，将y_3升维到三通道的y_4，把原始RGB图像x_r和y_4相加，则可以得到CSM的输出y_o。

2.3 轻量化特征增强自注意力模块

AGW[25]的研究工作证明在ResNet50中插入Non-Local自注意力模块[26]可以有效提升特征提取器的性能。然而，引入Non-Local自注意力模块所带来的巨大计算量不容忽视。那么，我们能否找到一种有效的替代方法，用较小的计算量就能达到和Non-local相似的效果呢？

在卷积神经网络中，感受野的大小决定着该层卷积关注的信息区域的大小，感受野越大，则越能关注到图像的整体信息。虽然增大卷积核的大小可以增大感受野，但同时也会带来更大的计算耗费。为了在不增加太多计算量的前提下增大卷积感受野，Yu等[27]提出了空洞卷积。此外，在“金字塔思维”的启发下，我们运用空洞卷积金字塔提出了轻量化特征增强自注意力模块，将其插入到ResNet50中，以提升特征提取器对图像外观的多尺度不变特征的关注。FSM的结构如图4所示。

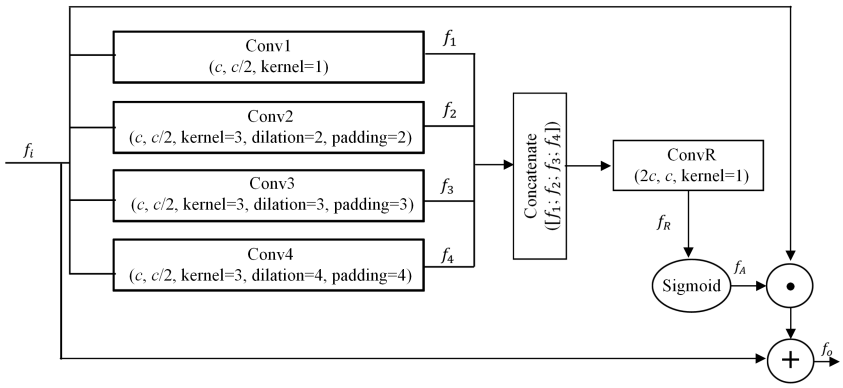

图4 轻量化特征增强自注意力模块结构

Fig. 4 Structure of lightweight feature-enhanced self-attention module

图4中，f_i表示FSM的输入，即基于ResNet50的特征提取器在某一阶段的输出获得的特征张量。如图中从上到下，FSM的金字塔结构由空洞率分别为1，2，3和4的空洞卷积$Conv1$，$Conv2$，$Conv3$和$Conv4$构成，它们各自的输出为f_1，f_2，f_3和f_4。确切地说，第一个空洞率为1的卷积即是传统卷积，设其卷积核的大小为1×1，其余的空洞卷积层的卷积核设为3×3。对于每个空洞卷积层，输出特征张量的通道数量首先被降维到输入特征张量的1/2，随后，4个空洞卷积层输出的特征张量在通道方向上级联在一起。为了保持FSM的输入和输出的维度一致，级联后的特征张量被降维到输入的维度；接着，对降维后的张量f_R做Sigmoid操作，得到其注意力掩码f_A；然后对f_i和f_A做点积运算，再加上f_i，实现特征的加权融合，即可得到整个FSM的输出f_o。

2.4 损失函数

本文提出的SSN可输出同身份不同衣着行人图像的两种特征向量。对于这两个不同分支输出的特征向量，为了提取到具有身份辨别性的特征表示，我们采取交叉熵损失函数来优化分类结果，并且将三元组损失函数应用于距离度量学习。

交叉熵损失函数的计算式为：

$$L_{CE}=-\sum_{i=1}^{N}p(x_i)\log q(x_i) \tag{1}$$

其中，x_i表示第i张输入图像，$p(x_i)$表示真实标签，$q(x_i)$是此图片输入网络后提取到的特征向量经Softmax层之后得到的预测标签。

三元组损失函数的公式为：

$$L_{Tri}=\frac{1}{N}\sum_{a\in batch}[\alpha+\max_{p\in A}d_{a,p}-\min_{n\in B}xd_{a,n}]_+ \tag{2}$$

式(2)意即，对于一个输入批次(batch)中的N个样本(a)，在训练过程中，拉近a与其同类样本($p\in A$)特征间的距离，拉远其与异类样本($n\in B$)特征间的距离。式(2)中α表示三元组损失的边界阈值，$d_{a,p}$，$d_{a,n}$表示特征距离(相似度)的度量函数。$[\ ]_+$的意思是，若方括号里的计算结果小于0，则记为0。

在两个分支输出特征以及级联后的特征上的损失函数的计算式如下：

$$Loss_b=L_{CE}^b+L_{Trip}^b(b=1,2,3) \tag{3}$$

其中，$b=1,2,3$表示在左分支、右分支输出特征向量，以及两分支的输出向量在通道维度级联后的向量。

此外，为了模拟生物视觉的双目叠加效应，除了特征级联实现的融合学习外，还需实现两分支输入特征的对比学习。因此，本文方法还计算了两特征间的换衣损失。换衣损失采用均方误差实现，计算式如下：

$$L_{CC}=MSE(v_1,v_2) \tag{4}$$

其中，v_1和v_2分别代表分支1和分支2输出的特征向量。

整体的损失函数计算式表示如下：

$$L_{\text{total}}=L_1+L_2+L_3+\lambda L_{CC} \tag{5}$$

3 实验与结果分析

本节首先介绍了实验所用到的数据集和评价指标，然后为本文方法的若干细节设计了性能分析实验，最后将本文方法和该研究领域的相关工作进行了对比实验。由于本文提出的SSN网络的结构左右对称，为了更加清晰地体现网络中各模块的性能，以下对细节的分析实验仅使用了网络的单个

分支(Single-Branch Net)。

3.1 数据集与评价指标

考虑到本文方法使用了行人的轮廓图像，而且没有使用面部信息，因此使用 PRCC(Person Re-id under moderate Clothing Change)数据集[17]来测试本文方法。该数据集包含 221 个行人的 33 698 张图片。这些图片分别采集自 3 个摄像头场景(A，B 和 C)。同身份行人在场景 A 和场景 B 中穿着同样的衣服，处于不同的环境中，场景 A 和场景 C 中的同身份行人穿着不同的衣服。与其他相关方法一样，我们用 PRCC 中的 150 个行人的图像作为训练样本，其余 71 个行人的图像用作测试。测试时，将场景 B 或 C 中的图像作为查询图像(Query Images)，将场景 A 中的图像作为库图像(Gallery Images)。测试过程中采用 single-shot 检索的方式，即库图像中每一身份行人仅使用一张随机选取的图像，使用累积匹配曲线(Cumulative Matching Characteristics，CMC)的 Rank-k 作为性能评价指标，记录实验“$B\rightarrow A$”(用 B 检索 A，不换衣)和“$C\rightarrow A$”(用 C 检索 A，换衣)两种条件下的测试结果。

3.2 实验环境与参数设置

本文方法实验所需的硬件条件是两张 NVIDIA GeForce RTX 2080Ti 显卡。代码依靠 Pytorch 1.7 机器学习库实现。输入的行人图像尺寸大小重置为 384×128。在训练过程中，数据读取的 batchsize 设置为 64，即每次选取 4 个行人，其中每个行人分配 16 张图像。三元组损失函数的阈值 α 设为 0.3，SGD 优化器的动量值设为 0.9。另外，在计算损失函数的过程中，参数 λ 设为 0.01。网络一共训练 80 轮，学习率 lr 随训练轮数(t)衰减的策略如下：

$$lr(t)=\begin{cases}0.035\times\dfrac{t}{10}, & t\leqslant 10\\ 0.035, & 10\leqslant t\leqslant 20\\ 0.0035, & 10\leqslant t\leqslant 20\\ 0.00035, & 10\leqslant t\leqslant 20\end{cases} \tag{6}$$

3.3 不同图像输入方式的性能分析

为探究 RGB 图像、灰度图像和轮廓图像这 3 种异质图像在换衣行人再识别系统中的表现，本实验在单分支网络上，用仅一种图像输入 ResNet50 的方法，以及 2 种或 3 种图像经过 CSM 模块输入 ResNet50，并在 ResNet50 的 Stage0 后融合特征的方法，进行了对比实验，结果如表 1 所列。

表 1 各种图像输入的性能分析

Table 1 Performance analysis of each image input method

(单位：%)

图像输入种类	方法性能(Rank-1)	
	$B\rightarrow A$	$C\rightarrow A$
RGB 图像	74.80	19.43
灰度图像	70.42	18.57
轮廓图像	37.25	18.39
RGB 图像＋灰度图像	76.34	31.26
RGB 图像＋轮廓图像	75.28	34.39
RGB 图像＋灰度图像＋轮廓图像	**83.23**	**36.21**

从实验结果可以看出，当行人再识别系统仅输入一种图像时，在不换衣的条件下，使用 RGB 图像的效果最好，我们推测这是因为 RGB 图像中的信息量较为丰富；而在换衣条件下，3 种输入方式的识别性能都大幅下降。但是在换衣条件下，信息量较少的轮廓图像和灰度图像仍可达到与 RGB 图像区别不大的效果。这表明：若不存在换衣，深度神经网络则会依赖衣服信息来辨别行人身份，而一旦出现换衣状况，衣服信息则成了干扰。当行人再识别系统输入 2 种或 3 种图像时，识别性能会有一定的提升。因此，我们认为综合使用 RGB 图像、灰度图像以及轮廓图像可以更好地利用异质图像的语义互补信息，避免以上问题。灰度图像和轮廓图像的获得来自 RGB 图像，其中蕴含的信息是 RGB 图像信息的子集，本文提出的 CSM 则是对该信息子集的增强，以达到自注意力机制的作用。表中“RGB 图像＋灰度图像＋轮廓图像”的实验结果也验证了我们的观点。

3.4 不同 DAF 策略的性能分析

在确定了图像输入为“RGB 图像＋灰度图像＋轮廓图像”并使用了 CSM 之后，为了寻求两种生成图像特征的双属性融合(DAF)位置，我们分别在 ResNet50 的 Stage0—Stage4 这 5 个位置做了尝试，相关实验结果如图 5 所示。

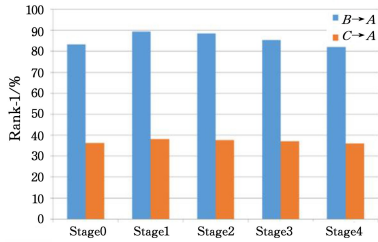

图 5 各种 DAF 策略的实验结果

Fig. 5 Performance of each fusion method for DAF

我们选取了在 Stage1 进行特征融合的网络设计。如图 2 所示，3 种图像经过 CSM 后，2 种生成图像被送进特征提取器 ResNet50 中，在 Stage1 之前提取各自分支的低级特征，然后在 Stage1 之后进行特征的双属性融合，提取高级特征。

3.5 不同 FSM 插入方式的性能分析

在确定图像输入方式和 DAF 位置之后，则需探索 FSM 的插入位置，本文在 MIMO 的单分支网络上设计开展了以下实验。对于 ResNet50 的 5 个阶段，首先分别单独在每个阶段之后插入一个 FSM，然后选取在换衣场景下行人再识别性能较好的前两个位置、前 3 个位置以及前 4 个位置插入 FSM，测试其识别准确率。此外，我们还在相同位置对比了 FSM 和 Non-Local 注意力模块的性能。相关的实验结果如表 2 所列。

表 2 各种 FSM 插入方式的实验结果

Table 2 Performance of each method of inserting FSM

(单位：%)

FSM 插入方法	方法性能(Rank-1)			
	$B\rightarrow A$		$C\rightarrow A$	
	FSM	Non-Local	FSM	Non-Local
Stage0	90.02	87.23	38.17	38.45
Stage1	90.81	86.31	38.25	37.94
Stage2	90.67	88.23	38.54	38.31
Stage3	90.73	89.27	38.90	37.35
Stage4	90.21	90.04	38.03	38.42
Stage2＋Stage3	90.97	—	39.02	—
Stage1＋Stage2＋Stage3	91.73	—	39.52	—
Stage1＋Stage2＋Stage3＋Stage4	91.20	—	38.93	—

从表 2 的实验结果中不难发现，当特征提取器中的各个

阶段单独插入一个注意力模块时，FSM 的性能优于 Non-Local 注意力模块。对于 FSM 的各种组合插入方式，“Stage1＋Stage2＋Stage3”可实现最佳性能。因此，我们在 MIMO 的两个分支设计了如图 2 所示的结构。

3.6 与其他方法的对比分析

为验证本文方法的有效性与先进性，我们在 PRCC 数据集上将其与当前一些代表性的方法进行了对比实验。这些方法包括手工特征方法 LBP[28]，HOG[29]，KISSME[30]和经典的深度方法 ResNet50[4](RGB)，PCB[31](RGB)，以及当前相关工作中的最佳方法 SPT＋ASE[17]。对比实验的结果如表 3 所列。

表 3 在 PRCC 数据集上的对比实验结果

Table 3 Comparative experimental results on PRCC dataset

（单位：%）

方法	C→A（换衣）		B→A（不换衣）	
	Rank-1	Rank-10	Rank-1	Rank-10
LBP[28]＋KISSME[30]	18.71	58.09	39.03	76.18
HOG[29]＋KISSME[30]	17.52	49.52	36.02	68.83
ResNet50[4](RGB)	19.43	52.38	74.80	97.28
PCB[31](RGB)	22.86	61.24	86.88	98.79
SPT＋ASE[17]	34.38	77.30	64.20	92.62
Single-Branch Net	39.52	80.02	91.73	98.67
SSN(无换衣损失)	40.41	82.30	92.35	98.73
SSN	**42.35**	**85.42**	**91.84**	**99.07**

表 3 中的“Single-Branch Net”指 SSN 网络的单个分支，“SSN”和“SSN(无换衣损失)”代表有无换衣损失(CC Loss)的网络结构。可见，本文方法的单分支网络比基线方法 ResNet50(RGB)在换衣和不换衣条件下的 Rank-1 都超出了大约 20%。与当前最佳方法 SPT＋ASE 相比，本文方法的性能也明显最优。具备换衣损失的 SSN 结构可以达到最佳效果。因此可以证明本文方法可有效应用于微换衣行人再识别任务中。

结束语 本文综合考虑了衣着相关信息和衣着无关信息对微换衣行人再识别任务的影响，以双目叠加仿生的思想提出了自注意力孪生网络。以同一行人不同衣着的两类图像作为双分支输入，然后在孪生网络架构中实现叠加效应。该方法以不损失原图像信息的方式，有效挖掘了与行人衣着无关的判别性特征，可实现一个较为鲁棒的微换衣行人再识别系统。在相关数据集上的实验结果表明，本文方法明显优于其他相关方法，并且为本领域的进一步研究提供了一种简单而有效的思路。

尽管本文方法在换衣行人再识别任务上具有良好的性能，但是该方法仍局限于常见的微换衣的场景。当行人换衣幅度较大时，仍需面部信息或视频序列的辅助表示才可有效建模。此外，虽然微换衣是出现最普遍的情况，但是目前该领域的公开数据集较少，因此本文仅能依靠一个数据集上的实验评价结果。下一步可以在数据集采集方面做一些工作，以及融合面部信息和动作信息来解决复杂换衣行人再识别问题。

参 考 文 献

[1] GHEISSARI N, SEBASTIAN T B, HARTLEY R. Person reidentification using spatiotemporal appearance[C]// 2006 IEEE-Computer Society Conference on Computer Vision and Pattern Recognition(CVPR'06). IEEE, 2006: 1528-1535.

[2] FARENZENA M, BAZZANI L, PERINA A, et al. Person re-identification by symmetry-driven accumulation of local features [C]// 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010: 2360-2367.

[3] GRAY D, TAO H. Viewpoint invariant pedestrian recognition with an ensemble of localized features[C]// European Conference on Computer Vision. Berlin: Springer, 2008: 262-275.

[4] 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.

[5] PAISITKRIANGKRAI S, SHEN C, VAN DEN HENGEL A. Learning to rank in person re-identification with metric ensembles[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1846-1855.

[6] SHEN Y, XIAO T, LI H, et al. End-to-end deep kronecker-product matching for person re-identification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6886-6895.

[7] ZHENG Z, YANG X, YU Z, et al. Joint discriminative and generative learning for person re-identification[C]// Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition. 2019: 2138-2147.

[8] WEI L, ZHANG S, GAO W, et al. Person transfer gan to bridge domain gap for person re-identification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 79-88.

[9] RISTANI E, SOLERA F, ZOU R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]// European Conference on Computer Vision. Cham: Springer, 2016: 17-35.

[10] ZHENG L, SHEN L, TIAN L, et al. Scalable person re-identification: A benchmark[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1116-1124.

[11] HOU R, CHANG H, MA B, et al. Temporal complementary learning for video person re-identification[C]// European Conference on Computer Vision. Cham: Springer, 2020: 388-405.

[12] FAN C, PENG Y, CAO C, et al. Gaitpart: Temporal part-based model for gait recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 14225-14233.

[13] LORENZO-NAVARRO J, CASTRILLÓN-SANTANA M, HERNÁNDEZ-SOSA D. An study on re-identification in RGB-D imagery[C]// International Workshop on Ambient Assisted Living. Berlin: Springer, 2012: 200-207.

[14] HAQUE A, ALAHI A, FEI-FEI LI F F. Recurrent attention models for depth-based person identification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1229-1238.

[15] WU A, ZHENG W S, LAI J H. Robust depth-based person re-identification[J]. IEEE Transactions on Image Processing,

2017，26(6)：2588-2603.

[16] JIA X，MENG Z，KATIPALLY K，et al. Clothing Change Aware Person Identification[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE Computer Society，2018.

[17] YANG Q Z，WU A C，ZHENG W S. Person Re-identification by Contour Sketch under Moderate Clothing Change[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2021，3(6)：2029-2046.

[18] LI Y J，LUO Z Y，WENG X S，et al. Learning shape representations for clothing variations in person re-identification[J]. arXiv：2003. 07340，2020.

[19] ZHANG J，LI Y，CHEN F Q，et al. X-Net：A Binocular Summation Network for Foreground Segmentation[J]. IEEE Access，2019，7：1412-71422.

[20] BLAKE R，WILSON H. Binocular Vision[J]. Vision Research，2011，51(7)：754-770.

[21] LOWE D G. Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision，2004，60(2)：91-110.

[22] REDMON J，FARHADI A. Yolov3：An incremental improvement[J]. arXiv：1804. 02767，2018.

[23] XIE S，TU Z. Holistically-nested edge detection[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015：1395-1403.

[24] HU J，LI S，ALBANIE S. Squeeze-and-Excitation Networks [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE，2018.

[25] YE M，SHEN J，LIN G，et al. Deep learning for person re-identification：A survey and outlook[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2021. https：//ieeexplore. ieee. org/abstract/document/9336268.

[26] XIA B N，GONG Y，ZHANG Y，et al. Second-order non-local attention networks for person re-identification[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：3760-3769.

[27] YU F，KOLTUN V. Multi-scale context aggregation by dilated convolutions[J]. arXiv：1511. 07122，2015.

[28] OJALA T，PIETIKÄINEN M，HARWOOD D. A comparative study of texture measures with classification based on featured distributions[J]. Pattern Recognition，1996，29(1)：51-59.

[29] DALAL N，TRIGGS B. Histograms of oriented gradients for human detection[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR'05). IEEE，2005：886-893.

[30] KOESTINGER M，HIRZER M，WOHLHART P，et al. Large scale metric learning from equivalence constraints[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2012：2288-2295.

[31] 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.

CHEN Kun-feng，born in 1995，postgraduate. His main research interests include computer vision and person re-identification.

PAN Zhi-song，born in 1973，Ph.D，professor. His main research interests include pattern recognition and machine learning.

（责任编辑：杨雪敏）