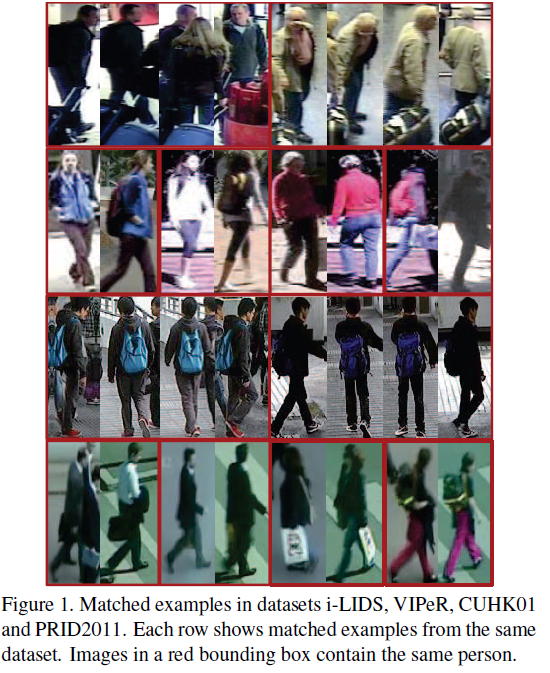


1. CVPR

**摘要**

跨摄像机的行人重新识别一直是一个非常具有挑战性的问题，特别是当摄像机之间没有重叠的视野时。在本文中，我们提出了一种新的基于三重框架的多通道身体部分的卷积神经网络模型，用于人员重新识别。 具体而言，所提出的CNN模型由多个通道组成，以共同学习输入人员的全局全身和局部身体部位特征。该CNN模型通过改进的三元组损失函数进行训练，在学习到的特征空间中，该函数用于拉近同一个人之间的距离，并且同时推开不同人之间的距离。大量的对比评估表明，我们提出的方法在具有挑战性的i-LIDS，VIPeR，PRID2011和CUHK01数据集方面明显优于许多最先进的方法，包括传统和深层网络方法。



**1. 引言**

行人再识别是多个摄像机之间匹配相同的个体或者单个摄像机内跨时间匹配的问题。由于其对视频监控、人机交互、机器人、基于内容的视频检索等许多应用的重要性，它在计算机视觉和模式识别研究界快速引起了越来越多的关注。

尽管经过多年的努力，由于以下原因，行人再识别仍然具有挑战性：（1）来自不同相机的不同视角引起的视觉外观和周围环境的显着变化;（2）行人在不同时间不同地点，显著的姿势变换；（3）背景杂乱和遮挡；（4）具有相似外观的不同个体。此外，在看不到或者看不清脸的时候，在许多情况下使用生物识别和软生物识别方法是不适用的。图1举例说明了在四个具有挑战性的行人再识别基准数据集i-LIDS [38]，VIPeR [13]，PRID2011 [17]和CUHK01 [24]中匹配对的一些例子。每个红色边界框中的图像来自同一个人。

给定查询人的图像，为了在由不同相机捕获的大量候选图像中找到正确的匹配，必须解决两个关键问题。首先，需要较好的表示出查询图像以及测试集中图片的特征。其次，合适的距离度量对于确定测试集中是否包含与查询图像相同的图像是必不可少的。许多现有的研究分开考虑这两个问题，并且更多地关注第一个问题，即开发更有辨别力和鲁棒性的特征表示来描述一个人的视觉外观。一旦特征提取完成，这些方法通常选择一个标准的度量方法例如基于L1范式的度量方法、Bhat距离度量、马氏距离度量等判断图片之间的相似性。

这种情况促使我们共同考虑特征和距离度量学习问题，以提高行人再识别的表现。为了为原始人物图像提取更好的特征，我们提出了一种新的多通道的卷积神经网络模型，该模型可以从输入人的全身和身体部位学习特征。全身和身体部位的特征连接在一起并送入最后的全连接层，以产生输入行人的最终表示。我们也借鉴了Wang等人的想法和FaceNet的工作使用三元组训练样本和改进的三元组损失函数来进一步增强学习特征的判别力。与仅要求类内特征距离小于类间特征距离的原始三元组损失函数相比，改进的损失函数还要求类内特征距离小于预定义的余量。我们的实验评估表明，与使用相同DCNN模型的原始三重损失函数相比，改进的三元组损失函数准确率提高了4%。

给定一张行人的图片，提出的卷积神经网络模型优于用800维特征去表示一张图片。所提出的CNN模型连同改进的三元组损失函数可以被认为是学习一个映射函数，该映射函数将每个原始图像映射到特征空间中，其中同一个人的图像之间的差异小于不同人的图像之间的差异。因此，所提出的框架可以为行人再识别任务共同学习最优的特征和距离度量。

本文的主要贡献有两个：1）一种新颖的多通道CNN模型，它学习全局全身和局部特征，并将它们集成在一起，以产生输入人的最终特征表示;2）改进的三元组损失函数，其要求类内特征距离不仅小于类间特征距离，还要小于预定阈值。实验评估结果表明，该方法在几个广泛采用的行人基准测试数据集上实现了最好的性能。

1. **相关工作**

有代表性的行人再识别系统包含两个主要的成分：用于描述检测图片以及测试集的特征提取的方法，以及用于比较图片中这些特征的距离度量。对行人再识别问题的研究通常侧重于构建鲁棒性和判别性特征，或者找到用于比较特征的改进的相似性度量，或两者的组合。

有大量的研究工作致力于发现对光照、姿势和视点变化至少部分不变的更好的特征表示。已经用于行人再识别任务的特征包括颜色直方图及其变形[41, 20, 21, 28, 23, 46], 局部二值模式（LBP）[41,20,21,28,23,46]，Gabor特征[23]，颜色名称[44]和其他视觉外观或情境线索[3]。相当一些作品还研究了多种视觉特征的组合，包括[41,20,23]。

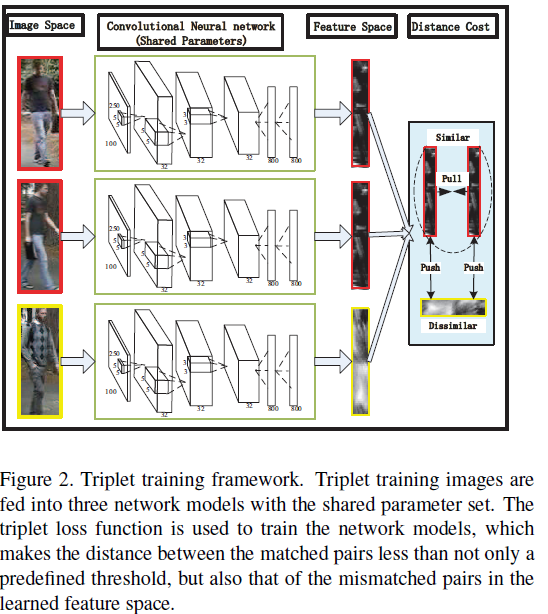
大量的度量学习和排序算法已经应用于行人再识别问题[43,31]。度量学习背后的基本思想是找到从特征空间到具有某些优点的距离空间的映射函数，例如来自同一个人的特征向量比来自不同人的特征向量更接近。这些度量学习方法主要包括Mahalanobis度量学习（KISSME）[21]，局部Fisher判别分析(LFDA) [41], 边际Fisher分析(MFA)[41],大边缘最近邻(LMNN)[41],局部自适应 决策函数(LADF) [26],和属性一致性匹配[20].

受到深度学习网络在各种计算机视觉和模式识别任务中取得巨大成功的启发[22,11,36,37,16]，将深度卷积神经网络（DCNN）模型应用于行人再识别问题变得越来越流行。值得注意的是，在被广泛使用的行人再识别数据集上，如i-LIDS，VIPeR，CUHK01等数据集上，近期取得的最好的表现都是应用深度卷积神经网络的网络。在下文中，我们将简要介绍与我们的工作相关或与之相比的基于深度学习的方法。王等人[39]使用三元组训练样本和三元组损失函数来学习细粒度图像相似性度量。FaceNet [34]和丁等人[6]分别将这种三元组框架应用于人脸和行人再识别问题。在本文中，我们还借鉴了[39]中的想法，并为行人再识别任务提出了改进的三元组损失函数。DeepReID [25]提出了一种新型的滤波器配对神经网络（FPNN），通过使用块匹配层来匹配跨视图的局部块的滤波器响应，以及其他卷积和最大池化层来模拟身体部位的位移，共同处理未对准，光度和几何变换，遮挡和黑色簇等问题。mFilter [48]还使用了局部块匹配方法，该方法学习中级过滤器以为行人再识别任务获得局部判别特征。艾哈迈德等人[1]提出了一种改进的深度学习架构，它采用成对图像作为输入，并输出表示两个输入图像是否描绘同一个人的相似度值。模型中的新元素包括计算交叉输入邻域差异以根据其中级特征捕获两个输入图像之间的局部关系的层，以及用于获取高级特征的块汇总层。Yi等人[45]构建了一个siamese神经网络（在我们的论文中表示为DeepM）来学习成对相似性，并且还使用身体部位来训练他们的CNN模型。在他们的工作中，人物图像被裁剪成三个重叠的部分，用于训练三个独立的网络。 最后，三个网络在得分处融合。

我们的卷积神经网络模型在网络结构和损失函数与上面说到的深度网络的方法不同。更具体地说，我们使用由多个通道组成的单一网络来学习全局全身和局部身体部分特征。我们在不同类型的通道中使用不同的卷积核大小来查看具有不同分辨率的全身和身体部分，这类似于DPM模型中的根/部分滤波器的想法[9]。此外，我们使用改进的三重损失函数使得同一个人的特征更接近，同时来自不同人的特征更远。在第4节中，与上述一些方法进行性能的比较将在我们的实验评估中。

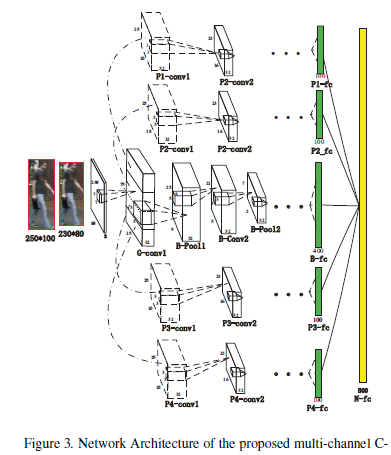
1. **The Proposed Person Re-Id Method**

在本节中，我们将详细介绍提出的行人再识别方法。我们首先描述我们的行人再识别方法的总体框架，然后详细阐述所提出的多通道卷积神经网络模型的网络架构。最后，我们提出了改进的三元组损失函数，用于训练所提出的卷积神经网络模型。



**3.1 The Overall Framework**

如图2所示，类似于[39,34]中的工作，所提出的行人再识别方法使用三元组示例来训练网络。表示由三张图片组成的三元组，其中，和表示相同的人，表示不同的人。三个卷积神经网络通过共享参数，如权重共和方差，我们将三元组从原始的图像空间映射到一个学习到的特征空间，其中表示为。图中的每个卷积神经网络是提出的多通道卷积神经网络模型，其能够提取全局全身和局部身体部分特征。当使用改进的三元组损失函数训练所提出的卷积神经网络模型时，学习的特征空间和之间的距离不仅小于和，也是预定义的边界值。改进的损失函数旨在学习道德特征空间中拉近同一个人的距离，同时推动属于不同人的距离彼此更远。



**3.2 Multi-Channel Parts-based CNN model**

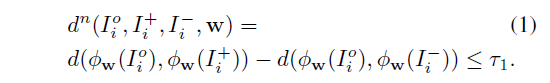
所提出的多通道卷积神经网络模型主要由以下不同层组成：一个全局卷积层，一个全身卷积层，四个身体部分的卷积层，五个通道全连接层和一个网络全连接层。如图3所示，全局卷积层是所提出的卷积神经网络模型的第一层。它由32个特征图组成，卷积核为7×7×3，步长为3像素。接下来，该全局卷积层被分成四个相等的部分，i = {1，...，4}，每个部分形成独立的身体部分通道的第一层，旨在学习相应身体部位的特征。还建立了以整个全局卷积层作为其第一层的全身通道，以学习输入行人的全局全身特征。四个身体部分通道与全身通道一起构成五个独立的通道，它们彼此分开训练。

全身通道配置如下：全局卷积层，最大池，全身卷积层，另一个最大池和全连接层。最大池化层的核函数大小为3\*3，全连接层生成的是400维。四个身体部分通道具有如下相同的配置：全局卷积层的四个等分部分之一，身体部分卷积层，无最大池和全连接层。全连接层生成100维的输出。由于全身卷积层和四个身体部分卷积层分别旨在学习全局全身和局部身体部位特征，我们前者用到的卷积核为5\*5，后者用到的卷积核为3\*3。这有助于为行人身体部分学习细粒度的局部特征。两种类型的卷积层都使用1的步幅。请注意，我们的卷积神经网络模型中的所有卷积层都包含一个用于生成其输出的relu层。

上述网络配置在相对较小的基准数据集上实现了最先进的行人精确度。在我们的实验中，我们发现对于一些较大的数据集，例如CUHK01，用两个卷积层构造五个独立通道中的每一个导致更好的结果。因此，我们分别使用两种网络配置来处理小型和大型基准数据集。除了每个单独通道中的卷积层数（一个或两个）之外，两个网络配置大致相同。在最后阶段，来自五个独立通道的全连接层的输出被连接成一个矢量，并被一起输送到最后的网络全连接层。上述多通道结构使得能够共同学习全局全身和局部身体部位特征，并且在最后阶段这两种类型特征的融合导致行人再识别准确率的显着改善。

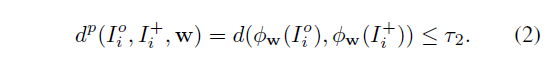
**3.3 Improved Triplet Loss Function**

如3.1节所述，我们用了三元组的样本去训练网络模型。给了一个三元组，网络模型映射到一个学习到的特征空间。三元组图像，，之间的相似性通过计算，，之间的L2范式的距离来度量的。原始的三元组损失函数要求之间的距离大于之间预定的边距，并且用以下的等式来强制这个要求：



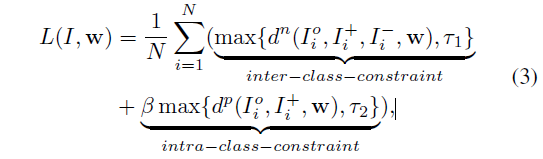
在等式中，是负的。然而，由于该损失函数没有规定对应该有多接近，因此，属于同一个人的实例可以在学习的特征空间中形成具有相对大的平均类内距离的大集群。显然，这不是一个理想的结果，并且不可避免地会伤害到行人再识别的表现。

基于以上的观察，我们在原始的三元组损失函数中加入了一个新的约束，以进一步要求的距离小于第二余量，并且远比。将次陈述翻译成等式，我们得到：



改进的损失函数目的在于，在学习到的特征空间中，拉近同一个人的图片，推动属于不同人的图片彼此更远。这与许多数据聚类和判别分析方法使用的原理更加一致。

总而言之，改进的三元组损失函数定义如下：



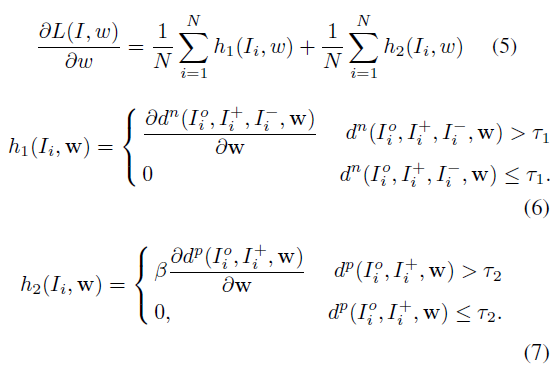
其中N是三元组训练样本的数量，是用于平衡类间和类内约束的权重。在我们的实现中，距离函数被定义为L2范数的距离，



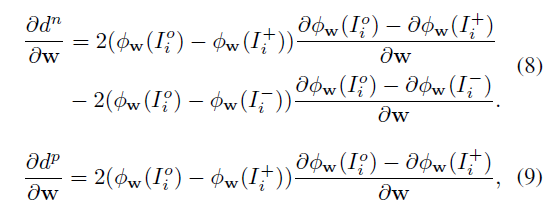
**3.4 The Training Algorithm**

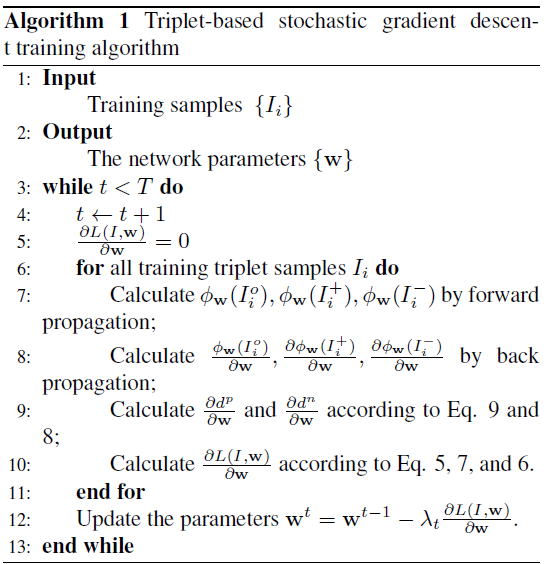
我们使用随机梯度下降算法来训练所提出具有改进的三元组损失函数的卷积神经网络架构模型。

方程（3）的导数可以如下计算：



通过对和，我们可以得到他们的梯度如下，



从上面的推导中可以清楚地看出，对于三元组示例中的每个图像，给定，，和，，的值可以很容易地计算每个输入三元组的梯度，这可以通过分别运行标准的前向和后向传播来获得。由于算法需要遍历每批中的所有三元组，以累积每次迭代的梯度，我们称之为基于三元组的随机梯度下降算法。算法1显示了训练算法的主要过程。

1. **Experiments**

**4.1 Setup**

**数据增强**：数据增加是增加训练数据量和缓解过度拟合问题的重要手段。 在我们的实现中，我们将所有图像调整为100×250像素。 在训练过程中，我们裁剪80×230像素的中心区域，每个图像有一个小的随机扰动，以增加训练数据。

**训练参数设置：**权重从两个零均值高斯分布初始化，标准偏差分别为0.01和0.001。偏差项设置为0。我们按如下方式生成三元组：对于每批100个实例，我们选择5个人，并在每次迭代中为每个人生成20个三元组。在每个三元组中，匹配的参考是从同一类中随机选择的，并且不匹配的参考也是随机选择的，但是来自其余的类。在我们的实验中，方程（3）中的参数，，β分别设置为-1,0.01和0.002。

**数据集：**我们使用四个常用的行人再识别基准数据集，i-LIDS，PRID2011，VIPeR和CUHK01，进行效果评估。所有数据集都包含一组人员，每个行人都有几张由不同摄像头捕获的图像。以下是这四个数据集的简要说明：

1. LIDS数据集：它是通过拍摄繁忙的机场到达大厅的视频图像构建的。 它包含来自119个人的479张图像，其被标准化为128×64像素。 每个人平均有四张图片。 这些图像由非重叠视域相机捕获，并受到大的光照变化和遮挡的影响。

PRID2011数据集：该数据集由两个静态监控摄像机记录的图像组成。 摄像机视图A和B分别包含385和749人，其中有200人在两个摄像机下出现过。

VIPeR数据集：此数据集包含632人的两个视图。每一对人的图片由具有不同视点，姿势和光照条件的不同相机捕获。由于其巨大的差异和差异，它是行人再识别任务最具挑战性的数据集之一。

CUHK01数据集：这是行人再识别任务的一个更大的数据集，其中包含从校园环境中的两个摄像机视图捕获的971个人。摄像机视图A捕获人的正面或背面视图，而摄像机B捕获行人的侧面视图。每个人有四张图像，其中每个摄像头各拍摄两张。

**评估协议** 我们采用广泛使用的累积匹配曲线（CMC）度量进行定量评估。对于每个数据集，我们随机选择大约一半的人进行训练，剩下的一半用于测试。对于具有两个摄像机的数据集，我们从摄像机A随机选择一个人的图像作为查询图像，并从摄像机B随机选择一个图像作为Gallery图像。对于多摄像机数据集，选择同一个体的两个图像：一个用作查询，另一个用作Gallery图像。Gallery集包括每个人的一个图像。对于query中的每个图像，我们首先使用L2距离计算训练网络生成的特征，计算query与所有gallery之间的距离，然后返回gallery中的前n个最近图像。如果返回的列表，包含与第k个位置图片相同的人的图像，则该查询被认为是rank-k的成功。我们重复该过程10次，并取平均作为评价结果。

**4.2 Experimental Evaluations**

我们提出的行人再识别方法包含两个新颖的成分：1）多通道卷积神经网络模型，能够学习全局全身和局部身体部位特征，2）改进的三元组损失函数，在学习的特征空间中，用于拉近同一个人的图片，同时将属于不同人的图片推开。为了揭示每种成分如何有助于提高性能，我们对所提出的行人再识别方法，进行了以下四种变形，并将它们与文献中的十几种代表性方法进行了比较：

变体1（表示为OursT）：我们从提出的卷积神经网络模型中移除提取四个身体部分的通道，并使用原始的三元组损失函数来训练网络。

变体2（表示为OursTC）：我们使用与OursT相同的网络模型，但使用改进的三元组损失函数来训练网络。

变体3（表示为OursTP）：我们使用所提出的多通道卷积神经网络模型的完整版本，并使用原始的三元组损失函数训练它。

变体4（表示为OursTPC）：我们使用与OursTP相同的网络模型，但使用改进的三元组损失函数训练它。

请注意，由于CUHK01数据集比其他三个数据集大得多，我们选择使用更大的设置对其进行建模，即在五个通道中都添加一个额外的卷积层。对应于变体1到4的导出模型分别表示为Ours3T，Ours3TC，Ours3TP和Ours3TPC。

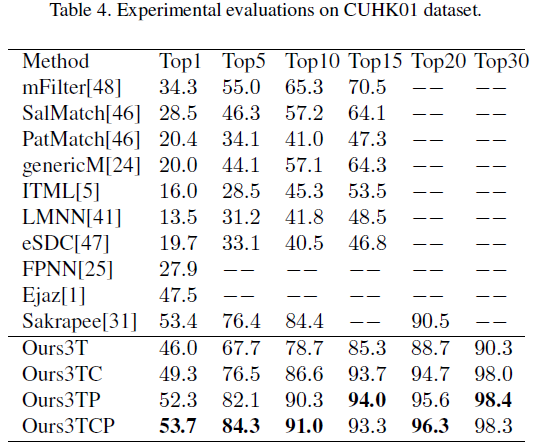
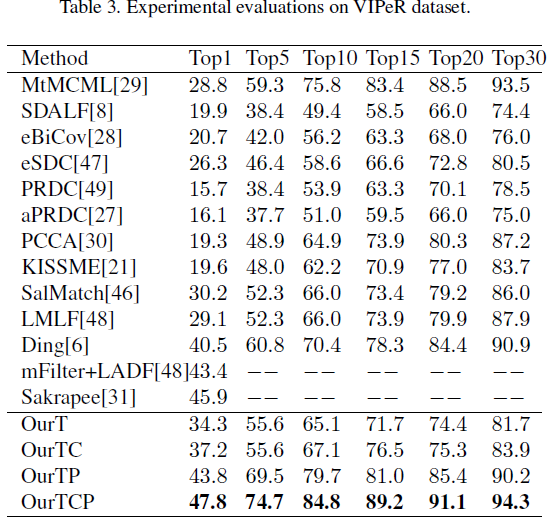
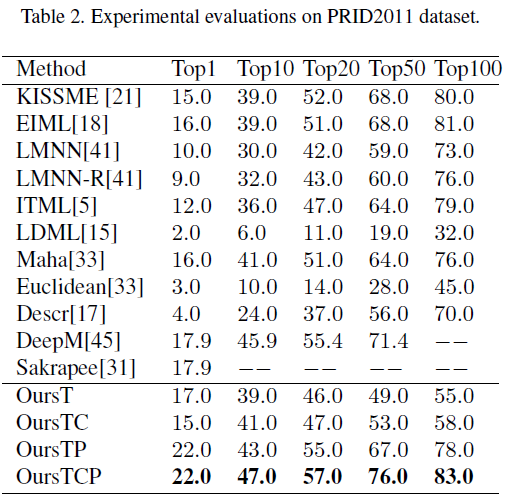
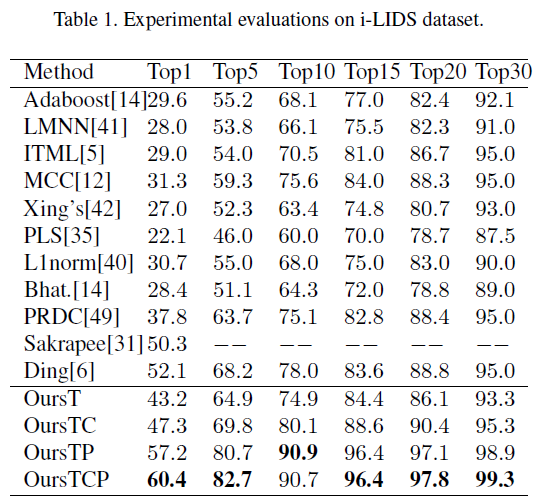
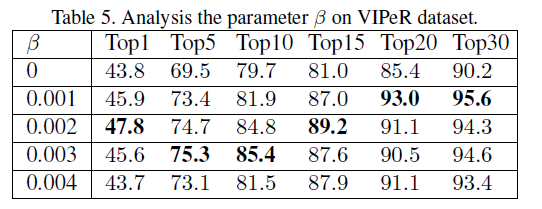


表1,2,3和4分别显示了在四个基准数据集的评估结果，使用排名前1,5,10,15,20和30的排名精度。每个表包括11到14个代表性方法，并且展示了这些方法在相应数据集的评估结果。这些表中的一些工作，如Ding的方法[6]，FPPN [25]，DeepM [45]，mFilter [48]和Ejaz的[1]都使用深度卷积神经网络模型来为行人再识别任务学习特征，他们的表现准确度接近榜单的前列。在这些工作中，DeepM还使用身体部位来训练他们的CNN模型。与我们具有多个频道的单个网络相比，这项工作将人物图像划分为三个重叠部分，并使用它们来训练三个独立的网络。三个网络将最后的得分融合。还有一些作品，如Sakrapee的方法[31]，mFilter + LADF [48]，它结合了几种不同的方法来提高性能的准确性。到目前为止，这些集合方法已经达到了最高的性能。与上述代表性作品相比，OursTCP模型已经在所有四个数据集上取得了最佳表现。四个表中显示的评估结果可总结如下。

•与Sakrapee基于集合的方法相比，这是迄今为止最先进的方法，OursTCP模型在CUHK01数据集上微微优于前者，但在其余三个数据集上的表现明显优于前者，在2％到10％之间。

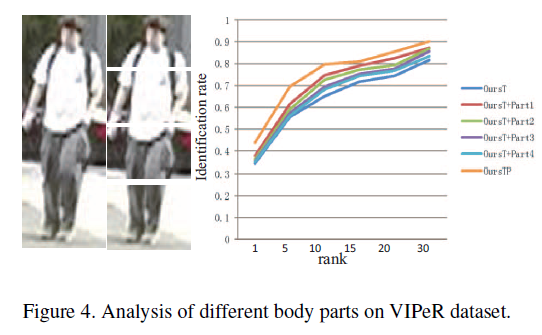
•改进的三元组损失函数可以提高单通道和多通道模型的性能准确度。 与使用原始三元组损失函数训练的相同模型相比，使用此损失函数训练模型可以使性能提高4％。

•探索全局全身和局部身体部位特征的多通道模型非常强大且有效地提高了性能准确性。与结构中没有身体部分信息的模型相比，它可以将行人再识别的精度提高多达13％。

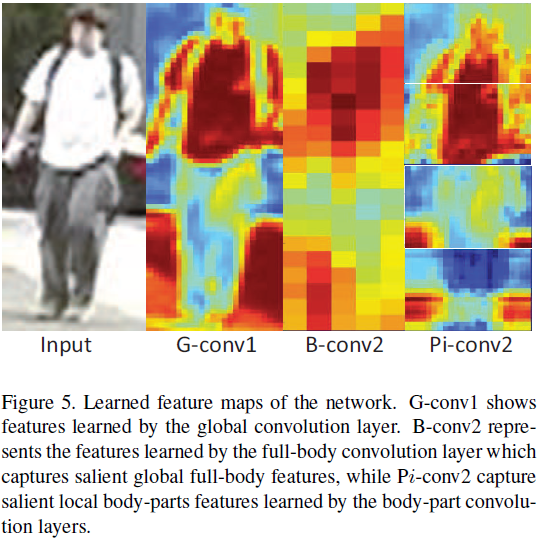


如公式（3）所定义，改进的三元组损失函数包含两个项：类内和类间约束。为了研究参数β对性能准确性的影响，我们在VIPeR数据集上使用交叉验证方法进行了实验，结果如表5所示。我们可以清楚地看到，当β在0.001到0.003的范围内时，我们提出的行人再识别的方法达到最佳性能。基于此观察，我们在所有实验评估中将β设置为0.002。

**4.3 Analysis of different body parts**



为了理解不同身体区域对行人再识别性能准确性的贡献，我们训练了四种不同的网络模型，其中包含全身通道和一个身体部分的通道，分别对应于身体部分1,2,3和4。这四个模型分别表示为Ours-Part1，Ours-Part2，Ours-Part3和Ours-Part4。我们还将OursT和OursTP的模型包括在内进行比较。实验在VIPeR数据集上进行，性能准确度如图4所示。有趣的是观察到包括人的面部和肩部的Ours-Part1得到了最好的性能表现。当我们向下移动身体时，性能改善逐渐降低，Ours-Part4（包括人的腿和脚）得到了最小的性能表现。这个结果并不令人惊讶，因为腿和脚是人的运动部分，其形状和姿势会发生显着变化。这些部件提供了最少的可靠特征，因此对行人再识别任务贡献很小。



我们已经可视化每个卷积层学到的特征，如图5所示。我们可以看到全身通道的第二个卷积层捕获每个人的全局信息，而四个身体部位通道的第二个卷积层捕获人的详细的局部身体部位特征。因此，全局全身和局部身体部位特征的这种联合表示和学习框架可以达到卓越的性能。

1. **Conclusion**

在本文中，我们提出了一种新的基于多通道部件的卷积网络，用于行人再识别问题，通过改进的三元组损失函数在三重框架下制定。在此框架中，我们构建了卷积神经网络架构，包括全局身体卷积层和局部部分身体卷积层。因此，我们模型学习的特征表示可以包含全局信息和局部细节属性。该体系结构通过一系列的三元组训练，目的是在将同一个人的样本拉近，同时通过有组织的三元组样本在学习到的特征空间中推动属于不同人的样本。我们的模型在大多数基准数据集上都具有最先进的性能。 将来，我们会将框架和方法扩展到其他任务，例如图像和视频检索问题。