基于时序多尺度互补特征的视频行人重识别

侯瑞兵[1),2)] 常 虹[1),2)] 马丙鹏[2)] 黄 锐[3)] 山世光[1),2)]

[1)](中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

[2)](中国科学院大学计算机科学与技术学院 北京 100049)

[3)](香港中文大学（深圳）理工学院 广州 518172)

摘 要 视频行人重识别在监控场景中起着非常重要的作用. 但是, 大多数现有方法没有充分利用行人视频序列的时空信息. 具体来说，这些方法以相同的分辨率和网络结构处理每一帧图像，造成连续帧特征的高度相似. 此外，现有方法通常通过引入各种复杂的操作提高精度，过多的计算开销使其不利于真实场景的部署. 针对上述问题，本文提出了一个时序多尺度互补网络，旨在高效地为视频的连续帧提取互补的特征. 具体来说，时序多尺度互补网络包含多个具有不同输入分辨率的分支. 其中，高分辨率分支处理原始分辨率帧，用于保留行人的细节线索；低分辨率分支处理以不同降采样率得到的低分辨率帧，用于捕捉更全局的行人信息. 通过将连续帧输入到不同分支中，连续帧能关注不同粒度的空间区域，生成互补的特征. 进一步，设计了一个多分支批量归一化层，保证了训练时分支之间的互补性. 最后，提出一个跨分支融合模块，将低分辨率分支的全局信息逐步传播到高分辨分支中，得到一个融合了多尺度全局粗粒度和局部细粒度互补信息的特征. 在 iLIDS-VID, MARS 和 LS-VID 三个数据集上的实验显示，本文提出的方法达到了比目前最好方法更好的性能，例如，在 LS-VID 上提升了 4.5% mAP 和 3.1% top-1 精度，证明了该方法的有效性. 此外，通过降低输入帧的分辨率和使用更小的网络处理低分辨率帧，本文方法大幅度降低了计算开销，仅需要大多数现有方法约 35%的计算开销.

关键词 视频行人重识别；多分支架构；多尺度特征表示

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2023.00031

Temporal Multi-Scale Complementary Feature for Video Person Re-Identification

HOU Rui-Bing[1),2)] CHANG Hong[1),2)] MA Bing-Peng[2)] HUANG Rui[3)] SHAN Shi-Guang[1),2)]

[1)](*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences Beijing* 100190)

[2)](*School of Computer Science and Technology, University of Chinese Academy of Sciences Beijing* 100049)

[3)](*School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen Guangdong* 518172)

Abstract Person re-identification (reID) aims to match the same person across multiple non-overlapped cameras, which plays an important role in surveillance video analysis. Recently, with the emergence of large video benchmarks and the growth of computational resource, video-based person reID has been attracting a lot of attention. However, most existing methods do not make full advantage of rich spatial and temporal clues in the videos. To be specific, the consecutive frames of a pedestrian video are highly similar, but the existing methods conduct the same operation with the same input resolution on each frame. As a result, the existing methods typically produce highly redundant features for consecutive frames. The

收稿日期：2021-12-08；在线发布日期：2022-06-09. 本课题得到国家自然科学基金面上项目（61976203，61876171）资助. **侯瑞兵**，博士研究生，主要研究领域为计算机视觉、模式识别、机器学习. E-mail: ruibing.hou@vipl.ict.ac.cn. **常 虹**，博士，研究员，中国计算机学会（CCF）会员，主要研究领域为计算机视觉、模式识别、机器学习. **马丙鹏**（通信作者），博士，副教授，中国计算机学会（CCF）会员，主要研究领域为计算机视觉、行人重识别. E-mail: bpma@ucas.ac.cn. **黄 锐**，博士，副教授，中国计算机学会（CCF）会员，主要研究领域为计算机视觉、行人重识别. **山世光**，博士，研究员，中国计算机学会（CCF）会员，主要研究领域为计算机视觉、模式识别、机器学习.

redundant frames easily focus on the same most representative local part, which may be difficult to distinguish the persons with seemingly similar local part. In addition, it is common for recent methods to improve the accuracy by introducing more operations, which is not conducive to deployment in many real scenes. In this paper, we present a powerful and efficient video person reID network, Temporal Multi-Scale Complementary Network (TMSCN), to extract complementary features for consecutive frames of a video efficiently. Specifically, TMSCN is built on a multi-branch architecture where each branch has a specific input resolution. High-resolution Branch processes frames at original resolution to preserve the detailed visual clues, and Low-resolution Branches with different down-sampling rates are employed to progressively enlarge the receptive field for capturing global information. By feeding consecutive frames into different branches, TMSCN can enforce consecutive frames to focus on the regions with different spatial scales. Such that the diverse visual features can be discovered for consecutive frames and finally form an integral characteristic of the target identity. Furthermore, since low-resolution images contain fewer details, a small network is sufficient to extract its discriminative features. While high-resolution images contain richer details and require a larger network to process. So TMSCN uses network structures of different capacities for different branches, which can reasonably allocate the computational resources. TMSCN also shares the parameters of the same stage for different branches, which needs no extra parameters over single-branch video reID networks. Furthermore, we design a Multi-Branch Batch Normalization (MBBN) which uses separated mean and variance for BN layer on each branch. MBBN can ensure stable training and maintain the complementarity among branches. At last, a Cross-Branch Fusion Module (CBPM) is proposed to successively propagate the global and coarse features of low-resolution branch to high-resolution branch. In this way, the final multi-scale feature contains both local details and global complementary information, showing stronger discrimination ability. Experiments on three benchmarks, MARS, LS-VID and iLIDS-VID, show that the proposed method achieves state-of-the-art performance. Specifically, it improves the performance of existing video reID methods by about 4.5% mean Average Precision (mAP) and 3.1% top-1 accuracy on LS-VID benchmark, validating the effectiveness of the proposed method. Moreover, by down-sampling some frames to low-resolution and using small network for low-resolution frames, TMSCN greatly reduces the computations, requiring about 65% less computation cost than most existing methods.

Keywords Video person re-identification; Multi-branch architecture; Multi-scale feature representation

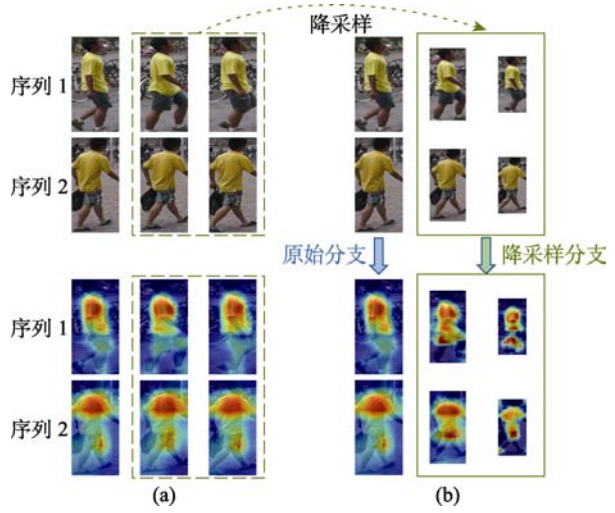
1 引 言

行人重识别（Person Re-identification, ReID）是指在多个非重叠摄像头的场景下，给定目标行人图像/视频，从大规模行人图像/视频库中检索出具有相同身份的行人. 由于不受识别距离和环境约束的特点，行人重识别可以广泛应用于视频监控、智能安防等领域. 因此，行人重识别研究具有很好的应用价值.

行人重识别根据数据源的不同分为基于图像的行人重识别[1,2]和基于视频的行人重识别[3,4]. 其中基于图像的行人重识别的样本为单帧图像；基于视频的行人重识别的样本为图像序列. 相比于单帧图像，视频数据具有天然的优势. 首先，视频数据具有更广泛的应用场景. 单帧的方法要求图像质量很高，这对相机的布置和使用的场景是非常大的限制. 而视频数据包含更丰富的信息，对图像质量要求相对较低，可以应用在复杂的监控场景中. 其次，视频数据包含更丰富的行人表观和时序信息. 一方面，行人视频序列的不同帧可能包含行人的不同姿态和视角，有助于建模丰富的行人表观特征. 另一方面，视频序列包含丰富的时序信息，有助于建模行人的动作特征，从而区分表观相似的行人. 基于以上优势，基于视频的行人重识别受到了研究者越来越多的关注.

近几年，随着深度学习的发展[5-7]，基于深度网

络的视频行人重识别方法[8-11]的性能不断提升，远优于传统的方法. 目前主流的基于深度学习的方法通常采用一个统一的框架：给定输入的行人视频序列，首先利用卷积神经网络独立地提取每一帧图像的特征，然后通过池化[4]或者时序注意力机制[12,13]融合所有帧的特征生成一个视频级别的行人特征表示. 通过使用强大的神经网络[14]以及大规模的标注数据集[4,15],这些基于深度学习的方法能够取得优于传统方法的性能.



（a）输入序列和现有方法[4]得到的特征激活图[16]（连续的三帧都关注到同一个局部区域）　（b）输入序列和我们的方法得到的特征激活图（通过将第二帧和第三帧以不同降采样率执行降采样操作，连续的三帧能关注到不同粒度的空间区域，联合这些互补的区域更容易区分这两个相似的行人）

图 1　结果比较

尽管视频行人重识别技术取得了很大的进展，但是大多数现有方法仍然没有充分利用视频中的时空信息. 具体来说，这些方法以相同的分辨率和网络结构处理每一帧图像，造成连续帧特征的高度相似和冗余. 这些高度相似的特征趋向于关注同一个显著但是局部的区域[9,16],导致其很难区分局部相似的不同行人. 如图 1（a）所示，在现有方法中，视频序列的连续帧都关注到黄色上衣区域,但是这个局部区域却难以区分图中的这两个行人. 因此，对于视频行人重识别任务,对视频的连续帧挖掘互补的视觉线索形成一个更丰富的完整行人表示是至关重要的.

此外，现有视频行人重识别方法通常引入各种复杂的操作来提高精度. 例如，VRSTC[17]引入了一个复杂的图像补全网络处理遮挡序列，增加了大约一倍的计算量；MGH[18]引入了一个多尺度图卷积网络建模时序信息，增加了大约 10%的计算开销. 但是，许多设备，例如机器人、自动驾驶汽车和智能手机这样的移动设备无法部署耗电的强大 GPU,因此有限的计算资源阻碍了先进视频行人重识别模型[11,18,19]的实际应用. 这启发我们开发一个高效的视频行人重识别模型，以尽可能少的计算开销充分挖掘行人序列的时空信息.

为了实现上述目标,本文提出了一种新型的视频行人重识别网络-时序多尺度互补网络（Temporal Multi-Scale Complementary Network, TMSCN）. 由于降低图像的分辨率能够有效提升网络在原始图像的感受野，从而关注到一个更全局的区域[20]，并且使用低分辨率的输入帧能够有效降低计算开销. 因此 TMSCN 通过以不同的分辨率处理连续的帧，达到高效互补建模的目的. 具体来说，TMSCN 包含多个具有不同输入分辨率的分支. 其中，高分辨率分支处理原始分辨率的帧，用于保留行人的细节视觉特征；低分辨率分支处理以不同降采样率得到的低分辨率帧，用于增大网络的感受区域从而捕捉到更全局的内容. 通过将视频的连续帧输入到不同的分支，TMSCN 能够使连续帧关注到不同尺度的空间区域，从而提取到帧间互补的特征. 如图 1 (b) 所示，通过降采样操作，连续的三帧关注到从局部到全局的互补区域（第一帧关注黄色上衣；第二帧关注黄色上衣 + 棕色/迷彩裤子，第三帧关注黄色上衣 + 棕色/迷彩裤子 + 运动鞋/拖鞋），从而更容易区分这两个相似的行人. 此外，Yang 等人[21]指出，低分辨帧含有较少的细节信息，使用一个小网络便足够提取其判别性信息，过大的网络反而增加了过拟合的风险；而高分辨率帧含有丰富的细节信息，需要一个更大的网络进行处理. 因此，TMSCN 提出在更低分辨率分支上使用更小的网络结构，更高效地为其提取判别性特征.

为了进一步提高帧间互补建模的能力，本文设计了一个多分支批量归一化层（Multi-Branch Batch Normalization, MBBN）. 具体来说，TMSCN 的最终目标是使不同的分支提取互补的特征，即不同分支上的特征应该具有不同的均值和方差. 而原始的批量归一化层（Batch Normalization, BN）会整合所有分支的特征更新均值和方差，并且使用这个混合的均值和方差进行归一化. 这导致对于特定分支来说，归一化使用的均值和方差并不是该分支上特征分布的均值和方差，使特征归一化出现偏差，扰乱了训练进程. 因此，MBBN 提出对每个分支使用独立的均值和方差，使每个分支能够正常训练，保证

了分支的互补性. 最后，设计了一个跨分支融合模块（Cross-Branch Fusion Module, CBPM），将低分辨率分支上的全局粗粒度特征传播到相邻的高分辨率分支中. 通过逐级跨分支融合，最终生成了一个整合所有分支信息的特征. 这个特征包含了行人多尺度的全局粗粒度和局部细粒度的互补信息，具有更强的判别性.

本文的方法在 iLIDS-VID[3], MARS[4]和 LS-VID[22]三个常用的视频行人重识别数据集上进行了对比实验，在多个指标上本文方法都达到了与已有方法同样优秀甚至更好的性能. 此外，通过对部分帧使用更低的分辨率和更小的网络进行处理，本文的方法大幅度降低了计算开销，仅仅需要大多数现有方法约 35%的计算开销.

本文第 2 节对已有的视频行人重识别方法进行综述，并讨论本文方法与已有方法的区别；第 3 节和第 4 节对本文方法进行详细的介绍；第 5 节在通用的大规模视频行人重识别数据集上进行了对比实验，验证本文方法的有效性；第 6 节对本文工作进行总结和展望.

2 相关工作

行人重识别旨在实现跨摄像头的目标检索，基于图像的行人重识别[2,23]已经被广泛研究，最近研究者开始关注基于视频的行人重识别[4,8]. 早期的视频行人重识别较多使用手工特征[24,25]，比如方向梯度直方图. 但是人工设计的特征严重依赖于设计者的先验知识和手工调参，很难适用复杂场景下的大规模任务. 近几年，以卷积神经网络[26]为代表的深度学习被广泛应用在视频行人重识别领域. 相比于人工设计的特征，深度网络能够自动从大数据中学习到特定的特征，减少了对特征工程的依赖，并且更适用于复杂的应用场景. 目前，基于深度学习的视频行人重识别方法已经在性能上大大超越了传统的方法. 基于深度学习的视频行人重识别方法一般可以分为两类：基于集合建模的方法[4,12,13,27]和基于时空建模的方法[8,28,29].

（1）基于集合建模的视频行人重识别. 基于集合建模的方法把视频中的每一帧看作独立的图像，用无序的图像集合表示视频序列. 这类方法一般可以整合到一个网络框架中，给定输入的行人图像序列，首先利用卷积神经网络独立提取每一帧的特征，然后将这个特征集合整合成一个统一的行人表示. Zheng 等人[4]使用时序的平均或者最大池化整合所有帧的特征. Li 等人[12]提出加权融合策略，使用一个质量估计模块预测每帧的分数，并进行加权求和得到最终的视频特征. Song 等人[13]和 Li 等人[27]进一步考虑局部特征的加权融合. 基于集合建模的方法可以直接使用已有的图像重识别模型，具有很强的适用性. 但是这类方法忽略了视频的时序关联，限制了视频特征建模能力.

（2）基于时空建模的视频行人重识别. 基于时空建模的方法保留了视频的时序关系，利用时序信息增强行人的表观特征或者提取运动特征. 早期的研究工作[8,30]经常采用光流编码相邻帧的短时动作信息. 但是手工提取光流的算法复杂度很高，计算量十分巨大. Maclaughlin 等人[8]第一次将循环网络（Recurrent neural network, RNN）用于视频行人重识别任务，将所有帧提取的卷积特征送入到 RNN 建模时序关系. 但是 Zhang 等人[31]发现打乱 RNN 输入帧的时间顺序后甚至能得到比正常输入更高的性能，说明循环网络可能无法很好地对序列信息进行有效建模. Liao 等人[28]提出使用三维卷积（3D Convolution）[32]提取行人视频的时空特征. Li 等人[29]提出多尺度三维卷积层，利用多个不同尺度的时序卷积同时捕捉行人序列的短时和长时时序关系. 最近，一些研究[19,22,32]采用非局部模块[33]或者图网络[34]建模行人序列的长时时序关联，取得了优越的性能.

基于帧间互补建模的视频行人重识别. 上述的方法都对每一帧执行相同的操作，由于相邻图像帧的高度相似性，卷积网络对连续帧提取的特征是高度相似的. 针对这个问题，一些工作提出利用时序信息对连续的视频帧提取互补的特征，丰富行人的特征表示. Hou 等人[9]提出了时序显著性擦除模块（Temporal Saliency Erasing, TSE）. 给定一个视频帧，首先在该帧中擦除掉相邻帧关注的区域，从而迫使该帧关注新的部件信息. 尽管 TSE 能够对相邻帧提取到互补的特征表示，但是擦除操作直接丢弃显著性特征，损害了特征表征能力. Hou 等人[35]进一步提出了一个双分支互补网络（Bilateral Complementary Network, BiCnet），两个分支分别处理原始分辨率和降采样后的帧. BiCnet 的双分支结构只能对相邻两帧提取两种尺度的特征，而 TMSCN 的多分支结构能够使连续的多帧关注多个尺度的互补区域. 此外，TMSCN 提出的多分支批量归一化层，有效保证了分支之间的互补性. 综合以上分析，相比于之前的方法，TMSCN 对更多的连续帧挖掘尺度互补的特征，具有更强的帧间互补建模能力，并最终得到一个更丰富的多尺度行人特征表示.

此外，大多数现有方法没有考虑计算开销，通

过引入各种复杂的操作来提高精度. TMSCN 通过对部分帧使用更低的分辨率和更小的网络进行处理，大幅度降低了计算开销，仅仅需要大多数现有方法约 35%的计算开销.

多模态融合方法. 多模态融合[36]是一种典型的联合多个模态的信息执行预测的方法. 在本文中，由于相邻帧通过不同分支进行处理，本文称它们为两个视图，以和模态（例如，图像、文本和音频）区分. 除了输入源不同，多视图的融合方法与多模态非常相似. 通常，多模态融合方法分为双线性融合和注意力融合两类. 双线性池化[37]是双线性融合的一个基础方法，该方法计算不同模态特征的外积构建一个联合表示空间，从而捕捉多模态特征之间的二阶交互. 但是双线性池化需要一个高维的投影矩阵和大量的模型参数，限制了其可用性. 后续工作[38-40]利用矩阵分解技术，将高维的投影矩阵分解为两个低秩矩阵，有效减小了参数量和计算时间. 在注意力融合方法中，早期工作[41-43]采用晚融合机制，即分别基于单模态的数据训练模型，然后根据最终特征联合推理出每个模态的全局注意分布. 最近，一些研究[44,45]使用 Transformer[46]结构实现多模态间的信息交互.

本文提出的跨分支融合模块 CBFM-A 可以归类到注意力融合. 相比于双线性融合方法[37-40], CBFM-A 考虑了相邻帧存在的空间不对齐问题，利用注意力机制实现相邻帧的特征对齐，从而在帧间融合相同部件的信息，产生更具表现力的融合特征；相比于早期的注意力融合方法[41-43]，CBFM-A 在特征学习阶段融合了另一模态信息，并且对每个局部位置自适应生成注意力图，融合与之相关的另一模态信息，有利于增强每个特征单元的表征能力；相比于复杂的 Transformer 结构[44,45], CBFM-A 具有更少的计算量和参数. 此外，Transformer 利用数据驱动的方式学习特征之间的关联值，在数据量较少的行人重识别任务中可能无法学习到准确的特征关联. 而 CBFM-A 直接计算特征之间的外观相似性，更容易定位到相邻帧的关联区域，实现更精准的跨分支融合.

3　时序多尺度互补网络

在视频行人重识别任务中，大多数方法无差别地处理每一个视频帧，导致连续帧的特征高度相似，甚至冗余. 如图 1 所示，这些冗余的特征趋向于关注到同一个局部区域，很难区分局部相似的不同行人. 针对这个问题，本文设计了时序多尺互补度网络（Temporal Multi-Scale Complementary Network, TMSCN），使连续帧关注多个粒度的空间区域，提高了帧间互补建模的能力.

3.1　多分支结构

如图 2 所示，TMSCN 构建在一个多分支架构上. 其中，每个分支具有不同的输入分辨率和不同大小的网络结构. 通过将连续的帧输入到不同的分支中，连续帧能够关注到不同粒度的空间区域，形成一个鲁棒的多尺度行人表示. 本节详细介绍了 TMSCN 的网络结构.

输入层. 给定一个分辨率序列 $\{(H_k,W_k)\}_{k=1}^{K}$，其中 K 是一个表示时序多尺度互补网络的分支个数；分辨率序列降序排列，即 $H_1>H_2>\cdots>H_K$，$W_1>W_2>\cdots>W_K$. 给定一个包含 K 帧的视频片段，$\boldsymbol{I}=\{\boldsymbol{I}_k\}_{k=1}^{K}$，其中 $\boldsymbol{I}_k$ 表示视频片段的第 k 帧，首先将 $\boldsymbol{I}_k$ 通过降采样操作调整到 (H_k,W_k) 分辨率，

$$\begin{aligned}\overline{\boldsymbol{I}}_1&=\boldsymbol{I}_1\\ \overline{\boldsymbol{I}}_k&=downsample[\boldsymbol{I}_k\,|\,(H_k,W_k)],(k>1)\end{aligned}\tag{1}$$

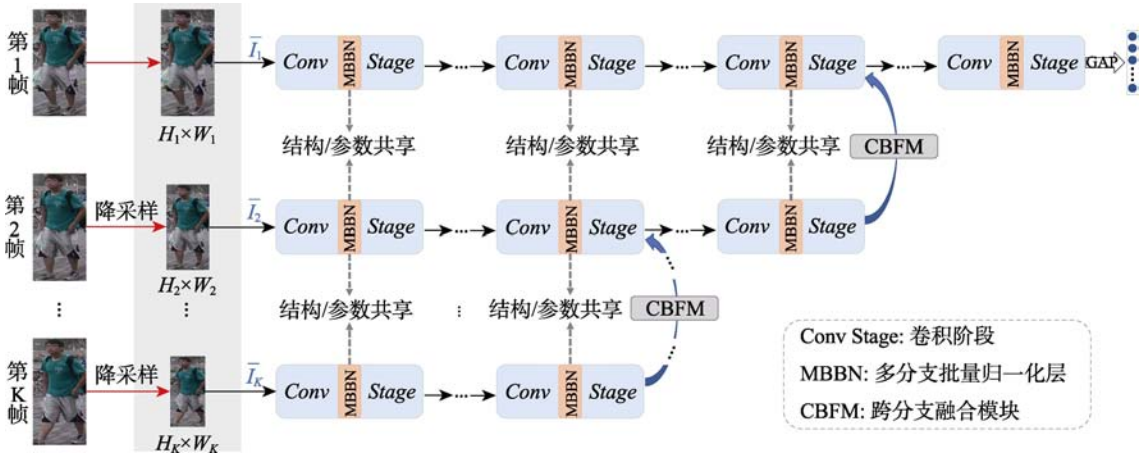


图 2　时序多尺度互补网络（Temporal Multi-Scale Complementary Network, TMSCN）结构

其中，$downsample[\boldsymbol{I}_k\,|\,(H_k,W_k)]$表示将$\boldsymbol{I}_k$降采样到$(H_k,W_k)$分辨率. 然后将降采样后的序列帧$\{\overline{\boldsymbol{I}}_k\}_{k=1}^{K}$送入到 TMSCN 的不同分支中.

分支网络. 如图 2 所示，TMSCN 包含 K 个卷积分支，分别处理不同分辨率的帧，其中每个分支都是一个包含多个卷积阶段（Conv Stage）的深度神经网络. Yang 等人[21]指出，低分辨率帧含有较小的细节特征，使用一个小网络便足够提取其判别性特征；而高分辨率帧包含丰富的细节内容，需要更大的网络进行处理. 因此，TMSCN 在不同的分支上使用深度不同的网络结构. 但是，多分支架构带来的直接问题是多个分支会引入几倍的参数量，增加了模型过拟合的风险. 因此，TMSCN 提出共享分支间相同位置卷积阶段的结构和参数，有效减小了参数的数量.

具体来说，假定 TMSCN 的最高分辨率分支有 L 个卷积阶段，低分辨率分支依次递减卷积阶段的个数，即第 k 个分支包含 $L+1-k$ 个卷积阶段. 不同分支间相同位置处的卷积阶段共享结构和参数. 给定输入的序列帧$\{\overline{\boldsymbol{I}}_k\}_{k=1}^{K}$，TMSCN 将$\overline{\boldsymbol{I}}_k$送入到第 k 个卷积分支中，提取其卷积特征. 通过分支间相同的卷积阶段结构和参数共享，TMSCN 具有与单分支行人重识别网络相同的参数量，有效规避了模型过拟合的风险.

3.2 多分支批量归一化的参数共享

但是分支间的参数共享带来了一个问题：不同分辨率的分支共享归一化层[47]（BN）是不合理的. 这是由于不同分辨率的分支上的特征趋向于有不同的分布. 为了比较不同分支下的特征分布，本节单独训练每一个分支，然后可视化最高分辨率和最低分辨率分支的均值和方差在不同维度上的取值. 从图 3 中可以观察到如下现象：（1）两个分支的均值分布不同，并且在不同的维度上，两个分支的均值有不同的大小关系. Zhang 等人[48]指出，卷积特征的每个通道维度一般都对应某个特定区域的响应. 因此，对于对应高分辨率帧响应区域的维度，由于高分辨率帧含有更多的细节，这个维度在高分辨率分支上会更具有表现力，因此会得到更高的均值统计量. 此外，由于低分辨率帧能够关注更多的区域，关注低分辨率额外响应区域的维度会在低分辨率分支上更具表现力，产生更高的均值. 比如，图 3 中的第 23 维度集中提取行人上衣的特征，对应的高分辨率分支关注的区域，此时高分辨率分支在这个维度上具有较高的均值；而第 56 维度集中提取行人下衣的特征，对应低分辨率引入的新区域，此时低分辨率分支在这个维度上具有较高的均值. 因此，在不同的维度上，不同分辨率分支的均值呈现不同大小的取值.（2）低分辨率分支的方差整体取值小于高分辨率分支. 这是由于低分辨率帧具有更多的低频信息，因此方差较小；而高分辨率帧含有更多的高频信息，因此方差较大. 综合以上分析，不同分辨率的分支会产生不同的均值和方差. 分支间共享 BN 会整合所有分支上的特征更新均值和方差，得到一个混合的统计量. 这个统计量不能代表特定分支上的特征分布，导致批量归一化出现偏差，扰乱了训练过程.

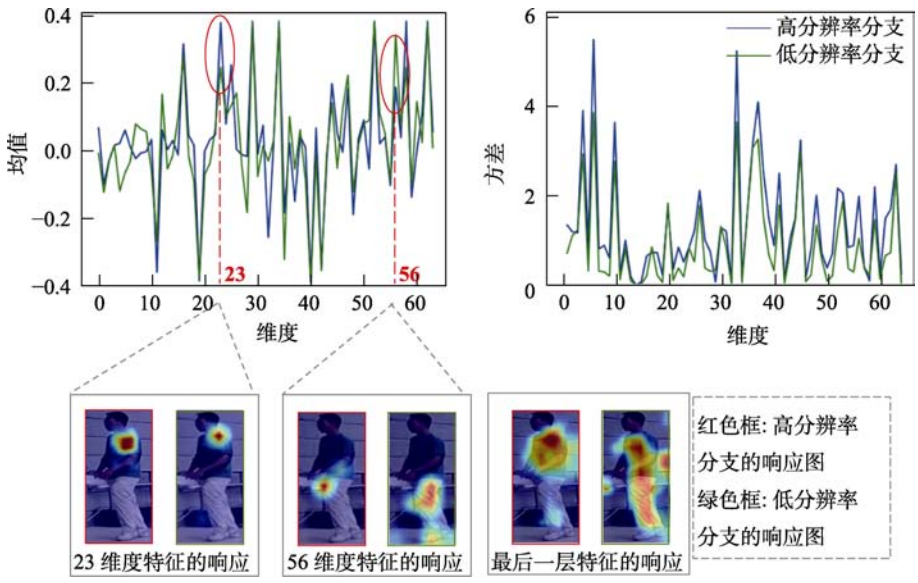


图 3 最高分辨率分支和最低分辨率分支 BN 层的均值/方差的可视化（不同的分支具有不同的均值/方差取值）

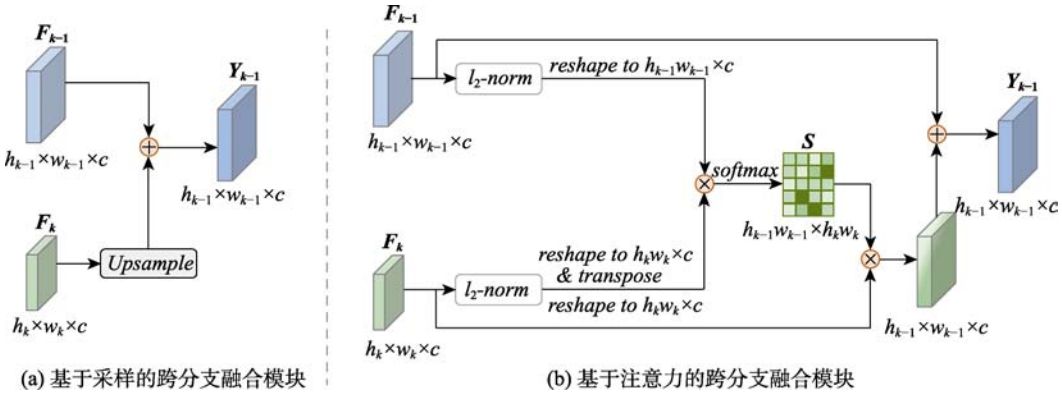


图 4　本文的跨分支融合模块

针对这个问题，本文提出多分支批量归一化（Multi-Branch Batch Normalization，MBBN），即每一个分支具有自己独立的均值和方差统计量. 具体来说，给定批大小为 m 的一批视频序列，首先将其输入到 TMSCN 中，得到 BN 层前的卷积特征 $\{\boldsymbol{x}_k^i\}_{i=1}^m$，其中 $\boldsymbol{x}_k^i$ 为第 i 个视频样本的第 k 帧的卷积特征. 然后，每个分支独立计算输入批次样本的均值和方差：

$$\boldsymbol{u}_k=\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_k^i$$

$$\boldsymbol{\sigma}_k^2=\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{x}_k^i-\boldsymbol{u}_k)^2 \tag{2}$$

$\boldsymbol{u}_k$ 和 $\boldsymbol{\sigma}_k^2$ 分别表示第 k 个分支的均值和方差. 每个分支根据自己的均值和方差对输入特征执行归一化操作：

$$\boldsymbol{y}_i^k=\gamma\left(\frac{\boldsymbol{x}_k^i-\boldsymbol{u}_k}{\sqrt{\boldsymbol{\sigma}_k^2+\epsilon}}\right)+\beta,(1\leqslant k\leqslant K) \tag{3}$$

其中 γ 和 β 是分支间共享的仿射变换参数. 分支间共享仿射变换参数可以将不同分支上的特征映射到相同的特征空间，更利于后续分支间共享卷积参数的训练. 因此 MBBN 采用分支间使用独立的均值和方差，但是共享仿射变化参数的设计.

3.3　跨分支融合模块

本文进一步引入跨分支融合模块（Cross-branch Fusion Module, CBFM），对不同分支上的特征进行跨分支融合. 如图 2 所示，CBFM 将低分辨率分支上的特征传播到相邻高分辨率分支上. 通过逐分支的特征融合，第一个（最高分辨率）分支最终整合了所有分支的尺度互补信息，生成一个多尺度行人特征表示.

Guo 等人[49]指出，不同深度的卷积阶段提取不同级别的行人特征，比如浅层的特征更关注纹理，颜色等低阶视觉信息，深层的特征包含了行人性别等高阶语义信息. 因此，CBFM 将低分辨率分支上的特征传播到相邻高分辨率分支相同位置的卷积阶段上，对相同级别的特征进行融合. 具体来说，给定第 k 个分支的最后一个 $L+1-k$ 卷积阶段得到的特征图. $\boldsymbol{F}_k^{L+1-k}\in\mathbb{R}^{h_k\times w_k\times c}$，其中 h_k，w_k 和 c 分别表示特征图的长、宽和通道个数，CBFM 将它与第 $k-1$ 个分支对应卷积阶段的特征（$\boldsymbol{F}_{k-1}^{L+1-k}$）进行融合. 形式化如下：

$$\boldsymbol{Y}_k^{L+1-k}=CBFM(\boldsymbol{F}_k^{L+1-k},\ \boldsymbol{F}_{k-1}^{L+1-k})\ (2\leqslant k\leqslant K) \tag{4}$$

更新后的特征 $\boldsymbol{Y}_k^{L+1-k}$ 经过第 $k-1$ 个分支上余下的卷积阶段，生成第 $k-1$ 个分支最终的特征.

本文设计了两个跨分支融合模块，基于采样的跨分支融合模块（Cross-branch Fusion Module based on Sampling, CBFM-S）和基于注意力的跨分支融合模块（Cross-branch Fusion Module based on Attention, CBFM-A）. 接下来分别予以介绍.

（1）基于采样的跨分支融合模块. CBFM-S 对不同帧同一位置的特征进行融合. 如图 4 (a) 所示，给定相邻分支的特征图 $\boldsymbol{F}_k\in\mathbb{R}^{h_k\times w_k\times c}$ 和 $\boldsymbol{F}_{k-1}\in\mathbb{R}^{h_{k-1}\times w_{k-1}\times c}$（为了简单起见，本节省略了 $\boldsymbol{F}$ 的上标 $L+1-k$），CBFM-S 首先对 $\boldsymbol{F}_k$ 执行上采样操作将其调整到 $h_{k-1}\times w_{k-1}\times c$ 大小，然后将上采样后的特征图与 $\boldsymbol{F}_{k-1}$ 按元素相加，得到整合后的特征.

$$\boldsymbol{Y}_{k-1}=unsample(\boldsymbol{F}_k)+\boldsymbol{F}_{k-1} \tag{5}$$

（2）基于注意力的跨分支融合模块. 尽管操作简单，CBFM-S 忽视了相邻帧存在的空间不对齐现象. Gu 等人[10]指出，由于视频行人重识别的数据来源于检测算法得到的行人裁剪图像，当检测算法存在误差时（检测框偏移、过大或者过小），相邻帧之间会出现空间不对齐现象. 例如，图 5（a）中相邻帧红框框住的同一位置的区域分别对应头部和背景. 当相邻帧存在空间不对齐时，CBFM-S 会把来

自不同部件的特征融合成一个值，在一定程度上破坏了行人的表观特征. 针对这个问题，本文设计了基于注意力的跨分支融合（CBFM-A）机制，对相邻帧对应部件的特征进行融合.

由于卷积网络的中间层特征图包含一些语义信息[33]. 因此，具有相同表观的特征通常具有较高的相似性，而具有不同表观的特征具有较低的相似性. 图 5（b）可视化了第 $k-1$ 帧标注位置与第 k 帧所有位置的特征相似性，可以观察到表观相同的区域呈现了更高的相似性. 因此，CBFM-A 根据帧间像素的相似性实现同一部件在不同帧的定位.

（a）由于检测误差导致相邻帧出现空间不对齐　（b）帧标注位置与 $k-1$ 帧所有位置的特征相似图（使用合理的 τ，CPFM-A 能在相邻帧中定位到同一部件的区域）

图 5　结果比较

CBFM-A 的结构如图 4（b）所示. 具体来说，给定相邻分支的特征图 $\boldsymbol{F}_k \in \mathbb{R}^{h_k \times w_k \times c}$ 和 $\boldsymbol{F}_{k-1} \in \mathbb{R}^{h_{k-1} \times w_{k-1} \times c}$，首先将它们分别调整为 $\boldsymbol{E}_k \in \mathbb{R}^{h_k w_k \times c}$ 和 $\boldsymbol{E}_{k-1} \in \mathbb{R}^{h_{k-1} w_{k-1} \times c}$ 大小，然后计算 $\boldsymbol{E}_k$ 和 $\boldsymbol{E}_{k-1}$ 任意两个空间位置的余弦相似性，之后使用 *softmax* 函数对结果沿着行进行归一化得到相似性矩阵 $\boldsymbol{S} \in \mathbb{R}^{h_{k-1} w_{k-1} \times h_k w_k}$.

$$(\boldsymbol{S})_{ij} = softmax\left(\tau \frac{(\boldsymbol{E}_{k-1})_i^T (\boldsymbol{E}_k)_j}{(\boldsymbol{E}_{k-1})_{i2} (\boldsymbol{E}_k)_{j2}}\right) \tag{6}$$

其中，τ 是一个调整相似性矩阵分布的超参数. 越大的 τ 会产生一个熵更小的相似性矩阵，相似性矩阵的值会集中在少数几个高响应的位置. 如图 5（b）所示，通过使用一个合理的 τ，相似性矩阵 $\boldsymbol{S}$ 能够精确定位到相邻帧中相同部件的对应位置.

最后，基于注意力的跨分支融合根据 $\boldsymbol{S}$ 整合 $\boldsymbol{E}_k$ 和 $\boldsymbol{E}_{k-1}$. 如等式 6 和图 5 所示，$(\boldsymbol{S})_{ij}$ 表示了 $\boldsymbol{E}_k$ 第 j 个位置与 $\boldsymbol{E}_{k-1}$ 第 i 个位置的特征相似性. 因此，对于 $\boldsymbol{E}_{k-1}$ 的第 i 个位置，以 $(\boldsymbol{S})_{ij}$ 作为 $\boldsymbol{E}_k$ 第 j 个位置的权重，对 $\boldsymbol{E}_k$ 的所有位置的特征进行加权求和，得到对应的融合特征：

$$\boldsymbol{z}_i = \sum_j (\boldsymbol{S})_{ij} (\boldsymbol{E}_k)_j \tag{7}$$

然后将所有 $\boldsymbol{z}_i$ 整合成一个特征图 $\boldsymbol{Z}_{k-1} \in \mathbb{R}^{h_{k-1} \times w_{k-1} \times c}$，并将其与 $\boldsymbol{E}_{k-1}$ 相加得到融合后的特征，再将其调整为原始特征的大小，

$$\boldsymbol{Y}_{k-1} = R(\boldsymbol{Z}_{k-1} + \boldsymbol{E}_{k-1}) \tag{8}$$

其中 R 为调整（Reshape）操作，将特征调整为 $h_{k-1} \times w_{k-1} \times c$ 大小.

4 基于 TMSCN 的视频行人重识别

整体流程. 在视频行人重识别任务中，给定一个包含 T 帧的视频序列，首先以 K 帧为间隔将其划分为多个子序列；然后分别将每个子序列输入到 TMSCN. 由于第一个（最高分辨率）分支整合了所有分支的信息，对第一个分支最终的卷积特征图进行全局平均池化（Global Average Pooling, GAP）得到子序列的特征向量；最后对所有子序列的特征向量取平均得到该视频的特征. 在训练阶段，根据网络输出的视频特征计算损失，利用梯度下降算法训练整个网络.

网络结构. 本文采用在大规模数据集 ImageNet[50] 上预训练的 ResNet-50[14]模型为主干网络. ResNet-50 包含四个连续的卷积阶段. 为了增大特征图的感受野，本文将 ResNet-50 网络最后一个卷积阶段的步长设置为 1. TMSCN 的分支构建在 ResNet-50 上，其中第一个（最高分辨率）分支使用完整的 ResNet-50 结构，余下的分支依次递减卷积阶段的个数. 本文将 TMSCN 的分支个数设置为 ResNet-50 的卷积阶段个数，因此最后一个分支个数仅包含一个卷积阶段. 本文将 ResNet-50 的 BN 都替换为提出的 MBBN，并将 CBFM 加在相邻分支上进行跨分支特征融合.

计算开销. 为了说明 TMSCN 的计算开销，本节考察一个常用的视频行人重识别基准模型（Baseline），使用 ResNet-50 以原始分辨率为每一帧提取特征. 假定 ResNet-50 处理原始帧的浮点运算次数（FLOPs）为 p，可以计算出 ResNet-50 的前 3/2/1 个卷积阶段处理原始帧的 FLOPS 分别为 $0.52p / 0.29p / 0.12p$. 给定一个包含 N 帧的视频序列，Baseline 需要的计算量为 Np. TMSCN 将不同帧降采样到不同的分辨率，用 s_k 表示第 k 帧的降采样率（$s_k = H_k / H$），TMSCN 需要的计算量约为 $\frac{N}{4}(p + 0.52s_2^2 p + 0.29s_3^2 p + 0.12s_4^2 p)$（CBFM 的计算量与特征提取的计算量相比可以忽略不计）. 相比于 Baseline，TMSCN 减小了 $(75 - 13s_2^2 - 7.25s_3^2 - 3s_4^2)\%$ 的计算量.

可以观察到，TMSCN 的计算复杂度随着 s_k 的减小而降低. 但是，当 s_k 过小时，输入的低分辨率图像会丢失大部分行人信息，引入干扰性信息，从

而造成模型性能的严重下降. 本文将 $\{s_2, s_3, s_4\}$ 设置为 $\{0.75, 0.5, 0.375\}$，可以在计算成本和精度之间取得良好平衡. 此时，TMSCN 的计算开销仅仅是 Baseline 的 35%，能够更有效地提取视频序列的特征表示.

损失函数. 本文实验使用的损失函数为交叉熵损失，困难样本三元组损失和对比损失. 行人重识别可以被视为分类任务. Zheng 等人[4]提出使用交叉熵损失训练视频行人重识别网络. 交叉熵损失通过行人重识别网络预测的行人类别概率与真实标签进行损失评估，公式如下：

$$L_1 = \sum_c y_c log(p_c) \tag{9}$$

其中，p_c 代表网络预测的输入序列属于 c 类的概率，y_c 代表真实概率.

三元组损失旨在一定距离上把正负样本分开. 本文使用困难样本三元组损失[51]，对于每个目标样本，在一个训练批次内挑选出距离最远的正样本和距离最近的负样本组成三元组计算损失，

$$L_2 = \left[\max_{p \in P(a)} d(\boldsymbol{f}_a, \boldsymbol{f}_p) - \min_{n \in N(a)} d(\boldsymbol{f}_a, \boldsymbol{f}_n) + \alpha, 0 \right]_+ \tag{10}$$

其中，$\boldsymbol{f}_a$，$\boldsymbol{f}_p$，$\boldsymbol{f}_n$ 分别为目标样本、正样本和负样本的特征向量；$P(a)$，$N(a)$ 分别表示 $\boldsymbol{f}_a$ 的所有正样本的索引集合和所有负样本的索引集合；$d(,)$ 表示余弦距离；α 为目标样本与正负样本的间隔参数，设置为 0.5.

对于每个目标样本，三元组损失仅仅考虑一个正负样本对. Hou 等人[35]在视频行人重识别任务中引入对比损失[52]. 对比损失在输入的批次样本中，考虑所有的正负样本对，定义如下，

$$L_3 = -\sum_{p \in P(a)} log \frac{exp(d(\boldsymbol{f}_a, \boldsymbol{f}_p))}{\sum_{n \in N(a) \bigcup P} exp(d(\boldsymbol{f}_a, \boldsymbol{f}_n))} \tag{11}$$

时序多尺度互补网络通过上述三个损失（交叉熵损失、三元组损失以及对比损失）联合训练，综合上述分析，时序多尺度互补网络的整体损失函数为

$$L = L_1 + \lambda_1 L_2 + \lambda_2 L_3 \tag{12}$$

λ_1 和 λ_2 是用来平衡各个损失的超参数.

5　实验验证

在本节中，通过在三个常用的视频行人重识别数据集上进行对比试验，验证本文方法相对于已有方法的优越性. 此外，本节采取了一系列消融实验，验证了本文方法各个模块的有效性，并对参数的敏感性进行分析.

5.1　数据集

数据集. 本文在三个极具挑战性的视频行人重识别数据集：MARS 数据集[4]、LS-VID 数据集[22]和 iLIDS-VID 数据集[3]进行实验.

MARS 数据集是第一个大规模的视频行人重识别数据集. 该数据集通过 6 个摄像头捕捉得到，包含 1261 个行人，共有 17503 个行人视频片段和 3248 个干扰视频片段. 其中，每个视频序列平均有 58 帧. 在 1261 个行人中，625 个行人用于训练，636 个行人用于测试.

LS-VID 是最近提出的一个大规模视频行人重识别数据集. 该数据集一共部署了 15 个摄像头，包含 3,772 个行人，共有 14,943 个行人视频片段，每个序列平均有 200 帧. 其中训练集包含 842 个行人，验证集包含 200 个行人，测试集包含 2,730 个行人.

iLIDS-VID 数据集包含 300 个行人和 600 个行人序列，其中每一个行人有来自两个摄像头的序列. 每一个序列平均含有 71 帧. 相比其他的数据集，iLIDS-VID 包含更多的遮挡样本，更具挑战性. 与文献[25]相似，本文随机选取 150 个行人用于训练，其余行人用于测试.

评价指标. 本文使用累计匹配特征曲线 CMC 和平均准确率 mAP 评估模型的性能.

实验环境. 本文使用 Pytorch[53]进行代码编写，在配置 Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz 和 NVIDIA 2080Ti GPU 的服务器进行实验.

5.2　实验细节

训练阶段. 本文采用 Adam 进行网络优化，共训练 150 个回合. 初始化学习率设为 3.5×10^{-4}，每训练 40 个回合后下降 10 倍. 训练使用的小批量大小为 64，其包含 16 个行人类别，每个类别包含 4 个视频片段. 在训练时，输入的视频片段的长度应该相同，并且每一帧应该以相同的概率被采样. 因此，对于每个原始视频，以 4 帧为间隔随机采样 8 帧形成一个视频片段. 本文采用水平翻转和随机擦除作为数据增广. 在时序多尺度互补网络中，网络的分支个数 K 设为 4，输入的分辨率序列设为 $\{256\times128, 92\times96, 128\times64, 96\times48\}$，$\tau$ 设为 6. 如果没有特殊说明，本文默认使用基于注意力的跨分支融合模块. 在 MARS 数据集上，使用交叉熵和三元组损失联合优化. 在 iLIDS-VID 和 LS-VID 数据集上，使用交叉熵和对比损失联合优化. 不同损失具有相同的权重. 本文的所有实验都是多次实验求平均的结果. 具体而言，对于每个实验，本文都采用固定的 5 个随机种子

进行了 5 次实验，并报告了这 5 次实验的平均结果.

测试阶段. 与现有视频重识别方法相同，本文使用视频的全部帧进行测试. 首先将每个原始测试视频等分为多个含有 K 帧的视频片段，然后利用训练好的 TMSCN 提取每个视频片段的特征表示，最后取所有片段的特征均值作为原始测试视频的特征表示.

对于每个要查询的视频（query），计算它与所有候选视频（gallery）特征的余弦相似性；然后，根据计算出的特征余弦相似性对所有的候选视频进行排序，选出最相似的候选视频，达到目标检索的目的.

5.3 与已有方法对比

本节将本文提出的方法与现有视频行人重识别方法进行比较. 对比方法主要是最近提出的深度学习的方法，包括基于集合建模的方法：QAN[12]，DRSA[27]，VRSTC[17]；基于时空建模的方法：M3D[29]，V3D[28]，GLTP[22]，I3D[54]，P3D[55]，STGCN[19]，AP3D[10]，MGH[18]，MGRAFA[11]；基于帧间互补建模的方法：TCLNet[9]，BiCnet-TKS[35]. 对比结果见表 1，本文提出的 TMSCN 与众多对比模型相比均达到了较高水平.

表 1 与现有视频行人重识别方法性能对比（本文方法（TMSCN）的结果见最后一行. 比较的方法分为以下三组：集合建模方法（S），时空建模方法（ST），帧间互补建模方法（C）. GFLOPs 表示模型平均处理一帧的浮点运算次数）

	方法	GFLOPs	MARS			LS-VID			iLIDS-VID	
			mAP	top-1	top-5	mAP	top-1	top-5	top-1	top-5
S	QAN[12]	1.6	51.7	73.7	84.9	–	–	–	68.0	86.8
S	DRSA[27]	4.1	65.8	82.3	–	–	–	–	80.2	–
S	VRSTC[17]	7.4	82.3	88.5	96.5	–	–	–	83.4	95.5
ST	M3D[29]		74.1	84.4	93.8	40.1	57.7	76.1	74.0	94.3
ST	V3D[28]		77.0	84.3	–	–	–	–	81.3	–
ST	GLTP[22]		78.5	87.0	95.8	44.3	63.1	77.2	86.0	**98.0**
ST	I3D[54]		83.0	88.6	–	33.9	51.0	70.1	–	–
ST	P3D[55]	≥4.1	83.2	88.6	–	35.0	53.4	71.2	–	–
ST	STGCN[19]		83.7	89.9	96.4	–	–	–	–	–
ST	AP3D[10]		85.1	90.1	–	73.2	84.5	–	86.7	–
ST	MGH[18]		85.8	90.0	96.7	–	–	–	85.6	97.1
ST	MGRAFA[11]		85.9	88.8	97.0	–	–	–	88.7	**98.0**
C	TCLNet[9]	4.1	85.1	89.8	96.5	70.3	81.5	92.7	86.6	96.9
C	BiCnet[35]	1.9	85.6	89.8	96.6	74.3	83.1	93.0	–	–
C	BiCnet–TKS[35]	2.0	86.0	90.2	96.6	75.1	84.6	94.5	–	–
C	TMSCN(ours)	1.4	**86.2**	**90.7**	**97.1**	**79.6**	**87.7**	**96.0**	**90.0**	**98.0**

与基于集合建模和基于时空建模的方法相比，TMSCN 具有更高的性能和更少的计算开销. 基于集合建模和基于时空建模的方法以相同的分辨率和网络结构处理视频序列的每一帧，忽略了相邻帧的冗余. TMSCN 以不同的分辨率和网络处理不同的连续帧，使其能够提取从局部到全局的互补特征，形成一个多尺度行人特征，从而取得更好的性能. 相比于上述结果最好的方法 MGRAFA[11]，TMSCN 在 MARS 数据集上的 top-1 提升了 1.9%，mAP 提升了 0.3%.

从表 1 可见，TMSCN 具有比 TCLNet[9]更好的性能. 在 LS-VID 数据集上，TMSCN 的 mAP 提升了 9.3%，top-1 精度提升了 6.2%. 性能的大幅度提升验证了 TMSCN 具有更强的帧间互补性建模能力. 具体来说，TCLNet 的显著性擦除操作直接丢弃掉显著性特征，不可避免地损害了特征的表示能力. 此外，TCLNet 只考虑了单尺度的特征提取，TMSCN 通过采用多个不同分辨率的分支，为连续的多帧提取尺度互补的特征，得到一个更鲁棒的多尺度表示.

从表 1 可见，TMSCN 具有比 BiCnet[35]更好的性能. 在 LS-VID 数据集上，TMSCN 的 mAP 提升了 5.3%，top-1 精度提升了 4.6%. 这是由于（1）BiCnet 仅仅考虑了两种尺度的输入帧，帧间互补性建模能力有限；TMSCN 的多分支结构为更多的连续帧提取多尺度互补的特征，具有更强的帧间互补性建模能力；（2）BiCnet 使用分支间共享的 BN 层，

导致模型训练不稳定；TMSCN 提出的 MBBN 稳定了模型的训练过程，保证了分支间的互补性；（3）BiCnet 采用简单的加和形式融合不同分支的特征；而 TMSCN 考虑帧间不对齐现象，设计了 CBFM-A 模块，进行更有效的跨分支特征融合.

本节进一步比较了 TMSCN 与已有方法的计算开销. QAN[12]使用 GoogleNet[56]作为主干网络，具有比较小的计算开销. 相比于 QAN，TMSCN 减小了 0.2GFLOPs 的计算量，并且在 MARS 数据集上提升了高达 30%的 mAP. 其余基于集合建模和时空建模的方法都使用 ResNet50 作为主干网络，并在 ResNet50 上添加提出的集合建模或时空建模模块. 因此这些方法的计算量都高于 ResNet50，需要大于等于 4.1GFLOPs 的计算量处理视频帧. TCLNet[9]在 ResNet50 上添加了一个非常轻量级的显著性擦除模块，具有与 ResNet50 相当的计算量. BiCnet[35]通过降低输入帧的分辨率减小了计算开销，处理一帧平均需要 2.0GFLOPs. 相比之下，TMSCN 通过使用小网络处理低分辨率帧，具有更少的计算开销，仅仅需要 BiCnet 模型 70%的计算量.

5.4　与动作建模方法的结合

本文 TMSCN 集中建模行人序列的表观特征，即对视频的连续帧提取互补的特征，形成一个完整的行人表观描述. 然而，相比于单帧图像，行人视频序列一方面包含更丰富的表观信息，另一方面包含动作信息，动作信息能够用来帮助区分表观相似的不同行人. 因此，可以将 TMSCN 与现有的动作建模方法相结合，使重识别模型具备同时提取行人序列表观特征和动作特征的能力. 本节在 MARS 数据集上测试了 TMSCN 与不同动作建模方法结合之后的性能.

本节考察了 4 个常用的动作建模方法，AP3D[10], TKS[35], NonLocal[33]和 TGCN[19]. 其中，AP3D 和 TKS 基于 3D 卷积建模时序关系；NonLocal 和 TGCN 利用图网络捕捉序列长期的动作信息. 为了简单起见，本文只考虑嵌入一个模块，即分别将这四种模块嵌入到 TMSCN 的第二个卷积阶段后. 结果如表 2 所示. 可以看到上述结合方法引入了很少的计算开销和参数，并且能够进一步提升模型性能. 例如，AP3D 和 NonLocal 在 mAP 上分别提升了 0.3%和 0.5%. 这个结果说明了 TMSCN 模型的鲁棒性和兼容性. 因此，后续的视频行人重识别工作能够以 TMSCN 作为表观建模的基准模型，集中研究更先进的动作建模方式.

表 2　TMSCN+动作建模模块在 MARS 数据集上的性能
（Param.表示模型的参数量）

模型	GFLOPs	Param.	mAP	top-1
TMSCN	1.44	23.5M	86.2	90.7
+TGCN[19]	1.44	23.8M	86.4	90.7
+TKS[35]	1.44	25.1M	86.4	90.9
+AP3D[10]	1.46	24.4M	86.5	**91.1**
+NonLocal[33]	1.46	24.0M	**86.7**	90.9

5.5　图像行人重识别的应用

本文的 TMSCN 能够应用到基于图像的行人重识别任务上. 实现上，对单张图像复制成四张，分别作为 TMSCN 四个分支的输入. 本节将 TMSCN 与一个基准模型（使用 ResNet-50 提取图像特征），以及最近的图像行人重识别方法进行比较. 对比结果见表 3. 在 Market-1501[63]和 MSMT[64]数据集上，相比于基准模型，TMSCN 的 mAP 分别提升了 3.1%和 4.4%，top-1 分别提升了 1.7%和 2.8%. 性能的提升归功于 TMSCN 可以建模图像的多粒度信息，形成一个更鲁棒的行人表征. 此外，TMSCN 与最近先进的图像行人重识别方法相比也达到了较高水平，进一步验证了本文方法对图像行人重识别任务的有效性. 但是 TMSCN 应用到图像重识别任务的一个缺陷是会引入额外的计算量.

表 3　与现有图像行人重识别方法性能的对比

方法	GFLOPs	Market-1501		MSMT17	
		mAP	top-1	mAP	top-1
PCB+RPP[2]	⩾4.1	81.6	93.8	40.4	68.2
IANet[57]		83.1	94.4	46.8	75.5
OSNet[58]	1.0	84.9	94.8	52.9	78.7
CtF[59]		84.9	93.7	–	–
CDNet[60]	⩾4.1	86.0	95.1	54.7	78.9
PAT[61]		88.0	**95.4**	–	–
Baseline[62]	4.1	85.2	93.5	54.5	76.8
TMSCN	5.6	**88.3**	95.2	**58.9**	**79.6**

5.6　TMSCN时序信息的利用

本文主要利用视频的时序性提取连续帧粒度互补的特征. 实现上，给定一个 L 帧的视频，将其等划分为多个包含连续 K 帧的视频片段. 对于每个视频片段，TMSCN 依次递减这 K 帧的输入分辨率进行处理. 通过这个操作，视频的连续 K 帧具有互补性，能够分别关注从局部到全局不同粒度的互补区

域，提取到粒度互补特征. 同时，视频的非连续帧（时间间隔大于 K ）具有独立性，利用远距离帧之间的姿态和视角等差异变化进一步丰富行人特征.

综上分析，本方法利用了以下的视频时序信息：视频的时序连贯性导致相邻帧更为相似，而时序间隔较远的帧由于姿态和视角变化具有一定的差异性. 图 6 展示了基准模型中第 1 帧与第 M 帧的特征余弦相似性. 可以观察到，随着时序间隔的增加，不同帧之间的特征相似性随之降低，验证了时序上连续帧的特征相似与远距离帧的特征差异. 因此基于上述时序信息，TMSCN 互补建模连续 K 帧，鼓励连续 K 帧的特征互补；同时独立建模间隔大于 K 的不同帧，全面捕捉远距离帧的差异化特征.

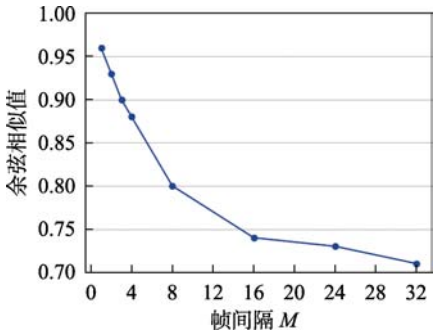


图 6 在 MARS 数据集上，基准模型中所有测试序列第 1 帧与第 M 帧的特征余弦相似性的均值

为了验证 TMSCN 中时序信息的有效性，引入了两个比较模型，TMSCN-PV 和 TMSCN-PC. 给定一个 L 帧的视频，TMSCN-PV 随机改变视频的所有 L 帧的时序顺序；TMSCN-PC 将原始视频等划分为多个包含连续 K 帧的视频片段，然后随机改变每个片段的 K 帧的时序顺序. 实验结果如表 4 所示. 可以观察到，TMSCN-PV 的 top-1 和 mAP 分别降低了 1.9%和 0.7%，验证了时序信息的有效性. 这是由于改变所有帧的时序顺序之后，模型可能对远距离的不同的帧关注不同区域，从而丢失了相同区域中由于姿态、视角变化带来的新的视觉线索. TMSCN-PC 的 top-1 和 mAP 分别降低了 0.4%和 0.5%，验证了在视频片段内保留时序信息的有效性. 这主要是由于相邻帧的特征更为相似（图 6 所示），将相邻帧输入到相邻分支时，CBFM-A 能够利用特征相似性更精确定位到相邻帧中同一部件的区域，产生更好的融合结果.

本节引入了两个比较模型：TMSCN-RV 和 TMSCN-RC. 给定一个 L 帧的视频，TMSCN-RV 逆序排列 L 帧；TMSCN-PC 将原始视频等划分为多个包含连续 K 帧的视频片段，然后逆序排列每个片段的 K 帧. 如表 4 所示，TMSCN-RV 和 TMSCN-RC 取得了与 TMSCN 相当的性能. 这是由于行人的运动具有周期性，逆序的视频保留了时序性，同时维持了相邻帧更相似的时序特点，因此不会影响重识别性能. 为了实现方便，本文维持了视频的原始时序顺序.

表 4 时序信息的有效性

模型	mAP	top-1
TMSCN-PV	85.5	88.8
TMSCN-PC	85.7	90.3
TMSCN-RV	86.0	90.6
TMSCN-RC	86.1	90.5
TMSCN	**86.2**	**90.7**

5.7 消融实验

本节通过在 MARS 数据集上采取一系列消融实验验证了 TMSCN 各个模块的有效性. 首先引入了一个常用的基准模型（Baseline），使用 ResNet-50 独立提取每一帧的特征，然后通过时序平均池化融合多帧特征得到视频特征. 基准模型以相同的分辨率（默认为 256×128 ）处理所有的输入帧，并且使用与 TMSCN 相同的损失函数进行训练. 消融实验结果如表 5 所示.

表 5 TMSCN 在 MARS 上不同模块的消融实验

模型	GFLOPs	Param.	mAP	top-1
Baseline	4.08	23.5M	85.2	89.6
BiCnet[35]	1.89	27.6M	85.6	89.8
TMSCN-S	1.98	23.5M	85.8	89.9
TMSCN-BN	1.44	23.5M	83.9	89.3
TMSCN-CBFM-S	1.41	23.5M	85.7	89.9
TMSCN-LF	1.46	26.3M	85.4	89.9
TMSCN(-CBFM-A)	1.44	23.5M	**86.2**	**90.7**

分支个数的影响. 本节测试了 TMSCN 的分支个数（ K ）对于模型性能的影响. 由于 ResNet-50 包含 4 个卷积阶段，且 TMSCN 多个分支依次递减一个卷积阶段，因此 TMSCN 的最大分支个数设置为 4. 实验结果如表 6 所示. 从表 6 可以观察到：（1）在合适的低分辨率上训练 ResNet-50 仍然能够提供较高的精度. 相比于原始分辨率，采用 192×96 分辨率，top-1 仅降低了 1.3%；采用 128×64 分辨率，top-1 仅降低了 2.6%. 同时采用低分辨率的输入大幅减小了计算开销，192×96 分辨率降低了 44%的计算开销，

128×64降低了 75%的计算开销；（2）太小的输入分辨率会导致严重的性能下降. 当输入分辨率为96×48时，基准模型的 mAP 降低了 12.7%. 这可能是由于当输入分辨率太小时，模型仅仅能够捕捉到全局信息，丢失了细节线索，从而很难区分类间变化小的不同行人；（3）多分支模型（$K>1$）的性能一致优于单分支模型（$K=1$），说明了多分支结构对于视频行人重识别任务的有效性. 当 $K=4$ 时，TMSN 取得了最高的性能，并且计算量只是基准模型的 35%. 因此，TMSCN 采用 4 分支结构.

表 6　分支个数 K 对时序多尺度网络（TMSCN）性能的影响（高度 H 表示分支的输入帧的分辨率为 $H\times H/2$）

K	高度 H				GFLOPs	mAP	top-1
	256	192	128	96			
1	√				4.08	85.2	89.6
		√			2.30	83.1	88.3
			√		1.02	78.8	87.0
				√	0.57	72.5	81.9
2	√	√			2.64	85.9	89.9
3	√	√	√		1.87	86.1	90.1
4	√	√	√	√	**1.44**	**86.2**	**90.7**

训练时采样的视频片段长度的影响. 训练时，本文对每个原始视频采样 8 帧形成一个视频片段，且每个批量包含 64 个视频片段. 使用 8 帧视频片段主要有以下两方面的原因. 一方面，一个短的视频片段的特征容易受到低质量帧的干扰，导致训练过程不稳定. 相反，一个长的视频片段能够整合更多帧的信息，得到一个更鲁棒的视频特征，有利于训练的稳定性. 另一方面，当采样的视频片段过长时，由于 GPU 显存的限制，无法使用一个大的批量大小进行训练，导致梯度震荡严重，不利于收敛. 表 7 展示了视频片段长度对模型性能的影响，由于网络分支个数为 4，视频片段长度应该是 4 的整倍数，且批量大小根据 GPU 显存进行调整. 可以观察到，相比于 4/16 帧片段，使用 8 帧片段进行训练能够取得最优的性能. 因此，TMSCN 使用 4 分支、8 帧片段的方式进行训练.

表 7　训练时采样的视频片段长度的影响（批量大小根据 GPU 显存进行调整）

帧长	批量大小	mAP	top-1
4	128	84.3	89.2
16	32	84.1	89.8
8	64	**86.2**	**90.7**

多分支结构的影响. 本节测试了不同多分支结构的影响，引入了一个比较模型 TMSCN-S：TMSCN-S 的不同分支采用相同的 ResNet-50 结构. 比较结果如表 5 所示. 可以观察到，TMSCN 具有更优的性能. TMSCN-S 的 mAP 和 top-1 分别降低了 0.4%和 0.8%. 这是由于 TMSCN-S 在低分辨率上使用一个大网络，增大了模型对于低分辨率输入过拟合的风险. 其次，TMSCN 具有更低的计算开销. TMSCN 在低分辨率上使用更小的网络，降低了计算开销， 计算开销仅是 TMSCN-S 的 72%. 综合上述分析，TMSCN 具有比 TMSCN-S 更高的性能和更低的计算开销. 因此，本文采用 TMSCN 的网络设计，即对低分辨率分支使用更小的网络结构.

相同位置卷积阶段结构/参数共享的有效性. 为了进一步验证不同分支相同位置上的卷积阶段共享结构和参数的有效性，本节引入了两个比较模型，TMSCN-WoSA 和 TMSCN-WoSP. 其中 TMSCN-WoSA 对不同的分支同一位置的卷积阶段使用不同的网络结构，即从高分辨率分支到低分辨率分支分别使用 ResNet-50，ResNet-34，ResNet-18 和更轻量级模型 MobileNetv2[64]；TMSCN-WoSP 对不同分支的相同位置卷积阶段使用相同的网络结构，但是分支间不共享参数. 比较结果如表 8 所示，可以观察到 TMSCN-WoSA 和 TMSCN-WoSP 都带来了性能的降低. 这是由于这两个模型都引入了过多的参数量，其中 TMSCN-WoSA 引入了几乎 2 倍的参数量，增加了模型过拟合的风险，导致性能降低. 相反，TMSCN 共享分支间相同位置处的卷积结构和参数，共享的参数能够通过多个分辨率的输入共同优化，避免了过拟合在单个分辨率上，增加了模型的泛化性，从而取得了更好的测试性能.

表 8　TMSCN 的多分支结构设计

模型	GFLOPs	Param.	mAP	top-1
TMSCN-S	1.98	23.5M	85.8	89.9
TMSCN-WoSA	1.69	60.1M	85.2	89.7
TMSCN-WoSP	1.44	33.7M	85.6	89.0
TMSCN	1.44	23.5M	**86.2**	**90.7**

多分支归一化层的有效性. 本节测试了多分支批量归一化（MBBN）的有效性，引入了一个比较模型 TMSCN-BN. TMSCN-BN 使用原始的批量归一化层（BN），即网络的不同分支共享均值和方差统计量. 如表 5 所示，TMSCN 的性能显著优于 TMSCN-BN，其中 mAP 提升 2.3%，top-1 提升 1.4%. 这个

结果验证了对不同分辨率的分支使用独立的均值和方差的合理性，说明了 MBBN 在多分支模型中的有效性.

接着进一步对 MBBN 进行消融实验，引入了三个比较模块：MBBN-u、MBBN-σ 和 MBBN-WoSA. 其中 MBBN-u 对不同的分支使用独立的均值和共享的方差；MBBN-σ对不同的分支使用独立的方差和共享的均值；MBBN-WoSP 对不同的分支使用独立的均值和方差，并且不共享仿射变换参数. 比较结果如表 9 所示. 可以观察到，仅仅使用独立的均值（MBBN-u）或者方差（MBBN-σ）都会造成性能大量降低，其中 MBBN-u 的 mAP 降低了 2.0%，MBBN-σ的 mAP 降低了 3.2%. 这个现象说明了同时对不同的分支使用独立的均值和方差的有效性. 此外，MBBN-WoSA 的性能略低于 MBBN，mAP 降低了 0.2%，top-1 降低了 0.3%. 这是由于分支间共享相同位置卷积的参数不同，当 MBBN 使用不同的仿射变化参数时，可能将不同分支上的特征映射到不同的特征空间，不利于后续分支间共享参数的训练. 因此，MBBN 采用分支间使用独立的均值和方差，但是共享仿射变化参数的设计.

表 9 MBBN 模块的消融实验

模型	GFLOPs	Param.	mAP	top-1
MBBN-u	1.44	23.5M	84.2	88.9
MBBN-σ	1.44	23.5M	83.0	88.3
MBBN-WoSP	1.44	23.6M	86.0	90.4
MBBN(TMSCN)	1.44	23.5M	**86.2**	**90.7**

跨分支融合模块的有效性. 本节测试了不同跨分支模块对模型性能的影响. 其中 TMSCN-CBFM-S 和 TMSCN 分别使用基于采样的跨分支融合模块（CBFM-S）和基于注意力的跨分支融合模块（CBFM-A）. 如表 5 所示，TMSCN-CBFM-S 和 TMSCN 的性能一致优于基准模型，说明了这两个跨分支融合模块的有效性. 此外，CBFM-A 能够取得比 CBFM-S 更高的性能，mAP 提升了 0.5%，top-1 提升了 0.8%. 这是由于 CBFM-S 忽略了相邻帧之间的空间不对齐现象，可能将不同部件的特征融合成一个值，破坏了行人的表观特征. 而 CBFM-A 利用注意力机制对相邻帧进行相同部件的融合，产生更具表现力的融合特征，因此具有更高的性能.

跨分支融合模块的放置位置. 本节进一步测试了 CBFM-A 放置位置，引入了一个变体，TMSCN-LF. TMSCN-LF 将低分辨率的卷积特征与相邻高分辨率分支最终卷积阶段的特征进行融合. 由于不同卷积阶段的特征通道数不一样，TMSCN-LF 首先对低分辨率特征采用1×1卷积将其映射到高分辨率分支最终特征的维度. 结果如表 5 所示，TMSCN-LF 的 mAP 为 85.4%，top-1 为 89.9%. 相比于 TMSCN，TMSCN-LF 的 mAP 和 top-1 分别降低了 0.8%和 0.9%. 这个结果说明了在分支间融合相同阶段的卷积更为有效. 这是因为：（1）相同卷积阶段的特征提取相同级别的信息，更具有可比性. 因此，将相同阶段的特征进行融合时，CBFM-A 能够更精确的定位到相邻帧中同一部件的区域，产生更好的融合结果；（2）由于更深的卷积特征一般具有更强的判别性，TMSCN-LF 将低分辨率的低层特征与高分辨率高层特征结合，可能损害了高层特征的判别性，导致性能降低.

多尺度多分支结构的有效性. 本节验证了 TMSCN 采用的多尺度多分支结构的有效性，引入了两个比较模型 TCN 和 TMS. 其中 TCN 采用同尺度多分支结构设计，不同的分支都使用相同的输入分辨率 256×128，并保持与 TMSCN 一致的多分支结构；TMS 采用多尺度同分支结构设计，不同的帧使用与 TMSCN 一致的不同分辨率处理，但是都经过同一个 ResNet50 分支结构. 实验结果如表 10 所示. （1）相比于 TCN，TMSCN 的 top-1 和 mAP 分别提升了 0.7%和 0.5%，验证了多尺度的有效性. 这是由于在多尺度结构下，大尺度的输入帧能够保留行人的细节视觉特征，小尺度的输入帧能够增大网络感受区域捕捉到更全局的内容，此时连续帧能够提取到尺度互补的特征，增强了视频特征的表征能力. 此外，TMSCN 仅仅需要 TCN 方法 56%的计算量，进一步说明了多尺度结构的优越性. （2）相比于 TMS，TMSCN 的 top-1 和 mAP 分别提升了 0.8%和 0.4%，验证了多分支的有效性. 这是由于 TMS 对不同的尺度都采用同一个 ResNet50 分支，增大了模型对小尺度输入的过拟合风险. 其次，TMSCN 具有更低的计算开销，仅是 TMS 的 72%. 综上，TMSCN 具有更少的计算开销和更高的性能，并且性能的提升来自于两个方面：不同帧的多尺度输入以及不同尺度下的多分支结构.

表 10 TMSCN 中多尺度/多分支的消融实验

模型	GFLOPs	Param.	mAP	top-1
TCN	2.59	23.5M	85.7	90.0
TMS	1.98	23.5M	85.8	89.9
TMSCN	1.44	23.5M	**86.2**	**90.7**

多帧特征融合的有效性. 为了验证多帧特征融合的有效性，本节比较了两个单帧特征融合方法，Wang 等人[65]提出的 DaRe 和本文方法的变体 TMSCN-SF. 其中，DaRe 在输入视频片段中随机抽取一帧图像，然后融合单帧图像在网络的不同深度层特征用于重识别；TMSCN-SF 在输入视频片段中随机抽取一帧图像，然后将单帧图像同时输入到 TMSCN 的多个分支中，进行单帧图像间的跨分支特征融合. 实验结果如表 11 所示. 可以观察到，TMSCN 的性能显著优于 DaRe 和 TMSCN-SF，在 mAP 上大约提升了 2%，验证了多帧特征融合的有效性. 这是由于视频的不同帧可能包含不同的姿态和视角，多帧特征融合更能充分利用视频丰富的行人信息，从而建模更鲁棒的行人表示，取得更高的性能水平.

表 11　多帧特征融合的有效性

模型	mAP	top-1
DaRe[65]	84.2	89.3
TMSCN-SF	84.5	89.4
TMSCN	**86.2**	**90.7**

5.8　参数分析

超参数 τ. CBFM-A 引入了超参数 τ（等式 6）. 在这一节中，进一步测试 τ 对模型的影响. 如图 5 (b) 所示，过小的 τ（$\tau=1$）得到的相似性矩阵的值过于分散，可能引入一些噪声区域；相反，过大的 τ（$\tau=8$）得到的相似性矩阵的值集中在少数几个高响应的位置，虽然过滤掉了噪声区域，但可能只定位到部分目标区域. 因此，需要选择一个合理的 τ 值实现帧间相同部件的精准定位. 图 7 是 TMSCN 采用不同的 τ 时在 MARS 和 LS-VID 上的性能. 如图 7 所示，当 $\tau=6$ 时，TMSCN 取得了最好的性能. 值得注意的是，当 τ 取不同的值时，TMSCN 的性能浮动较小（小于 1%）并且一致优于 Baseline 的性能，说明了 TMSCN 对超参 τ 的鲁棒性.

损失函数. 本节测试了不同损失函数对模型性能的影响. 在 MARS，LS-VID 和 iLIDS-VID 数据集上的结果如表 12 所示. 从表 12 可以得到如下结论. （1）交叉熵损失和三元组损失/对比损失的联合使用取得更好的实验效果. 这主要得益于两种损失的互补性. 交叉熵损失将每一个身份的行人当做一个类别，属于分类损失. 三元组损失/对比损失旨在拉近正样本对距离，推开负样本对距离，属于度量损失. 分类损失可以近似看作为学习样本在特征空间的分类面. 度量学习进一步缩小类内距离，增大

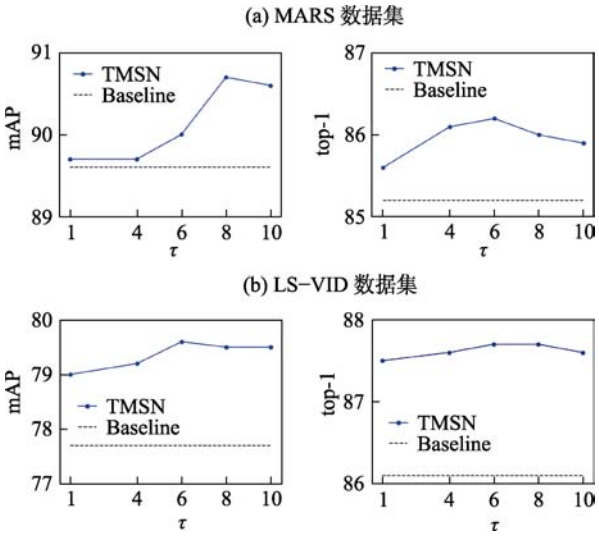


图 7　超参数 τ 对模型 TMSCN 性能的影响

表 12　损失函数对模型性能的影响（L_1 为交叉熵损失，L_2 为困难样本三元组损失，L_3 为对比损失）

损失函数			MARS		LS-VID		iLiDS
L_1	L_2	L_3	mAP	top-1	mAP	top-1	top-1
√			82.3	87.8	66.4	77.8	74.0
	√		78.8	86.7	72.5	82.6	84.7
		√	78.1	86.3	72.7	82.9	86.0
√	√		**86.2**	**90.7**	78.9	87.2	86.0
√		√	85.7	89.6	**79.6**	**87.7**	**90.0**
√	√	√	**86.2**	90.0	79.2	87.5	86.7

类间距离，可以近似看作样本在特征空间进行聚类. 因此，两种损失联合使用会取得更优的性能. （2）联合使用交叉熵损失和三元组损失在 MARS 数据集上性能更好；联合使用交叉熵损失和对比损失在另外两个数据集上性能更好. 这可能是由于对比损失考虑了更多的正负样本对，一般能取得比三元组损失更好的结果. 但是由于 MARS 数据集含有大量的标注错误[66]，对比损失会引入过多的错误样本对，损害了模型的性能.

损失函数的权重. 本节测试了损失函数的权重对模型性能的影响. 如等式 12 所示，λ_1 和 λ_2 分别为三元组损失和对比损失相对于交叉熵损失的权重. （1）本节首先评估三元组损失权重 λ_1 的影响. 图 8 是 TMSCN 采用不同的 λ_1 时在 MARS 数据集上的性能. 如图 8 所示，当 $\lambda_1=1$ 时，TMSCN 取得了最好的性能. 此外，当 λ_1 在一定范围内取不同的值时（$0.4\leqslant\lambda_1\leqslant2$），TMSCN 的性能浮动很少（小于 1%），说明了 TMSCN 对于三元组损失的权重是不敏感的. （2）本节进一步评估对比损失权重 λ_2 的影

响. 图 9 是 TMSCN 采用不同的 λ_2 时在 LS-VID 数据集上的性能. 如图 9 所示，当对比损失权重为 1 时，TMSCN 取得了最好的性能. 此外，当对比损失权重很小时，$\lambda_2=0.1$，引入对比损失依然能大幅度提升模型性能，在 mAP 和 top-1 上分别提升了 8.6% 和 5.2%. 同时，当 λ_2 在一定范围内取不同的值时（$0.6\leqslant\lambda_2\leqslant2$），TMSCN 的性能浮动小于 1%，说明了 TMSCN 对对比损失的权重是不敏感的. 基于上述分析，本章将不同损失函数设定为 1.

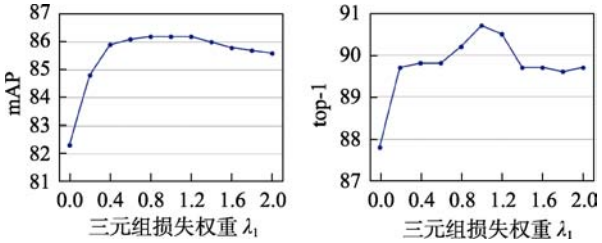


图 8 三元组损失权重 λ_1 对模型 TMSCN 性能的影响. MARS 数据集上的实验结果

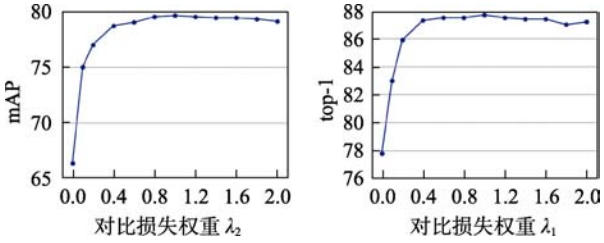


图 9 对比损失权重 λ_2 对模型 TMSCN 性能的影响. LS-VID 数据集上的实验结果

本节进一步测试了自适应权重下模型的性能. 采用文献[67]提出的多损失动态训练方法，在每一次迭代中，根据每个损失值减小的可能性动态确定该损失的权重. 实验结果如表 13 所示. 可以观察到，自适应确定不同损失的权重并不能带来性能的提升. 这是由于 TMSCN 对于损失函数的权重是非常不敏感的，因此在一定范围内波动时不会带来性能的提升. 此外，多损失动态训练方法根据批量数据（mini-batch）的损失值确定权重. 由于批量数据的随机性，该方法无法精确地估计损失值的变化趋势，导致动态估计的权重存在偏差，其性能可能低于本文采用的固定权重策略.

表 13 自适应确定损失权重对模型性能的影响

方法	MARS		LS-VID	
	mAP	top-1	mAP	top-1
固定权重为 1	**86.2**	**90.7**	**79.6**	**87.7**
自适应权重[67]	86.0	90.3	79.4	**87.7**

5.9 跨分支融合模块与多模态融合方法的对比

本节将 CBFM-A 与一些常用的多模态融合方法进行对比分析. 对比的方法包括双线性融合方法、MFB[39]、BAN[40]以及注意力融合方法 DAN[42]、DRA[43]、Transformer[45]. 在 MARS 数据集上的实验结果如表 14 所示. 相比于双线性融合方法[39,40]，CBFM-A 具有更高的性能和更少的计算开销和参数量. 不同于图像、文本等不同模态的特征差异较大，相邻帧具有相同的模态和相似的外观，特征差异较小，因此不需要高阶的双线性融合机制. 此外，双线性融合方法没有考虑相邻帧的空间不对齐现象，可能破坏行人的表观特征，导致性能降低. 相比于 DAN[42]和 DRA[43]，CBFM-A 在 mAP 和 top-1 上分别提升了大约 1%. 这是由于早期的注意力融合方法采用晚融合机制，没有在特征学习时考虑分支之间的关系. 而 CBFM-A 使用早融合机制，在特征学习时融合了相邻帧信息，因此能够学习到表征能力更强的特征. 相比于 Transformer 结构[45]，CBFM-A 在只需 10%的计算开销下取得了略高的性能水平. 这可能是因为 Transformer 引入多头注意力，以数据驱动的方式学习每个注意力捕捉的特征关联. 然而由于行人重识别任务数据量较小，Transformer 可能无法学习到有意义的帧间关联. 相反，CBFM-A 直接计算特征之间的外观相似性，更容易定位到相邻帧的关联区域，从而取得更优的性能.

表 14 CBFM-A 与常用的多模态融合模块的对比分析（GFLOPs 表示融合模块平均处理一帧增加的浮点运算次数，Param.表示融合模块增加的参数量）

	模型	GFLOPs	Param.	mAP	top-1
双线性方法	MFB[39]	0.33	7.5M	85.3	89.6
	BAN[40]	0.12	3.5M	85.2	89.4
注意力方法	DAN[42]	0.11	8.2M	85.3	89.3
	DRA[43]	0.15	17.9M	85.0	89.5
	Transformer[45]	0.34	8.3M	86.0	90.3
	CBFM-A	0.03	0M	**86.2**	**90.7**

5.10 可视化分析

为了定性分析，本节比较了 Baseline 和 TMSCN 特征图的可视化结果. 此外，本节还可视化了一个不含 CBFM 的多分支模型，TMSCN-wo-CBFM. TMSCN-wo-CBFM 的分支之间没有交互，能够清楚地可视化每个分辨率分支独自关注的区域. 如图 10 所示，Baseline 的特征仅仅集中在一个局部区域（黄色上衣），很难区分图中不同行人. 相反，多分辨率分支

结构能够对不同的帧挖掘互补的视觉线索. 从图 10 中可以观察到：（1）低分辨率分支通过降采样输入帧能够关注到更大的空间区域. 比如第三个分支开始关注行人的裤子区域. 通过低分辨率分支关注的互补区域，这些局部相似的行人更容易被模型区分；（2）不同的行人对可能需要不同的分支个数. 比如，图 10 中的（a）和（b）的下衣不同，利用前三个分支的信息便能够很好的区分；（a）和（c）序列的上衣和下衣都非常相似，需要引入第四个分支挖掘的鞋子信息才能够区分开. 这个现象也说明了，相比于 BiCnet 的双分支模型，TMSCN 的多分支结构具有更强的互补性建模能力，能够处理更多的难样本对；（3）尽管低分辨率分支能够关注更多的行人身体区域，但是也引入了噪声区域. 比如图（b）的第三个分支关注的区域扩散到了上衣周围的背景区域. 这是由于低分辨率帧丢失了细节信息，增大了前景和背景的区分难度. 因此，需要将低分辨率分支和高分辨率分支联合使用，一方面利用高分辨率分支更强的判别性，另一方面利用低分辨率分支更多的关注区域，提取到更具表现力的行人特征；（4）TMSCN 利用跨分支传播模块有效整合了不同分支的互补信息. 如图 10 所示，TMSCN 最终整合的特征能够关注到行人多个具有判别性的部件区域，并且不会扩散到背景区域，从而使得局部相似的不同行人变得容易区分.

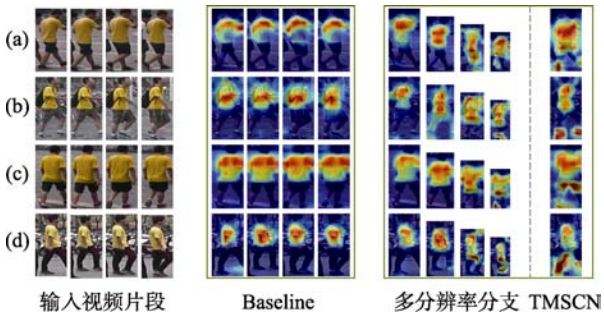


图 10　基准模型（Baseline）和本文方法（TMSCN）特征图的可视化（每个视频片段包含连续的四帧）

6　总　结

本文提出了一个用于视频行人重识别的时序多尺度互补网络. 在大规模的视频行人重识别评测数据集上的大量实验表明，本文提出的方法达到了目前最好的性能，并且相比于大多数现有方法，本文提出的时序多尺度网络没有引入额外的参数量，并且减小了大约 65%的计算开销，说明了本文方法的优越性和可行性. 由于时序多尺度互补网络用不同的输入分辨率和网络结构处理不同的连续帧，避免了目前大多数方法中存在的连续帧特征冗余的问题，因此可以达到更好的性能. 后续工作中，计划在更多的视频任务中进行测试，比如视频人脸识别和视频行为分类，检测本文方法的通用性.

参　考　文　献

[1] Du Yu-ning, Ai Hai-zhou. Learning quadratic similarity function for pedestrian re-identification. Chinese Journal of Computers, 2016, 39(8): 1639-1651 (in Chinese)
(杜宇宁, 艾海舟. 基于二次相似度函数学习的行人再识别. 计算机学报, 2016, 39(8): 1639-1651)

[2] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 480-496

[3] Jinjie You, Ancong Wu, Xiang Li, Wei-Shi Zheng. Top-push video-based person reidentification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1345-1353

[4] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, Qi Tian. Mars: A video benchmark for large-scale person reidentification//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 868-884

[5] Zhang Shun, Gong Yi-hong, Wang Jin-jun. The development of deep convolution neural network and its applications on computer vision. Chinese Journal of Computers, 2019, 42(3): 453-482 (in Chinese)
(张顺, 龚怡宏, 王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用. 计算机学报, 2019, 42(3): 453-482)

[6] Zhou Fei-yan, Jin Lin-peng, Dong Jun. Review of convolutional neural network. Chinese Journal of Computers, 2017, 40(6): 1229-1251 (in Chinese)
(周飞燕, 金林鹏, 董军. 卷积神经网络研究综述. 计算机学报, 2017, 40(6): 1229-1251)

[7] Feiniu Yuan, Lin Zhang, Jinting Shi, Xue Xia, Gang Li. Theories and applications of autoencoder neural networks: a literature survey. Chinese Journal of Computers, 2019, 42(1): 203-230 (in Chinese)
(袁非牛, 章琳, 史劲亭, 夏雪, 李刚. 自编码神经网络理论及应用综述. 计算机学报, 2019, 42(1): 203-230)

[8] Niall McLaughlin, Jesus Martinez Del Rincon, Paul Miller. Recurrent convolutional network for video-based person re-identification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1325-1334

[9] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, Xilin Chen. Temporal complementary learning for video person reidentification//Proceedings of the European Conference on Computer Vision. Virtual, 2020: 388-405

[10] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, Xilin

Chen. Appearance-preserving 3d convolution for video-based person reidentification//Proceedings of the European Conference on Computer Vision. Virtual, 2020: 228-243

[11] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Zhibo Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 10407-10416

[12] Yu Liu, Junjie Yan, Wanli Ouyang. Quality aware network for set to set recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, America, 2017: 5790-5799

[13] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, Shaofan Cai. Region-based quality estimation network for large-scale person reidentification//Proceedings of AAAI Conference on Artificial Intelligence. Lousiana, USA, 2018: 5790-5799

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[15] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, Yi Yang. Exploit the unknown gradually: One-shot video-based person reidentification by stepwise learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 5177-5186

[16] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba. Learning deep features for discriminative localization//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2921-2929

[17] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, Xilin Chen. VRSTC: Occlusion-free video person re-identification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 7183-7192

[18] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, Ling Shao. Learning multigranular hypergraphs for video-based person reidentification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 2899-2908

[19] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, Qi Tian. Spatial-temporal graph convolutional network for video-based person re-identification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 2899-2908

[20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie. Feature pyramid networks for object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 2117-2125

[21] Le Yang, Yizeng Han, Xi Chen, Shiji Song, Jifeng Dai, Gao Huang. Resolution adaptive networks for efficient inference// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 2369-2378

[22] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, Shiliang Zhang. Global-local temporal representations for video person re-identification//Proceedings of the International Conference on Computer Vision. Seoul, Korea, 2019: 3958-3967

[23] Hao Liu, Wei Jiang, Xing Fan, Sipeng Zhang. A survey on deep learning based person re-identification. Acta Automatica Sinica, 2019, 45(11): 2032-2049 (in Chinese)
(罗浩, 姜伟, 范星, 张思朋. 基于深度学习的行人重识别研究进展. 自动化学报, 2019, 45(11): 2032-2049)

[24] Damien Simonnet, Michal Lewandowski, Sergio A Velastin, James Orwell, Esin Turkbeyler. Re-identification of pedestrians in crowds using dynamic time warping//Proceedings of the European Conference on Computer Vision. Firenze, Italy, 2012: 423-432

[25] Taiqing Wang, Shaogang Gong, Xiatian Zhu, Shengjin Wang. Person re-identification by video ranking//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 688-703

[26] Alex Krizhevsky, Ilya Sutskever, Geoffrey EHinton. Imagenet classification with deep convolutional neural networks//Proceedings of the Advances in Neural Information Processing Systems. Neveda, USA, 2012:1097-1105

[27] Shuang Li, Slawomir Bak, Peter Carr, Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 369-378

[28] Xingyu Liao, Lingxiao He, Zhouwang Yang, Chi Zhang. Video-based person re-identification via 3d convolutional networks and non-local attention//Proceedings of the Asian Conference on Computer Vision. Perth, Australia, 2018:620-634

[29] Jianing Li, Shiliang Zhang, Tiejun Huang. Multi-scale 3d convolution network for video based person re-identification// Proceedings of AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019: 8618-8625

[30] Dahjung Chung, Khalid Tahboub, Edward J Delp. A two stream siamese convolutional neural network for person re-identification//Proceedings of the International Conference on Computer Vision. Venice, Italy, 2017: 1983-1991

[31] Le Zhang, Zenglin Shi, Joey Tianyi Zhou, MingMing Cheng, Yun Liu, Jia-Wang Bian, Zeng Zeng, Chunhua Shen. Ordered or orderless: A revisit for video based person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(4): 1460-1466

[32] Shuiwang Ji, Wei Xu, Ming Yang, Kai Yu. 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 221-231

[33] Xiaolong Wang, Ross Girshick, Abhinav Gupta, Kaiming He. Non-local neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 369-378

[34] Thomas N. Kipf, Max Welling. Semi-supervised classification with graph convolutional networks//Proceedings of the International Conference on Learning Representations. Toulon, France, 2017: 1-14

[35] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, Shiguang Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 2014-2023

[36] Chao Zhang, Zichao Yang, Xiaodong He, Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 478-493

[37] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250, 2017

[38] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325, 2016

[39] Zhou Yu, Jun Yu, Jianping Fan, Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering//Proceedings of the International Conference on Computer Vision. Venice, Italy, 2017: 1821-1830

[40] Jin-Hwa Kim, Jaehyun Jun, Byoung-Tak Zhang. Bilinear attention networks//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2018: 1571-1581

[41] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 2048-2057

[42] Hyeonseob Nam, Jung-Woo Ha, Jeonghee Kim. Dual attention networks for multimodal reasoning and matching//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 299-307

[43] Ahmed Osman, Wojciech Samek. Drau: dual recurrent attention units for visual question answering. Computer Vision and Image Understanding, 2019, 185: 24-30

[44] Hao Tan, Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv: 1908. 07490, 2019

[45] Aditya Prakash, Kashyap Chitta, Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 7077-7087

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin. Attention is all you need//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 5998-6008

[47] Sergey Ioffe, Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 448-456

[48] Shanshan Zhang, Jian Yang, Bernt Schiele. Occluded pedestrian detection through guided attention in cnns//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6995-7003

[49] Yiluan Guo, Ngai-Man Cheung. Efficient and deep person re-identification using multi-level similarity//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 2335-2344

[50] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei. Imagenet: A large-scale hierarchical image database// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami Beach, USA, 2009: 248-255

[51] Alexander Hermans, Lucas Beyer, Bastian Leibe. In defense of the triplet loss for person reidentification. arXiv preprint arXiv: 1703.07737, 2017

[52] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, Dilip Krishnan. Supervised contrastive learning//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020: 18661-18673

[53] Adam Paszke, Sam Gross, Francisco Massa, et al. Pytorch: An imperative style, high-performance deep learning library// Proceedings of the Advances in neural information processing systems. Vancouver, Canada, 2019: 8026-8037

[54] Joao Carreira, Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 6299-6308

[55] Zhaofan Qiu, Ting Yao, Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks//Proceedings of the International Conference on Computer Vision. Venice, Italy, 2017: 5533-5541

[56] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich. Going deeper with convolutions// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1-9

[57] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, Xilin Chen. Interaction-and-aggregation network for person re-identification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 9317-9326

[58] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, Tao Xiang. Omni-scale feature learning for person re-identification// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, America, 2019: 3702-3712

[59] Guan'an Wang, Shaogang Gong, Jian Cheng, Zengguang Hou. Faster person re-identification//Proceedings of the European Conference on Computer Vision. Virtual, 2020: 275-292

[60] Hanjun Li, Gaojie Wu, Wei-Shi Zheng. Combined depth space based architecture search for person re-identification// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 6729-6738

[61] Tianyu He, Xu Shen, Jianqiang Huang, Zhibo Chen, Xian-Sheng Hua. Partial person re-identification with part-part correspondence learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 9105-9115

[62] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian. Scalable person re-identification: A benchmark//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015:116-1124

[63] L. Wei, S. Zhang, W. Gao, Q. Tian. Person trasfer gan to bridge domain gap for person re-identification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 79-88

[64] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks//Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4510-4520

[65] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8042-8051

[66] Chih-Ting Liu, Jun-Cheng Chen, Chu-Song Chen, Shao-Yi Chien. Video-based person reidentification without bells and whistles//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 1491-1500

[67] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, Rongrong Ji. Pyramidal person reidentification via multi-loss dynamic training//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 8514-8522

HOU Rui-Bing, Ph.D. candidate. Her research interests include person re-identification, few-shot learning.

CHANG Hong, Ph.D., professor. Her research interests include computer vision, pattern recognition, and machine learning.

MA Bing-Peng, Ph.D., associate professor. His research interests include computer vision and person re-identification.

HUANG Rui, Ph.D., associate professor. His research interests include computer vision and person re-identification.

SHAN Shi-Guang, Ph.D., professor. His research interests include computer vision, pattern recognition, and machine learning.

Background

Person re-identification (reID) aims at retrieving a particular person across multiple non-overlapped cameras. It is widely regarded as a sub-problem of image retrieval. There are two common problems for recent video reID methods. Firstly, existing methods do not take full advantage of rich spatial-temporal clues in the videos. To be specific, the consecutive frames of a pedestrian video are highly similar but the existing methods conduct the same operation on each frame at the same input resolution. As a result, these methods typically produce highly redundant features for consecutive frames. Secondly, it is common for recent methods to improve the accuracy by introducing more operations. However, many devices cannot deploy powerful GPUs that are power-hungry, so the limited computational resources prevent the application of state-of-the-art video reID models. This inspires us to develop efficient video reID models that can fully mine the spatial-temporal clues in the video using as little computations.

In this paper, we propose a powerful and efficient video reID network, namely TMSCN. TMSCN aims to extract complementary features for consecutive frames of a video with less computations. In our design, TMSCN is built on a multi-branch architecture where each branch has a specific input resolution. High-resolution Branch processes frames at original resolution to preserve the detailed visual clues, and Low-resolution Branches with different down-sampling rates are employed to enlarge the receptive field for capturing global information. By feeding consecutive frames into different branches, TMSCN can enforce consecutive frames to focus on the diverse regions with different spatial sizes. Such that the complementary visual features can be discovered for consecutive frames and finally form an integral characteristic of the target identity. To further reduce the computations, TMSCN uses a smaller network for lower-resolution branch. It is reasonable since a lower- resolution image contains fewer details, a smaller network is sufficient to extract its discriminative feature. By down-sampling some frames to low-resolutions and using small network for these frames, TMSCN greatly reduces the computations, requiring about 65% less computation cost than most existing methods. Experiments on three widely studied datasets validate the effectiveness and efficiency of the proposed method.

This research was supported by the National Natural Science Foundation of China under grant numbers 61976203 and 61876171. Before this work, our group has worked on the topic of person reID for more than five years. Some methods proposed by our group have been published in top-tier computer vision conferences and journals such as TPAMI, TIP, TMM, Trans. Cybernetics, TCSVT, TNNLS, CVPR, ICCV, ECCV, etc.