南京大学

研究生毕业论文
(申请硕士学位)

论文题目 基于度量学习和注意力机制的步态识别研究

作者姓名 李雪健

学科、专业名称 计算机科学与技术

研究方向 人工智能

指导教师 申富饶 教授 宋方敏 教授

2021 年 5 月 27 日

学　　号：MG1833044

论文答辩日期：2021 年 5 月 20 日

指 导 教 师：　　（签字）

Research on Gait Recognition Based on Metric Learning and Attention Mechanism

by

LI Xue-Jian

Supervised by

Professor Shen Fu-Rao, Professor Song Fang-Min

A dissertation submitted to
the graduate school of Nanjing University
in partial fulfilment of the requirements for the degree of
MASTER
in
Computer Science and Technology

Department of Computer Science and Technology
Nanjing University

May 27, 2021

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：基于度量学习和注意力机制的步态识别研究

计算机科学与技术 专业 2018 级硕士生姓名：李雪健

指导教师(姓名、职称)：申富饶 教授 宋方敏 教授

摘　要

近年来，随着信息科学技术的不断发展，生物识别技术的应用也越来越广泛。其中，步态识别由于其具有非受控识别、远距离识别以及难以伪装和发现等特性，已经成为了学术界研究的热点领域。然而，步态识别也面临着很多的问题，由于人在三维空间中行走，服装变化，以及走路时与摄像机视角的变化都会对人走路的外表造成很大的影响，造成识别准确率下降。

在此背景下，本文从度量学习的角度出发，提出一种结合基于分类损失和基于距离损失的全新损失函数，利用该损失函数能够提取到更加具有区分性的特征。在此基础上，本文设计并提出了两种注意力机制的施加方式，并成功将所提方法应用到实际场景下。本文的主要内容如下：

1) 从度量学习的角度出发，我们提出了一种全新的损失函数，该损失函数既能够利用到三元组损失直接优化度量学习任务的目标，使得同类样本之间的距离小于不同类样本之间的距离，提取得到有区分性的特征；又能够利用到分类的损失，提取得到有代表性的特征。为了使得网络能够顺利优化，我们提出在将特征送入分类损失的全连接层之前，增加批归一化层，将输入特征的分布调整到一个可学习的分布下。在 CASIA-B 数据集和 TUM GAID 数据集上，我们的方法相比于学术界现有方法都达到了最好的效果。
2) 传统的卷积神经网络具有一定的局限性，我们从人大脑的工作方式中受到启发，提出了两种注意力机制的作用方式，分别是像素级别注意力机制与帧级别注意力机制，其中，像素级别注意力机制能够关注到特征图内部对于识别有帮助的信息，而帧级别注意力机制能够关注到整段序列内有区分力的信息。CASIA-B 数据集上的实验，证明了所提方法的有效性。

3) 我们构建了步态识别系统，将本文所提的方法成功运用在了实际生产环境中。该系统包括前端部分和后端部分，其中前端提供用户界面，并录制包含行人走路姿态的视频，后端对视频进行处理，提取特征，实现行人注册和识别的具体逻辑。

相关的实验表明，本文所提方法具有有效性，并且相较于已有的步态识别方法，识别准确率有了大幅的提升。在相关应用实践中，本文所提的方法也展现出了充分的应用价值。最后，按照本文的工作路线，可以继续开展相关的研究工作。

关键词： 步态识别；度量学习；注意力机制；特征提取；深度学习

南京大学研究生毕业论文英文摘要首页用纸

THESIS: Research on Gait Recognition Based on Metric Learning and Attention Mechanism

SPECIALIZATION: Computer Science and Technology

POSTGRADUATE: LI Xue-Jian

MENTOR: Professor Shen Fu-Rao, Professor Song Fang-Min

ABSTRACT

In recent years, with the rapid development of computer science and technology, the application of biometrics has become more and more extensive. Among them, gait recognition has received extensive attention from academia and industry. Due to its uncooperative and noninvasive manner, gait is very hard to imitate or counterfeit. However, gait recognition also faces many problems. People are walking in a three-dimensional space. Changes in clothing, and changes in the angle of view of the camera while walking will have a great impact on the appearance of the person, resulting in a decrease in recognition accuracy.

In this context, from the perspective of metric learning, we propose a new loss function combining classification-based loss and distance-based loss, which can be used to extract more discriminative features. On this basis, we design and propose two ways to apply the attention mechanism. The proposed methods are successfully applied to actual scenarios. The main content of this thesis is as follows:

1) From the perspective of metric learning, we propose a new loss function that utilizes the triple loss to directly optimize the metric learning goal, which is making the distance between samples of the same class smaller than the distance between samples of different classes; and utilizes classification-based loss to extract representative features. In order to make the network to be optimized successfully, we propose to add a batch normalization layer before sending the features to the fully connected layer of the classification loss. The distribution of features is adjusted to a learnable

distribution, and the network parameters can be successfully optimized at this time. On the CASIA-B data set and TUM GAID data set, our approach achieves the best results, which proves the effectiveness of the proposed approach.

2) The traditional convolutional neural network has certain limitations. Inspired by the mechanism of the human brain, we propose two ways to apply attention mechanisms, namely the pixel-level attention mechanism and the frame-level attention mechanism. The pixel-level attention mechanism can focus on the information that is helpful for identification within the feature map, while the frame-level attention mechanism can focus on the discriminative information in the entire sequence. Experiments on the CASIA-B data set prove the effectiveness of the proposed method.
3) We have built a gait recognition system and successfully applied the methods proposed in this thesis to actual environment. The system includes a front-end and a back-end. The front-end provides a user interface and records a video containing pedestrian walking gestures. The back-end processes the video, extracts features, and also implements the specific steps of pedestrian registration and recognition.

Related experiments show that the methods proposed in this thesis is effective. Compared with the existing gait recognition methods, the recognition accuracy has been greatly improved. The gait recognition system based on the proposed methods in this thesis also shows great application value. Finally, according to the work route of this thesis, related research can be continued.

KEYWORDS: Gait Recognition, Metric Learning, Attention Mechanism, Feature Extraction, Deep Learning

目　　次

插图清单

附表清单

第一章　绪论

1.1　研究背景和意义

生物识别技术在我们的日常生活中扮演着非常重要的角色。我们每一次进行交易，每一次乘坐火车或飞机，甚至每一次使用手机进行解锁，都在应用着生物识别技术。生物识别有很多种方式，可以通过识别的方式将其分为静态生理特征识别和动态行为特征识别[1]。静态生理特征识别方式通过人体的静态特征对人进行识别，包括指纹识别、掌静脉识别和人脸识别等；动态行为特征识别方式利用人体的动态行为特征对人进行识别，包括步态识别、声纹识别和签名识别等。

目前来说，应用范围最广泛以及发展最好的生物识别技术有指纹识别、人脸识别和步态识别。其中，指纹识别技术是通过人指纹的纹理特征进行识别的，它的识别速度非常快，精度也比较高，但是它对环境要求很高，当指纹不干净时无法识别成功；此外，由于部分人群天生指纹特征不明显，甚至没有指纹，这些因素都会导致指纹识别无法成功进行[1]。人脸识别主要是通过人的面部特征进行识别，它的识别不需要与被识别者直接接触，但这种识别方式由于需要高质量的人脸图像而存在很多缺陷。例如，人移动过快或者距离过远都会无法拍摄到清晰的人脸，导致识别效果大打折扣。此外，人脸与摄像头之间的角度、光照，以及不同的妆容都会对人脸识别造成很大的影响，导致无法成功识别人脸。步态识别是利用人走路的姿态对人进行识别。就像世界上没有两片完全相同的叶子，人的走路姿态由于身高、各个身体部位形状、体重、身体肌肉发力习惯，以及之前患过的伤病不同等因素，导致每个人的走路姿态都不尽相同，这些都为步态识别奠定了基础。

相较于其它识别方式，步态识别具有以下优势：首先，步态识别对目标个体的识别距离较远。一般来讲，人脸识别一般要求距离目标 3 米以内，虹膜识别要求至少距离目标 1 米，而步态识别可以在距离目标 50 米时对目标进行识别，这是步态识别相较于其余识别方式的显著优势。由于作用距离非常远，识别难

以被目标发现，能够达到隐蔽状态下识别的效果。同时，步态的识别也不需要被识别者的主动配合，而其余的识别方式或多或少都需要人的主动配合。其次，每个人都有着独特的行走方式，难以伪装。每个人的走路姿态都是长期培养出来的，而且人的身高、体重、体型和身体肌肉发力习惯等等因素都会很大程度上影响走路的姿态，所以很难模仿出其它人的走路姿态。此外，步态识别技术应用成本非常低。现如今，大街小巷都已遍布摄像头，这为采集数据以及分析数据提供了大量帮助，也为步态识别技术的研究与创新创造了方便的环境。最后，步态识别的适应能力更强。人脸识别会很大程度上受到人脸视角、光照和光源位置等外界因素的影响，而步态识别技术的研究一般会考虑到这些因素，期望能够在有这些影响因素时，仍然达到良好的识别准确率。

步态识别技术有如此多的优势，对于该项技术的研究也已经成为了目前学术界的热门方向。步态识别还具有非常大的实用价值。例如，可用于疫情防控下的身份验证领域、监控安防领域和智能家居领域等。

随着 2019 年末新型冠状病毒的爆发，非接触式生物识别技术变得越来越重要。为了避免被识别者与识别者的接触导致细菌的传播，指纹识别方式不再可行；同时，随着卫生意识的逐渐形成，人们前往公共区域都会主动佩戴口罩，这种情况下，人脸识别性能也会受到很大的影响。此时，步态识别作为一种远距离、非接触的识别方式，具有非常大的应用前景。

现在各种公共场所，包括商场，以及街道，都遍布摄像头，用来保证大家的财产安全和人身安全，以及定位锁定嫌疑人的踪迹。然而，由于视频监控太多，对于视频监控的查看需要耗费大量的人力资源。此时，可以利用步态识别技术，对比嫌疑人的走路姿态，找到监控视频中所有嫌疑人出现过的时间节点，从而快速获取嫌疑人的行动轨迹并实施抓捕。

此外，当前大多智能家居设备，识别家中不同用户的方式是通过人脸识别或者语音识别，然而由于两种方式都需要用户与设备距离比较近，或者说需要用户显式地进行配合。此时利用步态识别技术，可以远距离识别不同用户，从而对不同的家庭成员提供个性化服务，达到更加智能的效果。

1.2 国内外研究现状

在介绍研究工作之前，需要先了解一下步态识别领域中数据预处理的流程，以及为什么需要这样的预处理。因为步态识别的研究都会考虑到衣服和背景的变换，如果直接拿彩色图去识别，不同衣服或者不同背景的颜色信息，会对最终的识别造成很大的干扰。因此，在步态识别领域的研究中，一般都会先将图像转换为只包含前景和背景的二值图，其中，行人出现的地方赋值为 1，未出现的地方赋值为 0，这样的二值图，叫做轮廓图（silhouette）。通过这样的表示方式，步态的识别不会依赖于任何的颜色信息，能够得到更加广泛的应用。

如图 1-1 所示，步态识别可以分为训练与测试两个阶段。在训练阶段，主要的任务是得到一个特征提取器，该特征提取器输入预处理后的步态数据，能够有效提取出具有区分性的步态特征。在测试阶段，利用该特征提取器提取待识别的步态特征，与库中的步态特征进行比对，输出识别结果。

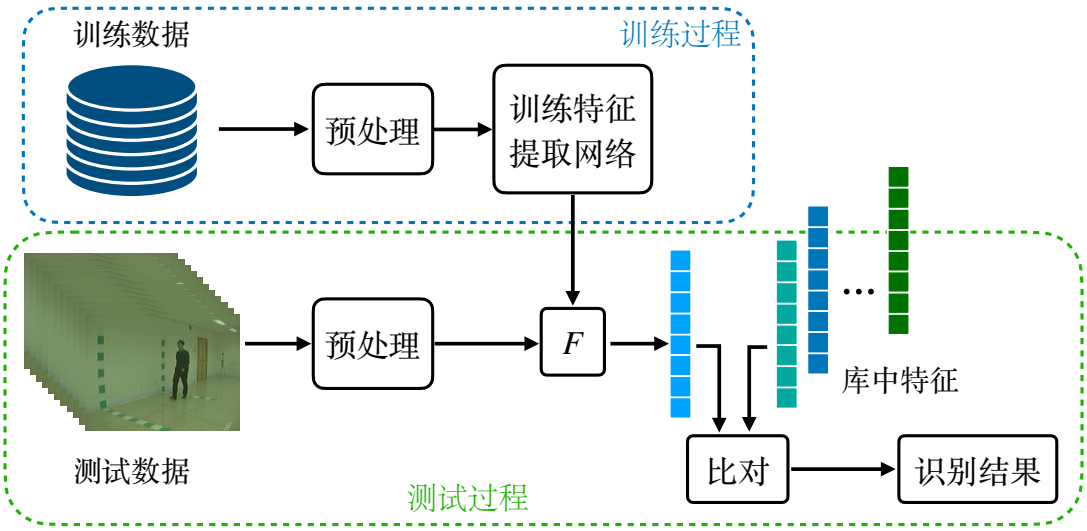


图 1-1: 步态识别整体流程

人的走路姿态可以被很多因素影响[2]，例如走路速度快慢、视角、携带背包情况、衣服变化情况，心情和走路表面变化等等。在实际应用中，视角[3]、携带背包情况[4] 和衣服变化情况[5] 是最需要考虑的因素。由于摄像头和行人前进方向视角的不同，会造成录制的步态视频带来很大的影响；而不携带背包、携带单肩包以及双肩包时都会有不同的走路姿势；同时，在识别中一般会很大程度上依赖腿部的信息，而一旦身着大衣无法得到腿部信息，也会使得识别性能大打

折扣。这些影响因素是学术界和工业界所重点研究的方向，也已取得很多研究进展[6-9]。

早期的步态识别一般都基于传统的机器学习方法，通过手工设计模板来对走路姿态进行匹配，或者利用手工的方式，从图中寻找一些特定的特征，比如腿长，步长等。其中，使用范围最广泛的方法为步态能量图（Gait Energy Image，GEI)[10]。将一个行走周期的轮廓图计算平均值作为行人的步态特征，然后通过主成分分析（Principal components analysis，PCA）算法进行降维，最终采用最近邻的方式识别行人。由于方法简单，计算速度快，当没有涉及视角变化时识别精度高，是最为流行的基于传统机器学习的方法。但是该方法求平均的过程中，会损失很多有用的信息。为了缓解这一问题，学术界提出了步态能量图的替代表示形式，分别是步态熵图（Gait Entropy Image，GEnI)[11]、步态流图（Gait Flow Image)[12] 和染色步态图（Chrono Gait Image，CGI)[13]。除了这一类整合整段序列信息提取特征的方式以外，还有另外一类提取特征的方式，称为基于模型的方法。该方法会将行人的身体部位建模为 2D 或者 3D 模型，然后从该模型中提取特征。典型的方法包括：Luo 等人[14] 通过 3D 行人姿态、体型和衣服估计的方法，使用多个视角的步态轮廓图序列来重建 3D 的步态模型；Wang 等人[15] 利用轮廓图中边缘与重心之间的距离，设计了一种步态特征，叫做区域平均距离（Area Average Distance，AAD），在此基础上，将多个特定视角下训练得到的隐马尔可夫模型学习器的输出结果集成起来，用以提升跨视角步态识别的性能。

随着 Alex 等人[16] 在 2012 年使用深度卷积神经网络大幅提升在 ImageNet 数据集[17] 上的表现，深度学习技术开始逐渐进入人们的视野。深度学习技术不断发展，越来越多计算机视觉任务都利用深度学习技术来解决[18-20]，步态识别也不例外。Wu 等人[21] 最先将深度卷积神经网络应用于步态识别领域，从此之后，越来越多的学者开始研究基于深度学习的步态识别方法。Yu 等人[22] 利用生成对抗网络（Generative Adversarial Network，GAN）结构，将任意视角的步态能量图转换到一个固定的视角，此后有几项工作也基于这个思路[23-25]。Zhang 等人[6] 为了使模型学习得到外表不变的特征，使用自编码器显式地将基于时序的步态特征和基于帧的外表特征拆解开来，达到了非常好的效果。

卷积神经网络虽然能够关注到局部的空间信息，但是无法捕捉到长期的时空依赖性，而这对于视频理解领域，尤其是步态识别任务来说，是至关重要的。

注意力机制[26] 是 2014 年提出的一种模拟人类解决信息过载问题时的方法。人类大脑会无时无刻面临非常多的信息，比如气味、颜色、声音等信息，但是人仍然能够专注于自己的工作，不被一些不相关的信息打扰。这个大脑的工作机制叫做注意力机制。自提出以来，由于其解释性强，且使得网络能够关注于对任务有效的部分，被广泛应用于自然语言处理领域[27] 以及计算机视觉领域[28]。当人类进行步态识别时，其实也是通过被识别个体某些有特点的部位来正确识别，因此，对于步态识别任务来说，使用注意力机制解决是一个合理的、高效的、具有实施性的方案。然而，目前国内外应用注意力机制解决步态识别的任务相对较少。

步态识别本质上是一个度量学习的任务，通过网络的学习，相同身份的样本在特征空间上的距离更加接近，不同身份的样本距离更加远离。而目前国内外对步态识别的研究，大多基于网络结构的改进，从度量学习的角度来看，还有非常大的改进空间。

1.3 本文主要工作

为了解决上述问题，本文从度量学习的角度，提出一种全新的损失函数，能够用来提取更加具有区分性的步态特征。此外，本文还分析了传统卷积神经网络的缺陷，受人脑工作原理的启发，将注意力机制作用于步态识别领域，提出了两种注意力机制的施加方式。最后，将本文所提方法应用于实际环境中，验证了其有效性。本文主要的研究内容总结如下：

- 本文从度量学习的角度考虑，提出了一个全新的损失函数。该损失函数既能够利用到三元组损失直接优化度量学习目标的优势，又能够利用到基于分类的损失函数能够提取得到有代表性步态特征的优势，最终可以学习到更加具有区分性的步态特征。为了能够使两种损失更好地共同优化，我们也对网络结构进行了修改。在 CASIA-B 数据集和 TUM-GAID 数据集上进行的实验结果，都大幅度领先于现有的步态识别方法，证明了我们所提方法的有效性。
- 本文从卷积神经网络的缺陷出发，提出了两种注意力机制的施加方式，分别是像素级别注意力机制和帧级别注意力机制，其中，像素级别注意力机制能够捕捉到空间中的全局信息，提取得到有区分力的行人部位进行识别；帧级

别注意力机制能够提取得到时间上的依赖关系，捕捉人运动过程中具有区分力的身体动作。两种注意力机制可以分开单独使用，也可以共同使用。在CASIA-B 上的实验证明了我们所提方法的有效性。

- 我们将所提方法应用于实际环境中，搭建了基于步态识别的身份验证系统。该系统包括前端和后端部分，具备视频录制、行人注册和行人识别的功能。该系统的实现，进一步验证了我们所提方法的有效性和实用性。

1.4 本文组织结构

本文主要研究了基于度量学习和注意力机制的步态识别方法，设计并提出了一种全新的损失函数，用以提取具有区分力的步态特征，以及提出了基于注意力机制的步态识别算法，最后，将所提方法成功应用于实际的步态识别系统。全文分为六章：第一章为绪论，介绍了研究背景和研究意义，对国内外研究现状进行了总结归纳，还介绍了本文的主要研究内容；第二章介绍了卷积神经网络的基础知识及相关的研究工作；第三章主要介绍结合识别与分类的步态识别损失函数的设计及实验验证；第四章主要介绍基于注意力机制的步态识别网络的设计及实验验证；第五章介绍使用基于本文所提方法的步态识别系统；第六章进行全文的总结，以及对未来工作进行了展望。

第二章　预备知识与相关工作

在深度学习技术中，视觉任务使用最多的网络结构，莫过于卷积神经网络。它通过权值共享的方式来捕捉图像中的局部信息，利用多个卷积层堆叠的方式来增大神经元的感受野，提取图像的特征。由于本文的工作很大程度上基于卷积神经网络，因此在本章会对相关知识进行介绍。

在本章中，我们将对步态识别的研究工作按照时间顺序进行梳理，希望能够为读者理解本文的具体工作铺垫足够的背景知识。在本章中，我们将解决步态识别的方法分为两类，分别是基于机器学习的方法和基于深度学习的方法。本文基于深度学习技术，相比较于传统机器学习技术，深度学习在步态识别任务上有很大的优势，所以本文侧重于介绍基于深度学习技术的方法相关知识。

2.1　卷积神经网络

一般来说，卷积神经网络中包括如下几个结构的堆叠重复：卷积层、池化层、激活层、批归一化层和全连接层，其中，卷积层通过权值共享的特性，提取输入数据的局部特征；池化层可以减少网络的参数，防止网络在数据集合上过度拟合；激活层可以在网络中引入非线性，增强网络的拟合能力；批归一化层通过将输入特征归一化到统一的维度，能够加速网络的训练，并且使网络训练过程变得稳定；最后通过全连接层的特征映射，将输入特征映射到类别对应的空间。接下来将对这些结构分别进行介绍。

卷积层是卷积神经网络中最为关键的网络结构。通过在输入特征图 $\boldsymbol{X}_I$ 上不断滑动不同的卷积核来得到输出特征图。下面通过一个 3×3 的卷积核，来说明其计算方法，这里为了便于描述，将滑动的步长设置为 1。在卷积核滑动到输入数据的某一个位置 (x,y) 时，输出特征图 $\boldsymbol{X}_O$ 对应位置的计算方式如式 (2-1) 所示。

$$\boldsymbol{X}_{O(x,y)}=\sum_{i=0}^{2}\sum_{j=0}^{2}\boldsymbol{X}_{I(x-i+2,y-j+2)}\boldsymbol{K}_{(2-i,2-j)}+b \tag{2-1}$$

其中，$\boldsymbol{K}$ 表示卷积核中的参数，b 代表偏置参数，一个卷积核对应一个偏置参数。

某一个输入特征图的位置，其周围对应的值与卷积核对应的值分别相乘，然后相加，得到这一位置对应的输出值大小。这一过程如图 2-1 所示。

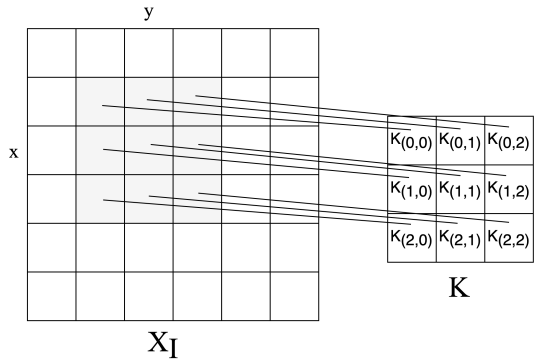


图 2-1: 卷积过程示意图

池化层也是卷积神经网络中的重要一环，有时也会叫做下采样。池化操作整合一个位置与周围相邻位置的信息，这样做的好处在于：（1）可以减少网络的参数量，节省计算资源的同时，防止过拟合；（2）虽然采集到的图像或者视频可能是各式各样的，但我们希望一个同样类型的物体，经过微小的平移、旋转和尺度的缩放，能够提取出同样的特征，通过池化层，可以在网络中引入平移、旋转以及尺度的不变性，达到这样的效果。

常用的池化方式包括最大池化和平均池化。其中，最大池化使用最大值函数，对一个位置及周围的元素进行整合，计算它们的最大值，作为输出值。最常使用的池化窗口大小为 2×2，池化步长为 2。最大池化的示意图如图 2-2 所示。平均池化原理与最大池化相似，只是将最大值函数换成了平均值函数。

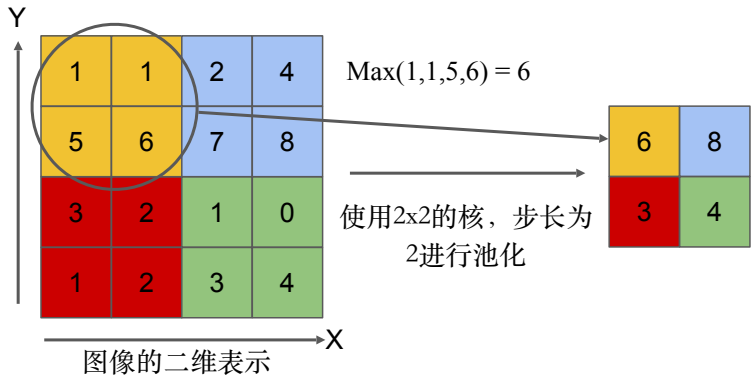


图 2-2: 最大池化示意图

激活层也是人工神经网络中非常重要的组成结构。激活函数的思想，最初借鉴于人类对大脑中神经元的研究。通过整合输入的信息，达到一定状态以后，

这个神经元就变为激活的状态。激活函数应用在卷积神经网络中，最主要的目的是为了在网络中引入非线性。如果网络中全都是线性的结构，那么网络即使结构再复杂，仍然是只能够拟合线性的目标函数。

表 2-1: 激活函数比较

激活函数	表达式	优势	缺陷
sigmoid	$f(x) = \frac{1}{1+\exp(-x)}$	输出在 $(0,1)$ 内，可以被描述为概率，一般会用在分类问题的最后一层	1. 存在梯度消失问题； 2. 函数输出不是以 0 为中心这样会使权重更新效率降低； 3. 指数运算比较慢
tanh	$f(x) = \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$	输出在 $(-1,1)$ 内，以 0 为中心，权重更新效率高，一般会用在隐含层	1. 存在梯度消失问题； 2. 指数运算比较慢
ReLU	$f(x) = \max(0,x)$	1. 在输入为正数的时候，不存在梯度消失问题； 2. 计算速度非常快	当输入是负数的时候，ReLU 是完全不被激活的，也叫做“ReLU 死亡问题”
PReLU	$f(x) = \max(\alpha x, x)$	1. 负数区域内，PReLU 有一个很小的斜率，这样也可以避免 ReLU 死掉的问题； 2. 只有线性运算，非常快	输出没有以 0 为中心

最常见的激活函数包括 sigmoid 函数、tanh 函数、ReLU 函数和 PReLU 函数。它们的优势及缺陷比较如表 2-1 中所示。Sigmoid 激活值可以被解读为概率值，一般用在分类神经网络的最后一层。tanh 激活函数的输出以 0 为中心，权重更新效率高，一般会用在隐含层中。ReLU 激活函数由于不涉及指数计算，同时能够缓解梯度消失的问题，是目前最受欢迎的激活函数，但是它在输入值为负数时输出为 0，这时神经元无法有效学习，这个现象也叫做“神经元死亡”问题。为了缓解这个问题，PReLU 在输入小于 0 时会乘上一个很小的斜率，这个斜率是一个可以学习的参数。

随着网络层数的加深，更新某一层的参数，可能会对整个网络的最终输出造成很大的影响。例如网络第一层的参数发生了细微的变化，但是经过不断的前向传播，最终输出的结果可能差距就会非常大。而在网络训练过程中，将一批数据经过网络以后，通过计算损失函数得到网络中所有参数应该更新的梯度方向，当所有参数都向着这个计算出来的方向更新时，不一定能够更新到使得最终损失更小的方向。这是因为在某一层更新网络参数时，其余层已经更新了参数，而当前层仍然是在其余层旧的参数基础上进行更新的。网络参数的变化导

致输入当前层的数据分布不同，因此之前的参数更新可能都是无效的。这样的现象，称为内部协方差偏移（Internal Covariate Shift）。

为了克服这样的问题，Ioffe 等人[29] 提出了批归一化层（Batch normalization，BN）的网络结构。将该网络结构置于某一层神经网络之前时，经过数据的归一化，能够使得该层网络的各维度输入数据始终位于统一的分布下，从而可以减少内部协方差偏移对该层网络造成的影响。

具体来说，用 x 表示某一个人工神经网络中某一层中一个神经元的输出，在经过归一化后会送往下一层神经网络，用 K 表示一个小批量的大小。首先，使用式 (2-2) 计算这一小批量数据的均值和方差，然后将这一批量的每一个数据，使用 (2-3) 进行归一化，其中 ϵ 是一个微小的正数，用来防止分子是 0 的情况出现。

$$\begin{aligned} \mu_B &= \frac{1}{K}\sum_{i=1}^{K} x_i \\ \sigma_B^2 &= \frac{1}{K}\sum_{i=1}^{K}(x_i-\mu_B)^2 \end{aligned} \tag{2-2}$$

$$\hat{x_i} = \frac{x_i-\mu_B}{\sqrt{\sigma_B^2+\epsilon}} \tag{2-3}$$

这批数据经过归一化后，成为了均值为 0，方差为 1 的正态分布。但是如果这样就送入下一层网络以后，会损失网络很大一部分的表达能力。一方面，这些数据经过 Sigmoid 激活函数时，大多数都会分布于 Sigmoid 函数的线性区域，从而限制了神经网络非线性的表达能力；另一方面，将数据强制规定到这一分布下，也是不太合理的，会失去原始分布的有用信息。因此，通过引入两个可学习的参数 γ 和 β，可以还原神经网络的表达能力，这一步的过程如式 (2-4) 所示。

$$y_i = \gamma\hat{x_i} + \beta \tag{2-4}$$

其中，y_i 是经过批归一化层以后得到的最终结果。

值得注意的是，由于在测试阶段可能只有一个测试样本需要提取向量，或者小批量内的样本数目很少，用这些样本计算出的均值和方差是有偏差的。因此对于批归一化层，在训练阶段和测试阶段的表现是不同的。训练阶段除了按照上述步骤进行前向传播以外，还需要通过每一个小批量内的均值和方差来估

计所有样本整体的均值和方差，用于测试阶段的归一化。

全连接层是一个线性映射层。它可以将一个向量从输入空间映射到输出空间，一般位于卷积神经网络最后面的部分。如图 2-3 所示是一个两层的全连接层，这个网络结构也可以叫做多层感知器（Multilayer Perceptron，MLP）。一般应用于卷积神经网络最后的部分，以将提取得到的特征，映射到训练集对应的类别数上。

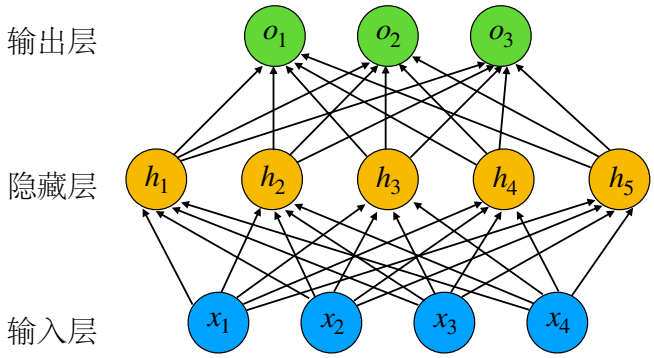


图 2-3: 全连接层示意图

假设某一层全连接层的输入特征维度为 u，输出特征的维度为 v。全连接层中可训练的参数包括两个部分：权重和偏置。首先，通过一个权重矩阵，将输入特征的每个维度按照权重相加，再加上一个偏置，得到一个输出的元素。一组权重，对应一个输出元素，所以一共有 v 组参数，每一组参数包含 u 个权重参数以及 1 个偏置参数，用公式表示如式 (2-5) 所示。

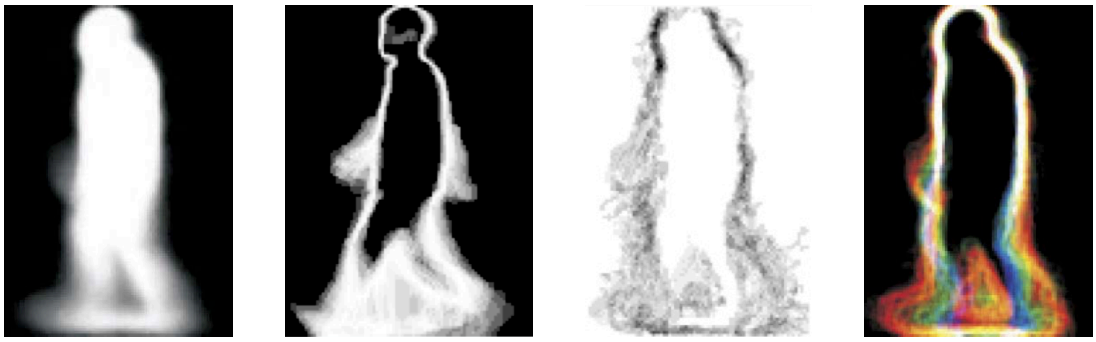
$$\boldsymbol{o}_i = \sum_{j=1}^{u} \boldsymbol{W}_{i,j}\boldsymbol{x}_j + \boldsymbol{b}_i \tag{2-5}$$

其中，$\boldsymbol{W}$ 代表该全连接层中的权重参数，$\boldsymbol{b}$ 代表该全连接层中的偏置参数，$\boldsymbol{x}$ 代表输入数据，$\boldsymbol{o}$ 代表输出的结果，$i \in [1, v]$。

2.2 基于传统机器学习的步态识别算法

早期的步态识别一般都是基于传统的机器学习方法，在本节将会对一些比较典型的算法进行介绍。

为了从轮廓图序列中有效提取步态特征，学术界提出了一种“时空模板”的表示形式，即将轮廓图序列经过运算，整合到一张图上，然后再从整合后的结



(a) 步态能量图 (b) 步态熵图 (c) 步态流图 (d) 染色步态图

图 2-4: 时空模板方法

果中提取特征，如图 2-4 所示，典型的时空模板方式包括步态能量图（GEI）[10]、步态熵图（GEnI）[11]、步态流图（GFI）[12] 和染色步态图（CGI）[13]。其中，步态能量图最早提出，由于其计算简单方便，效果也好，仍然是目前很多算法的基础特征[30-32]。

如图 2-5 所示，在一个步态周期中，将时间维度上的轮廓图计算平均值，即可得到步态能量图。图中每个像素值的亮度，可以表示该像素位置行人身体出现的概率。

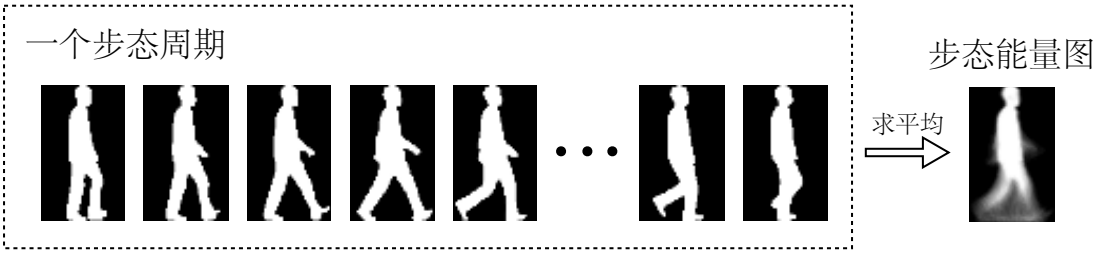
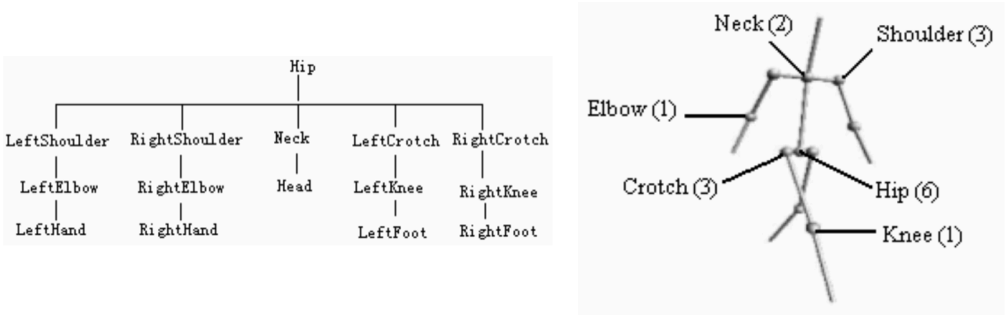


图 2-5: 步态能量图计算示例

从时空模板图中提取特征以后，就可以利用传统的机器学习方法进行训练，如 Han 等人[10] 使用 PCA 算法对特征进行降维，然后通过最近邻的方式进行识别。PCA 算法是一种非监督的降维方法，通过计算样本的协方差矩阵，来衡量每个输入特征维度的信息量，从而对输入特征进行降维，提升计算速度。

时空模板虽然计算速度快，实际表现不错，但是它对于走路姿态变化非常敏感，主要原因在于整合到一张图的过程中损失了大量的信息。为了缓解这个问题，学术界提出了基于模型的方法[33,14]。该类方法首先利用人体 2D 或 3D 的结构对人体进行建模，然后使用模型来获得有区分力的特征。

Zhao 等人[33] 定义了如图 2-6 所示的 3D 人体模型，包括骨架模型和树模型，



(a) 人体的树模型 (b) 人体骨架模型

图 2-6: 人体 3D 模型定义

其中，树模型也叫做外表模型，包括用来描述人体部位的形状参数，骨架模型则用来定义人体的姿势。该方法通过多个视角捕获到的视频建立出人体 3D 模型以后，提取利用下肢的运动轨迹，将其作为动态特征，然后使用线性时间规整的方法来进行识别。

(a) 骨架模型 (b) 3D 模型正面 (c) 3D 模型侧面 (d) 模型轮廓图

图 2-7: 人体 3D 模型模板

Luo 等人[14] 提出了与衣着和携带情况无关的 3D 行人姿势和体型估计方法，通过用如图 2-7 所示的参数化 3D 步态模型模板（template）得到轮廓图，去拟合多个视角录制得到的轮廓图序列，从而重建 3D 人体模型。

这类基于模型的方法对于外表的变化和视角的变化比较鲁棒，并且由于模型是手工设计的，识别结果的可解释性比较强；然而，这类方法一般会对模型的有效性非常敏感，且通常需要相机捕获到质量很好的视频来建立模型。

2.3 基于现代深度学习的步态识别算法

在本节中，我们将基于深度学习的步态识别方法分为两类：第一类是端到端的方式，通过设计不同的网络结构与损失函数，来有效提取步态特征，提高识别精度；第二类是先进行视角或者走路情况的转换，然后利用转换后的结果再进行识别，期望能够提升跨视角或者不同走路情况下的识别准确率。接下来将分别介绍这两类方法。

前者因为是端到端的形式，可以直接通过网络达成我们的目标，但是如果网络或者损失函数设计得不好，可能很难能够提取得到具有区分力的特征；而后者先进行转换的方式，能够一定程度上减轻识别网络的压力，但是最终识别的效果很大程度上依赖于转换后的效果，如果转换过程中损失了一定的身份信息，那么识别效果就会大打折扣。

在实践中，由于转换的过程中难以保留身份信息，相比较之下，端到端的方式简单直接，效果一般也好于先转换后识别的方式，被学术界和工业界广泛使用。

2.3.1 基于端到端的方法

Wu 等人[21] 最先提出使用深度卷积神经网络来解决步态识别问题。作者设计了如图 2-8 所示的三种神经网络来计算两个输入数据源之间的相似性，区别在于对比两个输入数据源的时机不同，LB 网络直接将两张 GEI 在通道上拼接起来，之后通过卷积来加权求和，模拟两张图相减计算相似度的过程，然后继续通过卷积神经网络；MT 网络先让两张图分别通过参数不共享的卷积网络来提取图像特征，之后将二者也通过拼接和卷积的方式融合起来；而 GT 网络是在特征层面将两者融合起来，再得到一个最终的向量。所有三个网络的输出，都是两个输入数据源属于同一个行人的概率值。通过这样的方式，能够使得网络学习到输入数据之间的相似性。在测试阶段，通过网络找到待识别目标在库中最为相似的样本。

Wu 等人提出的三种网络结构都利用卷积神经网络处理图像的优势来计算相似度，但是这样做的缺陷也是显而易见的：每次想要计算相似度，都需要将两个输入数据共同送入网络，而这是非常耗时的，当库中视频较多时，识别效率比

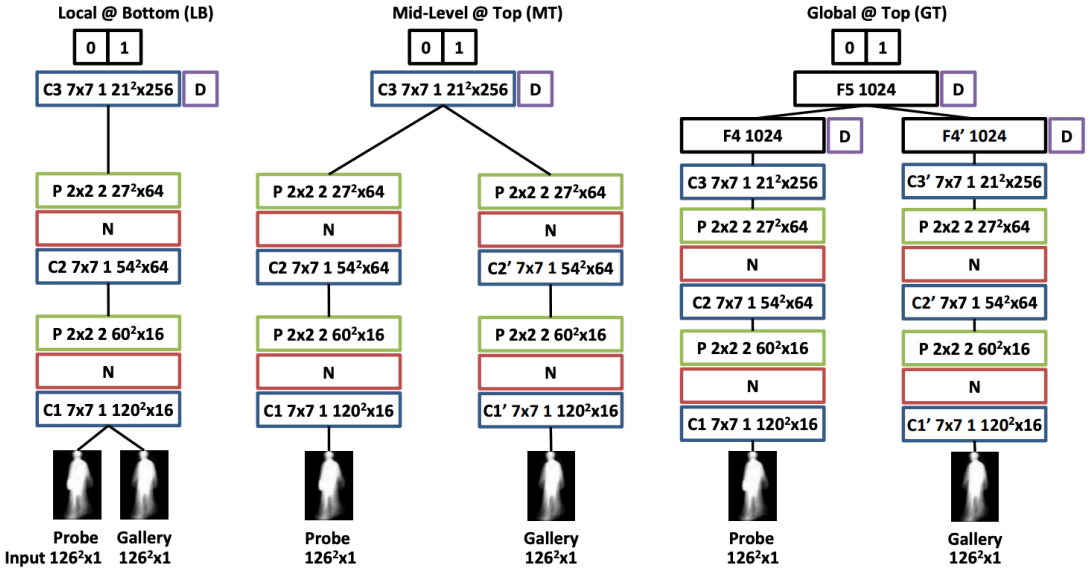


图 2-8: 三种网络结构

较低。但如果我们在库中保存的是提取后的步态特征，可以通过比对待识别的步态特征与库中的特征进行识别，达到非常高的识别效率。因此，后续的研究工作主要专注于提取得到很好的特征表达，而非通过网络计算相似度。

Wolf 等人[34] 提出了一种三维卷积神经网络（3DCNN）结构，该结构如图 2-9 所示，其中输入数据由灰度图和光流图在通道上拼接而成。该网络期望于通过 3DCNN 来整合时间上的信息，但是这种结构也存在着一些缺陷，如输入图像序列的帧数必须是固定的，以及该网络作为一个分类网络，只能够对训练集中出现过的类别达到较好的识别效果，而实际应用中，一般测试阶段待识别的样本身份是没有在训练集中出现过的。

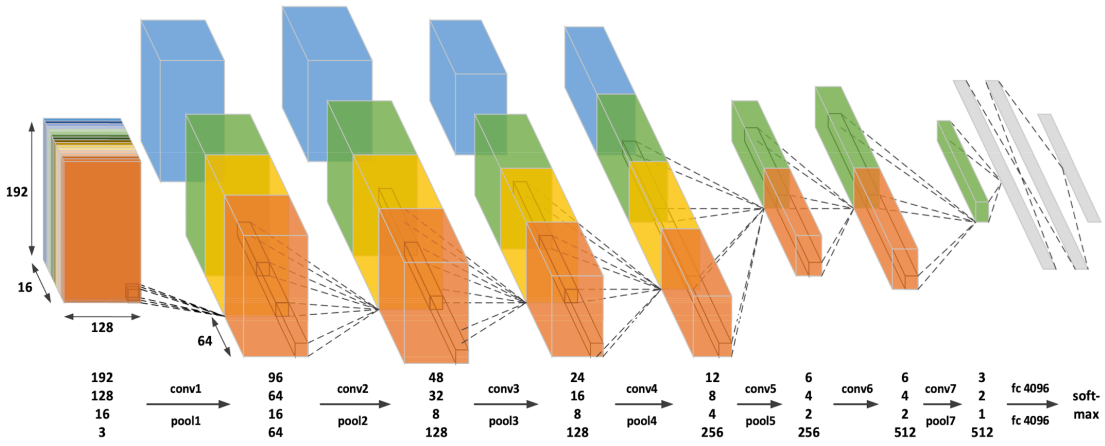


图 2-9: 3D 卷积神经网络结构

为了更好地利用彩色图中的信息，Zhang 等人[6] 提出通过编码器将输入图

像编码为外表特征和步态特征，期望能够将二者显式地拆解开来，单独使用步态特征进行识别。除此之外，网络还包括一个解码器，解码器的输入为两种特征的组合，输出为解码出来的图像。如图 2-10 所示，作者通过精心设计损失函数来引导网络达成特征拆解的目标，所设计的损失函数包括三个部分：(1) 交叉重建损失，基于整段视频都有相同步态特征，但每帧可能都有不同外表特征的假设，将某一帧提取出的步态特征与同一段视频中的其余帧提取出的外表特征输入解码器，使解码得到的结果与外表特征对应原始帧的结果尽量相似；(2) 步态特征相似度损失，不同视频中的同一行人提取出的步态特征应该尽量相似；(3) 身份分类损失，之前的损失都是基于单独一帧的，通过将一段序列的步态特征输入 LSTM 来提取特征，后跟全连接层进行分类，施加一个视频层面的分类损失。这是解决步态识别问题的一个非常好的思路。

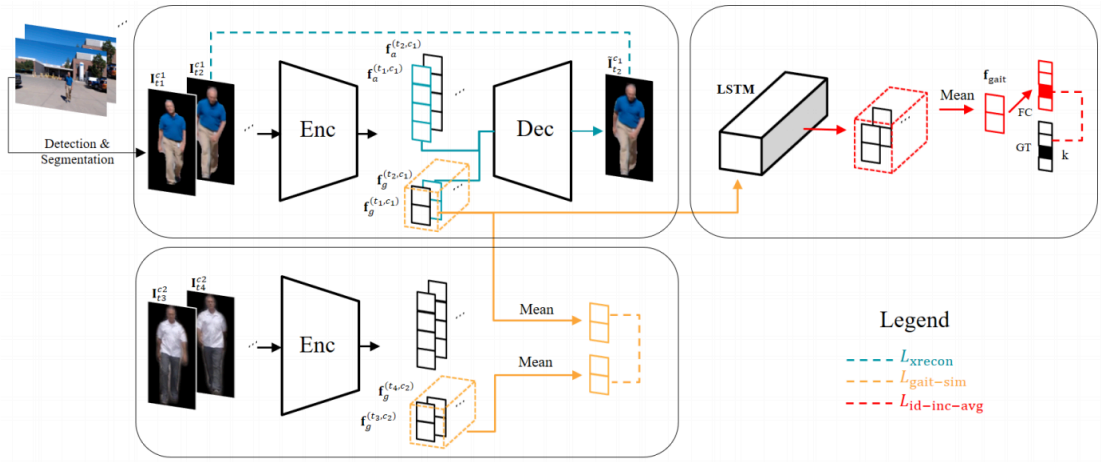


图 2-10: 特征拆解

Chao 等人[9] 基于步态轮廓图的外表中已经包含了彼此之间先后顺序关系的假设，提出将所有步态轮廓图当作一个集合考虑，直接从集合中提取步态特征。这样做的好处在于：输入轮廓图序列的个数以及顺序不再重要，甚至可以将多个视频片段提取出的轮廓图共同输入进网络来提取特征；相比于 3DCNN[34] 来说，有效利用多帧轮廓图中丰富信息的同时，不需要显性地提取时序上的信息，因此并没有引入太多的计算量。

GaitSet 的模型结构如图 2-11 所示。整体上来看，该模型使用卷积神经网络来提取图像特征，通过集合池化（Set Pooling）的方式来将每一帧得到的特征信息整合起来，最后，借鉴于行人重识别领域[35] 的思想，考虑不同行人具有区分力的部位不同，将特征按照不同高度拆分开来，分别进行优化。在测试阶段，将

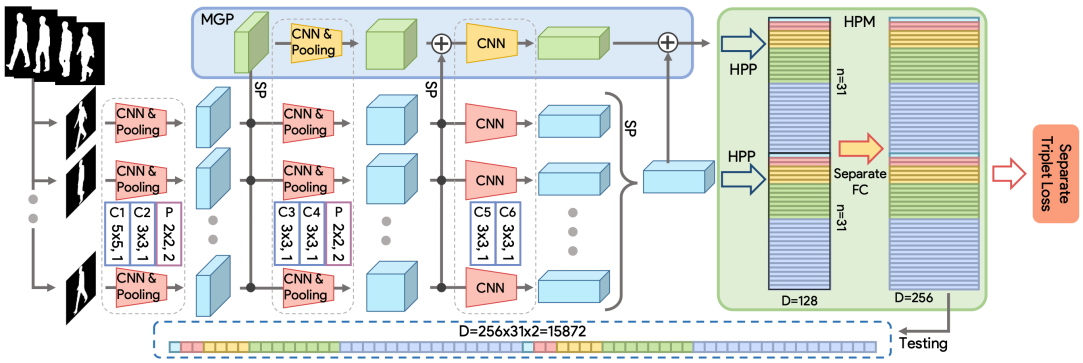


图 2-11: GaitSet 模型结构

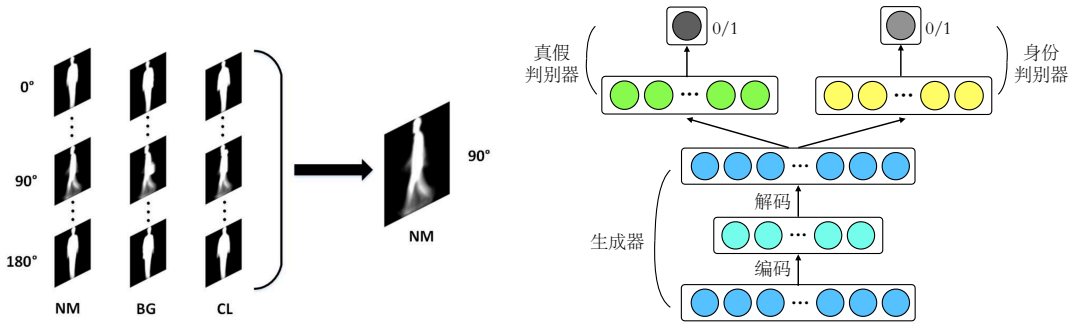
所有不同高度提取出的特征拼接起来，作为最终的步态特征，用于与库中特征进行比对，得到识别结果。

GaitSet 网络性能非常好，训练速度快，识别精度高，是目前步态识别领域最常用的方法，因此，我们第 3 章中将它作为骨架网络，来进行我们的实验。

2.3.2 基于视角转换的方法

另外一类基于深度学习的方法为先转换输入数据的视角，在转换后的数据上进行识别，目标是提升跨视角时的识别性能。

Yu 等人[22]提出了 GaitGAN 模型，如图 2-12 所示，利用生成对抗网络（GAN）将任意视角，以及任意走路情况的步态能量图，转化为一个共同的视角，并且是正常的走路情况下的步态能量图，即没有携带背包和衣服的变换。



(a) 转换为标准 GEI (b) GaitGAN 整体架构

图 2-12: GaitGAN 工作原理

GaitGAN 的模型结构包括三个部分，分别是生成器（G）、真假判别器（D_1）和身份判别器（D_2）。其中，生成器使用自编码器的结构，将一个输入 GEI 图像（记为 I_S）先编码为一个隐向量，再将该隐向量解码为一张生成的 GEI（记为

$\hat{I}_T$）；真假判别器是一个二分类的网络，当输入生成的 GEI 时，网络输出应该为 0，当输入是真实的 GEI 图像（记为 I_T）时，网络输出应该为 1。这个判别器的损失函数为

$$L_{D_1} = -t\log(D_1(I)) + (t-1)\log(1 - D_1(I)), \tag{2-6}$$

其中

$$t = \begin{cases} 1 & \text{if } I \in \{I_i\}, \\ 0 & \text{if } I \in \{\hat{I}_i\}. \end{cases}$$

其中，$\{I_i\}$ 为真实 GEI 图像的集合，$\{\hat{I}_i\}$ 为生成 GEI 图像的集合。

在视角转换的过程中，为了防止损失身份的信息，作者提出了身份判别器。该网络也是一个二分类网络，结构上类似于上文提到的 LB 网络结构[21]，网络给出输入的两幅 GEI 图像来自同一个人的概率。作者在训练过程中引入一个目标域上的且身份不同的 GEI 图像，记为 I_T^-，只有当该判别器的输入为 I_T 和 I_S 时，网络应该输出 1，否则网络都应该输出 0。身份判别器的损失函数表示为

$$L_{D_2} = -t\log(D_2(I_S, I)) + (t-1)\log(1 - D_2(I_S, I)), \tag{2-7}$$

其中

$$t = \begin{cases} 1 & \text{if } I = I_T, \\ 0 & \text{if } I = \hat{I}_T \text{ or } I = I_T^-. \end{cases}$$

通过生成器与判别器相互“博弈”，GaitGAN 模型能够生成看起来足够真实，并且也保留了身份信息的 GEI 图像。然而，直接将输入 GEI 转换为另外一个视角可能是非常困难的，因为视角差距比较大时，GEI 之间的差距是非常大的。为了减小问题的难度，Yu 等人[25] 提出使用多个堆叠的自编码器层，其中每一层只进行 18° 视角以内的转换，降低网络的难度，达到了更好的识别效果。

基于步态图像特征都位于一个低维流形上的假设，He 等人[24] 提出多任务生成对抗网络（MGAN）来进行视角的转换。如图 2-13 所示，该网络结构包括五个部分：编码器，将输入编码为特征；视角分类器：对视角进行分类；视角转换层：在特征的低维流形上进行视角的转换；生成器：将转换后的特征输入生成器来生成转换后的样本；判别器：同时进行三项任务的评价，分别是角度是否转换成功、输出数据是否与输入数据来自同样的分布，以及是否保留住了身份的

信息。

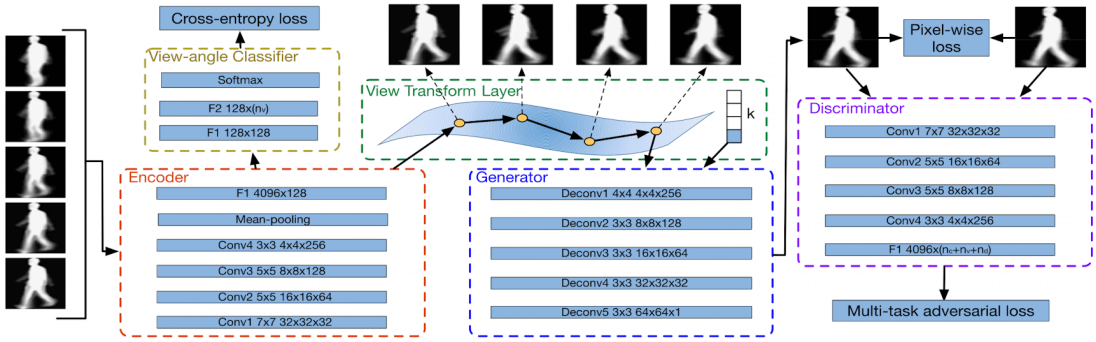


图 2-13: MGAN 模型结构

在低维流形上进行视角的转换时，作者为每个视角与前一个视角之间训练了一个转换向量，来自视角 u 的特征向量 $\mathbf{z}^u$ 可以通过式 (2-8) 转换为视角 v 的特征向量 $\mathbf{z}^v$。

$$\mathbf{z}^v = \mathbf{z}^u + \sum_{i=u}^{v-1} \mathbf{h}_i \tag{2-8}$$

其中，$\mathbf{h}_i$ 代表从视角 $i-1$ 转换为视角 i 的转换向量。

2.4 本章小结

在本章中，我们对步态识别算法进行了梳理，文中将步态识别算法分为了两类，分别是基于传统机器学习的方法和基于现代深度学习的方法。基于深度学习的方法，不再需要人手动设计特征，能够尽可能的挖掘数据的潜在特征，已经成为了主流的做法。其中，卷积神经网络作为视觉领域最重要的神经网络结构，同时也是我们所提方法的基础，我们也在本章中进行了介绍。

受限于篇幅，本章只是对一些典型算法进行了简略的介绍，旨在为读者提供本文工作相关的一些整体概念，为读者理解本文具体工作做好理论铺垫。

第三章　结合识别与分类的步态识别损失函数设计

步态识别，本质上是一个度量学习的任务。而在度量学习中，基于分类的损失函数，能够提取到有代表性的特征，但是对于训练阶段没有见过的类别泛化能力不强；而基于距离的损失函数，能够直接优化度量学习的目标，但是存在训练难以收敛，容易过拟合的缺陷。

我们从度量学习的角度出发，提出了一种全新的损失函数，该损失函数既能利用到三元组损失直接优化度量学习目标的特点，又能够利用到基于分类的损失提取有代表性特征的特点。

3.1　常用损失函数介绍

3.1.1　基于分类的损失函数

基于分类的损失函数，用于解决分类任务时，具有天然的优势，而去掉最后的全连接层，将网络输出作为特征向量，此时也可以用于解决度量学习的任务。特征提取器提取出样本的特征向量以后，首先将这个特征向量映射到训练集中样本身份数量的维度上，得到每个样本的分类结果。通过计算分类的结果与真实结果之间的差距得到损失值，使用反向传播来更新网络参数。

Softmax 损失函数是深度神经网络最常使用的损失函数，它是一种基于分类的损失函数，也叫做交叉熵损失函数。假设训练数据中包含 K 个类别，训练中的一个批量中包含 N 个样本，这时，网络的最后一层全连接层会把每个样本输出为 K 维，某个维度上的值越大，代表网络认为该样本有更大的可能属于该类别。要计算 Softmax 损失，首先需要将网络最后一层的输出归一化到 $[0,1]$ 的范围内，且归一化后的值和为 1，此时可以将输出视为网络将一个样本预测为某一个类别的概率值，然后通过最小化负对数似然，来使得网络的预测结果更接近

真实结果，公式定义如下：

$$\mathcal{L}_{\text{Softmax}} = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{\boldsymbol{W}_{y_i}^T \boldsymbol{x}_i + b_{y_i}}}{\sum_{j=1}^{K} e^{\boldsymbol{W}_{y_j}^T \boldsymbol{x}_i + b_{y_j}}}\right), \tag{3-1}$$

其中，y_j 为第 j 个样本的真实标签，$\boldsymbol{x}$ 代表网络最后一层全连接层的输入向量，一般被看作是提取得到的特征向量，维度为 d，$\boldsymbol{W}_j$ 为最后一层全连接权重参数中的第 j 列，b_j 为最后一层全连接偏置参数中的第 j 个元素，$\boldsymbol{W}_{y_j}^T \boldsymbol{x}_i + b_{y_j}$ 表示全连接层第 j 维的输出，通过单调递增的指数函数来将输出映射为非负数，之后进行归一化。

原始的 Softmax 损失函数，优化的是某一个特征与其对应类别的内积，它对应的决策边界，是一个高维欧式空间中的超平面。而两个向量的内积，天然具有相似度计算的特性。如果我们将输入特征 $\boldsymbol{x}$ 归一化到长度为 1，同时将 $\boldsymbol{W}_j, 0 \le j \le K$ 也归一化到长度为 1，此时内积计算就相当于计算两个向量的余弦相似度，内积越大，说明两个向量越相似。

想要在角度空间中优化 Softmax 损失函数，首先需要将网络最后一层全连接层归一化到长度为一，将偏置置为 0，如式 (3-2) 所示。

$$\begin{aligned} \tilde{\boldsymbol{W}}_j &= \frac{\boldsymbol{W}_j}{\|\boldsymbol{W}_j\|}, 0 \le j \le K \\ b_i &= 0, 0 \le i \le K \end{aligned} \tag{3-2}$$

其中，$\tilde{\boldsymbol{W}}_j$ 表示长度归一化为 1 后的全连接层参数。此时，将输入向量通过这一全连接层，结果如式 (3-3)，其中，$\tilde{\boldsymbol{x}}$ 为在 $\boldsymbol{x}$ 方向上的单位向量，有 $\|\tilde{\boldsymbol{x}}\| = 1$。这时，将输入特征向量 $\boldsymbol{x}$ 经过全连接层后的结果表示为了其与全连接中参数夹角的余弦值，这样的全连接层，叫做“角度线性层”。此时的分类边界变为了角度空间。

$$\begin{aligned} \tilde{\boldsymbol{W}}_j^T \boldsymbol{x} + b &= \tilde{\boldsymbol{W}}_j^T \tilde{\boldsymbol{x}} \, \|\boldsymbol{x}\| \\ &= \cos(\theta_{j,\boldsymbol{x}}) \, \|\boldsymbol{x}\| \end{aligned} \tag{3-3}$$

基于上述计算，可以得到修改后的 Softmax 损失函数的计算方法，如式 (3-4) 所

示。

$$\begin{aligned}\mathcal{L}_{\text{modified-softmax}} &= \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{\mathrm{e}^{\boldsymbol{W}_{y_i}^T \boldsymbol{x}_i}}{\sum_{j=1}^{K}\mathrm{e}^{\boldsymbol{W}_{y_j}^T \boldsymbol{x}_i}}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{\mathrm{e}^{\|\boldsymbol{x}_i\|\cos(\theta_{y_i,i})}}{\sum_{j=1}^{K}\mathrm{e}^{\|\boldsymbol{x}_i\|\cos(\theta_{j,i})}}\right),\end{aligned} \tag{3-4}$$

其中 $\theta_{j,i}$ 表示网络对第 i 个样例提取得到的特征向量 $\boldsymbol{x_i}$ 与全连接层参数的第 j 列向量之间的夹角，y_i 代表第 i 个样例对应的身份标签。

我们先分析一下如何增加角度间隔，这里使用一个二分类的例子。有一个属于第一个类别的样本提取出的特征 $\boldsymbol{x}$，修改后的 Softmax 损失函数在二分类的情况下，决策边界为

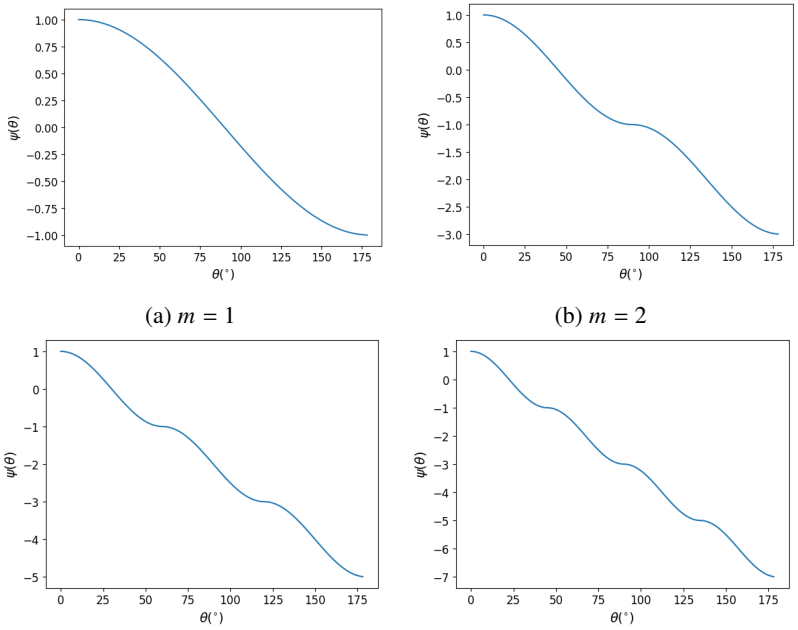
$$\|\boldsymbol{x}\|\left(\cos(\theta_1) - \cos(\theta_2)\right) = 0 \tag{3-5}$$

其中，θ_1 和 θ_2 为待识别样本的特征与全连接层权重两个列向量之间的夹角。要使得网络预测正确，需要有 $\cos(\theta_1) > \cos(\theta_2)$，$\cos(\theta)$ 函数在 $\theta \in [0,\pi]$ 时单调递减，先考虑夹角位于这个区间内的情况，那么需要有 $\theta_1 < \theta_2$。为了添加一个角度上的间隔，可以取一个整数 $m \geq 1$，要求有 $m\theta_1 < \theta_2$ 时才判定网络的分类是正确的。这就是带有角度间隔的 Softmax 损失函数（Angular Softmax, A-Softmax）的思想。

A-Softmax 损失由 Liu 等人提出[36]，它通过增加一个角度的间隔来使提取得到的特征更加区分开来，从而提高模型的泛化能力，得到更好的识别效果。

为了去掉 $\theta \in [0,\pi]$ 的限制，可以使用基于余弦函数修改的角度函数 ψ，这个函数可以始终保持单调递减的特性，并且在角度属于 $[0, \frac{\pi}{m}]$ 时和余弦值相等，这个角度函数的表达如式 (3-6) 所示，ψ 函数在不同的 m 取值下的函数图如图 3-1 所示，从图中可以直观看出，ψ 始终是单调递减的。

$$\psi(\theta_{y_i,i}) = (-1)^k\cos(m\theta) - 2k \tag{3-6}$$



(a) $m = 1$ (b) $m = 2$

(c) $m = 3$ (d) $m = 4$

图 3-1: ψ 函数示意图

由此，我们可以计算出施加角度间隔以后的损失函数，如式 (3-7) 中所示。

$$\mathcal{L}_{\mathrm{A}} = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{\mathrm{e}^{\|\boldsymbol{x}_i\|\psi(\theta_{y_i,i})}}{\mathrm{e}^{\|\boldsymbol{x}_i\|\psi(\theta_{y_i,i})} + \sum_{j\neq y_i}\mathrm{e}^{\|\boldsymbol{x}_i\|\cos(\theta_{j,i})}}\right). \tag{3-7}$$

3.1.2 基于距离的损失函数

另外一类是基于距离的损失函数，它们一般只在度量学习中有用到。这类损失函数的优化目标直接是度量学习的目标：学习到具有区分性的特征表达。在理想情况下，这个特征表达应该具有这样的特点：同一个身份得到的样本，经过特征提取器后提取得到特征，特征之间最大的距离，要小于不同身份的样本提取得到的特征之间的最小距离。换句话说，最大的类内距离，小于最小的类间距离。基于这个特点，可以在测试阶段使用最近邻分类器，对待测试样本进行识别。

一个非常常用的基于距离的损失为三元组损失。一个三元组包括三个从数据集中采样出来的样本：首先从数据集中采样出一个样本 x_a，作为锚点（anchor），然后采样出一个对应身份与锚点相同的样本 x_p，称作正例（positive），最后采样出一个对应身份与锚点不同的样本 x_n，称作负例（negative）。

如图 3-2 所示，为了达到度量学习的要求，三元组损失优化的目标是使得

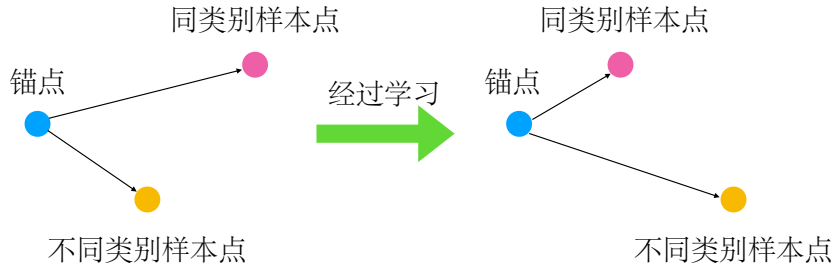


图 3-2: 三元组示意图

同类别对应样本之间的距离尽可能小，同时使得不同类别对应的样本之间的距离，要大于同类别对应样本之间的距离，当这两者之间的差值大于一个提前设定好的阈值以后，就不再更新这一个三元组。使用式 (3-8) 计算这一个三元组对应的三元组损失。

$$\mathcal{L}_{a,p,n} = \max(0, m + D(f(x_a), f(x_p)) - D(f(x_a), f(x_n))) \tag{3-8}$$

其中 f 为设计作为特征提取器的人工神经网络，$D(\cdot,\cdot)$ 代表计算两个向量之间的欧式距离，m 是一个需要提前设定好的阈值，该值控制同类样本之间的距离与不同类样本之间的距离相互之间的间隔。

接下来会介绍两个三元组采样的方式：小批量三元组全采样（Batch-ALL Triplet loss）和小批量困难三元组采样（Batch-Hard Triplet loss），这些采样方式最初是由 Hermans 等人[37] 提出。这两种三元组采样方式的共同之处是，在训练集的身份中随机地采样 P 个行人，其中每个行人有 Q 个包含步态的视频。对于小批量三元组全采样方式，是取这一批量的数据中，所有符合条件的三元组来计算损失，这里的条件指的是，一个样本作为锚点，一个样本对应身份与锚点身份相同，最后一个样本对应身份与锚点身份不同，这样构成一个三元组。

综上所述，小批量三元组全采样方式得到的三元组损失可以如式 (3-9) 进行计算。

$$\mathcal{L}_{\mathrm{BA}} = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{Q}}^{\text{all anchors}} \overbrace{\sum_{\substack{p=1\\p\neq a}}^{Q}}^{\text{all pos.}} \overbrace{\sum_{\substack{j=1\\j\neq i}}^{P}\sum_{n=1}^{Q}}^{\text{all negatives}} \left[m + d_{j,a,n}^{i,a,p}\right]_+, \tag{3-9}$$

$$d_{j,a,n}^{i,a,p} = D(f(x_a^i), f(x_p^i)) - D(f(x_a^i), f(x_n^j)), \tag{3-10}$$

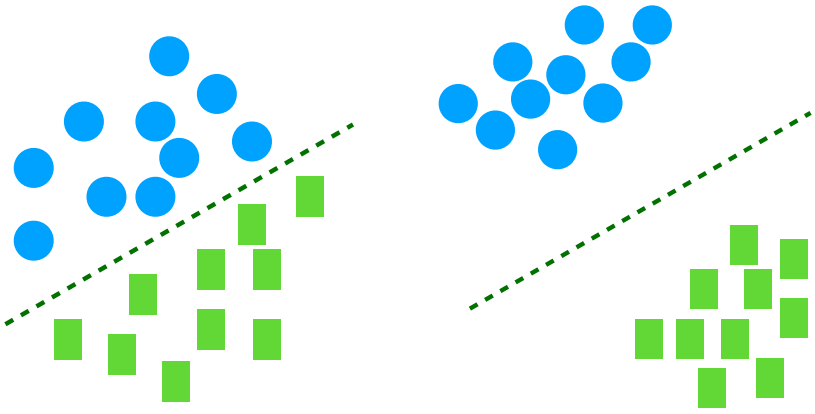
其中，x^i 是从身份 i 采样出来的一个样本。$[]_+$ 表示当中间的部分大于 0 时，取该值，否则取 0，这样做的目的是为了防止网络过度拟合简单样本，同时避免了负损失值的出现。

小批量困难三元组试图去寻找一个小批量中最困难的三元组，相比于在整个数据集中寻找到最困难的正例与负例可能会过于困难，在小批量中寻找困难三元组不会太难，用公式表达如式 (3-11) 所示。

$$\mathcal{L}_{\mathrm{BH}} = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{Q}}^{\text{all anchors}} [m + \overbrace{\max_{p=1\ldots Q} D(f(x_a^i), f(x_p^i))}^{\text{hardest positive}} - \overbrace{\min_{\substack{j=1\ldots P \\ n=1\ldots Q \\ j\neq i}} D(f(x_a^i), f(x_n^j))}^{\text{hardest negative}}]_+. \tag{3-11}$$

3.2 损失函数设计

在介绍我们的损失函数之前，需要先区分特征的两个属性：可分性（separable）与区分性（discriminative）。可分的特征是指，我们可以利用这个特征，以及通过后面的全连接层，可以正确对目标进行分类，即在超球体上，存在一个超球面，能够正确将特征进行分类。而具有区分性的特征，是指我们可以直接利用这个特征，通过最近邻的方式就能实现识别。特征的这两个属性对比如图 3-3 所示。



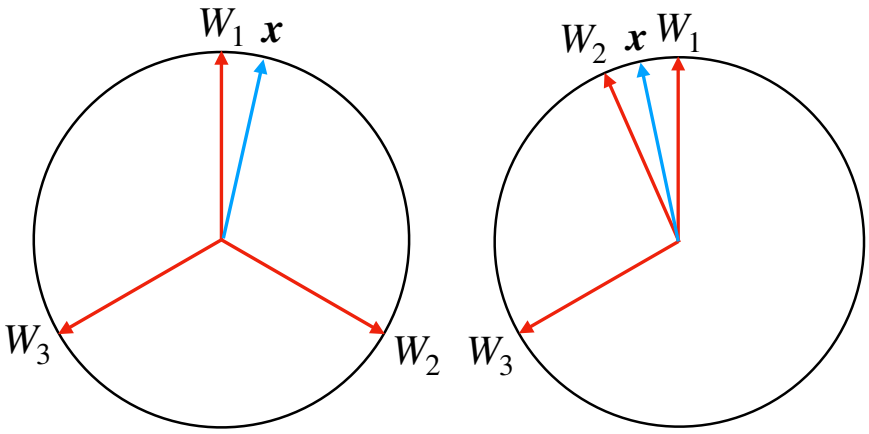
(a) 可分的特征　　(b) 具有区分性的特征

图 3-3: 可分与具有区分性的区别

对于基于分类的损失函数来说，虽然 Softmax 经过改进以后，期望能够在角度空间中增加特征之间的区分性，但是它并不是直接对这个目标进行优化，而

是通过优化特征与其对应类别的“代表向量”之间的相似度，来间接优化度量学习的目标，这里的“代表向量”指的是最后一层全连接网络中对应类别的权重向量。

用一个例子很容易解释这里的思想。考虑一个非常简单的二维特征空间，该空间上有三个类别，如图 3-4 所示。其中，图 3-4a 是我们期望达到的理想情况，在该特征空间下，类别互相之间都分得很开，模型通过计算夹角，很容易就能够正确判断每一个样本特征属于什么类别。然而，实际情况往往都如图 3-4b 所示，在该特征空间内，由于权重向量 W_1 和 W_2 过于接近，待分类的样本特征非常容易落在二者的分类边界附近，导致出现错误分类。实际中特征空间经常是成百上千维，由于特征空间维度过于庞大，这样的现象只会更加严重。这里只是一个非常简单的示例，但是足够说明基于分类损失的问题所在：用于分类的参数向量在特征空间中分布不均衡。



(a) 理想情况下的特征空间　　(b) 实际可能的特征空间

图 3-4: Softmax 损失函数不能直接优化特征距离

除此之外，使用基于分类的损失函数还会带来一个非常大的缺陷：在未知类别上泛化能力不强。由于在实际使用中，往往都是开集的步态识别任务，即测试时的类别在训练集中没有出现过。使用基于分类的损失函数，即使能够在训练集中的类别上很好地分类，当测试阶段有新的类别出现时，往往效果会大幅下降。

那么，为什么还要使用基于分类的损失函数呢？我们认为，基于分类的损失函数虽然单独使用可能并不适用于步态识别任务，但是由于它能够学到非常好的特征表达，即提取到特征以后，能够用该特征进行分类，达到非常好的分类效果，说明提取到的特征中蕴含了正确识别的一些关键因素。这些因素往往是非

常有用的，而将该损失函数与基于距离的损失函数相结合，可以克服掉该损失的一些缺陷，从而更好地应用于度量学习任务。

接下来分析一下三元组损失。使用三元组损失，由于是直接对度量学习的任务目标进行优化，所以能够学到具有区分力的特征，但是它存在一个很大的缺陷：训练非常困难。三元组的选取策略，会对网络参数的更新造成非常大的影响。过于简单的三元组，会使得网络训练过程中没有更新，浪费计算资源，而过于困难的三元组，可能是由于视角变化或者数据获取过程中造成的误差，会导致网络朝着错误的方向更新，因为网络参数还没有优化到能够“消化”这么困难的输入样本。写到这里，读者应该也能够看出来，最好的训练方式，其实是循序渐进，先使用一些比较简单的三元组训练网络，然后逐步提升输入数据的难度，同时也不能全部都是困难的样本，因为网络可能会造成对之前简单数据的“遗忘”现象，导致在简单数据下表现很差，我们在后续实验中也验证了这一观点。所以三元组的选取策略需要依赖大量的经验，三元组损失的训练结果往往波动很大，损失值表现非常不稳定。

因此，我们提出将基于分类的损失与基于距离的损失有效结合起来，期望可以发挥出这两种损失函数的优势，同时也能够减少分别使用两种损失时的劣势。即期望基于分类的损失函数，能够带领三元组损失找对优化的方向；而三元组损失可以直接对度量学习的目标进行优化，同时具有未知类别上的泛化能力，最终达到比较好的结果。我们最终提出的损失函数形式如式 (3-12) 所示。

$$\mathcal{L} = \mathcal{L}_{\text{tri}} + \alpha \mathcal{L}_{\text{A}}, \tag{3-12}$$

其中，α 是一个超参数，用于控制三元组损失和带有角度间隔的 Softmax 损失函数之间的比例，$\mathcal{L}_{\text{tri}}$ 代表三元组损失，$\mathcal{L}_{\text{A}}$ 代表 A-Softmax 损失。

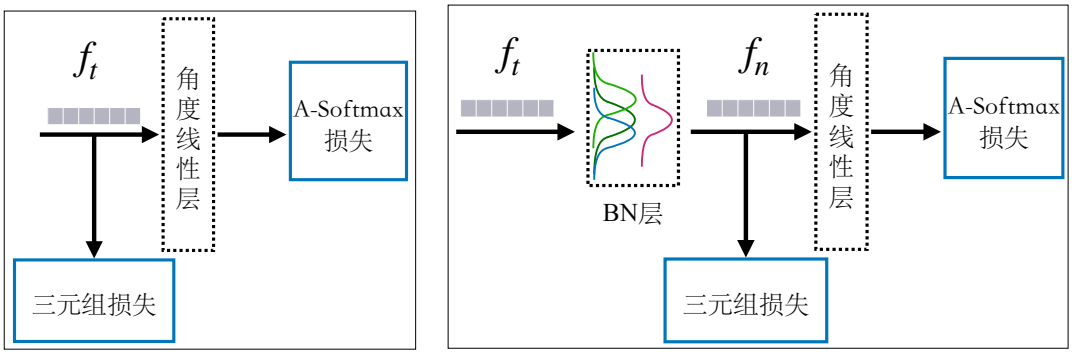
3.3 损失函数的协同优化

带有角度间隔的 Softmax 损失函数主要优化的是余弦空间中的距离，而三元组损失函数主要优化的是欧式空间中的距离。如果直接对两种损失同时优化，可能带来的是其中一种损失减少，另外一种损失增大，或者说两种损失都在不停波动，导致训练无法收敛的现象。

为了使得这两种损失能够共同优化，我们在提取完步态特征以后，使特征通过一个批归一化层。批归一化层通过减去小批量中的均值，并且除以小批量中的标准差，将前一层网络的输出归一化到标准正态分布，然后使用两组可以训练的参数来分别控制新的小批量数据中的均值和方差。如 2.1 节中所述，批归一化层可以减少内部协方差偏移的现象，从而加速深度神经网络的训练过程。增加了批归一化层以后，在该层之前层神经网络输出的变化，对于该层以及之后层的影响会比较小，因为数据都是位于同样的分布之下的。

在我们的模型中，两种损失函数是优化在不同的度量空间中的，通过增加批归一化层可以减少优化两种不同损失带来的影响，从而使得训练过程变得可行，损失能够顺利收敛。

具体来说，如图 3-5a 所示，我们将骨架网络提取出的特征标记为 $\boldsymbol{f}_t$，直接优化两种损失的方式为，使用经过特征 $\boldsymbol{f}_t$ 计算三元组损失，同时将特征送入用于分类的全连接层，将特征分类为对应的身份，计算基于分类的损失。这时，由于两种损失都直接对提取出来的特征施加限制，而这两种损失的更新方向可能相差很远，使得神经网络的参数无法正确更新，最终导致网络无法有效收敛到使两种损失都比较小的位置。



(a) 直接优化两种损失　　(b) 增加 BN 层

图 3-5: 增加 BN 层以优化两种损失

为了缓解这个问题，我们提出在全连接层对特征进行分类之前，先增加一层批归一化层，来将输入特征 $\boldsymbol{f}_t$ 都归一化到相同的分布 $\boldsymbol{f}_n$ 下。如图 3-5b 所示，此时，两种损失能够一起进行优化，从而达到比较好的效果：学习到的特征，通过基于分类损失的优化，能够经过分类得到目标对应的身份，说明这个特征里面包含了识别身份所需的信息；通过基于距离的损失的优化，同一个人特征之间的距离会比较接近，而不同人特征之间的距离会比较远离。

为什么增加 BN 层能够有效减少同时优化两种损失对彼此造成的影响呢?

对于一般的网络结构而言，增加 BN 层可以加速网络的训练，这是因为如果没有经过分布的归一化，网络中参数的变化对于最终输出可能会造成非常大的影响，尤其是网络比较深时，具体来说，当反向传播时，为了使得最终的损失减小，网络中所有层的参数都需要更新，更新的顺序为从后往前，而当前面（较浅层）的参数更新时，由于此时网络后半部分的参数已经改变了，其实已经是在一个全新的网络上面进行更新，此时仍然用一个旧网络上的计算的方向来更新，可能最终更新到一个错误的方向，这就是为什么一般网络必须设置很小的学习率。而经过归一化以后，由于归一化层之后的网络输入分布始终是统一的，它之前的网络参数改变对于它的影响会变得小很多，所以 BN 层能够加速网络的训练，使用更大的学习率。

对于我们的结构而言，其实也是同样的原理。在两种损失同时优化的过程中，增加 BN 层以后，即使两个损失当前想要网络参数朝着不同的方向更新，但是由于 BN 层的存在，输入角度线性层的特征，始终是处于同一分布下的特征，A-Softmax 损失只需要作用于这一分布下的特征，而非任意分布下的特征，三元组损失也是同理，只需要作用于 BN 输出分布下的特征。相当于降低了损失优化的难度，从任意分布的优化变成了某一特定分布的优化，从而使得两种损失能够同时优化。

3.4 步态识别模型的学习算法

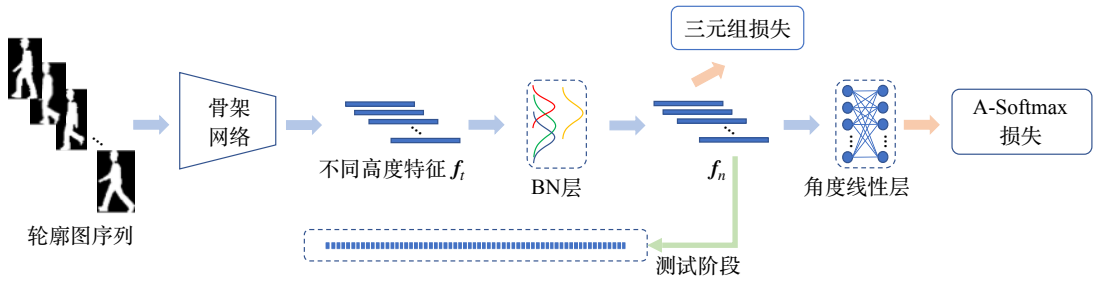


图 3-6: 总体流程

如图 3-6 所示是我们的整体流程，首先，将预处理后的步态轮廓图送入骨架网络来提取特征，我们这里选取目前表现效果最好的 GaitSet 网络[9] 作为骨架网络；然后，将骨架网络提取出的特征送入批归一化层中，将特征的各个维度进

行归一化，利用归一化后的特征计算三元组损失；同时将归一化后的特征送入角度线性层进行分类，利用分类后的结果计算基于角度的 Softmax 损失。在测试阶段，使用归一化后的特征，利用最近邻算法进行识别。

在本节中，我们将会详细介绍训练阶段与测试阶段具体的做法。

3.4.1 训练阶段

在训练阶段中，我们需要得到一个特征提取器，利用该特征提取器能够有效提取步态特征。训练阶段包括以下几个步骤：数据预处理、随机采样、前向传播、损失计算和更新网络权重。其中，由于数据集中已经有处理好的行人轮廓图，预处理步骤需要做的是将轮廓图中的行人对齐到统一大小，这是为了去除掉照相机与目标之间距离的影响。随机采样步骤涉及到三元组的采样，也与一般的深度神经网络不同。这几个步骤是不断重复的，每重复一次，称为一次迭代。

训练阶段的详细过程，参考算法 3.1。

3.4.2 测试阶段

由于实验中所进行的都是开集识别，即训练阶段的行人身份与测试阶段的行人身份之间没有交集。因此，测试阶段的数据需要分为两个部分，其中，第一部分的数据是注册阶段所用的轮廓图序列；另外一部分数据是待识别的数据，需要对这部分轮廓图序列进行身份的识别。

测试阶段包括以下几个步骤：数据预处理、特征提取和特征匹配。其中，数据预处理采用与训练阶段相同的处理方式。特征提取是指将待识别的轮廓图集合输入进训练得到的网络来提取得到步态特征；同时也要提取注册在库中轮廓图序列的步态特征。特征匹配是指使用最近邻算法，在注册的库中进行搜索，找到与待识别的步态特征之间距离最近的库中特征。

测试阶段的详细过程，参考算法 3.2。值得注意的是，在测试阶段，我们会去掉网络的最后一层角度线性层 AL，此时相比于所使用的骨架网络，我们所增加的计算量只有批归一化层的计算量，而在参数固定以后批归一化层的计算量非常少，相比于整个网络来说是微不足道的。

算法 3.1 训练阶段

输入： 训练数据 D，其中包括 K 个行人的 N 个行走轮廓图序列。小批量大小 (P, Q)，用于提取特征的神经网络 F，批归一化层网络 BN，角度线性网络 AL，训练迭代次数 e，两种不同损失函数之间的比例 α，行人对齐的函数 align。

输出： 经过训练后的网络参数 F_Θ，BN_Θ 和 AL_Θ。

for i in $[1, N]$ **do**
  $\tilde{D}_i \leftarrow \mathrm{align}(D_i)$
end for
for i in $[1, e]$ **do**
  从 N 个人中随机采样出 P 个人;
  for j in $[1, P]$ **do**
    从行人 P_j 中随机采样出 Q 段轮廓图序列;
  end for
  将采样出的轮廓图序列合并到一个张量 $\boldsymbol{T}$ 中;
  将张量 $\boldsymbol{T}$ 送入骨架网络 F 中，提取得到步态特征 $\boldsymbol{f}_t \leftarrow F(\boldsymbol{T})$;
  将步态特征 $\boldsymbol{f}_t$ 送入批归一化层 BN，得到批量归一化后的特征 $\boldsymbol{f}_n \leftarrow \mathrm{BN}(\boldsymbol{f}_t)$;
  将批量归一化后的步态特征 $\boldsymbol{f}_n$ 送入到角度线性层, $\boldsymbol{X} \leftarrow \mathrm{AL}(\boldsymbol{f}_n)$;
  使用 $\boldsymbol{f}_n$，根据式 (3-9) 来计算三元组损失 $\mathcal{L}_{\mathrm{tri}}$;
  使用 $\boldsymbol{X}$，根据式 (3-7) 来计算带有角度间隔的 Softmax 损失 $\mathcal{L}_{\mathrm{A}}$;
  计算最终的损失函数 $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{tri}} + \alpha\mathcal{L}_{\mathrm{A}}$
  使用反向传播算法，更新网络参数 Θ;
end for
return 训练后的网络参数 $F_\Theta, \mathrm{BN}_\Theta, \mathrm{AL}_\Theta$.

算法 3.2 测试阶段

输入： 训练后的网络参数 F_Θ，BN_Θ，库中的步态轮廓图序列 G，其中序列的个数为 N，所有库中数据的标签 L，待测试的步态轮廓图序列 q，距离度量方式 $D(\cdot, \cdot)$，行人对齐的函数 align。

输出： 待测试轮廓图序列的识别结果。

$\tilde{q} \leftarrow \mathrm{align}(q)$;
提取测试轮廓图序列的步态特征 $\boldsymbol{f}_{n_q} \leftarrow \mathrm{BN}_\Theta(F_\Theta(\tilde{q}))$;
for i in $[1, N]$ **do**
  $\tilde{G}_i \leftarrow \mathrm{align}(G_i)$;
  提取步态特征 $\boldsymbol{f}_{n_i} \leftarrow \mathrm{BN}_\Theta(F_\Theta(\tilde{G}_i))$;
  计算两者之间的距离 $d_i = D(\boldsymbol{f}_{n_q}, \boldsymbol{f}_{n_i})$;
end for
找到最小距离对应的身份 $m \leftarrow \arg\min_i d_i$;
return L_m.

3.5 实验与分析

在本节中，我们首先会介绍一下使用到的数据集，之后进行大量的实验对比分析。在实验中，我们首先得到跨视角时每个视角下的识别精度，分析不同角

度对于有效识别的重要性程度；我们使用不同三元组的选取策略进行实验，分析三元组选取策略造成的影响；我们使用不同的损失函数进行训练，以验证我们所提损失函数的有效性；此外，我们验证增加批归一化层，是否确实能够更好地共同优化两种损失，带来识别精度的提升；最后，我们与学术界方法进行比较。

3.5.1 数据集介绍

我们在两个广泛使用的步态数据集上进行了我们的实验，以说明所提方法的有效性，分别是 CASIA-B 数据集[4] 和 TUM GAID 数据集[38]。

CASIA-B 数据集[4] 是步态识别领域中使用最广泛的数据集。该数据集由中科院自动化研究所提出，共有 124 个行人，每个行人共采集了 10 段行走的视频。在这个数据集中，考虑到了三种走路情况的变化，分别是：正常情况（NM）、携带背包情况（BG），以及穿着外套情况（CL）。在正常情况下，有六段走路序列，携带背包情况下和穿着外套情况下各有两段走路序列。除此之外，CASIA-B 专注于研究摄像机视角带来步态外表变化的影响，因此每一段走路序列从 11 个视角拍摄。

(a) 正常行走 (b) 背包行走 (c) 穿着外套行走

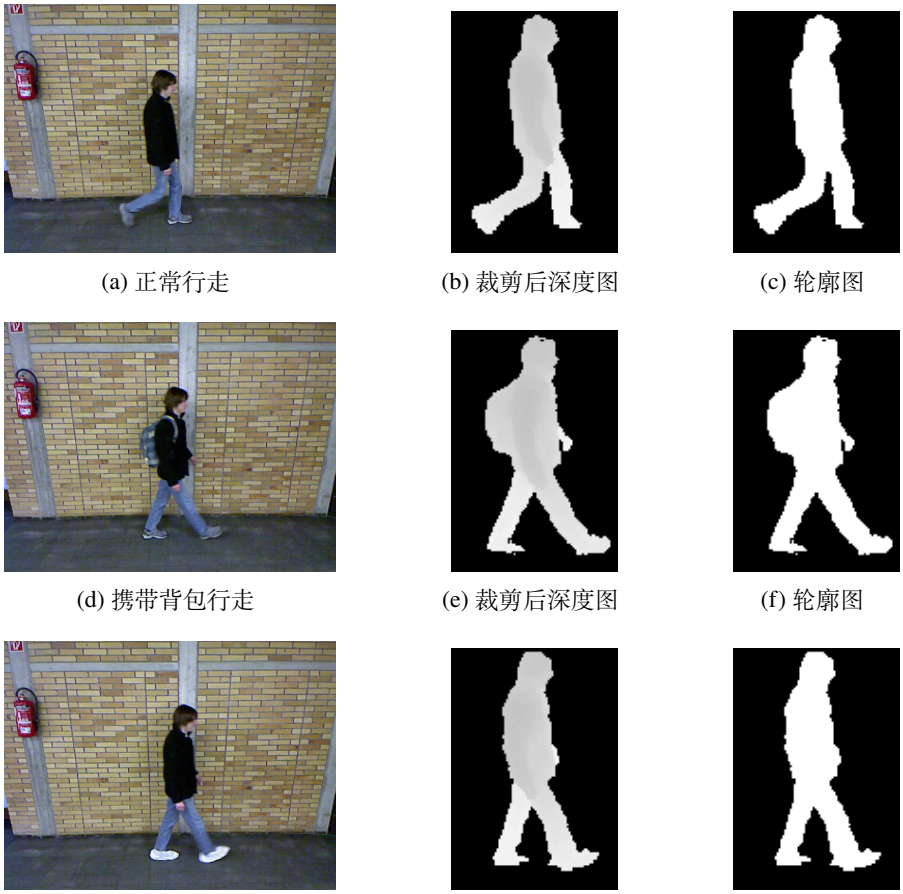
(d) 0° (e) 18° (f) 36° (g) 54° (h) 72° (i) 90° (j) 108° (k) 126° (l) 144°

图 3-7: CASIA-B 数据集示例

如图 3-7 所示为 CASIA-B 数据集的示例，其中，第一行为同一个行人的三种不同走路情况，分别为正常状态下的走路情况、背包状态下的走路情况以及穿着外套时的走路情况，这里展示了行走视频中的一帧画面；第二行为该行人正常走路情况下各个视角摄像机拍到的走路姿态，这里展示的是提取出的轮廓

图裁剪对齐到同样大小后的结果。从这里展示出的示例图像中可以看出，摄像机视角的变化造成走路轮廓图发生了很大程度的变化。

音频、图像和深度采集的 TUM 步态数据集（TUM Gait from Audio, Image and Depth (GAID)）是另外一个在步态识别中被广泛使用的数据集。该数据集考虑了携带情况的变化和鞋类型的变化，采集了 305 个行人的步态数据，其中，每个行人采集了 6 段正常走路的数据（N），两段携带约 5 公斤重的背包的数据（B）和两段穿着带有涂料的鞋的数据（S）。图 3-8 为 TUM GAID 数据集的示例。



(a) 正常行走 (b) 裁剪后深度图 (c) 轮廓图

(d) 携带背包行走 (e) 裁剪后深度图 (f) 轮廓图

(g) 穿着带有涂料的鞋行走 (h) 裁剪后深度图 (i) 轮廓图

图 3-8: TUM-GAID 数据集示例

为了便于处理，我们使用 TUM-GAID 数据集中提供的裁剪后的深度图，通过设定一个阈值来区分目标和背景，从而得到步态轮廓图。

3.5.2 实验细节

对于 CASIA-B 数据集，在训练阶段，我们使用该数据集的前 74 个人对模型进行训练，在测试阶段，将测试数据按照走路情况的变化分为三组。其中每一组

数据，都是使用正常情况下的前 4 段走路序列（NM01-04），作为库中的注册步态数据，分别使用正常情况下的后两段走路序列（NM05-06），携带背包情况下的走路序列（BG01-02），和穿着外套情况下的走路序列（CL01-02）进行识别。

对于 TUM-GAID 数据集，我们使用该数据集发布者描述的划分方案[38]，即 150 个行人用作开发集，其余 155 个行人用作测试集。在测试阶段，这 155 个行人的前四段正常走路序列（N1-N4）作为库中的注册步态数据，同样也有三组不同情况的实验：使用正常情况下的数据（N5-N6）进行测试，使用携带背包情况的数据进行测试（B1-B2），使用穿着带有涂料的鞋进行测试（S1-S2）。

在实验中，我们使用 2.3.1 节中描述的 GaitSet 网络结构作为我们的骨架网络，使用所提的损失函数，利用反向传播算法[39] 来训练网络。其中，我们使用 Adam 优化器[40] 来优化我们的网络参数，将学习率设置为 0.0003，CASIA-B 数据集训练网络 8 万轮，TUM-GAID 数据集训练 4 万轮，这里的每一轮次指对训练集中的数据进行一次采样。我们将采样的小批量大小设置为（8，4）。

3.5.3 跨视角识别精度

因为人在三维空间中的行走方向是不确定的，所以相较于人脸识别，步态识别一般会专注于研究跨视角的步态识别，即注册在库中的视角与待识别样本的视角是不同的，这里的视角指的是行人前进方向与摄像机之间的夹角。值得注意的是，由于视角变化会对行人外观变化造成非常大的影响，所以相比于非跨视角步态识别来说，跨视角步态识别是一个要困难得多的研究任务。

表 3-1: 各个角度的跨视角识别精度

测试集	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	平均精度
NM	92.5	99.0	99.5	98.5	95.8	94.2	95.8	97.9	98.7	98.0	90.9	96.4
BG	89.45	95.09	95.55	94.86	90.36	87.45	90.73	95.09	96.55	95.32	88.09	92.59
CL	70.64	81.00	83.82	79.91	75.00	73.18	74.18	77.73	78.91	78.73	61.91	75.9

表 3-1 中展示的是跨视角时的步态识别精度，表中的每一个数值都是将该列对应的角度作为测试时的角度，库中的角度分别为 0° 到 180° 之间，每隔 18° 的其中一个角度，得到的识别准确率。通过对表中的数据进行分析，我们可以得出以下几个结论：

(1) 0° 和 180° 的识别结果是最差的，这是因为当人向着摄像头走路或者背对向摄像头走路时，摄像头难以捕捉到行走时的步长，以及手臂的摆动幅度，而这些因素可能对于步态识别来说是至关重要的。

(2) 在 90° 时，识别可以利用到步长以及手臂摆动幅度的信息，但是这个视角下采集的数据，会非常容易受到携带背包情况变化的影响，因为这个视角下背包造成的外观变化是最明显的。

(3) 最好的识别结果，出现在视角为 36° 和 144° 时。这是因为，从这两个视角采集的步态数据，既能够捕捉到行走的步长以及摆臂幅度，又能够对于背包变化情况不是非常敏感，从而达到了最好的识别精度。

(4) 从 0° 到 180°，整体的识别精度呈现出了以 90° 为中心，两边对称的趋势，这是由于人正面对着摄像头走，或者背对着摄像头走，通过步态的方式能够捕捉到类似的信息，从而识别精度相差不多。除此之外，能够看出正对着摄像头行走时的识别效果要略微好于背对着摄像头的识别效果，这是由于当人朝着摄像头行走时，可能能够有效利用到人脸的信息共同进行识别，从而达到了更好的结果。

3.5.4 不同三元组损失计算方式

我们评估了不同的三元组损失计算方式对最终识别效果的影响。如表 3-2 中所示，其中“batch all”方式表示使用整个小批量内的所有符合条件的三元组计算三元组损失，“batch hard”表示使用每一个锚点下，最困难的一组三元组来计算三元组损失，“batch all + batch hard”表示先使用 batch all 损失优化网络，然后再使用 batch hard 损失在之前网络的基础上继续优化网络，“batch hard + batch all”表示相反的操作，即先使用 batch hard 优化网络，然后再使用 batch all 优化网络。

值得注意的是，为了防止后期训练中网络参数更新过快，使得网络无法收敛，使用另外一种计算方式微调网络时，我们将学习率调小到了 0.00003。

在表中可以看出，单独使用 batch all 和 batch hard 方式计算三元组，能够得到非常相似的性能，虽然 batch all 每次能够利用很多三元组进行训练，但是当训练轮次足够久，大部分简单的三元组上的损失已经降为了 0，此时继续进行优化，就相当于是在优化小批量内比较困难的三元组了，因此二者表现相近。但

表 3-2: 三元组损失计算方式的影响

计算方式	NM	BG	CL	平均精度
batch all	94.89	88.51	69.66	84.36
batch hard	94.67	88.44	69.18	84.10
batch hard + batch all	95.35	**89.05**	70.70	85.03
batch all + batch hard	**95.62**	88.77	**71.46**	**85.28**

是，当先使用 batch all 方式，再使用 batch hard 方式，或者是调换两者的顺序，相比于之前单独使用某一种损失都有了明显的效果提升。这是由于，batch all 三元组损失可以利用到小批量内整体之间的关系优化网络，而 batch hard 三元组损失利用小批量内最困难的三元组去优化网络，两者的有效结合，可以使得神经网络既能够学习到丰富的整体数据间的关系，又对于困难的样本进行了特殊优化，从而达到了比较好的效果。

3.5.5 损失函数有效性

为了验证本文所提损失函数的有效性，我们设计了三组实验：单独使用三元组损失、单独使用 A-Softmax 损失和两种损失相结合。在 CASIA-B 数据集上的跨视角识别精度如表 3-3 所示。从表 3-3 中可以看出，A-Softmax 损失对特征

表 3-3: 损失函数的影响

损失函数	NM	BG	CL	平均识别精度
三元组损失	95.42	88.54	68.88	84.28
A-Softmax	95.37	89.08	61.15	81.87
A-Softmax + Triplet	**96.10**	**91.88**	**74.36**	**87.45**

施加角度间隔以后再进行分类，在训练集上能够达到很好的分类精度，这说明 A-Softmax 损失函数能够提取得到很好的特征表达，该特征表达中已经蕴含了正确分类行人所需的信息。然而，使用 A-Softmax 损失训练的网络提取出来特征的泛化性能不够强，对于需要开集识别的步态识别任务来说，测试阶段的行人在训练阶段都没有出现过，想要对这些行人有效提取特征，单单使用 A-Softmax 的效果是比较差的。单独使用三元组损失的效果好于 A-Softmax 损失函数，这是

因为三元组损失直接优化的是度量学习的目标：将同类样本特征之间的距离拉近，不同类样本特征之间的距离拉远。但是由于三元组损失波动较大，训练难以收敛，仍然得不到满意的识别结果。

通过将两种损失有效结合起来，既利用到了三元组损失直接端到端优化目标的好处，又利用到了 A-Softmax 能够学习到行人很好的特征表达的优势，达到了比较好的效果。值得注意的是，在最困难的跨视角识别任务 CL 上，结合两种损失使得跨视角的识别精度提升了 5.5%，整体的跨视角识别精度相较于单独使用一种损失提升了 3.1%。

3.5.6 批归一化层的作用

为了验证批归一化层对于同时优化两种损失的作用，我们分别使用未增加批归一化层的网络和增加批归一化层的网络进行训练；除此之外，如图 3-9 所示，在网络中增加批归一化层以后，我们可以取批归一化之前的特征或者批归一化之后的特征计算三元组，同时，在测试阶段，也面临着这个问题，即取哪个特征当做描述行人步态的特征，这会带来非常大的影响。通过对这个影响进行分析，能够进一步说明加入批归一化层的作用。

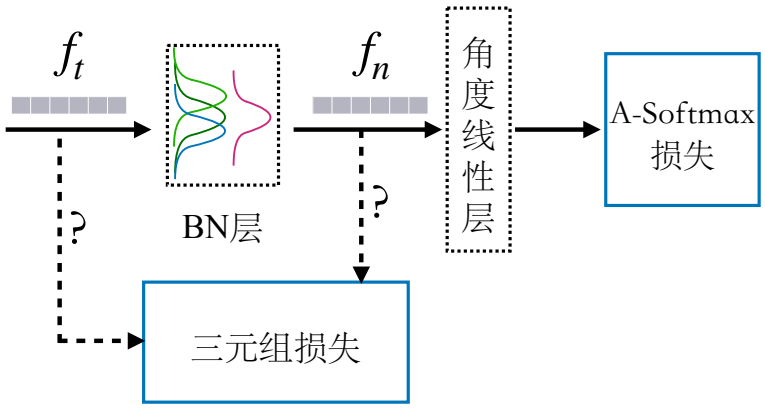


图 3-9: 使用不同特征训练与测试

具体来说，我们设计并进行了五个实验，分别为：

1. 网络中未加入批归一化层，记为 e1；
2. 计算三元组损失用归一化前的特征，测试时用归一化之前的特征，记为 e2；
3. 计算三元组损失用归一化前的特征，测试时用归一化之后的特征，记为 e3；
4. 计算三元组损失用归一化后的特征，测试时用归一化之前的特征，记为 e4；

5. 计算三元组损失用归一化后的特征，测试时用归一化之后的特征，记为 e5；

实验结果如表 3-4 中所示，当未使用批归一化层时，两种损失在不同的空间中进行优化，此时得到的结果很差，这是由于神经网络无法有效朝着使得两种损失共同减小的方向进行优化；而当加入批归一化层以后，尤其是当训练阶段使用归一化后的特征 f_n 计算三元组损失时（e4 和 e5），识别效果都有了显著的提升，这是因为在训练阶段，使用经过批归一化的特征计算三元组，同时，也使用批归一化后的特征通过角度线性层计算 A-Softmax 损失，此时，由于批归一化层的存在，降低了损失优化的难度，使得只需要对某一特定分布下的特征进行优化，而非任意分布下的特征。有效减少了同时优化两种损失对彼此造成的影响，达到了更好的效果。

表 3-4: 批归一化层的作用

实验	NM	BG	CL	平均识别精度
e1	94.96	89.65	69.89	84.83
e2	94.62	89.84	67.55	84.00
e3	95.84	91.66	70.64	86.05
e4	95.66	91.28	72.72	86.55
e5	**96.10**	**91.88**	**74.36**	**87.45**

而如果训练时使用归一化之前的特征 f_t 计算三元组损失，此时，三元组损失经过反向传播能够直接作用于归一化前的特征，而 A-Softmax 损失作用于归一化后的特征，这样导致的结果是三元组损失会对最终特征的形成会造成更大的影响，而失去了 A-Softmax 损失的优势，学习不到很好的特征表达，所以效果要差一些。

至于在测试阶段使用批归一化后的特征 f_n，达到的效果要好于批归一化之前的特征 f_t，这是由于批归一化层的存在，最终提取得到的特征各个维度都处于相同的分布下，此时再进行识别能够取得更好的效果，否则可能会某一个维度波动幅度比较大，计算出来距离值也比较大，这一维度的特征值可能会对最终识别造成很大的影响。

3.5.7 对比实验

在本节中，我们与目前学术界发表的一些效果最好的方法比较结果。如图 3-10，图 3-11 和图 3-12 所示，我们的方法相比于之前的方法，在正常情况下和之前的方法精度相差不多，但是在背包情况下相较于之前的方法提升了 4.6%，在穿着外套的情况下，相比于之前的方法提升了 3.9%，达到了非常好的跨视角识别效果。图中展示的是每个视角具体的识别结果。

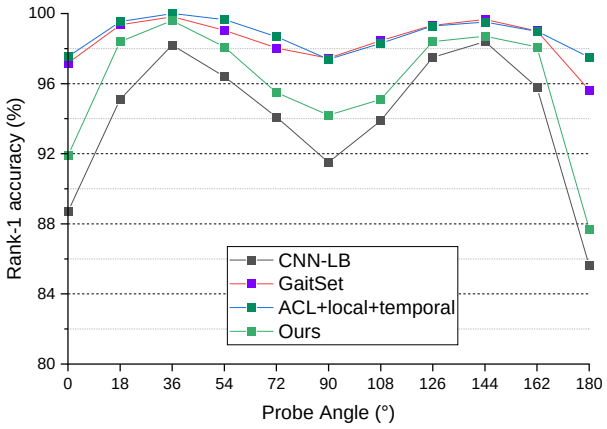


图 3-10: 正常情况的识别结果

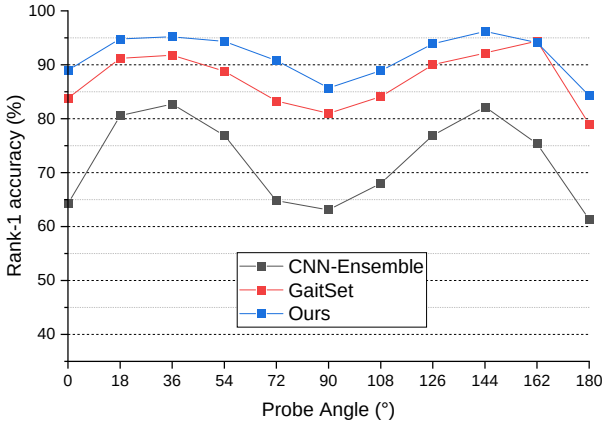


图 3-11: 携带背包情况的识别结果

为了进一步展示所提方法的有效性，我们比较了目前学术界最好的一些方法，包括 MGAN[24]、CNN-Ensemble[21]、CNN-LB[21]、DisGait[6]、GaitSet[9] 和 ACL+local+temporal[8]。如表 3-5 所示是将所有角度求平均的结果，从表中可以更加直观地看出识别的整体表现。这里由于跨视角步态识别任务是一个相对来

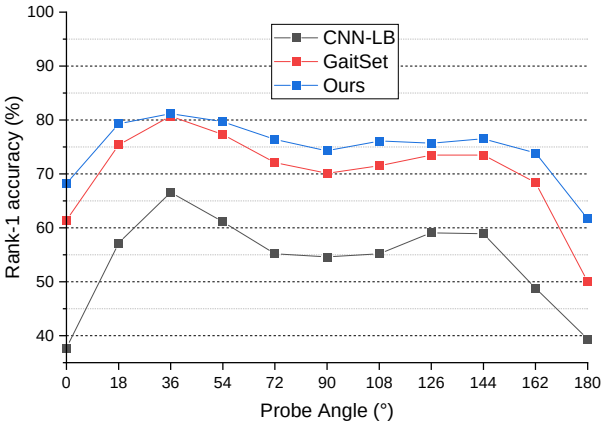


图 3-12: 穿着外套情况的识别结果

说比较困难的任务，所以图中的一些方法没有研究所有的任务，只有其中一部分的结果。

表 3-5: CASIA-B 数据集上的对比实验结果

Method	NM	BG	CL	平均识别精度
MGAN[24]	68.1	54.7	31.5	51.4
CNN-Ensemble[21]	94.1	-	-	-
CNN-LB[21]	-	72.4	54.0	-
DisGait[6]	93.9	82.6	63.2	79.9
GaitSet[9]	95.0	87.2	70.4	84.2
ACL+local+temporal[8]	96.0	-	-	-
Ours	**96.1**	**91.9**	**74.4**	**87.4**

除此之外，我们还在 TUM GAID 数据集上也进行了实验，结果如表 3-6 所示，从表中可以看出，我们的方法几乎能够正确分类 TUM GAID 数据集中的所有测试样本。具体来说，在 N 和 B 测试集中，我们的方法能够完全正确的分类数据集中的样本，对于 S 集合中的测试集，我们方法得到的结果也要好于之前方法的结果。需要注意的是，我们的方法仅仅使用到了数据集中提供的深度图像数据，通过阈值判断前景和背景，非常高效，而[41] 和[42] 中的方法都涉及到光流的计算，是非常耗费计算资源的。

表 3-6: TUM GAID 数据集上的对比实验结果

方法	N	B	S	平均识别精度
GEI[38]	99.4	27.1	52.6	59.7
Fusion Baseline[38]	99.4	59.4	94.5	84.4
TGLSTM[43]	-	-	-	98.4
2D-CNN[41]	99.4	97.7	96.1	97.7
3D-CNN[41]	98.7	91.1	94.5	96.7
CNN-SVM[42]	99.7	97.1	97.1	98.0
CNN-NN128[42]	99.7	98.1	95.8	97.9
Ours	**100.0**	**100.0**	**99.7**	**99.9**

3.6 本章小结

本章主要提出了一种结合识别与分类的损失函数，其中，基于角度的 Softmax 损失虽然能够有效在余弦空间中施加间隔，但是只能够正确分类训练集中出现过的类别，对于开集的步态识别任务来说表现不太好；而三元组损失虽然能够学习到具有区分性的特征，但是三元组损失训练比较困难，损失值容易上下波动。通过两种损失函数有效结合，使得网络能够在提取得到有代表性特征的同时，也考虑到在训练集中未出现过的身份的泛化性能。

基于角度的 Softmax 损失主要优化在余弦空间中，而三元组损失主要优化在欧式空间中，两种损失的优化方向可能是不一致的。为了使两种损失同时优化，我们提出在网络提取出特征之后，将特征先送入批归一化层，用归一化后的特征计算三元组损失与基于角度的 Softmax 损失，由于批归一化层能够使损失只对某一特定分布下的特征优化，而非任意分布下的特征，降低了网络训练的难度，减少了同时优化两种损失时对彼此造成的干扰，从而达到了更好的效果。

最后我们通过大量的实验，展现了所提算法的有效性，在 CASIA-B 数据集与 TUM GAID 数据集上，我们的算法都大幅超过了现有的步态识别算法，达到了非常好的识别精度。

第四章　基于注意力机制的步态识别网络的设计

卷积神经网络虽然能够关注到局部的空间信息，但是无法捕捉到长期的时空依赖性，而这对于视频理解领域，尤其是步态识别任务来说，是至关重要的。

本章提出了使用注意力机制来进行步态识别，该想法借鉴于人大脑的工作方式。具体来说，本章提出了两种注意力机制，分别是像素级别注意力机制与帧级别注意力机制。像素级别注意力能够关注到一张特征图内的关键信息，找到空间上比较有区分性的部位，用于后续识别；帧内注意力能够关注到整段序列内比较重要的信息，可以捕捉到时间上的依赖关系。两种注意力机制既可以分开单独使用，也可以共同使用。

4.1　注意力机制

4.1.1　基础注意力机制

注意力是人大脑中不可或缺的一种复杂认知功能，指人在面对过载的输入信息时，有能力自主选择关注某一些信息，而忽略掉另一些信息。我们会不断的通过听觉、视觉、触觉接触到大量的输入信息，但是我们仍然能够在这些巨量信息中有条不紊地进行我们的工作，正是由于我们有意或者无意中对大部分信息进行了忽略，从而能够专注于少部分的信息。在神经网络的设计中，面临大量的输入信息，也可以借鉴人脑中的注意力机制，选择其中一部分关键信息，提高神经网络的计算效率。

注意力机制最早由 Mnih 等人提出，应用于图像领域[26]，后来由于其可解释性强，而且专注于某一部分而非全局的信息带来了非常好的效果，被广泛应用于之前各种基于循环神经网络的深度学习任务中，包括视觉领域的看图说话[44]、自然语言处理中的阅读理解和文本分类[27] 等任务。注意力机制已经成为了当今研究的热点领域。

为了从 N 个输入向量 $[x_1, \cdots, x_N]$ 中找到与某个特定任务相关的信息，需要对该任务进行建模，引入一个与任务相关的表达，称作查询向量（Query Vector），然后通过打分函数来计算每个输入向量与查询向量之间的相关性。打分函数的计算方式包括加性模型、点积模型、缩放点积模型和双线性模型，其中最经常使用的是点积模型以及缩放点积模型。

点积模型是直接计算两个向量的点积：

$$s(\boldsymbol{x}_i, \boldsymbol{q}) = \boldsymbol{x}_i^T \boldsymbol{q} \tag{4-1}$$

相比较于加性模型，点积模型在实现时可以更好地利用矩阵乘积的优化，从而计算效率更高。但是，当输入向量的维度 d 比较高时，点积后的值通常会有比较大的方差，从而导致优化时计算的梯度值比较小。缩放点积的提出就是为了解决这个问题，它在计算时引入了 $\sqrt{d}$ 进行归一化：

$$s(\boldsymbol{x}_i, \boldsymbol{q}) = \frac{\boldsymbol{x}_i^T \boldsymbol{q}}{\sqrt{d}} \tag{4-2}$$

计算出查询向量与输入向量相似度后，可以通过 Softmax 函数来将所有的相似度归一化为概率值，把这个概率值作为查询向量选择该输入向量的概率，这里因为是使用概率值进行表示，所以是一种“软性”的信息选择机制。具体来说，给定查询向量 $\boldsymbol{q}$ 和输入向量 X 下，选择第 i 个输入向量的概率 α_i 为

$$\begin{aligned} \alpha_i &= p(z = i | X, \boldsymbol{q}) \\ &= \text{Softmax}(s(\boldsymbol{x}_i, \boldsymbol{q})) \\ &= \frac{\exp(s(\boldsymbol{x}_i, \boldsymbol{q}))}{\sum_{j=1}^{N} \exp(s(\boldsymbol{x}_j, \boldsymbol{q}))} \end{aligned} \tag{4-3}$$

选择到想关注的信息以后，还需要对选择要到的信息进行汇总，汇总的方式有两种：软性注意力机制（Soft Attention Mechanism）和硬性注意力机制（Hard Attention Mechanism）。其中，如式 (4-4) 所示，软性注意力机制是将所有向量按照上述关注每个向量的概率加权求和，以得到最终的输出，而硬性注意力直接

取输出概率最大的输入向量，如式 (4-5) 所示。

$$\text{att}(X, \boldsymbol{q}) = \sum_{i=1}^{N} \alpha_i \boldsymbol{x}_i \tag{4-4}$$

$$\text{att}(X, \boldsymbol{q}) = \boldsymbol{x}_j \tag{4-5}$$

其中 $j = \arg\max_{i=1}^{N} \alpha_i$，代表输入数据中模型最关注的向量下标索引。

软性注意力机制通过加权融合所有信息，相比于硬性注意力机制只能够保留一个输入信息，软性注意力机制能够使网络有能力关注于全部输入信息，具备更强的拟合能力，因此，在实际中，会更多地软性注意力机制。

4.1.2 键值对注意力机制

键值对注意力机制是基础注意力机制的泛化形式。对于每一个输入信息，使用键值对来表示（key-value pair）。

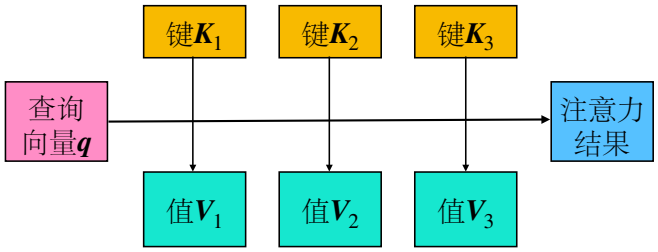


图 4-1: 键值对注意力机制

如图 4-1 所示，对于每一个需要计算注意力的特征 $\boldsymbol{x}$，通过可学习的参数，将其映射为三个向量：查询向量 $\boldsymbol{q}$，键向量 $\boldsymbol{k}$ 和值向量 $\boldsymbol{v}$。在基础注意力机制中，这里的向量 $\boldsymbol{k}$ 和 $\boldsymbol{v}$ 是同一个向量。映射为三个向量以后，就可以用向量 $\boldsymbol{q}$ 在所有的键向量 $\boldsymbol{K}$ 中进行查询，计算相似度，最后按照相似度作为权重，整合所有的值向量 $\boldsymbol{V}$ 中的信息。

将键值对注意力机制用公式表示为式 (4-6) 。

$$\begin{aligned} \text{att}\left((\boldsymbol{K}, \boldsymbol{V}), \boldsymbol{q}\right) &= \sum_{i=1}^{N} \alpha_i \boldsymbol{v}_i \\ &= \sum_{i=1}^{N} \frac{\exp(s(\boldsymbol{k}_i, \boldsymbol{q}))}{\sum_{j=1}^{N} \exp(s(\boldsymbol{k}_j, \boldsymbol{q}))} \boldsymbol{v}_i \end{aligned} \tag{4-6}$$

通过键值对的表示，可以使注意力机制中各个部分分工更加明确，将键和值拆分开来，键单独用来计算相似度，值单独用来表示其内的信息，能够增强网络的表达能力，同时使得网络更容易进行训练。

在实际使用中，为了使得网络具备更加强大的拟合能力，可以对键值对注意力继续进行扩充，将几个参数不共享的注意力机制的结果拼接起来，得到最终注意力的输出结果，这种注意力方式叫做多头注意力机制。经过不同初始化得到的注意力头，可以关注到不同的信息，而这些信息往往对于最终的任务来说，都是有用的。

4.2 基于注意力机制的网络结构设计

在本节中，我们从经典卷积神经网络的缺陷出发，提出基于注意力机制的网络结构。卷积神经网络只能够关注到局部的空间信息，而对于步态的识别，往往需要网络能够捕捉到全局的空间关系，以及时间上的一些依赖关系，此时使用卷积神经网络很难达到期望的目标。如图 4-2 所示，我们提出的结构可以分为三个部分，分别是：(1) 骨架网络，用于提取帧级别的特征，由卷积神经网络堆叠而成；(2) 像素级别注意力模块，用于提取特征图内的关键信息；(3) 帧级别注意力模块，用于提取所有帧中的关键信息。其中，像素级别注意力模块和帧级别注意力模块可以分开使用，也可以共同使用。接下来会分别介绍每个模块的具体组成部分。

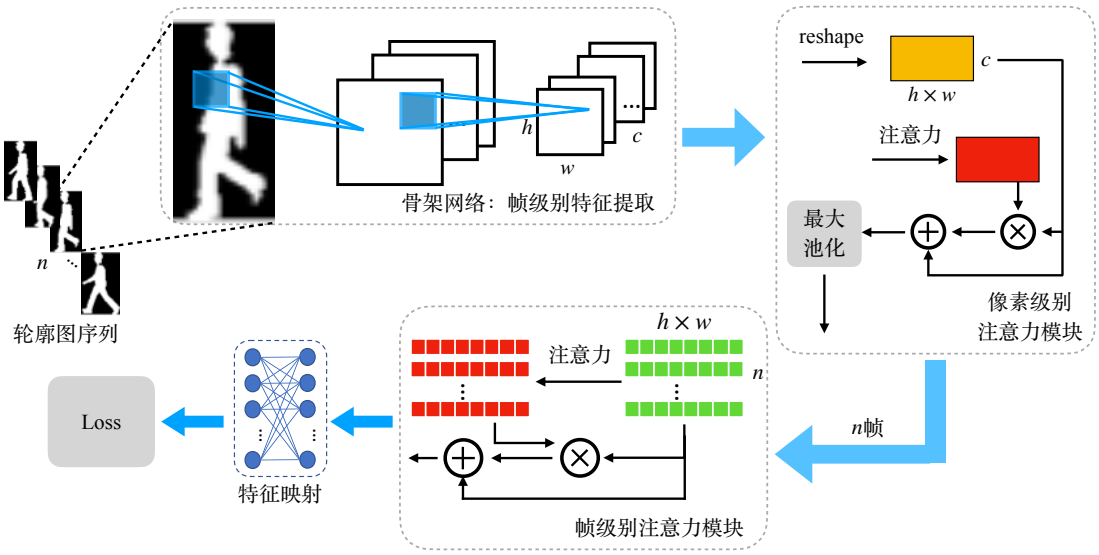


图 4-2: 基于注意力机制的步态识别网络结构

4.2.1 卷积神经网络的缺陷

在开始具体介绍我们提出的模型之前，先来分析一下经典卷积神经网络的缺陷。卷积只能够关注到局部的空间信息，无法捕捉到长期的时空依赖性，而这对于视频理解领域，尤其是步态识别任务来说，是至关重要的。

卷积的设计理念就是，通过卷积核捕捉空间中的局部信息，其中，每一个神经元做出是否激活的决策时，考虑到的空间范围，也就是该神经元的感受野，如图 4-3 所示为一个简单的示例，在该示例中，位于第二层的一个神经元（用黄色标识），其对应到输入图像上的感受野大小为 5×5。从图中可以看出，通过多层卷积网络的堆叠，可以逐步增加卷积核的感受野，但是这样局部信息的堆叠策略，仍然难以建模出长期的空间依赖；而且随着网络深度增加，带来的是表达能力更强，也更容易发生训练集上过拟合的现象，除此之外，更深的网络也会有更严重的梯度消失或者梯度爆炸现象。

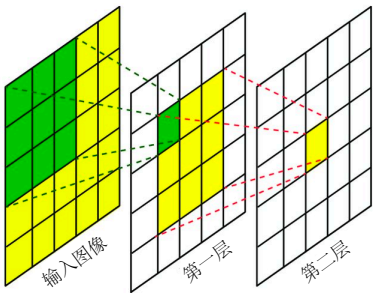


图 4-3: 卷积核的感受野

对于单张图像，卷积网络尚且无法提取到全局的空间信息，更不用说视频中时空信息的提取了。当涉及到视频的处理时，可以使用 3D 的卷积核，就是传统卷积核从三维（长，宽和输入通道数）变为四维（长，宽，时间长度和输入通道数）。但是相比于 2D 的卷积核，3D 的卷积核有更多的参数，也就更加难以训练，而且由于参数过多占用显存，实际处理时往往需要手动将视频进行分段，在分段的过程中其实已经丧失了很多空间上的有用信息。

卷积结构处理图像时，能够提取出局部空间中的有用信息，这一点是大家有目共睹的。因此，我们在网络架构设计中，仍然使用卷积神经网络来提取每一帧局部的空间信息，这一步主要是为了减少后续网络的参数量，避免过拟合的现象发生。

鉴于传统卷积神经网络提取时空信息时的缺陷，我们提出基于注意力机制

的网络架构。该架构首先使用卷积神经网络整合每一帧中局部的空间信息，之后通过注意力机制来整合空间或时间上的信息，挖掘对于身份识别更有价值的时空信息，最后利用整合到的结果，有监督地训练我们的网络参数。

4.2.2 特征提取模块

特征提取模块是我们网络中的第一个模块，用于提取每一帧输入轮廓图中的局部信息，通过局部信息的整合，从图像中提取有用信息，同时减少后续网络中的参数量。

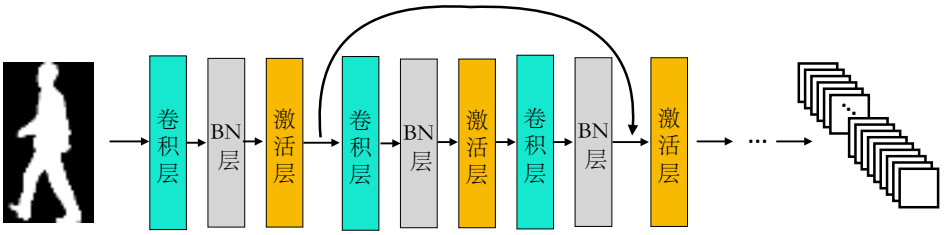


图 4-4: 特征提取网络设计

如图 4-4 所示是我们在本章中使用的特征提取网络，它一共包含有五个卷积层，这里只展示了前三层的网络结构，最后两层卷积的结构与图中的后两层结构相同。每一个卷积后都跟批归一化层，来将卷积得到的结果归一化到相同的分布下。批归一化层后跟激活层，来为网络中引入非线性，增加网络的表达能力。这里的激活函数均使用 ReLU 激活函数。

在网络设计中，我们参考了跨层连接的思想，将第一层卷积后的结果，与第三层卷积后的结果相加，经过激活层后送入第四层卷积层。同样，将第四层卷积层的输入，与第五层卷积的输出相加，经过激活层后作为特征提取模块最终的输出。跨层连接的思想最初由何恺明等人在图像识别领域提出[45]，主要目的是为了解决当网络较深时，梯度消失的问题。通过在网络中增加一个通路，将输入的复制经过该通路，直接作为该通路的输出，这条通路相当于是神经网络中的一个“捷径”。正是由于这条捷径的存在，反向传播时可以保证这条通路上的参数更新不会接近于 0，因此能够成功训练比较深的网络。

需要注意的一点是，对于步态识别来说，输入数据的形式是轮廓图序列，或者说是一个图像的序列，本章中使用的特征提取网络，是将图像序列中的每一帧图像都经过相同的网络结构，提取图像特征，即对于一个序列内的每张图像

来说，这个网络的参数是共享的。

4.2.3 像素级别注意力机制

人在进行步态的识别时，有时，我们能在静态的图像中可能能够判断出某一个人是谁，或者说，某个人有一些独特的外表特征，我们能够一眼认出来，比如肩膀比较宽，或者腿比较粗，仅通过观察图中的一部分，我们就有足够的把握能够正确识别。

从人的识别方式中受到启发，我们设计了像素级别的注意力机制，使得网络有能力关注到每一张特征图中不同位置的重要性，找到图像中的潜在特点，从而利用每一个人有代表性的部位进行识别。

我们使用 n，c，h 和 w 分别代表上节描述的特征提取网络对于每一帧轮廓图提取出特征的维度，其中 n 代表一个视频选取的帧数，c 代表一帧图像提取得到特征图的通道数，h 和 w 分别代表提取得到特征图的高度和宽度，像素级别注意力机制就是基于特征图中每一个位置上的元素做的。

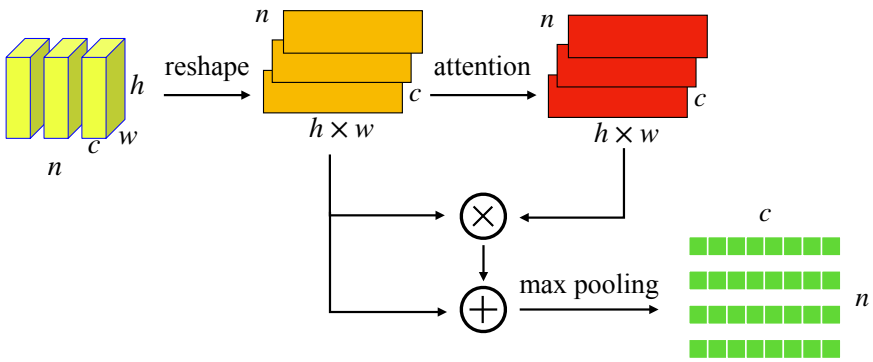


图 4-5: 像素级别注意力机制

如图 4-5 所示，将提取出来多帧的特征图，变形成为 n 个堆叠的二维张量，这个二维张量的宽度为 $h \times w$，代表该帧特征图中的每一个位置上的元素，高度为 c，为该帧特征图的通道数，这里将每一个元素上不同通道的特征，作为描述元素位置的特征。

在每帧图像内部，通过特征图中不同位置的元素互相作用，每一个位置的输出信息，都整合了所有位置的信息。此时，我们将注意力机制得到的结果，作为原始特征图每个元素重要性的一个权重，将权重与原始特征图相乘，同时加上原始特征图，这一步的目的也是为了跨越连接，使得反向传播时有一个直接

的通路，网络能够更好地训练。最后，通过一个最大池化函数，整合一个特征图内部的信息，得到每一帧提取出的特征。值得注意的是，这里提取每帧特征的过程中，利用到了特征图中不同元素点位置的信息，网络有能力去关注一个特征图中的某一个位置，然后提取特征。

4.2.4 帧级别注意力机制

除了能够通过静态的有区分力的部位进行识别以外，我们也可以通过动态的走路姿态来识别一个人，比如某一个人走路时手臂摆动比较快，或者头会不断偏向某一个方向，我们能够通过观察这一局部时间内的走路姿态，就有足够的把握对人进行识别。

从这一动态的识别过程中受到启发，我们设计了帧级别注意力机制。通过帧级别的注意力机制，网络有能力关注到整个图像特征序列中比较有代表性的某些帧。从这些有代表性的特征序列中进一步提取步态特征，然后进行识别。

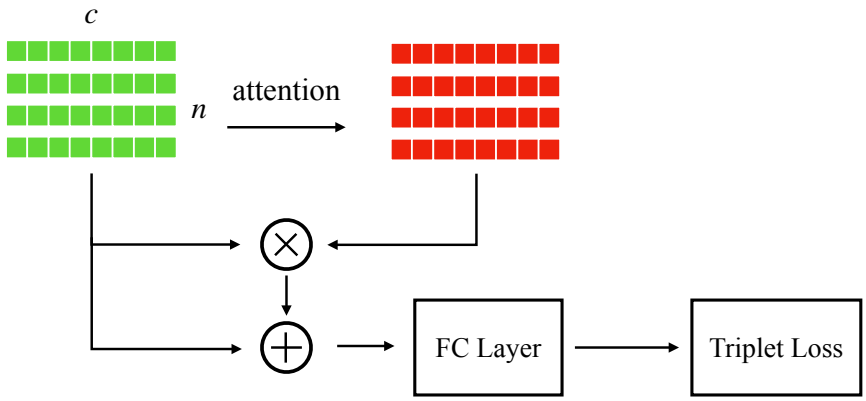


图 4-6: 帧级别注意力机制

如图 4-6 所示。在提取出像素级别注意力的基础上，进行帧级别注意力的提取。这一步的模型输入大小为 $n \times c$，其中，n 为所用的帧数，c 为每一帧提取出来的特征维度，这个特征是通过上一步对像素级别特征图所有位置的特征进行最大池化得到的，所以它的维度也为 c，与特征图的通道数相等。

通过帧级别注意力机制，网络能够关注到整段轮廓图序列中的重要信息，从而有能力根据动态的姿态轨迹，来对一个人进行识别。

在经过帧级别注意力机制以后，与像素级别注意力机制的做法类似，我们仍然将注意力机制得到的结果，作为该输入特征每个位置的权重，将该权重与

输入特征的对应位置相乘，然后加上该输入特征，得到注意力机制的最终结果。

为了使得提取出来的特征更加具有区分性，我们在帧级别注意力机制之后增加了全连接层，将特征映射到一个更有区分度的空间中。对映射到该空间中的特征向量计算三元组损失，来对网络参数进行优化，损失的计算方式在 3.1.2 节中进行了介绍，此处不再赘述。

4.3 实验与分析

在本节中，我们首先会介绍具体的实验设置，然后介绍注意力机制作用方式的实验结果，为了分析两种所提注意力机制的性能，我们在其基础上设计了四种作用方式；之后我们在本章所提网络的基础上，使用第三章提出的损失函数，进一步验证所提损失函数的有效性与通用性；最后，我们与学术界的一些方法比较结果，说明我们所提方法的有效性。

4.3.1 实验设置

对于一段步态序列，我们选择其中的三十帧用于提取特征，经过特征提取网络提取得到的特征维度大小为 $30\times32\times16\times11$)，其中，30 为视频的帧数，32 为每一帧图像提取出的通道数，16 和 11 分别为提取得到特征图的宽和高。

在像素级别注意力机制中，我们使用的多头注意力机制的头数为 4，层数为 1，输入注意力机制的特征维度，也就是图像中提取出来的通道数 32，在注意力机制中的特征维度数，也就是 $\boldsymbol{k}$ 和 $\boldsymbol{v}$ 的维度数为 16。在帧级别注意力机制中，我们使用的多头注意力机制参数与像素级别注意力机制参数相同。

网络最后用于特征映射的全连接层，输出维度为 256。在训练过程中，我们损失函数使用三元组损失。学习率设置为 0.00003，批量大小设置为每个批量中包含 8 个行人，每个行人包含 4 段走路序列。总共训练 80000 轮次，其中，一个轮次指从数据中采样出一个小批量的数据。

在实验中，我们使用了 CASIA-B 数据集[4]，该数据集已经在 3.5.1 节中进行了介绍，此处不再赘述。

4.3.2 注意力机制作用方式

在 4.2 节中，我们提出了两种注意力机制，分别是像素级别注意力机制和帧级别注意力机制，为了验证其效果，我们分别研究了单独使用像素级别注意力、单独使用帧级别注意力、像素级别注意力后跟帧级别注意力（连续），以及像素级别注意力与帧级别注意力结果拼接，这四种注意力的作用方式。其中，像素级别注意力代表仅使用像素级别注意力，然后将注意力后的结果取最大池化，再经过全连接得到每一帧的特征。帧级别注意力代表仅使用帧级别的注意力机制，是在得到特征图以后，先在帧内进行最大池化，然后经过帧级别注意力机制。连续注意力机制代表两种注意力同时使用，按照先像素级别注意力，后帧级别注意力的方式。拼接注意力机制代表将两种注意力机制得到的结果向量拼接起来，然后再去提取特征。

我们在 CASIA-B 上进行了相同视角的实验，实验结果如表 4-1 所示。经过我们实验发现，单独使用像素级别注意力达到了最好的效果，这是由于它的作用方式简单直接，能够按照权重对整张特征图内所有信息进行聚合，有非常好的识别效果。

表 4-1: 不同注意力机制实验结果

注意力作用方式	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	平均识别精度
像素级别注意力	**99.18**	**98.36**	**99.18**	96.72	**100.0**	**99.18**	**99.18**	**98.36**	98.36	96.72	**98.36**	**98.51**
帧级别注意力	96.72	**98.36**	97.54	95.90	98.36	97.54	95.90	97.54	95.90	97.54	95.90	97.02
连续注意力	98.36	96.72	97.54	**97.54**	**100.0**	**99.18**	96.72	96.72	**100.0**	**98.36**	96.72	97.99
拼接注意力	98.36	95.90	**99.18**	**97.54**	95.90	97.54	97.54	96.72	98.36	**98.36**	97.54	97.54

单独使用帧级别注意力，达到了最差的效果。我们分析后认为，帧级别注意力机制虽然有能力关注到所有特征图中的有用信息，但是却忽略了短期时间窗口内的顺序信息。这样的特点导致了帧级别注意力无法有效提取到具有区分力的步态特征，最终效果比较差。

使用拼接注意力机制和连续注意力机制，达到了相似的效果，但是都略差于像素级别注意力的效果。这样的结果可能是由两方面导致的：首先第一方面仍然是帧级别注意力机制无法关注到短期时间窗口的信息，导致无法有效融合序列内的特征；另外一方面，由于两部分注意力机制的存在，大幅增加了网络参数，而数据集中人数较少，过拟合现象比较严重。

总的来说，单独使用像素注意力机制来整合特征图内部的信息，后跟最大池化整合序列内的所有信息，达到了最好的效果，因此我们在后续实验中，均使用这样的配置。

4.3.3 损失函数实验

为了进一步说明我们在第 3 章中提出损失函数的有效性和通用性，我们使用不同的损失函数来训练本章中提出的网络。实验结果如表 4-2 中所示，从表中可以看出，将两种损失函数有效结合起来，确实能够提升识别的性能，达到更好的识别精度，最终在多个视角上，都达到了最优的识别精度。

表 4-2: 损失函数实验结果

损失函数	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	平均识别精度
三元组损失	98.36	97.54	**99.18**	97.54	98.36	97.54	99.18	94.26	97.54	**100.0**	**100.0**	98.14
A-Softmax 损失	98.36	**99.18**	98.36	98.36	**100.0**	**100.0**	**100.0**	**98.36**	98.36	98.36	97.54	98.81
第三章所提损失	**100.0**	98.36	**99.18**	**99.18**	**100.0**	99.18	99.18	**98.36**	**99.18**	99.18	99.18	**99.18**

我们从表 4-2 中还发现了一个有趣的现象：当视角接近于 180° 时，即人近似背对于摄像头方向行走，此时使用三元组损失有最好的效果；而当视角接近于 90° 时，即人的行走方向与摄像头的方向近似垂直时，此时使用 A-Softmax 会有更好的效果。将两种损失有效结合以后，在对应视角下的表现可能不如单独使用某一种损失函数，但是带来的是整体视角上的识别精度的提升，因此最终识别精度也有了一定的提升。总的来讲，不同的损失函数，在不同的视角下会有着不同的识别性能，而将两种损失有效结合起来，带来的是整体性能的提高。

为了更好地研究不同损失函数的效果，我们绘制了识别精度变化的曲线，如图 4-7 所示。在图中，我们展示了前一万次迭代时模型的识别精度，其中每 1000 次迭代进行一次测试。从图中可以看出，我们提出的损失函数，精度上升非常快，1000 次迭代以后，已经有 74% 左右的识别精度了。单独使用另外两种损失的效果要差于我们所提出的损失。值得注意的是，A-Softmax 损失虽然识别精度上升较慢，但是经过长时间训练以后，效果反而超过了三元组损失。这是由于三元组损失直接对最终目标进行优化，使得相同身份的行人特征之间距离更加接近，不同身份的行人特征之间距离更加远离，因此初期识别性能能够快速提升，但是经过一段时间的训练以后，由于大部分都是简单的三元组，三元组损失已

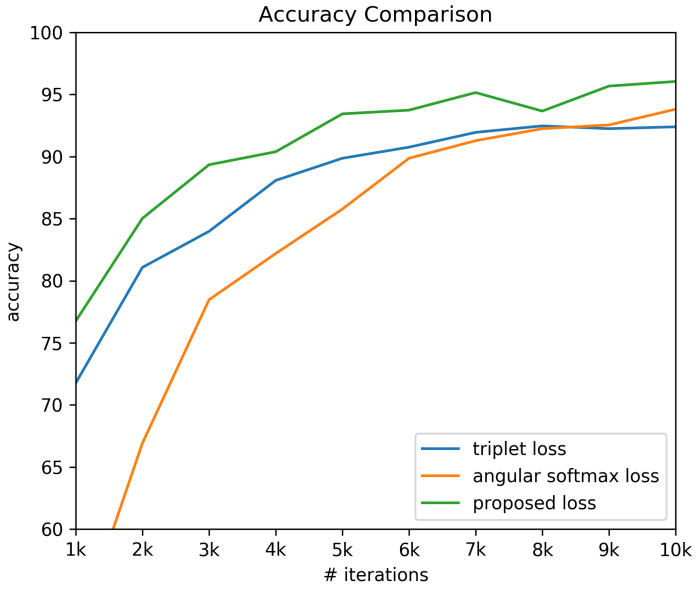


图 4-7: 识别精度变化曲线

经逐渐无法学习到有用信息，识别精度也趋于饱和。

4.3.4 对比实验

最后，我们比较了我们的方法与学术界一些方法在各个视角下的识别性能。具体来说，我们与 GaitGAN[22]、PTSN[46]、3DCNN[34] 和 SPAE[25] 方法进行了比较，如表 4-3 所示，我们的方法达到了最好的整体识别精度，尤其是当视角为 0°、54° 和 108° 时，我们方法的识别效果大幅超越了其余方法的识别效果，充分说明了我们所提方法的有效性。

表 4-3: 方法比较

方法	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	平均识别精度
PTSN[46]	96.77	**99.19**	98.39	98.39	94.35	96.77	95.97	95.97	96.77	98.39	95.16	96.92
3DCNN[34]	96.30	98.20	98.50	95.40	94.30	**99.90**	98.60	97.00	97.40	**99.20**	96.10	97.35
SPAE[25]	98.39	**99.19**	97.58	95.97	95.97	95.97	96.77	98.39	97.58	96.77	**100.0**	97.51
GaitGAN[22]	**100.0**	**99.19**	**99.19**	95.97	**100.0**	98.39	97.58	**99.19**	**99.19**	99.19	**100.0**	98.90
像素注意力 + 所提损失	**100.0**	98.36	99.18	**99.18**	**100.0**	99.18	**99.18**	98.36	99.18	99.18	99.18	**99.18**

4.4 本章小结

在本章中，我们分析了经典卷积神经网络的缺陷，即无法捕捉到长期的时空依赖性，而这对于视频的理解和分析是至关重要的。从这一点出发，受到人类大脑处理多源信息方式的启发，我们设计并提出了两种类型的注意力机制，分别是像素级别注意力机制和帧级别注意力机制。在此基础上，我们设计了整体的网络结构：先通过特征提取模块，提取图像局部特征的同时，减少网络的参数量；然后每一帧提取出的特征通过像素级别注意力机制，得到像素级别的注意力结果；将像素级别的注意力结果通过帧级别注意力机制，得到整段序列的注意力关注结果。

像素级别注意力机制能够关注到一张特征图内不同位置的特征重要性，通过网络学习到特征图内部的区分性信息；帧级别注意力机制能够使得网络关注到不同帧之间的重要性信息，通过对某些帧进行重点关注来完成识别。在实践中，可以将两种注意力方式结合起来，先关注特征图内的信息，再关注序列中帧间的信息。最后，我们在 CASIA-B 数据集上的实验，证明了所提模型的有效性。

第五章　步态识别系统设计

为了验证本文所提步态识别算法的有效性与实用性，我们将本文提出的算法成功地使用在了实际系统中。在本章中将会对我们设计并实现的步态识别系统进行详细的介绍。

5.1　系统概述

随着数字化以及信息技术的不断发展，对于身份识别的需求日益增加。相比于传统的指纹识别与人脸识别，步态识别能够远距离对人进行识别，同时可以无接触、非感知地进行识别，这就使得步态识别难以伪装和发现，由步态识别实现的身份验证系统，可以应用在需要高精度的身份识别场景，比如银行专属贵宾客户的身份识别，军用身份识别等场景。

为了系统能够顺利使用，我们需要提前离线训练好用于提取步态特征的模型，在实际使用中不再涉及模型的训练。整个系统可以分为两个部分，分别是注册过程与识别过程。在注册过程中，通过调用摄像头对待注册的行人走路姿态进行录制，然后将提取的特征保存在库中；在识别过程中，对于待识别的行人，同样调用摄像头录制其走路姿态，通过对走路姿态提取特征，与库中的所有特征进行比对，得到识别的结果。两个部分均涉及到数据的预处理过程。

由于视频的处理以及特征提取非常耗费计算资源，而在实际识别时一般需要在嵌入式设备或者计算资源很少的设备上进行。为了保证识别效率，我们将系统分为前端和后端，其中，前端完成的任务是调用系统摄像头，录制包含行人走路姿态的视频，并将该视频传递给后端；后端完成具体的注册和识别任务。注册过程会将提取得到的行人特征保存在库中，识别过程提取出待识别的行人特征，与库中的所有特征进行比对找到最相似的特征，然后验证两者是否来自于同一个行人。

5.2 系统设计

如图 5-1 所示，我们的整个系统包含三大模块：分别是数据预处理模块、注册模块和识别模块。拍到的视频首先会进行预处理，提取对齐后的行人轮廓图；在注册阶段，将利用轮廓图提取出的步态特征存入数据库中；在识别阶段，使用轮廓图提取出的步态特征与库中所有特征进行比对，得到识别结果。在本节中，将会对这三个模块具体介绍。

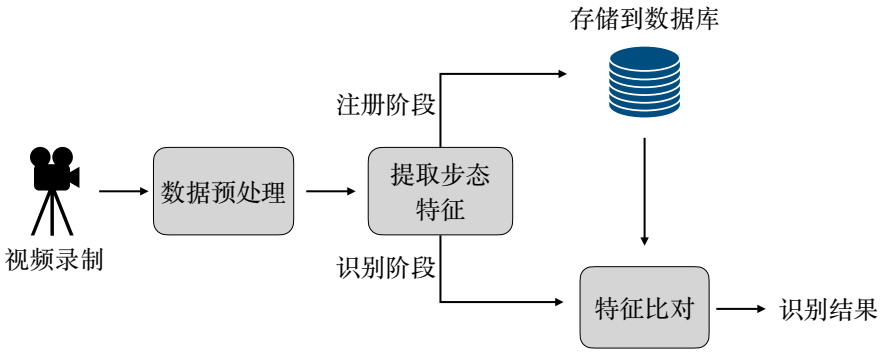


图 5-1: 系统整体流程

5.2.1 数据预处理

在实际的视频录制过程中，可能会存在很多问题导致录制的数据出现不一致的情况，比如光照变化，相机与行人的距离变化，行人衣服颜色信息或者背景颜色信息变化等。为了去除这些信息的干扰，获得与训练集中数据一致的步态轮廓图数据，需要先进行数据预处理的操作。这里用到的数据预处理包括两个步骤：提取步态轮廓图和行人对齐。

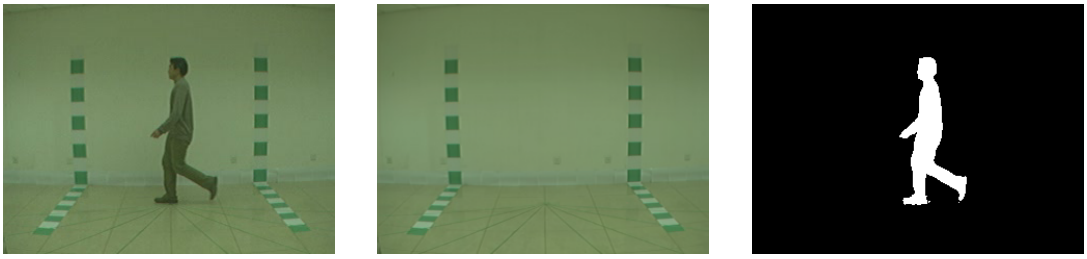
提取步态轮廓图是指通过将视频中行走的行人与背景分割开来，来得到行人所在的区域。将行人所在的区域像素值标记为 1，行人没有出现的区域，像素值标记为 0。通过对输入数据进行二值化，能够移除由于人衣服颜色信息以及背景的颜色信息对识别时造成的干扰，获得更加准确的识别结果。

我们这一步使用的方法为背景减除法，其由于运行速度快，容易实现，非常适用于摄像头和背景相对固定的场景下，是步态识别中最常使用的检测运动目标的方法。具体来说，该方法首先对背景进行建模，然后将视频中的每一帧与背景计算绝对差值，根据提前设定好的阈值，找到图像中与背景差距比较大的

像素，认为是检测到的运动目标所在位置。用 $I_k(x,y)$ 表示视频第 k 帧图像位于 (x,y) 处的像素值，用 $B(x,y)$ 表示建模出的背景位于 (x,y) 处的像素值，提前选取一个合适的阈值记为 T，在最后的检测结果中，$D_k(x,y)=1$ 代表对应位置是检测出的运动物体，$D_k(x,y)=0$ 代表对应位置是背景。采用如下方式计算检测结果：

$$D_k(x,y)=\begin{cases}1, \text{if } |I_k(x,y)-B(x,y)|>T\\ 0, \text{else}\end{cases} \tag{5-1}$$

如图 5-2 所示为背景减除的过程，其中，图 5-2a 是输入视频中的其中一帧，图 5-2b 是建模出来背景图像，图 5-2c 是最终检测到的目标。



(a) 输入视频帧　(b) 背景　(c) 检测结果

图 5-2: 背景减除法提取行人

这里背景建模的方法是采用同一个位置下的摄像头录制没有人走路时的视频，通过对背景视频中的每一帧计算均值即得到了背景图像。

提取出行人轮廓图以后，接下来需要进行的预处理步骤为行人对齐。由于每个人可能离相机距离不同，甚至说同一个人由于向着相机走动或者向着反方向走动，可能会造成人的高度变化非常大，因此需要将视频中的行人对齐。我们这里将所有的行人对齐到 64×64 像素，如果像素值太大，虽然人的轮廓信息会更清楚，但是带来的是计算资源的大幅提高，所需要的识别时间太久，导致用户体验比较差，而像素值太小又会造成识别结果比较差。经过实践，64×64 像素是一个比较合适的大小。

我们采用如下方式将检测到的行人对齐：（1）首先找到行人出现的最高的像素行坐标 y_1，以及最低出现的像素行坐标 y_2；（2）裁取出 y_1 与 y_2 中间的区域，把图像中的这一部分区域等比例拉伸到高度为 S 像素，这里取 $S=64$；（3）此时找到图像横坐标中央位置 x_{mid}，从这里出发左右各取 $S/2$ 个像素，如果左侧或右侧不足 $S/2$ 个像素，则补 0。该轮廓图对齐算法具体的伪代码如算法 5.1

中所示。对齐的过程如图 5-3 所示。

(a) 行人提取结果 (b) 对齐结果

图 5-3: 对齐行人大小

算法 5.1 轮廓图对齐算法

输入： 待对齐的轮廓图 D，其高为 h，宽为 w；图像放缩插值函数 $R(I, s)$，其中，I 为输入图像，s 为放缩后的图像大小；S 为对齐后的图像大小

输出： 对齐后的轮廓图 $\tilde{D}$

$y_{\text{top}} \leftarrow$ 最上像素所在行坐标
$y_{\text{bottom}} \leftarrow$ 最下像素所在行坐标
$\text{ratio} \leftarrow S/(y_{\text{bottom}} - y_{\text{top}})$
$w\_ = \text{ratio} \times w$
$D_{\text{resized}} = R(D[y_{\text{top}} : y_{\text{bottom}}, :], (S, w\_))$
$x_{\text{sum}} = 0$
$\text{count} = 0$
for $i \in \{1 \cdots S\}$ **do**
  for $j \in \{1 \cdots w\_\}$ **do**
    $x_{\text{sum}} = x_{\text{sum}} + j$
    $\text{count} = \text{count} + 1$
  end for
end for
$x_{\text{mid}} \leftarrow x_{\text{sum}}/\text{count}$
$\tilde{D} = D_{\text{resized}}[:, x_{\text{mid}} - S/2 : x_{\text{mid}} + S/2]$
return $\tilde{D}$

5.2.2 注册过程

如图 5-4 为注册过程的流程图。注册过程完成的任务是：前端点击“开始录制”按钮，调用摄像头开始录制包含走路姿态的视频，并在前端界面上实时显示录制画面，然后点击“结束录制”按钮，将录制好的视频传递给后端，此时等待后端返回注册成功的提示信息。在后端，会对这个包含走路姿态的视频进行步态特征的提取，将提取出来的特征储存下来，用于后续的识别。

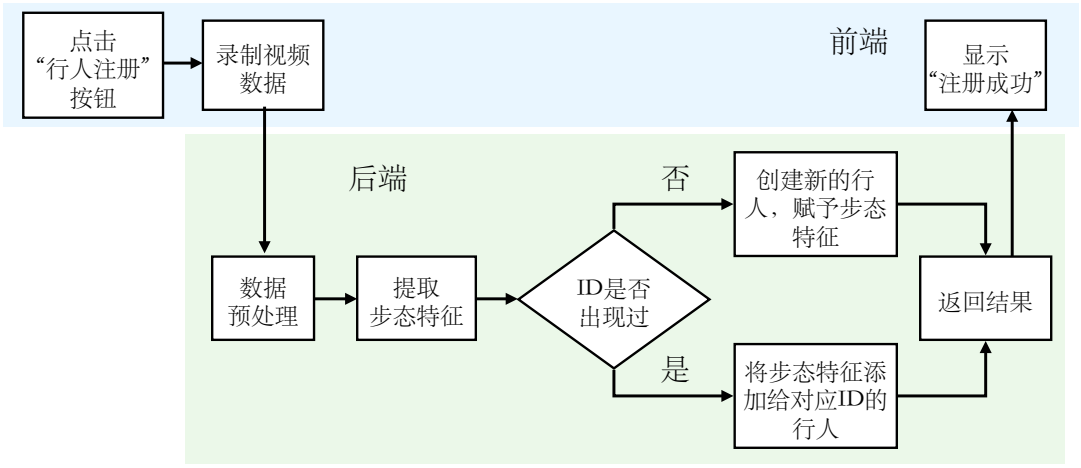


图 5-4: 注册过程流程图

在实际使用过程中，每个人可以注册多段走路姿态视频，最好是在不同的走路情况下，比如背着背包行走，或者是穿着不同类型的衣服鞋子行走。为了实现这个逻辑，我们对每个人有一个独特的 ID，当有一个新的注册需求时，通过检测将要注册的行人 ID 是否已经在库中注册过，如果注册过的话，就为这个人添加一条新的记录，否则就注册一个新的行人。

后端这些任务完成以后，将成功信息返回给前端，此时，会在前端界面上显示“注册成功”的信息。至此，注册过程完成。

5.2.3 识别过程

如图 5-5 所示为识别过程的流程图。识别过程需要完成的任务是：前端通过点击“行人识别”按钮，调用摄像头开始录制含有待识别行人走路姿态的视频，并在前端界面上实时显示录制信息，然后前端点击“结束录制”按钮，将录制好的视频传递给后端，等待后端返回识别的结果。

在后端，首先用训练好的模型提取待识别视频的步态特征，通过将该特征与注册在库中的所有行人步态特征进行对比，两两计算余弦距离，找到与待识别行人特征最接近的步态特征。然后进行步态验证的任务，即判断这两个步态特征对应的行人是否是同一个行人，具体的验证方式是比较两者之间的余弦距离，如果距离小于一个给定的阈值，那么就说这两个步态特征属于同一个行人，后端返回识别成功，以及该行人的信息；否则，后端返回该行人不属于库中行人的信息。

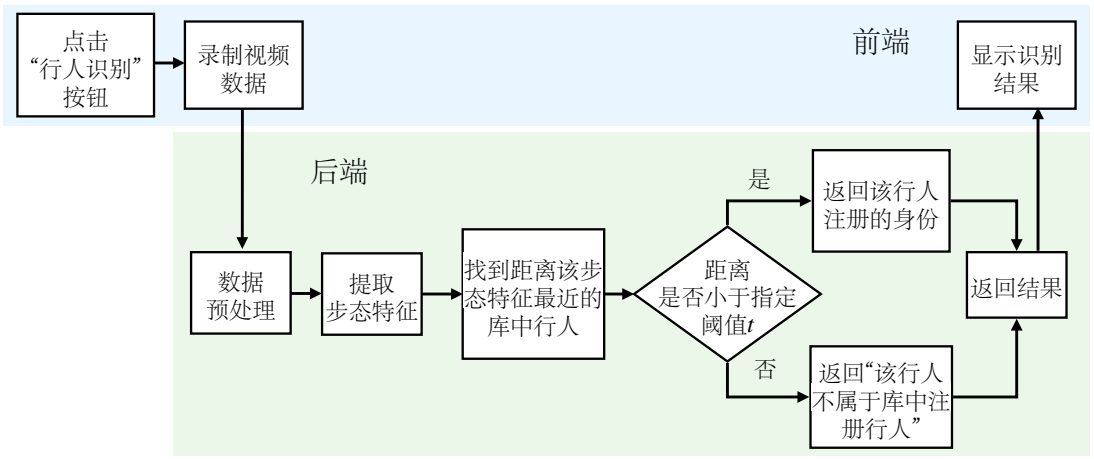


图 5-5: 识别过程流程图

后端这些任务完成以后，前端会在界面上显示“识别完成”的信息，同时将识别的结果展示出来。至此，识别过程完成。

5.3 系统实现

本节中会详细介绍该系统的实现，包括两个部分，分别为硬件实现与软件实现，此外，在本节中，还会展示我们系统实际运行的效果。

5.3.1 硬件实现

由于步态的识别需要涉及到对视频进行处理，非常耗费计算资源，因此我们选择了前后端的部署方式，前端通过调用摄像头录制包含行走姿态的视频，通过网络将需要完成的任务传递给后端，在后端中进行特征的提取与匹配。因此，前端主要注重灵活性和便携性，后端主要注重的是计算能力，前端和后端之间通过网络传输信息。

在实践中，我们选择使用树莓派 3b+ 来部署我们的前端系统，摄像头采用的是罗技 1080P 高清摄像头 C920e，其自身具备了一定的预处理能力，能够自动调节视频亮度。前端设备的选择，令我们的系统具有非常大的灵活性，可以部署于各种嵌入式设备上，使系统得到更加广泛的应用。

对于后端系统，我们选择基于 Linux 系统的服务器，CPU 为 4 核心的英特尔 i7-4790 CPU，主频为 3.60GHz，运行内存为 24GB。由于大多数计算都需要使用显卡加速，我们使用的显卡为 GeForce GTX TITAN Black，计算能力足以应

付后端计算需求，有 6GB 的显存。

5.3.2 软件实现

我们的系统完全基于 Python 开发。前端功能包括视频的录制，以及程序界面的展示，还有与后端进行通信，将录制的视频传递给后端。我们使用 Qt Creator 完成窗口、按钮和显示信息等的设计，基于 PyQt5[①]搭建程序界面，它是一个跨平台的前端界面显示库，支持 Windows，Linux，Android 和 iOS 等系统，使得我们的前端可以运行于各个操作系统，非常灵活。我们使用 opencv-python 库[②]进行视频的录制和保存，使用 numpy 库[③]将录制到的视频帧转换为字节流。在前后端通信中，我们使用 protobuf 序列化机制[④]将需要传输的数据序列化，该传输方式通讯速度快，支持自定义数据结构，灵活性非常高。

后端中，包括数据库模块、特征提取模块、特征比对模块。对于一个前端传输过来的视频，会附带有一个识别或者注册的请求，如果是注册，此时还会带有注册人的身份信息，后端会将这些信息，连同提取出来的特征向量，一同录入数据库，以便高效查询。此处的数据库，我们选择了开源的非关系型数据库 levelDB[⑤]来存储数据，该数据库直接通过键值对进行数据的增删改查，随机写和顺序读/写的效率都非常高。当前端传来识别的请求时，我们提取该视频的特征向量，然后和库中的特征向量进行比对，找到最相似的特征向量，返回识别结果。对于系统中使用到的深度学习模型，我们基于 pytorch 框架[47][⑥]进行搭建。相比于其它深度学习框架，该框架简洁灵活，可以很方便地定义网络内部操作，易于实现。

5.3.3 效果展示

我们这里将前端部署于一台 Linux 环境下的设备，以展示我们程序的运行效果。

[①] PyQt5: https://github.com/baoboa/pyqt5
[②] opencv-python: https://github.com/opencv/opencv-python
[③] numpy: https://github.com/numpy/numpy
[④] protobuf: https://github.com/protocolbuffers/protobuf
[⑤] levelDB 数据库：https://github.com/google/leveldb
[⑥] pytorch: https://github.com/pytorch/pytorch

程序运行时的界面如图 5-6 所示。在使用时，首先点击“打开相机”按钮，调用摄像头的接口，实时显示画面到窗口上方，然后点击“行人注册”或者“行人识别”按钮，进行行人的注册或者识别任务，如果点击“行人注册”，会跳出如图 5-7 所示的信息输入界面，录入待注册行人的身份信息，包括行人姓名和行人 ID，每个行人通过 ID 来唯一标识。

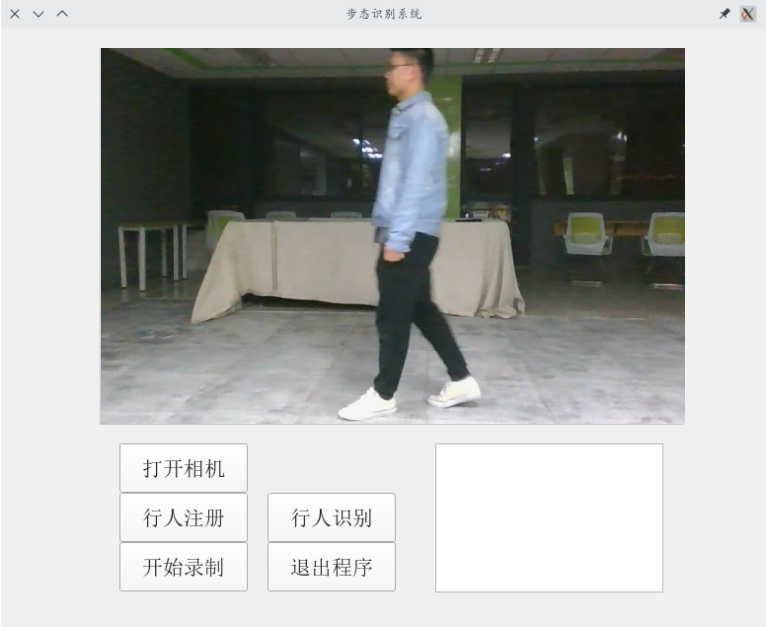


图 5-6: 程序界面

图 5-7: 信息输入界面

点击“开始录制”按钮，即开始视频的录制，录制完成后会自动传输到后端，进行注册或识别的任务。如果是行人注册的任务，会在完成任务后，将结果返回到前端显示。对于行人识别的任务，最终识别的效果，同样也会从后端中返回到前端，显示在窗口上。如图 5-8 所示，这些信息都会显示在右下角的文本框内。

图 5-8: 识别结果显示

5.4 本章小结

在本章中，我们利用所提算法，实现了用于身份验证的步态识别系统。通过系统的实现，可以验证本文所提算法的有效性和实用性，将这个系统稍加修改，即可应用于公共安全、智能家居和高精度身份识别领域，为人们的生活提供极大的便利。

第六章　总结与展望

随着信息科学技术的不断发展，生物识别技术得到了非常广泛的应用。其中，步态识别技术由于其可以在非受控环境下识别、可以远距离识别，以及识别难以伪装和发现的特性，成为了学术界研究的热点领域。但是，当前的步态识别研究，大多都单纯基于卷积神经网络或者 LSTM 神经网络，卷积神经网络可以提取得到空间中的局部信息，但是无法建模视频内的长期依赖关系。LSTM 虽然能够对时间信息进行建模，但是由于其网络运算无法并行化，导致运行效率比较低。

在本文中，我们从度量学习的角度考虑，设计了一种全新的损失函数，能够提取出更加具有区分性的步态特征。此外，本文还从人脑的作用机制出发，研究了将注意力机制引入步态识别领域的方式，设计并提出了两种注意力机制。我们成功地将本文所提方法应用到实际的步态识别场景中。具体来说，本文的主要贡献如下：

- 从度量学习的角度考虑，我们提出了一种全新的损失函数，该损失函数既能够利用到三元组损失的优势，直接优化度量学习的目标，使得相同身份样本特征之间的距离小于不同身份样本特征之间的距离；又能够利用到基于分类的损失函数的优势，学习到有代表性的步态特征。三元组损失函数优化在欧式空间中，而基于分类的损失函数优化在余弦空间中，为了使网络参数能够顺利优化，我们提出在提取出特征以后，加入一层批归一化层，调整特征的分布到一个可以学习的分布上，然后再计算损失。在 CASIA-B 数据集和 TUM GAID 数据集上进行的实验结果显示，我们所提的算法要好于现有的步态识别算法，证明了所提方法的有效性。
- 受人脑工作原理的启发，我们提出使用注意力机制来解决步态识别问题。在将轮廓图序列中的每一帧分别经过卷积神经网络提取特征后，我们提出了两种注意力机制，分别是像素级别注意力机制和帧级别注意力机制。其中，像素级别注意力机制能够使得网络关注到帧内的空间关系，利用每个人有代表性的部位进行识别；帧间注意力机制关注的是整个轮廓图序列内的信息，能

够关注到在行走过程中一些动态的特点。可以将两种注意力方式结合起来使用，也可以单独使用其中一种。经过我们在 CASIA-B 数据集上的实验，证明了在步态识别领域使用注意力机制的有效性。

- 我们在实际应用场景中，搭建了用于身份验证的步态识别系统。该系统实现上包括两个部分，分别是前端和后端。其中，前端展现系统的操作逻辑，还负责录制包含行人走路姿态的视频；后端负责具体的行人注册和行人识别步骤，同时还涉及到数据库的存取，需要在数据库中保存行人的信息。

按照本文的工作路线，可以继续开展相关的研究工作。首先，利用注意力机制，如何更加有效提取到序列内的信息仍然是一个待解决的任务，如果能够有效解决，效果应该是要好于像素级别注意力机制，这是因为人走路是一个动态的过程，运动中动态的信息是非常重要的。一个可行的方向是将轮廓图序列的位置编码信息也输入网络，使网络有能力关注到不同帧的先后顺序关系，可能会带来一定的性能提升。此外，神经网络的可视化，也是一个非常热门的研究领域，通过可视化能够更好地探寻注意力机制的作用原理，也可以从这个角度出发进行研究。最后，可以继续从度量学习的角度，精心设定不同类型的损失函数，以达到一些期望的目标，提取得到更加具有区分力的特征。

还有另外一些相关的研究可以进行：人走路时的外表信息可能对最终结果造成干扰，传统的做法是将图像转化为轮廓图，但是这样就无法利用目前学术界视觉领域的很多研究进展，比如无法使用预训练模型，因为输入数据之间相差太多。如何更好地利用到原图中的信息，这也是一个待解决的问题。一个可行的做法，就是将输入信息中的外表信息，与步态特征信息拆解开来，仅使用其中的步态特征信息进行识别，特征拆解也是目前计算机视觉中的热点研究领域。

参考文献

[1] DELAC K, GRGIC M. A survey of biometric recognition methods[C]// Proceedings. Elmar-2004. 46th International Symposium on Electronics in Marine. [S.l.]: IEEE, 2004: 184-193.

[2] NIXON M S, TAN T, CHELLAPPA R. Human identification based on gait: volume 4[M]. [S.l.]: Springer Science & Business Media, 2010.

[3] BEN X, GONG C, ZHANG P, et al. Coupled bilinear discriminant projection for cross-view gait recognition[J/OL]. IEEE Trans. Circuits Syst. Video Technol., 2020, 30(3):734-747. https://doi.org/10.1109/TCSVT.2019.2893736.

[4] YU S, TAN D, TAN T. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition[C]//Proc. 18th International Conference on Pattern Recognition: volume 4. [S.l.]: IEEE, 2006: 441-444.

[5] NANDY A, PATHAK A, CHAKRABORTY P. A study on gait entropy image analysis for clothing invariant human identification[J/OL]. Multim. Tools Appl., 2017, 76(7):9133-9167. https://doi.org/10.1007/s11042-016-3505-0.

[6] ZHANG Z, TRAN L, YIN X, et al. Gait recognition via disentangled representation learning[C]//Proc. the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2019: 4710-4719.

[7] ZHANG P, WU Q, XU J. VT-GAN: View transformation GAN for gait recognition across views[C]//Proc. International Joint Conference on Neural Networks. [S.l.]: IEEE, 2019: 1-8.

[8] ZHANG Y, HUANG Y, YU S, et al. Cross-view gait recognition by discriminative feature learning[J]. IEEE Transactions on Image Processing, 2019, 29:1001-1015.

[9] CHAO H, HE Y, ZHANG J, et al. GaitSet: Regarding gait as a set for cross-view gait recognition[C]//the AAAI Conference on Artificial Intelligence: volume 33. [S.l.: s.n.], 2019: 8126-8133.

[10] HAN J, BHANU B. Individual recognition using gait energy image[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 28(2):316-322.

[11] BASHIR K, XIANG T, GONG S. Gait recognition using gait entropy image[C]// International Conference on Imaging for Crime Detection and Prevention. [S.l.: s.n.], 2009: 1-6.

[12] LAM T H, CHEUNG K, LIU J N. Gait flow image: A silhouette-based gait representation for human identification[J/OL]. Pattern Recognition, 2011, 44(4): 973-987. https://www.sciencedirect.com/science/article/pii/S0031320310004954. DOI: https://doi.org/10.1016/j.patcog.2010.10.011.

[13] WANG C, ZHANG J, PU J, et al. Chrono-gait image: A novel temporal template for gait recognition[C]//European Conference on Computer Vision. [S.l.]: Springer, 2010: 257-270.

[14] LUO J, TANG J, TJAHJADI T, et al. Robust arbitrary view gait recognition based on parametric 3D human body reconstruction and virtual posture synthesis[J]. Pattern Recognition, 2016, 60:361-377.

[15] WANG X, YAN W Q. Cross-view gait recognition through ensemble learning[J]. Neural Computing and Applications, 2019:1-13.

[16] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C/OL]//PEREIRA F, BURGES C J C, BOTTOU L, et al. Advances in Neural Information Processing Systems: volume 25. Curran Associates, Inc., 2012. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[17] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database.[C/OL]//CVPR. IEEE Computer Society, 2009: 248-255. http://dblp.uni-trier.de/db/conf/cvpr/cvpr2009.html#DengDSLL009.

[18] FENG Y, MA L, LIU W, et al. Unsupervised image captioning[C/OL]//IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, 2019: 4125-4134. http://openaccess.thecvf.com/content_CVPR_2019/html/Feng_Unsupervised_Image_Captioning_CVPR_2019_paper.html. DOI: 10.1109/CVPR.2019.00425.

[19] CHEN J, LEI B, SONG Q, et al. A hierarchical graph network for 3d object detection on point clouds[C/OL]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. IEEE, 2020: 389-398. https://doi.org/10.1109/CVPR42600.2020.00047.

[20] LI F, LONG Z, HE P, et al. Fully convolutional pyramidal networks for semantic segmentation[J/OL]. IEEE Access, 2020, 8:229132-229140. https://doi.org/10.1109/ACCESS.2020.3045280.

[21] WU Z, HUANG Y, WANG L, et al. A comprehensive study on cross-view gait based human identification with deep CNNs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(2):209-226.

[22] YU S, CHEN H, REYES G, et al. GaitGAN: invariant gait feature extraction using generative adversarial networks[C]//Proc. the IEEE Conference on Computer Vision and Pattern Recognition Workshops. [S.l.: s.n.], 2017: 30-37.

[23] HU B, GUAN Y, GAO Y, et al. Robust cross-view gait recognition with evidence: A discriminant gait gan (diggan) approach[J]. arXiv preprint arXiv:1811.10493, 2018.

[24] HE Y, ZHANG J, SHAN H, et al. Multi-task GANs for view-specific feature learning in gait recognition[J]. IEEE Transactions on Information Forensics and Security, 2018, 14(1):102-113.

[25] YU S, CHEN H, WANG Q, et al. Invariant feature extraction for gait recognition using only one uniform model[J/OL]. Neurocomputing, 2017, 239:81-93. https://doi.org/10.1016/j.neucom.2017.02.006.

[26] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention [C/OL]//GHAHRAMANI Z, WELLING M, CORTES C, et al. Advances in Neural Information Processing Systems: volume 27. Curran Associates, Inc., 2014. https://proceedings.neurips.cc/paper/2014/file/09c6c3783b4a70054da74f2538ed47c6-Paper.pdf.

[27] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C/OL]// GUYON I, LUXBURG U V, BENGIO S, et al. Advances in Neural Information Processing Systems: volume 30. Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[28] WU B, XU C, DAI X, et al. Visual transformers: Token-based image representation and processing for computer vision[J/OL]. CoRR, 2020, abs/2006.03677. https://arxiv.org/abs/2006.03677.

[29] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//International conference on machine learning. [S.l.]: PMLR, 2015: 448-456.

[30] XU C, MAKIHARA Y, LI X, et al. Speed-invariant gait recognition using single-support gait energy image[J/OL]. Multim. Tools Appl., 2019, 78(18):26509-26536. https://doi.org/10.1007/s11042-019-7712-3.

[31] LI J, ZHANG J, NI J. 基于修正步态能量图和视角检测的步态识别方法 (Gait Recognition Method Based on Modified Gait Energy Image and View Detection) [J/OL]. 计算机科学, 2016, 43(8):300-303. https://doi.org/10.11896/j.issn.1002-137X.2016.08.061.

[32] VERLEKAR T T, CORREIA P L, SOARES L D. View-invariant gait recognition system using a gait energy image decomposition method[J/OL]. IET Biom., 2017, 6(4):299-306. https://doi.org/10.1049/iet-bmt.2016.0118.

[33] ZHAO G, LIU G, LI H, et al. 3d gait recognition using multiple cameras[C/OL]// Seventh IEEE International Conference on Automatic Face and Gesture Recognition (FGR 2006), 10-12 April 2006, Southampton, UK. IEEE Computer Society, 2006: 529-534. https://doi.org/10.1109/FGR.2006.2.

[34] WOLF T, BABAEE M, RIGOLL G. Multi-view gait recognition using 3D convolutional neural networks[C]//Proc. IEEE International Conference on Image Processing. [S.l.]: IEEE, 2016: 4165-4169.

[35] FU Y, WEI Y, ZHOU Y, et al. Horizontal pyramid matching for person re-identification[C/OL]//The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. AAAI Press, 2019: 8295-8302. https://doi.org/10.1609/aaai.v33i01.33018295.

[36] LIU W, WEN Y, YU Z, et al. Sphereface: Deep hypersphere embedding for face recognition[C]//Proc. the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2017: 212-220.

[37] HERMANS A, BEYER L, LEIBE B. In defense of the triplet loss for person re-identification[J]. arXiv preprint arXiv:1703.07737, 2017.

[38] HOFMANN M, GEIGER J, BACHMANN S, et al. The TUM gait from audio, image and depth (GAID) database: Multimodal recognition of subjects and traits [J]. Journal of Visual Communication and Image Representation, 2014, 25(1): 195-206.

[39] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. nature, 1986, 323(6088):533-536.

[40] KINGMA D P, BA J. Adam: A method for stochastic optimization[C/OL]// BENGIO Y, LECUN Y. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. 2015. http://arxiv.org/abs/1412.6980.

[41] CASTRO F M, MARÍN-JIMÉNEZ M J, GUIL N, et al. Evaluation of CNN architectures for gait recognition based on optical flow maps[C]//Proc. International Conference of the Biometrics Special Interest Group. [S.l.]: IEEE, 2017: 1-5.

[42] CASTRO F M, MARÍN-JIMÉNEZ M J, GUIL N, et al. Automatic learning of gait signatures for people identification[C]//International Work-Conference on Artificial Neural Networks. [S.l.]: Springer, 2017: 257-270.

[43] BATTISTONE F, PETROSINO A. TGLSTM: A time based graph deep learning approach to gait recognition[J]. Pattern Recognition Letters, 2019, 126:132-138.

[44] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C/OL]//BACH F, BLEI D. Proceedings of Machine Learning Research: volume 37 Proceedings of the 32nd International Conference on Machine Learning. Lille, France: PMLR, 2015: 2048-2057. http://proceedings.mlr.press/v37/xuc15.html.

[45] 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 (CVPR). [S.l.: s.n.], 2016.

[46] LIAO R, CAO C, GARCIA E B, et al. Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations[C]//Chinese Conference on Biometric Recognition. [S.l.]: Springer, 2017: 474-483.

[47] PASZKE A, GROSS S, MASSA F, et al. Pytorch: An imperative style, high-performance deep learning library[M/OL]//WALLACH H, LAROCHELLE H, BEYGELZIMER A, et al. Advances in Neural Information Processing Systems 32. Curran Associates, Inc., 2019: 8024-8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[48] BHANU B, HAN J. Model-based human recognition—2D and 3D gait[M]// Human Recognition at a Distance in Video. [S.l.]: Springer, 2010: 65-94.

[49] CUTTING J E, KOZLOWSKI L T. Recognizing friends by their walk: Gait perception without familiarity cues[J]. Bulletin of the Psychonomic Society, 1977, 9(5):353-356.

[50] DENG J, GUO J, XUE N, et al. Arcface: Additive angular margin loss for deep face recognition[C]//Proc. the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2019: 4690-4699.

[51] DING S, LIN L, WANG G, et al. Deep feature learning with relative distance comparison for person re-identification[J]. Pattern Recognition, 2015, 48(10): 2993-3003.

[52] DOCKSTADER S L, BERG M J, TEKALP A M. Stochastic kinematic modeling and feature extraction for gait analysis[J]. IEEE Transactions on Image Processing, 2003, 12(8):962-976.

[53] HE X, ZHOU Y, ZHOU Z, et al. Triplet-center loss for multi-view 3D object retrieval[C]//Proc. the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2018: 1945-1954.

[54] LIAO R, YU S, AN W, et al. A model-based gait recognition method with body pose and human prior knowledge[J]. Pattern Recognition, 2020, 98.

[55] LIU W, WEN Y, YU Z, et al. Large-margin softmax loss for convolutional neural networks.[C]//the 33nd International Conference on Machine Learning: volume 2. [S.l.: s.n.], 2016: 7.

[56] LUO H, GU Y, LIAO X, et al. Bag of tricks and a strong baseline for deep person re-identification[C]//Proc. the IEEE Conference on Computer Vision and Pattern Recognition Workshops. [S.l.: s.n.], 2019: 0-0.

[57] OHLYAN S, SANGWAN S, AHUJA T. A survey on various problems & challenges in face recognition[J]. International Journal of Engineering Research & Technology (IJERT), 2013, 2(6):2533-2538.

[58] SCHROFF F, KALENICHENKO D, PHILBIN J. FaceNet: A unified embedding for face recognition and clustering[C]//Proc. the IEEE conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2015: 815-823.

[59] TAKEMURA N, MAKIHARA Y, MURAMATSU D, et al. On input/output architectures for convolutional neural network-based cross-view gait recognition [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017.

[60] WANG H, WANG Y, ZHOU Z, et al. Cosface: Large margin cosine loss for deep face recognition[C]//Proc. the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2018: 5265-5274.

[61] WANG Y, SONG C, HUANG Y, et al. Learning view invariant gait features with two-stream GAN[J]. Neurocomputing, 2019, 339:245-254.

[62] WEN Y, ZHANG K, LI Z, et al. A discriminative feature learning approach for deep face recognition[C]//European Conference on Computer Vision. [S.l.]: Springer, 2016: 499-515.

致　　谢

时间如白驹过隙，弹指挥间，我的研究生生涯即将结束，人生的新阶段也即将开始。仍记得刚开学的第一天，我怀着忐忑又激动的心情踏进南京大学的校园，尽管我是一个本科非计算机专业的学生，但我对本专业的兴趣和热情让我对接下来的三年求学路充满憧憬。回首这三年，我遇到了很多课题上的困难，但最终都一一攻克，同时我也获得了很大的收获，这些都离不开老师和同学们的鼓励与帮助。在这里，我要真心感谢曾经给予我帮助的老师和同学们。

首先，要感谢我的导师申富饶教授和宋方敏教授。申富饶教授每周会在百忙之中抽出时间与学生单独交流，在每次交流过程中，我看待事物的方式，以及学习能力和科研能力，都得到非常大的提升，申富饶教授帮助培养了我独立思考和独立研究的能力，使我受益良多。此外，申富饶教授还为我们提供了非常好的科研环境，包括我们组内的图书馆、数据管理服务器、显卡服务器，以及我们所使用的电脑和显示器等，这些便利的设施为我们的科研提供了非常大的帮助，使得我们的科研可以顺利进行。还要感谢赵健老师，赵健老师为我们分享过多次报告，帮我们非常仔细地批改论文，也对我们的研究工作非常上心。

其次，感谢实验室同门，与大家的学术和生活上的讨论使我进步良多。此外，实验室内氛围非常好，每周组会报告上的交流讨论，都会让我感到有非常多的收获，特别是当科研没有什么进展的时候，大家也可以在组会报告上集思广益，提出自己的想法，帮助同学解决面临的问题。

最后，还要感谢我的家人和我的女朋友杨颖，没有他们对我日常生活上的关心照顾，我无法完成这篇论文，是他们的爱与支持让我可以安心在学校进行研究，不需要考虑其它任何生活上的问题，他们是我前进路上最坚强的后盾。

简历与科研成果

基本信息

李雪健，男，汉族，1997 年 1 月出生，山西省运城人。

教育背景

2018 年 9 月 — 2021 年 6 月 南京大学计算机科学与技术系 硕士

2014 年 9 月 — 2018 年 6 月 南京邮电大学测绘工程专业 本科

攻读硕士学位期间完成的学术成果

1. **Xuejian Li**, Feng Han, Jian Zhao, Furao Shen, “A More Robust Gait Recognition System Based on Metric Learning”, under-review.

攻读硕士学位期间的发明专利

1. 申富饶，**李雪健**，韩峰，赵健。一种结合两个向量嵌入空间的高精度步态识别方法。专利申请号：202110109320.0
2. 陈力军，梁雨，**李雪健**，申富饶，刘佳，张晓聪。一种基于深度图像的图书馆机器人障碍识别方法。专利申请号：201810644120.3

攻读硕士学位期间参与的科研课题

1. 国家自然科学基金‘‘基于深度感知增量式联想记忆神经网络的信息融合系统研究，Information fusion system based on deep perception and incremental associative memory neural networks”（课题年限 2019 年 1 月 ~ 2022 年 12 月），负责神经网络模型相关研究。

攻读硕士学位期间获得的比赛奖项

1. 韩峰，**李雪健**，赵加成，姜少魁。2019 首届 IKCEST “一带一路”国际大数据竞赛，国际二等奖

学位论文出版授权书

本人完全同意《中国优秀博硕士学位论文全文数据库出版章程》（以下简称“章程”），愿意将本人的学位论文提交“中国学术期刊（光盘版）电子杂志社”在《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》中全文发表。《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》可以以电子、网络及其他数字媒体形式公开出版，并同意编入《中国知识资源总库》，在《中国博硕士学位论文评价数据库》中使用和在互联网上传播，同意按“章程”规定享受相关权益。

作者签名：_____

_____年_____月_____日

论文题名	基于度量学习和注意力机制的步态识别研究				
研究生学号	MG1833044	所在院系	计算机科学与技术系	学位年度	2018
论文级别	☑ 硕士　☐ 硕士专业学位 ☐ 博士　☐ 博士专业学位　（请在方框内画勾）				
作者 Email	lixj@smail.nju.edu.cn				
导师姓名	申富饶 教授 宋方敏 教授				

论文涉密情况：

☑ 不保密

☐ 保密，保密期：_____年_____月_____日 至 _____年_____月_____日

注：请将该授权书填写后装订在学位论文最后一页（南大封面）。