心求至善

SOUTHEAST UNIVERSITY

学校代码: 10286
分 类 号: 000
密 级: 公开
U D C: 000
学 号: 220163496

東南大學
硕士学位论文

SEU THESIX 用户手册

手册

研究生姓名: SEU THESIX 开发组

导师姓名: 高德纳 教授

兰伯特 副教授

申请学位类别 TEX 学硕士 学位授予单位 东南大学

一级学科名称 TEX 论文答辩日期 2019 年 4 月 24 日

二级学科名称 LATEX 学位授予日期 2019 年 4 月 24 日

答辩委员会主席 高德纳 评 阅 人 Frank Mittlebach

David Carlisle

2019 年 4 月 24 日

SEU THESIX 用户手册 手册

SEU THESIX 开发组

东南大学

学校代码：10286

分 类 号：000

密 级：公开

U D C：000

学 号：220163496

東南大學

硕士学位论文

SEU THESIX 用户手册

手册

研究生姓名：SEU THESIX 开发组

导师姓名：高德纳 教授

兰伯特 副教授

申请学位类别 TEX 学硕士　　学位授予单位 东南大学

一级学科名称 TEX　　论文答辩日期 2019 年 4 月 24 日

二级学科名称 LATEX　　学位授予日期 2019 年 4 月 24 日

答辩委员会主席 高德纳　　评 阅 人 Frank Mittlebach

David Carlisle

2019 年 4 月 24 日

東南大學

硕士学位论文

SEU THESIX 用户手册

手册

专 业 名 称: TEX

研究生姓名: SEU THESIX 开发组

导 师 姓 名: 高德纳 教授

兰伯特 副教授

本课题的研究获 LaTeX project 赞助:www.latex-project.org

SEU THESIX USER MANUAL

SEU THESIX

A Thesis submitted to

Southeast University

For the Academic Degree of Master of TEX

BY

SEU THESIX developer group

Supervised by:

Prof. Donald E. Knuth

and

Associate Prof. Leslie Lamport

School of TEX

Southeast University

2019/4/24

东南大学学位论文独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得东南大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

研究生签名：_____ 日 期：_____

东南大学学位论文使用授权声明

东南大学、中国科学技术信息研究所、国家图书馆有权保留本人所送交学位论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权东南大学研究生院办理。

研究生签名：_____ 导师签名：_____ 日 期：_____

摘　要

本文介绍如何使用SEU Thesix 文档类撰写东南大学学位论文。

关键词： TEX, LATEX, 文档类，学位论文

Abstract

This work presents an introduction of how to use SEU Thesix document class to typeset the thesis/dissertation of Southeast University.

Keywords: TEX, LATEX, document class, thesis/dissertation

目录

插图目录

表格目录

算法目录

第一章 绪论

1.1 研究背景和意义

检索任务的定义是指根据用户特定的信息需求, 对这种特定的信息采用一定的方法、技术手段, 根据一定的线索与规则找到满足用户需求的信息。同款服装检索作为检索的子任务, 则是需要根据用户提供的需求信息, 在由多种款式、风格的服装图片组成的检索库中找到其同款服装。

几年前, 网页购物的快速便捷极大的促进了人们消费水平的进步, 随后, 各大电商平台进一步将它们的购物应用推广到了用户的手机里。也正是因为这样, 用户的购买习惯也在悄悄地发生着改变: 时间和地点不再是局限, 只要拥有一部连接互联网的手机, 就可以直接获取想要购买的商品。随着移动技术的迅速发展, 移动设备的安全、高速、便捷等特点越来越获得人们的认可, 这就使得移动购物的行为变得越来越普遍。然而目前 PC 和移动终端中, 用户多数情况下还是通过输入文本关键词以获取目标商品, 用这种单一的文本信息去描述商品有时很难表达用户的真实需求。

这种基于文本信息的由粗到精的检索方式, 一定程度上可以帮助用户定位到有具体标签的商品。然而, 当用户需求的商品的一些关键信息不明确时, 抽象出适合的关键词去进行检索就变得很困难了, 这种情况下以图搜图的检索方式能更好的表达用户的需求。CBIR (Content-Based Image Retrieval)[1], 即基于内容的图像检索, 是近十年来计算机视觉最关注的研究领域之一, CBIR 是通过分析提取图像的视觉或语义特征, 使用相似度度量算法, 从检索库中得到一组与其最为相似的图像。从根本上来讲,CBIR 是一种相似度度量技术, 它包含了图像处理、计算机视觉和图像理解等领域, 是国内外研究的热点。

如今, 各大电商平台已经提供了以图搜图的功能, 消费者可以上传实时照片以检索同款商品, 如图1.1所示。随着互联网及多媒体技术的迅速发展, 图像的来源不断扩大, 高速、大容量的存储系统为存储海量图像提供了保障, 图像信息在各行各业都越来越广泛的被使用, 因此, 图像信息资源的高效管理和高性能的检索算法变得日益重要。图像检索是图像数据研究的一项核心技术, 是近年来海量信息处理所面临的“瓶颈”。所以, 对基于图像内容的检索算法的研究有重要意义。

1.2 研究内容及贡献

本课题采用 DeepFashion[2] 数据集作为训练集, 阿里巴巴 2017 大规模图像搜索大赛的数据集作为测试集, 共 300 万张测试图片。评价标准为上装, 裙装和下装的检索 top4 命中率, 对应的目标性能分别为 85%、80%、75%。

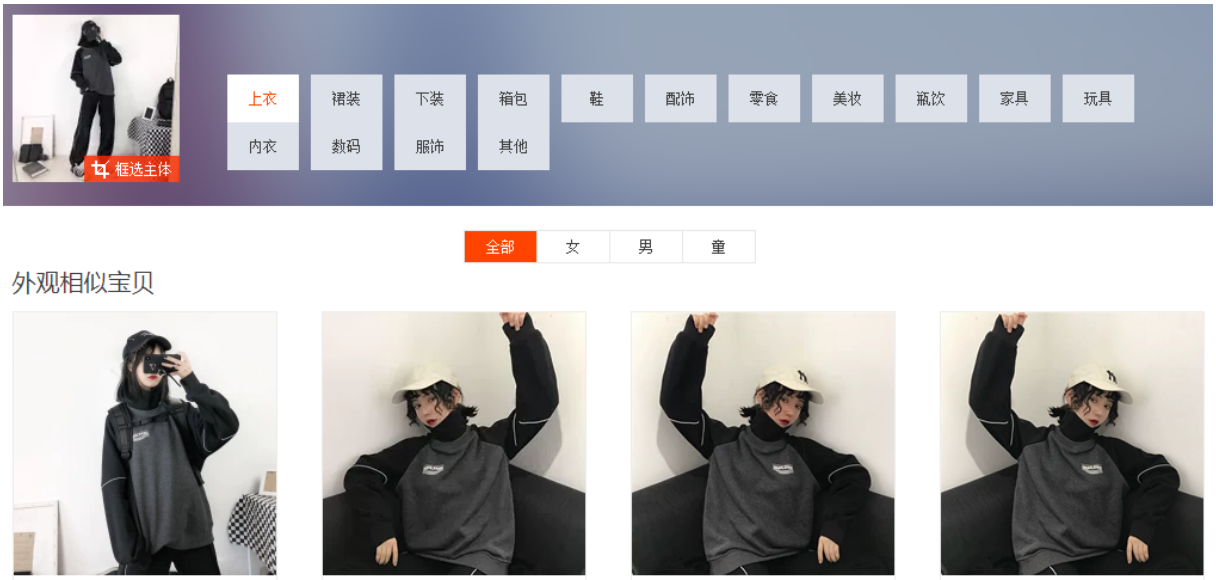


图 1.1: 淘宝提供的以图搜图功能，用于检索同款商品

对于服装图像检索这项任务, 局部特征的学习与对齐一直是研究的重要方向, 这也是本课题面临的最大考验。此外, 深度学习一直以来都有一个问题: 训练好的模型在另外一个域的性能会下降。基于这些问题，本课题的研究内容主要针对如下内容展开:（1）探究度量学习以及多任务学习，以提升模型表达能力;（2）研究模型在不同域数据集之间性能不一致的问题，提升模型泛化能力;（3）探索如何有效结合局部特征以及全局特征，以及更加准确的局部区域对齐方式。据此，本文的主要贡献有如下几个方面：

1. 引入注意力机制，让网络自适应的学习应该重点关注的局部区域。学习特征空间维度的权重分布，弱化图像中的噪音部分，且网络训练过程中除了类别标签外不需要额外的标注信息。采用多个分支并行的架构，可以使得每个分支关注的局部区域有所区别，进一步提升模型的表达能力。

2. 提出跨域样本挖掘的方法。模型在实际使用场景下的输入数据与训练数据集常常有一定的区别，这种情况之下模型的性能往往达不到预期，因为来自不同域的数据集的数据分布与训练集本身就有差别。基于此，本文提出一种跨域挖掘样本的方式：先在原有的训练集训练出一个模型，然后在不同域的测试集中随机选择一定量的图像作为模型输入并在测试集做检索，每张输入图像都会得到 Top K 的排序输出，我们取这 K 张图片中排序相对靠前的图像作为和输入图像的同款服装样本，排序靠后的作为非同款样本，最后把这些样本放入原有的训练集训练。通过这种方式，模型可以更好的适应测试集的数据分布，进而有效提升在测试集的表现。

3. 提出多粒度切分的方法，大幅度增强特征的局部表达能力。分类任务一般基于整幅图的全局特征去做，本文在保留全局特征分类做法的基础上，增加了局部特征分类的分支，采用切分的方式，对特征图切分为 N 等份，每一块局部特征经过 Average Pooling 之后用类别标签去做监督。另外，使用不同大小的 N 作切分，多粒度的切分分支并行训

练，进一步提升模型的性能。

4. 提出环形 Pooling 的策略。首先通过环形的切分方式，得到环形的特征块，对每快特征做环形的平均池化，可以有效解决实际场景常见的旋转问题，结合横向切分和纵向切分，使得网络对局部特征的学习更加全面。

第二章　相关工作

2.1　基于视觉特征的服装检索

服装检索的发展经历了两个阶段, 分别是基于文本和基于内容的服装检索。基于文本的服装检索通过关键字或者更加自由形式的文本信息来描述商品, 然后通过文本匹配算法进行服装商品的检索, 其本质是以文本搜图。目前, 像服装购物平台淘宝等主流电子商务网站检索服装图像都是以 TBIR(Text-based Image Retrieval) 技术为主,TBIR 一般通过关键字检索, 其特点是快速精准, 但是随着服装种类与数量的不断迅速增长, TBIR 的不足之处也开始显现出来: 现在人们对于服装细节的要求越来越高, 然而采用关键字去标注服装图像无法全面准确的表示服装的细节信息;TBIR 需要人工的对海量的服装图像做标注, 工作量很大; 人工标注有时会缺乏客观性, 这会直接导致检索结果的偏差。因此, 对基于内容的服装图像检索算法的研究很有必要性,CBIR 通过学习图像的视觉语义特征进行检索, 可以弥补 TBIR 的很多不足。

CBIR 算法首先会抽取图像的特征, 然后将该特征和检索库里的特征对比, 计算相似度, 按照相似性的大小来排序从而可以得到最终的检索结果。初期的 CBIR 算法一般致力于更好的提取服装的视觉特征, 服装是整个服装图像中的主体, 因其款式多样、颜色鲜明、细节突出等特点, 相对与自然景象或者生活中其他常见物体等背景信息更加具有区分度。从服装的这些特点中我们可以抽取服装的纹理、颜色、形状这三大视觉特征, 对这三种特征的提取是初期 CBIR 算法的主要方向:

1. 纹理特征: 服装面料最为明显, 最具区分性的特征就是纹理, 纹理一般情况下以图像的某种局部特性表现出来, 纹理的多样性决定了服装的外在美观程度, 更好的提取服装图片的纹理特征十分有利于检索得到相似度高的服装图像。传统提取纹理特征的方法主要有四种: 频谱法、统计法、结构法和模型法。纹理特征的提取算法最好具有旋转不变性,Ojala 等人提出的 LBP 纹理分析方法具有旋转不变性的优点, 但该方法只使用了服装图像的局部信息去提取特征, 没有很好的利用图像的全局信息[3] 。Manthalkar 等人的方法具有旋转不变性和尺度不变性, 然而该方法牺牲了纹理的方向信息[4]。

2. 颜色特征: 颜色是服装的重要构成部分, 是图像的一种显著的视觉特征, 其对于图像检索也一直是十分重要的特征。早期对颜色特征的提取主要通过统计图像像素点的值, 近阶段则是偏向于研究图像颜色信息空间分布的检索方法,Pass 等人统计图像中各颜色最大连续区域的像素值, 将其作为颜色特征[5];Strieker 等人将图像划分, 再分别对划分后的各个子区域进行颜色特征的统计提取[6]。总体的来说, 目前基于颜色信息的检索方法主要考虑对全局以及局部颜色特征的抽取。

3. 形状特征: 形状是决定服装款式多样性的重要因素, 蕴含了服装的设计理念和风

格, 其对于人的视觉感受也是十分重要的因素。形状特征的提取被广泛应用于 CBIR, 除了如何更好的提取形状特征之外, 不同形状特征之间的相似度计算也是近来被探索的课题。对形状特征的提取注重关注服装图像的轮廓信息, 以及更好的处理区域特征。轮廓特征指服装的边界形状信息所包含的特征, 相关参数有边界点、面积、周长等, 相关研究有 Livarinen 等人提出链码直方图[7];Berretti 等人提出基于平滑曲线分解特征等[8]。区域特征指的是图像服装区域内部所包含的信息, 常用矩的方法, Chin 等人提出几何不变矩，可以更好地提取形状特征[9]。对形状特相似性度量的研究:Peter 等提出 K 最近邻图[10];Bai 等人利用了形状特征之间的相似性与图之间的相互关系, 将形状特征间的相似性构建为图, 更有效得去度量形状间相似性[11]。

2.2　基于语义特征的服装检索

传统的 CBIR 方法采用颜色、形状、纹理等视觉特征, 这些特征较为底层, 使用的分类器大多是浅层分类器, 如支持向量机。这种基于底层视觉特征的检索系统和人类视觉体系对图像的理解存在着语义鸿沟, 即对于不同的图片, 机器从低级的可视化特征得到的相似性和人从高级的语义特征得到的相似性之间的不同[12]。所以, 即便图像检索领域在对底层视觉特征的提取有了很大进展, 提出了一系列不同的方法, 但由于语义鸿沟的存在, 图像检索依然面临着巨大的挑战, 我们希望机器可以像人一样理解识别图片内容, 则需要从更高层次去分析图片, 现阶段最有希望解决这个语义鸿沟的技术是机器学习。机器学习是一门涉及统计学、概率论、优化算法等领域的交叉学科, 旨在研究如何让计算机像人一样去学习新的知识, 模仿人的行为, 不断的通过学习提高自己。机器学习算法的学习流程是: 人为的将大量数据输入到计算机程序, 让计算机去处理这些海量数据, 使其发现并总结出这些数据背后所蕴含的规律等隐含信息, 机器学习的优势是可以凭借计算机的高性能计算从大数据中学习得到人类无法轻易总结出的规律。21 世纪以来, 机器学习技术不断发展成熟, 应用范围也从开始单纯的字符识别慢慢多样化, 比如生物信息中的基因大数据分析, 金融行业中通过对历史数据规律的分析预测市场走向。

在机器学习中, 深度学习技术近年来得到了爆发性的发展, 深度学习在计算机视觉、自然语言处理、多媒体、语音识别等方向均取得了巨大的成功, 极大的推动了人工智能领域的发展进程。深度学习作为机器学习领域的一个分支, 起源于 80 年代的 BP[13] 神经网络, 这是一种对误差逆向传播的多层前馈神经网络, 其核心思想是通过梯度下降法不断优化网络, 使得算法不断往误差的最小化方向参数调优。深度学习发展如此迅猛主要得益于两个方面的原因: 计算机的计算能力快速发展, 神经网络的训练可以部署在 GPU 上并且可以并行训练, 这使得大规模的神经网络可以训练; 另一个原因则是大数据的快速发展, 训练数据是机器学习算法的核心之一, 随着互联网的发展, 日常产生的数据呈指数级增长, 这些海量的标注数据促进了深度学习模型的训练。相对于传统机器学习, 深度学习可以自动的找出分析问题所需要的重要特征, 随着神经网络的加深, 可以抽取上层次的语义信息。深度网络中比较有代表性的是卷积神经网络,Lecun 等提出的 LeNet-

5[14] 在手写字符识别领域的成功应用引起了学术界对于卷积神经网络的关注,2012 年 Alex Krizhevsky 提出 AlexNet[15], 一举摘下了视觉领域竞赛 ILSVRC 2012 的桂冠, 在百万量级的 ImageNet[16] 数据集合上, 效果大幅度超过传统的方法 [7], 这成为卷积神经网络的一个历史性时刻,AlexNet 之后, 不断有新的卷积神经网络模型被提出, 从 VGG[17]、GoogLeNet[18]、Res-Net[19] 到近期的 Res-NeXt[20]、SE-Net[21]。

和基于视觉特征类似, 对于区域信息特征的学习是基于语义特征的服装检索一个重要研究方向,CVPR2016 的工作 Fashion-Net[2] 在服装检索中通过关键点信息来对局部特征进行对齐操作, 通过第一步预测出关节点, 然后通过关键点信息和池化操作获得对应的局部特征, 这种做法可以较为准确的定位到服装所在区域, 但是需要对于关键点的标注数据进行训练, 资源消耗较大。在服装检索的问题上, 同款服装的不同图片的相似度应大于不同款服装的相似度, 因此度量学习在网络训练时被广泛使用。常用的度量学习损失方法有对比损失 (Contrastive loss) 、三元组损失 (Triplet loss) 、困难样本采样三元组损失 (Triplet hard loss with batch hard mining, TriHard loss) 等。除了以度量损失函数作为网络训练的监督信息以外, 服装的低层信息, 比如颜色、纹理、形状等也常被作为辅助监督, 这种多任务学习的方式使得网络具有更强的表达能力, 学习得到语义信息更加丰富的特征。

2.3 深度卷积网络

目前情况下，包括图像检索在内的许多计算机视觉任务，在大部分情况下都会使用常用的基础分类网络作为其骨干网络充当特征提取器，这些主流的分类网络的性能已经在 ImageNet 得到证明。最经典的卷积神经网络是 Yann LeCun 在 1998 年设计并提出 LeNet，这个用于识别手写字符的网络规模较小，但是包含了现在卷积神经网络的最基本组件：卷积层，池化层，全连接层。Alex Krizhevsky 于 2012 年提出的 AlexNet 是卷积神经网络的一大步：AlexNet 使用 ReLU 取代 Sigmoid 作为激活函数，成功解决了 Sigmoid 在网络较深时的梯度弥散问题；训练时采用了 Dropout 策略，随机忽略一部分神经元，可以有效避免模型的过拟合；引入了最大池化层而不是像之前只使用平均池化层，这有效的提升了特征的丰富性；使用 CUDA 加速神经网络训练，有效利用了 GPU 的计算能力。AlexNet 被提出之后，深度学习飞速发展，越来越多性能优异的基础网络随之被提出，下面简要分析几个主流的基础网络：

1. VGG：相比 AlexNet，VGG 的一个改进是采用连续的几个 3x3 的卷积核代替 AlexNet 中的 11x11 或者 5x5 的较大卷积核。对于给定的感受野，采用堆积的小卷积核是优于采用大的卷积核，因为多层非线性层可以增加网络深度来保证学习更复杂的模式，此外小的卷积核也意味着更少的参数量，更低的计算复杂度。总结性的来说，VGG 在控制计算量增长的同时将网络架构做的更深更宽，分类效果显著提升。

2. Inception：又被称为 GoogLeNet，实际上 Inception 指的是 GoogLeNet 的核心结构。一般来说提高网络表达能力最直接的方法就是增加网络的深度和宽度，但是直接这

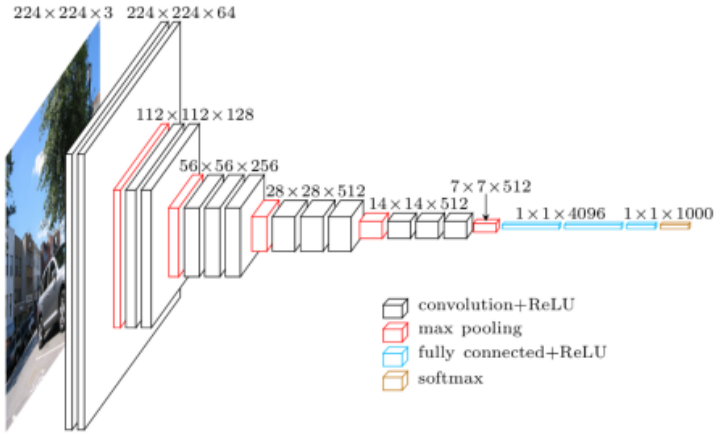


图 2.1: VGG 整体结构

么做会带来一些问题：

(1) 相对更深更宽的网络不可避免的会增加参数量，而过多的参数量容易导致过拟合。

(2) 随着网络的深度的增加，反向传播时会出现梯度消失的问题，导致网络很难优化。

(3) 计算量增加，会消耗更多的计算资源。

Inception 结构针对限制神经网络性能的主要问题对传统的卷积策略不断改进，Inception v1 设计出了多路并行的卷积模块,不同分支的卷积核大小不同,分别有 1x1,3x3,5x5 三种尺度，这么做的好处是可以以不同大小的感受野去学习得到不同的特征，卷积操作之后，不同分支的特征通过拼接的方式做融合。Inception v2[22] 基于 Inception v1 做了进

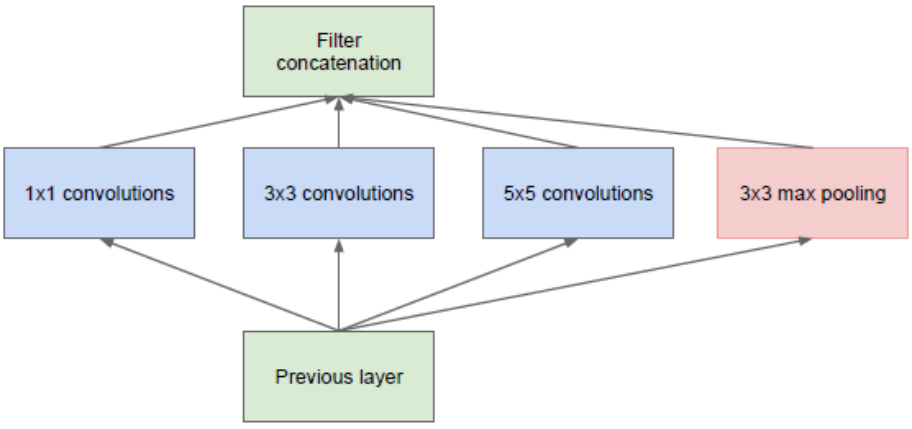


图 2.2: Inception v1

一步的改进，提出了有重大意义的 BN(Batch Normalization)。训练深度神经网络时，作者抛出一个叫做“Internal Covariate Shift”的问题，这个问题指在训练过程中，第 n 层的输入就是第 n-1 层的输出，在训练过程中，每训练一轮参数就会发生变化，对于一个

网络相同的输入，但 n-1 层的输出却不一样，这就导致第 n 层的输入也不一样，BN 的提出就是为了解决这个问题。在传统机器学习中，对图像提取特征之前，都会对图像做白化操作，即对输入数据变换成 0 均值、单位方差的正态分布，卷积神经网络的输入就是图像，白化操作可以加快收敛，对于深度网络，每个隐层的输出都是下一个隐层的输入，即每个隐层的输入都可以做白化操作，BN 就是在训练中的每个 mini-batch 上做了白化，可以有效防止梯度消失并加速网络训练。

3. ResNet：由微软研究院的 Kaiming He 等提出的 ResNet，通过使用 ResNet Unit 成功训练出了 152 层的神经网络，其效果非常优异，在 ILSVRC2015 比赛中夺得头筹。

提出残差学习的思想。传统的卷积网络或者全连接网络在信息传递的时候或多或少会存在信息丢失，损耗等问题，同时还有导致梯度消失或者梯度爆炸，导致很深的网络无法训练。ResNet 在一定程度上解决了这个问题，通过直接将输入信息绕道传到输出，保护信息的完整性，整个网络只需要学习输入、输出差别的那一部分，简化学习目标和难度。VGGNet 和 ResNet 的对比如下图所示。ResNet 最大的区别在于有很多的旁路将输入直接连接到后面的层，这种结构也被称为 shortcut 或者 skip connections。

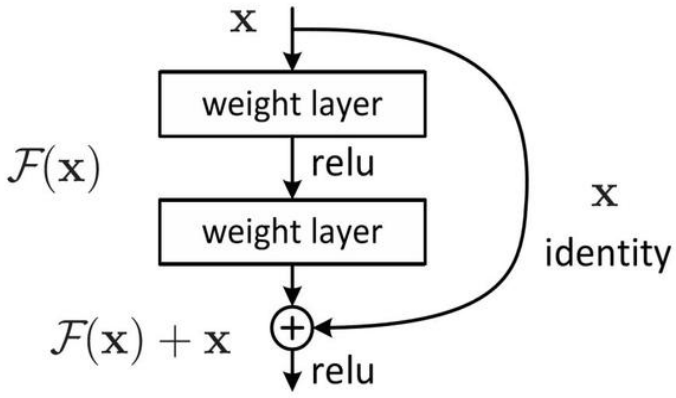


图 2.3: ResNet 残差模块

4. SENet: 近年来，为了提升网络性能，多数工作从空间纬度展开，比如 Incepion 使用多尺度的卷积核以获取并聚合不同感受野的特征，另外比较具有代表性的有将注意力机制引入到空间维度上去，这些工作都取得了很好的效果。而 SENet 则引入了另一种思路：是否可以考虑特征通道之间的关系以提升网络性能？SE 是 Squeeze-and-Excitation 的缩写，Squeeze 和 Excitation 则是 SENet 核心模块的两个关键操作，模块具体流程如下：

(1) 对一个三维的特征图做 Global average pooling，得到特征维度为 $c \times 1 \times 1$，其中 c 为特征图的通道数，这个操作成为 Squeeze。

(2) 随后为两个 FC 层（Fully-connected-layer）去学习通道之间的相关性，其中第一个 FC 层将输入维度降低至原来的 1/16，并经过 ReLu，第二个 FC 层再将特征升至原来的维度。用两个 FC 层的好处是可以增加非线性以更好的建模通道相关性，并且可以大

幅度减少参数量。

(3) 最后通过 Sigmoid 将特征归一化至 0 到 1 之间，代表每个通道的重要程度，并将权重点乘至原特征图上。

2.4　目标检测

服装检索这个任务在执行过程中常常需要用到目标检测，这是由任务的实际场景所决定的，训练深度模型的训练样本一般来自商家或者买家的拍摄图像，而拍摄得到的图像难免会包含除了服装信息之外背景噪声，这个时候就需要目标检测在图像输入网络之前对其做预处理。目标检测是计算机视觉的基础任务之一，相较于分类任务，检测不仅需要识别图像中“有什么”，还要判断所检测物体“在哪里”。以图2.4为例：这是来自买家拍摄的一张图，除了买家身上的裙子之外，图像中还有镜子、桌子、窗帘以及人体等不相关因素，这种情形之下需要一个检测群装的检测器，检测的结果如图中的红色检测框所示，那么我们就可以去除绝大部分的噪声，得到只包含服装的图像。

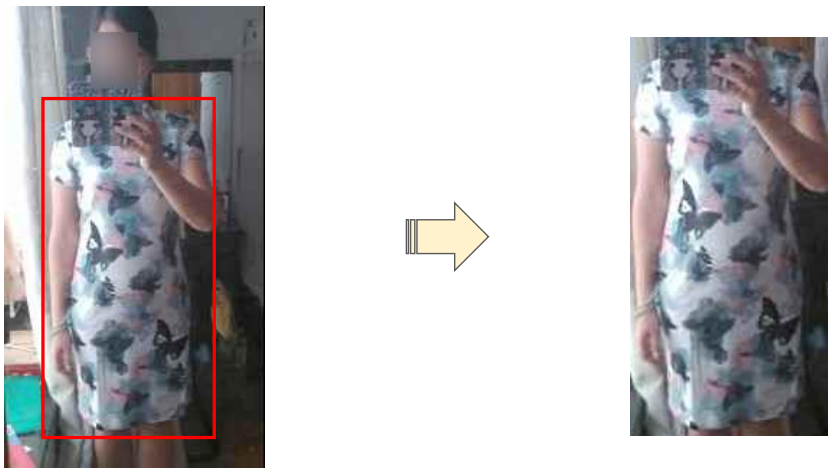


图 2.4: 使用目标检测对检索图像预处理

在深度学习诞生之前，目标检测还是基于人工构建的特征，因此特征的表达能力十分有限，多数工作的重点集中在设计更好的检测算法上，从而弥补特征表达能力的不足。当时比较有代表性的算法有：Viola-Jones 检测器[23;24]，在当时的计算条件下，首次实现了实时的人脸检测；针对行人检测问题提出的 HOG 行人检测器[25]；DPM 算法及其改进[26;27;28] 将对整个物体的检测拆分为对多个部件的检测，最后再整合在一起，DPM 算法里的很多思想一直延续至今，比如上下文信息以及困难样本等。

在深度学习诞生的前几年，目标检测的进度几乎停滞不前，首次将深度学习引入目标检测任务的工作是 R-CNN[29]，这也标志着目标检测进入高速发展阶段。R-CNN 策略简单易懂，首先对输入图像提取了大量候选框（Object proposal），根据每个 proposal 对原图裁减，并缩放至统一之大小后输入 AlexNet 提取特征，最后对特征使用 SVM 分类。

R-CNN 相较于 DPM 系列算法有着大幅度的提升，但是 R-CNN 算法的不足之处也较为明显：需要先提取特征，再用 SVM 分类，无法端到端的训练；不同的 proposal 可能会有大面积的重合区域，重复提取特征造成了资源浪费而且导致检测速度较慢。

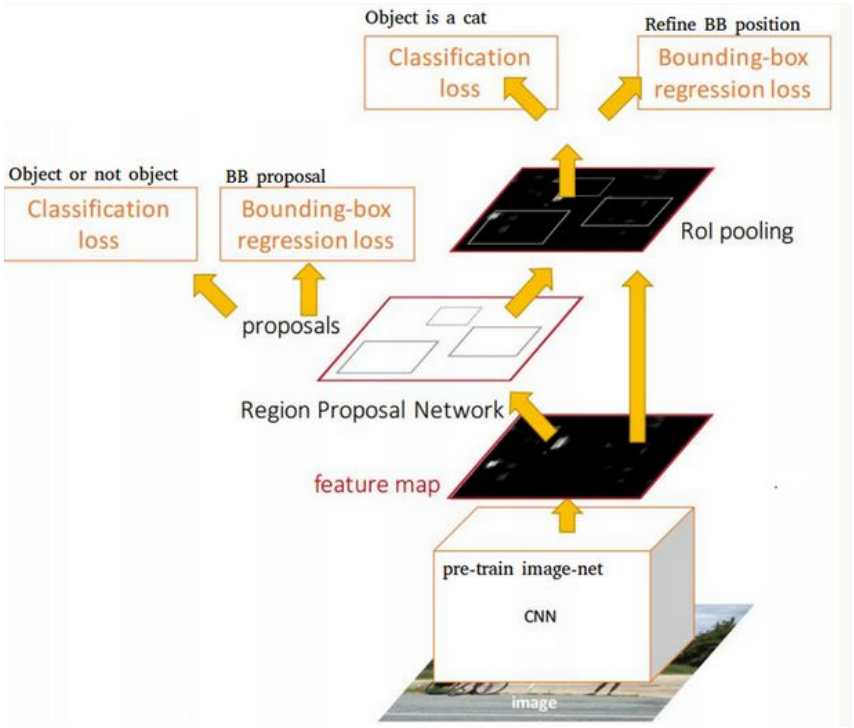


图 2.5: Faster R-CNN 结构图

在 R-CNN 的基础上，Girshick 等人于 2015 年提出了 Fast R-CNN[30]，实现了分类和回归的多任务学习，提升了性能的同时大幅度的加快了网络的训练速度以及检测速度，Fast R-CNN 的检测框生成方式依然是外部算法产生，所以并未实现端到端的训练。后来的 Faster R-CNN[31] 将候选框的提取也交给了网络去实现，这也是 Faster R-CNN 的最大创新之处，即候选区域生成网络（RegionProposal Network），实现了真正的端到端的深度学习目标检测框架，对 640 × 480 的输入图像可以做到 17 帧每秒的检测速度。

第三章　基于注意力机制的局部对齐网络

3.1　引言

基于深度学习的服装检索网络一般可以概括为两个子网络：表示和匹配。表示网络即特征提取器，一般使用在 ImageNet 预训练的主流深度卷积网络，比如 VGG、ResNet 等主流骨干网络；匹配网络对所提取特征进一步学习，以获得更适合服装检索这个任务的特征。特征提取器输出的特征经过 Pooling 之后得到一个向量之后，一般都会引入度量学习去做监督，使同款的服装特征相似度提高，不同款的服装特征相似度降低, 特征的相似度或者距离衡量常用余弦相似度。常用的度量学习损失有 Contrastive loss[32]、Triplet loss[33] 等。

大家在早期对服装检索的研究主要关注在全局信息上，即对整幅图提取一个全局的特征，但是后来这种方式遇到了瓶颈，大家开始慢慢把研究的方向转向对局部特征的学习与表示。对于服装图像来说，全局信息包含了更加丰富的语义信息，但是对局部信息的抽取也非常有帮助，因为不同款式的服装有的时候仅仅有着细微的差别，在全局特征中，这些重要但是不明显或者所占区域比例较小的局部信息会被 Average Pooling 稀释掉，如果可以通过某种方式将这个区域的特征单独提取出来，或者强化这个局部区域的特征强度，对检索结果将会带来巨大的收益。比较常用的局部特征提取方法有对输入图像的切分[34]，或者网格[35] 等，这种方式直接获取局部特征比较简单直接，但是也有其相应的不足之处：同款服装的不同拍摄图像摆放位置及形状不一定相同，两幅图的局部特征并不能很好的对齐。Fashion-Net 使用关键点信息协助局部的定位，网络第一阶段先生成对关键点位置的预测，第二阶段根据关键点位置对局部信息 Pooling。通过关键点的方式可以很好的解决局部部件对齐的问题，但是这个方法需要训练样本有对应关键点的标注信息，带来的资源消耗较大。

近年来，注意力机制（Attention Mechanism）在深度学习的各个领域都被广泛的使用，从自然语言处理到语音识别再到计算机视觉，都很容易看到注意力模型的存在。深度学习中的注意力机制其实借鉴自人类视觉系统的注意力机制。人类的视觉注意力机制的本质是我们大脑的一种信号处理机制，首先对眼睛观测到的图像做全局的扫描和理解，分析之后会把注意力放在需要重点关注的区域，从而抓取更多的细节信息，一定程度上屏蔽相对无关的信息，人的视觉注意力机制有效的提高了对视觉信息的理解效率和效果。在计算机视觉领域，注意力模块往往指一个额外的网络模块，这个模块可以给输入的信息分配不同的权重之后再输出，特殊条件下，如果权重大小只能是 0 或者 1 时，就包括了切分或者网格的处理方式，本节中的注意力机制特指软注意力机制（Soft Attention），即注意力权重为 0 到 1 之间的任意值。注意力模块可以直接嵌入到神经网

络之中的，Soft Attention 的权重输出是可微的，所以整个模型可以进行端到端的训练。

3.2　方法与实现

3.2.1　度量学习

服装检索任务的目标是在由很多款式服装图像组成的检索库中找到和检索图片（query）相同款式的服装。这个任务可以被看作一种排序问题：给定一个 query，那么检索库中和 query 相同款式的服装相对于与其不同款式的服装应当和 query 更加相似。

基于此，本方法引入度量学习训练模型，训练样本组成方式如下：对于一个批次（Batch）的样本 $\mathcal{B}=\{I_1, I_2, \cdots, I_N\}$，我们从中组成一系列三元组，$\mathcal{T}=\{(I_a, I_p, I_n)\}$，其中 (I_a, I_p) 是一对正样本对，表示这是来自同一款服装的两幅照片；(I_a, I_n) 是一对负样本对，表示来自不同款式服装的两幅照片。

由于检索任务的本质是一个排序问题，我们使用三元组损失（Triplet loss）函数优化网络，其数学表达式为：

$$l(I_a, I_p, I_n) = \max\{d(h(I_a), h(I_p)) - d(h(I_a), h(I_n)) + m, 0\} \tag{3.1}$$

这里 $(I_a, I_p, I_n) \in \mathcal{T}$，$m$(margin) 是我们认为负样本对之间的距离和正样本对之间应该有的距离差值，借鉴已有的工作[33]，在我们的实现中，m 取了 0.2。$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2$，代表了欧几里得距离，即欧式距离。$h(I)$ 代表将图像 I 输入网络并提取得到其特征。所以我们可以得到如下完整的损失函数定义：

$$\mathcal{L}(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{(I_a, I_p, I_n)\in\mathcal{T}} l(I_a, I_p, I_n) \tag{3.2}$$

公式里的 $|\mathcal{T}|$ 代表这个 Batch 里所包含所有三元组的个数。

3.2.2　局部对齐网络

网络的第一阶段是一个特征提取器，这个特征提取器是一个深度全卷积神经网络（FCN），其输出是一个特征图，随后将其作为局部特征提取网络的输入，由其提取并输出局部特征。与直接对输入图像做空间上的水平、竖直切分或者使用网格切分的方式不同，我们的目的是提取对齐之后的局部特征。

局部对齐网络如图3.1所示，包含了几个分支，每个分支都以 FCN 的输出作为输入，并检测到一个具有判别力的独立的局部区域，最后将这个区域的特征提取并输出。我们将 FCN 所提取的特征图用一个三维的张量 $\mathbf{T}$ 表示，其每个维度大小为 $w \times h \times c$，代表宽度为 w，长度为 h，通道数为 c 的张量。局部对齐网络的每个分支都会生成一个二维的掩膜 M_i，i 代表第 i 个分支，其大小为 $w \times h$，$M(x, y)$ 的大小表示 (x, y) 这个区域的特征对应的权重。那么对于第 i 个分支来说，其输出特征 $\mathbf{T}_i$ 可表示为如下形式：

$$\mathbf{T}_i(x, y, c) = \mathbf{T}(x, y, c) \times M_i(x, y) \tag{3.3}$$

单个分支除了学习二维的掩膜 M_i 以外，还会同时生成通道维度的权重分配向量 C_i，那么 $\mathbf{T}_i$ 现在为：

$$\mathbf{T}_i(w,h,k) = \mathbf{T}_i(w,h,k) \times C_i(k) \tag{3.4}$$

其中 k 表示张量的第 k 个通道。经过两次变换后得到 $\mathbf{T}_i$ 之后，会通过全局平均池化（Global average pooling）操作得到一个向量，$\mathbf{f}_i = AveragePooling(\mathbf{T}_i)$。随后用一个线性降维层对这个向量降维，降维层采用全连接层实现，$\bar{\mathbf{f}}_i = \mathbf{W}_{FC_i}\mathbf{f}_i$。接下来，我们将来自所有分支的局部特征拼接起来：

$$\mathbf{f} = [\bar{\mathbf{f}}_1^{\top}, \bar{\mathbf{f}}_2^{\top}, \cdots, \bar{\mathbf{f}}_N^{\top}]^{\top} \tag{3.5}$$

最后对拼接后的特征做 L_2 归一化，得到公式3.1中的 $h(I)$。

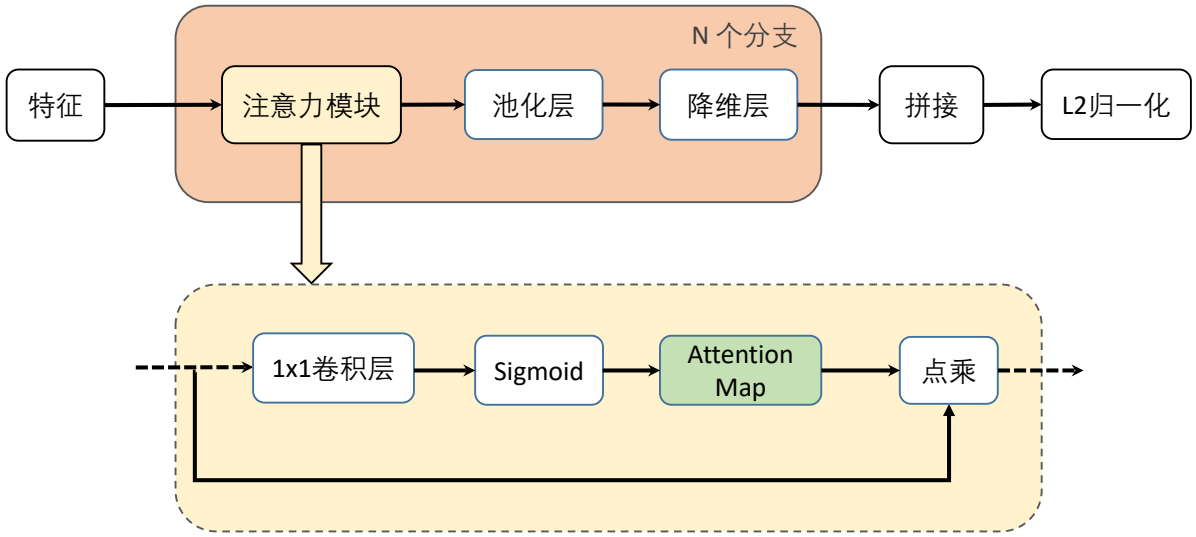


图 3.1: 局部对齐网络的整体架构

3.2.3 注意力模块

本小节介绍如图3.1中所示的注意力模块。在局部对齐网络中，每一个分支都会生成一个掩膜 M_i，这个掩膜的生成也是局部对齐的关键，其生成过程基于注意力机制，我们称之为注意力模块。提出的注意力模块包括两个子模块：空间注意力模块和通道注意力模块。空间注意力模块学习空间维度的注意力分布；通道注意力模块则借鉴 SE-Net 的思想，认为通道间具有互相依赖的关系，学习通道维度上的权重分布可以有效挖掘这种依赖关系以提升网络表达能力。这两个子模块有着相同的输入，即张量 $\mathbf{T}$。

3.2.3.1 空间注意力

对空间注意力的学习旨在挖掘对网络表达能力有益的局部区域特征，根据输入 $\mathbf{T}$ 学习得到一个二维的掩膜 M 并根据 M 更新输入 $\mathbf{T}$，整个过程分为三个步骤：（1）使

用 1×1 大小的卷积核做 **T** 做卷积操作，将 **T** 压缩为单通道的特征图。（2）对得到的单通道特征归一化处理，使得所有的特征值分布区间为 0 到 1 之间，归一化后的特征称为注意力分布图（Attention map），即掩膜 M。Attention map 每个值的大小代表着学习到的注意力分布在这个区域的权重大小，特征值越大，注意力权重越大。（3）将 Attention map 应用到的输入 **T** 之上，完成对空间局部信息的挖掘。

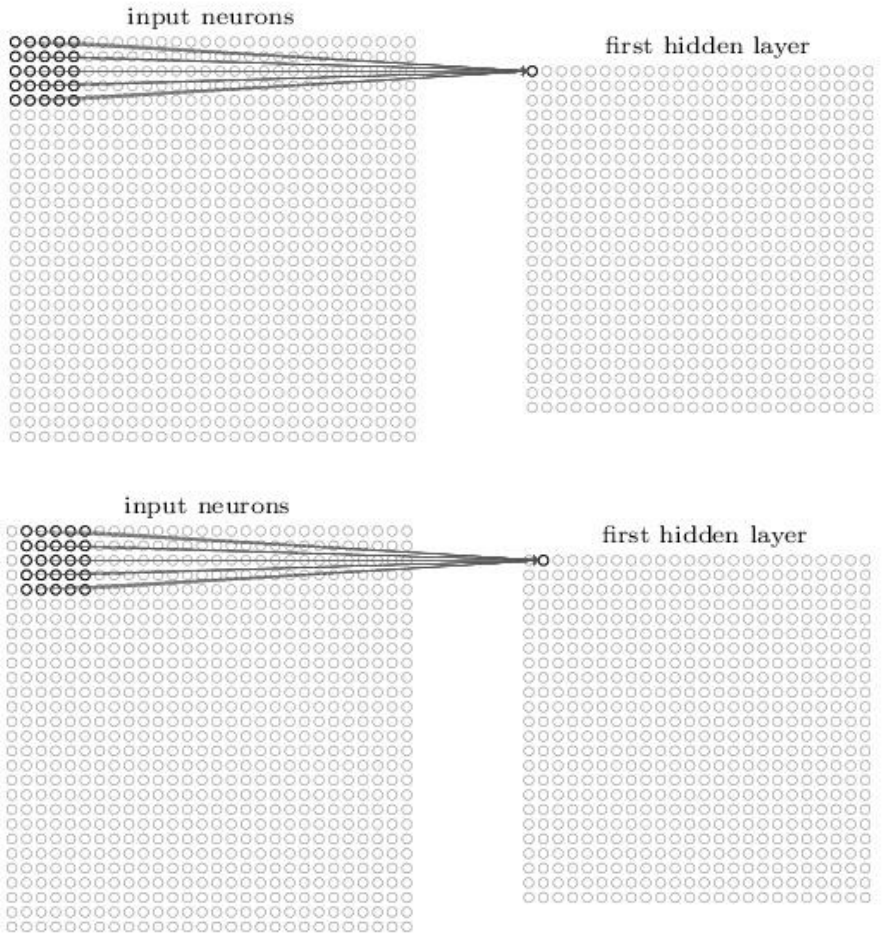


图 3.2: 对卷积操作的简要说明

在卷积神经网络出现之前，深度神经网络的每一层网络都和全连接层相连接，下一层的每个输入节点都和当前层的所有输出节点相连接。对于计算机视觉来说，由于其输入是图像，其空间位置信息对图像的语义理解有着非常重要意义，而全连接层由于和所有输入像素点连接，所以无法捕捉图像的空间信息，这样很不合理。卷积操作是卷积神经网络的核心，如图 3.2所示，对于一个 28×28 维度的输入，采用 5×5 大小的卷积核通过滑窗的方式逐步计算输出值，这个 5×5 的区域称之为感受野，每次只计算感受野内部的特征且全局共享这个卷积核的参数，最后得到 24×24 的输出。5×5 大小的卷积核包含 25 个权重参数 w 以及一个偏置参数 b，每次滑窗的计算即卷积核所有权重和对应感受野特征值的相乘后取和。本方法所提出的空间注意力模块第一步使用的卷积层采

用 1×1 大小的卷积核，这个做法的出发点基于 1×1 卷积核的两个特性：（1）进行跨通道的特征交互与整合。（2）对输入特征的维度变换。Attention map 的生成应基于输入 **T**，且 Attention map 的空间维度应与 **T** 相同，即 $w \times h$。1×1 卷积层以 **T** 为输入，且与图3.2中不同的是，卷积的输出可以保留特征的原空间大小，并将 c 个通道压缩至一个。

卷积层输出的特征图在维度上已经和我们需求的 Attention map 尺度一致，接下来需要对特征值做归一化，保证每个值的大小都在 0 至 1 之间，这里我们使用 Sigmoid 函数去实现归一化，其表达式为：$\sigma(z) = \frac{1}{1+e^{-z}}$。一般情况下，Sigmoid 在深度神经网络中被用作激活函数，激活函数的意义在于引入非线性，而我们使用 Sigmoid 的原因在于它的一个重要特性：输出范围为 0 到 1，如图3.3所示。由于这个特性的原因，Sigmoid 除了作为激活函数外，也常常被用作二分类的概率输出层，其输出值表示类别为真的概率。在我们提出的注意力模块中，我们可以把 Attention map 的每个值看作一个二分类的概率，表示当前区域是否是为需要分配更大权重的局部信息。

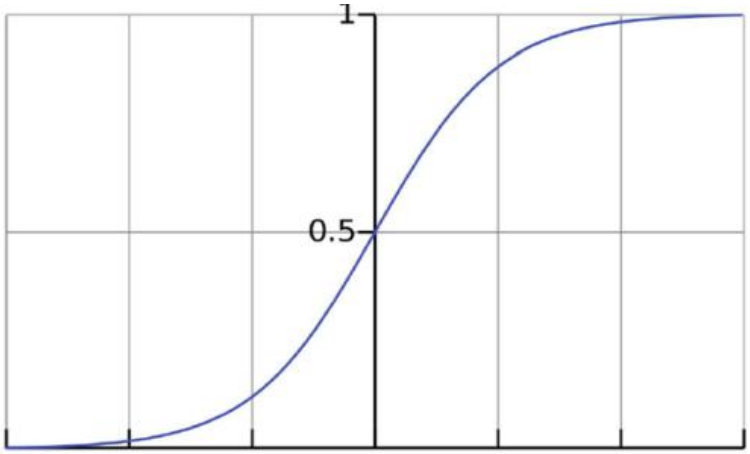


图 3.3: Sigmoid 函数

得到的 Attention map 为 $w \times h$ 大小的二维特征图，而输入 **T** 的为 $w \times h \times c$ 大小的张量，因此需要将 Attention map 复制 c 份，得到与 **T** 相同的尺寸，两个张量做哈达马积（element-wise product）引入注意力权重，得到空间注意力模块的输出。

3.2.3.2 通道注意力

通道注意力模块借鉴自 SE-Net，这个 ImageNet 2017 竞赛图像分类的冠军架构在 ImageNet 数据集上将 Top-5 error 降低到 2.251%。近两年来，SE-Net 因其强大的性能在深度学习领域被广泛使用，也为众多研究人员提供了新的思路和参考。由于 SE-Net 考虑的是特征图不同通道之间关联性，所以其初衷并不是为了做视觉注意力，此处视觉注意力特指图像在空间维度的注意力。

Sun 等人认为 SE-Net 通过调节通道间的权重分配也可以强化或者抑制特征图空间区域的局部响应强度，采用多个分支并行的方式，每个分支可以得到不同的局部高响

应[36]。这种同构多分支的网络设计方式本质是一种多特征融合，这种策略的有效性也已经在很多任务得到证明：谷歌于 2017 年提出的 Multi-head attention[37] 将这种结构用于机器翻译任务，使得网络性能得到有效提升；Hu 等人将相同的理念应用在目标检测任务[38]，也取得了成功。

通道注意力模块单分支的设计与 SE-Net 一致，如图3.1：首先通过全局平均池化（Global average pooling）将输入 **T** 压缩 c 维的向量，然后经过全连接层（FC layer）编码学习通道注意力特征，最后经过 Sigmoid 归一化。得到的通道注意力向量与空间注意力模块的输出相乘后完成当前分支注意力模块的输出。

3.2.4　跨域样本挖掘

传统的机器学习方法要求源域数据和目标域数据具有相同的分布，从而保证在源域数据集训练的模型性能具有较高的可信度，然而在大多数情况下，训练网络模型的数据集和实际应用场景之下的数据集很难具有同分布。因此，在已有源域数据训练出的模型的情况下，如何借助模型提取的特征知识，实现模型从源域到相似数据域的迁移具有重大的意义。本节介绍一种结合在线三元组损失（Online triplet loss），通过已有模型在目标域挖掘无标签样本，并学习其数据分布，达到模型迁移的方法。

3.2.4.1　网络训练

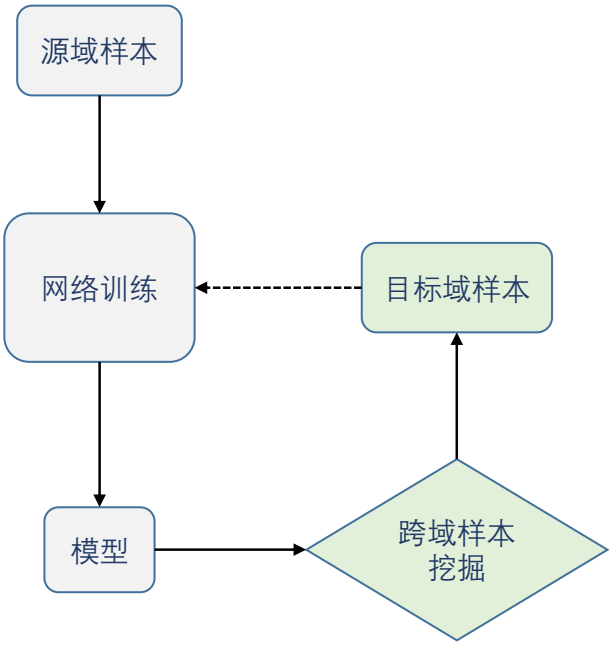


图 3.4: 网络训练流程图

本文所提出的跨域样本挖掘方法是一种可以复用，因此基于此方法的网络训练也是一种迭代的方式，如图3.4所示，网络的训练流程分为如下步骤：

1. 使用源数据域的数据组成训练集，并在此基础上训练出一个模型，训练用的深度网络采用上节所述的局部对齐网络。

2. 根据步骤 1 训练所得模型在目标域做样本挖掘，这一过程会在目标域中获取一定量的样本，并附有标签信息。

3. 将在目标域获取的样本和源域样本一起组成新的训练集，重新训练网络，得到新一轮的模型。

4. 重复执行步骤 2-3 多次，直至模型性能稳定。

结合企业的实际应用场景来说，用于模型训练的数据一般来自商家拍摄的服装图像，这些数据也就是源域。而模型在实际使用时，输入是用户拍摄的服装图像，这些由用户拍摄的图像数据集合可以看作是目标域。来自商家的样本一般拍摄质量较高，其中多数为模特摆排，光线等环境条件也比较稳定；而来自用户的检索样本随机性较强，衣服的形状、摆放位置、光照条件、是否有遮挡或者旋转等因素都十分不稳定。因此，来自商家和来自用户的数据分布具有一定程度的区别。此外，来自用户上传的检索图像没有标签信息，无法直接用于训练，采用人力标注的方式也不便执行，并且会耗费大量人力资源。

本节提出的跨域样本挖掘的方式可以在目标域挖掘无标签样本，并且可以为其生成伪标签用于网络训练。另外，用户上传的图像数据是一个不断增加的过程，随着目标域数据量的逐步递增，其数据分布也是一个渐变的过程。基于跨域样本挖掘的网络迭代训练的模式则十分适合这个应用场景，在对过去所学习到的知识的基础上，不断迭代优化以适应新的数据分布，本质上来说也是一种增量学习。

3.2.4.2 样本挖掘

图 3.5: 样本挖掘流程图

本小节详细介绍跨域样本挖掘的具体细节，挖掘流程如图3.5所示：

1. Query 采样。我们把模型的输入，即待检索图像称为 query，每次基于源域训练集训练得到一个模型之后，从目标域随机采样一部分图像 query，剩下的图像作为检索库。

2. 相似度比对。检索任务的本质是基于相似度或者距离的排序任务，比对的对象是输入的 query 以及整个检索库里的所有样本。以步骤 1 得到的所有 query 作为模型的

输入，在目标域做检索，根据检索库里图像特征和当前 query 对应的特征的相似度排序，我们保留 top N 的结果，即前 N 个和输入 query 最为相似的图像。

3. 伪标签生成。对与得到的 top N 的结果，认为前 m 个样本与 query 是同款服装，前 i 至 j 个样本与 query 是非同款样本，其中 $0 < m < i < j < N$。

挖掘到的样本对应的标签是上述算法得到的伪标签，伪标签最然没有绝对的准确性，但是是由模型的检索结果决定，所以按照伪标签去训练不会使模型的参数分布产生巨大变动。

3.2.4.3　在线三元组损失

如上文所述，训练时引入度量学习作为网络的监督，即 Triplet loss。传统的三元组（triplet）组成采用离线的方式，在网络训练前将样本配成三元组，训练时依次将所有的三元组送入网络。这种方法较为笨拙，预先组成的三元组数量非常庞大，用文本的方式保存也非常不灵活。我们采用在线组成三元组的方式训练网络，有效简化了这一问题。

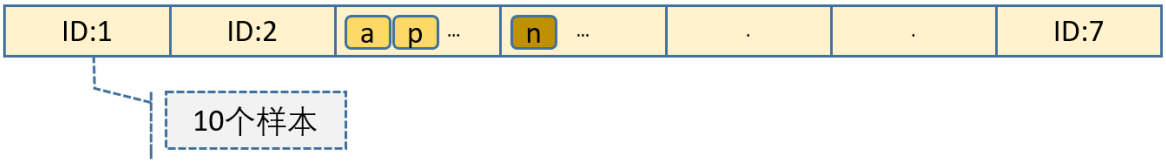


图 3.6: 在线三元组的采样方式

在图3.6中描述了网络在训练过程中每个 Batch 的样本构成方式，Batchsize 为 70，共包含 7 个不同的 ID，每个 ID 又由 10 个样本组成。由三元组的定义可知，任意两个来自相同 ID 的样本，如图中的 a(anchor) 和 p(positive)，以及另一个来自任意不同 ID 的样本，如图中的 n(negative) 便可以组成一个三元组，网络在训练时在线遍历了当前 Batch 所有的三元组并统计当前 Batch 所有三元组的损失的均值，每个 ID 的损失计算如公式3.6:

$$\mathcal{L}_{ID} = \sum_{a=1}^{|ID|} \sum_{p=1}^{|ID|} \sum_{n=1}^{|\bar{ID}|} \{l(I_a, I_p, I_n); p \neq a\} \tag{3.6}$$

其中 $|ID|$ 代表当前 ID 的样本总数，$|\bar{ID}|$ 代表除了当前 ID 外，Batch 里其余所有样本的数量。

图3.6所描述的样本组成及损失计算方式适用于训练样本仅来自于源域的情况，当在源域的数据集下训练出第一个模型并在目标域挖掘出跨域的训练样本之后，后续网络训练使用的在线三元组损失的计算策略作出了对应的调整。其中与图3.6描述的方式有如下几点区别：

1. 训练样本来自两个域：源域和目标域。其中来自源域的训练样本包含 M 个 ID；来自目标域的训练样本由跨域样本挖掘算法得到，包含 N 个 ID，每个 ID 由挖掘算法中提及的 query 以及其检索对应的 top m 的样本组成。

2. 在每个 Batch 的组成过程中，随机从 M+N 个 ID 中采样，若采样得到的 ID 来自目标域的 N 个 ID，则将这个 ID 的 query 对应的 top i-top j 组成第 M+N 个 ID 与之一起送入当前 Batch，这种情况下的 Batch 构成与图3.7一致，三种颜色分别代表了来自源域和目标域以及额外的 ID：M+N。

3. 不统计属于 ID：M+N 的样本作为三元组里的 a 或者 p 时所产生的损失。其出发点也很容易理解：我们认为 top i - top j 的样本含有大量噪声，应属于 query 的非同款样本，且他们本身不应看作同一款服装样本，但是当它们作为一个 ID 与 top m 对应的 ID 样本组成三元组时时可以使得 top m 的样本之间的距离越来越近，且 top m 的样本和 top i - top j 之间的距离越来越远。

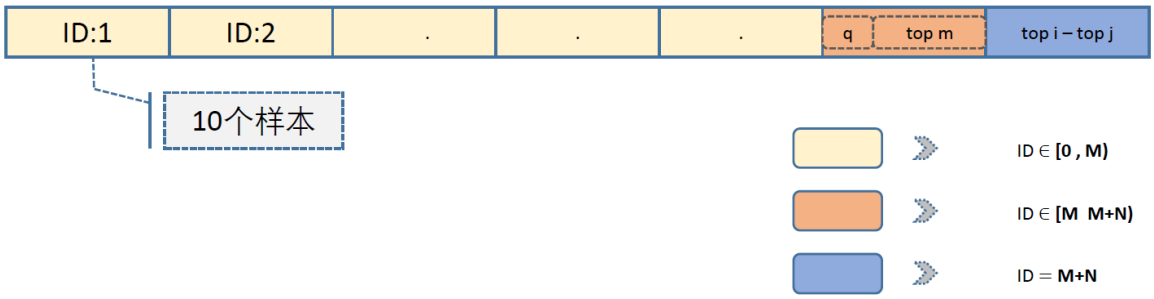


图 3.7: 包含跨域训练样本的 Batch 组成方式

3.3 实验与分析

3.3.1 数据集介绍

全部实验在 DeepFashion 数据集上训练，DeepFashion 是一个于 2016 年推出的大规模时尚服装数据库。整个数据集包含超过 80 万类的服装款式，覆盖了上装、下装、群装三大类服装，其中的图像来源分为买家和卖家。来自卖家的图像拍摄质量较高，多数为在服装商店的摆拍；来自买家的图像则随机性较大，完全由消费者拍摄，没有摆放方式等规则限制。数据集含有丰富的服装标注信息，除了款式标注之外还包括了检测框标注、关键点标注以及多种属性标注。检测框用于图像输入网络前的预处理；属性信息一般指颜色、风格、尺寸、适宜人群等，在网络训练时常使用属性信息做监督以提高特征的语义表达能力。

用于测试模型性能的测试集为阿里巴巴大规模图像搜索竞赛数据集（ALISC2015），这个数据集由三个部分组成：检索集（3,195,334 张无标注图像）、验证集（1417 张 query，以及这些 query 对应的同款服装图像及标注）、测试集（未公开，用于比赛提交时在线测试算法性能）。在 ALISC2015 数据集中，共包含十类图像类别：背包、饮料、家具、箱包、群装、上装、下装、饰品、零食、鞋子。我们从验证集中挑出上装、下装以及群装三个类别，验证集中除了 query 以外的有标注的数据和检索集一起构成检索库，用于模

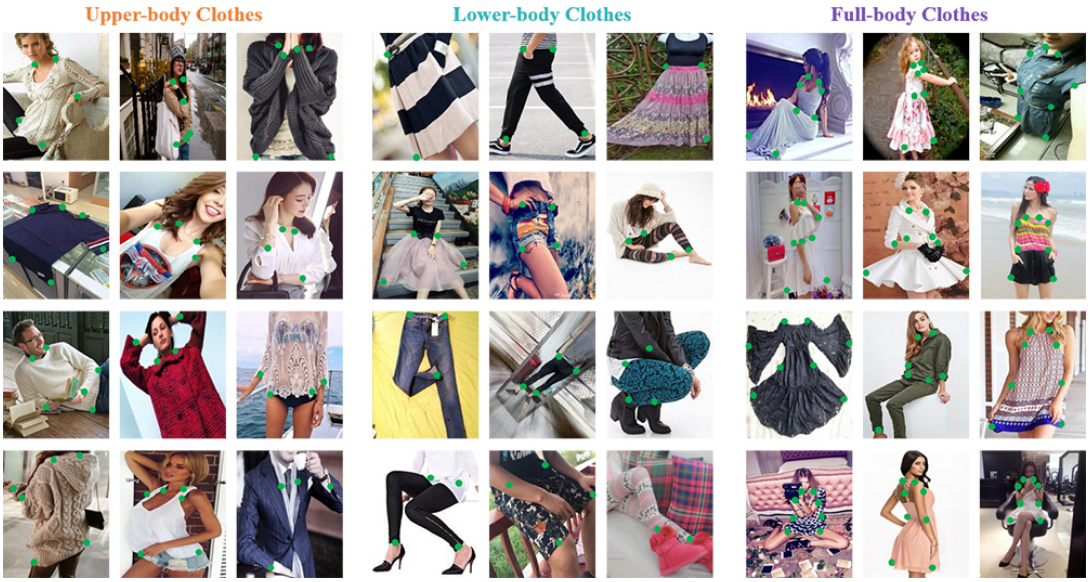


图 3.8: DeepFashion 数据集

型性能的测试，我们可以看出 ALISC2015 数据集的检索库中带有标注的样本占着非常小的比例，可以较为准确的体现模型的性能。

3.3.2　实验设置

3.3.2.1　评价指标

模型检索性能的评价指标使用 top 4 的准确率，其定义为：检索返回的排序结果的 top 4 中至少出现一张同款服装图像的 query 数目所占所有测试 query 数目的比例。

3.3.2.2　预处理

由于 ALISC2015 中没有检测框的标注，我们使用方等人所标注的检测框[39]。他们为 ALISC2015 中的每个类别标注了 1000 张图像，我们用这些标注图像训练 Faster R-CNN，使用 AlexNet 作为预训练模型。在对输入图像做检测时，每幅图最多输出一个检测框，若打分超出 Faster R-CNN 域值的检测框有多个，则取分数最高的那个。

我们基于目标检测预测的检测框对图像进行裁剪，对得到的图像较长的边缩放至统一尺寸（227 个像素点），而较短的边按照与长边相同的缩放比例缩放，并在两侧 Padding，即填充值为零的像素点，这样所有的图像都可以统一为尺度 227×227 的输入。

3.3.2.3　数据增强

对于电商平台使用基于图象内容的检索算法的实际应用场景来说，买家上传的检索图像（query）质量无法保障，常常会有旋转、遮挡、光线等的难以避免的因素使得模型

图 3.9: 对输入图像的预处理

的检索结果产生一定的偏差。针对这些因素，我们在模型训练阶段对训练数据在以下几个方面做了数据增强，以提升模型的鲁棒性：

1. 以一定概率对输入图像做水平翻转、竖直旋转（180 旋转）以及随机小角度旋转。水平翻转可以使模型具有识别镜像中服装的泛化能力，因为有一部分买家拍摄的图像是对着镜子的自拍照；竖直旋转的数据增强是考虑到部分图像是买家将衣物摆放在床上拍摄，这种情况就会有上下颠倒的可能性；随机小角度的旋转是针对人体的姿势或者相机角度对拍摄图像角度产生的小幅度干扰。

2. 以一定概率对输入图像进行亮度和对比度的调整。光照因素对基于深度学习的计算机视觉任务一直都是十分重要的问题，不同的光线条件对模型性能的影响很大，如果只用室内的图像训练模型，那么模型对于室外的输入图像就很难识别。不同的服装图像可能拍摄于多种光照强度之下，因此对输入进行调整亮度、对比度的样本扩充很有必要。

3. 以一定概率对输入图像生成随机位置、随机大小的矩形掩膜，即将原始图像的某个矩形区域擦除，对应的像素值填充为 0。Jon 等人指出，这种随机的擦除策略对行人重识别算法的性能有明显的提升[40]，其原因是对原图的随机的擦除类似于深度网络中常见的 Dropout 层，强制网络基于缺失部分信息的图像去学习，可以有效防止过拟合，提升网络泛化能力。在我们的实验中，掩膜的长和宽取在原图长和宽的 1/16 至 1/8 之间的随机数。

3.3.2.4 特征提取网络

实验使用 SE-Inceptionv2 作为特征提取网络，我们针对上装这一类对多个基础网络的基础性能（Baseline）做了对比，即用不同的基础网络训练分类任务，测试阶段取网络的 Pool5 层特征做检索测试，实验结果显示 SE-Inceptionv2 的 Baseline 性能要高与其他的基础网络，如 VGG、Inceptionv2、SE-ResNet152、ResNet152，SE-Inceptionv2 指的是在 Inceptionv2 的网络架构中，嵌入 SE-Net 的核心模块，这种结构得益与 SE 模块设计的灵巧性，SE-ResNet 也是采用了同样的方式。具体实验结果如图3.10所示:

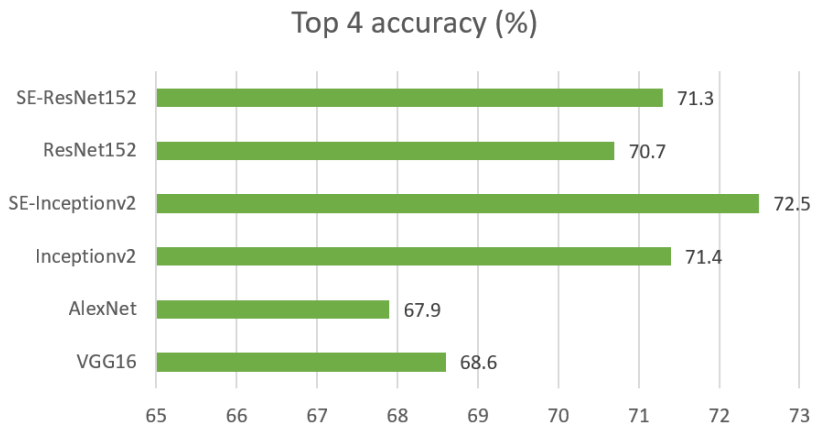


图 3.10: 不同基础网络的 Baseline 性能对比

3.3.2.5　损失函数

本小节全面探究了度量学习与多任务学习对本方法性能产生的影响，使用 SE-Inceptionv2 作为特征提取网络，取 Pool5 的特征作为输入图像的最终表示。表3.1中 ID loss 指 Softmax loss，用服装的类别信息监督；Attribute loss 指使用 DeepFashion 中的属性标注作为辅助监督，每个类别的属性都对应一个多分类的任务，比如颜色属性下对应多种具体的颜色类别。我们从实验结果可以得出以下结论：

表 3.1: 度量学习与多任务学习的有效性分析

损失函数	上装	下装	群装
ID loss	72.5	68.8	61.2
ID loss + Attribute loss	73.0	69.2	62.3
Triplet loss	74.4	**71.0**	**63.5**
Triplet loss + ID loss	**74.8**	70.5	63.2
Triplet loss + ID loss + Attribute loss	**74.8**	70.3	63.2

1. 实验表明在不引入 Triplet loss 时，Attribute loss 作为辅助监督对网络性能有一定帮助。

2. Triplet loss 作用最为明显，相较于 ID loss 在每个服装类别都有两个点左右的提升，这充分证明了度量学习对于检索任务的有效性。

3. 在引入 Triplet loss 之后，ID loss 和 Attribute loss 都不再能为网络的性能带来提升。

基于以上实验分析，本方法采用 Triplet loss 作为网络的损失函数，且不再使用额外的监督，如服装类别和属性信息等。

3.3.3 注意力模块

局部对齐网络的注意力模块包含空间注意力模块和通道注意力模块，如图3.1，我们分别验证了这两个子模块的效果。

表 3.2: 空间注意力及通道注意力的有效性分析

实验	上装	下装	群装
Baseline	74.4	71.0	63.5
空间注意力	75.0	71.3	64.2
通道注意力	74.8	71.0	63.9
空间注意力 + 通道注意力	**75.3**	**71.3**	**64.7**

实验设置为单分支，即图3.1中的分支数 K 取 1，表3.2中，Baseline 指的是表3.1中第三组实验，即使用完整的 SE-Inceptionv2 网络。使用空间注意力模块或者通道注意力模块替换 SE-Inceptionv2 最后一层卷积层之后的部分，网络性能均有提升，且结合空间与通道注意力的完整的注意力模块可以进一步提升性能，这说明空间与通道的注意力学习可以互相促进。

3.3.4 局部对齐网络

本节探讨局部对齐网络分支数以及每个分支输出特征的维度对实验结果带来的影响。对于 K 值的对比实验如图3.11所示，K 在取 4 时得到最好的实验结果，且 K 在取 1

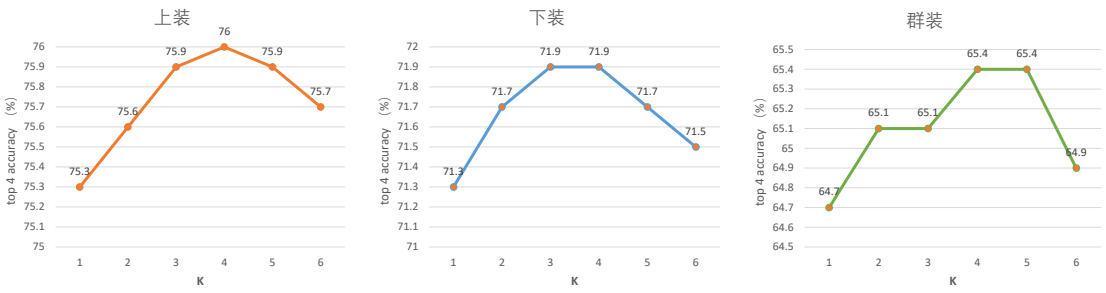


图 3.11: 局部对齐网络分支数（K）对网络性能的影响

到 4 时对应的性能呈现递增的趋势，K 大于 4 时性能会逐渐下降。

所提出的局部对齐网络之所以设计为多个分支，是因为我们希望每个分支通过自己的注意力模块学习得到自己关注的局部区域，且不同的分支可以学得不同的注意力分布。为了更直观的验证网络是否达到这个设计思路预期，对每个分支的 Attetion map 做了可视化分析。每个分支的 Attention map 均为 7×7 大小，我们将 Attention map 采用最

近邻差值的方法上采样至原图大小 (227×227)，再将其复制为三份（RGB 三个通道）并与原图做哈达马积，得到图3.12的效果：其中第一列为输入网络的原图，右侧四列分别

图 3.12: 局部对齐网络不同分支 Attention map 的可视化

是局部对齐网络四个分支的注意力分布。我们可以看出即使网络的四个分支结构相同，但是其注意力分布却有所不同，有的分支注意力偏向于服装的整体，仅仅抑制掉了图像的背景部分；有的分支注意力分布在服装的上半部分或者下半部分；还有的分支侧重于服装的细节部分，比如 Logo 图案等，这种细节可能是区别这种款式服装与其他款式的关键信息。

表 3.3: 局部对齐网络单个分支特征维度对网络性能的影响

单分支特征维度	上装	下装	群装
128	75.2	71.3	64.5
256	76.0	71.9	65.4
512	76.4	71.9	65.5

每个分支的输出维度由分支最后一层全连接层（图3.1中的降维层）维度决定，默

认为 256 维，这也是由相关对比实验得出的结论。单分支特征输出维度取 256 维时对比 128 维有较为明显的提升，而 512 维提升较为微弱，综合权衡检索算法消耗的时间与性能，降维至 256 维最为合适。

3.3.5 跨域样本挖掘

本小节讨论跨域样本挖掘算法的有效性，在所提出的样本挖掘算法中，由三个核心的参数，即决定目标域样本伪标签的 m、i、j 的取值。由于这三个参数的组合总数量比较庞大，我们采取固定其中两个参数调节另一个的方式确定所调节参数的取值：首先令 $i = 30$，$j = 40$，采用初始值为 5，步长为 5 的方式调节 m 的取值，根据不同的 m、i、j 的取值组合进行跨域样本的挖掘，并分别将从目标域挖掘得到的样本与来自源域的训练样本一起组成新的训练集进行下一轮的模型训练，对比模型的性能挑选出最具优势的参数组合；确定下来 m 之后，j 依然取 40，用同样的方式，i 在 $m + 5$ 至 $j - 5$ 之间调节，根据下一轮模型的性能确定 i 的取值。实验结果见图3.13，由于上装、下装、群装

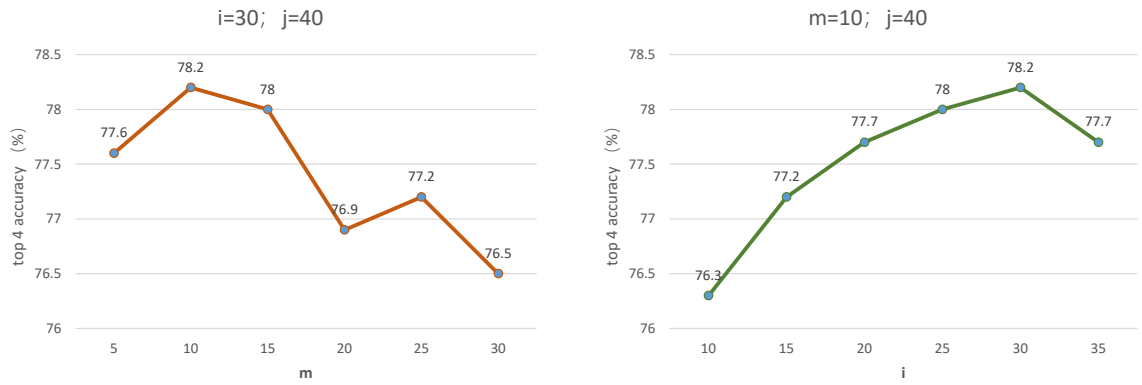


图 3.13: 伪标签生成相关参数（m、i、j）组合与第一轮挖掘后模型性能的对比试验

三个类别的实验结果分布基本一致，图中只展示了上装的实验结果。可以看出不同的伪标签生成区间性能有所区别，但相对于挖掘样本参与训练前的模型（只有源域数据参与训练的 Baseline 模型）的性能（76%），第一轮样本挖掘后训练的模型会有较为可观的提升。基于模型性能的分布，选择 top 10 的检索样本为 query 的同款，top 30 至 top 40 的检索样本为 query 的非同款。

本文所提出的基于跨域样本挖掘算法的网络训练流程是一个可迭代的过程，每次迭代得到的模型都可以重新用来做新一轮的样本挖掘，理论上来说迭代的次数并没有上限，所以迭代停止的判断标准应是模型性能是否趋于稳定。我们将跨域样本挖掘一次并重新训练模型成为网络的一次迭代，网络的迭代次数与模型性能的变化如表3.4所示：迭代次数 0 作为 Baseline，代表只由源域训练集训练的模型性能。网络迭代的前两次对模型性能提升都较为明显，这充分说明了所提出跨域样本挖掘算法的有效性。从第三次迭代以后，模型的性能趋于稳定，这表明新的模型已经在一定程度匹配目标域的数据分

表 3.4: 对基于跨域样本挖掘算法的网络训练不同迭代轮数的对比试验

迭代次数	上装	下装	群装
0	76.0	71.9	65.4
1	78.2	72.7	66.5
2	79.8	74.1	67.8
3	**80.2**	**74.8**	68.4
4	79.8	74.6	**68.7**

布。

3.4 本章小结

本章提出了一种创新性的局部区域对齐网络，用于解决服装检索局部信息的定位与匹配问题。所提出的网络设计方式的核心方法基于注意力机制，这是一种借鉴自人类视觉信息处理方式的机制，这种网络设计方式在训练时只需要服装样本的类别（款式）信息，不需要有关局部信息的任何标注去监督。与将服装图像切分为网格状或者条状不同，本章所提出的方法自适应的定位输入图像的局部区域，这种对齐方式相对来说更为可靠和准确。

此外，基于在线三元组损失，本章还提出一种对跨域样本的挖掘算法。模型的跨域表现多数情况下都会有一定程度的下降，这是因为模型不能很好的匹配目标域的数据分布，所提出的跨域样本挖掘算法基于源域训练的模型在目标域挖掘无标签样本，并给样本打上伪标签与源域训练集一起重新训练模型。算法可以迭代使用，通过挖掘目标域的无标签样本，使得模型不断拟合目标域的数据分布，泛化能力也逐渐提升。这种迭代训练的方式类似于增量学习，可以很好的迎合企业的实际需求，随着用户上传的无标签数据的增加，可以每隔一段时间实施一次样本挖掘保持模型的精度。

第四章　基于多粒度切分的局部对齐网络

4.1　引言

近年来，局部信息在计算机视觉的各种任务中越来越经常被用到，这些任务包括但不限于：细粒度分类[41]、行人重识别[42]、视觉问答、图像检索等做好服装检索这个任务的关键就是尽可能的去优化图像的局部信息表示，而学习局部特征最为重要的问题就是局部区域的定位与对齐。以对局部特征不同的学习方式为区分，现阶段一些比较好的方法可以分为两种路线：基于关键点[43] 和基于语义信息[44]。基于关键点的做法需要训练数据具有关键点的标注，而直接基于语义信息去做则不需要额外的标注。

在第三章中提出的基于注意力机制的局部对齐网络就是用语义信息去做局部的定位，基于注意力的做法比传统的对图像按照一些规则切分更加准确和灵活，在本章，我们介绍一种具有创新性的切分方式去学习局部特征。

4.2　方法

传统的基于切分去学习局部特征的做法大多数是对原图做切分[45]

参考文献

[1] Kato, Toshikazu, Kurita, Takio, Otsu, Nobuyuki, et al. A sketch retrieval method for full color image database-query by visual example[C]. In: [1992] Proceedings. 11th IAPR International Conference on Pattern Recognition. 1992. 530–533.

[2] Liu, Ziwei, Luo, Ping, Qiu, Shi, et al. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. 1096–1104.

[3] Ojala, Timo, Pietikäinen, Matti, and Mäenpää, Topi. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, (7):971–987.

[4] Manthalkar, Ramchandra, Biswas, Prabir K, and Chatterji, Biswanath N. Rotation and scale invariant texture features using discrete wavelet packet transform[J]. Pattern recognition letters, 2003, 24(14):2455–2462.

[5] Pass, Greg, Zabih, Ramin, and Miller, Justin. Comparing Images Using Color Coherence Vectors.[C]. In: ACM multimedia. 1996. 65–73.

[6] Stricker, Markus and Dimai, Alexander. Spectral covariance and fuzzy regions for image indexing[J]. Machine vision and applications, 1997, 10(2):66–73.

[7] Iivarinen, Jukka, Peura, Markus, Särelä, Jaakko, et al. Comparison of Combined Shape Descriptors for Irregular Objects.[C]. In: BMVC. 1997.

[8] Berretti, Stefano, Del Bimbo, Alberto, and Pala, Pietro. Retrieval by shape similarity with perceptual distance and effective indexing[J]. IEEE Transactions on multimedia, 2000, 2(4):225–239.

[9] Teh, C-H and Chin, Roland T. On image analysis by the methods of moments[C]. In: Proceedings CVPR'88: The Computer Society Conference on Computer Vision and Pattern Recognition. 1988. 556–561.

[10] Kontschieder, Peter, Donoser, Michael, and Bischof, Horst. Beyond pairwise shape similarity analysis[C]. In: Asian conference on computer vision. 2009. 655–666.

[11] Bai, Xiang, Yang, Xingwei, Latecki, Longin Jan, et al. Learning context-sensitive shape similarity by graph transduction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(5):861–874.

[12] 卢兴敬. 基于内容的服装图像检索技术研究及实现 [D]: [PhD thesis]. 哈尔滨: 哈尔滨工业大学, 2008.

[13] Rumelhart, David E, Hinton, Geoffrey E, Williams, Ronald J, et al. Learning representations by back-propagating errors[J]. Cognitive modeling, 1988, 5(3):1.

[14] LeCun, Yann, Bottou, Léon, Bengio, Yoshua, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278–2324.

[15] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks[C]. In: Advances in neural information processing systems. 2012. 1097–1105.

[16] Deng, Jia, Dong, Wei, Socher, Richard, et al. Imagenet: A large-scale hierarchical image database[C]. In: 2009 IEEE conference on computer vision and pattern recognition. 2009. 248–255.

[17] Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[18] Szegedy, Christian, Liu, Wei, Jia, Yangqing, et al. Going deeper with convolutions[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. 1–9.

[19] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, et al. Deep residual learning for image recognition[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. 770–778.

[20] Xie, Saining, Girshick, Ross, Dollár, Piotr, et al. Aggregated residual transformations for deep neural networks[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. 1492–1500.

[21] Hu, Jie, Shen, Li, and Sun, Gang. Squeeze-and-excitation networks[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. 7132–7141.

[22] Ioffe, Sergey and Szegedy, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.

[23] Viola, Paul, Jones, Michael, et al. Rapid object detection using a boosted cascade of simple features[J]. CVPR (1), 2001, 1:511–518.

[24] Viola, Paul and Jones, Michael J. Robust real-time face detection[J]. International journal of computer vision, 2004, 57(2):137–154.

[25] Dalal, Navneet and Triggs, Bill. Histograms of oriented gradients for human detection[C]. In: international Conference on computer vision & Pattern Recognition (CVPR'05). 2005. 886–893.

[26] Felzenszwalb, Pedro F, McAllester, David A, Ramanan, Deva, et al. A discriminatively trained, multiscale, deformable part model.[C]. In: Cvpr. 2008. 7.

[27] Felzenszwalb, Pedro F, Girshick, Ross B, and McAllester, David. Cascade object detection with deformable part models[C]. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2010. 2241–2248.

[28] Felzenszwalb, Pedro F, Girshick, Ross B, McAllester, David, et al. Object detection with discriminatively trained part-based models[J]. IEEE transactions on pattern analysis and machine intelligence, 2010, 32(9):1627–1645.

[29] Girshick, Ross, Donahue, Jeff, Darrell, Trevor, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. 580–587.

[30] Girshick, Ross. Fast r-cnn[C]. In: Proceedings of the IEEE international conference on computer vision. 2015. 1440–1448.

[31] Ren, Shaoqing, He, Kaiming, Girshick, Ross, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]. In: Advances in neural information processing systems. 2015. 91–99.

[32] Hadsell, Raia, Chopra, Sumit, and LeCun, Yann. Dimensionality reduction by learning an invariant mapping[C]. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). 2006. 1735–1742.

[33] Schroff, Florian, Kalenichenko, Dmitry, and Philbin, James. Facenet: A unified embedding for face recognition and clustering[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. 815–823.

[34] Varior, Rahul Rama, Shuai, Bing, Lu, Jiwen, et al. A siamese long short-term memory architecture for human re-identification[C]. In: European conference on computer vision. 2016. 135–153.

[35] Li, Wei, Zhao, Rui, Xiao, Tong, et al. Deepreid: Deep filter pairing neural network for person re-identification[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014. 152–159.

[36] Sun, Ming, Yuan, Yuchen, Zhou, Feng, et al. Multi-attention multi-class constraint for fine-grained image recognition[C]. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018. 805–821.

[37] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, et al. Attention is all you need[C]. In: Advances in neural information processing systems. 2017. 5998–6008.

[38] Hu, Han, Gu, Jiayuan, Zhang, Zheng, et al. Relation networks for object detection[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. 3588–3597.

[39] Fang, Zhiwei, Liu, Jing, Wang, Yuhang, et al. Object-aware deep network for commodity image retrieval[C]. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. 2016. 405–408.

[40] Almazan, Jon, Gajic, Bojana, Murray, Naila, et al. Re-ID done right: towards good practices for person re-identification[J]. arXiv preprint arXiv:1801.05339, 2018.

[41] Wang, Yaming, Morariu, Vlad I, and Davis, Larry S. Learning a discriminative filter bank within a CNN for fine-grained recognition[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. 4148–4157.

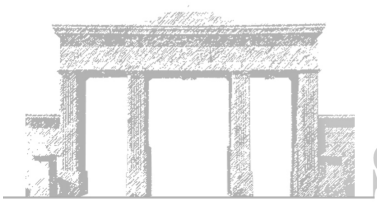
[42] Sun, Yifan, Zheng, Liang, Yang, Yi, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018. 480–496.

[43] Wei, Shih-En, Ramakrishna, Varun, Kanade, Takeo, et al. Convolutional pose machines[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. 4724–4732.

[44] Wei, Longhui, Zhang, Shiliang, Yao, Hantao, et al. Glad: Global-local-alignment descriptor for pedestrian retrieval[C]. In: Proceedings of the 25th ACM international conference on Multimedia. 2017. 420–428.

[45] Li, Dangwei, Chen, Xiaotang, Zhang, Zhang, et al. Learning deep context-aware features over body and latent parts for person re-identification[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. 384–393.

心求至善



SOUTHEAST UNIVERSITY