# 華中科技大學

## 研究生（650.801 文献阅读与选题报告）报告

**题 目：基于深度学习和上下文语义的视觉内容识别与分析研究**

学　　号　D201377750

姓　　名　欧 新 宇

专　　业　计算机应用技术

指 导 教 师　凌 贺 飞

院（系、所）　计算机科学与技术学院

华中科技大学研究生院制

# 填 表 注 意 事 项

一、本表适用于攻读硕士学位研究生选题报告、学术报告，攻读博士学位研究生文献综述、选题报告、论文中期进展报告、学术报告等。

二、以上各报告内容及要求由相关院（系、所）做具体要求。

三、以上各报告均须存入研究生个人学籍档案。

四、本表填写要求文句通顺、内容明确、字迹工整。

# 目 录

# 1. 绪　论

## 1.1. 课题来源

本课题的研究主要来源于以下科研项目：

1. 国家自然科学基金重点项目：网络大数据环境下的多媒体敏感内容感知、识别、检索与分析研究（U1536203）
2. 国家自然科学基金：人像图片的语义理解方法研究（61572493）
3. 湖北省是自然科学基金创新项目：基于云计算的监控视频大数据智能分析与检索关键技术研发及应用（2015AAA013）
4. 国家自然科学基金：面向社交网络的数字指纹技术研究（61272409）

## 1.2. 研究背景和意义

21 世纪是数据信息时代，移动互联网、社交网络、电子商务、云计算、物联网等技术大大拓展了互联网的疆界和应用领域，由此而产生的各类数据呈爆炸式增长。在各类数据中，图像视频由于其直观性的特点，一直在人类社会生活中占据着重要的地位，是人类获取信息最主要的途径之一，在全球图像视频数据爆炸式增长的今天，图像视频已经成为当今互联网无处不在的资源，在互联网中每分钟都有无数的图像被相互分享。曾经主导各大网站的文本资源，目前也逐渐转变为丰富的图像和视频资源，在我国，爱奇艺、腾讯视频、优酷土豆、QQ 空间、微信朋友圈等互联网应用的数据量已经占据全网 90%以上的数据量。图像视频大数据的分析与处理成为保障国家和公共安全的战略高技术和电子信息产业新的增长点，具有很大的发展潜力和广阔的应用前景。同时，它使我们获取的资源更加丰富，形式更加多样化，极大地丰富了人民群众的文化生活，为人民群众参与文化建设提供了新的渠道。但是，由于图像和视频大数据本身的特性，在处理和分析它们时依然有很多困难和挑战。主要体现在以下几个方面：

第一，**效率。**海量的数据对于模型的性能和效率都具有更高的要求，特别是在当前移动互联网和移动终端快速发展的环境中，如何保证在能够处理大量数据的同时，大幅降低数据的处理时间是不可回避的问题。

第二，**可用性。**面对海量的数据，对特定用户有价值的数据通常比较少，即数据价值密度比较低，这就要求模型具有较强的特征提取能力和过滤筛选能力，能够从海量的数据中发掘出具有价值的内容和高层语义信息。

第三，**多样性。**大数据环境下样本通常都具有较明显的多样性，如何合理地处理多样化的数据是提高整个系统性能的一大难题。

第四，**有用性。**面对充斥于互联网的各种资源和数据，如何高效地发现非法

和不良信息，净化网络空间是促进社会稳定与和谐发展的急迫性和基础性问题。

随着近年来深度学习的巨大成功，多种基于深度学习的视觉内容识别与分析方法也快速地发展起来，这些方法为我们解决上述问题提供了有效途径，也为我们更好地使用互联网中的图像大数据提供可能。这些技术包括：图像分类、目标检测、场景解析和基于内容的图像检索等。

从另一个角度来看，利用视觉目标对象的局部信息、邻域信息、对象与对象间的交互信息以及目标所处的场景信息等各种类型的上下文信息，能极大地丰富目标本身的信息表达，有效地改进以对象或对象语义为中心任务的性能。这几个结论在近年来若干重大的国际竞赛中被证明。例如：利用局部上下文信息的部件可变性的部件模型[1]（Deformation Part Model，DPM）获得了 Pascal VOC 2011 竞赛的第一名，和利用多全局上下文融合的 Overfeat[2]、VGG-Net[3]、GoogLeNet[4] 等模型在 ILSVRC 竞赛上也都名列前茅。

**本课题拟在深度学习的框架下，结合层次化语义关系、全局与局部语义关系等多种上下文关系，针对深度学习模型在图像视觉内容识别、场景解析与图像检索上的不足展开研究工作。**论文的研究具有三大优势：（1）基于深度学习的特征提取架构，不仅回避了传统方法特征选择的困难，同时能够获得更鲁棒的特征和高层语义信息；（2）利用层次化语义的差异性，一方面可以充分利用不同视角特征的互补性，另一方面也大大提高了算法在整个数据集上的执行效率；（3）通过整合多种类型的上下文语义信息，充分挖掘了样本的内在属性，不仅提高了算法对样本内数据的性能，同时也大大提高了算法的泛化性能。

本课题的研究非常具有挑战性的，研究内容涉及到计算机视觉、多媒体处理、机器学习、深度学习、优化理论、并行计算等理论与方法。论文的实现不仅丰富了计算机视觉和深度学习相关理论和技术，而且对相关领域的学科发展也起到促进作用。更重要的是，对其开展研究，不仅能够推动我国互联网多媒体应用的进一步发展，造福大众，更是保障国家互联网安全，进化网络空间的有力技术措施。

## 1.3. 国内外研究现状

视觉内容的语义理解是视觉内容识别和分析的基础，它涉及到机器学习、深度学习、计算机视觉以及认知心理学等多个学科领域，是一个非常重要的研究领域。下面我们将从基于上下文语义的视觉内容分析和基于深度学习的视觉内容分析两条线来进行文献综述。

### 1.3.1. 基于上下文语义的视觉内容识别

(1) 视觉语义理解概述

我们使用文献计量法对基于图像的视觉语义的研究现状进行了统计分析。在

图 1-1 中，我们以“图像”和“语义”作为关键字在 Web of Science (SCI 检索)、Engineering Village (EI 检索)、CNKI（期刊+会议）、CNKI（硕士博士论文）四个数据库的计算机及相关领域中，检索了 2000 年以来关于图像语义研究的国内外文献。

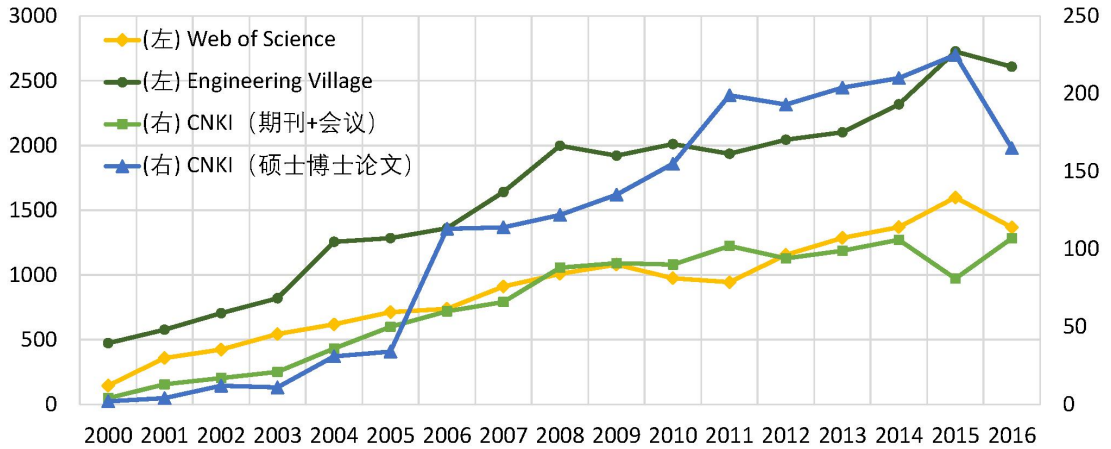


图 1-1 2000 年以来与“图像+语义”相关的文件统计图

从图像语义理解的研究发展进程来看，2003 年以前相关文献相对较少，这主要是因为国内外语义理解，甚至是图像处理的研究仍然处于初级阶段。随后的几年，文献数量逐年上涨，可见基于语义的分析方法对于理解图像变得越来越重要。有趣的是到 2016 年，各大数据库的文献数量有一定的下降，据我们所知，主要的原因可能是因为基于监督的方法已经暂时达到一个瓶颈，研究的主力开始向无监督、半监督、生成对抗网络 GAN 和增强学习方面转变。但可以预见的是，当这些学习方法有一定的稳定度之后，基于语义的研究依然会迎来新的春天。

当前，在语义理解的研究领域中，国外的科研机构在图像语义研究中取得大量研究成果的机构主要有：新加坡国立大学、南洋理工大学、微软研究院、卡内基梅隆大学、加州大学、IBM 华盛顿研究中心等。而国内的研究者在图像语义研究方面也占举足轻重的地位，虽然中文文献远少于英文文献，但从机构分布看，无论是中文还是英文文献几个主要的研究机构依然名列前茅，它们是：微软亚洲研究院、清华大学、中国科学院自动化所、浙江大学、中国科学技术大学、香港中文大学等。

Szummer 等人[5]最早提出用基于底层特征的图像分类方法来理解图像的语义，它们将图像进行分块，然后再提取每个子块的颜色特征和纹理特征，并用 K 近邻分类器对每个图像子块进行聚类，最后利用统计学的方法实现整幅图像的分类。该方法虽然提取的是底层特征，但是它源于图像的语义识别。因为一些简单的底层视觉特征可以利用颜色和纹理将图像块分类到特定的场景，例如：深蓝的大海，绿色的草地、浅蓝的天空等。虽然该方法只能处理简单的场景分类，但是它开创

了基于语义识别图像的先河。

借助机器学习方法，计算机可以在底层的视觉特征和高层语义类别之间建立映射模型，并通过学习分类器实现图像的分类识别。在这类方法中最重要的模型是基于词袋[6]（Bag of Feature，BoF）的模型和费舍尔核[7-9]（Fisher Kernel，FK）。BoF 利用 K-Means 等聚类算法对图像局部特征（如 SIFT[10, 11]，HOG[12]和 LBP[13]）进行聚类得到视觉语义词典，并使用一定的编码方案（如：矢量量化[14]、稀疏编码[15]和高斯混合模型[16]）完成编码、量化，然后利用金字塔匹配模型（SPM）[17]、VLAD[18]等方法实现局部特征到样本全局特征的转换，最后利用直方图来表征图像。费舍尔核[19]通过组合生成方法和判别方法来将将线性不可分的样本空间映射到一个高维的线性可分的特征空间，再转换为视觉直方图来表征图像。这些方法在视觉识别任务中取得了很大的进步，但距离人类理想的性能还有很大的差距。这主要是因为这些传统方法基本都是通过人工技巧性地选取适合某个样本集的鲁棒性的特征，并结合分类器进行物体识别，这对于具有多样性的大数据来说效果并不好。

随着深度学习的发展，基于卷积神经网络[2-4, 20-23]（Convolutional Neural Network，CNN）的方法可以实现端到端（End-to-End）地从原始像素到高层语义的转换，这不仅简化了识别过程，更重要的是它避免了特征选择过程中的语义损失。

(2) 基于上下文的特征表达

从视觉内容的特征表达的角度来看，上下文特征主要包括全局上下文特征和局部上下文特征。全局上下文特征是指包含图像整个场景的特征。例如，基于整个图像统计信息来描述场景的 Gist 特征[24]，通过级联局部特征来构建全局特征，并用来描述图像整体的纹理信息。局部上下文的特征主要可以分为星座模型[25, 26]（Costellation Model）、视觉词袋模型[27-29]（Bag of Feature，BoW）、空间金字塔模型[17, 30, 31]（Spatial Pyramid）和部件模型[1, 32, 33]（Part Model）。

星座模型[25, 26]利用目标区域和邻域部件的相对尺度、相对位置和外观信息构建几何关系来描述目标。它通常更注重视觉区域间的相互关系，并通过不同的局部区域来构建基于交互关系的上下文信息。在星座模型中，目标部件通常会被限制在兴趣点所决定的稀疏的位置集合中，并通过高斯分布来描述部件的几何分布。

视觉词袋模型[27-29]将图像描述成视觉单词的无序集和，它忽略了图像的全局结构和不同图像块之间的空间约束，这使图像的表达有很大的限制性。为了解决这个问题，对象空间关系[34-36]和全局场景[37]通常被引入用来提供局部上下文联系和全局上下文联系。

空间金字塔模型[17, 30, 31]是一个层次化的上下文模型，它在不同分辨率下将图像划分为多个子区域，并将这些子区域进行局部连接，构建视觉直方图表征图像。

层次化的上下文融合隐式地引入了空间信息，比单纯的 BoF 模型具有较大的优势。金字塔 Hog[30]（Pyramid HoG）通过将不同分辨率的 HoG 特征进行级联和归一化获得具有空间属性的金字塔 Hog 特征。

部件模型[1, 32, 33]描述的通常是一些具有明确语义的局部区域（如人的头，汽车的轮子等），它不但需要描述这些区域的特征，也需要描述这些区域间的拓扑关系。可变性的部件模型[1]（Deformable Parts Models，DPM）通过一个弹簧模型来融合多个局部上下文关系，使模型具有较好的抗形变和遮挡的特性。Zhang 等[32]提出了基于部件的 R-CNN[38]，利用 CNN 强制学习整个对象和部件之间的几何约束，来提高整个对象的特征性能，这种方法较好地保持了局部部件之间的空间关系，对于同一对象的不同姿态有较好的不变性。

(3) 多上下文建模分析方法

最近几年，很多研究者[39-45]都考虑采用多种上下文来改进不同任务的性能。基于上下文的分类方法，常见的多上下文建模方法包括：全局-全局上下文建模、局部-局部上下文建模和全局-局部上下文建模。

全局-全局上下建模是一种最通用的方法。Ciresan 等人[42]提出了一种多栏 CNN（Multi-column DCNN）用于图像分类，在这个算法中，输入图像被采用多种不同的策略进行预处理，然后被分别送入多个独立的卷积神经网络中单独进行训练。最后的预测结果，通过平均这些独立的预测而得。由于不同的数据其统计属性和物理解释总是具有多样性的特点，单视角可能无法很好地获得一致的判决信息，Yu 等人[46]提出一种基于高阶距离的多视图随机学习方法用于从不同的视角来学习特征。深度多模距离度量学习方法是由 Yu 等人[47]提出的另外一种解决多样化样本的方法，该方法通过使用多个模型共同学习一个样本，从而有效地降低样本间的语义鸿沟。另一个方面，大多数参加 *Imagenet* 大规模视觉识别挑战赛[48]（Imagenet Large Scale Visual Recognition Chellenge，ILSVRC）的参赛队伍[39, 43, 45, 49]也都使用加权平均融合不同模型来实现性能的提升。事实上，从最近两年 ILSVRC 的比赛结果[50, 51]来看，几乎所有名列前茅的算法都是采用了多模型融合的算法。

局部-局部上下文建模通常用来衡量对象间的相互关系或部件间的几何关系。Zhang 等人[52]提出了一种对偶卷积网络，该网络通过组合候选区域建议网络和定位网络来同时生成对象位置和对象类别信息，该方法生成的局部特征的精度和效率都很不错。Fergus 等人[53]提出一种融合形状、外观、相对尺寸的概率表达来构建类似星座图谱的对象关系图，基于选定的中心对象，其他相关区域由基于熵的特征选择器获得。这种方法通过局部区域间的约束关系实现非监督尺度不变的目标识别。视觉词袋模型[27-29]也是典型局部-局部上下文建模方法，它通过组合不同的局部区域为视觉词典，并利用统计方法来衡量整个样本。基于部件的 R-CNN[38]

使用几何约束将对象的部件组合在一起来表征完整的对象，它较好地保持了对象的局部上下文的不变性，对遮挡和视角有较强的鲁棒性。DPM[1]通过一个弹簧模型来融合多个局部上下文关系，使得模型具有较好的抗形变和遮挡的特性。此外，Liu 等人[54-56]为了利用时序信息，使用视频不同帧作为不同的全局上下文进行融合。

全局-局部上下文建模旨在充分利用全局上下文和局部上下文的互补性。全局上下文建模致力于生成鲁棒的全局特征，局部上下文建模则被设计用来发现细节的信息。Zhao 等人[44]使用了两条独立的 CNN 支路，分别用来训练全局上下文和局部上下文，最后将两条支路的全连接层串联在一起，用于共同生成显著性图。Karpathy 等人[41]在视频分类任务中，将输入帧分为两种上下文流，一个支路用于产生低分辨率特征，另外一个支路用于生成高分辨率特征，两条支路最后也是通过一个全连接层来进行串联，并输出最终的预测。

通常情况下，标准的分类任务都比较关注于从整幅图片去学习鲁棒的特征，这些方法可以被认为更加关心的是全局上下文信息；而检测任务通常是处理一个区域的特征，它可以被认为是更关注局部上下文信息。然而，常见的多上下文建模的方法，无论是哪一种上下文融合的方法，通常最后都是采取多个特征直接融合的方式来加强整体特征的鲁棒性。这些方法虽然可以有效地利用不同特征的独特性来提高性能，但是我们认为这并不是最好的解决方法。

(4) 层次化分析方法

层次化对于数据理解和数据管理具有举足轻重的作用。*WordNet*[57]是基于层次化语义结构最重要的成果之一，它由语言学和自然语言处理社区发起并完成。现有的很多重要的数据集都按照 *WordNet*[57]的层次化语义进行组织，例如：大规模图像数据集 *Imagenet*[48]和 *TinyImage*[58]。很多工作表明[58, 59]层次化结构对于分类任务的精度有积极的影响，并且有文献证明利用层次语义关系来改进分类任务的性能时只需要更少的训练样本即可实现较好的评估结果[60]。层次化技术也被应用到其他的视觉任务中[61, 62]。在真实世界中，敏感的隐私类别很容易湮没在庞大的类别空间，Yu 等人[63]通过集成深度卷积神经网络的特征表达和一种判决树的分类方法，提出了一种多任务的学习算法用于识别这些敏感对象。

空间金字塔策略也是一种有效的层次化方法。Ivan 提出基于多分辨率的 HoG 特征提取方法 PHoG[30]，他在不同的分辨率下将图像划分为多个子区域，并将这些子区域进行局部连接来构建视觉直方图。这种方法隐式地引入了空间信息，有效地增加了特征的强度。He 等[64]提出了基于空间金字塔池化的深度神经网络 SPPNet，这种方法的核心是利用空间金字塔去替代卷积网络的最后一个卷积层。该方法不仅融合了多个尺度的特征，更使网络可以实现任意尺寸的输入。

粗到细策略是一种典型的层次化方法，它被广泛使用在了多种计算机视觉的

任务[65-68]中。Eigen 等人[67]首先使用 FCN[67]生成像素级的语义分割结果，然后利用粗到细的策略，将已生成的结果当做一种粗分割送入到另一个全新的 FCN 网络中用于生成更细粒度的像素级预测。Ling 等人[66]利用粗到细策略实现精确的图像检索。他们首先将与待查询样本具有相似高级语义的样本收集在一起组合成候选集，然后再使用深度中级特征表达进行过滤。

### 1.3.2. 基于深度学习的视觉内容识别

图像分类、目标检测和场景解析是图像识别的三个核心问题，也可以被认为是图像识别的三个不同粒度的任务。图像分类关注的是如何对整个图像进行语义类别判定；目标检测则定位图像中特定物体出现的区域并判定其语义；场景识别处理的是像素级的分类问题，它为每个像素都指定一个语义标签。三项技术在信息检索、广告投放、用户分析、商品推荐等互联网应用中都大有用武之地。此外，基于内容的图像检索是大数据互联网时代搜索引擎发展的必然产物，它可以为用户提供个性化的资源服务，这种技术的实现通常以不同粒度的识别任务为基础，并且支持用户能够以多模态、多属性的形式搜索不同类型的媒体数据。随着深度学习的可快速发展，图像分类、目标检测、场景解析和基于内容的图像检索也得到了快速发展，下面简要回顾这些技术的传统方法，并着重从深度学习的角度进行综述。

(1) 图像分类

传统图像分类算法中具有代表性的是 Yang[70]等人在 2009 年提出的采用稀疏编码技术表征图像，并用支持向量机[71]（Support Vector Machine，SVM）进行图像分类的方法。另一类具有代表性的识别框架是基于词袋[27-29]的模型。它利用人工进行特征提取（如：SIFT[10, 11]，HOG[12]和 LBP[13]），并使用一定的编码方案（如：矢量量化[14]、稀疏编码[15]和高斯混合模型[16]）完成编码，最后用金字塔匹配模型（SPM）[17]、VLAD[18]等方法构建视觉直方图。虽然稀疏编码和词袋模型在视觉识别任务中取得了很大的进步，但是距离人类理想的性能还有很大的差距。因为这些传统方法都是通过人工技巧性地选取适合某个样本集的鲁棒性的特征，并结合分类器进行物体识别，这对于具有多样性的大数据来说效果并不好。

图像分类领域根本性的变革来源于 2012 年的 ILSVRC[72]挑战赛，Alex Krizhevsky 等人[22]将 *Imagenet* 数据集[73]Top5 分类识别错误率从过去的 25%降低到 15%，引起了人们对深度学习的广泛关注。随后，以卷积神经网络为代表的各种深度学习算法被广泛应用于图像识别中，并不断刷新记录。截至 2015 年，*Imagenet* 图像 Top5 分类的识别错误率已经降低到 3.1%[23]，超越了人的识别能力 5.1%。同时，在其他一些数据库上，卷积神经网络也展现了其强大的识别性能，在很多视觉识别任务中均已超过了人类的识别能力，包括交通信号识别[74, 75]，人

脸识别[76-78]，自然图像分类[20, 23, 79]和手写字体识别[74, 80]等。

卷积神经网络在视觉识别领域获得如此巨大的性能改进，主要归功于两个方面的巨大进步：一是构建了更加强大的模型，二是设计了更有效的策略来抵抗过拟合问题。一方面，神经网络越来越能够更好地拟合训练数据，这主要是因为网络复杂性的增加（例如：深度的增加[3, 4, 20]，宽度的增大[4, 81, 82]和使用更小的步长[2-4, 79, 82]），新的线性激活单元[81, 83-87]的使用和复杂层的设计[4, 23, 64]。另一方面，有效的正则化技术[9, 80, 84, 88]，积极的数据扩展技术[3, 4, 22, 81]和大规模的已标记的数据集[73, 89, 90]实现了神经网络模型更好的泛化能力。

(2) 目标检测

大多数目标检测系统都包含两个重要的组件：特征提取器和分类器。传统的对象检测方法，特征抽取器通常是基于一些手动特征建模，例如 HOG[12]特征和 SIFT[10, 11]特征。分类器通常是使用一个线性支持向量机（SVM，Support Vector Machine）、一个非线性 boosted 分类器[91]或者一个带核的 SVM[92]。更复杂有效的检测算法，如可变性的部件模型[1]（Deformable Parts Models，DPM）或者一些非线性多核方法[93]也取得了较好的成绩。

进两年来，目标检测领域获得了巨大的进展，这主要是由于深度学习，特别是卷积神经网络[2-4, 20-23]模型的快速发展。目标检测的精度瓶颈也由识别精度转变成目标定位精度，良好的定位精度可以有效地改善目标检测的性能。候选建议区域生成算法中，比较有代表性的算法包括：Selective Search[92]、BING[94]、Objectness[95]、MCG[96]、EdgeBoxes[97]、DeepBox[98]等。

在基于 CNN 的检测系统中，最重要的三个工作分别是基于建议框的检测方法 Overfeat[2]、R-CNN[38, 64, 99, 100]框架和基于回归的方法 SSD[101]框架。

Overfeat[2]设计了两个 CNN 模型，并将它们以滑动窗口的模式在一幅图像上以不同尺度进行密集地扫描，一个利用 *Softmax* 分类器对区域进行分类，另一个通过回归预测对象的边界框。这些密集的分类和定位预测通过贪婪合并算法以投票机制生成一个对象检测的集合，作为最终的输出。

R-CNN[38]是一个非常成功的检测算法，它首先利用 Alex Krizhevsky 所设计的 CNN 模型 AlexNet[22]预训练了一个基于分类任务的卷积神经网络，然后使用 Selective Search[92]算法生成带定位信息的候选建议区域，并利用这些候选建议区域对预训练好的卷积神经网络进行微调训练，得到最终的检测网络。在进行检测的时候，整个过程是一个端到端的过程，通过检测网络提取的特征，最终利用若干个特定类的线性支持向量机实现基于类别的识别。尽管 R-CNN 获得了很好的识别性能，但它也面临识别时间过长的困境。对于一副彩色图片，在 GPU 的帮助下通常至少需要花费 10-20s 的时间来进行区域建议和特征提取（而在 CPU 的环境中更是需要长达 60s 以上的时间）。He 等人[64]提出了基于空间金字塔池化的

深度神经网络（SPPNet，Spatial Pyramid Pooling-net），该网络通过区域映射来实现卷积特征的共享，重复使用卷积特征图大大提高了推理阶段的运算速度。Fast RCNN[99]是由 Girshick 提出的快速版的 R-CNN，通过引入 RoI 池化层，Fast RCNN 实现了完全端到端的运行机制。不但允许利用 CNN 同时输出分类和定位信息，还可以不依赖额外的磁盘空间来用于中间特征的存储。Fast RCNN 实现了比 SPPNet 更高的精度，同时在训练的时候提速 3 倍，测试的时候提速 10 倍。Faster RCNN[100]将建议框生成的步骤集成到了整个网络中，称为区域建议网络（Region Proposal Network，RPN），不但彻底抛弃了额外的建议框生成过程，而且再一次使系统的速度和精度都得到了提升。

SSD[101]在 YOLO[102]的基础上发展而来，它结合了 YOLO 中的回归思想和 Faster R-CNN 中的 anchor 机制，使用全图各个位置的多尺度区域特征进行回归，既保持了 YOLO 速度快的特性，也保证了窗口的预测跟 Faster R-CNN 一样精准。SSD 在 VOC2007 上 mAP 可以达到 72.1%，速度在 GPU 上达到 58 帧每秒。

RCNN 系列和 SSD 系列给我们提供了优秀的目标检测底层框架，除此以外，还有一系列的技巧被提出来改进目标检测的性能。（1）难样本挖掘[103]（Online Hard Example Mining，OHEM）。它通过反向传播损失最大的一些样本的误差替代所有样本，不但实现正负样本的平衡，还加速了训练过程。（2）多层特征融合。RCNN 系列利用的都是最后一层卷积层的特征来进行目标检测，但是高层特征由于多次池化操作，已经丢失了不少细节信息，会产生定位不准的问题。HyperNet[104]等一些方法通过整合多个卷积特征层，不但利用了高层特征的也语义信息，还考虑了底层特征纹理特征，使得目标定位的更加准确。（3）上下文信息。除了从建议区域内提取特征，利用上下文信息[105, 106]对于提高检测框的类别信息的判断也非常有意义。

(3) 语义分割

与图像分类类似，大多数成功的语义分割（Sementic Segmentation）系统都依赖于手工特征和一个简单的分类器，例如：推进分类器[107, 108]（Bootsing），随机森林[43]（Random Forests）或支持向量机[44]（Support Vector Machines，SVM）。受益于集成丰富的上下文信息[109]和结构化预测技术[110, 111]，传统的分割网络性能有了长足的进步。然而，这些系统的性能依然受限于传统手工特征的表达能力。

过去几年在图像分类领域取得巨大成功的深度学习技术也被快速迁移到了语义分割任务中。由于语义分割，同时涉及分割和分类，因此，一个核心的问题是如何组合这两种任务。为了处理这个问题，三种基于深度神经网络的方法体系被提出。

第一个流派采用一个级联的自低向上的图像分割算法生成建议区域，然后再使用深度卷积神经网络对区域进行识别。如将边界框算法 Selective Search[92]或遮

罩算法 MCG[96]生成的候选区域引入到分割网络的 RCNN[38]和 SDS[112]。类似的，Mostajabi 等人[113]依赖超分辨率表达来生成区域建议。

第二个流派也是在分割区域中实现局部对象的识别，但与上一个流派使用额外预处理方法生成区域不同，第二流派个直接利用卷积特征图来生成区域。Farabet 等人[114]使用卷积神经网络生成多种图像分辨率。Hariharan 等人[115]使用忽略层将输入和中级特征级联起来，用于生成像素级分类。Dai 等人[116]提出使用区域建议来池化中级特征图。虽然利用卷积网络生成区域建议改进了性能，但仍然是基于分割算法来生成区域，再进行分类。然而，生成建议区域和分割可能是不可靠的。

第三个流派放弃了区域预分割和融合的步骤，直接使用深度卷积神经网络生成具有类别信息的像素级预测。最重要的工作是 Shelhamer 等人提出的全卷积网络[69]（Fully Convolutional Network，FCN），它使用升采样的方式处理每一个中间层的卷积特征图，然后将这些升采样后的卷积特征图组合起来生成包含多尺度信息卷积特征图，然后再进行全画幅的像素级预测。DeepLab[117]将多尺度池化技术融入到全卷积网络中，并在网络的顶端用密集连接的条件随机场[118]（Conditional Random Filed，CRF）来优化对象的边缘，从而生成细腻的像素级分割。随着 DeepLab[117]的公开，语义分割领域受到了极大的推动。很多研究组都取得了巨大的进步，特别是在 Pascal VOC 2012 语义分割竞赛上，很多排名前列的算法[119-126]都或多或少地基于 DeepLab 完成自己的算法。特别是 Deeplab 提出的 Atrous 卷积和全连接 CRF 几乎成为图像分割的标准配置。

随着基于对象的图像分割的发展，场景解析[117, 127-130]和人脸解析[131-134]也成为图像分割研究领域的重要目标。

场景解析[117, 127-130]是理解场景的基础，它可以被应用到如自动驾驶、机器人导航等重要领域，同时它也可以为常规的目标识别、对象检测任务的提供大量辅助信息。与语义分割不同的是，场景解析不仅仅要识别和分割出场景中的对象或者显著性物体，同时也要识别出所有的背景元素。也就是说，场景中的每一个像素，都是场景解析所关心的内容。

人脸解析[135-138]可以认为是人脸识别[76-78, 139, 140]之后一个更高级的应用，它也是人体解析[141-143]的一个分支领域。一方面它可以为传统的人脸识别提供更强大助力，如利用局部上下文信息处理遮挡和视角变换问题。另一方面，人脸解析也扩展出很多实用的应用系统，如自动化妆系统，美颜软件以及现今各种直播系统中的人脸插件。更重要的是，对人脸的解析、分析和应用，可以为现今各种安全和安防系统提供极大的助力，一方面方便人民生活，另一方面也为可以为社会安全做贡献。

(4) 基于内容的图像检索

目前，比较主流的图像检索技术，包括传统基于词袋模型的图像检索方法、基于哈希的图像检索方法和基于深度学习的图像检索等。

传统图像检索方法依赖于手工特征，例如编码成词袋直方图（Bag of Words，BOW）的 SIFT 描述子[10, 11]、GIST 描述符[24]和费舍尔向量[19]（Fisher Vector，FV）。

哈希编码通过简短的二进制编码来缩小特征表达的维度，并通过哈希表查询来加速搜索，从而大幅提高查询效率。根据哈希生成过程中是否利用数据的特性，可以将哈希方法分为数据独立哈希和数据感知哈希。数据独立哈希方法中最著名的是局部敏感哈希[144]（Locality-Sensitive Hashing，LSH），该算法通过随机映射的方式将对象从特征空间映射成二进制码字。类似的，最小哈希[145]（min-Hash）采用随机序列的方式进行编码，通过大量哈希表逼近搜索条目之间的杰卡德相似系数。数据感知哈希利用机器学习工具对数据样本进行学习，从而自动地得到高效、紧凑的编码，例如：谱哈希[146]（Spectral Hashing，SH）、基于熵编码的相似性保护算法[147]（SPEC Hashing）、二进制重构嵌入[148]（Binary Reconstructive Embeddings，BRE）算法、自学习哈希[149]（Self-Taught Hashing，STH）等。

基于哈希的方法有效地加速了检索的速度，但是它对样本的语义保持却显得无能为力。CNN 在图像分类[2-4, 20-23]中获得了巨大的改进，作为一种通用的图像表达，CNN 特征较好地保持了高层语义信息，同时，它也能够被应用到检索任务中，并获得良好的性能。Gong 等人提出了多尺度无序池化[150]（Multi-scale Orderless Pooling，MOP）方法，将高层的 CNN 特征与 VLAD 进行融合，这些高层激活特征都是通过多尺度的滑动窗口机制从 CNN 中抽取获得，实验表明这些特征实现了较好的检索结果。神经编码[151]（Nerual Codes）通过在一组与查询图像相似的额外的地标数据集上进行了微调训练，毫无悬念地获得了优秀的检索性能。不幸的是，收集这些相似地标的训练样本并重新训练整个 CNN 模型需要消耗大量的人力和计算资源，这使得应用这种方法具有较大的限制。Wan 等人[152]通过模型重训练和相似性学习方法全面地研究了 CNN 特征在真实世界的图像检索问题，得到了令人鼓舞的实验结果，CNN 特征可以有效地弥补低层视觉特征和高层概念之间的语义鸿沟。Ng 等人[153]的工作受 MOP[150]将 CNN 特征应用到 VLAD 的启发，从 CNN 模型的每一层的卷积特征图中都抽取一次特征，并使用 VLAD 进行编码。Ou 等人[154]提出了传导迁移深度哈希，可以对图像进行深层次特征的学习与表达，并通过近邻结构保持将特征映射为区分度强的哈希码，进行大规模图像近似搜索。

## 1.4. 当前视觉内容识别与分析存在的问题

从大数据环境的挑战和国内外研究现状可以看出，图像识别和分析目前仍然存在很多值得深入研究的问题。

**问题 1：大数据环境中样本多样性问题**

在处理大规模视觉内容识别任务时，样本的多样性问题是首先要面对的难题。这种多样性可能会来源于图像的内容、尺度、分辨率、拍摄角度和图像质量等多个因素。自然图像的内容可能会涉及从专业摄影到手机自拍，从人的行为到器官特写，从模糊图像到高分辨率图像，从小图像到大图像，从二进制灰度图到全彩图，从卡通和手绘图到相机拍摄的图像等等。这些都是我们在大数据环境下必须要考虑的问题。此外，在某些特定的识别任务中，由于任务本身的性质导致类别空间较小，可能会因为大数据环境中样本多样性问题导致分类困难。以本课题研究的成人内容识别为例。成人内容识别通常是二分类问题（即：“是成人”或“不是成人”），在大规模样本的环境中，很多同类样本的差异可能会远高于不同类样本的差异，我们称这个现象为类内距大于内间距。这种类内距大于内间距的现象会严重影响分类器的性能，即使图像的特征具有完美的表达能力，这个问题依然无法避免。

**问题 2：场景解析中难目标识别问题**

场景解析是图像分割任务的一个分支，与对象语义分割、实例分割不同的是，场景解析不仅要处理场景中的对象，还要处理场景中背景；并且场景解析所要处理的样本通常比以对象为中心的语义分割任务要复杂得多，一个场景中通常包含很多类别的样本，并且很多对象具有尺度小、交互性多（易存在遮挡、重叠、共生等现象）、隐藏性强（易湮没在周围较相近的背景像素中）等特性。我们称这些对象为难目标，对难目标的识别通常是场景解析中最困难的问题。

**问题 3：额外背景类造成的误判问题**

在检测和分割任务中，有一些区域始终很难判定它们的类别归属。很多算法都会设置一个额外的背景类[1]，在训练中收集这些负样本或边缘样本来提高训练模型的健壮度。这个策略帮助训练一个更好的模型，但是它也导致一些像素在推理阶段被错误地分配成额外背景类。这个问题对于目标检测和语义分割来说，并不是大问题，至少从视觉上看并不显著。因为它们关注的是特定的类，它们可以将其他像素都归结为“额外背景”。换句话说，非目标区域都可以识别为额外背景，包括不需要识别的小对象，以及场景中真实存的背景。相对而言，场景解析必须要处理每一个像素，并且给他们都分配一个类别。在训练中增加额外背景类后，推理阶段会使一些像素被认定为额外背景类。然而，额外背景是为了训练而手工添加的，它并不是真实存在的，这造成了这些像素的错误分类。在必须增加“额外背景”类的任务中，如何在推理阶段避免将像素分配到这个类别是一个需要考虑的问题。

**问题 4：上下文融合时语义保持困难的问题**

[1] 此处，我们将新增加用于改善性能的背景类定义为“额外背景”类，而且原始场景解析中的天空、地板、草地等真实的背景定义为“背景”类。

在基于全局上下文和多个不同区域的局部上下文融合的生成网络中，由于不同区域的差异性，同时基于这些区域来生成样本时，会产生生成不同步的问题，特别是区域的边缘会产生明显的差异性。这主要是由于不同区域尺度不同，关注的特征类型也不同，以同样的方式进行迁移就会产生明显的不同步问题。例如本课题研究的妆容迁移任务，脸型、粉底、眼影和唇彩都有各自的特性，如果在统一网络中采取同样的方式来生成，显然是不科学的。因此，如何保持这些不同的上下文区域在融合后的语义不变，是首先需要面对的困难。

**问题 5：大数据环境下搜索空间太大引起的效率降低的问题**

大数据环境下的图像搜索，首先要面对的问题是执行效率，如果这个问题得不到解决，搜索引擎将失去其意义。然而，传统的基于内容的图像检索都需要将待查询样本和数据库中所有图像进行逐对的相似性计算。随着数据规模的增长，执行时间也会线性增加。对于过去几千、甚至几万的小规模数据集，延迟问题并不明显。但是，面对数百万、甚至上千万的数据库，这显然无法接受。因此，找到一种高效的相似性计算方法来缩小搜索空间，对于海量数据的搜索问题尤为重要。

# 2. 主要研究内容、预计需达到的要求和技术指标

本课题针对大数据环境下视觉内容识别与分析对高性能和高效率的要求，在深度学习框架下，基于成人内容识别、自然场景解析、人像妆容迁移和基于内容的检索四个典型的应用，开展基于多种上下文语义的视觉内容识别和分析的研究。

## 2.1. 基于深度学习的特征学习和表达

大数据环境下图像数据纷繁复杂，大量信息都潜伏在大数据中，要对这些海量数据进行分类、识别与检索，就需要能能够从图像样本中获得较准确的特征和语义表达。一方面要求特征表达更加全面和深层次，以应对超大规模数据集的表示；另一方面要求计算复杂性很低，能够应对数据量的飞速增长。传统方法大多通过人工技巧性地选取局部不变特征描述，但随着数据规模的不断扩大，其性能将越来越差。基于深度学习的特征学习和表达是进一步实现分类、检测、语义分割和检索的基础。

(1) 基于深度学习的特征提取框架研究

构建能够承载大数据的深度学习特征提取框架，是本课题的基础部分，也是核心部分。无论是实现对样本的分类识别、语义分割，还是实现内容检索；也无论是针对样本整体的处理，还是样本局部对象的处理；一个完善的深度学习特征提取框架都是不可或缺的。幸运的是，我们可以通过设计一个统一的基准框架完成上面所有任务的特征提取工作，同时通过引入迁移学习技术，在不需要重新修

改网络结构的前提下，轻松地将该基准框架快速地迁移到相关或相似的任务中，实现快速部署。

为了实现这个目标，该框架应该具有很好的开放性和兼容性，能够适应不同类型和不同分布的图像数据，满足特征提取的鲁棒性和可辨识性的要求；此外还需要兼顾紧凑性和易于计算等特点；更重要是通过模块化的设计思路，在更新整个网络的部分组件的时候实现整体性能的提升。

(2) 基于 CNN 的特征提取和筛选融合

通过完整的特征提取框架，我们可以实现全局特征和局部特征的同时获取。全局特征主要反应的是样本的整体信息，对象的大轮廓信息，以及对象间的相互关系；局部特征反应的通常是对象的局部信息，或者特写信息。针对样本中获取的大量不同类型的特征，我们需要考虑的是如何合理利用这些特征，通过何种筛选和融合的方式实现比单一特征更强大的表达能力。

## 2.2. 基于多种上下文语义的视觉内容的分析

深度学习相对于传统学习有一个很大的优势，它既可以获取样本低层的粗糙特征和中层特征，也可以获取高层的语义信息。丰富的特征和语义信息，为我们处理视觉内容提供了极大的方便，使我们利用这些信息改进各种计算机视觉任务成为可能。本课题旨在充分利用高层语义信息，并结合不同的策略实现精度和效率的同时提升。层次化语义分析方法和多上下文语义联合决策是多上下文语义分析的两个重要方向，前者采用纵向的思路，通过逐层过滤以筛选出最符合目标的样本和信息，后者则是采用横向的思路，通过多种上下文信息的联合判别获得最终目标。

(1) 层次化分析方法

层次化分析方法目标是在一个大规模样本的任务中，将复杂的问题向简单化转变。然而，如何将问题简化需要对任务本身有较深刻的分析，换句话说，层次化方法通常比较适合于特定任务。通过对任务和样本的分析，找出数据本身的共性和差异性，将样本从一个较难的问题空间，迁移到一个较为容易的问题空间。

针对成人内容识别任务，正负样本类别空间狭窄，通常是一个“是或不是”的二分类问题，但样本复杂多样（可能包括肖像、全身图、器官特写、猫、狗、桌子等多种类别），因此子类别可能存在较大的相似性，要直接将样本分为两类是一件很困难的问题。**如何处理还类别空间和样本复杂性的冲突，是本课题需要重点研究的问题。**

针对大规模图像检索任务，效率和性能是两个最关键的因素。传统的图像检索方法需要使用待检索特征与图像库中的所有样本的特征进行一一比对，对于数百万检索任务这显得很不现实。因此，**如何利用强壮的深度特征保证精确匹配的**

**同时，降低检索时间是本课题需要重点解决的问题。**

(2) 多上下文语义联合决策

如前所述，深度卷积神经网络可以得到基于全局和基于局部两种类型的特征和语义信息。通过对样本全局上下文和局部上下文的联合可以大大改善系统性能。

对于成人内容识别任务，需要识别的正样本具有极大多样性，它涉及到人的整体形象、姿态、视角和光照的影响，也受到多人之间的不同的交互行为影响；同时，很多样本是以局部器官特写的形式出现，具有很大的特异性；此外，对象的尺度和场景的混乱和复杂多样性也是成人内容识别的一个难点。显然，采用单一的卷积神经网络无法覆盖所有的情况。针对不同类型样本，使用不用的识别器变得尤为重要。因此，**如何较好地处理样本多样性识别的问题，是本课题需要重点研究的内容。**

对于场景识别任务，相比传统语义分割和实例分割任务，难点主要有两点：（1）场景中同时充斥着复杂多样的对象需要去识别，同时检测器还需要去识别那些对象周围不同类型的复杂背景；（2）相对于以对象为中心的语义分割任务，在场景任务中，对象通常会很小，而这些小尺度的对象很容易湮没在复杂的背景中。因此，**如何同时处理好对象和背景的识别问题，如何处理小对象的问题，如何处理对象之间、对象与背景间边缘的清晰性，都是我们在场景任务中需要重点考虑的问题**。

在人脸妆容迁移任务中，有三个较难点需要处理：（1）如何得到关于人脸部件的准确解析结果，用于进行特定妆容的迁移。（2）如何提取这些区域的特征，以及提取什么样的特征用于迁移。（3）在迁移的过程中，如何既保持原始人像的全局上下文特性（即：脸型，五官轮廓等），又能很自然地将参考妆容局部特征（如：眼影和唇彩）迁移到待化妆人脸上。因此，**精确的人脸部件解析和自然的上下文特征融合是本课题研究的重点。**

# 3. 课题研究的技术关键和技术方案

## 3.1. 技术关键

本课题主要有以下几个关键技术需要实现：

### 3.1.1. 大数据环境下利用深度学习实现视觉内容语义的特征表达

大数据环境下图像数据纷繁复杂，大量信息都潜伏在大数据中，要对这些海量数据进行分类、识别与检索，就需要能能够从图像样本中获得较准确的特征和语义表达。一方面要求特征表达更加全面和深层次，以应对超大规模的数据集的表示；另一方面要求计算复杂性很低，能够应对数据量的飞速增长。传统的方法大多通过人工技巧性地选取局部不变特征描述，但随着数据规模的不断扩大，其性能将越来越差，对目标对象的识别与搜索将越来越难。

深度学习模型，特别是卷积神经网络在基于视觉的任务中展现了其强大的性能，然而现有的深度学习模型也存在一些问题，有待研究，例如：（1）现有的深度学习模型提取的特征大多是基于整个画面的，属于全局特征，在表达局部对象方面存在不足，这与人类视觉感知系统侧重于目标对象有所不同，因此如何提取基于 CNN 的局部对象特征表达成为一个新问题；（2）提取多个局部 CNN 特征势必带来巨大的计算消耗，而大数据环境要求算法要具有较高的运行效率，如何兼顾高效的计算效率和强辨识力的深度表达能力，成为又一难题；（3）要实现检测和检索，现有的检测算法通常会产生很多建议框，这对于基于目标的分类、识别和检索效率较低，因此需要对多个目标建议框进行筛选、聚集和融合等。

以上几点都是本课题在利用深度学习技术对图像进行特征提取时需要重点研究的问题，也是实现本课题的技术关键。

### 3.1.2. 大数据环境下特定对象的多样性处理

大数据环境下特定视觉对象的多样性主要体现在：（1）对象的多样性：一个局部对象可能会存在多种表现状态，例如：尺度、位置、光照、姿态、表情、方向、仿射变换、遮挡等；（2）样本多样性：在实际应用中，数据库中的图像样本可能会同时包含多种不同的分布，例如：单一对象场景、多对象场景、自然风景场景等。传统的方法，通常对于样本数据集有较高的要求，比如：手写字体识别、交通标志识别、人脸识别等，但是对于大数据时代的应用，通常一个数据集可能会混合多种不同类型的数据，这些内容表现形式的多样化和任务的多样化，给基于内容的感知、识别与检索带来的巨大的挑战。

具体而言，要实现大数据环境下特定对象的多样性处理，就必须要考虑到目

标对象检测的精确定位、整体特征与局部特征的融合、原始数据的增强与扩展、目标检测建议框的高召回率以及建议框的筛选与融合等方面的关键技术。

### 3.1.3. 大数据环境下特定对象的高效检索

大数据环境下，要实现基于对象的检索面临很多挑战性的问题。首先，传统的浅层次的方法需要手工选择特征，这不仅存在特征选择的困难，也存在特征表达能力不足的问题；其次，对于细粒度的任务，对于模型的特征提取能力具有更高的要求，如何排除非对象区域对对象区域的影响是基于对象的检索需要着重考虑的问题；最后，特征的维度是检索系统最重要的组成部分，过高的特征维度会严重影响检索的速度。因此，在深度学习的框架下实现细粒度的特征提取，并采用尽量小的特征维度来表示一个样本，是实现大数据下特定对象高效检索的技术关键。

### 3.1.4. 大数据环境下检索系统的自适应能力和扩展性

如何提高系统的适应性和扩展能力，使训练好的模型能够被快速地部署到不同的应用而不仅仅局限于某个专门的应用中，是系统是否能被广泛使用的关键因素。通常训练一个深度学习模型需要大量带标签的数据，而在大数据环境下，由于数据种类复杂多样，带有标签的数据非常难以获得。此外，大数据环境下，数据具有来源不一致，更新快的特点，因而在多数情况下，训练源任务的数据和目标任务的数据，在类别和分布上存在较大差别。迁移学习可以在一定程度上解决目标任务标签数据不足的问题，同时可以实现目标任务模型的快速收敛。然而利用迁移学习的时候，也面临一些挑战性问题。如何保证从源任务向目标任务迁移的时候是一种正迁移而非负迁移；当迁移发生的时候，源任务的模型需要做如何的调整以确保从源任务迁移的知识能被目标任务较好地利用（网络的深度、宽度、结构是否需要进行调整?哪些层次的参数需要做调整？）；进行迁移训练的时候，需要多少针对目标任务的标注样本才能确保模型不发生过拟合和欠拟合问题，从而学到针对性较好的特征；对于一个动态更新的样本集（例如一个在线的图像库或者视频库），迁移训练是否可以实现动态地更新模型。因此，如何结合迁移学习和卷积神经网络实现基于对象的特征提取，并能够快速且有针对性地将特征从源任务迁移到目标任务，是实现检索系统扩展性需要解决的技术关键。

## 3.2. 技术方案

### 3.2.1. 多级多尺度图像表达

CNN 已经被证明它有很强大的能力通过训练不同尺度的样本来隐式地表达不同尺度特征。此外，显示地考虑多尺度也被证明可以同时提高大尺度和小尺度

对象的识别性能。图 3-1 展示了我们多尺度特征提取网络的共享卷积层部分。

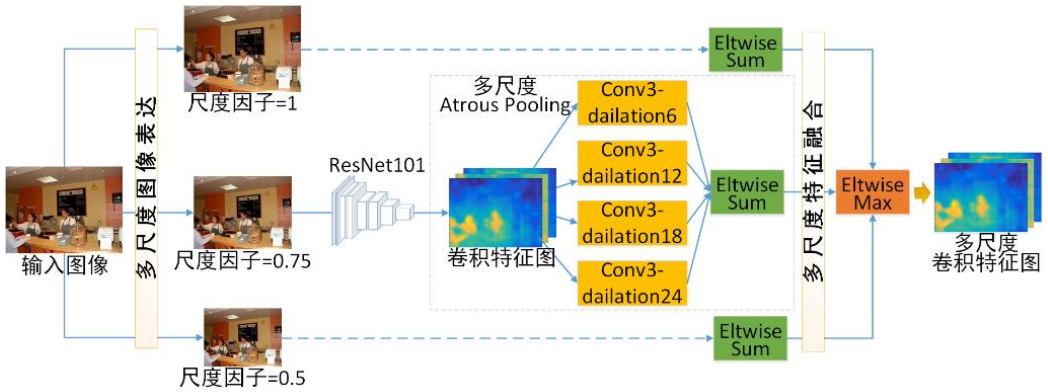


图 3-1 多级多尺度图像表达

我们使用了两种方法来实现多尺度策略。在第一阶段中，一个通用的多尺度训练方法被应用到输入图像上，三个不同的尺度被用来抽取卷积特征图。具体来说，原始的图像由三个不同的尺度因子进行初始变换，然后分别送入三个不同的的 CNN 支路进行，三个支路共享同样的结构和超参数。为了产生更细腻的特征图，三个支路的卷积特征图首先通过双线性插值还原为原始图像分辨率，然后对于每个空间位置都以最大概率的方式进行逐点融合，融合后的概率图作为最终的输出。这个操作在训练和测试的时候同时运行。在第二个阶段中，一个基于“Atrous”卷积层的空间金字塔池化[64]（Spatial Pyramid Pooling，SPP）用于合成多个不同感知域尺度的特征。基于同一个单尺度特征图，我们采用 4 种不同的膨胀率实现卷积特征图的多感知域变换，这 4 种卷积特征图最终通过逐元素相加的方式生成多感知域尺度的卷积特征图。它可以公式化为：

$$\mathcal{F}_M = \max_{m\in[1,\dots,M]} \sum_{n=1}^{N} \mathcal{F}_{(m,n)} \qquad (3\text{-}1)$$

在图 3-1 中，支路 $M$ 和 $N$ 分别为 3 和 4。我们用 $m = 1,2,\dots,M$ 表示分辨率为$R_m$ 的第 $m$ 条支路，每个分辨率$R_m$都有一个固定的尺度因子$f = \{1,0.75,0.5\}$，则$R_m = R_{in}(f \times \text{width}, f \times \text{height}, 3)$，其中$R_{in}$为输入分辨率。$n = 1,2,\dots,N$表示第 $n$ 个感知域的支路。在本章的工作中，感知域的膨胀步长 $k = \{6,12,18,24\}$。$\mathcal{F}_M$ 和$\mathcal{F}_{(m,n)}$分别是多尺度卷积特征图和单尺度卷积特征图。

多级多尺度处理可以有效地改进了性能，但是它也增加了前向推理的时间和大大增加了 GPU 显存的消耗。

### 3.2.2. 基于 CNN 的局部特征学习与提取方案

通过前面文献综述的分析，我们知道：在大数据环境下，由于表达能力弱，用传统的局部特征来表示可视内容，对大规模的视觉对象识别与搜索越来越难。采用深度学习提取视觉中高层语义特征表达，可以获得更强的表达能力，能充分表示和发掘海量数据中蕴藏的语义信息。利用 CNN 对大规模的图像进行特征提取已经被证明是目前最有效的方法，在基于特征的各种应用中，如图像分类、目标检测、人脸识别，图像搜索都达到了目前最好的性能。但是目前采用深度学习的特征往往是基于整个画面的，属于全局特征，在进行多标签分类、局部目标识别与搜索时存在不足。因此，本节将侧重于设计和实现一个可以用于迁移训练并且能够进行局部特征提取的基准 CNN 模型结构。

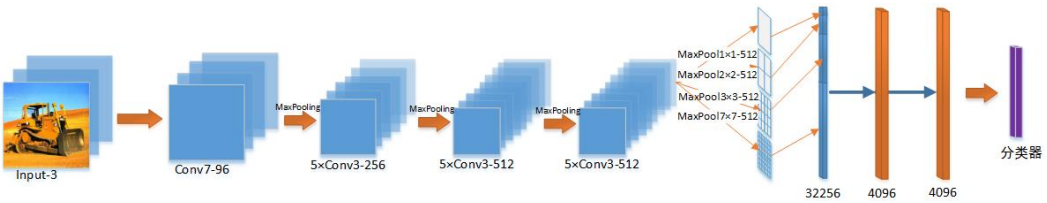


图 3-2 基准 CNN 模型结构示意图

如上图所示，我们设计了一个深层的 CNN 模型来进行特征提取，其中蓝色的模块是卷积层（其中 5×Conv3-256，表示 5 个连续的卷积层，卷积核大小为 3×3，特征图数量为 256）；红色箭头表示 Max-Pooling 层；在 16 个卷积层和 3 个 Max-Pooling 层之后，由 4 个并联的 Max-Pooling 层（7×7，3×3，2×2，1×1）组成一个空间金字塔 Pooling（SPP 层）用于进行多尺度的特征提取，并生成定长特征向量；最后由 2 个全连接层和 1 个基于全连接的分类器层组成。参数化的修正线性单元（PReLU）被应用到每一个卷积层和全连接层之后，用于生成非线性的激活输出。参数化的线性修正单元 PReLU 可以用公式：$f(x_i) = max(0, x_i) + a_i \min(0, x_i)$来表示，相较于目前常用的线性修正单元 ReLUs 和 Sigmoid 函数来说，PReLU 可以加速收敛，并在不增加计算开销的前提下提高大约 1%左右的识别性能。CNN 特征的提取可以从多个激活层产生，包括融合后的 SPP 层、卷积层、Pooling 层和各个全连接层等，也可通过将这些层进行不同程度的融合。

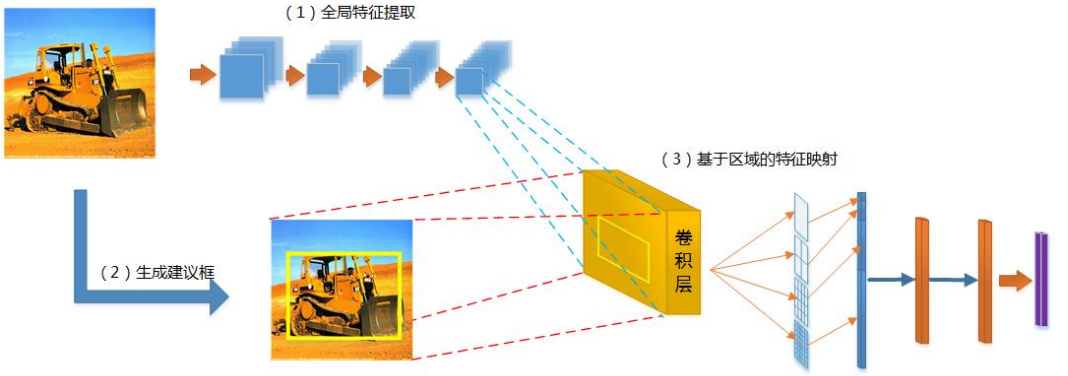


图 3-3 基于建议区域的局部特征提取框架

如上图所示，要实现基于目标对象的特征提取，主要包括三个步骤：

**(1) 基于 CNN 的全局特征提取**

要实现基于对象的特征映射，首先要完成基于样本的全局特征学习，我们利用图 3 中训练好的基准 CNN 模型的卷积部分来学习样本的全局深度特征，在该模型中，我们去掉了全连接层和 Softmax 分类器层，并以 SPP 层作为最后的局部特征提取器。在进行特征映射前，将最后一个卷积层获得的特征保留，用于进行基于对象的特征映射，我们通过 CNN 模型为样本$I(W,H,C)$（其中，$W$, $H$ 表示输入样本的宽和高，$C$ 表示颜色通道的数量）生成一个三维的特征矩阵$F_I(w,h,n)$，其中 $w$, $h$ 表示特征提取层的宽和高，$n$ 表示这个特征层特征图的数量。相对于特征矩阵$F_I(w,h,n)$，原始的输入样本$I(W,H,C)$与其具有稳定的空间位置对应关系，这也是我们实现基于对象的特征映射的基础。

**(2) 基于对象的建议框生成**

受益于计算机视觉的发展，有很多对象分割、对象检测的算法被提出，我们可以利用这些算法来实现在样本中查找可能存在对象的区域，生成对象建议框。目前比较优秀的算法包括 Selective Search、Edge Boxes 和 BING 等。但由于这些算法都同时存在建议框过多的问题，导致后续的特征提取性能下降，因此，为了高效地实现基于对象的特征提取，在本课题，一方面寻求减少建议框数量的方法；另一方面探索实现一些更好的建议框生成机制，如基于对象的概略图等。

为了保证足够的物体检测召回率，Selective Search、Edge Boxes 和 BING 等算法都会生成数千个建议框，这对于基于对象的目标检测是很大的问题，因为过多的建议框将会带来计算复杂性的上升和计算资源的浪费，还会由于时间过长而导致算法的可用性降低。根据我们实验观察，对于明显存在对象的区域，算法总是会生成大量的建议框来匹配这些区域，这些区域可能是对象的整体，也可能是

对象的部件，或者是部件混合了周围比较临近的其他物体；而对于存在对象不明显的区域，建议框较少，或者根本没有建议框。另一方面，当我们获取建议框的一些统计信息的时候，例如梯度信息、颜色信息、纹理信息，明显存在对象的区域总是表现出异质性的特性，范围内的统计信息是杂乱无章的模式，而背景区域总是表现出同质性的特性，范围内的统计信息是比较单一和纯净的。基于这个分析，对于统计性比较杂乱的区域，由于具有较多的建议框将其推荐出来，我们可以通过 max-pooling 的方法来加强这种杂乱的特性，然后对于这些杂乱的区域采用基于几何约束的方式加以统一，从而将多个区域融合成一个区域，这样的区域通常存在对象的概率将大大上升；而对于统计性比较单一的区域和建议框较少的区域，我们可以通过提高阈值的方式，进一步对其排除，因为这些区域通常是以背景的形式存在。通过这种方式，可以大大地减少建议区域的数量，实现建议框的筛选。具体的方法包括：积分投票筛选机制、基于图论的筛选和基于聚类的筛选等。

**(3) 基于区域的特征映射与提取**

通过以上两个步骤，我们将获得基于原始输入样本$I(W,H,C)$的特征矩阵$F_I(w,h,n)$和若干建议区域$R_j(x,y)$。由于输入样本$I$和特征矩阵$F_I$有稳定的空间位置对应关系，因此只需要利用 SPP 层将由输入样本 $I$ 生成的全局特征矩阵$F_I$根据建议区域 $R$ 映射成基于区域的特征矩阵$F_R$，然后再进行特征提取即可。

### 3.2.3. 基于局部特征与全局特征联合的特征提取

相对于常见的类别间差异较大的对象的识别，细粒度对象的目标识别面临更大的挑战。例如：人脸识别，鸟类识别，花草识别等。这些目标从整体上来看具有更加相似的特征，很难通过一个全局的模型就对其进行精确的识别。但仔细观察可以发现，即使是对于同类物体，当我们将其进行一定的尺度变换，在其细节部分仍然可以找到很大的区别用于进行对象的区分。受益于基于对象的局部特征提取方法，一方面，我们可以用预先训练好的基于部件的分类器和基于对象的特征提取来完成部件的精确定位和特征提取；另一方面，我们可以通过提高检测窗口的敏感性来实现细节区域的特征加强和特征提取。

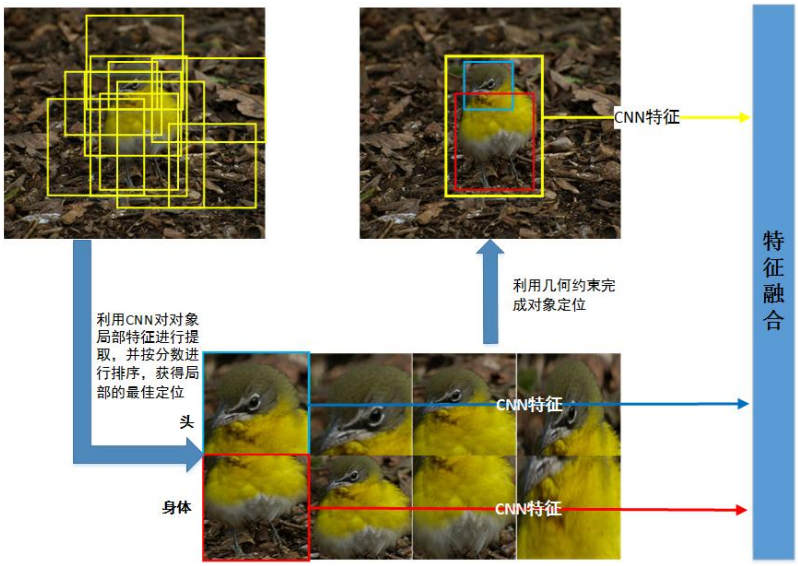


图 3-4 基于对象局部特征与全局特征联合的特征提取

大量实验表明，联合不同局部特征，对于整个样本的识别是有很大帮助的，因此我们提出了基于对象部件和全局特征融合的特征提取方法，如图 3-4 所示，具体过程如下：

1) 利用前面介绍的基于局部对象的特征提取方法，完成样本全局特征的提取和样本局部对象的检测和定位和筛选，然后获得基于局部部件的特征提取。

2) 将对象部件通过几何约束生成完整对象的定位，并结合局部对象的特征提取方法，生成基于对象的深度特征提取。

3) 将对象和对象部件的特征进行融合，并用预先训练好的分类器，完成融合特征的提取，并作为对象最终的特征。

实验证明，由于进行建议框筛选后的区域，总是一些包含对象，或者比较显著的区域，这些区域中的像素对于表达整个样本的独特性，通常具有较强的作用，因此，单独对这些区域进行特征提取，并融合到最后对象的特征表达中，通常能够在复杂场景中实现对重要信息的强化，并进一步提高分类的准确性

### 3.2.4. 基于局部上下文的区域增强

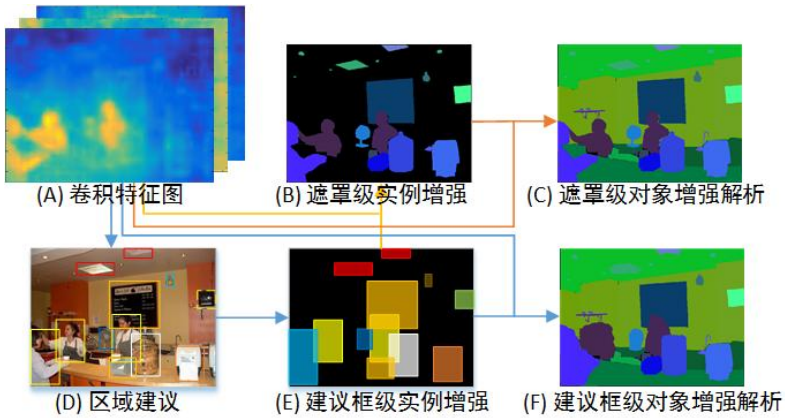


图 3-5 对象增强的流程图

基于局部上下文的区域增强以多尺度特征作为输入，实例感知的解析结果作为输出。它包含三个阶段：建议框级实例增强，遮罩级实例增强和全连接 CRF。

(1) 建议框级实例

建议框级实例是对象的一种粗糙反映，如前所述，我们可以用建议区域$R_i = \{t_i, p_i, c_i\}$ 表示一个对象的定位和分类信息，它粗糙地考虑了一个关于类别$c_i$的矩形区域 $t_i = \{x_i, y_i, w_i, h_i\}$。这个区域可以被转换为建议框级的实例，表示为：

$$B_i(W,H,p) = \begin{cases} 1 & p \in R_i(t_i) \\ 0 & \text{otherwise} \end{cases} \quad (3\text{-}2)$$

其中 $p$ 是区域$B_i(W,H)$中的一个像素。

(2) 遮罩级实例

显而易见，建议框级实例的定义是有缺陷的。对于一个分类或者检测任务，这个粗糙的定义并没有太大问题，然而对于像素级的场景解析任务可能会带来很多麻烦。大自然中的对象通常都是不规则的，使用一个矩形来描述对象将会有大量的背景元素被混合进去。而简单的使用建议框级的实例去增强场景解析可以改进对象的识别，但是也可能会影响背景的识别。我们可以将建议框级实例和卷积特征组合在一起生成了遮罩级实例。对于每一个建议框级实例$B_i(W,H,p)$，我们都可以得到一个遮罩级实例$M_i(W,H,p)$。这个过程可以公式化为：

$$M_i(W,H,p) = \begin{cases} 1 & p \in F(W,H,c_i) \cdot R_i(t_i,c_i) > t \\ 0 & \text{otherwise} \end{cases} \quad (3\text{-}3)$$

其中 $p$ 是遮罩内的像素，$F(W,H,c_i)$ 是关于类别$c_i$的特征图，它由 CNN 前向

推导得到的卷积特征图的第$c_i$个通道在双线性插值和全卷积 CRF 过滤后获得，它的宽度 $W$ 和高度 $H$ 与输入图像一致。阈值 $t$ 控制遮罩的大小，它的上界为建议框级实例。我们可以发现，遮罩只涉及索引和建议框类别相同的通道。此外，由于同一个类别可能会包含多个实例对象，为了方便，我们可以将这些具有相同类别的特征遮罩合并在一起统一进行计算。最终，我们可以通过迭代所有的区域$R_i$获得完整的特征图 $M(W,H,C)$，其中$C=0,1,\dots,c_i$是类别空间。

(3) 基于实例的增强

在获得了建议框级和遮罩级实例之后，我们可以将他们应用到多尺度卷积特征图上用来生成对象级增强特征。实例特征增强在权重 w 和建议框类别概率$p_i$作用下共同完成。其中遮罩级增强特征$F_{ME}$可以表示为：

$$F_{ME}(W,H,C)=F(W,H,c_i)\cdot M\left(W,H,c_j\right)\times w\times p_i\times c^* \tag{3-4}$$

如果 $c_i=c_j$，则 $c^*=1$，否则$c^*=0$，这意味着只有当特征图和实例具有同样的类别标签时，对象区域增强才被激活。在公式(3-4)中，通过替换 $M\left(W,H,c_j\right)$ 为$B_i\left(W,H,c_j\right)$可以获得建议框级实例增强特征$F_{BE}$。

(4) 全连接 CRF

由于多个叠加的最大池化层（max-pooling layer）的存在，不断增加的不变性和较大的感知域会导致输出的激活响应过于平滑，而导致场景解析中的小范围内的分类过趋于一致。这对于复杂的场景显然是一种有害的影响。为了尽量克服这个限制，CRF 是一种有效的方法，它以能量函数进行优化：

$$\begin{aligned}E(x)=&\sum_i(-\log P(x_i))\\&+\sum_{ij}\mu(x_i,x_j)\left[\lambda_1 exp\left(-\frac{\|p_i-p_j\|^2}{2\delta_\alpha^2}\right)\right.\\&\left.+\lambda_2 exp\left(-\frac{\|p_i-p_j\|^2}{2\delta_\beta^2}-\frac{\|c_i-c_j\|^2}{2\delta_\gamma^2}\right)\right]\end{aligned}\tag{3-5}$$

其中 $x$ 是图像中每个像素的标签。一元项$P(x_i)$是像素 $i$ 通过 CNN 获得的推理概率。后续的二元项允许一对连接的图像执行全连接的图推理。我们使用 $i$ 和 $j$ 分别表示卷积特征图和原始输入图上一个像素的位置。如果$x_i=x_j$，则$\mu(x_i,x_j)=$

1；否则，$\mu(x_i, x_j) = 0$。也就是说，只有当节点具有不同标签的时候才会执行惩罚。如公式(3-5)所示，两个高斯核被应用到不同的特征空间上。前面的一项表示只利用了像素位置的信息（表示为 $p$），它在执行平滑的时候只考虑了近邻空间的信息；后面的一项表示的是一个双线性核，它同时依赖于 RGB 颜色空间（表示为 $c$）和位置空间（表示为 $p$），它强制使具有相似颜色和相似位置的像素具有相似的标签。超参数$\delta_\alpha$，$\delta_\beta$和$\delta_\gamma$控制高斯核的尺度，权重参数$\lambda_2$和$\lambda_2$用来平衡两个特征的重要性。

### 3.2.5. 基于层次化语义的样本过滤策略

(1) 相似性策略

相似性检索的目标是从一个图像集合中找到与给定图像最相似的图像或者图像的子集。但是，如何定义最相似呢？为了回答这个问题并利用层次化的知识来实现图像检索，我们考虑两个子进程来评估两个图像间的相似性。首先，如何有效地表达图像，尽量避免不必要的语义损失；其次，如何快速地计算相似性。

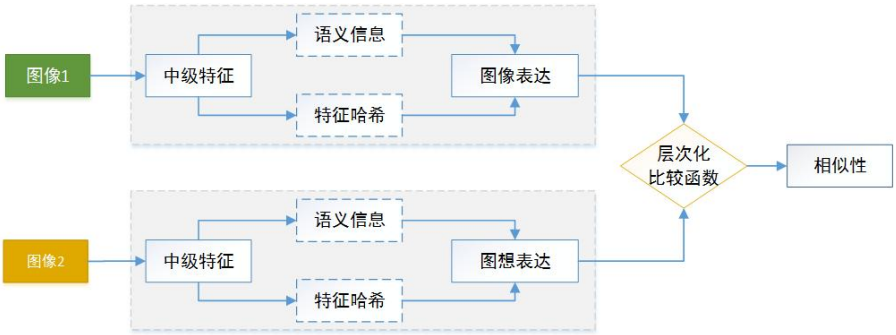


图 3-6 两个图像相似性评估方法示意图

值得注意的是，我们并不需要像传统方法一样去精确和细致地设计哈希函数，来获样本的哈希值。通过深度卷积神经网络的前向传播过程和层次化融合策略，我们可以直接得到样本的哈希值。

(1) 基于标签的语义级相似性

我们提出的方法的核心是利用语义属性和中级特征之间的层次性的先验知识去计算相似性，从而实现图像检索。我们首先考虑一个基于标签的语义级相似性来进行相似性的衡量，该方法是一个非概率的版本，它使用二进制属性集$\{1, \ldots, K\}$来描述图像 $a$。一般来说，属性可以是对象的颜色（“是红色”），纹理（“是木质”），类别（“是一辆汽车”），部件（“是头部”），或者是其他任意的关于图像 $a$ 的信息。在本课题的研究中我们主要集中于对象的类别属性，但是该方法也可

以很容易地扩展到任意属性。

给定语义标签集 $L=\{1,...,C\}$，两个图像 $a$ 和 $b$ 之间的相似性可以使用它们标签集的匹配程度来衡量。我们用符号 $\delta_i(a)\in\{0,1\}$ 标识图像 $a$ 是否包含语义 $i$，那么图像 $a$ 的语义信息可以表示为 $L_i^C(a)=\{\delta_i(a)|i=(1,...,C)\}$，其中有且仅有一个 $\delta_i(a)=1$。我们可以基于图像 $a$ 和 $b$ 的语义标签定义它们的相似性 $Sim(a,b)=\sum_{(i,j)}\delta_i(a)\boldsymbol{S}_{ij}\delta_i(b)$，其中 $\boldsymbol{S}\in\mathbb{R}^{(C\times C)}$，同时$\boldsymbol{S}_{ij}$可以被认为是语义 $i$ 和 $j$ 之间的“匹配分数”矩阵。这是一个非常通用的形式，它需要一个非常庞大的类别语义空间来计算相似性。然而这种基于标签的语义级相似性非常依赖于人的先验知识，也就是说必须要对每一个样本都能够精确定义它所属的类别。举个特殊的例子，当数据集的语义标签是互斥的，并且 $S$ 是一个标识矩阵，那么相似性 $Sim(a,b)$ 可以被用来标识图像 $a$ 和 $b$ 是否属于同一类别。换句话说，我们可以用“相同”或“不相同”来衡量图像 $a$ 和 $b$ 之间关于语义 $i$ 的相似性。具体说，我们可以重新定义图像 $a$ 和 $b$ 之间的相似性 $Sim(a,b)=1\{L_i^C(a)==L_i^C(b)\}\in\{0,1\}$，也就是说，$Sim(a,b)=1$ 表示图像 $a$ 和 $b$ 相似，而$Sim(a,b)=0$ 表示图像 $a$ 和 $b$ 不相似。基于这个设置，两个图像之间的语义信息只能用来表达它们是相同或者不相同的类别，检索系统就无法通过语义信息来衡量图像 $a$ 和 $b$ 关于查询样本哪一个更相似，即无法实现相似性的排序。尽管如此，依然有很多方法基于这个设置来实现图像检索。与这些方法不同的是，我们将这个设置作为一个过滤器来改进我们的检索速度，而不是直接用它来衡量图像的相似性。

(2) 基于概率的语义级相似性

基于标签的语义级相似性是一种很容易想到并实现的方法，然而，仅仅利用语义类别来衡量相似性对于图像检索来说是一个巨大的挑战。一方面，自然界的语义类别总会存在重叠，类别也经常会出现二义性，仅仅使用硬类别去进行识别很容易失败。另一方面，完美的分类语义也是不现实的。例如，通常我们认为“环尾狐猴”和“冠美狐猴”是相似的，这意味着他们都属于狐猴。但是，如果只需要从数据集中搜索“环尾狐猴”，那将会变得十分地困难。

为了解决这个问题并改进检索性能，我们提出使用基于概率版本的语义级相似性替代简单的基于标签的语义级相似性。给定语义标签集 $L=\{1,...,C\}$，图像 $a$ 和 $b$ 的相似性可以通过他们的匹配程度来衡量。符号 $\delta_i(a)\in\{0,1\}$ 标识图像 $a$ 是否包含语义 $i$，类似地，我们也可以使用概率 $Y_i^C=P(\delta_i(a)=1|a)$ 来标识图像 $a$ 包含语义 $i$ 的可能性。显然，索引 $i=max(Y_i^C)$ 是图像 $a$ 的语义标签。在本课

题的工作中，概率$Y_i^C$可以通过卷积神经网络的前向传播获得，它的值等于 *Softmax* 分类器的输入向量。和文献[155]类似，我们采用查询—图像的信息相关性来定义概率版的相似性。

我们考虑两个图像相关，只发生在它们的联合概率超过截断阈值 *t* 时，则：

$$R_{II}\big(I(a),I(b)\big)=\big[P\big(I(a),I(b)\big)\big]_t \tag{3-6}$$

也就是说：

$$x=\begin{cases}[x]_t, & x>t\\ 0, & otherwise\end{cases} \tag{3-7}$$

其中$x=P\big(I(a),I(b)\big)$。

基于上述讨论，我们可以发现相关矩阵是一个对角矩阵，因为图像相关仅仅发生在两幅图像同时拥有语义 *i*。此外，在图像所涉及的语义子集中，大多数语义的概率都非常低，因此，通过截断阈值 *t* 的设置，大多数概率被强制置 0。因此，相关矩阵具有较强的稀疏性，这也使基于概率的语义级相似性能够过滤掉大量的不相关图像。

(3) 哈希级相似性

在本节中，我们讨论哈希级相似性。给定图像 *I*，我们首先抽取全连接层的输出作为图像的中级特征表达，它可以用一个 *D* 维的特征向量 $g(I)$ 来表示，其中 $g(\cdot)$ 是输出层之前所有层次关于输入图像的卷积变换。然后，我们可以通过一个简单哈希函数 $h(\cdot)$ 将这个 *D* 维的特征向量转换为 *q* 比特的二进制编码。对于每一个比特 $i=1,...,q$，我们可以通过如下公式输出它的二进制哈希编码：

$$H=h(x)=\begin{cases}1, & f(x_i)-Avg_i^q\big(f(x_i)\big)>0\\ 0, & f(x_i)-Avg_i^q\big(f(x_i)\big)<0\end{cases} \tag{3-8}$$

其中 $x=g(I)$ 是卷积层输出的 CNN 特征，$x_i(i=1,...L)$，$L=\{1,...,C\}$。$f(x)$ 是 *Sigmoid* 函数，它的定义公式为 $Sigmoid(v)=\frac{1}{1+e^{-v}}$，符号 $Avg(\boldsymbol{u})$ 是均值函数，对于求取向量 $\boldsymbol{u}$ 中所有元素的平均值。这里 *Sigmoid* 函数用于将输出的 CNN 特征归一化到区间 $[0,1]$。

假设 $I=\{I_1,I_2,...,I_n\}$ 是由 *n* 个图像构成的图像集和，$H=\{H_1,H_2,...,H_n\}$，

$H_i \in \{0,1\}^q$ 是关于图像集 $I$ 中每一个图像的二进制编码。给定一个查询图像$I_q$，我们可以使用它的二进制编码形式$H_q$去标识这幅图像。那么，查询图像$H_q$和图像集中的图像$H_i \in H$之间的哈希级相似性可以用它们之间的欧式距离来衡量：

$$d(q,i) = Dist(I_q, I_i) = ||H_q - H_i|| \tag{3-9}$$

对于两个图像来说，它们之间的欧氏距离越小，意味着它们越相似。此时，图像集中的每个图像$I_i$可以依据相似性进行降序排列，从而得到前 $k$ 个最相似的图像。

(4) 语义级和哈希级融合的相似性

至此，我们定义了图像 $a$ 和 $b$ 之间的语义级相似性和哈希级相似性，下一步，我们可以将它们组合起来形成我们的层次化的相似性：

$$\begin{aligned} Sim(a,b) &= \sum_{i}^{C} (p(a), H_a)\, S(p(b), H_b) \\ &= \sum_{i}^{C} [P(I(a), I(b))]_t \times (1 - d(a,b)) \end{aligned} \tag{3-10}$$

其中，$\boldsymbol{R}_{II}(I(a), I(b)) = [P(I(a), I(b))]_t$是基于概率的语义级相似性，$1 - d(a,b)$是哈希级相似性。公式前面的项是一个对角矩阵，后面的项是一个值。操作“×”将两个相似性联合在一起组成一个确定的值，用来衡量图像 $a$ 和 $b$ 之间的相似性。实际上，$\boldsymbol{R}_{II}$是非常稀疏的，这非常有利于去创建一个和查询图像相关的图像列表。给定一个查询图像 $q$，我们遍历整个数据集去查找所有和查询图像 $q$ 语义相关的图像。对于查询图像 $q$ 来说，它可能会和多个语义或者近似语义相关。我们将涉及到这些相关语义的所有图像组合成一个图像列表，然后再利用哈希特征来求取它们和查询图像之间的相似性序列。

# 4. 课题研究进展计划

本课题研究进展大致分为以下几个阶段：

1. 2013 年 7 月至 2014 年 7 月，收集相关资料，进行深度学习理论基础的学习；熟悉实验环境，并搭建实验平台；研究基本的基于 CNN 的特征提取基准模型框架，并进行基于 CNN 的图像分类、检测和检索等应用的基础性研究；撰写相关论文。
2. 2014 年 8 月至 2015 年 7 月，研究层次化语义方法，并应用到基于内容的图像检索上；撰写相关论文。
3. 2015 年 8 月至 2016 年 7 月，结合层次化的思路，考虑多上下文融合的策略，并应用到特定视觉内容的识别上；撰写相关论文。
4. 2016 年 8 月至 2017 年 1 月，研究基于多上下文的场景解析；撰写相关论文。
5. 2017 年 2 月至 2017 年 3 月，撰写毕业论文。
6. 2017 年 4 月至 2017 年 6 月，完善毕业论文，并进行毕业论文答辩。

# 参考文献

[1] Felzenszwalb, Pedro F.; Girshick, Ross B.; Mcallester, David A., et al. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. 2010. 32(9):1627-1645.

[2] Sermanet, Pierre; Eigen, David; Zhang, Xiang, et al. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *CoRR*. 2013.

[3] Simonyan, Karen; Zisserman, Andrew. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*. 2014.

[4] Szegedy, Christian; Liu, Wei; Jia, Yangqing, et al. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2015. 1-9

[5] Szummer, Martin; Picard, Rosalind W. Indoor-Outdoor Image Classification. In *International Workshop on Content-Based Access of Image and Video Databases (CAIVD)*. Bombay, India. 1998. 42-51

[6] Sivic, Josef; Zisserman, Andrew. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *IEEE International Conference on Computer Vision (ICCV)*. Nice, France. 2003. 1470-1477

[7] Perronnin, Florent; Dance, Christopher R. Fisher Kernels on Visual Vocabularies for Image Categorization. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. Minneapolis, Minnesota, USA. 2007.

[8] Nchez, Jorge S. A.; Perronnin, Florent; Mensink, Thomas, et al. Image Classification with the Fisher Vector: Theory and Practice. *International Journal of Computer Vision (IJCV)*. 2013. 105(3):222-245.

[9] Hinton, Geoffrey E.; Srivastava, Nitish; Krizhevsky, Alex, et al. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*. 2012.

[10] Lowe, David G. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*. 2004. 60(2):91-110.

[11] Lowe, David G. Object Recognition from Local Scale-Invariant Features. In *IEEE International Conference on Computer Vision*. 1999. 1150-1157

[12] Dalal, Navneet; Triggs, Bill. Histograms of Oriented Gradients for Human Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2005. 886-893

[13] Ojala, Timo; Inen, Matti Pietik A.; A, Topi M. A. Enp. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2002. 24(7):971-987.

[14] Nasrabadi, Nasser M.; King, Robert A. Image coding using vector quantization: a review.

*IEEE Transactions Communications*. 1988. 36(8):957-971.

[15] Wright, John; Ma, Yi; Mairal, Julien, et al. Sparse Representation for Computer Vision and Pattern Recognition. *Proceedings of the IEEE*. 2010. 98(6):1031-1044.

[16] Hedelin, Per; Skoglund, Jan. Vector quantization based on Gaussian mixture models. *IEEE Transactions Speech and Audio Processing*. 2000. 8(4):385-401.

[17] Lazebnik, Svetlana; Schmid, Cordelia; Ponce, Jean. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR}*. New York, NY, USA. 2006. 2169-2178

[18] Arandjelovic, Relja; Zisserman, Andrew. All About VLAD. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2013. 1578-1585

[19] Perronnin, Florent; Liu, Yan; Nchez, Jorge S. A., et al. Large-scale image retrieval with compressed Fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2010. 3384-3391

[20] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing, et al. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2016. 770-778

[21] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing, et al. Identity Mappings in Deep Residual Networks. In *European Conference on Computer Vision*. 2016. 630-645

[22] Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E. ImageNet classification with deep convolutional neural networks. In *Annual Conference on Neural Information Processing Systems*. 2012. 1097-1105

[23] Szegedy, Christian; Ioffe, Sergey; Vanhoucke, Vincent, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*. San Francisco, California, USA. 2017. 4278-4284

[24] Torralba, Antonio. Contextual Priming for Object Detection. *International Journal of Computer Vision (IJCV)*. 2003. 53(2):169-191.

[25] Fergus, Robert; Perona, Pietro; Zisserman, Andrew. Object Class Recognition by Unsupervised Scale-Invariant Learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. Madison, WI, USA. 2003. 264-271

[26] Weber, Markus; Welling, Max; Perona, Pietro. Towards Automatic Discovery of Object Categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Hilton Head, SC, USA. 2000. 2101

[27] Breiman, Leo. Random Forests. *Machine Learning*. 2001. 45(1):5-32.

[28] Cao, Yang; Wang, Changhu; Li, Zhiwei, et al. Spatial-bag-of-features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. San Francisco, CA, USA. 2010.

3352-3359

[29] Csurka, Gabriella; Dance, Christopher R.; Fan, Lixin, et al. Visual categorization with bags of keypoints. *European Conference on Computer Vision*. 2004. 44(247):1-22.

[30] Laptev, Ivan. Improvements of Object Detection Using Boosted Histograms. In *Proceedings of the British Machine Vision Conference (BMVC)*. Edinburgh, UK. 2006. 949-958

[31] Zhang, Jianguo; Marszalek, Marcin; Lazebnik, Svetlana, et al. Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *International Journal of Computer Vision (IJCV)*. 2007. 73(2):213-238.

[32] Zhang, Ning; Donahue, Jeff; Girshick, Ross B., et al. Part-Based R-CNNs for Fine-Grained Category Detection. In *European Conference on Computer Vision (ECCV)*. Zurich, Switzerland. 2014. 834-849

[33] Felzenszwalb, Pedro F.; Huttenlocher, Daniel P. Pictorial Structures for Object Recognition. *International Journal of Computer Vision (IJCV)*. 2005. 61(1):55-79.

[34] Harzallah, Hedi; Jurie, Fr E. D. E.; Schmid, Cordelia. Combining efficient object localization and image classification. In *IEEE 12th International Conference on Computer Vision (ICCV)*. Kyoto, Japan. 2009. 237-244

[35] Song, Zheng; Chen, Qiang; Huang, Zhongyang, et al. Contextualizing object detection and classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Colorado Springs, CO, USA. 2011. 1585-1592

[36] Russakovsky, Olga; Lin, Yuanqing; Yu, Kai, et al. Object-Centric Spatial Pooling for Image Classification. In *European Conference on Computer Vision (ECCV)*. 2012. 1-15

[37] Chen, Qiang; Song, Zheng; Hua, Yang, et al. Hierarchical matching with side information for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Providence, RI, USA. 2012. 3426-3433

[38] Girshick, Ross B.; Donahue, Jeff; Darrell, Trevor, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2014. 580-587

[39] Szegedy, Christian; Liu, Wei; Jia, Yangqing, et al. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015. 1-9

[40] Liu, Si; Feng, Jiashi; Song, Zheng, et al. Hi, magic closet, tell me what to wear!. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*. Nara, Japan. 2012. 619-628

[41] Karpathy, Andrej; Toderici, George; Shetty, Sanketh, et al. Large-Scale Video Classification with Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Columbus, OH, USA. 2014. 1725-1732

[42] Ciresan, Dan C.; Meier, Ueli; Schmidhuber, J. U. Rgen. Multi-column deep neural networks

for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Providence, RI, USA. 2012. 3642-3649

[43] Sermanet, Pierre; Eigen, David; Zhang, Xiang, et al. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *CoRR*. 2013.

[44] Zhao, Rui; Ouyang, Wanli; Li, Hongsheng, et al. Saliency detection by multi-context deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA. 2015. 1265-1274

[45] Simonyan, Karen; Zisserman, Andrew. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*. 2014.

[46] Yu, Jun; Rui, Yong; Tang, Yuan Yan, et al. High-Order Distance-Based Multiview Stochastic Learning in Image Classification. *IEEE Transactions Cybernetics*. 2014. 44(12):2431-2442.

[47] Yu, Jun; Yang, Xiaokang; Fei, Gao, et al. Deep Multimodal Distance Metric Learning Using Click Constraints for Image Ranking. *IEEE Transactions on Cybernetics (ToC)*. 2016. 1-11.

[48] Russakovsky, Olga; Deng, Jia; Su, Hao, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*. 2015. 115(3):211-252.

[49] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing, et al. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. 770-778

[50] Imagenet. Imagenet Large Scale Visual Recognition Challenge 2015 (ILSVRC2015). 2015. http://image-net.org/challenges/LSVRC/2015/results

[51] Imagenet. Imagenet Large Scale Visual Recognition Challenge 2016 (ILSVRC2016). 2016. http://image-net.org/challenges/LSVRC/2016/results

[52] Zhang, Fan; Du, Bo; Zhang, Liangpei, et al. Weakly Supervised Learning Based on Coupled Convolutional Neural Networks for Aircraft Detection. *IEEE Transactions Geoscience and Remote Sensing*. 2016. 54(9):5553-5563.

[53] Fergus, Robert; Perona, Pietro; Zisserman, Andrew. Object Class Recognition by Unsupervised Scale-Invariant Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Madison, WI, USA. 2003. 264-271

[54] Liang, Xiaodan; Xu, Chunyan; Shen, Xiaohui, et al. Human Parsing with Contextualized Convolutional Neural Network. In *IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile. 2015. 1386-1394

[55] Liu, Si; Liang, Xiaodan; Liu, Luoqi, et al. Fashion Parsing With Video Context. *IEEE Transaction Multimedia*. 2015. 17(8):1347-1358.

[56] Liu, Si; Wang, Changhu; Qian, Ruihe, et al. Surveillance Video Parsing with Single Frame Supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

2017.

[57] Fellbaum, C.; Miller, G. WordNet : an electronic lexical database. *The Library Quarterly: Information, Community, Policy*. 1999.

[58] Deng, Jia; Berg, Alexander C.; Li, Kai, et al. What Does Classifying More Than 10, 000 Image Categories Tell Us? In *European Conference on Computer Vision (ECCV)*. Heraklion, Crete, Greece. 2010. 71-84

[59] Branson, Steve; Wah, Catherine; Schroff, Florian, et al. Visual Recognition with Humans in the Loop. In *European Conference on Computer Vision (ECCV)*. Heraklion, Crete, Greece. 2010. 438-451

[60] Fergus, Robert; Bernal, Hector; Weiss, Yair, et al. Semantic Label Sharing for Learning with Many Categories. In *European Conference on Computer Vision (ECCV)*. Heraklion, Crete, Greece. 2010. 762-775

[61] Liu, Si; Liang, Xiaodan; Liu, Luoqi, et al. Matching-CNN meets KNN: Quasi-parametric human parsing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA. 2015. 1419-1427

[62] Yang, Cong; Tiebe, Oliver; Shirahama, Kimiaki, et al. Object matching with hierarchical skeletons. *Pattern Recognition*. 2016. 55(183-197.

[63] Yu, Jun; Zhang, Baopeng; Kuang, Zhengzhong, et al. iPrivacy: Image Privacy Protection by Identifying Sensitive Objects via Deep Multi-Task Learning. *IEEE Transactions Information Forensics and Security*. 2017. 12(5):1005-1016.

[64] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2015. 37(9):1904-1916.

[65] Figueroa, Nadia; Dong, Haiwei; Saddik, Abdulmotaleb El. A Combined Approach Toward Consistent Reconstructions of Indoor Spaces Based on 6D RGB-D Odometry and KinectFusion. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2015. 6(2):11-14.

[66] Lin, Kevin; Yang, Huei Fang; Hsiao, Jen Hao, et al. Deep learning of binary hash codes for fast image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015. 27-35

[67] Eigen, David; Fergus, Rob. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture. In *IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile. 2015. 2650-2658

[68] Li, Xiaoyan; Liu, Tongliang; Deng, Jiankang, et al. Video Face Editing Using Temporal-Spatial-Smooth Warping. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2016. 7(3):31-32.

[69] Shelhamer, Evan; Long, Jonathan; Darrell, Trevor. Fully Convolutional Networks for Semantic Segmentation. *CoRR*. 2016.

[70] Yang, Jianchao; Yu, Kai; Gong, Yihong, et al. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. Miami, Florida, USA. 2009. 1794-1801

[71] Cortes, Corinna; Vapnik, Vladimir. Support-Vector Networks. *Machine Learning*. 1995. 20(3):273-297.

[72] Russakovsky, Olga; Deng, Jia; Su, Hao, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*. 2015. 115(3):211-252.

[73] Deng, Jia; Dong, Wei; Socher, Richard, et al. ImageNet: A large-scale hierarchical image database. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. Miami, Florida, USA. 2009. 248-255

[74] Ciresan, Dan C.; Meier, Ueli; Schmidhuber, J. U. Rgen. Multi-column deep neural networks for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*. Providence, RI, USA. 2012. 3642-3649

[75] Ciresan, Dan C.; Meier, Ueli; Masci, Jonathan, et al. Multi-column deep neural network for traffic sign classification. *Neural Networks*. 2012. 32(333-338.

[76] Parkhi, Omkar M.; Vedaldi, Andrea; Zisserman, Andrew. Deep Face Recognition. In *Proceedings of the British Machine Vision Conference*. Swansea, UK. 2015. 4101-4112

[77] Sun, Yi; Chen, Yuheng; Wang, Xiaogang, et al. Deep Learning Face Representation by Joint Identification-Verification. In *Advances in Neural Information Processing Systems*. Montreal, Quebec, Canada. 2014. 1988-1996

[78] Sun, Yi; Wang, Xiaogang; Tang, Xiaoou. Deep Learning Face Representation from Predicting 10, 000 Classes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Columbus, OH, USA. 2014. 1891-1898

[79] Chatfield, Ken; Simonyan, Karen; Vedaldi, Andrea, et al. Return of the Devil in the Details: Delving Deep into Convolutional Nets. In *British Machine Vision Conference (BMVC)*. Nottingham, UK. 2014.

[80] Wan, Li; Zeiler, Matthew D.; Zhang, Sixin, et al. Regularization of Neural Networks using DropConnect. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*. Atlanta, GA, USA. 2013. 1058-1066

[81] Zeiler, Matthew D.; Ranzato, Marc'aurelio; Monga, Rajat, et al. On rectified linear units for speech processing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vancouver, BC, Canada. 2013. 3517-3521

[82] Zeiler, Matthew D.; Fergus, Rob. Visualizing and Understanding Convolutional Networks. In *European Conference Computer Vision (ECCV)*. Zurich, Switzerland. 2014. 818-833

[83] Srivastava, Rupesh Kumar; Masci, Jonathan; Kazerounian, Sohrob, et al. Compete to Compute. In *Advances in Neural Information Processing Systems (NIPS)*. Lake Tahoe, Nevad. 2013. 2310-2318

[84] Goodfellow, Ian J.; Farley, David Warde; Mirza, Mehdi, et al. Maxout Networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*. Atlanta, GA, USA. 2013. 1319-1327

[85] Min Lin, Qiang Chen And Shuicheng. Network in network. *arXiv preprint*. 2013.

[86] Nair, Vinod; Hinton, Geoffrey E. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*. Haifa, Israel. 2010. 807-814

[87] Ng, Al Maas Ay Hannun. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*. 2013.

[88] Srivastava, Nitish; Hinton, Geoffrey E.; Krizhevsky, Alex, et al. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*. 2014. 15(1):1929-1958.

[89] Lin, Tsung Yi; Maire, Michael; Belongie, Serge J., et al. Microsoft COCO: Common Objects in Context. In *European Conference Computer Vision*. Zurich, Switzerland. 2014. 740-755

[90] Everingham, Mark; Van Gool, Luc J.; Williams, Christopher K. I., et al. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*. 2010. 88(2):303-338.

[91] Wang, Xiaoyu; Yang, Ming; Zhu, Shenghuo, et al. Regionlets for Generic Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2015. 37(10):2071-2084.

[92] Uijlings, Jasper R. R.; van de Sande, Koen E. A.; Gevers, Theo, et al. Selective Search for Object Recognition. *International Journal of Computer Vision (IJCV)*. 2013. 104(2):154-171.

[93] Vedaldi, Andrea; Gulshan, Varun; Varma, Manik, et al. Multiple kernels for object detection. In *IEEE International Conference on Computer Vision (ICCV)*. Kyoto, Japan. 2009. 606-613

[94] Cheng, Ming Ming; Zhang, Ziming; Lin, Wen Yan, et al. BING: Binarized Normed Gradients for Objectness Estimation at 300fps. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Columbus, OH, USA. 2014. 3286-3293

[95] Alexe, Bogdan; Deselaers, Thomas; Ferrari, Vittorio. Measuring the Objectness of Image Windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2012. 34(11):2189-2202.

[96] Ez, Pablo Andr E. S.; Tuset, Jordi Pont; Barron, Jonathan T., et al. Multiscale Combinatorial Grouping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Columbus, OH, USA. 2014. 328-335

[97] Zitnick, C. Lawrence; R, Piotr Doll A. Edge Boxes: Locating Object Proposals from Edges. In *European Conference on Computer Vision (ECCV)*. Zurich, Switzerland. 2014. 391-405

[98] Kuo, Weicheng; Hariharan, Bharath; Malik, Jitendra. DeepBox: Learning Objectness with Convolutional Networks. In *IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile. 2015. 2479-2487

[99] Girshick, Ross B. Fast R-CNN. In *IEEE International Conference on Computer Vision*. 2015. 1440-1448

[100] Ren, Shaoqing; He, Kaiming; Girshick, Ross B., et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*. 2015. 91-99

[101] Liu, Wei; Anguelov, Dragomir; Erhan, Dumitru, et al. SSD: Single Shot MultiBox Detector. In *European Conference on Computer Vision (ECCV)*. Amsterdam, The Netherlands. 2016. 21-37

[102] Redmon, Joseph; Divvala, Santosh Kumar; Girshick, Ross B., et al. You Only Look Once: Unified, Real-Time Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA. 2016. 779-788

[103] Shrivastava, Abhinav; Gupta, Abhinav; Girshick, Ross B. Training Region-Based Object Detectors with Online Hard Example Mining. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA. 2016. 761-769

[104] Kong, Tao; Yao, Anbang; Chen, Yurong, et al. HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA. 2016. 845-853

[105] Bell, Sean; Zitnick, C. Lawrence; Bala, Kavita, et al. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA. 2016. 2874-2883

[106] Gidaris, Spyros; Komodakis, Nikos. Object Detection via a Multi-region and Semantic Segmentation-Aware CNN Model. In *IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile. 2015. 1134-1142

[107] Tu, Zhuowen; Bai, Xiang. Auto-Context and Its Application to High-Level Vision Tasks and 3D Brain Image Segmentation. *IEEE Transactions on Software Engineering*. 2009. 32(32):1744-1757.

[108] Shotton, Jamie; Winn, John M.; Rother, Carsten, et al. TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context. *International Journal of Computer Vision (IJCV)*. 2009. 81(1):2-23.

[109] Carreira, Jo A. O.; Caseiro, Rui; Batista, Jorge, et al. Semantic Segmentation with Second-Order Pooling. In *European Conference on Computer Vision (ECCV)*. Florence,

Italy. 2012. 430-443

[110] Carreira, Jo A. O.; Sminchisescu, Cristian. CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. 2012. 34(7):1312-1328.

[111] Hl, Philipp Kr A. Henb; Koltun, Vladlen. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *Advances in Neural Information Processing Systems (NIPS)*. Granada, Spain. 2011. 109-117

[112] Hariharan, Bharath; Ez, Pablo Andr E. S.; Girshick, Ross B., et al. Simultaneous Detection and Segmentation. In *European Conference Computer Vision*. Zurich, Switzerland. 2014. 297-312

[113] Mostajabi, Mohammadreza; Yadollahpour, Payman; Shakhnarovich, Gregory. Feedforward semantic segmentation with zoom-out features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA. 2015. 3376-3385

[114] Farabet, Cl E. Ment; Couprie, Camille; Najman, Laurent, et al. Learning Hierarchical Features for Scene Labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2013. 35(8):1915-1929.

[115] Hariharan, Bharath; Ez, Pablo Andr E. S.; Girshick, Ross B., et al. Hypercolumns for object segmentation and fine-grained localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA. 2015. 447-456

[116] Dai, Jifeng; He, Kaiming; Sun, Jian. Convolutional feature masking for joint object and stuff segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA, USA. 2015. 3992-4000

[117] Chen, Liang Chieh; Papandreou, George; Kokkinos, Iasonas, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *CoRR*. 2016.

[118] Hl, Philipp Kr A. Henb; Koltun, Vladlen. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *Advances in Neural Information Processing Systems*. Granada, Spain. 2011. 109-117

[119] Chen, Liang Chieh; Yang, Yi; Wang, Jiang, et al. Attention to Scale: Scale-Aware Semantic Image Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA. 2016. 3640-3649

[120] Dai, Jifeng; He, Kaiming; Sun, Jian. BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation. In *IEEE International Conference on Computer Vision*. Santiago, Chile. 2015. 1635-1643

[121] Zheng, Shuai; Jayasumana, Sadeep; Paredes, Bernardino Romera, et al. Conditional Random Fields as Recurrent Neural Networks. In *IEEE International Conference on Computer Vision*

*(ICCV)*. Santiago, Chile. 2015. 1529-1537

[122] Lin, Guosheng; Shen, Chunhua; van den Hengel, Anton, et al. Efficient Piecewise Training of Deep Structured Models for Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA. 2016. 3194-3203

[123] Noh, Hyeonwoo; Hong, Seunghoon; Han, Bohyung. Learning Deconvolution Network for Semantic Segmentation. In *IEEE International Conference on Computer Vision*. Santiago, Chile. 2015. 1520-1528

[124] Liu, Ziwei; Li, Xiaoxiao; Luo, Ping, et al. Semantic Image Segmentation via Deep Parsing Network. In *IEEE International Conference on Computer Vision*. Santiago, Chile. 2015. 1377-1385

[125] Chen, Liang Chieh; Barron, Jonathan T.; Papandreou, George, et al. Semantic Image Segmentation with Task-Specific Edge Detection Using CNNs and a Discriminatively Trained Domain Transform. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA. 2016. 4545-4554

[126] Papandreou, George; Chen, Liang Chieh; Murphy, Kevin, et al. Weakly- and Semi-Supervised Learning of a DCNN for Semantic Image Segmentation. *CoRR*. 2015.

[127] Lin, Guosheng; Shen, Chunhua; van den Hengel, Anton, et al. Exploring Context with Deep Structured models for Semantic Segmentation. *CoRR*. 2016.

[128] Zhou, Bolei; Khosla, Aditya; Lapedriza, A. Gata, et al. Learning Deep Features for Discriminative Localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA. 2016. 2921-2929

[129] Zhou, Bolei; Khosla, Aditya; Lapedriza, A. Gata, et al. Places: An Image Database for Deep Scene Understanding. *CoRR*. 2016.

[130] Zhou, Bolei; Zhao, Hang; Puig, Xavier, et al. Semantic Understanding of Scenes through the ADE20K Dataset. *CoRR*. 2016.

[131] Zhou, Yisu; Hu, Xiaolin; Zhang, Bo. Interlinked Convolutional Neural Networks for Face Parsing. In *International Symposium on Neural Networks (ISSN)*. Jeju, South Korea. 2015. 222-231

[132] Yamashita, Takayoshi; Nakamura, Takaya; Fukui, Hiroshi, et al. Cost-alleviative Learning for Deep Convolutional Neural Network-based Facial Part Labeling. *IEEE Transactions on Computer Vision and Applications*. 2015. 7(99-103.

[133] Tang, Wei; Huang, Yongzhen; Wang, Liang. 1000 Fps Highly Accurate Eye Detection with Stacked Denoising Autoencoder. In *Chinese Conference on Computer Vision (CCCV)*. Xi'an, China. 2015. 237-246

[134] Luo, Ping; Wang, Xiaogang; Tang, Xiaoou. Hierarchical face parsing via deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. Providence, RI, USA. 2012.

2480-2487

[135] Guo, Dong; Sim, Terence. Digital face makeup by example. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Miami, Florida, USA. 2009. 73-79

[136] Tong, Wai Shun; Tang, Chi Keung; Brown, Michael S., et al. Example-Based Cosmetic Transfer. In *Proceedings of the Pacific Conference on Computer Graphics and Applications*. Maui, Hawaii, USA. 2007. 211-218

[137] Liu, Si; Ou, Xinyu; Qian, Ruihe, et al. Makeup Like a Superstar: Deep Localized Makeup Transfer Network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*. New York, NY, USA. 2016. 2568-2575

[138] Liu, Luoqi; Xing, Junliang; Liu, Si, et al. Wow! You Are So Beautiful Today!. *ACM Transactions on Multimedia Computing, Communications and Applications*. 2014. 11(1s):20-21.

[139] Taigman, Yaniv; Yang, Ming; Ranzato, Marc'aurelio, et al. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2014. 1701-1708

[140] Sun, Yi; Wang, Xiaogang; Tang, Xiaoou. Deeply learned face representations are sparse, selective, and robust. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA. 2015. 2892-2900

[141] Liu, Si; Song, Zheng; Liu, Guangcan, et al. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Providence, RI, USA. 2012. 3330-3337

[142] Liu, Si; Yan, Shuicheng; Zhang, Tianzhu, et al. Weakly Supervised Graph Propagation Towards Collective Image Parsing. *IEEE Transactions on Multimedia*. 2012. 14(2):361-373.

[143] Liang, Xiaodan; Liu, Si; Shen, Xiaohui, et al. Deep Human Parsing with Active Template Regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. 2015. 37(12):2402-2414.

[144] Datar, Mayur; Immorlica, Nicole; Indyk, Piotr, et al. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the 20th ACM Symposium on Computational Geometry (SCG)*. Brooklyn, New York, USA. 2004. 253-262

[145] Chum, Ondrej; Philbin, James; Zisserman, Andrew. Near Duplicate Image Detection: min-Hash and tf-idf Weighting. In *Proceedings of the British Machine Vision Conference (BMVC)*. Leeds, Brtish. 2008. 1-10

[146] Weiss, Yair; Torralba, Antonio; Fergus, Robert. Spectral Hashing. In *Advances in Neural Information Processing Systems (NIPS)*. Vancouver, British Columbia, Canada. 2008. 1753-1760

[147] Lin, Ruei Sung; Ross, David A.; Yagnik, Jay. SPEC hashing: Similarity preserving algorithm for entropy-based coding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. San Francisco, CA, USA. 2010. 848-854

[148] Kulis, Brian; Darrell, Trevor. Learning to Hash with Binary Reconstructive Embeddings. In *Advances in Neural Information Processing Systems (NIPS)*. Vancouver, British Columbia, Canada. 2009. 1042-1050

[149] Zhang, Dell; Wang, Jun; Cai, Deng, et al. Self-taught hashing for fast similarity search. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. Geneva, Switzerland. 2010. 18-25

[150] Gong, Yunchao; Wang, Liwei; Guo, Ruiqi, et al. Multi-scale Orderless Pooling of Deep Convolutional Activation Features. In *European Conference on Computer Vision*. 2014. 392-407

[151] Babenko, Artem; Slesarev, Anton; Chigorin, Alexander, et al. Neural Codes for Image Retrieval. In *European Conference on Computer Vision*. 2014. 584-599

[152] Wan, Ji; Wang, Dayong; Hoi, Steven Chu Hong, et al. Deep Learning for Content-Based Image Retrieval: A Comprehensive Study. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*. Orlando, FL, USA. 2014. 157-166

[153] Ng, Joe Yue Hei; Yang, Fan; Davis, Larry S. Exploiting local features from deep networks for image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2015. 53-61

[154] Ou, Xinyu; Yan, Lingyu; Ling, Hefei, et al. Inductive Transfer Deep Hashing for Image Retrieval. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*. Orlando, FL, USA. 2014. 969-972

[155] Chechik, Gal; Sharma, Varun; Shalit, Uri, et al. Large Scale Online Learning of Image Similarity Through Ranking. *Pattern Recognition and Image Analysis*. 2009. 11-14.

研　究　生　签　字 _____

指　导　教　师　签　字 _____

院（系、所）领导签字 _____

年　　月　　日