分 类 号 _____ 学号 D201377750
学校代码 10487 密级 _____

华中科技大学

博士学位论文

基于深度学习和上下文语义的视觉内容识别与分析研究

学位申请人： 欧新宇
学 科 专 业 ： 计算机应用技术
指 导 教 师 ： 凌贺飞 教授
答 辩 日 期 ： 2017 年 5 月 5 日

A Dissertation Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in Engineering

Research on Visual Content Recognition and Analysis based on Deep Learning and Context Semantics

Ph.D. Candidate : Xinyu Ou

Major : Computer Application Technology

Supervisor : Prof. Hefei Ling

Huazhong University of Science and Technology

Wuhan 430074, P.R. China

May 2017

独创性声明

本人声明所呈交的学位论文是我个人在导师的指导下进行的研究工作及取得的研究成果。尽我所知，除文中已标明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本论文属于　保密□，　在_____年解密后适用本授权书

不保密□。

（请在以上相应方框内打“√”）

学位论文作者签名：　　指导教师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

摘 要

随着互联网技术的飞速进步以及深度学习展现出强大的性能，基于图像和视频的各种应用也得到了前所未有的发展。然而，伴随着这些应用给日常生活带来便利的同时，也给社会带来了许多潜在的负面影响。因此，如何高效、准确地从这些纷繁复杂的海量数据中甄别出有用的信息和过滤有害的信息，已经是大数据环境下亟待解决的问题。随着深度学习的发展，计算机视觉任务的应用领域也得到了空前的扩展，包括：图像分类、目标识别、目标检测、图像分割、对象跟踪等。

本文将在深度学习的框架下，以四个计算机视觉的典型应用为基础，通过结合多种不同的上下文关系，开展面向大数据的视觉内容的识别与分析研究。这四个任务分别是：成人内容识别、特定图像检索、自然场景解析和人像妆容迁移。

首先，针对成人内容识别任务中类别空间稀少和正负样本空间内样本多样化导致的分类难的问题，提出基于高层语义的细到粗策略和基于多上下文混合建模的联合决策方案。传统成人内容识别通常都是二分类问题（“是成人”或“不是成人”），而复杂的样本会导致部分样本类内距大于类间距，增大分类器训练的困难。本文提出的细到粗策略，通过在训练中细化类别来改善分类器的性能。此外，通过全局上下文、局部上下文和跨上下文等多种上下文建模方式，从不同的角度去理解样本，最大限度地解决样本多样化问题。与传统特征融合方式不同，策略融合并不直接融合特征，它在最大限度保证基于分类的全局上下文准确性的同时，利用基于检测的局部上下文信息生成置信度较高的决策来尽力修正被误判的样本，从而实现召回率和准确率的同时提高。此外，模块化的设计方案，允许通过更新全局上下文建模或局部上下文建模实现整个网络性能的提升。

其次，针对场景解析任务中对象尺度较小、交互性多（遮挡）、隐藏性强（易湮没于复杂的背景中）等特性带来的对象识别困难的问题，提出一种基于深度学习的对象区域增强网络。该网络集成了针对任务设计的两个核心模块：对象区域增强策略和黑洞填充策略。前者将检测到的语义置信度较高的对象区域直接对应到卷积特征图

的特定类别通道上的局部区域，并通过加权特征来改进上下文关系，完成对困难对象区域的识别；后者通过屏蔽额外背景类来避免解析网络将部分困难区域判定为额外背景类的错误。此外，模块化的设计方案使模型不但可以通过更换模块实现整体解析性能的提升，还可以将两个策略应用到其他现有的场景解析网络中。

然后，针对以人脸解析为基础的典型应用—妆容迁移中的两个难点问题：（1）如何获得精确的人脸解析结果；（2）如何按需保持（如：脸型、五官）和迁移（如：唇彩、眼影）人像的特征，提出了对称加权交叉熵损失和深度局部妆容迁移网络。前者对特定的局部上下文区域进行加权，并强制对眼影、嘴唇等特殊区域进行对称性约束；后者利用不同类型的特征分别描述形状敏感和纹理敏感两种局部区域，最后通过迭代算法逐渐将局部妆容特征从参考人像迁移到未化妆的人像上。端到端的生成网络，不但可以产生自然的妆容迁移效果，还可以实现妆容浓淡程度的自由调节，这使得该系统的可用性大大增强。

最后，针对大数据环境下图像检索效率和性能的问题，提出一种基于深度学习的层次化深度语义哈希方案。该网络可以端到端地同时输出样本的高层语义和哈希编码。通过基于概率的语义级相似性和哈希级相似性的融合相似性计算方案，首先利用几乎零开销的高层语义信息过滤大量语义不相关的样本，然后再利用哈希编码在小很多的候选建议集中完成相似性检索。该方案在百万级的 *Imagenet* 数据集上，可以保证在检索性能不降低的前提下，实现大约 150 倍的速度提升。

综上所述，本文所研究的多种上下文语义融合策略，不但在计算机视觉的理论层面具有一定的参考价值，更关键的是本文的研究对于设计和开发鲁棒、实用的应用系统也具有一定的借鉴意义。

关键词：深度学习；层次化语义；多上下文语义；图像识别；图像检索；场景解析

Abstract

With the rapid development of Internet technology and the great performance of deep learning, image and video applications also got significant evolution. However, accompanied with the convenience these applications, it has brought some negative influence on the society. Therefore, how to identify useful information and filer the harmful information effectively and exactly from the massive and complex data sea, are realistic problem that needs to be solved in big-data environment. With the development of deep learning, the application fields of computer vision has been expanding rapidly, including image classification, object recognition, object detection, image segmentation, tracking, etc. This paper will aim at four typical applications, such as, adult content identification, scene parsing, makeup transfer and content based image retrieval. These works are based on the deep learning framework and integrate the hierarchical context and multi-context semantic information.

To solve the hard problem of classification caused by the diverse samples, this paper proposes a high-level semantics based fine-to-coarse strategy and multi-context semantics based joint decision strategy. The adult context recognition is usually a binary classification problem, but complex samples will result in the intra-class distance maybe larger than the inter-class distance for some images, which increase the difficulty of training a classifier. The fine-to-coarse strategy improves the performance of the classifier by refining the categories in training. In addition, the diversity problem can be relaxed by multi-context modeling, which consists of global-context modeling, local-context modeling and cross-context modeling. Different from traditional feature fusion, policy fusion not combines the features directly. It is designed to ensure the accuracy that is produced by classification based global-context modeling, and uses detection based local-context modeling to fix wrongly discriminating samples. This strategy can improve the recall and precision simultaneously. Moreover, the modular design allows to improve overall system by upgrading the individual global-context modeling component or local-context modeling component.

To solve the difficulty of scene parsing caused by the hard object that involves object

scale (too small), interactive (occlusion), and hiddenness (easy obliterated in complex background region), this paper proposes a deep objectness region enhancement network (OENet). This network includes two core components: objectness region enhancement network and black-hole filling. The former uses the high confidence proposal regions to weight the areas of the specific channel of convolutional feature maps. The latter is used to avoid pixels are judged to nonexistent category by shielding extra background. In addition, the modular design makes the two modules not only can be updated by replacing a high performance one, but also can be applied to other existing scene-parsing network.

Makeup transfer is a very interesting work. It starts with face parsing, and uses generative network to produce a natural-looking makeup. To solve two challenge problems: (1) how to get a precise face parsing, (2) how to keep (facial sharp and features) and transfer (lip-gloss and eye shadow) the feature of human on demand, this paper proposes a weighted cross-entropy loss and a deep localized makeup transfer network. The former is used for weighting specific local-context regions, and enforces the symmetric prior on some special areas, such as eyes and lips. The latter uses different feature to describe sharp-sensitive region and texture-sensitive region respectively. This generation network not only produces natural-looling makeup, but also controls the lightness of the makeup.

To solve the problems of precision and efficiency in large-scale image retrieval, this paper proposes a hierarchical deep semantic hashing scheme. This network can produce high-level semantic and hash codes simultaneously. With probability-based semantic-level similarity and hashing-level similarity, the unrelated samples are filtered in advance by zero-cost high-level semantic information, and then the retrieval is achieved in a small candidate proposal set with hash codes. This scheme can achieve accelerating about 150 times with similar accuracy in imagenet dataset.

In summary, the multi-context semantic fusion strategy and the deep learning methods are discussed in this paper. They not only have reference value, but also have reference meaning in design, development a practicable and robust application system.

Key Words：Deep learning；Hierarchical semantic；Multi-context semantic；Image recognition；Image retrieval；Scene parsing

目 录

插图目录

表格目录

1 绪论

1.1 课题来源

本论文的研究主要来源于以下科研项目：

1. 国家自然科学基金—联合基金重点项目：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 研究背景与意义

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，它使各种应用中获取到资源更加丰富，形式更加多样化，这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，这些方法为解决上述问题提供了有效途径，也为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。

本论文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。

本论文的研究非常具有挑战性，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 视觉内容上下文语义的定义和分类

在介绍基于上下文语义的视觉内容识别与分析前，首先要明确几个问题，即：什么是视觉内容的语义理解、什么是视觉内容的上下文、上下文包含的主要类型以及上下文的相互关系有哪些。

1.3.1 视觉内容的语义理解

视觉内容语义理解是指以图像或视频[1]为研究目标，利用人工智能的方法完成对图像内容的语义解释，通过研究图像中包含的对象、对象间的关系、对象所处的场景以及对象的行为，使计算机自动地获取或图像所蕴含的信息和知识，并可以用语义对图像内容进行描述。人类的视觉系统可以轻松、快速地获得图像所表达的所有信息，但对于计算机来说，这却是一个极具挑战的问题。

图 1.1 给出了一个视觉内容多种层次的语义理解范例。在计算机视觉识别的第一层次分类任务上，计算机会通过图像的前景主体目标对图像进行分类，得到这是一幅关于“人”的图像这个高级语义；在第二层次检测任务上，计算机可以发现图像包含两个“人”和一个“足球”等关于图像内容更详细的信息；在第三层次的场景解析任务上，计算机可以从像素级的识别中得到诸如“草地”、“观众群”和“广告牌”等更多的背景信息。当然，更高层次的视觉语义理解，还包括去理解场景更深层次的内容，

[1] 为了简便除特殊指明，本文后续内容将使用“图像”替代所有有关“图像和视频”相关的描述。

得到如“两个运动员正在足球场上进行一场足球比赛”这样具备交互性和总结性知识的语义。从另一个层面上看，还可以得到有关对象属性或部件的局部语义信息，例如：“9 号运动员的黄色球衣”、“2 号运动员的白色球袜”、“绿色的草地”等。

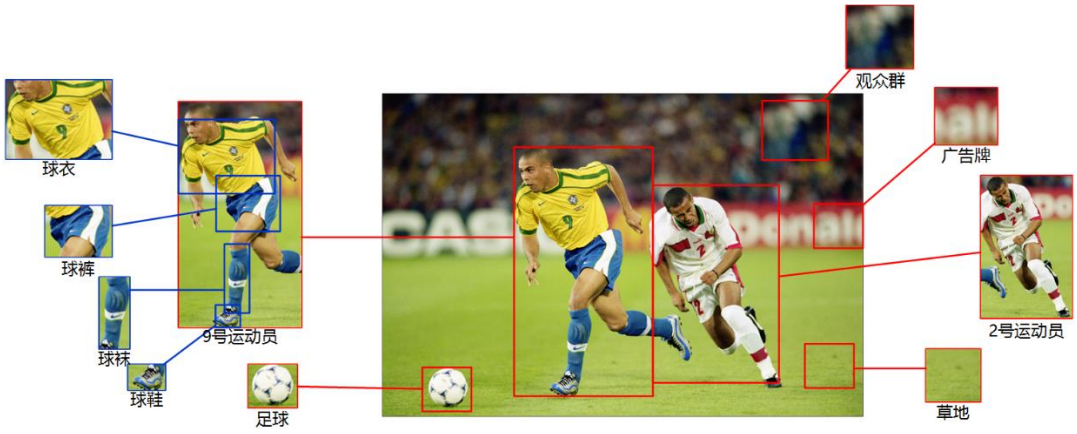
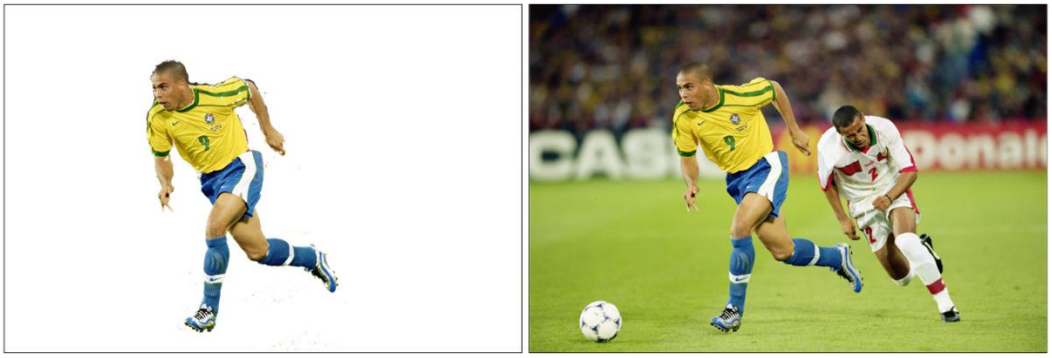


图 1.1 视觉内容多种层次的语义理解

1.3.2 视觉内容的上下文

根据视觉内容语义理解，可以发现对象并不是孤立地存在场景中。而那些能够直接或间接影响某个视觉对象感知与识别的信息则可以被称为上下文信息。比如，在图 1.1 中，关于“正在参加足球比赛的 9 号运动员”这条语义信息相关的信息就包括：交互对象（“2 号运动员”，“足球”等）、场景信息（“足球场”、“草地”、“观众群”等）、部件信息（“球衣”、“球裤”等）。这些与目标视觉内容存在一定共生、包容和位置关系的元素对于理解视觉内容都具有重要的意义。

如图 1.2（a）所示，在缺少上下文信息的时候， 仅依靠一个单一目标很难准确地判断对象的真实行为，也很难很好地理解整幅图像所要表达的信息。对于图 1.2（a）中的对象，可以将其理解为在参加足球赛、篮球赛，或者是跑步和游戏等不确定的信息；而从图 1.2（b）中，根据丰富的上下文信息，比如：交互关系（追逐的人）、其他对象（足球）、场景信息（足球场）等，则可以准确地给出“9 号运动员正在进行一场激烈的足球赛”这个答案。



(a)缺少上下文信息　　(b)包含上下文信息

图 1.2 上下文对理解视觉内容的重要性

1.3.3 上下文的主要类型

基于上下文的视觉理解是人类认知事物的基本方法，而上下文的分类也自然依赖于人的认知过程。借助于认知心理学的研究，人们发现人类利用视觉认知世界的过程是一个从全局到局部，从简单到复杂的层次化过程。前者可以认为是一个横向的层次化认知过程，它可以划分为：全局上下文和局部上下文；后者是一个纵向的层次化认知过程，它是一个从粗到细，逐层提炼的过程，可以称为层次化上文。它们的具体描述如下：

1. 全局上下文

全局上下文通常包含整个场景的全局信息，是对图像的整体描述。对于一个具体的视觉对象来说，全局上下文刻画了它与全局场景、场景内所有对象以及背景的依赖关系。一方面，它暗示了视觉对象之间的内在联系和交互信息；另一方面，不同的全局场景信息也暗示了不同视觉对象出现的可能性，及其出现的位置和方式。例如：一个港口场景，意味着可能会存在大量的轮船，并且大量的船员正在努力地将货物搬运到甲板；而一个运动场，则暗示了可能会有大量的观众在观看运动员的足球比赛。

2. 局部上下文

局部上下文通常描述的是视觉对象间更具体的相互关系，以及场景中视觉对象自身的局部关系。视觉对象间更具体的相互关系是指视觉对象和其他视觉对象或其邻域像素间的相互关系，在语义层面上可以理解为尺度关系、位置关系、交互关系和共生关系等。视觉对象自身的局部关系主要是指对象内部的部件间的尺度关系、依存关系、位置关系和相互关系等。

3. 层次化上下文

层次化上下文可以是自底向上的认知过程，也可以是自顶向下的认知过程。在视觉语义的层面上，通常是按照自然界物质的类别的固有的层次关系来定义，例如：生物分类学的七层关系（界门纲目科属种），对象的从属关系（整体—部件）等。

4. 上下文的相互关系

从认知心理学[5]的角度，视觉内容上下文的相互关系通常可以分为：（1）位置关系，描述视觉内容在场景中出现的位置以及内容间的相对位置关系；（2）从属关系，描述视觉内容的层次化信息；（3）尺度关系，刻画视觉内容受视角、远近等因素影响带来的尺度变化；（4）干涉关系，描述视觉内容与视觉内容之间和视觉内容与背景间的相互干扰关系，例如遮挡等。上下文的这几种关系影响着计算机对视觉内容的识别和分析，但它们也提供了除外观之外更丰富的信息与支持，对于表征视觉内容有积极的意义，可以有效地辅助视觉内容的识别与分析。

1.4 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。下面本文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。

1.4.1 基于上下文语义的视觉内容识别

1. 视觉语义理解概述

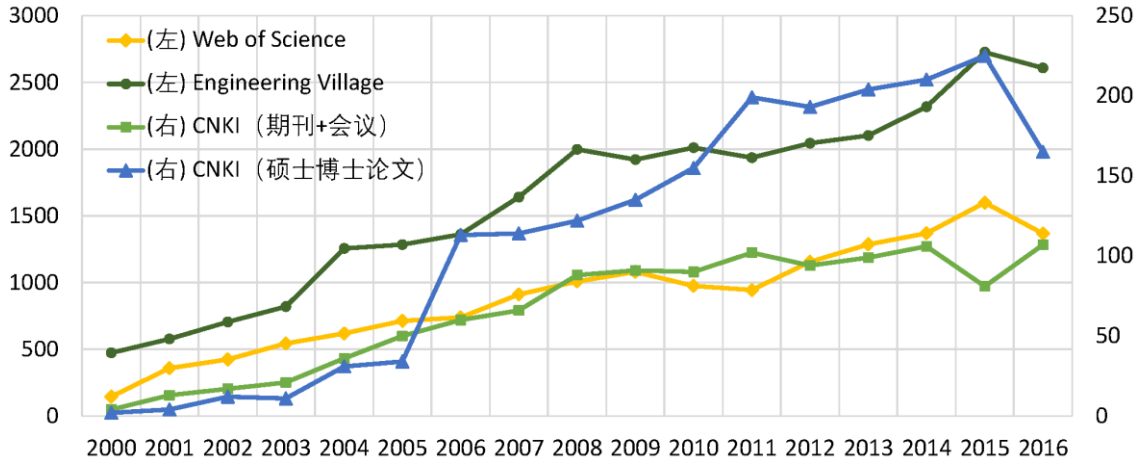


图 1.3 2000 年以来与“图像+语义”相关的文件统计图

本文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 1.3 中，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。

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，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、斯坦福大学、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[6]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[7]（Bag of Feature，BoF）的模型和费舍尔核[8-10]（Fisher Kernel，FK）。BoF 利用 K-Means 等聚类算法对图像局部特征（如 SIFT[11, 12]，HOG[13]和 LBP[14]）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[15]、稀疏编码[16]和高斯混合模型[17]）完成编码、量化，然后利用金字塔匹配模型（SPM）[18]、VLAD[19]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[20]通过组合生成方法和判别方法来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, 21-24]（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[25]，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[26, 27]（Costellation Model）、视觉词袋模型[28-30]（Bag of Feature，BoW）、空间金字塔模型[18, 31, 32]（Spatial Pyramid）和部件模型[1, 33, 34]（Part Model）。

星座模型[26, 27]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[28-30]将图像描述成视觉单词的无序集合，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[35-37]和全局场景[38]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[18, 31, 32]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[31]（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, 33, 34]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 等人[33]提出了基于部件的 R-CNN[39]，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[40-46]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—局部”上下建模是一种最通用的方法。Ciresan 等人[43]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[47]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[48]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[49]（Imagenet Large Scale Visual Recognition Chellenge，ILSVRC）的参赛队伍[40, 44, 46, 50]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[51, 52]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。此外，Liu 等人[53-55]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。

“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[56]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[57]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[28-30]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[39]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。

“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[45]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[42]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，但是这并不是最好的解决方法。因此，本文提出了一系列基于策略的多层上下文语义的建模方法，该方法并不直接将不同的特征进行简单的融合，而是充分考虑特征的独特性，在尽量保留单一特征的优势的基础上，利用不同特征的互补性来修正单一特征的性能缺陷。第 2 章在细到粗的语义筛选策略的基础上，实现局部上下文、全局上下文和跨上下文等三种不同视角的上下文联合决策的方法来识别特定类别的样本；第 3 章保留了全局上下文的场景解析结果，并利用带语义的局部上下文信息来加强局部对象的特征强度，从而提高场景解析的能力；第 4 章将一幅参考人像图片上的局部上下文信息当做一种个性化的特征来学习，并将其迁移到另外一幅未化妆人像图片的对应区域中，在迁移的时候充分考虑旧人像局部上下文的特点和新人像原有的全局和局部上下文的特点，使得最后生成的妆后人像能够按需获得原来两幅人像图像所有个性化的语义特征；第 5 章提出了一种基于高层语义的过滤策略，该方法缩小了搜索空间，大大提高了检索速度。

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*[58]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*[58]的层次化语义进行组织，例如：大规模图像数据集 *Imagenet*[49]和 *TinyImage*[59]。很多工作[59, 60]表明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[61]。层次化技术也被应用到其他的视觉任务[62, 63]中。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[64]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 等人提出一种基于多分辨率的 HoG 特征提取方法 PHoG[31]，它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 等人[65]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[66-69]中。Eigen 等人[68]首先使用 FCN[70]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[67]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。

本文第 5 章的基于层次化语义哈希的图像检索中，也使用了粗到细策略实现初步过滤，从而加速检索。另一方面，本文提出了一种细到粗的层次化方法，利用高级语义的层次化关系，首先将样本按照细粒度的类别进行分类，之后再依据语义之间的层次化关系，将细粒度类别映射成粗粒度类别。这种方法非常适合于本文第 2 章的

工作，后续的章节中将详细介绍如何利用这种细到粗的策略实现成人图像的识别。

1.4.2 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 等人[71]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[72]（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[28-30]的模型。它利用人工进行特征提取（如：SIFT[11, 12]，HOG[13]和 LBP[14]），并使用一定的编码方案（如：矢量量化[15]、稀疏编码[16]和高斯混合模型[17]）完成编码，最后用金字塔匹配模型[18]（SPM）、VLAD[19]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[73]挑战赛，Krizhevsky 等人[23]将 *Imagenet* 数据集[74]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%[24]，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[75, 76]，人脸识别[77-79]，自然图像分类[21, 24, 80]和手写字体识别[75, 81]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, 21]，宽度的增大[4, 82, 83]和使用更小的步长[2-4, 80, 83]），新的线性激活单元[82, 84-88]的使用和复杂层的设计[4, 24, 65]。另一方面，有效的正则化技术[10, 81, 85, 89]，积极的数据扩展技术[3, 4, 23, 82]和大规模的已标记的数据集[74, 90, 91]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[13]特征和 SIFT[11, 12]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[92]或者一个带核的 SVM[93]。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[94]也取得了较好的成绩。

近两年来，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, 21-24]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[93]、BING[95]、Objectness[96]、MCG[97]、EdgeBoxes[98]、DeepBox[99]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[39, 65, 100, 101]框架和基于回归的方法 SSD[102]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[39]是一个非常成功的检测算法，它首先利用 Krizhevsky 所设计的 CNN 模型 AlexNet[23]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[93]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[65]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[100]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[101]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[102]在 YOLO[103]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 系列算法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[104]（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[105]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语义信息，还考虑了底层纹理特征，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[106, 107]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[108, 109]（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[110]和结构化预测技术[111, 112]，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底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[93]或遮罩算法 MCG[97]生成的候选区域引入到分割网络的 RCNN[39]和 SDS[113]。类似的，Mostajabi 等人[114]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，第二流派直接利用卷积特征图来生成区域。Farabet 等人[115]使用卷积神经网络生成多种图像分辨率。Hariharan 等人[116]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[117]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[70]（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[118]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[119]（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[118]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[120-127]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[118, 128-131]和人脸解析[132-135]也成为图像分割研究领域的重要目标。

场景解析[118, 128-131]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[136-139]可以认为是人脸识别[77-79, 140, 141]之后一个更高级的应用，它也是人体解析[142-144]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[11, 12]、GIST 描述符[25]和费舍尔向量[20]（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[145]（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[146]（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[147]（Spectral Hashing，SH）、基于熵编码的相似性保护算法[148]（SPEC Hashing）、二进制重构嵌入[149]（Binary Reconstructive Embeddings，BRE）算法、自学习哈希[150]（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, 21-24]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[151]（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[152]（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[153]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[154]的工作受 MOP[151]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[155]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。

1.5 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。以本文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，这种现象称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。这些对象通常被称为难目标，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[2]，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

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。例如本文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

[2] 此处，本文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。

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。

1.6 研究内容与目标

本文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图像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。

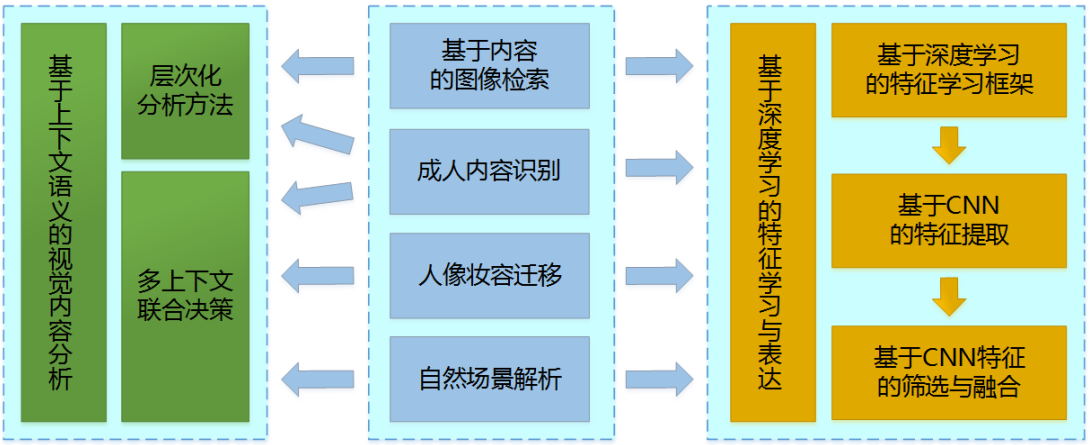


图 1.4 研究内容框架图

图 1.4 展示了本文的主要研究内容框架图。从构建特征学习框架到实现基于 CNN 的特征提取，再到对这些特征的筛选和融合，四个典型的应用都在深度学习的框架下进行研究。如何找到和利用强大的 CNN 深层特征是本文研究的四个任务的基本前提。此外，两种基于上下文的分析方法分别被应用到四个应用中，其中基于内容的图像检索和成人内容识别使用层次化的上下文分析方法；成人内容识别、人像妆容迁移和自然场景解析使用了多上下文联合决策的方法。具体研究内容在本节的后续部分进行详细介绍。

1.6.1 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学习框架研究

构建能够承载大数据的深度学习特征提取框架，是本论文的基础部分，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，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，可以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 的特征提取研究

通过完整的特征提取框架，可以实现全局特征和局部特征的同时获取。全局特征主要反映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；局部特征反映的通常是对象的局部信息，或者特写信息。借助于典型的深度学习方法——卷积神经网络（CNN），可以获得比传统手工特征更鲁棒的特征表达。然而，如何获得最优全局和局部特征，并不是自然而然的过程，这个问题是本文需要重点考虑的。

3. 基于 CNN 特征的筛选与融合研究

在获取了样本大量不同类型的特征后，接下来需要考虑的是如何合理利用这些特征。随着对特征的进一步研究，可以知道，单一的特征通常不能很好地表征样本，

特别是那些复杂的样本；同时，并不是所有的特征对识别和分析样本都是有积极意义的。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，是本文需要重点研究的内容。

1.6.2 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具有很大的优势，它既可以获取样本低层的粗糙特征和中层特征，也可以获取样本的高层的语义信息。丰富的特征和语义信息，为处理视觉内容提供了极大的方便，同时使得通过利用这些不同级别的信息来改进各种计算机视觉任务成为可能。本文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或多种上下文信息的融合来获取最终的信息。

1. 层次化分析方法

层次化分析方法的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，迁移到另一个较为容易的问题空间。

对于成人内容识别任务，通常是一个“是或不是”的二分类问题。该任务具有典型的类别空间狭窄的特性。然而，对于每个类别来说，样本确是复杂多样的（可能包括肖像、全身图、器官特写、猫、狗、桌子等多种类别）。因此，不同子类别之间可能存在较大的相似性，要直接将所有的样本都分为两个类别是一个很困难的问题。**如何处理类别空间较少和样本复杂性的冲突，是本文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**

检索时间是本文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，是本文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文研究的重点。**

1.7 论文组织结构

在论文组织结构上，全文共分为 6 章，为了便于理解全文总体框架和行文思路，图 1.5 给出了各章的组织结构图，具体描述如下：

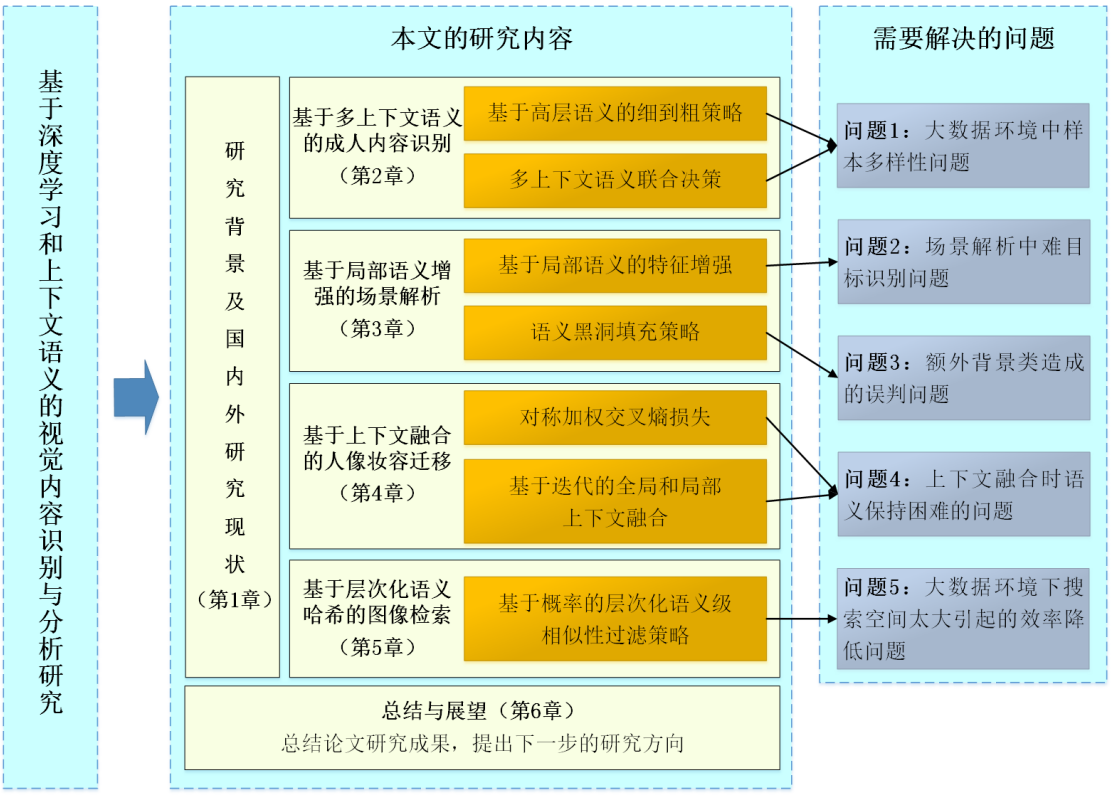


图 1.5 论文组织结构图

第 1 章：绪论。从本文的研究背景和意义出发，顺序介绍了视觉内容上下文语义的定义和分类、基于上下文语义的视觉内容识别和基于深度学习的视觉内容识别的国内外研究现状，然后提出了当前视觉内容识别存在的问题，也即本文需要重点解决的问题，基于这些问题提出了本文主要的研究内容和创新点。

第 2 章：基于多上下文语义的成人内容识别。本文以成人内容识别为例，重点分析了成人内容识别与传统内容识别的差异性，提出了一种针对性的解决方案深度多上下文网络（Deep Multi-Context Network，DMCNet），通过对全局上下文决策和局部上下文决策的整合实现成人内容的准确识别。同时，提出细到粗的层次化语义过滤策略方法，用来解决成人内容识别中，类别稀少带来的分类困难的问题。

第 3 章：基于局部语义增强的场景解析。为了解决场景解析任务中，一些对象因为湮没在复杂背景中而无法解析出来的问题，提出了一种对象区域增强网络（Objecteness Region Enhancement Network，OENet）。该网络通过检测发现一些特定

的对象区域，并将这些区域应用到特定的卷积特征图中，并通过增强这些区域的特征强度，实现丢失对象的召回。

第 4 章：基于上下文融合的人像妆容迁移。本文研究的内容是在人像语义分割的基础上，将人像局部区域的特征迁移到新的人像上，实现妆容的迁移。基于这个思路，本文提出一种新颖的深度局部妆容迁移网络（Deep Localized Makeup Transfer Network，DLMTN），用于实现妆容推荐、人像解析和妆容迁移。

第 5 章：基于层次化语言哈希的图像检索。提出一种基于概率的语义级相似性和哈希级相似性融合的相似性策略，用于解决大规模图像检索中检索精度和检索效率的问题。其中基于概率的语义级相似性用于过滤大量不相关的样本，实现缩小搜索空间；哈希级相似性用于快速计算两个样本间的距离。该方法在超大规模的数据集上验证了其精确度、效率和泛化性能。

第 6 章：总结与展望。对全文的研究工作做一个总结，并对未来的研究工作做进一步展望。

2 基于多上下文语义的成人内容识别

2.1 引言

Internet 作为全球信息中心，允许全世界所有的人自由地浏览、分享和交换他们的资源和信息。无数的网站都致力于提供各种各样的服务，例如图像、视频分享网站（Flickr，YouTube，QQ 空间等）和搜索引擎（谷歌，百度，Bing 等）。虽然互联网带来了便利，但是有害网站和非法内容仍然广泛存在，例如：成人内容。识别成人内容，对于净化互联网络空间，构建安全的互联网文化具有重要意义。但是，这也是一个非常具有挑战性的问题。

本章提出一种基于高层语义的细到粗策略和多上下文联合决策的深度多上下文网络（Deep Multi-Context Network，DMCNet）用于从大规模的样本集中识别出非法的成人图像和视频文件。本章的相关研究工作发表文献[156, 157]中。

2.2 问题描述

成人内容识别是图像识别的一个具体应用，由于任务的特殊性，它涉及到图像分类和对象检测两个方面的关键技术，在处理成人内容识别时，有几个问题是必须要考虑的：

首先，如何定义“成人”这个概念是首先要面对的重要问题。大多数国家将色情（淫秽）定义为“成人”并加入到他们的法律中，例如：美国法律[158]和中华人民共和国刑法[159]。这些法律一致认为淫秽信息主要是指在整体上宣扬淫秽行为、挑动人们性欲，导致普通人腐化、堕落的文字、图片、音频、视频等信息内容。然而，这些法律都没有严格地规定如何区分成人内容，以及如何对成人内容进行分级。这使得如何详细地划定和过滤敏感图像和视频成为很多网站管理者和多媒体管理员最为头痛的问题。虽然很多网站都明文规定禁止成人内容，然而它们并没有合适的系统去过滤和防止用户上传和分享涉及成人内容的文件。更糟糕的是，有的网站甚至主动提供成

人图像和视频以供用户访问。这些问题导致互联网上充斥着大量不优雅的视频和图像，严重影响了互联网的健康发展。

其次，由于互联网中的图像规模巨大，并且具有广泛的多样性，自动检测和识别成人内容也是非常具有挑战性的问题。（1）规模：对于任意一个网站，拥有数百万，甚至上亿的图片和视频并不是一件令人吃惊的事。（2）多样性：互联网上的图像在内容、尺度、分辨率和图像质量等方面都呈现明显的多样性。自然图像的内容可能会涉及从专业摄影到手机自拍，从人的交互行为到局部器官特写，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。面对如此海量和复杂的多媒体资源，任何网站都不可能完全依靠人工来进行识别和过滤。因此，全自动的识别成人视觉内容（图像/视频）变得非常重要。

2.3 基于高层语义的细到粗策略

利用层次化关系来分析数据，对于处理大规模的图像识别问题是具有重要意义的。“粗到细”的推导方法在很多视觉任务中被广泛使用，例如：图像检索[160]，对象检测[161]等。这个策略最大优点是它能够通过对类别空间的层次化处理，显著降低在大规模数据集上相似图像匹配的搜索空间，从而加速处理过程。对于成人内容识别任务，通常类别数量被严格限制，甚至只是一个二分类问题（“是成人”或“不是成人”）。然而，对于真实图片来说，每个类别中的图像仍然有较大的差异。例如，成人图像可能包含诸如裸体图、性器官、性行为等种类，而正常图像更是可能包含丰富的对象和场景，例如：猫，人，蛋糕，汽车或者游泳馆。将这些复杂的概念直接分配到两个类别中是非常困难的。因此，首先将样本通过分类器分配到一个较大的细粒度的类别空间，然后再依据类别间语义的层次化关系，将样本的细粒度语义类别映射到一个较小的粗粒度的类别空间，是解决这个难题的有效方法之一。本文将这个方法称为“基于高层语义的细到粗策略”。图 2.1 是本文提出的“基于高层语义的细到粗策略”的示意图。给定一个输入图像 I，它首先通过一系列的前向推导，被卷积神经网络分类到一个细粒度的类别空间，然后再根据高级语义间的对应关系被映射到一个粗粒度的

类别空间。需要注意的是，这里的细粒度类别空间和粗粒度类别空间的对应关系按照高级语义关系进行事先规划。

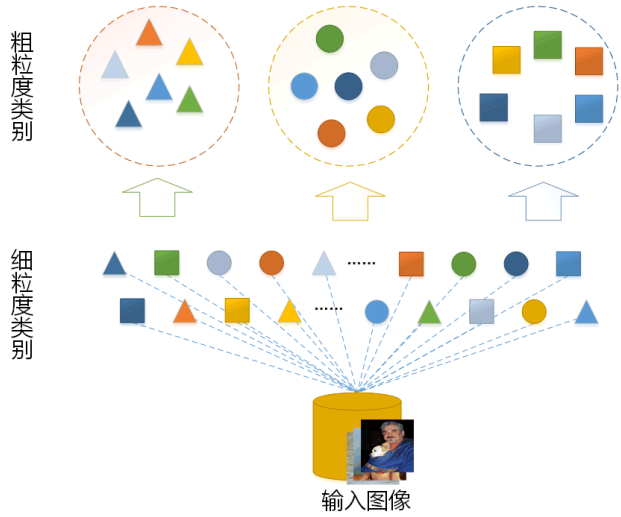


图 2.1 基于高层语义的细到粗策略

下面将给出这个算法的公式化形式。假设 $y=(y_c,\ y_f)$ 是输入图像 I 的标签集。图像 I 的粗粒度类别可以表示为 $y_c,(y_c \in C,\ c=(1,2,\ldots,M))$，同时 I 的细粒度类别可以表示为$y_f,(y_f \in F,\ f=(1,2,\ldots,N))$。此处，$M$ 和 N 分别表示粗粒度类别 C 和细粒度类别 F 中的类别总数。因此，在高层语义上，F 是属于 C 的一个子集，因此可以用$y=(y_c,\ y_f)$表示图像 I 同时属于粗粒度类别y_c和细粒度类别y_f。而粗粒度类别y_c和细粒度类别y_f之间的关系根据语义间的层次关系被事先指定。例如：细粒度的类别“猫”、“狗”属于粗粒度类别“正常图像”，细粒度类别“裸体”、“性器官”属于粗粒度类别“成人图像”。具体来说，在本文构建的 *Sensitive* 数据集中，可以将正常图像的 997 个细粒度的类别（995 个类由 *Imagenet* 数据集定义，2 个类由 *Sensitive* 数据集定义）组合成第一个粗粒度类别，称为“正常图像”；然后将 9 个色情和裸体的细粒度类别（例如裸体、性器官、性行为等）组合成第二个粗粒度类别，称为“成人图像”；另外的 3 个边缘类（例如内衣、泳装、腿模等）组合成第三个粗粒度类别，

称为“少儿不宜图像”。分类的推理过程是一个将图像 I 在 CNN 中所获得的分类概率最高的标签指定为预测类别的过程，可以用如下公式表示：

$$y_f = g(I) \tag{2.1}$$

$$y_c = T(y_f) \tag{2.2}$$

其中，$T: y_f \mapsto y_c$ 是一个映射函数，符号“$\mapsto$”表示直接将细粒度类别y_f映射为粗粒度类别y_c。函数$g(I)$是深度神经网络的前向推理结果，例如 CNN[3]或者 FasterRCNN[101]的输出类别。在本文的工作中，它由 DMCNet 生成。这个定义可以被扩展到特征级，用于将细粒度特征转换为粗粒度特征。第 2.4.4 节将使用这个扩展版的定义。图 2.5 给出了细到粗策略的评估结果。

2.4 基于多上下文联合的深度网络

2.4.1 全局上下文建模

图 2.2 给出了深度多上下文网络（Deep Multi-Context Network, DMCNet）的网络体系结构图。整个框架的上半部分（分支（1））是基于深度 Faster RCNN[101]构建的局部上下文建模（见第 2.4.2 节），用于实现局部对象的检测；框架的中间部分（分支（2））是跨上下文建模（见第 2.4.3 节），用于实现具有全局补偿的局部对象识别；框架的下半部分（分支（3））是基于深度 CNN 的全局上下文建模，用于实现全画幅的特征提取。在图 2.2 中，可视化了每一个全连接层和它们对应的维度，卷积层可以使用多种当前著名的模型替代，例如：Alexnet[23]、VGG16[3]和 GoogLeNet[4]。所有的三个分支都共享同一个卷积层框架（具有同样的参数，并且在 *Imagenet* 数据集上进行预训练，在 *Sensitive* 数据集上进行微调训练，推理阶段使用同一个卷积层框架输出的卷积特征图作为三条支路的输入）。一个实用的替代训练算法用于训练这些参数（如算法 2.1 所示）。在推理阶段，三个分支合并为一个统一的框架，用于生成特征并完成联合判决，从而实现成人内容识别。该图以 VGG16 网络为例。

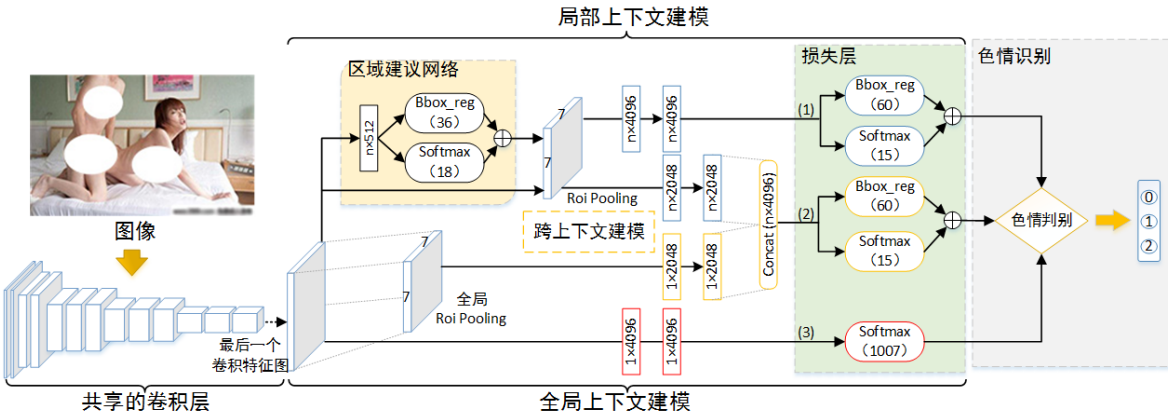


图 2.2 深度多上下文网络体系结构图

本文按照标准的训练过程[23]来训练图 2.2（分支（3））中的全局上下文建模网络。网络的输入采用固定尺寸（Alexnet[23]使用分辨率 227×227，VGG16[3]和 GoogLeNet[4]使用分辨率 224×224）并减去 RGB 值为[104，117，123]的均值图像；网络的输出是一个 1007 路的 *Softmax* 分类器，它的维度与 *Sensitive* 数据集的类别总数一致。该网络使用 *Imagenet* 预训练模型进行初始化，并在 *Sensitive* 数据集上进行微调训练。这个简单的初始 CNN 被用来初始化其他网络的卷积层。

受益于模块化的设计，DMCNet 可以很容易的集成当前最著名的一些深度卷积神经网络模型，例如：VGG16[3]、VGG19[3]、NIN[86]、GoogLeNet[4]和 ResNet[21, 22]等，本文的工作也使用了这些网络用来验证算法的扩展性。众所周知，越深的网络，通常具有更强的特征表达能力，这个结论在表 2.3 中得到验证。

2.4.2 局部上下文建模

当图 2.2（分支（3））的全局上下文建模致力于生成鲁棒的全局特征时，图 2.2（分支（1））的局部上下文建模则被设计用来发现细节的信息。局部上下文建模更关注于用一些较小的上下文区域来优化整个预测，例如一个局部器官或隐私部位。在本章中，一个重新实现的 Faster RCNN[101]模型被用于完成局部上下文建模。局部上下文建模由共享卷积层和两个主要的功能模块构成，一个是区域建议网络（Region

Proposal Network，RPN），一个是检测网络（Detection Network）。本章使用一个 5 层的 Alexnet 模型和 13 层 VGG16 模型构建共享卷积层，图 2.2（分支（1））可视化了基于 VGG16 模型构建的局部上下文建模。

区域建议网络使用卷积特征图作为输入，并输出一系列带对象性分数的长方形对象建议框。为了生成区域建议，DMCNet 在最后一个卷积层后构建了一个小型的网络，该网络使用 3×3 的卷积核。卷积运算的每一个滑动窗口都被映射到一个低维（Alexnet 使用 256 维，VGG16 使用 512 维）的特征向量，然后送入到一个双支路的全连接网络。一条支路用于建议框回归（*Bbox_reg*，36 维），另一条支路用于二进制分类（*Softmax*，18 维）。它们的维度由建议框生成锚点的个数决定（默认为 9）。与区域建议网络类似，检测网络同样被分离成两个分支层。但是，它们的维度由数据集的类别数决定（*Softmax*：15 维，*Bbox_reg*：60 维），第 2.5 节的数据集介绍部分将详细描述这些参数。

公式 2.3 使用一个多任务损失函数 L 来训练局部上下文网络用于类分类和建议框回归：

$$L(k,k^*,t,t^*)=\frac{1}{N_{cls}}L_{cls}(k,k^*)+\lambda\frac{1}{N_{reg}}k^*L_{reg}(t,t^*) \tag{2.3}$$

在公式 2.3 中，对于每一个兴趣区域 RoI，$k=(k_1,k_2,\dots,k_K)$ 是一个基于 K+1 个类的离散的概率分布，而$L_{cls}(k,k^*)=-\log p_{k^*}$是标准的基于两类（对象或非对象）的交叉熵损失。公式 2.3 的第二项k^*L_{reg}是针对类别k^*的建议框回归损失，当锚点（anchor）为正时被激活（Groundtruth 标签$k^*=1$）；反之，如果锚点为负时被抑制（$k^*=0$）。换句话说，只有分类正确的区域（与 Groundtruth 一致）才计算回归损失。此外，$t^*=\left(t_x^*,t_y^*,t_w^*,t_h^*\right)$ 表示类别k^*的 Groundtruth 边界框，而预测的对象边界框由 $t=\left(t_x,t_y,t_w,t_h\right)$ 表示。最后，$L_{reg}(t,t^*)=\sum_{i\in x,y,w,h}R(t_i^*-t_i)$，其中，$R(*)=smooth_{L_1}(x)=\begin{cases}0.5x^2, & if\ |x|<1\\ |x|-0.5, & otherwise\end{cases}$，是一个鲁棒的平滑$L_1$损失[100]函数，对于离群点它不像$L_2$损失[39]那么敏感。两个损失函数由参数$N_{cls}$和$N_{reg}$进行规范化，并由权

重参数λ进行平衡。其中N_{cls}是 mini-batch 的大小，N_{reg}是定位锚点的数量。例如，在 VGG16 模型中，$N_{cls}=64$，$N_{reg}=2400$。默认情况下，设置$\lambda=10$，因为对于区域建议网络，位置回归的重要性要高于类别判定。虽然其他设置可能会更适合训练区域建议网络，但是，作者认为这并不是决定性的参数。

在区域建议网络之后，是一个 RoI 池化层，它用于将 RPN 网络生成的区域特征映射成一个固定大小的特征图。将输出映射成一个固定尺度的向量，主要是为了兼容后续的全连接层的计算（例如，Alexnet 设置为 $H'=W'=6$，VGG16 设置为$H'=W'=7$）。RoI 池化层以 N 个特征图和 n 个感兴趣区域作为输入，通常$n \gg N$。N 个特征图由共享卷积层的最后一个卷积层获得，每一个特征图都是一个多维矩阵（$H \times W \times C$），其中 H，W 和 C 分别表示卷积特征图的行，列和通道数。对于每一个 RoI，RoI 池化层执行最大池化将区域内的特征映射到一个固定的空间尺度$H' \times W' \times C$（$H' \leq H$，$W' \leq W$）。

2.4.3 跨上下文建模

全局上下文基于全局外观的相似性进行学习，局部上下文则更关注区域特写。然而，由于拍摄角度和视点的多样性，人的局部区域或器官很容易被错误地识别。图 2.3 给出了四组容易混淆识别系统的图像范例，第一行是全尺寸图像，第二行是第一行原始图像的一个局部区域图，其中第一、二列是正常图像，第三、四列是成人图像。在该图中，第一、二列的图像很容易被识别系统判定为成人图像，因为它们的一些局部区域具有较强的二义性，局部上下文建模很容易输出较高的“色情”判别分数。然而，第三、四列的图像总是被识别系统判定为正常图像，因为主要人物占据了较大的范围，通过全局上下文建模将输出较高的“正常”判别分数。设计跨上下文建模主要是期望能够通过跨图像区域的特征融合从另外一个不同的角度来弱化这些判别冲突的影响。

图 2.3 四组易混淆图像的范例

如图 2.2（分支（2））所示，在最后一个卷积层后，卷积特征图的数据流被分为两个不同的分支。一个分支通过共享的卷积特征和区域建议架构实现局部上下文建模；另外一个分支，通过全局 RoI 池化将整幅图像的卷积特征映射到分类器，实现全局上下文建模。其中，全局 RoI 池化可以用现有的“RoI 池化层”[65]来实现。与传统的 RoI 池化相似，全局 RoI 池化同样也需要 RoI 区域作为对象建议区域去提取特征。在标准的检测网络中，RoI 区域利用 RPN 网络在整幅图像中进行区域回归得到，生成的 RoI 区域通常会标识一个具有语义信息的局部区域或局部对象；而对于全局 RoI 池化，整幅图像都将作为 RoI 区域。也就是说，整幅图像有且仅有一个 RoI 区域，而这个 RoI 区域将覆盖整个图像。全局 RoI 池化同样需要设置一个固定的H'和W'以兼容第一个全连接层，本文用与局部上下文建模同样的参数来设置全局 RoI 池化层。由图 2.2 可知，跨上下文特征将由原始的全局区域特征和所有建议区域特征组合而成，然后紧跟一个包含分类层和区域建议层的双支路网络。这个过程可以用下列公式表示：

$$\mathcal{F}_{cross}(n, D) = \mathcal{F}_{Local}(n, D_L) \oplus g(\mathcal{F}_{Local}(1, D_G), n) \tag{2.4}$$

其中，n 表示区域建议网络生成的建议区域的个数，D 是融合后的特征的维度，D_L和D_G分别表示局部上下文建模和全局上下文建模的全连接层的维度，其中$D_L = D_G = 2048$。此外，符号“$\oplus$”是一个连接操作，它将两个小的矩阵串联成一个大的矩阵。

函数$g(A,n)$用于实现矩阵复制，也就是说，给定一个二维矩阵 $\boldsymbol{V}$，函数$g(\boldsymbol{V},n)$将 $\boldsymbol{V}$ 转换为三维矩阵 $\boldsymbol{U}$。此处，矩阵 $\boldsymbol{U}$ 的前两维与矩阵 $\boldsymbol{V}$ 相同，第三维等于 n。换句话说，$\boldsymbol{U}$ 是一个 n 通道的 $\boldsymbol{V}$。这个操作可以将局部特征$\mathcal{F}_{Local}$转换为和全局特征$\mathcal{F}_{Global}$相同的结构，进而实现它们的串联操作。

2.4.4 多上下文决策和联合训练

在本小节中，将讨论多个模型的融合和决策问题。考虑以下三种多上下文决策的方法。

- ***策略 1***：一种直观而简单的方法是将多个不同级别的上下文模型组合成一个统一的框架，然后端到端地进行优化训练，所有的权重参数都在一个框架下同时进行优化学习。（统一优化）
- ***策略 2***：将多个不同且独立的上下文模型的输出概率组合起来，按照与类别相关的对应特征，计算它们的平均值或最大值生成新的与类别相关的融合概率以完成识别。（平均融合/最大融合）
- ***策略 3***：充分考虑不同上下文模型的异同点和互补性，使用层次化选择算法过滤非法样本。（策略融合）

端到端学习是计算机科学领域的一个经典理论方法，它最早由 Saltzer 等人[162]在 1981 年提出，并在最近几年随着 CNN 的发展被引入到了深度学习领域。它能够从输出层将误差反向传播到输入层，并同时更新所有层的参数。此外，它不需要为特征缓存而耗费存储空间，这不但节省了磁盘空间也加速了训练过程。然而，设计一个端到端的系统并不总是一件很容易的事。最重要的限制如文献[101]所描述，对于一些复杂的多功能网络一些梯度的传递过程很难被处理，例如：本文提出的 DMCNet。另外一个关键问题是，多通路的并发网络的复杂性和当前处理设备之间的矛盾。例如，同时训练三条支路的模型，会产生大量的参数（大约 350M），这远远超过了本文所使用的 GPU（NVIDIA Titan X 12G）的处理能力。因此，很不幸的是，***策略 1*** 并不能在实验中完成。

在过去的几年，很多种特别的融合方法被提出用于改进性能。一种简单但是通用

的方法是平均每一个支路的特征[75, 163]。此外，Lin 等人[164]提出了一种双线性融合方案，它们使用内积运算去将两个不同的深度特征转换为双线性特征。更复杂的是 Zhao 等人[165]提出的将两条用于处理不同上下文信息而独立训练的 CNN 的输出，合并后送入到一个二进制分类器中去实现显著性检测。受多栏 CNN[43]的启发，给定一些输入模式，***策略 2*** 可以将多个路径组合起来并通过简单的平均操作形成新的多上下文特征：

$$\mathcal{F}_{DMCN} = \frac{1}{N} \sum_{k=1}^{N=\#branches} \psi\big(\phi(\mathcal{F}_k)\big) \tag{2.5}$$

此处，使用 $k = 1,2,\dots,N$ 表示 DMCNet 由 N 个支路构成，第 k-th 条支路的特征用$\mathcal{F}_k$表示。函数 $\phi(*)$ 依据第 2.3 小节介绍的细到粗策略将细粒度特征映射成粗粒度特征。为了公平处理每一条支路，函数$\psi(*)$将特征归一化到区间[0, 1]。在本文中，支路的数量 N 固定设置为 3，特征$\mathcal{F}_k$，$(k = 1,2,3)$分别表示全局上下文建模，局部上下文建模和跨上下文建模的输出。这些特征的维度分别为：1007、15、15，也就是说这些支路的类别数分别为 1007、15、15（这将在第 2.5.2 节中介绍）。由于各条支路的特征维度不同，无法直接进行特征融合，因此映射函数$\phi(*)$变得至关重要，它可以实现将多个不同维度的特征变换成一致的维度。$\phi(\mathcal{F}_k)$的输出维度等于$\mathcal{F}_{DMCN}$的维度。如果维度等于 3，则表示 *L3* 识别；如果维度等于 2，则表示 *L2* 识别（细节见第 2.5.2 节）。

事实上，***策略 2*** 还存在一个隐含的问题。在图 2.2 中不同支路输出的概率，不仅仅是特征维度不一致，更重要的是它们的内涵也不同。全局上下文建模通过一个标准的 1007 维的 *Softmax* 分类器生成概率，每个节点的概率值表示的是输入图像究竟属于哪一个类，这意味着所有类别的概率之和应该等于 1。然而，局部上下文建模和跨上下文建模的输出概率表达的是输入图像可能包含某个确定类别对象的可能性，每一个神经元节点的值都可以被看作是一个二类的分类器。换句话说，全局上下文建模估计的是输入图像属于什么类别，而局部和跨上下文建模预测的是输入图像包含什么对象。

表 2.1 使用细到粗策略时全局上下文建模上的类别冲突矩阵

	类别	S00	S01	S02	S01+S02	总和
Alexnet	S00	56126	97	28	125	**56251**
	S01	251	2836	423	-	**3510**
	S02	422	735	2353	-	**3510**
	S01+S02	673	-	-	6347	**7020**
VGG16	S00	56161	64	26	90	**56251**
	S01	139	3085	286	-	**3510**
	S02	225	585	2590	-	**3510**
	S01+S02	474	-	-	6546	**7020**

注：1. S00：正常图像，S01：*L3* 成人图像，S02 少儿不宜图像

2.第二列的类别是标注的 Groundtruth 类别，第一行是预测得到的类别。数值反映的是识别性能，例如，“735”（第 4 行，第 4 列）表示有 735 个属于类别 S02 的图像被分类到了类别 S01 中。

表 2.1 给出了使用细到粗策略时全局上下文建模上的类别冲突矩阵，从数据中可以发现，全局上下文建模对于普通图像具有较高的识别精度，普通图像被错误的分类到成人图像类的仅仅只有 125 幅（Alexnet）和 90 幅（VGG16），大约 0.22%和 0.16%。相对于普通图像，成人图像类（类别 S01 和 S02）有较高的错误率，即使是使用 VGG16 模型，仍然有 4.0%和 9.5%的成人图像被错误地识别为正常图像，合计 6.8%（类别 S01+S02）。受这些结果的启发，并充分考虑不同推理概率的内涵和它们之间的异同，本文提出了一种新颖的层次化选择算法——***策略 3***——用于识别非法样本。本文的目标是尽量保留全局上下文正确的识别结果，并利用局部上下文建模和跨上下文建模来优化这个结果。很容易想到的是，简单地平均这些具有不同含义的输出概率将破坏这种现有的优点。具体说，将检测网络的 *nms* 阈值设置到一个较高的水准（在 *Sensitive* 数据集上通常高于 0.95），这样设置的目的是尽量确保那些被检测出来的对象具有较高的置信度。之后，本文考虑使用类似“排斥门[22, 166]”的方法去权衡选择不同类型的输出概率。局部上下文和跨上下文建模分别由权重w_1和w_2来调节，而全局上下文

由权重$(1-w_1-w_2)$进行调节。最终的 DMCNet 可以用公式 2.6 表示：

$$\mathcal{F}_{DMCN}=max\left((1-w_1-w_2)\cdot\tilde{\mathcal{F}}_{global},w_1\cdot\tilde{\mathcal{F}}_{local},w_2\tilde{\mathcal{F}}_{cross}\right) \tag{2.6}$$

特征 $\tilde{\mathcal{F}}_{global}=\phi(\mathcal{F}_{local})$ 表示进行决策的全局特征$\tilde{\mathcal{F}}_{global}$经由全局上下文建模和细到粗策略生成，而$\tilde{\mathcal{F}}_{local}=\psi\big(\phi(\mathcal{F}_{local},t_1)\big)$和$\tilde{\mathcal{F}}_{cross}=\psi\big(\phi(\mathcal{F}_{cross},t_2)\big)$ 则由局部上下文建模和跨上下文建模经过细到粗策略和正则化后获得。与公式 2.5 相同，函数$\psi(*)$和$\phi(*)$分别表示正则化运算和细到粗策略。参数t_1,t_2是检测网络的非极大抑制（*nms*）阈值，控制通过检测模块输出的对象的数量。如表 2.4 所示，局部上下文建模和跨上下文建模的结果比较相似，因此，可以设置$w=w_1=w_2$，且$t=t_1=t_2$，第 2.5.4 节将讨论不同的权重 *w* 和阈值 *t* 对性能的影响。为了简明起见，本文使用“local”来描述特征$\tilde{\mathcal{F}}_{local-cross}$。基于这个设置$\mathcal{F}_{DMCN}$可以被简化为：

$$\mathcal{F}_{DMCN}=max\left((1-w)\cdot\tilde{\mathcal{F}}_{global},w\cdot\tilde{\mathcal{F}}_{local}\right) \tag{2.7}$$

需要注意的是，全局上下文和局部上下文的物理含义是不同的。虽然公式 2.6 与很多“最大输出”的方法相同，也是在求取不同特征的最大值，但是本文中所描述的局部上下文是非常特殊的。如表 2.1 所示，由于采用强大、鲁棒的卷积神经网络，全局上下文具有较高的识别精度，相比之下，基础的局部上下文建模可信度要低很多。如表 2.4 所示，局部上下文的识别能力远远低于全局上下文建模。直接使用公式 2.6（或其他直观的方式）计算融合特征可能会损害整体的性能。基于这个原因，作者希望能够尽量保留全局上下文识别正确的图像，并进一步从被全局上下文判断为正常图像的图片中召回一些应该属于成人图像的图片。具体地说，本文将通过一个精确的检测模型来完成这个召回的过程。为了获得这个精确的检测模型，本文通过调节 *nms* 阈值 *t* 来增加准确率。众所周知，较高的 *nms* 阈值将降低检测器的召回率，但是对于基于多上下文融合的***策略 3*** 来说，这并不是一个大问题。首先，检测器被设计用来尽量找到那些敏感的对象或区域，而不是所有的对象。较高的精度意味着检测到的对象或区域具有较高的置信度被认定为是“成人”信息。即使检测系统丢失了很多的对

象，但是检测器是可信的。其次，从整个策略的设计逻辑来看，局部上下文建模并不是被设计来执行关键性的决策，而是作为辅助决策。受益于强大的全局上下文建模，这种设计不但不会降低召回率，反而还可以有效地提高整个系统的召回率。表 2.5 的结果证明了这个结论。值得注意的是，本文的目标是改进成人内容的识别率。当然，较高置信度并不能保证完全正确，因此联合决策依然是需要的。

2.5 实验与分析

为了更好地评估和比较本文提出的算法，本文在四个各具特色、且很具挑战性的基准数据集上完成对比实验，包括 *Sensitive*、*NDPI*、*DMCV* 和 *SPD*。在实验中，首先评估了基于高层语义的细到粗策略的有效性，然后讨论多上下文决策对性能的影响。最后，为了衡量识别算法的泛化性能，使用 DMCNet、基准模型和两个优秀的对比方法（AGNet[167]和增量学习[168]）在三个数据集上完成对比实验。泛化性能实验只在 *Sensitive* 数据集上进行训练完成，没有在其他三个数据集上进行微调训练。也就是说，只有 *Sensitive* 数据集包含训练图像，而其他数据集仅包含验证图像和测试图像。其中，验证图像用于调节决策系统的超参数，而测试图像用来做评估。

2.5.1 实现细节

本文使用开源的框架 CAFFE[169]来实现深度多上下文网络模型 DMCNet。所有的实验都在一台配置酷睿 E5 3.0GHz 的 CPU 和 NVIDIA TitanX 12G GPU 的计算机上完成。基于该配置，DMCNet 可以实现识别一幅图像低于 0.3 秒的效率。为了训练一个基于***策略 3*** 的 DMCNet，本文使用如算法 2.1 所示的 5 步训练方法。实验中仅使用单尺度训练和测试方法[65, 100]，并没有使用特征金字塔方法来实现多尺度训练。所有的图像都被调整到 224×224 的分辨率（在 Alexnet 上，所有的图像都被调整到 227×227 的分辨率）。为了训练区域建议网络，本文将与 Groundtruth 的 IoU 大于 0.5 的锚点设置为正，其他设置为负。在训练局部上下文建模网络时，采用以图像为中心[65, 100]的训练方法，设置每个 mini-batch 包含 1 个图像，并随机采样 128 个锚点用于计

算损失。每个 mini-batch 的正负样本的比例为 1:3。所有的模型都采用“step”学习策略，初始学习率为0.001。每次学习率衰减时的系数都为0.1，衰减步长60,000、25,000、30,000 分别对应于区域建议网络、全局上下文建模和局部上下文建模（跨上下文建模）。同时，设置动量为 0.9，权重衰减率为 0.0005。在训练过程中，使用阈值为 0.7 的非极大抑制算法（non-maximum suppression，nms）去过滤建议区域。在测试过程中，*nms* 阈值 *t* 是一个可调的超参数。细到粗策略始终被用做后处理。如图 2.2 所示，DMCNet 的三个输出维度分别为 15、15 和 1007。对于每一个分支，特征的输出维度都通过类别分组方案转换为 2 或 3 用来进行 *L2* 或 *L3* 评估。具体来说，在每个组中，使用最大概率作为该组的输出概率完成细粒度类别到粗粒度类别的转换。组的定义和转换规则将在下一小节中进行描述。最后，一个 one-hot 的 *Softmax* 分类器别用来做最终的决策。

算法 2.1：深度多上下文网络（DMCNet）训练过程

步骤 1：在 *Sensitive* 数据集上使用 *Imagenet* 预训练模型[3]训练一个新的深度模型，该模型作为基准 Baseline 模型和全局上下文模型，同时使用该模型去初始化*步骤 2* 和*步骤 3* 中的卷积层。这些卷积层称为共享卷积层。

步骤 2：使用*步骤 1* 中的网络作为预训练模型，训练区域建议网络，其中共享卷积层的参数固定不变。

步骤 3：使用*步骤 2* 中训练好的区域建议网络作为建议区域生成器，训练一个可表征局部上下文信息的对象检测网络。该检测网络仍然使用*步骤 1* 中训练好的基准模型作为初始权重。共享卷积层和区域建议网络的参数固定不变。

步骤 4：保持共享卷积层，区域建议网络的参数固定不变，组合局部上下文特征和全局上下文特征，并微调新的混合层用于生成跨上下文信息。建议区域仍然由*步骤 2* 中生成的网络产生。

步骤 5：将*步骤 2*、3 和 4 中训练好的模型组合起来形成统一的多上下文框架，作为最终的 DMCN 模型。

2.5.2 数据集及评估指标

Sensitive 数据集通过互联网收集获得，并按照 *Imagnet*[73]数据集的组织方法进行分类和标注，大体上包含 30,000 个非优雅的图像。“成人”图像在这个数据集中被定义为包含裸体男人或女人的图像，具体涉及不同的姿态、尺寸、行为和局部特写。这些图像被分为 14 个类用于区分不同的色情内容，例如：裸体、隐私部位、性行为、内衣照、泳装照和腿模照等。其中有两个例外类，包括：面部特写和正装图。为了进一步增加数据集的多样性和复杂性，作者从 *Imagenet*[73]数据集和 *Pascal VOC* 数据集[91]中抽取了大量的图像用于创建“正常”图像类。众所周知，*Imagenet* 数据集包含 1,000 个对象类（由于 *Imagenet* 数据集有 7 个类别和收集的样本比较接近，例如泳装照和内衣照，因此这些图像将从负样本中被移除，最后保留了 993 个类别），它能显著增加类别空间的多样性；而每一个 *Pascal VOC* 的图像都包含多个对象，它可以有效地增加数据集的复杂性。至此，正常图像将包含自然场景、正常对象、良性人像图和混合图像。此外，数据集中大多数人都是高加索人种和亚洲人种，少量图像为黑人种。最终，整个 *Sensitive* 数据集包含 1,413,765 幅图像，其中 22,657 幅成人图像和 1,391,108 幅正常图像。合计有 1,300,144 幅训练集图像，50,350 幅验证集图像和 63,271 幅测试集图像。为了训练 DMCNet，训练和验证图像都分为两个子集：分类数据集和检测数据集。后者只包含互联网收集的图像，不包含额外的数据。为了训练全局上下文模型（即分类任务），总共有 1,300,144 幅训练图像和 50,350 幅验证图像，它们分别属于 1,007 个类（993 个类由 *Imagenet* 数据集定义，14 个类由 *Sensitive* 数据集定义，每个类大约 1300 幅图像）。为了训练局部上下文模型（即检测任务），17,637 幅图像用于训练，700 幅图像用于验证。在检测数据集中，每幅图像都包含多个对象，所有的对象都属于 15 个类（14 个类由 *Sensitive* 数据集定义，另外一个是额外背景类）。所有实验都使用相同的测试图像，包含 7,020 幅成人图像和 56,251 幅正常图像。

此外，本文定义了两种评估标准来评估模型性能：两级评估模式（用“*L2*”表示）和三级评估模式（用“*L3*”表示）。为了使数据集更适合这些评估标准，可以将所有的正常图像归类到一个类，称为“正常图像”，用“S00”表示；9 个明显的色情和裸

露的类别组合成一个类，称为“成人图像”，用“S01”表示；3 个边缘类组合在一起称为“少儿不宜图像”，用“S02”表示。使用类别“S00”、“S01”和“S02”去完成“*L3*”评估，然后用类别“S01”和“S01+S02”完成“*L2*”评估。在本文的工作中，所有的实验都使用该设置完成。其中，*Sensitive* 数据集同时使用“*L2*”和“*L3*”评估标准，其他三个数据集仅使用“*L2*”评估标准进行评估。为了便于理解和简便本文使用“*L3* 成人图像”来表示“*L3*”评估中的“成人图像”S01；使用“成人图像”来表示“*L2*”评估中的“成人图像”（S01+S02）。

NDPI[170]成人数据集包含近 80 个小时的 400 个成人视频和 400 个非成人视频。它由巴西米纳斯吉拉斯联邦大学（Universidade Federal de Minas Gerais，UFMG）的 NPDI 工作组收集。在成人类中，该数据集包含多个族裔在不同场景下不同行为的 400 个视频样本。非成人类由两个子集构成，200 个视频由互联网随机收集，标注为“简单”，另外 200 个视频依据关键字“海滩”、“摔角”、“游泳”在互联网收集获得，这些视频大多含有裸露的皮肤（但并不属于成人图像），因此标注为“困难”，这些对于检测器来说是极具挑战的设置。从这 800 个视频中，总计截取出 16,272 个关键帧，并使用这些关键帧来估计一个视频是否属于成人内容。与原始数据集定义的标准交叉验证协议[170]不同，整个数据集划分为验证集和测试集两部分。其中，200 个视频（100 个成人，100 个非成人）用于调节超参数，剩下的 600 个视频（300 个成人，300 个非成人）用于评估模型的性能。非常值得注意的是，与文献 AGNet[167]不同，本文重新实现的 AGNet 模型并没有在 *NDPI* 数据集上进行微调训练，它仅仅只在 *Sensitive* 数据集上完成预训练。

DMCV（Dynamic Magnificent Colorful Video）数据集利用一个流行的视频软件收集获得，它包含 99 个成人视频和 100 个正常视频。这个数据集最大的挑战来源于拍摄时用户大量使用的华丽的滤镜和特效。这些辅助拍摄工具使成人视频和正常视频都具有极大的多样性，这给识别系统带来了巨大的干扰。在该数据集中，其中的 40 个视频被用来调节超参数，剩下的 159 个视频用于测试和评估模型。在整个数据集中，成人视频和正常视频是严格平衡的。对于每个视频，分别抽取 20 个镜头，并从每个镜头中抽取 1 个关键帧用来表征该镜头片段。与 *NDPI* 数据集类似，*DMCV* 数

据集也没有训练样本用来训练模型，仅用作泛化性能评估。

SPD（Sensitive Poster Dataset）数据集是一个小型但很困难的数据集，总共包含1,074 幅成人图像和 8,926 幅正常图像。其中，2,000 幅图像用做验证，8,000 幅图像用于测试。与其他数据集不同的是，*SPD* 包含大量复杂的图像（称为“海报”），它们都包含拥挤的场景，且主体对象通常很小。

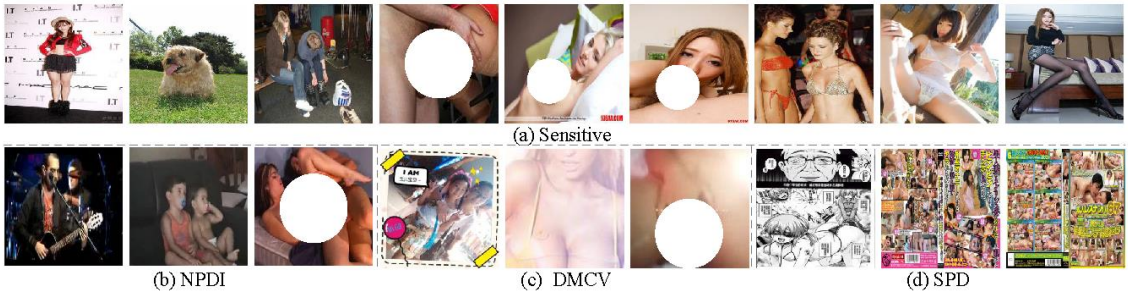


图 2.4 四个成人数据集的范例图像

为了更直观地理解这 4 个数据集的细节，图 2.4 中可视化了若干从数据集中抽取出来的图像或帧。子图（a-d）中的图像分别从 *Sensitive*、*NDPI*、*DMCV* 和 *SPD* 数据集中获得。其中，（a）*Sensitive* 的前三幅图像是“正常图像”，中间的三幅图像是“*L3* 成人图像”，最后的三幅图像属于“少儿不宜图像”；（b）*NDPI* 中的第一幅和第二幅图像分别是“容易”和“困难”的正常图像，第三幅是“成人图像”；（c）*DMCV* 的第一幅图像是正常图像，其余两幅是“成人图像”，所有的三幅图像都使用特殊的特效滤镜进行拍摄；（d）*SPD* 中所有的图像都是“成人图像”，其中第一幅是卡通素描，另外两幅是混乱的“海报”图像。从图 2.4 中可以清楚地看到，每一个数据集都有它的独特性。

评估指标。在本章中，四种评价指标被用来衡量模型的性能，分别是：召回率（Recall）、准确率（Precision）、F1-Score 和精确度（Accuracy），其表达式如下所示：

$$Recall = \frac{TP}{TP + FN} \tag{2.8}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.9}$$

$$F1-Score=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{2.10}$$

$$Accuracy=\frac{TP+TN}{Total} \tag{2.11}$$

其中，TP、FN、TN、FP 的关系表 2.2 所示：

表 2.2 相关样本关系表

	相关的样本	不相关的样本
检测到的样本	TP (True Positive) 检测到的相关的样本数	FN(False Negative) 检测到的不相关的样本数
未检测到的样本	FP(False Positive) 未检测到的相关的样本数	TN(True Negative) 未检测到的不相关的样本数

2.5.3 基于高层语义的细到粗策略的分析

根据***策略 3*** 的设置，所有的多上下文模型都基于全局上下文的扩展实现。因此，细到粗的策略评估也都基于全局上下文来实现。在这一节的实验中，全局上下文分支被从 DMCNet 中分离出来，并将最终的输出维度设置为 2 和 3 两种模式，用于完成 *L2* 和 *L3* 评估。在本节中，本文设计了两组不同的实验来分析和评估细到粗策略。首先，本文基于三种流行的深度 CNN 模型，在 *Sensitive* 数据集上对细到粗策略进行评估，这主要用来衡量该策略对于不同 CNN 模型的适应性和稳定性。其次，为了验证该策略对于不同数据集都有较好的性能，本文在所有四个数据集上都进行了评估验证。其中，在 *NPDI*、*DMCV* 和 *SPD* 数据集上使用 *L2* 评估协议；在 *Sensitive* 数据集上同时使用 *L2* 和 *L3* 两种评估协议。为了训练基准 Baseline 模型，本文将 *Sensitive* 数据集进行两个类别（“普通图像”S00、“成人图像”S0102）和三个类别（“普通图像”S00、“*L3* 成人图像”S01、“少儿不宜图像”S02）两种设置，然后构建二分类和三分类两个分类任务用于评估深度模型。

首先，DMCNet 框架可以很灵活地整合各种深度卷积神经网络模型。为了方便，本文直接将这些深度模型用来替换全局上下文模型，用来进行评估的模型结构包括：

Alexnet[23]、VGG16[3]和 GoogLeNet[4]。表 2.3 给出了细到粗策略在这三个模型下构建的全局上下文建模中的性能评估结果。所有的基准 Baseline 模型都没有使用细到粗策略，而全局上下文模型 Global-Context 都使用了细到粗策略。

表 2.3 细到粗策略在全局上下文建模中的性能评估

		S00	S01	S02	S0102	时间消耗（毫秒）
Baseline	Alexnet	99.0	73.23	57.9	91.8	32
	VGG16	99.3	80.0	72.9	93.9	145.4
	GoogLeNet	98.6	67.0	72.8	92.3	101.2
Global-Context	Alexnet	99.3	79.0	74.5	94.1	47
	VGG16	**99.5**	**85.2**	**80.8**	**95.9**	158.5
	GoogLeNet	99.4	81.7	77.5	94.8	117.6

注：S00：正常图像，S01：*L3* 成人图像，S02：少儿不宜图像，S0102：*L2* 成人图像（S01+S02）

从实验结果来看，较深的模型（VGG16 和 GoogLeNet）相对于较浅的模型（Alexnet）始终具有一定的优势，这也验证了即使使用了细到粗策略，模型仍然满足 CNN 模型的基本规则——越深越好。令人高兴的是，在使用了细到粗策略之后，所有基准框架的 F1-Score 性能都得到了提高。从结果来看，性能提升主要发生在类别 S01 和 S02 上，这主要有两个原因。（1）对于普通图像来说，目标对象通常具有较强的语义信息，一个好的特征表达和一个好的分类器总是能够很好地完成高层语义信息的识别。然而，受姿态、尺度、视点、行为、遮挡、对象完整性和其他不确定的因素影响，每一幅成人图像的语义信息都有可能是模棱两可的。将他们作为一个统一的类别，并在大规模的数据集上进行识别是非常困难的。（2）一些成人图像，特别是局部特写，在视觉上可能非常接近正常图像。如果没有一个细粒度的分类定义，对于一些图像，类内距离可能会比类间距离更大。

另一个值得高兴的是，引入细到粗策略后，时间负担增加并不明显。

其次，图 2.5 给出了细到粗策略在四个成人数据集上 F1-Score 的评估结果。（a-

b）分别显示了正常图像（S00）和成人图像（S0102）在 4 个数据集上的性能。其中，*Sensitive* 数据集中的类别 S0102 是类别 S01 和 S02 的组合。（c-d）反映的是 *Sensitive* 数据集上类别 *L3* 评估下，成人图像（S01）和少儿不宜图像（S02）的性能。从实验结果看，细到粗策略在所有的数据集上的性能都表现良好。将细粒度类转换为粗粒度类之后，所有类别的识别性能都得到了提高。特别是在 *DMCV* 数据集上，细到粗策略大约将 F1-Score 值提高了 32%，这主要是因为经过特效滤镜处理的图像大大增加了二分类分类器的识别难度。

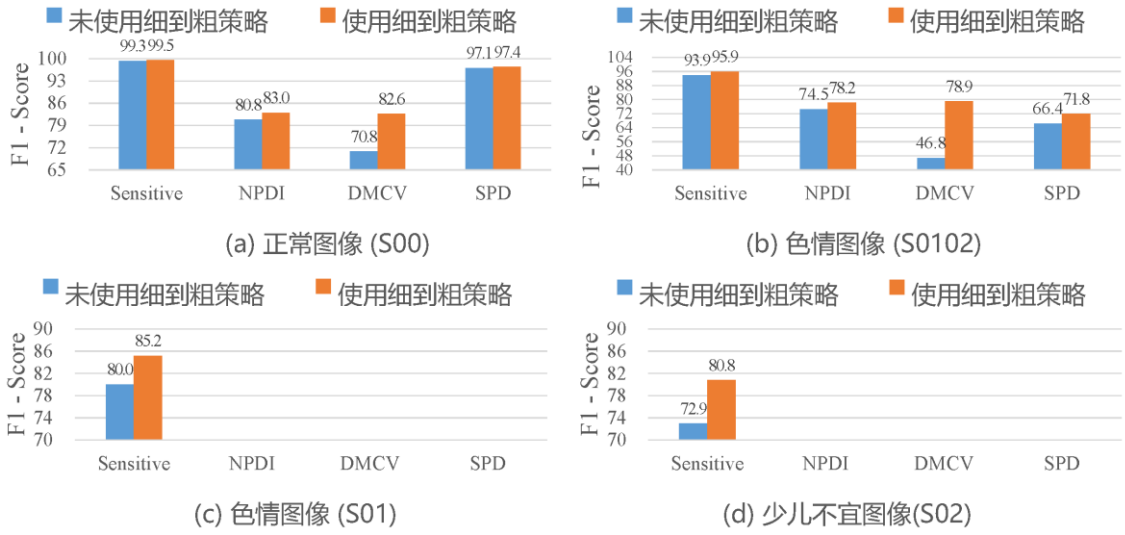
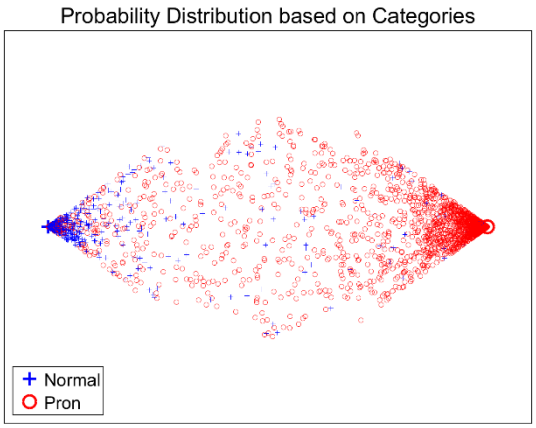
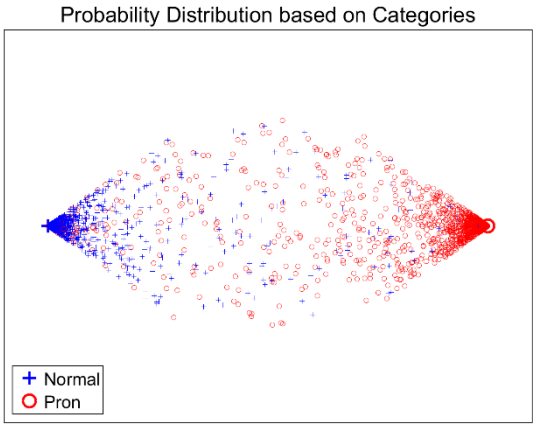


图 2.5 细到粗策略在四个成人数据集上 F1-Score 评估结果

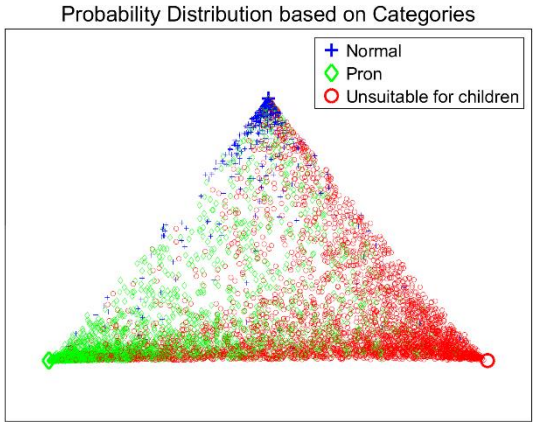
为了更好地理解为什么细到粗策略能够在“成人识别”任务中改进分类性能，可以通过每个图像的分类概率密度图来进行理解。图 2.6 给出了细到粗策略在 *Sensitive* 和 *DMCV* 数据集上的可视化评估结果。（a）、（b）、（c）显示了未使用细到粗策略的样本类别概率分布图；（d）、（e）、（f）显示了使用细到粗策略的样本类别概率分布图。图中的每一个符号都表示一幅图像，符号的颜色由图像的 Groundtruth 类别决定，位置由识别系统关于每一个类的分类概率决定。聚类中心使用较大的符号表示，它表征的是样本的类别中心；同类的图像使用相同形式但是较小的符号表示。假设图像 I 属于类别 C，那么关于类别 C 的分类概率越高，它与中心 C 的距离也就越近。类别中心由较大的符号表示，如果图像 I 的分类概率为 1，那么它将和类别中心重合。逻辑



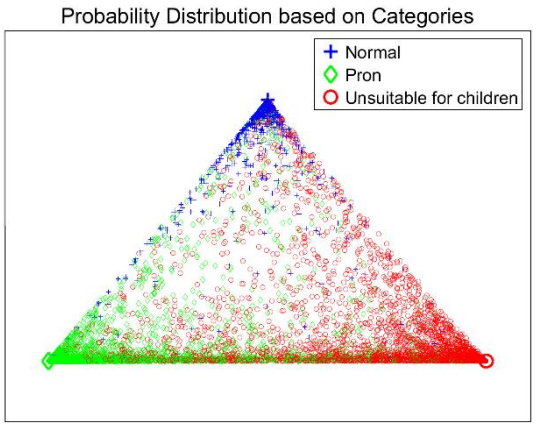
（a）*Sensitive*, *L2* 标准，未使用细到粗策略



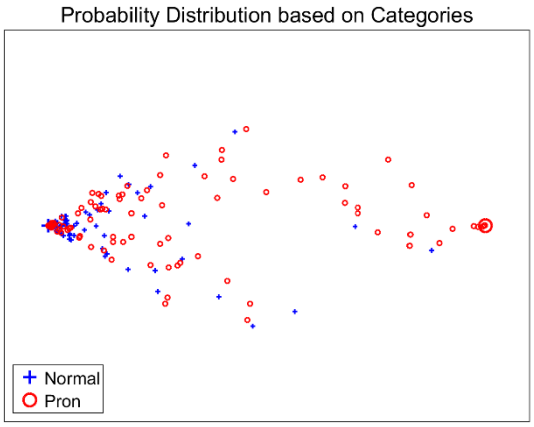
（d）*Sensitive*, *L2* 标准，使用细到粗策略



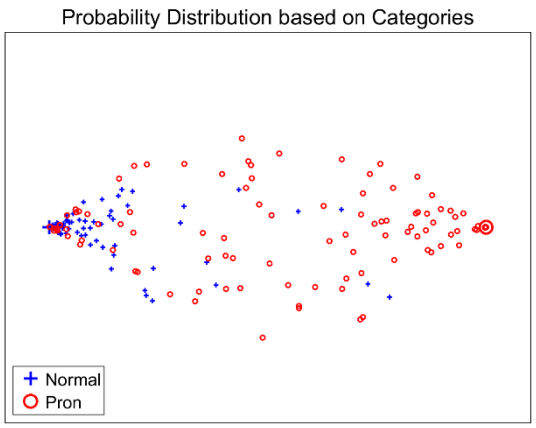
（b）*Sensitive*, *L3* 标准，未使用细到粗策略



（e）*Sensitive*, *L3* 标准,使用细到粗策略



（c）*DMCV*, *L2* 标准，未使用细到粗策略



（f）*DMCV*, *L2* 标准，使用细到粗策略

图 2.6 细到粗策略在 *Sensitive* 和 *DMCV* 数据集上的可视化评估结果

上，可以认为符号越密集意味着特征的判别能力越强，如果分类器是完美的，那么所有较小的符号都将和较大的符号重合在一起，即所有的样本都以概率 1 的值被准确地分配到 Groundtruth 类别中。在密度图中，首先手工指定每一个聚类中心的坐标以标识确定的类别，然后依据每幅图像在 DMCNet 中前向传播后获得的分类概率值绘制每个样本相应的坐标位置。子图（b）和（e）显示的是 *L3* 分类问题的概率分布，其他 4 个子图显示的是 *L2* 分类问题的概率分布。从图 2.6 中不难发现，在引入了细到粗策略之后，更多的符号涌向了类别中心，特别是成人图像。此外，可以发现类别边缘区域的样本明显减少。以图 2.6（a）和图 2.6（d）为例，它们展示的都是 *Imagenet* 数据集的 *L2* 分类问题。相比图 2.6（a），图 2.6（d）中大量的红色圆圈都聚集在右边的类别中心，这意味着细到粗策略的引入使成人图像更加容易地被分配到正确的类别中。与图 2.6（b）不同，图 2.6（e）中的绿色菱形大多集中在左下角，而红色圆圈主要集中在右下角。此外，另一个有趣的例子是图 2.6（c）和图 2.6（f）。在图 2.6（c）中，大量红色圆圈都集中在左边的蓝色十字中心附近，这可以认为是非常差的分类结果。通过比较图 2.5（b）中的量化数据，也可以发现 *DMCV* 数据集在不使用细到粗策略时，仅仅只有 46.8%的 F1-Score 值，大多数样本都被错误地分类，这也证明了本文的结论。相反，在引入了细到粗策略之后，F1-Score 值提高到了 78.9%。这个结果也可以在图 2.6（f）中清晰观察到，大量的红色圆圈跑向了右边的红色圈圈中心。

总的来说，增加类别空间可以显著地增强模型的判别能力，进而提高成人图像的识别能力。如表 2.3 所示，VGG16 总是具有最好的 F1-Score 性能，因此该模型被用来完成后续的实验。

2.5.4 多上下文建模策略的分析

在这一小节中，首先比较不同的上下文模型在四个数据集上的 F1-Score 性能，然后，深入分析超参数 t 和 w 对模型的影响。在这本节的实验中，细到粗策略被默认启用。

表 2.4 给出了多上下文建模在四个数据集上的性能评估结果。大多数情况下，多上下文模型在 4 个数据集上都比单上下文模型表现得更好。相对而言，多上下文模

型在普通图像的识别上并没有获得令人满意的结果，甚至在 *DMCV* 数据集上比 Baseline 模型还要更差（82.6 vs. 82.3）。尽管如此，多上下文模型在 *NPDI*、*DMCV* 和 *SPD* 数据集上的成人图像识别能力都优于 Baseline 模型。不幸的是，多上下融合策略在 *Sensitive* 数据集上并没有表现出太优秀的能力，尽管在所有类的识别上，它并不比任何 Baseline 模型更差。以上现象主要有以下几点原因。（1）深度卷积神经网络具有很强的特征学习能力。当一个样本具有明显的语义时，特征的判别能力将会非

表 2.4 多上下文建模在四个数据集上的性能评估

		F1-Score			
		S00	S01	S02	S01+S02
Sensitive	Global-Context	**99.5**	85.2	80.8	**95.9**
	Local-Context (t=0.99)	94.6	**19.8**	9.8	15.2
	Cross-Context (t=0.99)	95.5	20.0	**9.8**	15.5
	Multi-Context (w=0.47)	**99.5**	85.3	80.9	**95.9**
NPDI	Global-Context	83.0			78.2
	Local-Context (t=0.3)	77.7			79.2
	Cross-Context(t=0.3)	78.6			80.1
	Multi-Context (w=0.32)	**85.2**			**85.3**
DMCV	Global-Context	**82.6**			78.9
	Local-Context (t=0.2)	66.0			70.4
	Cross-Context (t=0.2)	66.6			71.1
	Multi-Context (w=0.27)	82.3			**80.4**
SPD	Global-Context	97.4			71.8
	Local-Context (t=0.6)	96.6			63.1
	Cross-Context (t=0.6)	**97.5**			63.7
	Multi-Context (w=0.48)	**97.5**			**74.7**

注：S00：正常图像，S01：*L3* 成人图像，S02 少儿不宜图像，S0102：成人图像（S01+S02）

常强大，特别是那些只有单目标，或者主体目标非常明显的样本。*Sensitive* 数据集中的普通图像就是典型例子。全局上下文建模在图像识别上已经能够处理得非常好，因此，增加额外的局部上下文信息很难明显改进识别性能。（2）成人图像通常包含大量的局部特写，局部上下文有助于发现这些信息，这可以有效地弥补全局上下文的缺陷，从而实现成人图像识别性能的提升。（3）受检测模型性能的限制，过分依赖于局部上下文信息可能会损害整个系统的性能，特别是当全局上下文建模已经具有很强的能力而有用的局部信息又很少的时候。例如 *Sensitive* 数据集，它包含大量各种各样的图像，但成人图像所占的比例又非常小。这也是为什么在 *Sensitive* 数据集上，*nms* 阈值 *t* 比其他数据集大很多，但是性能改进却非常有限。图 2.3（第 3-4 列）给出了一些局部上下文建模帮助改进识别的例子，由于组合了全局和局部上下文建模，多上下文模型修正了单纯使用全局上下文建模的一些预测误差。

此外，超参数的选择对于融合性能也非常重要。接下来将详细分析超参数 *w* 和 *t* 对模型的影响。为了公平可靠，所有的参数选择均在验正集上完成，最终的结果报告将在测试集上完成。事实上，由于训练、验证和测试集是随机分配的，因此验证集和测试集上的结果是非常接近的。如公式 2.6 所描述的，*w* 是一个排斥门参数。如果$w = 1$，则$\mathcal{F}_{DMCN}$由局部上下文决定，也就是说$\mathcal{F}_{DMCN} = \tilde{\mathcal{F}}_{local}$；当$w = 0$，则$\mathcal{F}_{DMCN}$由全部上下文决定，也就是说$\mathcal{F}_{DMCN} = \tilde{\mathcal{F}}_{global}$。从图 2.7 中可以得到以下几点主要的结论。（1）F1-Score 的极值分布通常位于 0.2~0.5 之间。由此，可以推断，全局上下文信息对于整个系统的性能更加重要。与此同时，在全局上下文信息中增加一定量的局部上下文信息能够改进整体的识别性能，特别是对于成人图像。如果图像包含很多的小对象或者图像明显是性器官的特写，那么局部上下文信息将更有效。例如，权重 *w* 在 *SPD* 数据集上要比其他数据集更大，这主要是因为它包含大量的“海报”图像。（2）超参数 *w* 的选择至关重要。由于检测网络的可靠性明显不如分类网络，参数 *w* 变得非常敏感。一个更小的 *w* 将增加召回率，但是准确率会快速下降。另一方面，较大的权重参数 *w* 能够改进准确度，但是它仅仅只能召回少量的图像，并且这些图像中的大多数已经由全局上下文建模正确判别。这也是为什么 *Sensitive* 数据集和所

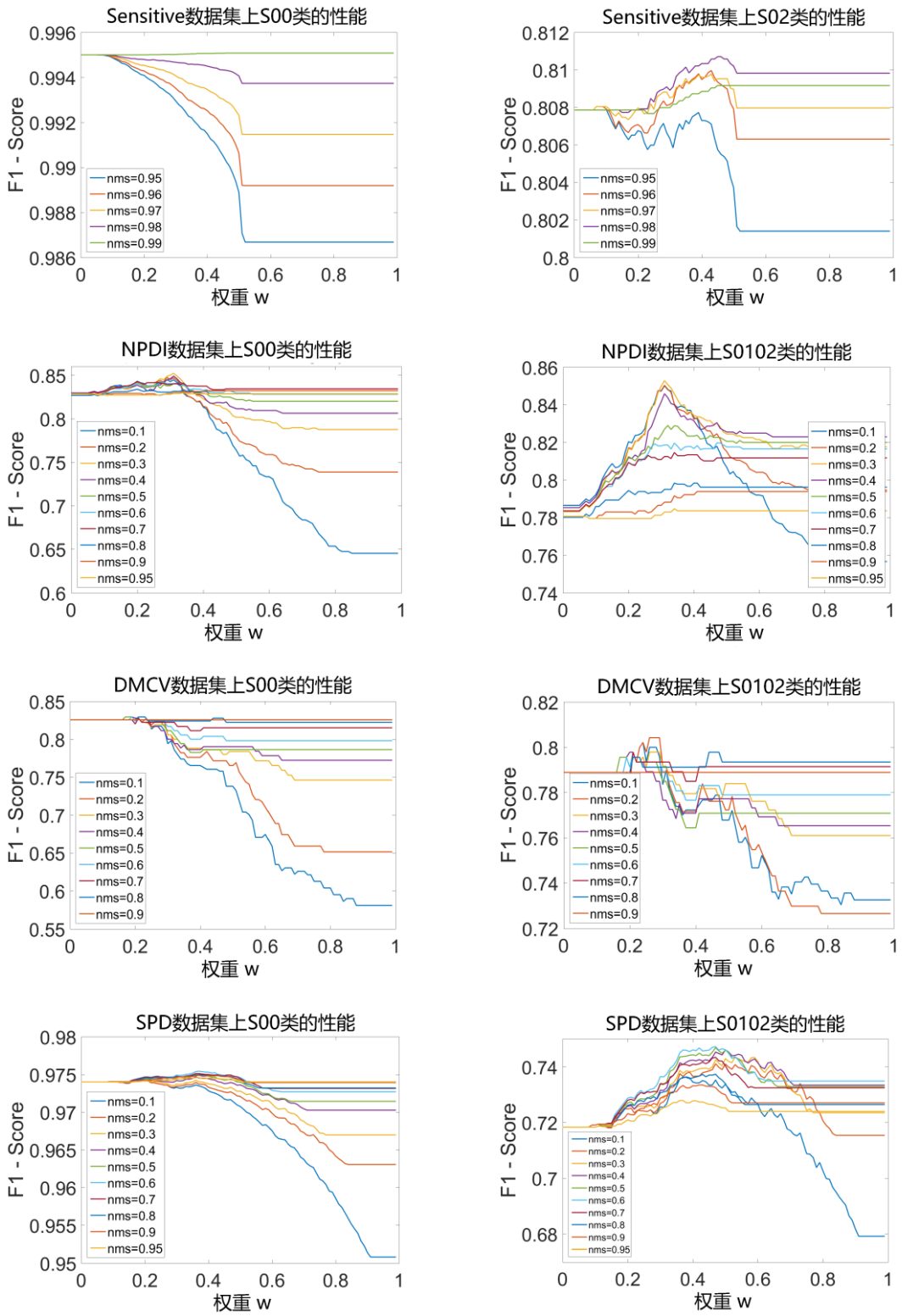


图 2.7 多上下文建模在四个数据集上的评估结果

有的正常图像的 F1-Score 曲线的增长受到限制。因此，如何权衡局部上下文的引入比例（通过 w）以及检测框的数量（通过 t），对于融合特征是极为重要的。本文在四个数据集的验证集上做了大量的实验来选择合适的参数。（3）部分曲线对于权重 w 并不是很敏感，例如类别 S00 可期望的性能改进主要来源于那些被全局上下文建模错误分配到正常图像的成人图像能够被局部上下文重新召回。然而，此时一些正常的图像也会被错误的识别成成人图像。也就是说，经过局部上下文修正之后，正常图像类的准确度可以被改进，但是召回率会下降。如图 2.7 所示，*Sensitive* 数据集的 S02 类的性能曲线也趋于平缓。这主要是因为 *Sensitive* 数据集是一个非常大的数据集，而且图像普遍都是高分辨率且以对象为中心，这使得全局上下文在细到粗策略的作用下将变得非常具有判别性，留给多上下文决策改进的空间变得非常小。事实上，最终只有几十个成人图像最终被局部上下文建模正确地召回，这个数量很容易被湮没在庞大的数据海洋中。因此，*Sensitive* 数据集的 S02 类别的性能曲线显得非常平缓。（4）根据结果来看，对于不同分布图像的数据集来说是非常敏感的，通过在验证集上进行调节是一个有效的方式。然而，更好方法是进一步改进检测模型的性能，一个更为强大局部上下文建模方法能够使参数变得更加稳定，从而降低模型对数据集先验知识的依赖性。

2.5.5 整体性能对比

表 2.5 和表 2.6 比较了两种当前成人识别比较优秀的算法，包括基于深度学习的 AGNet[167]和基于增量学习[168]（Incremental Learning）的传统方法。其中表 2.5 给出了 *Sensitive* 数据集上四个算法在各种指标下的评估结果；表 2.6 给出的是四个算法在其他三个数据集上的泛化性能测试，即：所有算法都没有在 *NPDI*、*DMCV*、*SPD* 数据集上进行训练，而是直接进行测试。本文重新实现的对比模型和文献中的原版略有不同。对于 AGNet[167]，本文仅仅只训练了一个网络，而不是五个网络的合成，因为 DMCNet 方法和增量学习的方法都没有采用 5-fold 方法来进行学习和验证。对于增量学习方法[168]，本文只在初始化的时候完成覆盖中心的选择。因为，本文作者发现对于大规模数据集（例如，本文用于训练的 *Sensitive* 数据集）来说，利用错误样本

来调节覆盖中心是无效甚至有害的。对于 DMCNet 模型，细到粗策略和多上下文融合策略在最后的测试中都作为默认配置被使用。

从表 2.5 的实验结果看，深度学习方法的性能非常强大，基于深度学习的三个对比方法一致优于基于传统增量学习的方法。除此以外，DMCNet 在成人内容识别上具有很明显的优势。DMCNet 在四个数据集上都获得了较高的召回率、F1-Score 和准确率。在准确率上，AGNet 略高于 DMCNet，但其召回率远低于 DMCNet。这主要是因为在处理成人内容识别任务时，本文旨在尽力而为地识别出更多的成人样本，而不仅仅是提高识别样本的准确率，这使得在提高召回率的过程中，准确率会有一定的下降。幸运的是，准确率的降低并不是很明显。

表 2.5 所有模型在 *Sensitive* 数据集上的性能评估结果

方法	召回率			准确率			F1-Score			精确度	
	S01	S02	S0102	S01	S02	S0102	S01	S02	S0102	L2	L3
AGNet	80.2	56.4	86.0	55.9	**92.9**	**98.8**	65.9	70.2	92.0	98.3	96.3
Incremental Learning	58.7	26.5	68.5	14.3	30.0	35.8	23.0	28.1	47.0	75.1	72.3
Baseline-VGG16	83.4	62.6	89.2	76.9	87.4	99.1	80.0	72.9	93.9	98.7	96.9
DMCNet	**88.0**	**73.9**	**93.4**	**82.6**	89.3	98.7	**85.3**	**80.9**	**95.9**	**99.1**	**97.8**

注：S01：*L3* 成人图像，S02 少儿不宜图像，S0102：成人图像（S01+S02）

表 2.6 进一步验证了本文提出方法的泛化性能（这也是为什么在 *NPDI* 数据集上，本文报告的 AGNet 的精度只有 79%，远低于原始论文的 94%[167]）。基于在 *Sensitve* 数据集上同样的原因，在 *NPDI* 数据集中，AGNet 在准确率上获得了最高的性能，它比 DMCNet 略高（85.6 vs. 85.1），然而它的召回率远低于 DMCNet（69.8 vs. 85.5）。表 2.6 也证明了，成人内容识别任务，对于不同的数据集具有一定的通用性。因此，一个训练好的强大的标准识别模型可以被轻松、无缝地推广到大多数需要进行成人内容识别的环境中，而不需要再次重复训练。这个性质为将成人内容识别系统推广和部署到不同环境中提供了基础保证。

表 2.6 所有模型在三个泛化数据集上的性能评估

数据集	方法	召回率		准确率		F1-Score		精确度
		普通	成人	普通	成人	普通	成人	
NPDI	AGNet	88.3	69.8	74.7	**85.6**	80.8	76.9	79.0
	Incremental Learning	76.2	51.6	71.8	57.2	73.9	54.3	63.9
	Baseline-VGG16	**92.2**	64.0	71.9	89.2	80.8	74.5	78.1
	DMCNet	85.0	**85.5**	**85.4**	85.1	**85.2**	**85.3**	**85.3**
DMCV	AGNet	87.0	65.7	71.9	83.3	78.7	73.5	76.4
	Incremental Learning	23.0	86.9	63.9	52.8	33.8	65.7	54.8
	Baseline-VGG16	**91.0**	33.3	58.0	78.6	70.8	46.8	62.3
	DMCNet	86.0	**76.8**	**78.9**	**84.4**	**82.3**	**80.4**	**81.4**
SPD	AGNet	99.8	49.2	94.4	96.9	97.0	65.8	94.4
	Incremental Learning	78.9	40.3	91.8	18.4	84.9	25.2	74.8
	Baseline-VGG16	**99.9**	49.9	94.4	**99.3**	97.1	66.4	94.6
	DMCNet	99.3	**63.1**	**95.8**	91.7	**97.5**	**74.7**	**95.4**

注：S01：*L3* 成人图像，S02 少儿不宜图像，S0102：成人图像（S01+S02）

2.6 小结

在本章中，提出一种基于深度学习的多上下文框架（Deep Multi-Context Network，DMCNet）用于成人图像和视频的识别。本章工作主要有以下几点贡献：

1. 设计了一种基于深度卷积神经网络的多上下文体系结构，用于同时学习样本的全局上下文信息、局部上下文信息和跨上下文信息。
2. 提出了一种精心设计的多上下文融合策略，该策略充分考虑了成人图像识别任务的特殊性。与常见的卷积特征融合不同，DMCNet 没有直接平均不同的特征，而是采用一种加权的层次化的联合决策策略。该策略的目的是尽量保持全局上下文建模的识别精度，然后使用局部上下文建模去修正少量的错误。

3. 由于成人图像的多样性，设计了一个基于高级语义的细到粗策略去发现更多的细节特征和语义信息，使对于成人对象的识别更加稳定和可靠。
4. 模块化的设计方案允许通过更新全局上下文建模和局部上下文建模组件来实现整体系统性能的提升。
5. 收集整理了三个专门用于成人内容识别的具有不同特色的数据集。在合适的时候，将通过正式的申请和授权公开数据集的下载和使用。

值得注意的是，虽然本文致力于解决成人图像和视频识别这样一个特殊的任务，但是这样一个特征学习和融合的方向可以被考虑用来完成细粒度识别任务和特殊目标识别的任务。

3 基于局部语义增强的场景解析

3.1 引言

场景解析是计算机视觉的另一个重要任务，它属于图像分割的一个分支，与语义分割、实例分割有很大的相似性，但又不完全相同。最大的区别在于场景解析不仅关注对象，同时也非常关注背景区域。换句话说，场景解析需要处理样本的每一个像素。因此，它需要有比语义分割和实例分割更细致的处理能力。场景解析在自动驾驶、互联网视频搜索、卫星遥感图像分析、图像搜索、机器人导航、室内三维建模等多个领域都有较广泛的潜在应用。同时，场景作为一种全局上下文信息，在其他诸如图像分类、目标检测、视频分类等任务中，都可以起到强相关性，极大地辅助这些任务的分析与处理。

本章提出一种利用局部语义来增强全局语义理解的场景解析方法——对象区域增强网络（Objectness Region Enhancement Network，OENet）。本章的相关研究工作即将发表在文献[171]中。

3.2 问题描述

场景解析，即识别和分割对象与背景，是理解场景的关键问题之一。作为一个重要的计算机视觉的任务，它可能影响到日常生活的很多方面。例如，在一个餐厅的场景中，一个服务机器人可以很容易的识别出它所处的位置和场景的类别。然而，想要自由地在场景中进行导航和使用物品，机器人还需要理解更多、更复杂的信息。例如，它不仅仅需要识别和定位大型对象（如：桌子，椅子和人），它还需要找到更小的对象（如：胡椒瓶、盘子、糖罐等）以及他们的部件（如：杯子的把手、桌子的表面）以完成一些潜在的交互任务；与此同时，还需要理解很多背景区域（如：墙，地板与门）用来进行空间导航。在过去的两年里，受益于全卷积网络[70]（Fully Convolutional Network, FCN）的发展，语义分割取得了前所未有的发展。通过重用图像的特征，

FCN 避免了在图像中计算每一个像素的类别时的冗余计算问题。它已经变成了实现密集预测的事实标准，很多方法都基于该方法进行改进，例如 DeepLab 模型[118]和 Adelaide 上下文模型[128]。

然而，FCN 的像素级预测是通过利用大跨度的双线性插值来完成粗糙的上采样卷积特征来实现。因此在分割中对象的边缘会过于平滑，同时由于固定尺寸的感知域可能会使一些前景对象湮没在大量复杂的背景区域，特别是那些较小的对象。对于语义分割和实例分割任务来说，这或许并不是太大的问题，因为它们主要关注的是如何将主要对象从背景中分离出来。甚至于在较为复杂的 *MSCOCO* 数据集[90]中，要找到大多数对象也不是很困难的事情。这是因为，在以对象为中心的分割任务中，对象的尺度通常足够大到让识别器认出它来。同时，并不需要关心背景是什么。基于这两点，分割任务的难度就会有一定的降低。

然而，在场景解析任务中，复杂的场景使很多对象都很小，且场景中所包含对象的种类和数量都会变得很多、很复杂。此外，场景解析不仅仅要将对象从场景分离出来，还需要去识别对象后面的背景究竟是什么。因此，为了较好处理场景解析问题，需要一种方法去从场景中将这些对象找出，特别是那些很小或者视觉上模棱两可的对象。许多研究者提出使用检测系统[172, 173]来帮助分割对象。这些方法首先利用检测系统来生成对象建议区域，然后在这些区域内执行图像分割。基于检测的方法有利于召回一些在原始的分割网络中难于发现的对象。然而，在场景解析任务中，从建议框中直接进行分割，可能会使一些背景被错误地识别成对象，特别是当分类器并不是很精确，或者建议框内的背景和对象的相似度较高的时候。此外，这种方法还需要一个额外的网络来处理背景的分割问题，因为区域建议无法覆盖所有的像素区域，并且这些区域并不关心背景是什么。与此不同，本文并没有直接在建议区域中执行场景解析，而是利用这些区域去增强原有场景解析结果的局部区域的特征。具体来说，本文只是通过对特定的特征通道的特定区域进行特征加权，而加权的通道标签必须等于检测区域的类别标签。此外，本文使用对象轮廓的内接区域替代检测的内接矩形框。这个策略避免了对象区域增强对背景的影响。本文认为加权特定的特征通道可以最小化错误匹配。甚至于当一部分背景区域被错误地进行加权，也不会过分影响这些区

域的判别。因为在对象通道中，这些背景区域的特征概率强度相对于与背景相关的通道中的特征强度并不会很高，这些区域最高的概率值应该出现在具有与它们真实类别相同索引值的特征通道内。换句话说，在与对象相同索引值通道内的背景区域，即使受到对象特征加权的影响，最后获得的概率值，可能仍然小于与背景相同索引值通道内的概率值，在经过求最大值的操作后，最后生成的像素类别预测仍然由背景相同索引值通道的索引号决定。为了更好地理解这个策略，图 3.1 可视化了对象区域增强的过程。在示意图中，左图表示经过区域建议子网后生成的带类别信息的对象区域建议；中图是基于区域增强的过程；右图为经过建议框级区域增强后的图像解析结果。值得注意的是，区域增强只发生在特征图的通道索引和建议区域类别索引相同时。

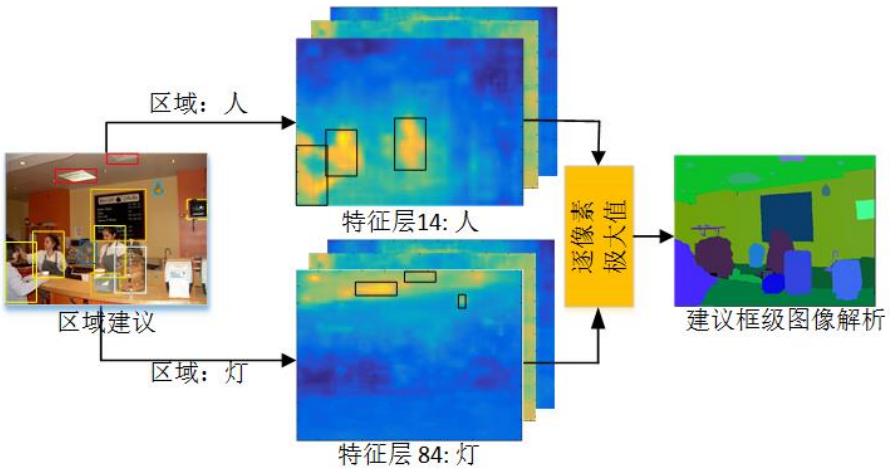


图 3.1 对象区域增强过程示意图

另一方面，无论是在检测还是在分割任务中，有一些区域始终很难决定他们究竟是什么内容。很多算法[100, 101, 118, 174, 175]通过增加一个额外的背景类（注：此处，本文将新增加用于改善性能的背景类定义为“额外背景”类，而原始样本中的天空、地板、草地等真实的背景定义为“背景”类），在训练中收集这些负样本或边缘样本，用于提高训练模型的健壮度。这个策略帮助训练一个更稳定的模型，但是它也会导致一些像素在推理阶段被错误地分配成额外背景类。这个问题对于语义分割和实例分割来说，并不是大问题，至少在视觉上来看并不是很显著的问题。因为语义分割和实例分

割主要关注的是识别特定类的样本，它们将其他像素都归结为“额外背景”。换句话说，额外背景可以被认为是一个真实存在的类，非目标区域都可以识别为额外背景，包括预先选择的额外背景区域，也包括不需要识别的小对象和场景中真实存在的背景。相对而言，在场景解析中，必须要处理每一个像素，并且都给它们分配一个类别，这包括所有的对象和所有的背景。这就意味着，如果一个像素被预测为“额外背景类”，那么它就是一种错误的分配，因为在 Groundtruth 中，它是有确定的类别信息的。因此，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，即使使用 CRF[119]来优化解析结果，也无法将所有预测为额外背景的区域修正为正确的类别。对于额外背景类，通常会使用“0”来进行编码，这些区域在视觉上看起来就像是一个个黑色的区域，因此本文称这个现象为“黑洞”。为了解决这个问题，本文使用概率值第二高的类别去替代额外背景类。因为这个类别应该具有比其他所有类别更加接近真实类别的特性。这个简单的策略称为“黑洞填充”。

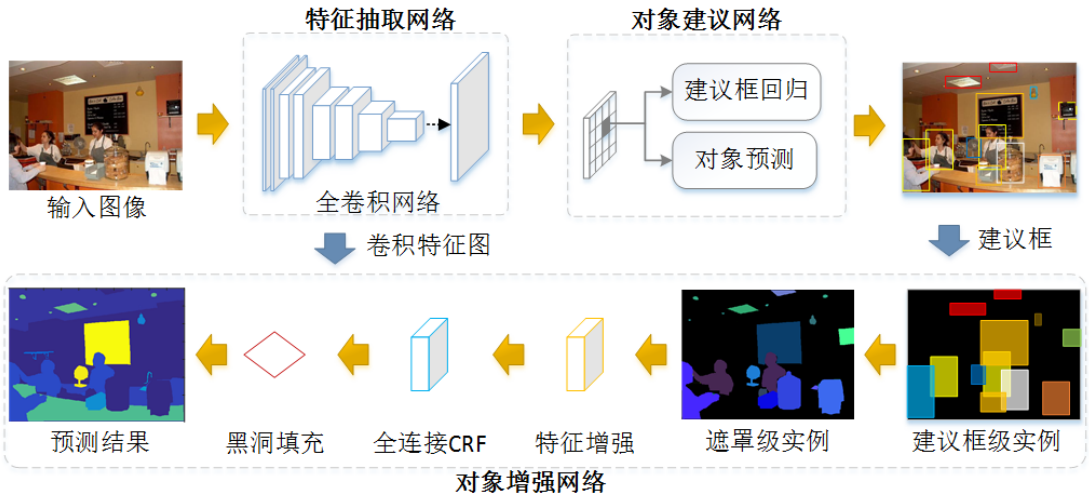


图 3.2 对象区域增强流程图

为了在场景解析中实现这两个策略，本文基于 DeepLab[118]模型构建了一个统一的框架。如图 3.2 所示，该模型包含三个子网络。第一个是特征提取网络（Feature Extracting Network, FEN）,它用于生成关于输入图像的卷积特征图；第二个是对象建议网络（Objectness Proposal Network，OPN），它用于定位和识别图像中的对象；最

后一个是对象增强网络（Objectness Enhanced Network, OEN），它用于像素级的预测进而实现场景解析。在这个网络的基础上，本文利用检测技术和黑洞填充策略来提高场景解析的性能。具体说，检测技术利用建议框级实例增强和遮罩级实例增强来完成对象区域增强。在实例增强之后，本文使用全连接 CRF[119]来完成对象边界的修复和精细化。黑洞填充策略被应用到网络的最后部分，用于处理那些被预测为额外背景类的对象和背景。

3.3 基于对象区域增强的深度网络

随着深度学习的发展，在基于 CNN 的语义分割和其他像素级的预测任务中，使用 FCN [118, 124, 174, 176, 177]的方法被证明是当前最有效的方法。受“Atrous 卷积”[118, 177]方法的启发，本文使用一个修改版本的 Resnet-101[21, 22]模型作为基准模型。首先，使用一个 151 路（150 个语义类和一个额外背景类）的 *Softmax* 分类器替换原始模型的 1000 路的 *Imagenet*[23]分类器。损失函数计算 CNN 输出特征图中每个空间位置的交叉熵的和。特征提取网络基于 *SceneParsing150*[131]数据集进行微调训练。

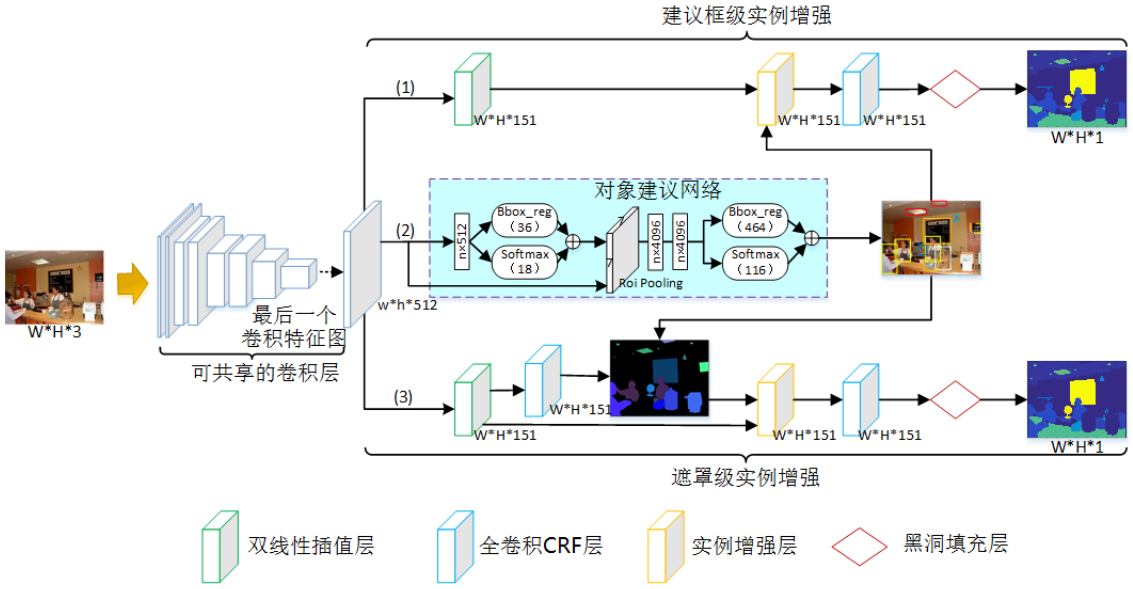


图 3.3 对象区域增强网络详细网络结构图

图 3.3 给出了对象区域增强网络（OENet）的详细网络结构图。整个 OENet 由三部分组成。输入图像首先被送入卷积层用于生成卷积特征图，该网络使用多级多尺度框架，并基于 Resnet-101 模型进行构建（第 3.3.1 节）。随后，生成卷积特征图被送入对象建议网络（分支 2）用于生成带类别信息的建议区域（第 3.3.2 节）。接下来这些特征图和建议区域被送入对象增强子网（第 3.3.3 节）（建议框级实例增强网络（分支 1）或遮罩级实例增强网络（分支 3））用于生成经过对象区域增强后置信度更高的卷积特征。最后，全连接 CRF 和黑洞填充策略被用来优化最终的解析结果（第 3.3.3 和 3.3.4 节）。

3.3.1 特征抽取网络

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。通过多级多尺度策略，本文将原始的基于 Resnet-101[21, 22]的基准模型扩展为多尺度的基准模型，并用该模型实现原始图像的卷积特征的提取。图 3.4 展示多尺度特征提取网络的共享卷积层部分。

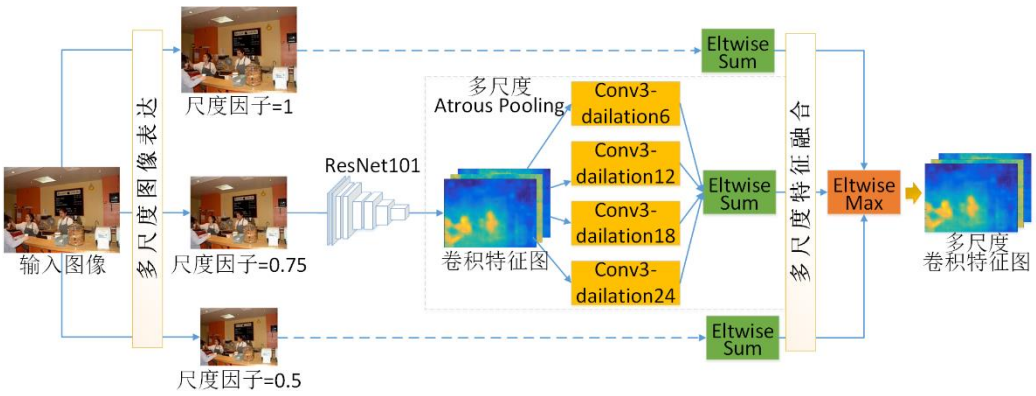


图 3.4 多级多尺度图像表达

本文使用了两种方法来实现多尺度策略，并用于场景解析。在第一阶段中，一个通用的多尺度训练方法[75]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 CNN 支路进行传输，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，三个支路所产生的卷积特征图，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[65]（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，本文采用 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。这个两级多尺度 CNN 特征提取网络称为特征提取网络（Feature Extracting Network, FEN）。它的输出可以被用来同时生成区域建议（第 3.3.2 节）以及对象实例（第 3.3.3 节）。最终的多尺度特征可以公式化为：

$$\mathcal{F}_M = \max_{m\in[1,\dots,M]} \sum_{n=1}^{N} \mathcal{F}_{(m,n)} \tag{3.1}$$

在本文中，支路 M 和 N 分别为 3 和 4。 $m = 1,2,\dots,M$ 用来表示分辨率为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。在本文的工作中，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。

如表 3.1 所示，多级多尺度处理有效地改进了性能，但是它也增加了前向推理的时间，并且大大增加了 GPU 显存的消耗。

3.3.2 对象建议网络

识别错误和边界错误是场景解析最关键的两个问题[121]。识别错误指对象被误认为错误的类别，或没有找到类别。图 3.5 给出了对象区域增强改进场景解析结果的示意图。其中，第一行树丛（Tree）和公共车（Bus），第三行的三轮车（Tricycle）在基准模型中丢失了，而椅子（Chair）和桌子（Desk）在第二行中被识别成错误的类别。图 3.5 的示意图展示了对象增强策略改进了树丛、公交车、椅子、桌子和三轮车的解

析结果。由此可见，对象区域增强策略可以为找回这些丢失的对象提供帮助。另一方面，全卷积 CRF 被用来处理边界错误，这个错误主要是发生在语义标签在对象边缘处产生的过平滑的错误预测。为了保证完整性，第 3.3.3 节将简要介绍全连接 CRF[119]技术。

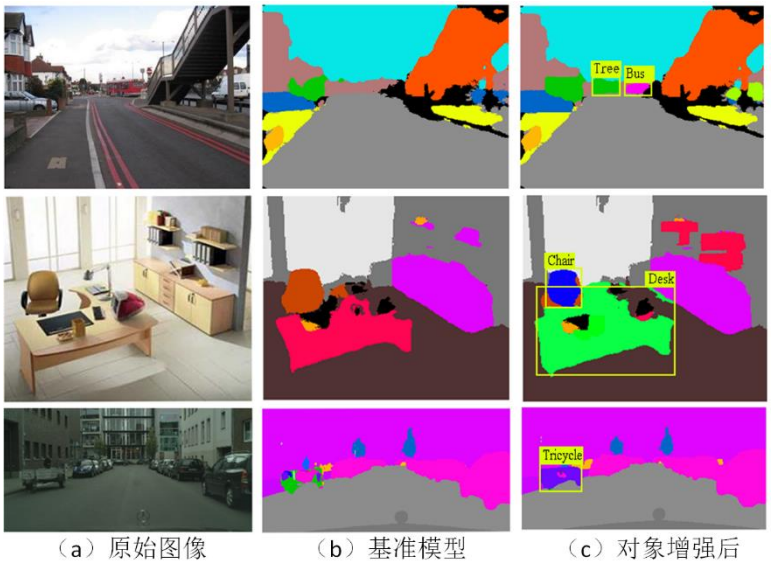


图 3.5 对象区域增强改进场景解析结果的示意图

受文献[65, 100, 101]启发，对象建议网络用于从复杂的背景元素和重叠的对象中找到特定的对象，以优化分割结果。不同于文献[178]，本文使用的“Objectness”不仅需要标识出区域是否为对象，还要识别出是什么对象。对象建议网络紧跟在特征提取网络之后，并使用后者的多尺度卷积特征图作为输入。不同感知域用来表示图像不同的区域，并同时执行分类和边界框回归，用来估计对象的区域定位和类别概率。为了训练对象建议网络，本文直接使用分割对象的 Groundtruth 的外接矩形框作为检测 Groundtruth，没有使用其他的信息。在本章的工作中，*SceneParsing150* 数据集中的 115 个离散的对象（例如：汽车，人，桌子）被用来训练对象建议网络。加上另外的 35 个背景类，总共 150 个类别被用来评估模型。这个定义遵循 *SceneParsing150*[131]数据集的评价标准。

对象建议网络的结构和损失函数设置遵循区域建议网络（Region Proposal

Network，RPN）[101]和快速 R-CNN[100]检测网络，这里简要介绍与这两个网络相关的内容。PRN 和检测网络都使用全卷积形式预测边界框的位置和对象分数。唯一不同的是 RPN 的边界框是没有类别信息的，它们只用来识别一个区域是否为对象。图 3.3（分支 2）显示了对象建议网络的框架结构。在共享卷积层的顶部，一个 3×3 的卷积层用于实现维度降低，接下来一个双支路的 1×1 的卷积层用于实现边界框回归和分类。RPN 使用一个多任务损失来训练区域建议网络，它可以用公式 3.2 表示：

$$L(R(k,k^*,t,t^*)) = \frac{1}{N_{cls}} L_{cls}(k,k^*) + \lambda \frac{1}{N_{reg}} k^* L_{reg}(t,t^*) \tag{3.2}$$

此处，使用k^*和$t^* = (t_x^*, t_y^*, t_w^*, t_h^*)$表示 Groundtruth 标签和 Groundtruth 边界框，关于类别k^*的预测使用元组$t = (t_x, t_y, t_w, t_h)$表示。此外，$k = (k_0, k_1, \ldots, k_K)$ 是一个基于 K+1 个类的离散的概率分布（额外的 1 表示额外背景类），$L_{cls}(k,k^*) = -\log p_{k^*}$是标准的两类（对象或非对象）交叉熵损失。第二项损失$L_{reg}(t,t^*) = \sum_{i\in x,y,w,h} R(t_i^* - t_i)$，其中$R(*) = smooth_{L_1}(x) = \begin{cases} 0.5x^2, & if\ |x|<1 \\ |x|-0.5, & otherwise \end{cases}$，是一个鲁棒的平滑的$L_1$损失[100]函数，对于离群点它没有$L_2$损失[39]敏感。两个损失函数由参数$N_{cls}$和$N_{reg}$进行规范化，并由权重参数$\lambda$进行平衡。其中$N_{cls}$是 mini-batch 的大小，$N_{reg}$是定位锚点的数量。例如，在 VGG16 模型中，$N_{cls} = 64$，$N_{cls} = 2400$。默认情况下，设置$\lambda = 10$，因为对于区域建议网络，位置回归的重要性要高于类别判定。

最后，可以得到一系列的边界框作为输出，其中每个边界框可以用一个多元组$R_i = \{x_i, y_i, w_i, h_i, p_i, c_i\}$ 来表示，其中 i 是R_i的索引。边界框R_i的坐标中心为(x_i, y_i)，宽为w_i，高为h_i，同时对于类别c_i其分类概率为p_i。

3.3.3 对象增强网络

对象增强网络（Objectness Enhancement Network，OEN）以多尺度特征作为输入，实例感知的解析结果作为输出。级联的 OEN 网络包含三个阶段：建议框级实例增强，遮罩级实例增强和全连接 CRF[119]。三个阶段可以被认为是三个独立的模块，既可以

独立存在，也可以顺序叠加到网络中，以优化基本的解析结果。OPN 输出的基于类别的区域建议被作为辅助输入与卷积特征图共同用来计算最终的解析结果。如图 3.6（c）和图 3.6（f）所示，OENet 可以产生两种不同的解析结果。（1）建议框级增强解析由建议框级实例和多尺度卷积特征生成；（2）遮罩级增强解析由遮罩级实例和多尺度卷积特征生成。其中，遮罩级实例可以通过建议框级实例和多尺度卷积特征计算获得。在这一小节中，将顺序介绍这几种技术：建议框级实例，遮罩级实例，实例增强和全连接 CRF。

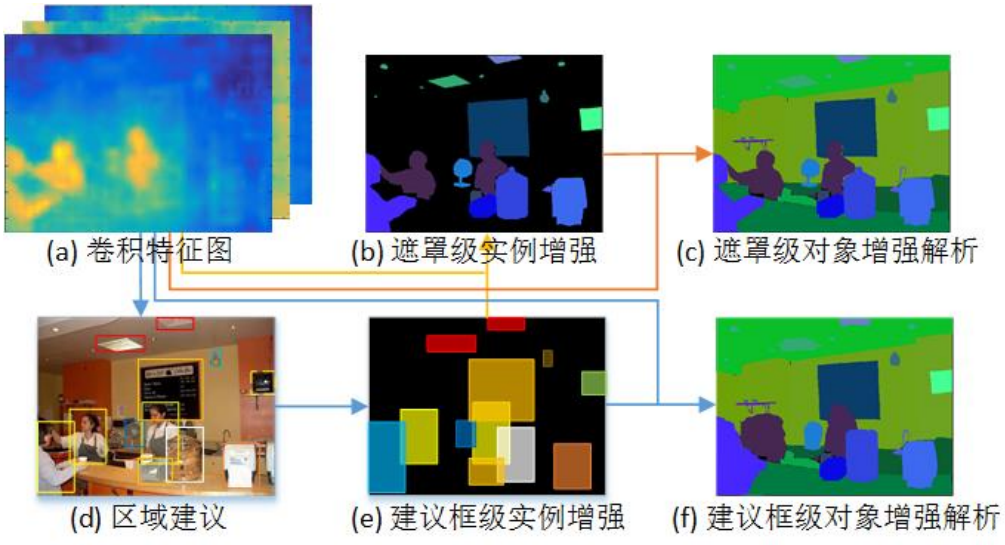


图 3.6 对象增强的流程图

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，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：

$$B_i(W, H, c_i) = \begin{cases} 1 & p \in R_i(t_i) \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

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。如图 3.3 所示，可以将建议框级实例和多尺度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$B_i(W,H,p)$，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：

$$M_i(W,H,c_i)=\begin{cases}1 & p\in F(W,H,c_i)\cdot R_i(t_i,c_i)>t\\ 0 & \text{otherwise}\end{cases} \tag{3.4}$$

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。从公式 3.4 中可以发现，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，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，通过迭代所有的区域R_i可以获得完整的特征图$M(W,H,C)$，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，可以将它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：

$$F_{ME}(W,H,C)=F(W,H,c_i)\cdot M\left(W,H,c_j\right)\times w\times p_i\times c^* \tag{3.5}$$

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 3.5 中，通过替换 $M\left(W,H,c_j\right)$为

$B_i(W, H, c_j)$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结果趋于一致。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，本文将基于全连接的条件随机场[118, 119]（Conditional Random Fields，CRF）集成到 OENet 模型中，作为一种后处理操作。该操作使用以下能量函数进行优化：

$$E(x) = \sum_i (-\log P(x_i)) + \sum_{ij} \mu(x_i, x_j) \left[\lambda_1 exp\left(-\frac{\|p_i - p_j\|^2}{2\delta_\alpha^2} \right) + \lambda_2 exp\left(-\frac{\|p_i - p_j\|^2}{2\delta_\beta^2} - \frac{\|c_i - c_j\|^2}{2\delta_\gamma^2} \right) \right] \tag{3.6}$$

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。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 3.6 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δ_α，δ_β和δ_γ控制高斯核的尺度，权重参数λ_2和λ_2用来平衡两个特征的重要性。

在本章的工作中，所有的实验都将全连接的 CRF 作为后处理操作。

3.3.4 “黑洞”填充策略

“额外背景”类作为一个特殊类别，经常被附加到检测和分割任务的类别空间中。由于这个策略可以从正样本集中移除那些模棱两可的样本，从而使分类器更健壮。所以，它能够显著地改进不同系统的性能。然而，在场景解析任务中，在推理阶段，一些像素可能会被分配到这个并非真实存在的额外背景类。这是一个显而易见的分类错误，这个问题可以称为“黑洞”。如公式 3.7 所示，本文设计了一种非常简单的方法来克服这个问题。

$$\mathcal{O}_{L(i,j)} = \underset{c\in[2,\ldots,N]}{Argmax}\, \mathcal{O}_{F(i,j,c)} \tag{3.7}$$

其中，$\mathcal{O}_{F(i,j,c)}$是 CNN 计算获得的特征，i 和 j 是像素的坐标位置，c 是卷积特征图的特征通道数量，它等于真实的类别总数加一个额外背景类。通常情况下，某个像素在所有通道中最大概率的通道索引值将被定义为该像素的标签。黑洞填充策略首先将通道$c = 1$（额外背景类）排除，然后再计算卷积特征图每个像素位置的最大概率，最后将最大概率的通道索引设置为预测标签。换句话说，如果预测的概率类为 1，那么就指定第二大的概率通道索引为标签，以此来修复黑洞问题。这个策略，同样显著改进了场景解析的精度。

3.4 实验与分析

3.4.1 MIT ScenParsing150 数据集

1. 数据集

本文使用 MIT *SceneParsing150*[131]场景解析数据集来评估各种模型的性能，原始的数据集包含 20,201 个训练图像，2,000 个验证图像，所有的图像都包含一个基于像素的标注图像。根据标准评估指标，使用平均 IoU 和像素分类精度来衡量性能。为了训练区域增强网络，本文在分割标签图像上，采用对象的外接矩形抽取指定对象的检测标签。数据集总共有 150 个类，其中 115 个定义为对象类（如：汽车、人、桌

子），其他 35 个为背景类（如：墙、天空、道路）。边界框 Groundtruth 只由这 115 个对象类产生。

2. 实现细节

为了训练 OENet，*MSCOCO* 数据集[90]被用于预训练 ResNet101 模型，一个 5 步的训练优化方案如**算法 3.1** 所示。

算法 3.1：对象区域增强网络（OENet）训练过程

步骤 1：在 *MSCOCO* 数据集[90]上预训练 ResNet101[21, 22]深度卷积神经网络模型。

步骤 2：使用*步骤 1* 中预训练的模型作为初始化网络，并用交叉熵损失训练多级多尺度特征提取网络（FEN）。该网络用于初始化其他网络，同时用做基准模型。

步骤 3：依据文献[118]的建议，采用交叉验证方式搜索全连接 CRF 参数。

步骤 4：使用*步骤 2* 中预训练模型进行初始化，训练对象建议网络 OPN。在该步骤中，首先训练 RPN 子网络，然后使用 RPN 生成的建议框训练检测网络。整个过程中，共享卷积层始终保持固定不定。

步骤 5：按照图 3.3 的结构输出统一的 OENet 模型，该模型通过集成第 2 步、第 4 步训练生成的模型，CRF 模块、区域增强模块和黑洞填充策略模块得到。

对于 FEN 和 OEN 网络，使用“ploy”学习率策略（学习率 $lr_{iter} = lr_{iter-1}\left(1-\frac{iter}{\max\_iter}\right)^{power}, power = 0.9$）进行参数学习，设置 mini-batch 为 10，初始学习率为 0.0025。初始动量为 0.9，权重衰减指数 0.0005。特征抽取网络在训练集上完成微调，并使用 10 均值迭代的方式交叉验证[118]CRF 的超参数。固定$\lambda_1 = 3$和$\delta_\alpha = 3$，然后使用交叉验证方式在 200 幅图片上搜索最优的$\lambda_2, \delta_\beta, \delta_\gamma$。在搜索过程中，使用粗到细的策略，首先使用参数范围$\lambda_2 \in [3:6], \delta_\beta \in [3:6], \delta_\gamma \in [30:10:100]$，然后在第一轮最好的值范围内，细化参数$\delta_\gamma$的搜索范围，从而获得最后的参数。接下来使用“step”学习率策略训练 OPN，初始学习率为 0.001，对于 RPN 和检测网络，分别在迭代 40,000 和 80,000 次时调节一次学习率。两个网络都使用 0.9 的动量和 0.005

的权重衰减。在训练多尺度网络时，对所有原始图像统一进行降采样处理，将长边固定为 500 像素，短边按比例缩放。为了扩展数据集，根据训练和测试协议，分别将图像裁剪为 513 像素和 321 像素。所有实验都基于开源工具包 CAFFE[169]，并且使用 NVIDIA Titan X GPU。

3. 消融研究

本节将对 OENet 所包含四个重要组件的有效性以此进行评估，包括：多级多尺度特征表达、建议框级区域增强、遮罩级区域增强和黑洞填充策略。基于这四个策略的 5 个变种的 OENet 模型在 *SceneParsing150*[131]数据集的 150 类别的验证集上的测试结果如表 3.1 所示。使用同样的训练协议，多级多尺度显著改善了性能，分别获得了平均 IoU 和像素准确率 4.1%和 1.5%的性能提升。如前所述，对象区域增强用于召回那些丢失的局部困难对象，它在一定程度上可以提高系统的解析性能。但是，建议框级对象增强，由于使用的是方形的建议框，它不仅包含了对象区域，也同时包含了部分对象周围的背景区域。因此，它不仅会召回丢失的对象，同时也会将一些背景区域错误地指派为对象类，从而损害整幅图像像素准确率的平均值。如表 3.1 第三行所示，通过增加建议框级对象增强，平均 IoU 得到了 0.6%的改善，但是像素准确率反而下降了 0.4%。遮罩级对象区域增强有效地解决了这个问题，该策略使平均 IoU 上升了 0.9%，像素准确率提高了 0.6%。黑洞填充策略通过修正额外背景带来的错分类问题，作为最后一个被添加的模块，获得了 1.9%和 2.2%的性能提升。

表 3.1 OENet 的各种策略在 *SceneParsing150* 上的性能评估

多级多尺度	建议框级增强	遮罩级增强	黑洞填充	平均 IoU	像素准确率
				30.9	74.0%
√				35.0	75.5%
√	√			35.6	75.1%
√	√	√		36.5	75.7%
√	√	√	√	38.4	77.9%

表 3.2 OENet 各种策略在 *SceneParsing150* 每个类上的性能评估

	mIoU	Acc.	wall	build	sky	floor	tree	ceili	road	bed	pane	grass	cabin	side	pers	earth	door	table	moun	plant	curt	chair	car	water	paint
Base	30.9	74	66	75.7	92.3	69.7	67.2	73	73.8	73.9	48.7	**65.6**	47.9	47.5	67.4	27.8	20.4	38.3	51.7	41.9	60.1	36.7	72	45.4	57.4
MMC	35	75.5	69.4	77	93.2	73.7	**69.2**	77.3	77.5	**81.7**	**54.6**	62.8	50.8	55.9	70.7	22.9	26.4	**47.6**	**49.9**	41.5	65.6	44	78.2	47.1	63.8
Box	35.6	75.2	68.9	76.8	93.2	71.3	69	77.4	76.8	72.9	52.2	63.4	**54.5**	55.7	66.5	22.5	**34.1**	42.3	49.2	**44.6**	64.1	45.1	73.2	46.9	62.9
Mask	36.5	75.7	**69.7**	76.9	93.2	**73.8**	68.7	**77.4**	77.5	81.3	53.5	62.9	52.9	**55.9**	**71.2**	22.8	29.2	46.9	49.5	40.5	**65.9**	**45.8**	78.7	47.1	**63.5**
BH	38.4	77.9	69.4	**77.7**	**93.3**	73.3	68	77.2	**77.7**	81.2	53.4	63.6	53	55.5	70.6	**29**	30.9	46	49.1	43	65.6	45.4	**79**	**47.8**	63.2

	sofa	shelf	house	sea	mirro	rug	field	arm	chsea	fence	desk	rock	ward	lamp	bath	rail	cush	base	box	col	sign	chest	count	sand	sink
Base	45.5	28.6	43.9	46.2	39.2	36.4	**30.8**	17.6	37.6	25.4	31.4	31.8	39.7	35.4	54.2	24.7	26.2	11.2	9.1	33.5	22	36.8	28.3	18.9	44.1
MMC	54.5	34.6	34.7	52.2	52.7	39.6	23.8	29.6	50	26	38.6	34.9	42.2	44.7	63.2	23.6	38	16.8	7	**38.9**	23.3	41.6	27.9	28.1	52.7
Box	54	34.7	34.7	51.6	54.4	37.3	22.7	**36.1**	50	25.6	34.3	**37.5**	43	42.1	59.7	22.4	31.1	16.6	10.8	36.7	22.8	46	27.6	28.7	53.1
Mask	**54.9**	34.5	34.8	52.1	54.2	39.6	22.9	33.2	51.6	25.8	**39.9**	33.9	44.6	**44.5**	62.9	23.6	38.3	16.8	9.8	38.2	23.3	48.1	**28.2**	28.1	57.1
BH	54.5	**34.9**	**39.3**	**55**	**54.6**	**39.7**	21.3	32.9	**53**	**28**	38.9	32.6	**45.5**	44.4	**63.2**	**24.8**	**39**	**18.8**	**13**	37.9	**24.2**	**48.8**	27.9	**34**	**59.1**

	skys	fire	refri	grand	path	sta	runw	case	pool	pill	scree	stair	river	brid	book	blind	Coff	toil	flow	book	hill	benc	count	stove	palm
Base	54.7	48.6	43.3	30.4	15.2	**27**	**63.3**	32	86.5	34.2	30	21.1	10.5	18.1	30.7	13.1	34.4	63.8	20.2	28.1	**6.3**	34.4	39.2	47.4	**40.3**
MMC	66.1	63.5	67.1	31	20.6	22.5	57.1	29.9	**88.4**	**38.3**	35.1	22.6	14.7	23.4	27.6	14.9	48.2	77.7	25.6	33.6	5.8	33.6	42.1	**55.6**	35.9
Box	66.2	66.2	64.7	29.3	20.6	22.4	57.1	28.3	88.1	21	30.3	22.6	14.3	23.4	**32.2**	**24.3**	42.5	73.2	24.7	20.3	5.8	33.1	40.3	49.3	31.3
Mask	**66.2**	**64.4**	**73.9**	31	**20.6**	22.5	57.2	29.4	**88.4**	37.5	32.8	22.6	14.6	23.5	30	19.9	48.8	**78.7**	25.2	33.9	5.8	**34.5**	44.5	51	32.6
BH	65	64.3	73.2	**35.6**	20.1	21.9	62.4	**35**	88.3	37.2	**36.8**	**22.7**	14.1	**40.7**	29.9	19.9	**48.8**	77.9	**28.6**	**36.6**	5.8	33.6	**46.6**	52.6	32

	kitc	comp	swiv	boat	bar	arca	hovel	bus	towel	light	truck	tower	chan	dawn	light	booth	tv	airpl	dirt	app	pole	land	bann	esca	otto
Base	29	44.9	31.1	42	27.4	27.7	**21.4**	63.7	38.8	15.4	8.5	26.1	43.9	10.4	3.6	30.6	47.4	50.4	0	20	4	0	4	**32.1**	21.1
MMC	24.5	54.9	31.5	41.4	24.5	25.5	4.8	83.3	39.5	21.2	21.1	32.9	52.5	12.5	**9.1**	37.7	57.4	52.1	0	17.7	5.5	2.6	2	5.7	27
Box	27.6	45.9	37.6	**57.1**	24.3	31.1	4.8	82.4	40.3	**21.2**	**25.5**	**33**	48.3	**23.6**	7.5	37.4	**62.4**	46.7	0	14.5	**10.5**	2.6	2	5.9	**31.2**
Mask	**29.8**	**56.2**	38.9	46.3	26.3	25.6	4.8	83.4	40.5	21.2	24.5	32.9	54.9	20.8	8.9	37.9	60.2	**52.4**	0	17.1	8.8	2.5	2	5.7	30
BH	29.2	55.6	**40**	51.9	**27.8**	**39.4**	10.1	**84.4**	**41.4**	21.2	24	32.7	**54.9**	20.6	9	**44**	58.1	46.6	0	**20.6**	8.9	**5.4**	**6**	5.7	29.7

	bot	buff	post	stage	van	ship	fount	conv	cano	wash	play	pool	stool	bar	bask	water	tent	bag	mini	crad	oven	ball	food	step	tank
Base	2.1	29.3	1	1.3	23.3	26.7	17.4	35.8	4.7	50	**18.6**	**18**	13.8	12.4	6	**67.1**	73.1	2.6	27.6	57.5	4.7	36.2	22.4	6	27.2
MMC	16.2	38.7	9.5	4.3	29.8	4.3	1.5	37.5	13.8	36	10.5	17.7	26	37.9	8.9	34.6	74.8	4	54.4	75.5	18.6	38.5	4.7	0	30
Box	20	**40.9**	**21.7**	4.6	21.8	18.9	19.9	38.3	**20.8**	34	4.4	17.9	**32.1**	**41.2**	**18.1**	34.5	**74.9**	2.5	42.5	67.5	**19.2**	31.8	17.8	4.7	34.6
Mask	18.2	39.2	12.8	4.3	42	4.6	19.9	37.5	14.4	35.9	10.4	17.7	28	38.7	13.9	34.7	74.8	1.9	52.2	75.6	17.2	38.4	15.7	7.3	37.2
BH	**26.7**	39.1	12	**6.4**	**42.3**	**27.2**	**19.7**	**50.3**	14.2	**50.8**	15.6	17.5	29.2	37.4	16.4	40.2	67.6	6.3	**55**	**75.6**	14.8	**40.4**	**54.4**	**8**	**40.2**

	trade	micro	pot	anim	bicy	lake	dish	scr	blank	sculp	hood	scon	vase	traf	tray	ash	fan	pier	crt	plate	moni	bull	show	radi	glass
Base	14	28.7	19.6	29.4	34.1	2.4	29.1	60.9	0	11.2	23.1	6.5	8.3	9.5	0	9.3	**34.3**	**26.8**	0	11.3	6.5	29.2	0	15.7	1.9
MMC	**13.4**	43.2	13.4	29.2	37.7	41.9	47.1	68.9	2.6	32.3	35	19.9	21.8	18.5	2.1	25.4	30.3	12.7	21.9	18.5	5	32	1	30.2	1.3
Box	12.4	44	11	33.3	**45**	41.9	**49**	**69.8**	0	34.5	24.4	30.3	23.8	26	3.6	22.3	29.2	12.7	22.2	25.5	13.7	**38.8**	**2.5**	34.1	**5**
Mask	12.7	44.7	13.4	33.1	38.9	**41.9**	48.5	68.9	4.1	34	35.1	31.8	32.7	**20**	4	24.8	30.2	12.7	**23**	26.6	18.3	32.3	2	**43.3**	1.5
BH	12.7	**56.8**	**20.7**	**52**	40.5	41.4	48.6	68.9	**4.4**	**41.9**	**35.3**	**31.9**	**33.2**	19.9	**6.5**	**28.1**	30	12.2	22.9	**39.5**	**19.3**	29.4	2	43.1	4

注：蓝色的表示 35 个背景类。下划线表示该类中性能最好的结果。

为了进一步分析本文算法的可信度，本文进一步详细研究了每个类的IoU性能。表 3.2 给出了这方面研究的数值结果。首先，可以发现性能改善并没有明显的偏差，也就是说对象和背景同时得到了改善，虽然本文的算法主要关注于召回对象。第二，总共有 109 个类在多尺度模型的作用下获得了性能提升，大约占 72.7%，这个结果也和类别的分布无关。第三，经过对象区域增强后，有 70.7%的类别超过了多尺度模型。例外的类别包括 15 个背景类和 29 个对象类，但是这些对象类实际上也可被认为是背景类，例如：树丛、栅栏、栏杆、浴盆等。这些结果验证了对象区域增强的可行性。最后，经过黑洞填充策略，60%的类别获得了性能提升。总的来说，经过 OENet 的优化，有 88%的类别的性能都超过了基准模型。

4. 区域建议评估

为了评价不同区域建议方法对场景解析的影响，本文集成了不同的对象检测网络来统一评估 OENet。用于对比的对象建议网络分别基于 Faster RCNN[101]和基于 RPN[101]的 RFCN[175]网络来实现，它用于产生建议框级对象实例。两个网络都使用同样的共享卷积层，只有检测部分不同。此外，Groundtruth 建议框被作为上界标准加入到评估实验中。从表 3.3 可以看出，基于 Faster RCNN[101]的解析结果比基于 RFCN[175]的解析结果略好一些，这基本符合两个检测器检测精度的规律。值得鼓励的是，运行在 Groundtruth 建议框上的解析结果达到了 47.8 和 80.3%的好结果。这个上界意味着，本文提出的基于对象区域增强的方法仍然有较大的改进空间。

表 3.3 不同区域建议方法在 *SceneParsing150* 上的性能评估

方法	检测精度	平均 IoU	像素准确率
Groundtruth	100%	47.8	80.3%
Faster RCNN[101]	84.4%	38.4	77.9%
Region FCN[175]	82.3%	38.0	77.7%

5. 整体性能对比

为了在 *SceneParsing150*[131]数据集上进行场景解析的性能评估，本文将文献[118,

131, 174, 179, 180]和一个基于 ResNet101 模型构建的分割网络作为基准模型。FCN[174]在像素级分割任务上多次对激活特征进行升采样操作；SegNet[180]构建了一个编解码的图分割网络；DilatedNet[179]在 VGG16 的基础上丢弃了 *pool4* 和 *pool5* 层，并将后续的卷积层使用 *Dilated* 卷积进行替换。Cascade-SegNet[131]和 Cascade-DilatedNet[131]构建了一个多支路级联模型用于同时识别背景、对象和部件，多个支路共享一个卷积结构，这意味着他们混合了多种不同类型的语义信息用于场景解析。DeepLabv2[118]是一个多尺度 ResNet101 模型，它融合了 atrous 空间金字塔池化和全卷积 CRF。本文的基准模型也基于 DeepLab 模型原型完成，但是没有使用多尺度的 ASPP 方案。因此，基准模型明显比 DeepLab 模型要差，但是在融合了区域增强和黑洞填充策略模块后，基于多级多尺度的 OENet 整体性能要略优于 DeepLab[118]原型模型。

表 3.4 给出了各种算法在 *SceneParsing150* 上的性能评估。结果表明，融合了所有模块的最终版模型，完成了 38.4 的评价 IoU 和 77.9%的像素精准率，这个结果在两个指标上显著超越了基准模型 7.5%和 3.9%。同时这个结果也优于其他对比模型。

表 3.4 各种算法在 SceneParsing150 上的性能对比

方法	平均 IoU	像素准确率
SegNet[180]	21.6	71.0%
Cascade-SegNet[131]	27.5	71.8%
FCN8s[174]	29.4	71.3%
DilatedNet [179]	32.3	73.6%
DeepLabv2[118]	34.3	75.3%
Cascade-DilatedNet[131]	34.9	74.5%
Baseline	30.9	74.0%
OENet	38.4	77.9%

6. 定性分析

在图 3.7 中可视化了基准模型以及融合了遮罩级对象区域增强策略和黑洞填充策略后的场景解析结果。其中每个子图分别表示：（a）原始图像，（b）Groundtruth 参考标准图，（c）基准模型完成的解析图，（d）在基准模型上使用多尺度表达和对象增强策略后的解析图，（e）在基准模型上使用多尺度表达，对象增强策略和黑洞填充策略后的解析图。

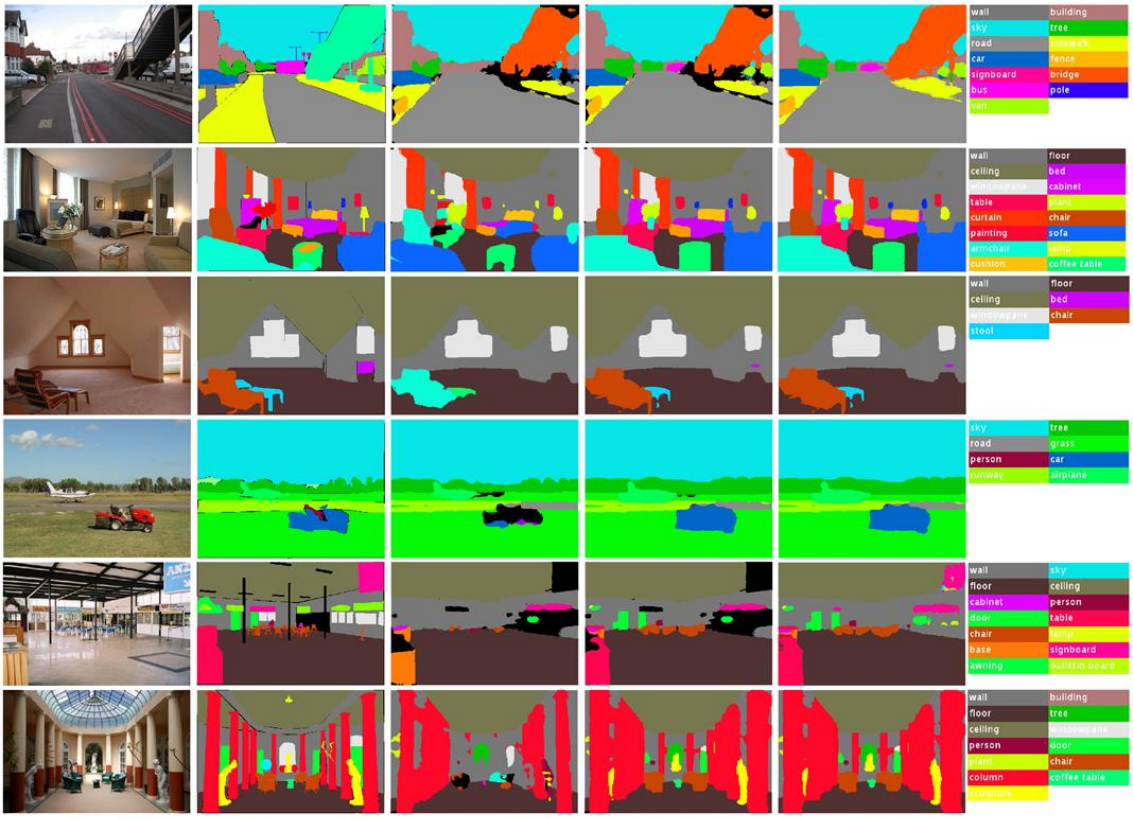


图 3.7 *SceneParsing150* 数据集上场景解析结果示意图

如图 3.7 所示，基准模型丢失了一些对象，而对象区域增强可以找回这些丢失的对象。受益于上下文信息，对象建议网络可以收集那些相关的像素，并将它们组合在一起构成对象，特别是去发现那些比较小的对象（例如：第二行的灯第六行的茶桌）以及那些被湮没在大量背景中的对象（例如：第一行的红色公共车和白色货车，第三

行的椅子）。黑洞填充策略的优点也非常明显。它能够找到那些被分配到额外背景类的像素，这些错误看起来就像是图片上的一个个黑色窟窿，黑洞填充策略试图找到一个合适的类别去填充它。由于 CNN 是一个基于概率的预测方法，因此，如果预测结果是额外背景类，那么可以用概率第二高的类别去替换额外的背景类。重新发现的第一行中的地面、第四行的汽车和第五行的门就是背景填充策略有效性的有利证据。

7. 失败案例分析

图 3.8 给出了一些 *SceneParsing150* 数据集上错误案例示意图。每个子图分别表示：（a）原始图像，（b）参考标准图，（c）基准模型完成的解析图，（d）在基准模型上使用多尺度表达和对象增强策略后的解析图，（e）在基准模型上使用多尺度表达，对象增强策略和黑洞填充策略后的解析图。

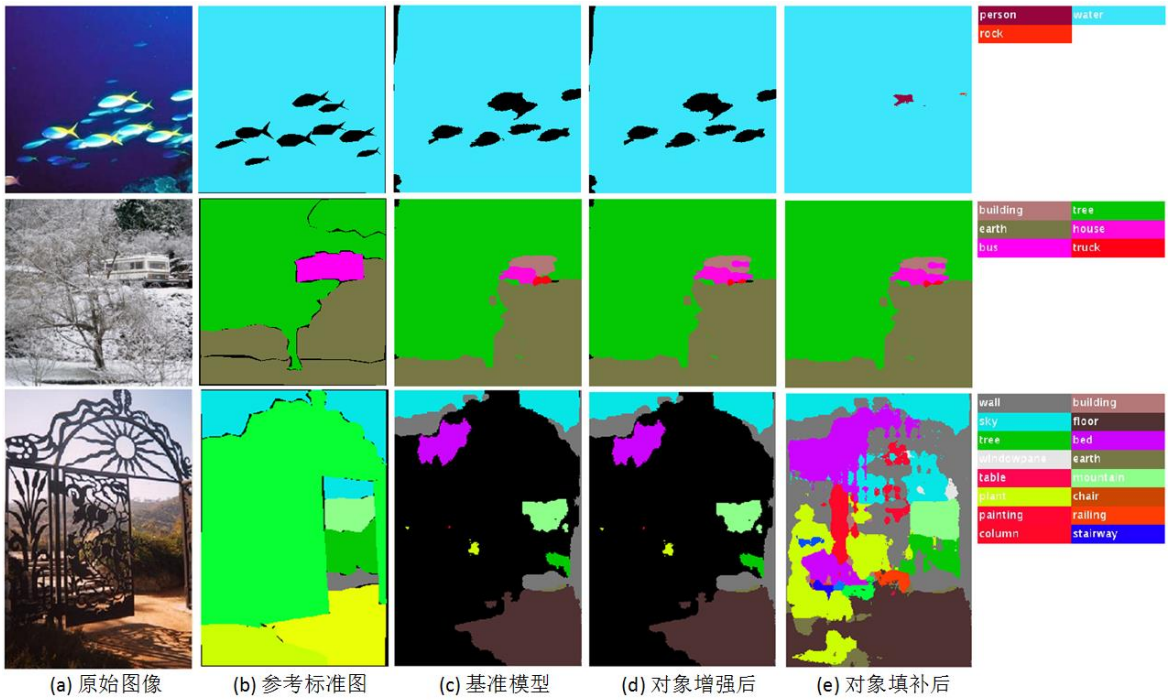


图 3.8 *SceneParsing150* 数据集上错误案例示意图

如前所述，对象增强和黑洞填充策略具有很多有用的属性，但它们也并非总是有效。主要的包括三个方面的问题。1）自然界的物种通常具有多样性和复杂性，对于

那些并没有存在于训练集中的对象，无论是分类器还是检测器都无法识别它们。这些区域可能会被识别成周围的区域或者相似的对象。例如：图 3.8 第一行所示，在使用黑洞填充策略之后，海中的鱼从背景海水中消失了。2）在一些场景中，由于对象和背景非常相似，分类器和检测器可能会被视觉感官所欺骗。一方面，分割网络可能会输出错误的分类结果；另一方面，检测网络无法找到任何目标对象。例如：图 3.8 第二行所示，公共汽车没有被分割网络完全识别出来。而检测器也没有检测到公共汽车，最终导致三个输出结果在视觉上基本一致。3）对于一些复杂的场景，特别是那些有对象交叉或者对象覆盖的场景，分类器和检测器都会感到非常困惑。例如图 3.8 第三行所示，由于后面的风景被铁门遮挡，导致了整个解析结果变得非常混乱不堪。

3.4.2 Cityscapes 数据集

1. 数据集

Cityscape[181]数据集是最近推出的一个大规模的、高分辨率的用于理解场景的数据集。它收集了 50 个城市的公路场景，包含 5,000 个精细的细粒度的像素级标签。整个数据集总共定义了 19 个类别，包括不同的对象和背景。此外，有 20,000 个粗粒度标注的图像提供。本章的工作没有使用这些样本。总的来说，训练集、验证集和测试集分别包含 2975，500，1525 个精细标注的图像。

2. 实现细节

与 *SceneParsing150*[131]数据集类似，ResNet101 网络同样被用于提取特征，训练过程与**算法 3.1** 一致。由于 GPU 显存的限制，在 *Cityscape*[181]数据集上没有使用多尺度方式表达图像，但是保留了基于多感知域的 atrous 池化。原始图像所使用的 2048×1024 的高分辨率对于使用有限的 GPU 来训练深度网络是一个较大的挑战，但同时也有利于生成更精细的结果。为了不降低解析精度，本文没有直接对原始图像执行尺度变化来压缩图像的分辨率，而是将原始图片按照 705×705 的分辨率进行有覆盖的裁剪，每幅图像都裁剪为 8 个图像片，这些图像片覆盖了原始图像的所有区域，彼此

间有一定的重复区域。这些图像片在训练的时候不进行降采样操作。为了进行数据扩展，训练和测试输入尺寸，分别按照分辨率 545×545，705×705 进行裁剪。其他超参数的设置与 *SceneParsing150*[131]数据集相同。

3. 整体性能对比

为了验证性能，本文分别在验证集和测试集上与一些优秀的算法进行了性能对比，表 3.5 是在验证集上运行的结果。从结果上看，高分辨率在使用对象区域增强前后，分别带来了 2%和 3.1%的性能提升。相比基于 VGG16 的模型，基于更深的 Resnet101 模型的 Baseline 的解析结果有较大的提升。整个本文提出的多种策略后，OENet 在 *Cityscapes* 上的性能有了进一步的提升。OENet 对于使用高分辨率图片前后，分别带来了 0.7%和 1.8%的性能提升。这主要是因为更高的分辨率有利于对象检测网络发现更多、更小的对象，从而提高整体的性能。因此，OENet 在高分辨率下优势更加明显。

表 3.5 各种算法在 *Cityscapes* 验证集上的性能对比

方法	平均 IoU
VGG16	
DeepLabv2-VGG16[118]	62.9
FCN[174]	63.4
Pixel-level Encoding	64.3
DPN[125]	66.8
DiatedNet[179]	67.1
Adelaide[128]	68.6
ResNet-101	
DeepLabv2-Resnet101[118]	71.4
Baseline	69.3
Baseline-高分辨率	71.3
OENet	70.0
OENet-高分辨率	73.1

此外，作者将基准模型和使用高分辨率图像的最优模型上传到官方的评估服务器。如表 3.6 所示，是各种算法在 *Cityscapes* 测试集上的性能对比结果。该结果显示，在结合了对象区域增强策略和黑洞填充策略后，OENet 的性能都比基准模型有一定的提升。虽然略低于 Adelaide[128]模型，但是优于大多数模型。值得注意的是，本文的模型仅在细粒度标签上进行训练，并没有使用 20,000 个粗粒度标签进行训练。此外，上报的数据结果均基于单模型完成。

表 3.6 各种算法在 *Cityscapes* 测试集上的性能对比

方 法	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	Bicycle	平均 IoU
FCN8s[174]	97.4	78.4	89.2	34.9	44.2	47.4	60.1	65	91.4	69.3	93.9	77.1	51.4	92.6	35.3	48.6	46.5	51.6	66.8	65.3
DPN[125]	97.5	78.5	89.5	40.4	45.9	51.1	56.8	65.3	91.5	69.4	94.5	77.5	54.2	92.5	44.5	53.4	49.9	52.1	64.8	66.8
DilatedNet[179]	97.6	79.2	89.9	37.3	47.6	53.2	58.6	65.2	91.8	69.4	93.7	78.9	55	93.3	45.5	53.4	47.7	52.2	66	67.1
DeepLab[118]	97.9	81.3	90.3	48.8	47.4	49.6	57.9	67.3	91.9	69.4	94.2	79.8	59.8	93.7	56.5	57.5	57.5	57.7	68.8	70.4
Adelaide[128]	98	82.6	90.6	44	50.7	51.1	65	71.7	92	72	94.1	81.5	61.1	94.3	61.1	65.1	53.8	61.6	70.6	71.6
Baseline	97.4	78.3	90	45.8	46.3	40.6	53.4	65.6	91.3	67.6	94.5	79.8	59.2	93.9	62.8	69.3	62.5	59.1	68.5	69.8
OENet	97.5	79.4	90.5	48.5	49	43.5	55.7	67.3	91.7	69.2	94.8	80.8	61.2	94.2	64.6	70.8	64.4	61.1	70	71.3

注：OENet 实验中未使用粗粒度标签和多尺度训练策略。

4. 定性分析

和 *Sceneparsing150* 数据集一样，本文同样可视化了 *Cityscape* 数据集的运行结果。图 3.9 给出了 Cityscape 数据集上场景解析结果示意图。每个子图分别表示：（a）原始图像，（b）标准参考图（Groundtruth），（c）基准模型完成的解析图，（d）在基准模型上使用高分辨率融合的解析图，（e）OENet 基于高分辨率融合的解析图。受益于对象增强策略，对象的完整性得以更好地保持（如：第二行的人，第三行的汽车、第四行的卡车和第六行的三轮车）。此外，可视化图中，可以发现高分辨率始终有利于改善对象增强的性能。如第一行的汽车和第五行的摩托车，高分辨率方法可以找到

一些更小的对象。

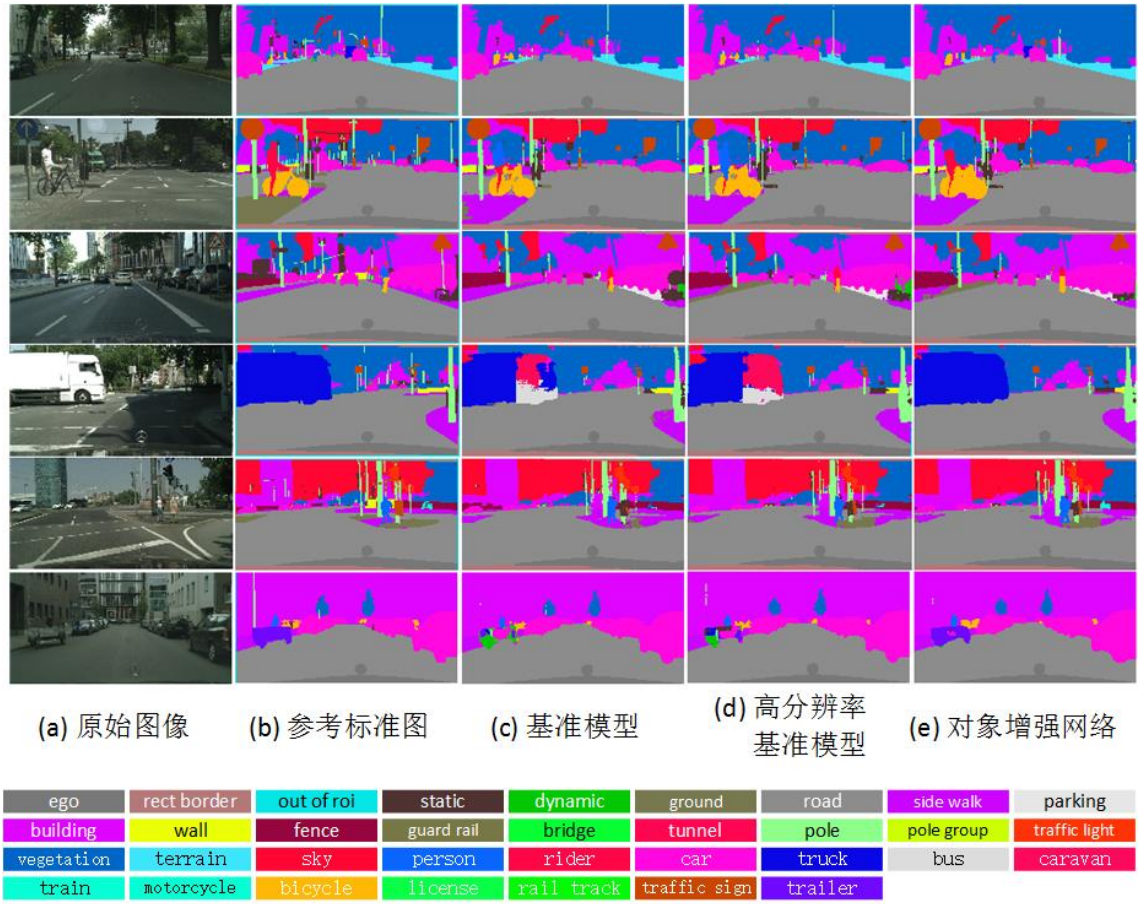
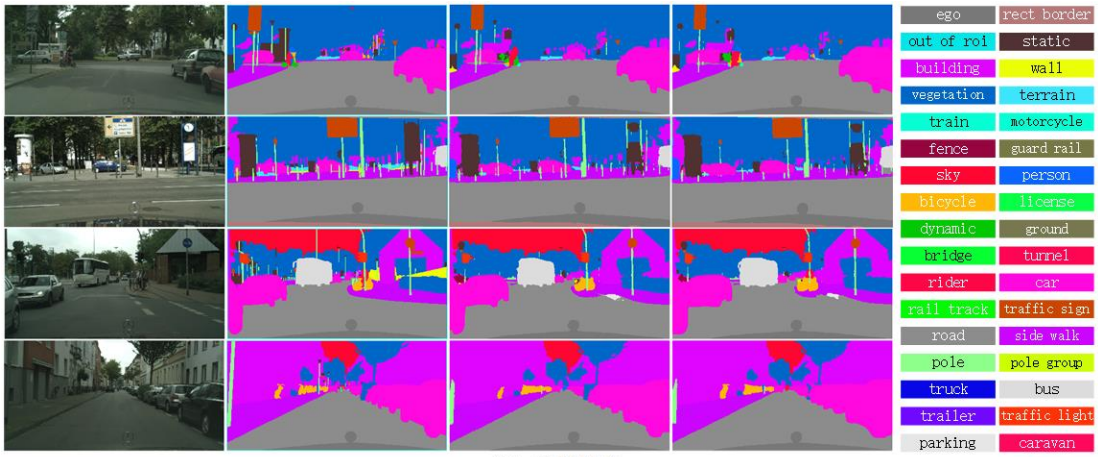


图 3.9 *Cityscape* 数据集上场景解析结果示意图

5. 失败案例分析

正如前面描述的，对象区域增强能够保持对象的完整性。然而，这种增强也会破坏一些有交叉的对象。例如，在图 3.10 的 *Cityscape* 数据集上 OENet 失败案例的示意图中，灯柱（第 1-3 行）和植物群（第 4 行）都丢失在了背景对象中。在图 3.10 中，每个子图分别表示（a）原始图像，（b）标准参考图，（c）在基准模型上使用高分辨率融合的解析图，（d）OENet 基于高分辨率融合的解析图。



(a) 原始图像　(b) 参考标准图　(c) 高分辨率基准模型　(d) 对象增强网络

图 3.10 *Cityscape* 数据集上 OENet 失败案例的示意图

3.5 小结

本文提出了一种利用上下文语义互补性的对象区域增强网络（Objectness Region Enhancement Network，OENet），利用传统的图像分类网络来实现场景分割任务。为了重新召回丢失的对象，一个基于对象建议网络的对象增强方法被用来产生建议框级实例和遮罩级实例。利用加权增强策略，一些丢失的对象被找回。建议框级实例和遮罩级实例都可以被认为是对象，不同的是，遮罩级实例可以被认为是细粒度的建议框级实例。为了获得语义上更精确和细致的解析结果，全连接 CRF 和黑洞填充也被集成到推理网络中。其中，黑洞填充策略可以有效地处理一些像素被错误分配到不存在的额外背景类的尴尬。本文的算法在 *SceneParsing150*[131]和 *Cityscape*[181]数据集上被验证是也有效的，它可以同时提高对象和背景解析的性能。值得注意的是，两个核心方法，对象区域增强和黑洞填充策略并非只能应用于 OENet，它们可以被分离出来嵌入到任何其他的场景解析模型中用于改进对象区域的解析能力。将来，将进一步考虑如何更好地进行对象区域建议，以及如何更好地融合局部区域，进而最终改进场景解析性能。

总的来说，本章主要有以下三个贡献：

1. 提出了一个统一的框架用于处理场景解析任务。受益于模块化设计方案，本文提出的算法不仅仅可以通过更换卷积或检测模块来提高整体性能，也可以将对象增强和黑洞填充应用到其他系统以提高系统对对象的解析能力。
2. 提出了一种基于对象区域增强的方法用于召回那些在标准的分割网络中无法识别的对象。
3. 提出了一种称为黑洞填充的技术，用于解决那些被错误地分类到不存在的额外背景类的像素。

4 基于上下文融合的人像妆容迁移

4.1 引言

随着互联网的飞速发展，以及照相机、摄像机、监控摄像头在日常生活中应用愈加普及，广大群众的生活方式也发生了巨大改变，比如电子商务、视频监控和社交网络的广泛普及。具体来说，电子商务网站的蓬勃发展使得国民可以足不出户，就可以挑选和购买到自己需要的日常用品。人类进入 Web 3.0 时代以后，社交网络也蓬勃地发展起来，大量用户自主创造的多媒体信息在社交网络中得到了广泛的传播，彻底改变了人类的社交方式。其次，越来越多的城市都覆盖大量的“电子眼”，高质量、实时的监控系统极大地提升了城市的安全性。在这些大数据中，人像图片占据着很大的比率和最为重要的地位。人像图片的语义理解（也可简称为人像解析），是一个对经济和社会有广泛影响的科学和工程中的重大基本问题。

化妆已成为目前大众日常生活的一部分，这给基于人脸的身份验证系统带来了巨大挑战。比如在公安系统中，无论罪犯进行了何种面容上的伪装，系统都应该能断定罪犯的身份。研究一种自动化妆系统对理解化妆后人脸身份验证具有重要意义。本章从一个有趣的任务出发，通过研究人像妆容迁移实现对人像面部信息的理解和分析。为了实现妆容迁移，研究了妆容推荐和合成这一科学问题，本文提出了一种新颖的深度局部妆容迁移网络（Deep Localized Makeup Transfer Network，DLMTN）。本章的相关研究工作即将发表在文献[138, 182]中。

4.2 问题描述

化妆使人更具有魅力，市场上也出现了越来越多的商业人脸化妆系统。Vitual Hairstyle 提供手工的发型试穿功能。Virtual Makeover TAAZ 允许给用户化上一些预定义妆容，例如：红色的唇彩和黑色的眼线。然而，所有的这些软件都依赖于预先定义的妆容，无法满足用户个性化的需求。与这些现有的工作不同的是，本文的目标是

设计一种实用的应用系统去自动推荐适合用户的妆容，并将推荐的妆容迁移到用户提供的未化妆的人脸照片上。如图 4.1 所示，本文模拟了化妆的两个功能。（1）第一个功能是妆容推荐，这更加符合个性化的需求。具体来说，具有相似的脸型、眼睛和嘴巴的形状，可能会更适合使用相似的化妆[139]。基于这个目的，给定一个妆前的脸，推荐系统可以从数据中找到视觉上最像的参考脸。相似性通过两幅人像的欧氏距离计算得到，而人像的特征由现有的深度人脸识别网络[77]生成。综上所述，推荐是一个个性化的、数据驱动的和易于实现的过程。（2）第二个功能是从参考脸上将妆容迁移到未化妆的脸上。这个妆容迁移功能，需要满足以下五个标准。1）全面的化妆功能：本文考虑了三种通用的妆容，包括：粉底、眼影和唇彩。值得注意的是，基于模型的扩展性，其他类型的妆容也可以很容易被扩展实现；2）妆容定制：不同的妆容采取不同的迁移方式。例如：眼影比较关注形状的迁移，而唇彩比较关心纹理和色彩的迁移；3）局部化：所有的妆容都对应于它特定的人脸区域。例如，唇彩只出现在嘴唇上，而眼影通常出现在眼皮附近；4）自然：妆容需要无缝地融合到未化妆的脸上。换句话说，化妆后的脸看上去很自然；5）妆容强度可控：可以根据需要调节每种妆容的浓淡。

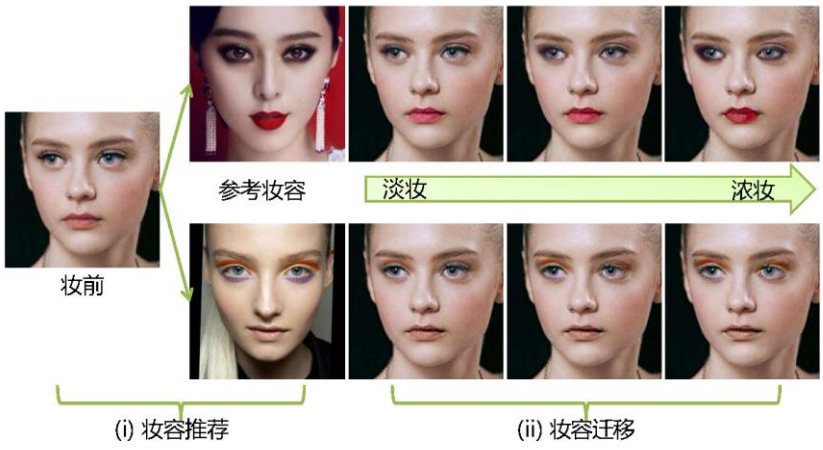


图 4.1 妆容迁移示意图

为了适应上述提到的五种标准，本文提出一个深度局部妆容迁移网络，如图 4.2 所示，网络将一个推荐的妆容迁移到了另一个未化妆的脸上。在整个迁移的过程中，

主要包含两个步骤：（1）建立人脸部件（妆前图）和妆容（参考妆容）的相关性；（2）眼影、唇彩和粉底的妆容根据人脸解析的结果迁移到指定位置，全局平滑被用来使上妆后的图更真实。首先，妆前图和参考妆容图都被送入人脸解析网络用于生成相关的标签图。解析网络基于全卷积网络[174]实现，它需要获取化妆的相关人脸部件，例如眼影，同时要考虑人脸正面器官的对称性。基于人脸解析结果，妆前图的局部区域（如：嘴）将对应到参考妆容响应的区域（如：唇）。其次，三种最通用的妆容（眼影、唇彩、粉底）以它们各自的方式完成迁移。例如，保持形状对于迁移眼影尤为重要；而皮肤纹理的平滑性对于粉底特别重要。所以眼影的迁移可以通过直接替换相关的深度特征图实现[183]，而粉底需要规范特征的内积获得[184]。妆前人像图将作为生成妆后人像图的初始化输入，然后通过随机梯度下降算法（Stochastic Gradient Descent，SGD）逐渐更新产生自然的结果。通过调节每一个妆容的权重，将生成一系列可控妆容强度的妆后效果图。

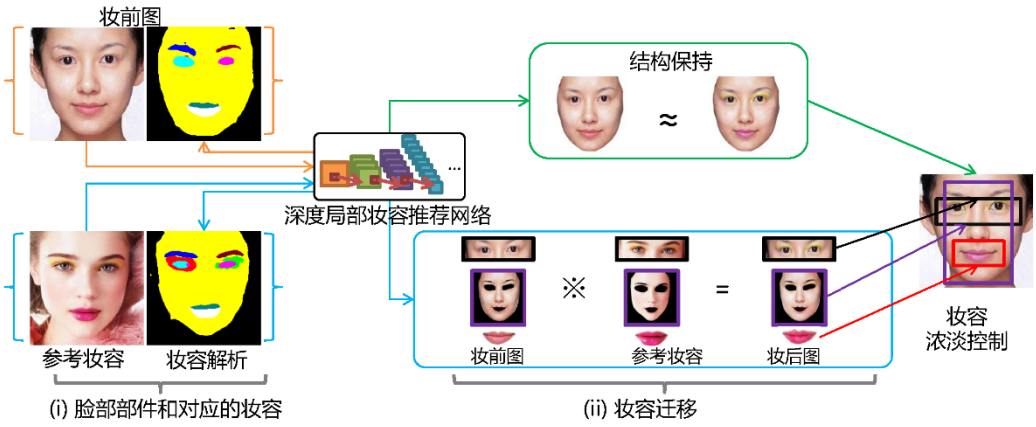


图 4.2 深度局部妆容迁移网络流程图

与传统的化妆迁移方法[136, 137, 139, 184]相比，本文的方法不需要复杂的数据预处理和数据标注，并且能够生成更好的化妆效果图。本文的贡献主要包括：（1）这是首次基于深度学习框架实现的妆容迁移系统，并且产生非常自然的结果。本文的系统可以迁移粉底、眼影和唇彩。同时所有妆容的强度是可控的。（2）本文提出了一个端到端的深度局部妆容迁移网络用于构建人脸局部区域和特定妆容的相关性，并完成妆容

迁移。不像 NeuralStyle[184]融合的是两个全局的图像，本文的方法可以实现局部妆容和相关人脸区域的对应融合。因此，可以避免大量不自然的妆容迁移。

4.3 深度局部妆容迁移网络

4.3.1 妆容推荐

妆容推荐最重要的标准是个性化[139]。看起来长得像的女性，使用相似的妆容通常效果也会比较合适与自然。给定一个妆前人脸图像，算法可以从参考妆容数据集中查找到若干与妆前人脸图像最相似的带妆的人脸图像。两个人脸图像的相似性由欧氏距离进行衡量，在计算人脸特征的时候，妆容推荐使用深度模型 VGG-Face[77]来完成，该模型基于 VGG16[3]模型微调训练获得，而深度特征则由 VGG-Face 模型中经过l_2正则化的两个 4096 维全连接层串联而成。VGG-Face 旨在识别不同的人脸，无论她是否经过化妆，这个性质正好满足本文的需求。因此，抽取到的特征能够很好地捕捉人脸的结构信息，以尽量保证搜索出来的人脸图像和妆前人脸图像在视觉上看起来很像。最后，检索结果中的妆容被作为参考妆容，迁移到未化妆的人脸上。图 4.3 显示了两个推荐妆容的例子，第一列是未化妆的人脸图像，其他列是推荐的参考妆容图像。它充分显示了推荐的参考妆容的脸型和妆前的脸从脸型到五官都非常相似，也因为这个过程，使得本文的妆容推荐具有个性化的特色。

图 4.3 两个妆容推荐的例子

4.3.2 面部器官解析

为了能够实现局部妆容迁移，需要构建未化妆人脸的部件和参考妆容带妆区域的对应关系，这种对应关系是一种一一对应的关系。大多数的对应关系都可以通过人脸解析来完成，例如：“嘴唇”和“唇彩”。唯一的例外是眼影迁移，因为未化妆的人脸没有任何一个区域可以很好地对应到参考妆容的眼影区域，而且每张人脸图像上的眼影通常是不同的，它们可以以任意的形状、任意的尺度和任意的范围出现。为了解决这个问题，本文以眼睛作为参考基准，通过一定仿射变换实现眼影和待化妆眼影区域的位置匹配。

1. 人脸解析

本文的人脸解析模型基于全卷积网络 FCN[174]进行构建。该网络利用妆前人脸图像和参考妆容人脸图像两类图共同进行训练，总有 11 个类别部件参与解析。该网络可以实现任意尺寸的样本输入，然后产生与输入图像相同尺寸的人脸解析图。在训练人脸解析模型时，本文更关注那些与化妆有关的人脸部件。例如，相比“背景”类，“左眼眼影”更加重要。因此，本文提出了一种加权的交叉熵损失函数，它描述的是最后一层的每个空间位置的加权和：

$$\ell(x;\theta)=\sum_{ij}\ell'\left(y_{ij},p\left(x_{ij};\theta\right)\right)\cdot w\left(y_{ij}\right) \tag{4.1}$$

其中，ℓ'是每个像素的交叉熵损失。y_{ij}和$p\left(x_{ij};\theta\right)$分别是每个像素的 Groundtruth 和预测值，$w\left(y_{ij}\right)$对应该像素的权重。权重的设置通过在验证集上求取最大 F1 值确定。

由于训练集中，所有的人脸都通过人脸检测和人脸对齐算法实现变换，所以所有的人脸图像最终都是正面视角。因此，在测试阶段，可以通过强制执行对称先验来替换原有的像素置信度预测。原有的坐标点p和其水平镜像点$f(p)$，可以用它们所在位置的所有通道的平均值来替代：$x_{p,c}=\frac{1}{2}\sum\left(x_{p,c}+x_{f(p),c}\right)$，其中 c 表示卷积特征图的通道数。与文献[127, 177]类似，当前对称先验只在测试阶段被使用。如何将类似的结构先验应用到训练中将作为将来的研究内容。图 4.4 展示了原始 FCN、加权 FCN

和对称加权 FCN 的人脸解析结果。

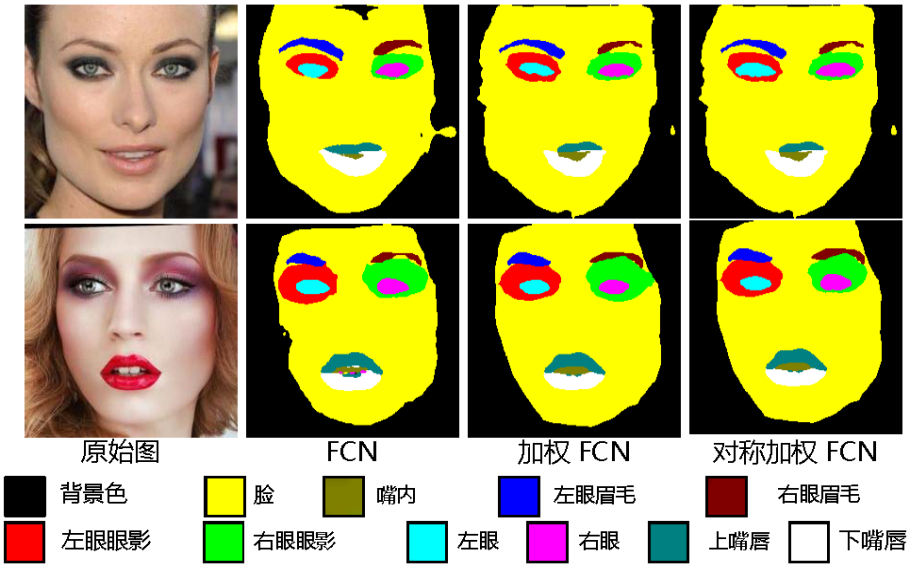


图 4.4 两组图像的人脸解析效果图

2. 眼影变形

基于人脸解析，大多数的区域相关性可以被建立，例如“脸”和“粉底”。然而，在化妆前的人脸图像上没有与眼影对应的脸部器官，因此，需要先在妆前的人脸图像上手工生成眼影遮罩。幸运的是，眼影通常与眼睛的形状和位置有一定的关联性。因此，可以根据眼睛的信息来构建眼影区域的遮罩。然而，妆前人脸图像和参考妆容图像的人脸的眼睛和眉毛的区域通常不可能严格一致，所以需要通过一定的形变操作将他们严格对齐。具体来说，本文在眼睛和眉毛区域选取了 8 个坐标点，包括，眼睛和眉毛的内部、上中、下中和转角。然后，眼影区域可以通过薄板样条插值[185]（Thin Plate Splines，TPS）完成变形。

4.3.3 妆容迁移

妆容迁移通过图像对之间的相关关系，特别是局部区域的相关关系完成。在这一小节中，将详细描述如何迁移眼影、唇彩和粉底。同时，在目标函数中本文的算法增

加了保持整个人脸结构形状的因素。

1. 眼影迁移

图 4.5 两个眼影迁移的例子

眼影的迁移需要同时考虑形状和颜色的影响。此处以左眼的眼影为例。假设s_r是由参考妆容生成的左眼眼影的二进制遮罩，s_b是妆前人脸图像经过形变后生成的待化妆区域。值得注意的是，理论上经过眼影变形，s_r和s_b应该具有相同的形状和大小。从技术上来看，眼影迁移可以认为是使用s_r去替换由确定的卷积层（本章中使用的是卷积层 *conv1-1*）生成的深度特征s_b'。左眼眼影迁移的损失函数，可以用公式$R_l(A)$来表达：

$$\begin{aligned} A^* &= \underset{A \in R^{H \times W \times C}}{\operatorname{argmin}} R_l(A) \\ &= \underset{A \in R^{H \times W \times C}}{\operatorname{argmin}} \left\| P\left(\Omega^l\left(A(s_b')\right)\right) - P(\Omega^l\left(A(s_r')\right)) \right\|_2^2 \end{aligned} \tag{4.2}$$

其中，H，W 和 C 分别是输入图像的高、宽和通道数量。Ω^l: $R^{H \times W \times C} \to R^d$是人脸解析模型中卷积层 *conv1-1* 的 D 维的特征表达。A 和 R 分别是化妆后的人脸图像和参考妆容的人脸图像。s_b'和s_r'分别是利用卷积特征遮罩方法[65, 117]从数据层映射到卷积层 *conv1_1* 的特征图。相似地，右眼眼影的损失函数可以定义为$R_r(A)$。图 4.5 是两个人像中眼影迁移的例子，从图中可以看到形状和颜色都得到了很好的迁移。

2. 唇彩和粉底迁移

粉底和唇彩的迁移主要注重的是颜色和纹理。其中粉底的迁移可以定义为：

$$A^* = \underset{A \in R^{H \times W \times C}}{\operatorname{argmin}} R_f(A)$$

$$= \underset{A \in R^{H \times W \times C}}{\operatorname{argmin}} \sum_{l=1}^{L} \left\| \Omega_{ij}^{l}\left(A(s_b')\right) - \Omega_{ij}^{l}\left(A(s_r')\right) \right\|_2^2 \tag{4.3}$$

其中，L 是卷积特征图的的层的数量。技术上，本文同时使用五个卷积层来共同实现多尺度的特征提取，包括 *conv1-1*，*conv2-1*，*conv3-1*，*conv4-1*,和 *conv5-1*。*Gram* 矩阵 $\Omega^l \in R^{N_l \times N_l}$ 由公式 4.4 定义，其中N_l是第 l 层的卷积特征图的数量，Ω_{ij}^{l}是第 l 层的向量化特征 i 和 j 的内积：

$$\Omega_{ij}^{l} = \sum_{k} \Omega_{ik}^{l} \Omega_{jk}^{l} \tag{4.4}$$

粉底迁移的结果，如图 4.6 所示。两个女孩面部的皮肤在迁移后，仍然很细腻，并且在色彩和纹理上融合了参考妆容的样式。

图 4.6 两个粉底迁移的例子

上嘴唇唇彩的损失$R_{up}(A)$和下嘴唇唇彩的损失$R_{low}(A)$的定义方式和公式 4.3 类似。唇彩迁移的实例如图 4.7 所示。经过唇彩迁移，妆前嘴唇的颜色已经变成了参考妆容嘴唇的颜色。

图 4.7 两个唇彩迁移的例子

3. 结构保持

结构保持项$R_s(A)$的定义和公式 4.2 相似，唯一的区别是s_r和s_b的每个元素都等于 1。结构保持项强制整个人脸在变化前后保持完全一样，该项损失的主要作用是不希望因为妆容迁移而导致原始脸的变形或色彩变得过分突出。换句话说，如果$R_s(A)$设置较大权重，上妆的过程将变得非常缓慢，甚至停止。而设置的权重较小，上妆将会变得很快，但是过快的权重可能会导致像素更新超过 RGB 色彩的上限值 255，导致颜色过曝。因此，为损失$R_s(A)$设置合适的权重对于妆容迁移的效果优劣非常重要。

4. 综合妆容迁移

整体的妆容迁移需要考虑眼影，唇彩，粉底，同时也需要考虑人脸的结构：

$$A^* = \underset{A\in R^{H\times W\times C}}{\operatorname{argmin}} \lambda_e\big(R_l(A)+R_r(A)\big)+\lambda_l\Big(R_{up}(A)+R_{low}(A)\Big) + \lambda_f R_f(A)+\lambda_s R_s(A)+R_{V^\beta}(A) \tag{4.5}$$

为了使结果看起来更自然，使用总方差项 $R_{V^\beta}=\sum_{i,j}\Big(\big(A_{i,j+1}-A_{i,j}\big)^2+\big(A_{i+1,j}-A_{i,j}\big)^2\Big)^{\frac{\beta}{2}}$用于强制平滑处理。其中，$R_l(A)$，$R_r(A)$，$R_f(A)$，$R_{up}(A)$，$R_{low}(A)$和$R_s(A)$分别是关于左右眼眼影，粉底，上下唇彩和人脸结构的损失。λ_e，λ_f，λ_l和λ_s分别是不同妆容的平衡权重参数。通过调节这些参数，可以调节妆容的浓淡。例如，增大λ_e将会使眼影变得更深。

公式 4.5 的能量函数可以通过带动量的随机梯度下降算法[183]（SGD）来优化：

$$\mu_{t+1} \leftarrow m\mu_t - \eta_t \nabla E(A)$$
$$A_{t+1} \leftarrow A_t - \mu_t \tag{4.6}$$

其中，μ_t是最终的梯度加权平均值，其衰减因子$m = 0.9$，A_0是初始的未化妆的人脸图像。

4.4 实验与分析

4.4.1 实验设置

1. 数据集和参数设置

迁移算法的性能评估在本文新收集的数据集上完成，其中包括 1000 幅妆前人脸图像和 1000 幅妆后人脸图像。其中，妆前人脸图像都是裸妆或淡妆。在 2000 幅人脸图像中，100 幅妆前人脸图像和 500 幅妆后人脸图像用于测试；剩下的 1300 幅人脸图像和 100 幅人脸图像分别用于训练和验证。给定一个妆前的测试人脸，最相似的人脸图像将从 500 个妆后的参考妆容人脸图像中被选择出来用于妆容迁移。权重参数 $[\lambda_s, \lambda_e, \lambda_l, \lambda_f]$ 分别设置为 $[10, 40, 500, 100]$。FCN 的不同标签的权重分别设置为 $[1.4, 1.2, 1]$，对应三组类别{眉毛，眼睛，眼影}，{嘴唇，嘴内}，{脸，背景}。这些权重在验证集上验证获得。

2. 基准模型

截止参考论文[138, 182]发表时，只有 Guo 等人[136]做过妆容迁移相关的工作。此外，本文使用 Gatys 等人[184]发表的神经艺术的两个变种方法生成了新的对比算法，它们都利用 CNN 来合成新的图像。在进行图像合成时，分别使用一张图的风格特征和另外一张图的内容特征。第一个变种称为 NerualStyle-CC，它将化妆前的人脸图像和参考妆容人脸图像都视为内容；另外一个变种称为 NeuralStyle-CS，它使用化妆前的人

脸图像作为内容，参考妆容人脸图像作为风格。其他的一些方法，如 Tong 等人[137]、Scherbaum 等人[186]或电子美妆系统[139]，都需要固定妆前-妆后人像对，3D 信息或者额外标注的人脸属性做辅助。为了公平，本文没有和这些算法进行对比。本文建议的算法在 NVIDIA Titan X GPU 上可以实现 6 秒内对 224×224 的人脸图像进行上妆。

4.4.2 妆容迁移程度分析

为了显示本文算法可以生成不同浓度的妆后图，即从浓妆到淡妆的变化，可以通过逐渐增加几个妆容权重λ_e，λ_l和λ_f来实现。四个可控浓淡妆容的例子如图 4.8 所示。前两行使用的是同一组妆前人脸图像和参考妆容人脸图像，从图中可以发现，第一行的女孩眼影逐渐变深，而第二行的唇彩变得越来越红。第三、四行是另外一组例子，第三行的眼影和第四行的唇彩同样也逐渐变深。

图 4.8 可控的妆容迁移示意图

4.4.3 整体性能对比

按照文献[139]的评估标准，本文的算法和 Guo[136]、NerualStyle-CC[184]、NeuralStyle-CS[184]算法分别进行了定性和定量的对比（此处，特别感谢两篇论文的作者提供了完整的代码，以供对比实验）。

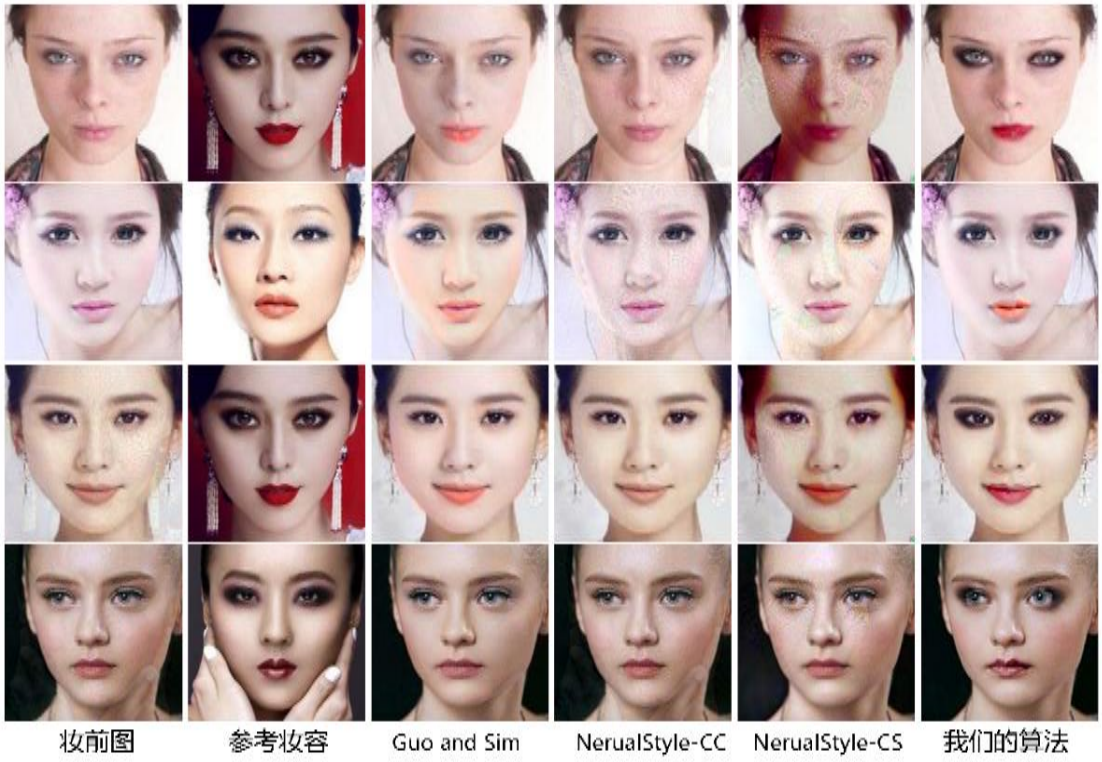


图 4.9 多种算法的定性对比示例图

定性实验的结果如图 4.9 所示，Guo 等人[136]和本文的算法产生的结果看起来都比较自然。 但是，本文算法所迁移的美妆无论是唇彩还是眼影都和参考妆容基本保持一致，然而，Guo 等人 [136]所迁移的美妆都有普遍偏亮的问题。例如第一行中，参考妆容和本文算法上妆后女孩的嘴唇都是深红色的，而 Guo 等人[136]上妆的女孩是橘红色的。此外，参考妆容人脸图像和本文上妆的女孩的眼影都是较深的烟熏妆，而 Guo 等人 [136]仅仅是产生了非常亮的眼影。与 NerualStyle-CC[184]和 NeuralStyle-CS[184]算法比起来，本文上妆的女孩人工雕琢的痕迹并不是很明显，因为本文采用的是局部区域的妆容迁移，而对比方法使用的是全局迁移。全局妆容迁移，很容易在图像对之

间产生误匹配的问题，最显著的后果是，在整个人脸上都产生一些本来只应该出现在局部的特征。例如：NeuralStyle-CS 算法第三行的女孩在额头两边出现了嘴唇的色彩；而 NeuralStyle-CC 和 NeuralStyle-CS 算法在第一行和第二行的女孩脸上的大部分区域都出现了色彩噪声，这主要是因为基于全局上下文信息的源深色区域的色彩被迁移到同样是全局上下文的目标人像图片上的浅色肤色区域，从而导致错误迁移。

表 4.1 多种算法的定量对比结果

	好很多	更好	相同	更差	差很多
Guo 等人[136]	9.7%	55.9%	22.4%	11.1%	1.0%
NeuralStyle-CC[184]	82.7%	14.0%	3.24%	0.15%	0%
NeuralStyle-CS[184]	82.8%	14.9%	2.06%	0.29%	0%

定量比较主要反映的是迁移妆容的质量和协调程度。对于所有的 100 个妆前测试人脸图像，五幅最相似的参考人脸图像被推荐出来进行妆容迁移。因此，对于每一种妆容迁移方法，都可以获得 100×5 种化妆后的人脸图像。本文顺序和三种基准方法进行了逐对对比，评价相对优劣。每一组对比图由一个 4 元组构成，包括：{妆前人脸图像，参考妆容人脸图像，妆后人脸图像（本文的方法），妆后人脸图像（对比方法}。这些结果由 20 个志愿者进行人工评价。所有的参与者都用基于五级的评价指标对每个四元组进行评价，五个等级分别是：“好很多”，“更好”，“相同”，“更差”，“差很多”。每个级别对比的百分比如表 4.1 所示。本文的算法相比 Guo 等人 [136]的算法分别有 9.7%和 55.9%的样本获得了“更好”和“好很多”的评价。对于 NeuralStyle-CC 和 NeuralStyle-CS 更是有 82.7%和 828%的样本获得了 “好很多” 的评价。

4.4.4 多重妆容迁移结果

图 4.10 不同女孩使用相同的妆容推荐进行上妆

在图 4.10 和图 4.11 中，给出两种多重妆容迁移的例子。在图 4.10 中，对于每一个参考妆容，图中都选择了 4 个最相似的妆前人脸图像进行妆容迁移。同样的妆容被同时应用到了这 4 个女孩的脸上。从效果图中可以发现，眼影、唇彩和粉底都被成功地迁移到了眼皮、嘴唇和脸的区域。值得注意的是，本文的妆容迁移方法可以处理不同脸部表情的妆容迁移。例如，图 4.10 中，左图的第二个女孩是露齿微笑的，而参考妆容的女孩并没有微笑的表情。受益于局部上下文的特性，唇彩并没有被迁移到牙齿上。

图 4.11 同一个女孩使用不同的妆容推荐进行上妆

在图 4.11 中，对于每一个妆前人脸图像，本文同样都选择了 4 个最相似的参考妆容人像。这个功能非常有利于真实的应用系统，因为用户可以尝试多种自己喜欢的参考妆容。

4.5 小结

本文提出了一种新颖的深度局部妆容迁移网络（Deep Localized Makeup Transfer Network，DLMTN）去自动地从一个带妆人脸图像上将妆容迁移到素颜图。提出的方法具有五个很好的属性，全面的妆容迁移、妆容定制、局部化、产生自然的上妆结果以及浓淡可控的妆容迁移过程。下一步，将扩展这个工作，例如，从两个或多个不同的参考妆容人脸图像迁移指定妆容到一个素颜人脸图像上。总的来说，本章有如下三个贡献：

1. 提出了一个统一的深度学习生成框架，实现从人脸推荐、人脸解析、人脸全局和局部特征提取和特征迁移等多项功能。基于这个功能，实现了一个有趣的应用——人像妆容迁移。
2. 充分利用人脸全局和局部的上下文信息，在保持人脸形状的基础上，实现了全局粉底和局部眼影、局部唇彩的特征迁移，在迁移的过程中很好地保持了人脸部件的结构、形状、色彩和纹理，并且在融合特征的过程中，尽量做到自然、真实。特征迁移的过程是一个生成学习的过程，利用迭代学习的方法，网络可以较好地处理了多种不同区域的不同特征的融合和生成。
3. 在特征迁移的过程中，通过对不同上下文信息的分离处理，在权重的控制下，可以定制不同局部和整体特征的强度，从而实现不同妆容的浓淡调节。

5 基于层次化语义哈希的图像检索

5.1 引言

在前面的三个章节中，本文提出了多种基于上下文语义信息实现不同粒度的视觉内容识别的方法，它们有效地识别出了给定样本不同级别的高层语义信息。但是，有时候用户可能不仅希望知道一个样本是什么，更希望从一个庞大的图像库中找到和给定图像相似的图像，这就需要用到基于内容的图像检索。在这一章，作者提出了一种快速、准确的，并且可以应用到大规模图像集的图像检索算法——层次化深度语义哈希（Hierarchical Deep Semantic Hashing, HDSH）。本章的相关研究工作发表文献[155, 187, 188]中。

5.2 问题描述

在计算机视觉领域，相似性图像检索的主要目标是：给定一个查询图像，从一个庞大的图片集中找到视觉上与之最相似的图像。图 5.1 是两个典型的图像检索的例子，最左边的图像是待检索的图像，右边上下分别列出了两个不同方法（基于层次化深度语义哈希的 HDSH 和未使用该策略的普通 CNN 方法）得到的检索结果。第一幅查询图像从 Holidays 验证集中选择，第二幅图像从 *Imagenet* 的验证集中选择。相应地，特征提取模型分别在 *Holidays* 和 *Imagenet* 数据集的训练集上完成训练。从查询结果可以看出，利用层次化语义关系可以更好地保持语义信息，使查询结果与查询图像在视觉上更加相似，从而提高检索性能。从结果来看，层次化关系可以显著提高系统的判别能力。比较有意思的是第二组基于 *Imagenet* 数据集的检索范例，在未使用高层语义过滤的时候，所有返回的结果都错误地被识别成了建筑，而经过高层语义的修正，即使仍然有错误，但是在大类上已经没有明显的偏差了。

图 5.1 使用 HDSH 前后的图像检索结果对比图

在基于内容的图像检索任务中，特征表达能力和计算开销是两个关键的指标。图像的特征表达是图像检索的引擎，它已经推动了计算机视觉的发展很多年。在过去的十几年中，特征革命主要基于 SIFT[11, 12]，HOG[13]，LBP[14]，GIST[189]和 Bag-of-Features（BoF）。然而，将这些特征表达映射到二进制编码用于图像检索，对于具有复杂语义结构的样本来说是不充分的，因为这些基于手工选择的特征很难从低级特征中捕捉到图像的高层语义信息。事实上，如何选择合适的特征来表征对象，并用这些特征（或者特征的组合）去实现检索本身就是一件非常困难的事情。因此，找到一种更有效的特征来描述图像变得极其重要。

此外，由于互联网上图像的爆炸性增长，快速的相似性检索方法对于大规模的图像检索任务越来越重要。很多研究都瞄准如何从大规模的数据集中高效地检索相关的图像。但由于较高的计算开销，传统的线性搜索的方法正面临着性能不足的窘境。基于哈希的方法[190-192]和近似最近邻搜索（Approximate Nearest Neighbor, ANN）方法被提出来替代线性搜索的方法以加速检索。这些方法通过将高维特征映射到低维特征空间，然后生成二进制哈希编码。受益于紧凑的二进制编码，通过计算汉明距离或二进制模式匹配的算法实现的快速图像检索方法，显著地降低计算开销，并进一步提高了搜索的效率。然而，这些方法都存在一个弊端，对于每一个查询图像，特征相似

性的计算都必须遍历到每一个待查询库中的样本。

受近年来端到端学习方法的启发，可以自然想到如何直接利用一种卷积神经网络实现紧凑二进制编码的生成。本章提出的层次化深度语义哈希很好地回答了这个问题，它不仅提高了检索性能，提高了检索效率，同时还可以被应用到大规模的图像检索任务中。这种方法可以利用 CNN 同时学习高层语义信息和二进制表达。它展现了如何有效地将图像中的高层语义信息集成到层次化的结构中，使检索结果更加符合人的先验知识和期望。例如，给定一个包含对象“猴子”的图像，从一个预先定义好的层次化模型可以得知，一个包含“狐猴”的图像会比包含“马”的图像更加接近原始查询图像。基于这种层次化的语义信息，可以实现在计算图像特征相似性之前就过滤掉那些语义不相关的类别。

在下面的小节中，将详细介绍这种层次化的深度语义哈希算法，描述算法是如何定义图像间的相似性，以及如何利用这种相似性来实现性能和效率的同时提升。

5.3 基于层次化语义的相似性算法

5.3.1 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，应该如何定义“最相似”这个概念？为了回答这个问题并利用层次化的知识来实现图像检索，下面先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。一般来说，最常见的哈希函数 $h(\cdot)$: $\mathbb{R}^D \rightarrow \{0,1\}$ 被认为是一个特征映射的过程，它将一个 D 维度的输入样本映射成一个紧凑的哈希编码。假设 $I = \{(x_n, c_n)\}_{n=1}^N$ 是一个包含标签的图像集合，其中每一个图像 $x_n \in \mathbb{R}^D$ 都对应一个语义标签 $c_n \in L = \{1, \dots, C\}$。算法的目标是学习一个哈希函数 $h(\cdot)$，并将图像x_n映射成二进制哈希编码 $h(x_n)$，同时保持语义c_n。利用新学习到的哈希函数，可以找到一组新的二元组 $H = \{(h(x_n), c_n)\}_{n=1}^N$ 用于替代原来的二元组 I，新的二元组 H 在表征图像和其对应标签的时候不会产生任何的语义损失。在过去的工作中，学习一个相似性函数，并使用该

函数将图像的低级特征映射成一个相似性值是检索的主要目标。给定一组图像 a 和 b，可以用符号 $f_l(a)$ 和 $f_l(b)$ 来表示它们的低级特征。由此，可以得到它们的相似性表达式 $Sim(a,b) = Sim(f_l(a), f_l(b))$。

图 5.2 给出了两个图像进行相似性评估的示意图。从逻辑上看，本文通过两个部分来学习图像表达，一方面通过一个概率估计来匹配语义标签，另一方面通过哈希函数来映射图像特征。最后的相似性通过一个层次化的比较函数计算获得。

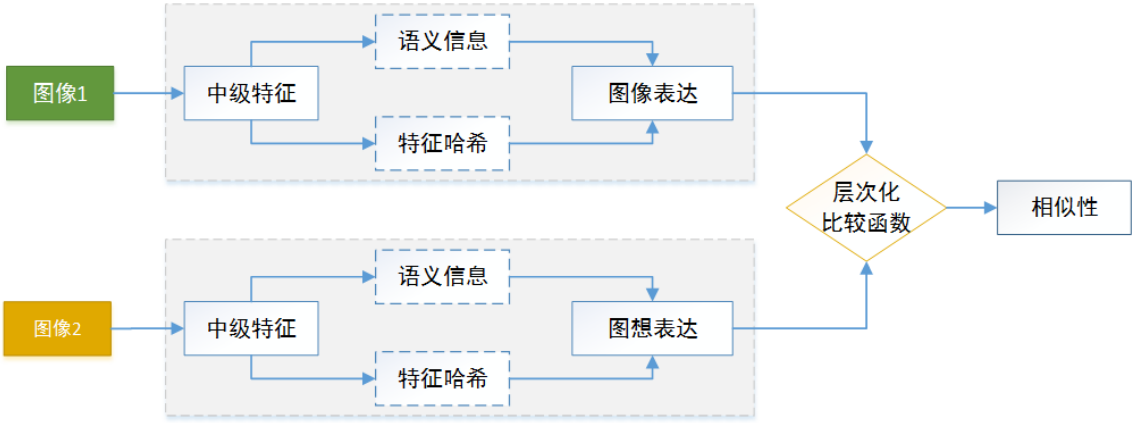


图 5.2 两个图像相似性评估方法示意图

检索系统首先通过卷积神经网络的前向传播获得原始图像 a 和 b 的中级特征 $f(a)$和$f(b)$，而不是传统方法所使用低级特征$f_l(a)$ 和$f_l(b)$。本文使用这些中级特征去估计关于标签的语义概率 $p(a) = (L(a)|f(a))$ 和 $p(b) = (L(b)|f(b))$，同时计算哈希编码 $h(a) = h(f(a))$ 和 $h(b) = h(f(b))$。接下来联合使用语义概率$p(a)$，$p(b)$和哈希编码$h(a)$，$h(b)$去计算层次化的相似性。此时，可以得到相似性计算函数：$Sim(a,b) = (p(a), h(a))\boldsymbol{S}(p(b), h(b))$，这里 $\boldsymbol{S}$ 是层次化相关矩阵，它可以通过语义级相似性和哈希级相似性计算获得。这种相似性函数计算方法不仅仅允许利用层次化信息来提高精度，也可以提高在整个数据集上进行相似性计算的速度。

值得注意的是，本文的算法并不需要像传统方法一样去精确和细致地设计哈希函数，来获取样本的哈希值。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，可以直接得到样本的哈希值。下一节将介绍这个过程。

1. 基于标签的语义级相似性

本文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。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$\{1,\dots,K\}$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。本文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 $L=\{1,\dots,C\}$，两个图像 a 和 b 之间的相似性可以使用它们所属的语义标签之间的匹配程度来衡量。假设使用符号 $\delta_i(a)\in\{0,1\}$ 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 $L_i^C(a)=\{\delta_i(a)|i=(1,\dots,C)\}$，其中有且仅有一个$\delta_i(a)=1$。基于图像 a 和 b 的语义标签可以定义它们的相似性 $Sim(a,b)=\sum_{(i,j)}\delta_i(a)\boldsymbol{S}_{ij}\delta_i(b)$，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，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，图像 a 和 b 之间的相似性可以重新定义为 $Sim(a,b)=1\{L_i^C(a)==L_i^C(b)\}\in\{0,1\}$，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，本文并没有直接使用语义来进行图像检索，而是将语义信息作为一个过滤器来改进检索的性能和速度。

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，通常可以认为“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，本文提出了一种使用基于概率版本的语义级相似性来替代简单的基于标签的语义级相似性。给定语义标签集 $L=\{1,\dots,C\}$，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 $\delta_i(a)\in\{0,1\}$ 可以标识图像 a 是否包含语义 i，类似地，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。在本文的工作中，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[193]类似，本文采用“查询—图像”的信息相关性来定义图像间的概率版相似性。

对于每一个 $i\in L$, 令 I_i^+ 表示与语义 i 相关的图像子集，I_i^- 表示不相关的子集。“语义—图像”之间的相关矩阵可以定义为 $\boldsymbol{R}_{IL}: I\times L\to\mathbb{R}^+$，对于所有的 $i\in L$，$I_i^+\in I_i$，$I_i^-\in I_i$，都满足 $\boldsymbol{R}_{IL}(i,I_i^+>0)$ 和 $\boldsymbol{R}_{IL}(i,I_i^-=0)$。为了计算“图像—图像”的相关矩阵$\boldsymbol{R}_{II}: I\times I\to\mathbb{R}^+$，假设图像 a 和 b 关于语义 i 是条件独立的，即 $P(I(a),I(b)|i)=P(I(a)|i)P(I(b)|i)$。此时，可以通过计算“图像—图像”的联合概率来衡量图像 a 和 b 的相关性：

$$\begin{aligned}P\big(I(a),I(b)\big)&=\sum_{i\in L}P(I(a),I(b)|i)P(i)\\&=\sum_{i\in L}P(I(a)|i)P(I(b)|i)P(i)\end{aligned}\tag{5.1}$$

为了改进稳定性，可以定义两个图像相关只发生在它们之间的联合概率超过截断阈值 *t* 时，则：

$$R_{II}\big(I(a),I(b)\big)=\big[P\big(I(a),I(b)\big)\big]_t \tag{5.2}$$

也就是说：

$$x=\begin{cases}[x]_t, & x>t\\ 0, & otherwise\end{cases} \tag{5.3}$$

其中，$x=P\big(I(a),I(b)\big)$。

基于上述讨论，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。在 5.4.4 节的实验部分详细分析了截断阈值 *t* 对检索性能的影响。

3. 哈希级相似性

本节将详细讨论哈希级相似性。给定图像 *I*，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，通过一个简单哈希函数 $h(\cdot)$ 可以将这个 *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 $i=1,\dots,q$，都可以通过如下公式输出它的二进制哈希编码：

$$H=h(x)=\begin{cases}1, & f(x_i)-Avg_i^q\big(f(x_i)\big)>0\\ 0, & f(x_i)-Avg_i^q\big(f(x_i)\big)<0\end{cases} \tag{5.4}$$

其中，$x=g(I)$ 是卷积层输出的 CNN 特征，$x_i(i=1,\dots L)$，$L=\{1,\dots,C\}$。$f(x)$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 $I=\{I_1, I_2, \ldots, I_n\}$ 是由 n 个图像构成的图像集合，$H=\{H_1, H_2, \ldots, H_n\}$，$H_i \in \{0,1\}^q$ 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，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：

$$d(q,i)=Dist\left(I_q, I_i\right)=||H_q-H_i|| \tag{5.5}$$

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，在定义了图像 a 和 b 之间的语义级相似性和哈希级相似性后，下一步，可以将它们组合起来形成本文提出的层次化相似性：

$$\begin{aligned} Sim(a,b) &= \sum_{i}^{C}(p(a), H_a)\, S(p(b), H_b) \\ &= \sum_{i}^{C}[P\left(I(a), I(b)\right)]_t \times (1-d(a,b)) \end{aligned} \tag{5.6}$$

其中，$\boldsymbol{R}_{II}\left(I(a), I(b)\right)=\left[P\left(I(a), I(b)\right)\right]_t$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 5.6 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，通过遍历整个数据集，可以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。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即可获得最终的查询结果。在本文的实验中，与一个查询图像语

义相关的语义大约有 1~10 个。因此，即使是在大规模的图像数据集中，相似性 $Sim(a,b)$ 的计算也将会非常的高效。

5.3.2 学习相似性

在真实世界中，检索系统可能会面临海量的检索数据，同时涉及数千的语义类，因此稳定性和效率是检索系统最需要考虑的问题。图 5.3 给出了基于层次化深度语义哈希的图像检索体系结构图。

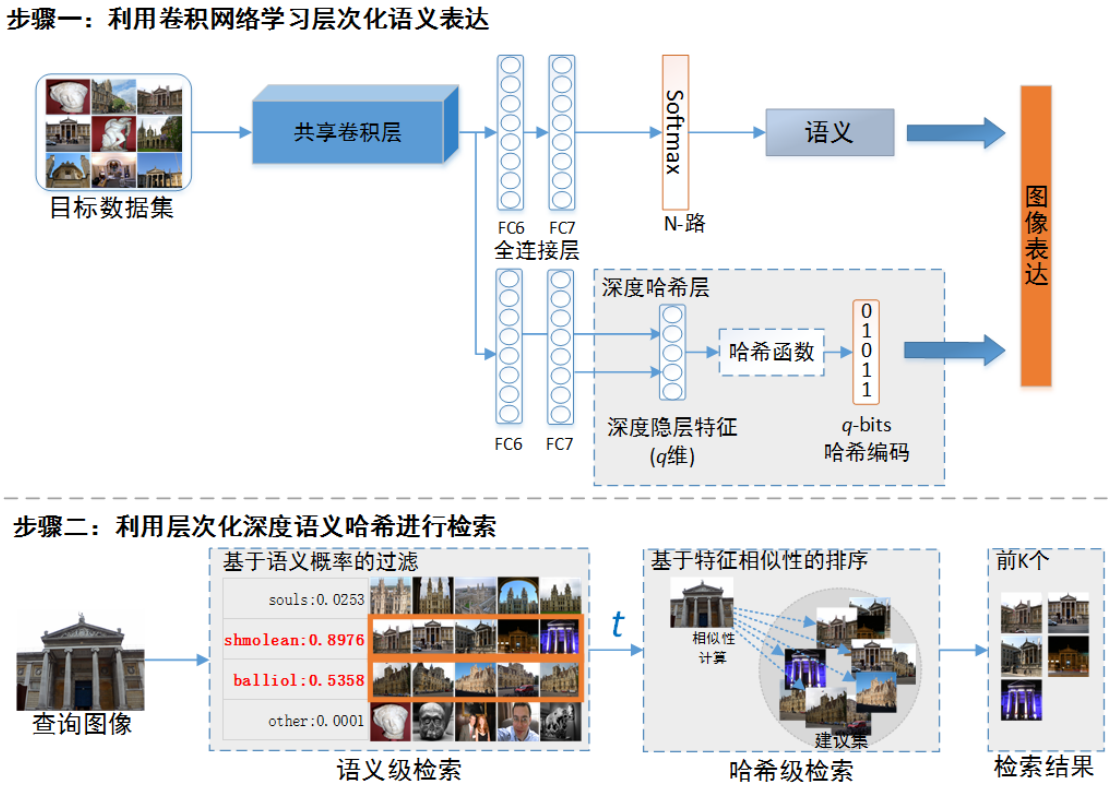


图 5.3 基于层次化深度语义哈希的图像检索体系结构图

本文提出的方法主要包含两个步骤。第一步，如图 5.3（上）所示，基于像素的原始图像被直接送入一个卷积神经网络，用于生成图像强健的特征表达。该网络首先需要在 *Imagenet*[73]数据集上进行预训练，然后再在检索的目标数据集上进行微调训练。预训练 CNN 模型由 Krizhevshy 等人[23]提出，它包含 5 个卷积层和 2 个全连接层

以及一个 1000 路的独热编码 *Softmax* 分类器。通过集成的 CNN 模型来构建哈希函数也在最近的一些文献[194]中被提到。与文献[194]不同的是，本文直接利用最后一个全连接层来计算语义标签概率，而不是用哈希层。因为从紧凑的哈希编码层抽取语义特征会增加语义信息的损失。为了实现这个目的，可以将最后一个卷积层后的全连接层分成了两个完全独立的全连接支路。一条支路为一个 n 路的 *Softmax* 分类器（n 的大小与目标数据集的类别数一致）提供输入特征，用于生成语义信息；另一条支路，通过全连接层构建一个深度哈希层，该层的功能类似传统的哈希函数，用来将 CNN 特征转换为指定维度的哈希编码。第二步，如图 5.3（下）所示，图像检索基于层次化的语义哈希，并使用粗到细的策略完成检索，它主要分为两部分：语义级检索和哈希级检索。

为了获得基于概率的语义表达，所有输入到 CNN 的图像都被调整大小到227×227 的分辨率，并减去均值图像，然后通过若干层的卷积运算和池化运算获得强健的中级特征。最后一个全连接层抽取的中级特征被送入一个学习好的 *Softmax* 分类器用于生成语义信息。值得注意的是，本文调节 *Softmax* 分类器的输出为关于每一个语义类别的概率，而不是使用 one-hot 生成的唯一类别标签。在实验中，*Softmax* 关于每个类别的分数值被作为输入图像关于每个语义的概率。

为了获得哈希级的特征表达，如图 5.3（上）所示，全连接层 *FC6* 和 *FC7* 被串联起来同时输入到一个 q 维的深度哈希层中，用于生成 q-bit 的哈希编码。相比使用 *FC6* 和 *FC7* 的联合特征，仅使用最后一个全连接层来生成哈希编码，不利于获得较鲁棒的特征，因为最高层的特征对于语义类别具有较强的不变性，它不利于捕捉一些细微的语义特性。

5.3.3 检索过程

获得了关于图像 I 的特征表达之后，就可以利用层次化的深度语义哈希方法来实现图像检索。现有的相似性学习算法通常是利用图像的底层特征来生成图像的相似性，例如：颜色特征、纹理特征、边缘特征、SIFT 特征和 HOG 特征等；而本文的相似性是组合了图像的高层语义信息和由中层特征生成的哈希级特征。因此，本文设

计了一个层次化的策略来完成检索。如图 5.3（下）所示，本文的算法首先计算查询图像 q 和目标图像集中每一个图像 b 的语义相关性$\boldsymbol{R}_{II}(I(q),I(b))$，如果语义相关性$\boldsymbol{R}_{II}=0$，则丢弃图像 b，反之，如果语义相关性$\boldsymbol{R}_{II}\neq 0$，则将图像 b 加入到候选图像集中。在语义相关性检查之后，将获得一个包含 m 幅图像的候选建议集 $P=\{I_1,I_2,\dots,I_m\},m\ll n\in R$。接下来的哈希级相似性可以在候选建议集 P 上计算获得。

5.3.4 效率分析

效率是大规模检索系统最主要的挑战。因此，对于大规模的图像数据集，例如 Imagenet 数据集来说，通过直接计算查询图像和数据库中的每一幅图像的距离来衡量相似性是不现实的。

给定一个查询图像I_q和包含语义标签集$L=\{1,\dots,C\}$的图像数据集$I=\{I_1,I_2,\dots,I_n\}$。假设计算查询图像I_q和一个待查询图像I_i之间的相似性的时间复杂度为$\mathcal{O}(1)$，那么计算整个数据集的时间复杂度将为$\mathcal{O}(n)$。本文提出的层次化方法可以大大降低整个查询过程的时间复杂度。与一个查询图像具有相关性的语义类别数量 C^+ 通常少于 10（大多数时候甚至少于 5）。事实上，计算语义开销几乎是零开销的，因此，系统的复杂性主要依赖于候选建议集 P 的大小 m。如果每个类别的数据分布是均匀的，那么时间复杂性将可以用$\mathcal{O}(m)$来表示，其中$m\ll$ n, $\frac{n}{C}=\frac{m}{C^+}$，总的加速比 $Rate_{up}=\frac{\mathcal{O}(n)}{\mathcal{O}(m)}=\frac{C}{C^+}$。例如，在 *Holidays* 数据集上的加速比大约为$\frac{500}{5\sim10}=50\sim100$倍。这是一个惊人的结果。

5.4 实验与分析

这一节中，本文基于CNN实现的检索方法将与现有的最好的检索方法进行对比，包括传统的基于 SIFT[19, 20, 195-198]特征的方法和一些最新的基于 CNN[151, 152, 154, 194, 199, 200]的方法。为了公平，所有的对比实验都只比较在进行检索时，特征提取和相似度计算的性能和效率。

5.4.1 数据集和评价指标

本文使用当前最著名的深度学习开源工具包 CAFFE[169]在多个知名的实例级图像数据集上进行算法评估。其中，*Holidays*[201]和 *Oxford5k*[202]数据集用于和现有的相似性学习算法进行对比，*Oxford105k*[202]和 *Imagenet*[73]数据集用于验证算法在大规模数据集上的有效性，*Caltech256*[203]和 *Imagenet*[73]数据集用于评估跨类泛化性能。

- *Holidays*[201]数据集包含 1491 幅图像，总共涉及 500 个不同的场景或对象。其中 500 幅图像作为验证集，用来调整模型参数和执行检索，其余的 991 幅图像作为训练集用来训练 CNN 模型。
- *Oxford5k*[202]数据集包含从 Flickr 收集的 5062 幅图像，每一幅图像都标识一个牛津大学（Oxford）地标建筑。所有的图像都由人工进行标注，分配到 11 个不同的 Groundtruth 地标建筑类中（部分图像可能会包含复杂的结构或多个建筑，但主体建筑都和地标建筑相关），同时每个地标都选出 5 幅图像作为查询图像。在所有的图像中，有 512 幅图像被标注为“优秀”或“好”，这意味着这些图像能够明确地标识出 11 个指定的地标建筑。为了获得较好的模型，这仅将这 512 副图像用于训练 CNN 模型；55 幅用于检索的图像作为验证样本。
- *Oxford105k*[202]数据集由 *Oxford5k* 数据集和额外的 100k 的负样本组成，这些负样本都与 11 个地标建筑不相关。*Oxford105k* 数据集用来评估检索系统在大规模数据集的环境下的检索性能。
- *Caltech256*[203]数据集包含 30,607 幅图像和 256 个对象类。在本文的工作中，为了验证检索系统的泛化性能，随机选择了 200 个类的图像作为训练集，其他的 56 个类作为验证集。在验证集中，30%的图像用于超参数选择，70%的图像用于测试和检索。
- *Imagnet*[73]数据集（ILSVRC 2012 竞赛版）包含 1,200,000 训练样本和 50,000 验证样本，大体上被均分为 1,000 个类别，每个类别 1,300 幅图像。为了调节参数，将原始的验证集拆分为两个部分，其中 10,000 幅图像用于超参数调节，40,000 幅图像用于测试和性能评估。

对于 *Holidays*，*Oxford5k* 和 *Oxford105k* 数据集，一律遵循数据集的标准测试协议，使用平均准确度（mAP）进行性能评估。对于 *Caltech256* 和 *Imagenet* 数据集，按照文献[204]的评价标准来进行评估。

5.4.2 数据扩展和预训练方法

由于 *Holidays* 和 *Oxford5k*/*Oxford105k* 数据集的训练样本极为缺乏，因此直接用来训练CNN模型是不太现实的。因此，本文考虑了两种有效的方法来解决这个问题，从而避免过拟合问题的产生。首先，迁移学习的方法被作为默认配置，该方法在大规模的数据集 *Imagenet* 上进行预训练，然后在目标检索数据集上进行微调。其次，本文提出了一种新颖的数据扩展方法来降低过拟合的影响。默认情况下，*Holidays* 和 *Oxford5k*/*Oxford105k* 数据分别包含 991 和 512 幅原始的训练图像，每个类的数据不平衡问题非常严重。如表 5.2 所示：*Oxford5k* 数据集的“pitt_rivers”类只有 1 个训图像，但是“radcliffe_camera”类有 216 个训练图像。为了解决这个问题，水平翻转、二维旋转和亮度变换三种策略被用于将每个类的样本都统一扩展到大约 1000 幅图像左右。首先为每个类别都定义一个独立的扩展率$n = floor(1000/N)$，其中 N 是类别 c 中所有原始图像的数量，函数$floor(\cdot)$执行向上取整操作。然后，针对每幅图像都执行 n/3 倍随机旋转，同时将旋转后的图像做水平翻转，旋转角度从区间 [-25, 25] 中随机选择。类似的，针对每幅图像，使用函数$f(\cdot)$执行 n/6 倍的随机亮度变换和水平翻转。假设 L 是原始图像的亮度，其取值范围为$(low_{in}，high_{in})$，那么新的亮度可以用L'表示，其中$L'(low_{out}，high_{out}) = f\left(L(low_{in}，high_{in})\right)$，函数$f(\cdot)$是一个映射函数，它将亮度值从原始亮度空间$(low_{in}，high_{in})$等比例地映射到新的亮度空间$(low_{out}，high_{out})$。亮度 L 的最大取值范围定义为 0 到 1，然后low_{out}从 0 到 0.2 之间进行随机选择，$high_{out}$从 0.8 到 1 之间随机选择。使用这种数据扩展方法，不但增加了训练样本，同时也平衡了每个类样本数量。此外，类似文献[23]，在训练过程中也执行了图像切割和水平镜像操作。

5.4.3 模型训练

受 Faster-RCNN[101]训练方法的启发，本文采用了一种 2 步的训练算法来学习共享卷积特征。第一阶段，在基于 *Imagenet* 数据集的预训练 CNN 模型上，针对目标数据集进行端到端的训练。第二阶段，固定所有全连接层作为语义分支，并创建一个新的全连接分支作为哈希分支，用来生成哈希编码。哈希分支包含两个全连接层，一个深度哈希层，一个哈希函数层和一个新的 *Softmax* 分类器。其中，哈希函数层，由一个 *Sigmoid* 激活函数和一个减均值函数构成，细节信息如第 5.3.2 节所示。这个新的网络采用端到端的方式进行训，直到收敛。值得关注的是，哈希分支的 *Softmax* 分类精度并不需要特别地关注。如图 5.3 所示，在推理阶段，将两个分支组合为一个统一的输出框架，用于同时获取语义级特征和哈希级特征。

5.4.4 层次化性能分析

1. 特征层分析

在这一小节中，首先研究基于层次化深度语义哈希方法（Hierarchical Deep Semantic, Hashing，HDSH）不同层的特征对性能的影响。为了简便，符号 *Conv5*、*FC6*、*FC7* 分别用来表示最后一个卷积层、第一个全连接层和第二个全连接层，*Softmax* 分类器的输出概率用符号 *FC8* 表示。这些层的特征被分别抽取作为深度 CNN 特征用于表达图像。此外，使用 *FC6* 和 *FC7* 特征去生成三种合成特征，计算它们的平均值和最大值作为平均特征和最大特征，分别用 *Mean* 和 *Max* 表示；同时，使用 *Cas* 表示串联 *FC6* 和 *FC7* 后形成的新特征。另一方面，抽取深度哈希层的特征作为紧凑的哈希特征（使用H_q表示，其中 q 是哈希编码的长度，它等于深度哈希层的神经元的维度），哈希特征可以有效地节省存储空间并加速检索。总的来说，可以得到了 7 种类型的全尺寸特征（HDSH-*Conv5*（*FC6*, *FC7*, *FC8*, *Mean*, *Max*, *Cas*））和 6 种类型的紧凑编码特征（HDSH-*H16*（*H32*, *H64*, *H128*, *H256*, *H512*））。图 5.4（左）比较了基于不同特征层的层次化方法（HDSH-0.2）和非层次化方法（HDSH-0）在 *Holidays* 和 *Oxford5k* 数据集上的性能。其中，层次化方法使用截断阈值 0.2 作为默认

参数，实线和虚线分别表示层次化方法和非层次化方法。所有方法的 mAP 曲线都具有相似的趋势，随着层次的加深，性能也逐渐上升到一个峰值，这是因为更深层的特征能够完成更好的不变性。然而，当使用最后一层的特征时性能反而急剧下降，这主要是因为 *FC8* 层的特征过分依赖于类别空间，使其丢失了很多中低层的语义信息。从图中可以看到，在 *Holidays* 和 *Oxford5k* 数据集上最好的性能都来源于 HDSH-*Mean* 特征，它由 *FC6* 和 *FC7* 的对应维度进行均值运算获得。不过 *FC6*，*FC7*，*Max* 和 *Cas* 特征都具有一定的竞争力，并不算太差，这体现了基于 CNN 的深度神经网络具有强大的特征表达能力。在 *Holidays* 数据集中，层次化的 *Mean* 特征性能显著优于 *Conv5* 和最后一层的特征（0.885 vs. 0.823 和 0.857）。*Oxford5k* 数据集也有类似的趋势。随后，在非层次化特征上的实验也得到了类似的结论。这证明了，对于不同的方法和不同的数据来说，深度 CNN 特征的表达能力和鲁棒程度是基本一致。值得高兴的是，层次化特征在 *Holidays* 和 *Oxford5k* 数据集上总是优于非层次化的特征。在 *Holidays* 数据集上，对于所有的特征层，层次化带来 6~10%的性能提升；在 *Oxford5k* 数据集上更是达到了 30%左右的性能提升。这主要是因为 *Oxford5k* 数据集中每个类具有更多的样本图像，这对于 HDSH 方法更加有利，因为它可以过滤更多不相关样本，从而提高检索系统的性能。图 5.4（左）的实验证明了，层次化的方法对于不同层次的深度特征和不同的数据集都是有效的。

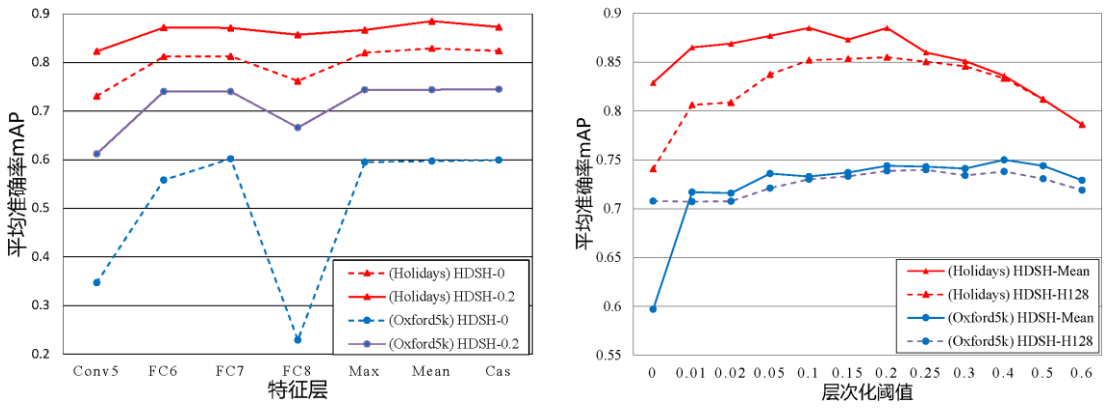


图 5.4 *Holidays* 和 *Oxford5k* 数据集上不同特征的性能比较

2. 参数分析

如第 5.3.1 节所描述，截断阈值 *t* 的主要功能是过滤那些具有较低语义相似性的类别，从而改进检索精度。由于阈值 *t* 决定了被过滤类别的数量。因此，在基于概率的语义级相似性中，需要确定一个合适的阈值 *t* 来过滤那些和查询图像 Q 语义不相似的类别。

为了选择合适的阈值 *t*，本文定义了一组固定的阈值来执行交叉验证，阈值空间可以表示为 $T = [0, 0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6]$，$t \in T$。由于 *Holidays* 和 *Oxford5k*/*Oxford105k* 数据集较小，因此阈值搜索和性能评估都在验证集上完成，没有设定专门的测试集。此外，*Imagenet* 和 *Catech256* 数据集的阈值搜索在验证集上完成，而性能评估在测试集上进行。在阈值搜索空间中，$t = 0$ 意味着查询图像 Q 和图像 I 之间的相似性仅仅通过计算特征之间的欧氏距离获得，也就是说这是一种非层次化的方法，因为$t = 0$时没有任何语义不相关的类别被预先过滤。较高的阈值可以过滤更多的语义不相似的图像，从而增加检索的精度并提高检索的速度。然而，较高的阈值也有可能会排除掉一些潜在的语义相似的类，特别是一些复杂的图像。极端情况下，只有与查询图像具有相同语义标签的图像会被保留。但是，由于基于 CNN 的分类器并不能保证语义标签的分类是 100%的准确，如果单纯的基于语义标签的异同来进行过滤，结果是不可信的。这也是基于标签的语义级相似性的缺陷所在。因此，本文希望经过语义级相似性的过滤之后，仍然有一些近似语义的类别被保留下来，进行后续的特征匹配。

图 5.4（右）展示了 *Holidays* 和 *Oxford5k* 数据集在不同阈值 *t* 下的 mAP 曲线，实线和虚线分别表示 *Mean* 特征和 128 比特的深度哈希编码。与前面的实验类似，全尺寸的特征 HDSH-*Mean* 和紧凑特征 HDSH-*H128* 同时用来评估阈值。其中，紧凑特征使用 128 比特的哈希码来标识一幅图像。从图 5.4（右）中可以发现，随着阈值的增大，mAP 也跟着增加直到一个峰值然后逐渐下降。阈值 $t = 0.2$ 是一个理想的取值，因为它给出了较高的 mAP 值，同时过滤图像的数量也较为合适。事实上，对于不同的数据集阈值 *t* 可能会不相同，但是在一定的范围内它是趋于稳定和可接受的。

例如：*Holidays* 数据集：0.01~0.2，*Oxford5k* 数据集：0.05~0.6。为了方便起见，在后续的实验中，统一设置阈值 $t = 0.2$。值得注意的是，层次化方法的性能始终要优于非层次方法。在 *Holidays* 数据集上，层次化方法 HDSH-*Mean*-0.2 带来了 6.8%的精度提升，从 0.829 上升到 0.885。对于紧凑编码，HDSH-*H128*-0.2 相对于非层次化方法 HDSH-*H128*-0 有 15.4%的性能提升，从 0.74 上升到 0.855。在 *Oxford5k* 数据集上也获得了惊人的结果，HDSH-*Mean*-0.2 和 HDSH-*H128*-0.2 分别获得了 24.6%和 4.4%的性能提升。

除了可以提高检索精度以外，设置阈值 *t* 的另外一个重要原因是，它大大缩小了搜索空间。在传统方法中，相似性的计算是在整个数据集上完成，也就是说所有的类别和所有的图像都会被扫描一遍。通过设置截断阈值 *t*，相似性搜索只需要在很少的类别中进行，通常少于 10 个类，甚至只有一个类别。这个策略极大的提高了检索的速度。表 5.4 显示了检索速度方面的实验结果。

5.4.5 整体性能对比

鉴于本文的方法是基于卷积神经网络所生成的特征来实现哈希级相似性和语义级相似性的融合相似性来衡量图像间的相互关系，本节重点对比了一些基于 CNN 特征和经典的基于 SIFT 特征的检索方法。

1. 未压缩特征性能分析

如表 5.1 所示，首先在三个数据集上进行未压缩的全尺寸特征的检索性能对比。在图 5.4（左）中，HDSH-*Mean* 特征获得了最好的检索性能，因此后续实验均使用这个特征的层次化版本（HDSH-*Mean*-0.2）和非层次化（HDSH-*Mean*-0）版本。虽然本文并没有致力于在未压缩的特征上产生最好的性能，但是 HDSH 仍然获得了相当有竞争力的结果。具体来看，HDSH-*Mean*-0.2 显著地超过了大多数基于 SIFT 特征的方法，即使这些方法使用了 BoW 等优秀的编码方案，这个结论证明了 CNN 特征的强大表达能力。最好的结果由两个对比方法获得，它们是基于对偶几何匹配、汉明嵌入和多重分配融合的传统方法[197]，以及基于 CNN 的多分辨率空间搜索算法[151]。但

值得注意的是，本文的框架并不仅仅局限于当前的配置，它可以很容易地融合到其他的重排序算法和多分辨率算法中。此外，与其他基于 CNN 的方法相比，HDSH-*Mean*-0.2 的性能也超出了大多数竞争对手，即使它们使用了额外数据[152]，或者使用了耗时的多尺度滑动窗口技术[151]。CNNaug-s[199]方法和 Spatial Pooling[200]方法产生了和 HDSH-*Mean*-0.2 较为相近的性能，但对比方法在多尺度空间上进行特征搜索的方法同样可以改进 HDSH 的性能。最重要的一点，本文基于 CNN 的特征学习框架使用的是标准的 Alexnet[23]网络，在后续的研究中，大量的新网络结构被证明具有更强的特征表达能力。可以预见的是，使用 VGG[3]网络或者 ResNet[28, 29]网络替换现有的 Alexnet[23]将大幅提高特征的表达能力，从而提高检索性能。

表 5.1 未压缩特征的性能比较

方法	Holidays	Oxford5k	Oxford105k
基于 SIFT 特征的方法			
BoW 200k-D[195]	0.54	0.364	–
Improved FV[20]	0.626	0.414	–
VLADintra[19]	0.653	0.558	–
LCS+RN[198]	0.658	0.517	0.456
CVLAD[196]	0.827	0.514	-
HE+MA+PGM[197]	**0.892**	**0.737**	–
基于 CNN 特征的方法			
Neural Codes[152]	0.793	0.545	0.512
MOP-CNN[151]	0.808	–	–
LFDN[154]	0.840	0.581	0.542
CNNaug-ss[199]	0.843	0.68	–
Spatial Pooling[200]	**0.896**	**0.843**	**0.795**
DHRS[194]	0.858	0.712	0.603
HDSH-Mean-0	0.829	0.597	0.523
HDSH-Mean-0.2	**0.885**	**0.744**	**0.712**

注：粗体字标识出实验中的最好结果

为了进一步证明层次化深度语义哈希算法的性能，表 5.2 中对比了 *Oxford5k* 数据集每一个类的 mAP 性能。因为 CNN 是一种典型的数据驱动算法，因此，足够的原始图像对于训练一个好的模型显得尤为重要。如表 5.2 所示，有 4 个类的 mAP 性能低于平均性能，但是不难发现这些类中原始图像都极其稀少。特别是类别“pitt_rivers”和“keble”分别仅仅只有 1 幅图像和 2 幅图像。虽然手工地将每个类别的图像扩展到了 1000 幅，但是原始图像的数量仍然比扩展图更为重要。因为从原始图像中，模型可以学习更多针对类别的不变性。这个结果暗示了，通过增加原始图像的数量能有效地改进这些类的检索性能。此外，大多数情况下，层次化的 HDSH 算法相比非层次化的算法具有较大的优势。

表 5.2 *Oxford5k* 数据集上的每个类的检索结果

方法	Souls	Ashm	Ball	Bodle	Christ	Corn	Hert	Keble	Magd	Pitt	Red	**平均**
HDSH-Mean(0)	0.57	0.49	0.44	0.80	0.60	0.39	0.94	0.43	0.29	0.68	0.96	0.60
HDSH-Mean(0.2)	0.91	0.79	0.38	0.96	0.91	0.25	0.99	0.48	0.84	0.68	1.00	0.74
HDSH-H128(0)	0.96	0.92	0.52	0.80	0.83	0.29	1.00	0.44	0.74	0.28	1.00	0.71
HDSH-H128(0.2)	0.97	0.94	0.49	0.91	0.94	0.26	1.00	0.44	0.91	0.29	1.00	0.74
样本数	72	20	7	19	73	4	49	2	49	1	216	512

2. 压缩特征性能分析

为了权衡检索精度、检索速度和存储空间，大多数方法将低级特征转换为低维特征表达。与这些方法不同，HDSH 直接从 CNN 特征生成哈希编码。如图 5.3（上）所示，本文聚焦于使用端到端的方法同时生成紧凑的哈希码和语义特征。

在表 5.3 中，HDSH 的阈值 $t = 0.2$，并在所有的数据集上和当前最好的检索方法进行比较，获得了非常有竞争力的性能。多种不同的低维特征被用来验证算法的有效性。对于所有的基于 SIFT 的方法，HDSH 都算法远远超过了它们的性能，这再一次证明了 CNN 特征的强大性能。此外，HDSH 在所有尺度的降维特征上都超出了 Neural Codes[152]方法，即使 Neural Codes[152]在额外的图像数据集上进行了微调训练。

最有趣的是最大的性能鸿沟主要出现在较低维的版本，如：16bit，32bit 和 64bit。Spatial Pooling[200]方法使用复杂的编码方式捕捉卷积特征的局部信息来替代简单的最大池化，然而，相比全尺寸的特征，较低维的 256bit 的特征性能下降明显，这说明空间搜索估计方法在低维特征空间中降低了 CNN 的特征表达能力。此外，HDSH 的性能也超过了使用 VLAD 编码的 512 维的 MOP-CNN[151]方法。从实验结果可以发现，从深度哈希层抽取的卷积特征在降低哈希编码维度的时候精度并没有明显降低，这证明了 HDSH 的稳定性，即使使用低维表达也能够保持较好的判别能力。

表 5.3 低维压缩特征的性能比较

方法	维度	Holidays	Oxford5k	Oxford105k
LCS+RN[198]	16	0.323	0.27	0.222
Neural Codes[152]	16	0.609	0.418	0.354
HDSH-H16-0.2	16	**0.815**	**0.722**	**0.665**
Neural Codes[152]	32	0.729	0.515	0.467
HDSH-H32-0.2	32	**0.858**	**0.723**	**0.665**
Neural Codes[152]	64	0.777	0.548	0.508
HDSH-H64-0.2	64	**0.856**	**0.737**	**0.671**
FV + T[205]	128	0.617	0.433	-
VLADintra[19]	128	0.625	0.448	-
LCS+RN[198]	128	0.335	0.322	0.262
Neural Codes[152]	128	0.789	0.557	0.523
LFDN[154]	128	0.836	0.558	52.900
HDSH-H128-0.2	128	**0.858**	**0.74**	**0.676**
Neural Codes[152]	256	0.789	0.557	0.524
DHRS[194]	256	0.818	0.574	0.49
Spatial Pooling[200]	256	0.742	0.533	0.511
HDSH-H256-0.2	256	**0.858**	**0.754**	**0.688**
MOP-CNN[151]	512	0.784	-	-
Neural Codes[152]	512	0.789	0.557	0.522
DHRS[194]	512	0.838	0.672	0.563
HDSH-H512-0.2	512	**0.86**	**0.768**	**0.693**

注：粗体字标识出实验中的最好结果

5.4.6 基于 *Imagenet* 大规模数据集的性能分析

接下来，大规模的 *Imagenet*[73]数据集上被用来评估各种相似性算法的有效性和稳定性。为了学习层次化深度语义哈希，HDSH 需要按照图 5.3 的 CNN 模型来训练和抽取二进制编码特征和语义特征。本文采用多种优秀方法[204]和 HDSH 算法进行了性能对比，它们主要包括：（1）B-Hie：一种使用先验矩阵的双线性相似性哈希算法；（2）Cosine-Hie：与 B-Hie 类似，但是用 *L2* 正则概率向量作为先验矩阵；（3）B-Flat：一种没有使用层次化编码的双线性相似性算法；（4）Cosine-NoCal：一种基于语义的余弦相似性算法，但没有使用概率校正；（5）Cosine-Flat：一种没有使用概率校正的余弦相似性算法；（6）SPM：一种非学习的算法，它通过将低级特征进行相位调制并编码成视觉词典，然后通过交互核心来进行排序的相似性算法；（7）Hard-Assign：将查询图像直接分类到最相关的类别，并根据分类概率进行相似性排序。

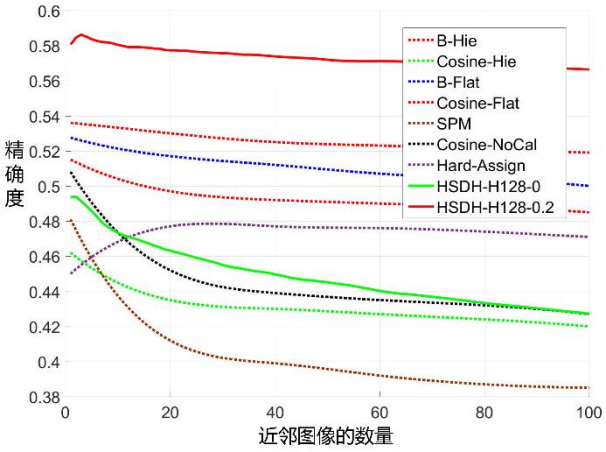


图 5.5 *Imagenet* 数据集上检索精度对比图

图 5.5 中给出了这些对比算法和 HDSH 算法的检索性能曲线。HDSH 在阈值 $t = 0.2$ 的设置下获得了最好的性能。从图中可以发现层次化的应用对于 HDSH 的哈希编码也非常重要，它大约带来了超过 20%的精确度的提升。与排名第二的算法相比，HDSH-*H128*-0.2 总是有 7%以上的优势。此外，这个性能是 HDSH 方法在 128bits 的低维版本上获得，相比数千维度的全尺寸特征，在检索速度上有较大的优势。

5.4.7 层次化的效率分析

值得注意的是，层次化深度语义哈希方法最大的改进是显著降低了搜索时间。本文中，所有的算法都使用欧式距离来计算两幅图像的相似性。表 5.4 中，HDSH-xxx-0 表示直接使用 CNN 特征计算相似性，HDSH-xxx-0.2 表示使用阈值 $t = 0.2$ 的层次化方法来计算相似性。结果显示，层次化方法对于图像检索是非常有用的，在所有的五个数据集上都有明显的速度提升。此外，从实验结果来看，速度改进对于不同的数据集并非完全一致，显而易见的是，加速比非常依赖于数据集总的类别数量。事实上，计算语义相关性几乎是零开销的，这使检索时间主要依赖于建议集的规模。因此，当数据集的类别越多的时候，HDSH 能够过滤更多的不相关的类别，从而显著减小搜索空间。

表 5.4 所有数据集上的检索时间对比（单位：毫秒）

方法		Holidays	Oxford5k	Oxford105k	Caltech256	Imagenet
类别数	维度	500	12	12	257	1000
HDSH-mean-0	4K	138	693	3504	1121	45558
HDSH-mean-0.2	4K	0.83	167	666	13	333
加速比		167.3	4.14	5.26	87	134.9
HDSH-H128-0	128	8.1	114	613	198	8900
HDSH-H128-0.2	128	0.15	28	140	5.4	54
加速比		54	4.15	4.37	36.7	165.1

此外，本文从资源消耗的角度研究了 HDSH 算法的性能。与其他通用的基于 CNN 的方法类似，HDSH 使用 GPU 完成前向传输和特征提取。在这个过程中，其他资源的消耗可以被忽略，内存消耗是最主要的问题。在实际运行的检索系统中，图像库中的图像的特征可以被事先计算和保存。图 5.6 显示了每 10,000 幅图像的内存空间消耗情况。纵坐标为每 10,000 幅图像所需要内存空间量，横坐标为不同长度哈希码的 HDSH 方法以及未使用 HDSH 而直接计算特征欧氏距离的方法。增加哈希码的长度能够提高性能（如表 5.3 所示），但是也增加了不期望的内存消耗。值得注意的是，在 *Holidays* 数据集上，未使用层次化策略的方法消耗了 240M 字节的内存空间，但是只完成了 0.829 的 mAP，这个结果甚至比更低维的 HDSH 方法还差，例如：128bits

的 HDSH 方法完成了 0.855 的 mAP。同样的现象在 *Oxford5k*、*Oxford105k* 和 *Imagenet* 数据集上也得到了验证。

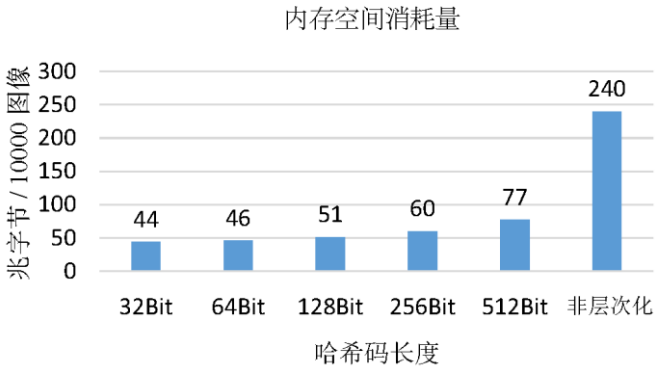


图 5.6 不同方法的内存空间消耗对比

5.4.8 跨类泛化性能分析

层次化的深度语义哈希还有一个潜在的优势，它将识别能力更好地泛化到那些在训练中不存在的类别。本文使用两个大规模的数据集 *Imagenet* 和 *Caltech256* 去完成这个任务。对于 *Imagenet* 数据集，首先从验证集和测试集中随机选择了其中的 100 个类别作为检索目标，对于训练集，制定两种不同的实验策略：（1）非泛化验证，即使用完整的 1000 个类的图像进行训练，以“seen”作为标示；（2）泛化验证，即使用除 100 个类以外的其他 900 个类的图像进行训练，以“unseen”作为标示。从图 5.7（左）可以看出，虽然“unseen”训练的模型性能远低于“seen”训练的模型，但 HDSH 远超过其他同等设置的基准模型[204]，即使使用的是 128bit 的压缩编码。深受鼓舞的是，层次化深度哈希方法始终优于未使用层次化的方法 HSDH-*H128*-0-xxx。这意味着，层次化语义哈希对于跨类的泛化性能是有积极意义的。类似于 *Imagenet* 数据集，*Caltech256* 数据集也分为两个部分，200 个类用于训练，56 个类用于检索。换句话说，这 56 个类在训练中是不可见的。图 5.7（右）展示了 *Caltech256* 数据集上的检索结果。“unseen”模型的性能远低于“seen”模型的性能，但是仍然远远优于其他方法。这再一次证明了 HDSH 方法的泛化能力，它能够有效地识别那些训练中不可见

的类别。值得注意的是，在所有实验中，训练图像和测试图像是互斥的。

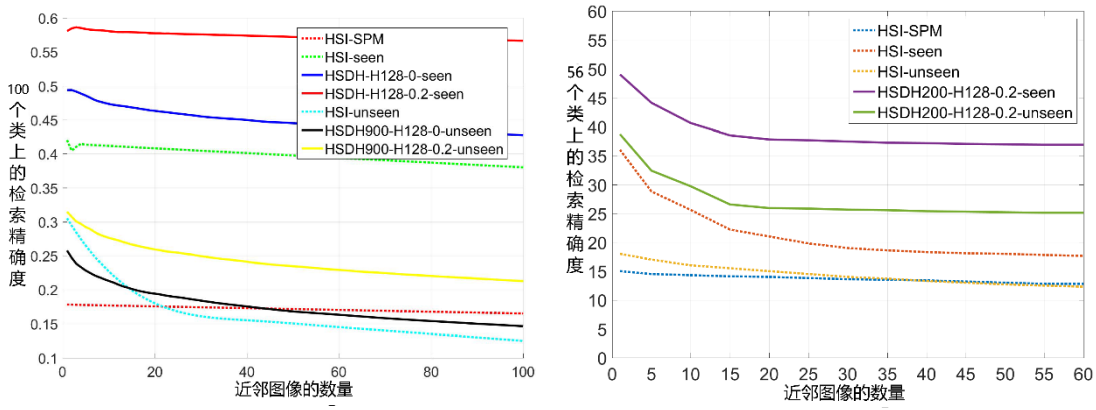


图 5.7 *Imagenet*（左）和 *Caltech256*（右）数据集上的跨类检索性能对比

在真实世界的应用中，在一个开放的环境中进行检索是非常有挑战性的。例如，如果我想要从 *Imagenet* 数据集中检索一个概念“墨西哥猪肉汤”，而这个概念并不是 *Imagenet* 数据集中存在的一个类别。感谢于 *Imagenet* 数据集提供丰富的类别空间，HDSH 系统可能会返回一个相关容器类中的所有图像，例如“汤碗”。这应该是一个非常接近真实需求的答案，因为“汤碗”中很可能装的正是“墨西哥猪肉汤”。但是，如果查询图像的语义信息和模型训练时已有的类别差别很大，例如：传说中的怪兽“美杜莎”，那么查询结果可能就会变得很尴尬，即使系统返回的候选建议集是“最相似”的类别。因此，由于 HDSH 是基于监督训练的方法，它将更适用于指定类别空间的检索环境，即使它的泛化能力远优于大多数的传统方法和深度方法。

5.5 小结

本章提出了一种新颖的快速图像检索算法—层次化深度语义哈希（Hierarchical Deep Semantic Hashing, HDSH），该算法有效地解决了基于内容的图像检索中最重要的两个问题：检索精度和检索效率。该工作主要有四个方面的贡献：

1. 提出了一个完整的端对端的特征学习和提取框架，它可以同时输出基于概率的语义级特征和哈希级特征。

2. 通过组合语义级相似性和哈希级相似性，HDSH 不仅为计算图像的距离提供了强大的先验知识，提高了检索的性能，也大幅缩小了检索空间，使算法可以被应用到大规模的数据集上。
3. 大量的实验证明，HDSH 算法具有较强的稳定性，即使特征维度降低到一个较小的尺度，例如 128bit 的哈希编码，它仍然具有较强的判别能力。
4. 本章还提出了一种简单但是有效数据扩展方法，用来解决数据集样本和样本类别不平衡问题。

事实上，本章的工作主要是验证层次化的上下文语义是否可以帮助提高检索的性能，因此，通过将原始的 Alexnet[23]模型替换为更强大的深度模型（例如 VGG16[3]或 ResNet[21, 22]）将进一步改善系统的性能。

6 总结与展望

6.1 工作总结

在互联网络飞速发展的今天，图像、视频内容也迎来了快速增长和传播的契机。这一方面丰富了人民群众文化生活、为人民群众日常生活带来了极大的便利；但是另一方面，也为人们使用资源和社会发展带来了负面的影响。因此，面对爆炸式增长的资源，在处理诸如内容识别、目标检测、图像检索、语义分割等方面应用的时候，如何从海量数据中获得所需的内容，进而进行分析、处理，都是大数据环境下亟待解决的问题。本文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，以四个应用任务为基础，深入研究了如何利用上下文语义信息来辅助视觉内容的识别、检索和解析。本文的主要工作与创新成果主要归纳如下：

1. **针对成人识别任务中样本多样性问题，提出基于高层语义的细到粗策略和多上下文联合决策**

对于成人内容识别任务的两个主要问题：类别空间稀少导致的类内距大于类间距和大数据环境中样本多样性导致的分类困难的问题，分别提出基于高层语义的细到粗策略和基于多上下文语义联合决策的方案。前者用于处理类别空间狭窄的问题，传统的成人内容识别通常是一个二分类问题，而复杂多样的样本可能会产生某些样本类内距离大于类间距离的问题。本文通过基于高层语义的细到粗策略，首先用一个较大的类别空间去执行分类任务，更细的类别划分能够较好地区分大部分样本的高层语义，接下来再将细粒度的类别直接映射到二类问题上。该策略可以有效地改进成人识别分类器的性能。在处理正负样本多样性的问题时，本文结合全局上下文、局部上下文和跨上下文信息来进行联合决策。与传统特征融合的方法不同，本文使用的是一种策略融合的方式，这种方法最大限度地保证了基于分类的全局上下文的准确性，然后利用基于检测的局部上下文信息来尽力修正被误判的样本，从而实现召回率和准确率的同时提升。为了将这两种策略应用到统一的模型中，本文提出了一种多上下

文深度学习框架（Deep Multi-Context Network, DMCNet）用于实现端到端的推理。此外，模块化的设计方案，允许通过更新全局上下文建模或局部上下文建模完成整个网络性能的提升。

2. **针对场景解析任务中难对象识别问题和额外背景类造成的误判问题，提出基于局部语义的特征增强策略和语义黑洞填充策略**

与通用的以对象为中心的语义分割和实例分割任务不同，场景解析任务需要面对的是场景中种类繁多的对象和纷繁复杂的背景，而场景中的对象通常还具有尺度较小、交互性多（容易产生遮挡）、隐藏性强（容易被背景元素所湮灭）等特性。本文提出两种策略基于局部上下文语义的对象区域增强和黑洞填充策略来缓解这些难点对场景解析的影响。对象区域增强策略主要用来解决难对象的解析问题，利用检测网络生成一些置信度较高的对象区域，然后直接用这些区域去增强全局上下文模式下得到的卷积特征图的特定类别通道的局部区域，从而通过增强这些难对象的特征强度，实现对全局上下文模式下场景解析的准确性。黑洞填充策略用于处理额外背景类导致的解析黑洞问题，这些额外的背景类被用于在训练中处理边界像素区域和难像素区域，从而降低训练难度，提高深度模型的性能。为了实现这两个策略，本文提出了一种基于深度学习的对象区域增强网络（Objectness Region Enhancement Network，OENet），同样模块化的设计方案使模型不但可以通过更换模块实现整体解析性能的提升，还可以将两个策略应用到其他现有的场景解析网络中。

3. **针对人像妆容迁移任务中上下文融合时语义保持困难的问题，提出对称加权交叉熵损失和基于迭代的全局上下文和局部上下文融合网络**

妆容迁移是一个非常有趣的任务，它基于人脸解析的结果，并涉及到生成网络的应用。该任务的难点主要是：（1）如何获得精确的人脸解析结果；（2）如何按需地保持（如：待化妆脸的脸型、五官）和迁移（如：唇彩、粉底和眼影）人像的特征。针对妆容迁移任务，提出了一种基于对称加权的交叉熵损失用于完成人脸解析任务，一方面保证强化指定局部区域的解析结果，另一方面也保证眼影，嘴唇等特殊区域在人

像上的对称性要求。为了实现妆容迁移，提出了一种深度局部妆容迁移网络（Deep Localized Makeup Transfer Network，DLMTN），该网络是一个基于随机梯度下降的迭代式的生成网络。该网络可以为不同的区域定制不同的妆容特征迁移。针对形状敏感的眼影、脸型等区域，采用多层深度特征融合的特征进行迁移；针对纹理敏感的粉底、唇彩等区域，采用二阶的 Gram 特征进行迁移。通过端到端的 DLMTN 生成网络，不但可以产生自然的妆容迁移效果，还可以实现妆容浓淡程度的自由调节，这使得该系统的可用性大大增强。

4. **针对大数据环境下搜索空间太大引起的效率降低问题，提出基于概率的层次化语义级相似性过滤策略**

查询的准确性和查询的效率是大数据环境下，基于内容的图像检索最难也是最需要解决的问题。产生这个两个问题的根本原因，一是特征的鲁棒性问题，二是样本的规模问题。对于数百万的样本进行逐对相似性计算使得大多数传统的图像检索都面临无效的窘境。本文提出了一种基于层次化语义的过滤方法来解决这个问题。通常情况，相似的样本应该具有相同或相近的高层语义，通过深度卷积神经网络生成这种高层语义信息，并在一定范围内进行样本过滤，这个过程接近零开销，但是可以排除大量的不相关样本，进而使后续的基于哈希的相似性计算变得更轻松。为了同时生成高层语义信息和哈希编码，提出了一种基于深度学习的框架层次化深度语义哈希（Hierarchical Deep Semantic Hashing，HDSH），通过该网络，可以端对端地输出样本的高层语义和哈希编码。结合层次化语义过滤和哈希相似性计算，使 HDSH 方法非常适用于大数据环境下的图像检索任务。

6.2 研究展望

尽管本文基于多种上下文语义信息在视觉内容的识别与分析上取得了一定的研究成果，但其效果还有很大的改进空间，对于希望广泛应用到实际应用中的计算机视觉任务，依然还有很多方面可以探索。下一步，我们期望从以下几个方面继续这方面的工作。

1. 图像描述和全面的理解场景

随着机器学习的发展，计算机视觉领域已经取得了巨大的成果。计算机视觉最初的任务是感知与识别，例如图像分类、目标检测等，然而，要让计算机更加智能，就必须要从感知过度到认知。场景解析是视觉任务发展的第二个阶段，它不仅仅区分图像是什么，它还需要去了解图像里有什么内容，这些内容位于图像的什么位置，但是这对于实现机器认知依然不够。下一步，还希望给场景中的对象和背景赋予更多的元素和知识，更好地去理解场景中上下文的含义和他们之间的关系。例如：每一个对象的属性，对象的行为，对象与对象之间的交互关系，对象的统计信息等。进而，可以用大量形式化的语言去描述一个场景，甚至进行人机交互。2016 年，斯坦福人工智能实验室推出了视觉基因 Visaul Genome 项目，旨在推动场景理解的进程。相信终有一天，计算机可以完美地回答我们这些问题：“这幅图片中有几个孩子？它们在做什么？他们是好朋友吗？他们下一步想要做什么？”。为了实现这个目标，场景描述和更深入全面的理解场景，都是我们希望去研究的内容。

2. 生成对抗网络

深度学习飞速发展的这 5 年，几乎在所有的领域都获得了巨大的成功。然而，时至今日，这些巨大的成功后面都离不开庞大的数据库的支持。监督学习，或者半监督、弱监督学习已经成深度学习必不可少元素之一。然而，现实中不可能对于任何任务都象图像分类一样去建立如 *Imagenet* 这样数百万样本的庞大数据集，用来执行监督训练。对于很多领域，并没有那么精力和财力去完成这项工作。事实上，可以想象一下一个 5 岁小孩学习的过程，在他的认知过程中，并不是每一次数据的接收都可以得到父母确定的答案引导（监督学习），在他的大脑中充斥着数以亿计的画面，而这些画面通常都没有确定的答案（标签），大多数情况下，都是通过他已有的认知不停地去判断和修正新接收到信息。这个过程既是无监督的，也是博弈的。生成对抗网络[206]（Generative Adversarial Network，GAN）的提出正好切合这个认知的过程。它通过同时存在的生成模型和判别模型的不断博弈，在努力达到纳什平衡的过程中，逐渐修正模型，甚至产生新的损失函数，以达到最优。幸运的是 GAN 的思想几乎可以用到

所有的基于深度学习的模型中，这也是我下一步期望去研究的方向。

3. 预测学习

2016 年巴塞罗那的 NIPS2016 国际会议上，Yann LeCun 给大家做了一场生动和激动人心的报告，Yann 在这次大会上第一次提出了预测学习（Predictive Learning）这个概念。在人工智能（Artifical Intelligence，AI）发展的过程中，我们已经做了很多的基础性工作，然而一直都错过了一个关键的因素，那就是预测！它是指机器给真实环境建模、预测可能的未来，并通过观察和演示来理解世界适合运行的能力。在过去的几年中，我们一直在通过给机器增加更多的能力（如：数据、知识和智能体（算法模块））来建立监督式的学习方法，然而这种人为的增加能力的方法，并不符合可持续发展的需求。在人工智能的发展中，机器不仅需要学习大量的背景知识，还需要去学习和理解世界是如何运行的，观察世界的状态，更新并记忆对世界状态的评估，更重要的是能够实现推理和规划。预测学习，不仅仅能够在无监督的状态下进行学习，更重要的是它能够习得一种预测和推理世界的模型。在 AI 领域，如果一个机器能够实现生成具有高度真实感的数据，那么它就发展出了对预测模型的理解。这是 GAN 的范畴，但也是预测学习的基础。由此可见，预测学习很可能也会落脚于 GAN。不管如何，预测学习都是 AI 的未来和发展方向，它也是我下一步期望去学习和研究的方向。

致谢

光阴荏苒，岁月如梭，转眼之间我博士求学生涯也即将画上圆满的句号。对于一个年近 35 岁的人来说，这四年是漫长而艰辛的，这期间有留下了很多的遗憾、辛酸与血泪，但更多的是收获和成长。在整个博士生涯中，得到了许多人长期的帮助与支持，在此郑重地说一声谢谢！

本文的研究工作是在导师凌贺飞教授的精心指导下完成的，从论文的选题到各阶段研究方向的把握和工作计划的制定，以及论文撰写与修改的每一个细节都倾注了导师的汗水和心血。导师严谨认真的治学态度，诲人不倦的师者风范，深厚的学术造诣，敏锐的学术洞察力深深地影响了我。这段珍贵的学习经历令我终生难忘，是我一生都享用不尽的财富。我能取的今天这样的成就，与导师长期的谆谆教诲是分不开的。在论文完成之际，表示衷心的感谢和崇高的敬意！

感谢华中科技大学的邹复好副教授、李平老师在不同方面给予我的帮助。同时还要感谢实验室的刘聪博士、严灵毓博士、雷洁博士以及全体硕士成员和计算机学院 2013 级博士班的所有同学在枯燥的学习生活中的互相扶持、关心和帮助。同时要感谢考博群的曾晶博士、陈琼英博士、蓝山博士、雨滴博士以及全体群友，是大家的互相帮助、扶持与鼓励才让我获得了博士学习的机会。

衷心感谢我在中国科学院信息工程研究所客座期间，亦师亦友的刘偲副研究员在我的科研道路上给予我的极大帮助，让我学会如何更好地进行学术工作；同时感谢她在我论文修改与润色上的帮助，论文质量的提高离不开她的努力。感谢中国科学院信息工程研究所的操晓春研究员和孙瑶副研究员在平时科研中的指导和帮助。此外，还要感谢信工所的王瑋博士、张三义博士、陈智勇、魏震、钱瑞和、廖越、庾涵、高佳俊、任乐健、朱德发、包仁达、任广辉在共同学习和工作期间结下了深厚的友谊和相互的支持，让我圆满地完成了学习任务。

感谢中科院计算所的山世光研究员和他领导创办的 VASLE，让我在迷茫的学术道路上找到了曙光，在每周一次的 Webinar 活动中学习到了大量最顶尖、最前沿的学

术知识。

感谢云南开放大学的李昭明副校长、田云鹏院长、吕恬副院长和吕椽副院长，在我离岗求学的过程中给予了我极大的支持和帮助，也让我毫无后顾之忧地全身心地投入到学术工作中。

衷心感谢我的父亲和母亲。感谢他们一直以来对我的关心、支持、照顾与养育之恩。在我成长和学习的道路上，他们是我最坚强的后盾，给予了我源源不断的精神动力。感谢我的岳父、岳母对我求学的支持、理解和帮助，我的学习离不开他们一如既往的无私奉献和默默支持。

特别感谢我的妻子王艺瑾和儿子欧卓轩多年来对我的支持、信任、理解和包容，每当我遇到困难与挫折和感到彷徨的时候，是他们给予了我坚持下去的信心和决心。同时，他们也是我克服困难、积极进取的动力源泉。在我人生路上的每一份成果都与他们密不可分，都有他们的付出与汗水。我为能有如此优秀的人生伴侣和机灵、可爱的儿子而感到骄傲和幸福！

衷心感谢在百忙之中抽出时间审阅本论文的各位专家教授！

最后，感谢我生命中的每一个人！

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] Biederman Irving, Mezzanotte Robert J., Rabinowitz Jan C. Scene perception: Detecting and judging objects undergoing relational violations. Cognitive Psychology. 1982. 14(2):143-177

[6] 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

[7] 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

[8] 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

[9] 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

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

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

[12] Lowe David G. Object Recognition from Local Scale-Invariant Features. in: IEEE International Conference on Computer Vision. 1999. 1150-1157

[13] Dalal Navneet, Triggs Bill. Histograms of Oriented Gradients for Human Detection. in: IEEE Conference on Computer Vision and Pattern Recognition. 2005. 886-893

[14] 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

[15] Nasrabadi Nasser M., King Robert A. Image coding using vector quantization: a review. IEEE Transactions Communications. 1988. 36(8):957-971

[16] 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

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

[18] 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

[19] Arandjelovic Relja, Zisserman Andrew. All About VLAD. in: IEEE Conference on Computer Vision and Pattern Recognition. 2013. 1578-1585

[20] 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

[21] 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

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

[23] 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

[24] 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

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

[26] 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

[27] 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

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

[29] 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

[30] 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

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

[32] 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

[33] 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

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

[35] 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

[36] 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

[37] 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

[38] 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

[39] 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

[40] 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

[41] 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

[42] 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

[43] 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

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

[45] 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

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

[47] 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

[48] 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

[49] 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

[50] 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

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

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

[53] 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

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

[55] 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

[56] 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

[57] 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

[58] Fellbaum C., Miller G. WordNet : An Electronic Lexical Database. The Library Quarterly: Information, Community, Policy. 1999

[59] 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

[60] 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

[61] 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

[62] 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

[63] Yang Cong, Tiebe Oliver, Shirahama Kimiaki, et al. Object Matching with Hierarchical Skeletons.

Pattern Recognition. 2016. 5(5):183-197

[64] 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

[65] 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

[66] 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

[67] 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

[68] 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

[69] 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

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

[71] 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

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

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

[74] 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

[75] 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

[76] Ciresan Dan C., Meier Ueli, Masci Jonathan, et al. Multi-column Deep Neural Network for Traffic Sign Classification. Neural Networks. 2012. 32(1):333-338

[77] Parkhi Omkar M., Vedaldi Andrea, Zisserman Andrew. Deep Face Recognition. in: Proceedings of the British Machine Vision Conference. Swansea, UK. 2015. 4101-4112

[78] 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

[79] 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

[80] 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

[81] 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

[82] 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

[83] Zeiler Matthew D., Fergus Rob. Visualizing and Understanding Convolutional Networks. in: European Conference Computer Vision (ECCV). Zurich, Switzerland. 2014. 818-833

[84] 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

[85] 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

[86] Min Lin Qiang Chen And Shuicheng. Network in network. arXiv preprint. 2013

[87] 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

[88] Ng Al Maas Ay Hannun. Rectifier Nonlinearities Improve Neural Network Acoustic Models. in: International Conference on Machine Learning. 2013

[89] 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

[90] 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

[91] 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

[92] 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

[93] 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

[94] 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

[95] 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

[96] 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

[97] 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

[98] 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

[99] 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

[100] Girshick Ross B. Fast R-CNN. in: IEEE International Conference on Computer Vision. 2015. 1440-1448

[101] 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

[102] 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

[103] 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

[104] 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

[105] 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

[106] 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

[107] 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

[108] 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

[109] 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

[110] 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

[111] 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

[112] 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

[113] 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

[114] 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

[115] 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

[116] 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

[117] 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

[118] 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

[119] 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

[120] 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

[121] 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

[122] 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

[123] 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

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

[125] 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

[126] 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

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

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

[129] 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

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

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

[132] 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

[133] 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(1):99-103

[134] 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

[135] 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

[136] 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

[137] 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

[138] 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

[139] 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

[140] 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

[141] 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

[142] 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

[143] 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

[144] 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

[145] 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

[146] 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

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

[148] 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

[149] 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

[150] 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

[151] 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

[152] Babenko Artem, Slesarev Anton, Chigorin Alexander, et al. Neural Codes for Image Retrieval. in: European Conference on Computer Vision. 2014. 584-599

[153] 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

[154] 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

[155] 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

[156] Ou Xinyu, Ling Hefei, Yu Han, et al. Adult Images and Videos Recognition by Deep Multi-Context Network and Fine-to-Coarse Strategy. ACM Transactions on Intelligent Systems and Technology (TIST). 2017

[157] Ou Xinyu, Wei Zhen, Ling Hefei, et al. Deep multi-context Network for Fine-Grained Visual

Recognition. in: 2016 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). 2016. 1-4

[158] Legaldictionary. Pornographic. 2015. http://legal-dictionary.thefreedictionary.com/pornography

[159] China Prc. Criminal law. 2015. http://www.lawtime.cn/faguizt/23.html

[160] 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. 2015. 27-35

[161] Pedersoli Marco, Vedaldi Andrea, Lez Jordi Gonz A., et al. A Coarse-to-Fine Approach for Fast Deformable Object Detection. Pattern Recognition. 2015. 48(5):1844-1853

[162] Saltzer Jerome H., Reed David P., Clark David D. End-to-End Arguments in System Design. in: Proceedings of the 2nd International Conference on Distributed Computing Systems. Paris, France. 1981. 509-512

[163] 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. Columbus, OH, USA. 2014. 1725-1732

[164] Lin Tsung Yu, Chowdhury Aruni Roy, Maji Subhransu. Bilinear CNN Models for Fine-Grained Visual Recognition. in: IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 1449-1457

[165] Zhao Rui, Ouyang Wanli, Li Hongsheng, et al. Saliency Detection by Multi-Context Deep Learning. in: IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 1265-1274

[166] Srivastava Rupesh Kumar, Greff Klaus, Schmidhuber J. U. Rgen. Highway Networks. CoRR. 2015

[167] Moustafa Mohamed. Applying deep learning to classify pornographic images and videos. arXiv. 2015

[168] Wang Chao, Zhang Jing, Zhuo Li, et al. Incremental Learning for Compressed Pornographic Image Recognition. in: IEEE International Conference on Multimedia Big Data. 2015. 176-179

[169] Jia Yangqing, Shelhamer Evan, Donahue Jeff, et al. Caffe: Convolutional Architecture for Fast Feature Embedding. in: ACM International Conference on Multimedia. 2014. 675-678

[170] Caetano Carlos, de Avila Sandra, Eliza Fontes, et al. Pornography Detection using BossaNova Video Descriptor. in: 22nd European Signal Processing Conference. Lisbon, Portugal. 2014. 1681-1685

[171] Xinyu Ou Hefei Ling Si Liu. Objectness Region Enhancement Networks for Scene Parsing. Journal

of Computer Science and Technology (JCST). 2017

[172] Dai Jifeng, He Kaiming, Sun Jian. Instance-Aware Semantic Segmentation via Multi-task Network Cascades. in: IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 3150-3158

[173] Pinheiro Pedro H. O., Collobert Ronan, R Piotr Doll A. Learning to Segment Object Candidates. in: Advances in Neural Information Processing Systems. Montreal, Quebec, Canada. 2015. 1990-1998

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

[175] Dai Jifeng, Li Yi, He Kaiming, et al. R-FCN: Object Detection via Region-based Fully Convolutional Networks. in: Advances in Neural Information Processing Systems. Barcelona, Spain. 2016. 379-387

[176] Hong Seunghoon, Noh Hyeonwoo, Han Bohyung. Decoupled Deep Neural Network for Semi-supervised Semantic Segmentation. in: Advances in Neural Information Processing Systems. Montreal, Quebec, Canada. 2015. 1495-1503

[177] Chen Liang Chieh, Papandreou George, Kokkinos Iasonas, et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. in: International Conference on Learning Representations. 2015

[178] Wang Yuhang, Liu Jing, Li Yong, et al. Objectness-aware Semantic Segmentation. in: Proceedings of the 2016 ACM Conference on Multimedia Conference. Amsterdam, The Netherlands. 2016. 307-311

[179] Yu Fisher, Koltun Vladlen. Multi-Scale Context Aggregation by Dilated Convolutions. CoRR. 2015

[180] Badrinarayanan Vijay, Kendall Alex, Cipolla Roberto. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. CoRR. 2015

[181] Cordts Marius, Omran Mohamed, Ramos Sebastian, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding. in: IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 3213-3223

[182] Ou Xinyu, Liu Si, Cao Xiaochun, et al. Beauty eMakeup: A Deep Makeup Transfer System. in: Proceedings of the 2016 ACM Conference on Multimedia Conference (ACMMM). Amsterdam, The Netherlands. 2016. 701-702

[183] Mahendran Aravindh, Vedaldi Andrea. Understanding Deep Image Representations by Inverting Them. in: IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 5188-5196

[184] Gatys Leon A., Ecker Alexander S., Bethge Matthias. A Neural Algorithm of Artistic Style. CoRR. 2015

[185] Bookstein Fred L. Principal Warps: Thin-Plate Splines and The Decomposition of Deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1989. 11(6):567-585

[186] Scherbaum Kristina, Ritschel Tobias, Hullin Matthias B., et al. Computer-Suggested Facial Makeup. Computer Graphics Forum. 2011. 30(2):485-492

[187] Ou Xinyu, Ling Hefei, Yan Lingyu, et al. Convolutional Neural Codes for Image Retrieval. in: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Chiang Mai, Thailand. 2014. 1-10

[188] Ou Xinyu, Ling Hefei, Liu Si, et al. Hierarchical Deep Semantic Hashing for Fast Image Retrieval. Multimedia Tools and Applications. 2016. 1-22

[189] Oliva Aude, Torralba Antonio. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. International Journal of Computer Vision. 2001. 42(3):145-175

[190] Norouzi Mohammad, Fleet David J. Minimal Loss Hashing for Compact Binary Codes. in: International Conference on Machine Learning. 2011. 353-360

[191] Xia Rongkai, Pan Yan, Lai Hanjiang, et al. Supervised Hashing for Image Retrieval via Image Representation Learning. in: Proceedings of the AAAI conference on artificial intelligence. 2014. 2156-2162

[192] Liu Wei, Wang Jun, Ji Rongrong, et al. Supervised hashing with kernels. in: IEEE Conference on Computer Vision and Pattern Recognition. 2012. 2074-2081

[193] 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

[194] Zhao Fang, Huang Yongzhen, Wang Liang, et al. Deep Semantic Ranking Based Hashing for Multi-Label Image Retrieval. in: IEEE Conference on Computer Vision and Pattern Recognition. 2015. 1556-1564

[195] Gou Herv E. J. E., Perronnin Florent, Douze Matthijs, et al. Aggregating Local Image Descriptors into Compact Codes. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012. 34(9):1704-1716

[196] Zhao Wan Lei, Gravier Guillaume, Gou Herv E. J. E. Oriented Pooling for Dense and Non-Dense Rotation-Invariant Features. in: British Machine Vision Conference. 2013

[197] Li Xinchao, Larson Martha, Hanjalic Alan. Pairwise Geometric Matching for Large-Scale Object Retrieval. in: IEEE Conference on Computer Vision and Pattern Recognition. 2015. 5153-5161

[198] Delhumeau Jonathan, Gosselin Philippe Henri, Gou Herv E. J. E., et al. Revisiting the VLAD Image Representation. in: ACM Multimedia Conference. 2013. 653-656

[199] Razavian Ali Sharif, Azizpour Hossein, Sullivan Josephine, et al. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. in: IEEE Conference on Computer Vision and Pattern Recognition. 2014. 512-519

[200] Razavian Ali Sharif, Sullivan Josephine, Maki Atsuto, et al. Visual Instance Retrieval with Deep Convolutional Networks. CoRR. 2014

[201] Jegou Herve, Douze Matthijs, Schmid Cordelia. Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search. in: European Conference on Computer Vision. 2008. 304-317

[202] Philbin James, Chum Ondrej, Isard Michael, et al. Object Retrieval with Large Vocabularies and Fast Spatial Matching. in: IEEE Conference on Computer Vision and Pattern Recognition. 2007

[203] Griffin Gregory, Holub Alex, Perona Pietro. Caltech-256 Object Category Dataset. California Institute of Technology. 2007

[204] Deng Jia, Berg Alexander C., Li Fei Fei. Hierarchical Semantic Indexing for Large Scale Image Retrieval. in: IEEE Conference on Computer Vision and Pattern Recognition. 2011. 785-792

[205] Gou Herv E. J. E., Zisserman Andrew. Triangulation Embedding and Democratic Aggregation for Image Search. in: IEEE Conference on Computer Vision and Pattern Recognition. 2014. 3310-3317

[206] Goodfellow Ian J., Abadie Jean Pouget, Mirza Mehdi, et al. Generative Adversarial Nets. in: Advances in Neural Information Processing Systems (NIPS). Montreal, Quebec, Canada. 2014. 2672-2680

附录 1 攻读博士学位期间发表的学术论文目录

[1] **Xinyu Ou**, Hefei Ling, Han Yu, Fuhao Zou, Ping Li, Si Liu. Adult Images and Videos Recognition by Deep Multi-Context Network and Fine-to-Coarse Strategy. ACM Transactions on Intelligent Systems and Technology (TIST). 2017.（JCR-Q1、SCI/EI，已接收，署名单位：华中科技大学）

[2] **Xinyu Ou**, Hefei Ling, Si Liu, Jie Lei. Hierarchical Deep Semantic Hashing for Fast Image Retrieval. Multimedia Tools and Applications (MTAP). 1-22, 2016.（JCR-Q2、CCF-C，SCI/EI，署名单位：华中科技大学）

[3] **Xinyu Ou**, Zhen Wei, Hefei Ling, Si Liu, Xiaochun Cao. Deep Multi-Context Network for Fine-Grained Visual Recognition. in: 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Seattle, WA, USA. July 2016. 1-4.（CCF-B、EI，署名单位：华中科技大学）

[4] **Xinyu Ou**, Lingyu Yan, Hefei Ling, Cong Liu Maolin Liu. Inductive Transfer Deep Hashing for Image Retrieval. in: Proceedings of the ACM International Conference on Multimedia (ACMMM). Orlando, Florida, USA. November 2014. 969-972.（CCF-A、EI，署名单位：华中科技大学）

[5] **Xinyu Ou**, Hefei Ling, Lingyu Yan, Maolin Liu. Convolutional Neural Codes for Image Retrieval, in: Annual Summit and Conference of Asia Pacific Signal and Information Processing Association (APSIPA). Angkor, Cambodia. December 2014. 1-10.（EI，署名单位：华中科技大学）

[6] **Xinyu Ou**, Hefei Ling, Si Liu. Objectness Region Enhancement Networks for Scene Parsing. Journal of Computer Science and Technology (JCST).（JCR-Q4、CCF-B、SCI/EI，已接收，署名单位：华中科技大学）

[7] **Xinyu Ou**, Si Liu, Xiaochun Cao, Hefei Ling. Beauty eMakeup: a Deep Makeup Transfer System. in: Proceedings of the ACM International Conference on Multimedia

(ACMMM). Amsterdam, the Netherlands. October 2016. 701-702.（CCF-A、EI，署名单位：中国科学院信息工程研究所、华中科技大学）

[8] Si Liu, **Xinyu Ou**, Ruihe Qian, Wei Wang, Xiaochun Cao. Makeup like a superstar: Deep Localized Makeup Transfer Network. in: The 25th International Joint Conference on Artificial Intelligence (IJCAI). New York City, USA. July 2016. 1-7.（CCF-A、EI，署名单位：中国科学院信息工程研究所、华中科技大学）

[9] Lingyu Yan, **Xinyu Ou**, Hefei Ling. Local Search Optimized Hashing for Fast Image Copy Detection, in: Annual Summit and Conference of Asia Pacific Signal and Information Processing Association (APSIPA). Angkor, Cambodia. December 2014. 1-10.（EI，署名单位：华中科技大学）

[10] Cong Liu, Hefei Ling, Fuhao Zou, Lingyu Yan, Yunfei Wang, Hui Feng, **Xinyu Ou**. Kernelized Neighborhood Preserving Hashing for Social Network Oriented Digital Fingerprints. in: IEEE Transactions on Information Forensics and Security (TIFS). 2014. 9(12):2232-2247.（JCR-Q1、CCF-A、SCI/EI，署名单位：华中科技大学）

[11] Jin Liu, Hefei Ling, Lingyu Yan and **Xinyu Ou.** Feature Fusion based Hashing for Large Scale Image Copy Detection. in: Intelligent Control and Information Processing (ICICIP), 2014 Fifth International Conference on. Dalian, China. August 2014. 307-312.（EI，署名单位：华中科技大学）

[12] Lingyu Yan, Hefei Ling, Cong Liu, **Xinyu Ou**. Hashing based Feature Aggregating for Fast Image Copy Retrieval. in: Signal and Information Processing (ChinaSIP), 2014 IEEE China Summit & International Conference on. Xi'an, China. June 2014. 441-445.（EI，署名单位：华中科技大学）

[13] Cong Liu, Hefei Ling, Fuhao Zou, Lingyu Yan, **Xinyu Ou**. Efficient digital fingerprints tracing. in: Signal and Information Processing (ChinaSIP), 2014 IEEE China Summit & International Conference on. Xi'an, China. June 2014. 441-445.（EI，署名单位：华中科技大学）

[14] Si Liu, Zhen Wei, Yao Sun, **Xinyu Ou**, Jizhong Han, Ming-Hsuan Yang. Semantics

Preserving Deep Image Retargeting. IEEE Transactions on Image Processing (TIP).（JCR-Q1、CCF-A、SCI/EI，评审中，署名单位：中国科学院信息工程研究所、华中科技大学）

[15] Sujing Wang, Bingjun Li, YongJin Liu, Wenjing Yan, **Xinyu Ou**, Xiaolan Fu. Micro-expression Recognition with Small Samples Size by Transferring Long-term Convolutional Neural Network. IEEE Transactions on Image Processing (TIP).（JCR-Q1、CCF-A、SCI/EI，评审中，署名单位：华中科技大学）

附录 2 攻读博士学位期间参与的科研课题

[1] 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，国家自然科学基金-联合基金重点项目（项目编号：U1536203）

[2] 面向社交网络的数字指纹技术研究，国家自然科学基金（项目编号：61272409）

[3] 人像图片的语义理解方法研究，国家自然科学基金（项目编号：61572493）

[4] 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，湖北省自然科学基金创新项目（项目编号：2015AAA013）

附录 3 攻读博士学位期间所获的奖励

[1] **欧新宇**. 华中科技大学 2017 年优秀博士毕业生

[2] **欧新宇**. 中国科学院信息工程研究所 2016 年所长特别奖. (2015 年 10 月-2017 年 2 月客座期间)

[3] **欧新宇**，魏震，凌贺飞，刘偲，操晓春. IEEE ICME & MSRA 2016 视觉识别挑战赛. 第三名.

[4] 刘偲，**欧新宇**，钱瑞和，王瑋，操晓春. 论文《Makeup like a superstar: Deep Localized Makeup Transfer Network》获信息安全国家重点实验室 2016 年度优秀论文二等奖. 中国科学院信息工程研究所信息安全国家重点实验室. (2015 年 10 月-2017 年 2 月客座期间).

[5] **欧新宇**，刘洋，李深. 2016 年第一届特定音视频分析系统评测资格大赛（特定标示）入围奖. 中国网络空间安全协会.

[6] 柳茂林，**欧新宇**，李叶，余成跃，陆竭. 2016 年第一届全国网络舆情（音视频）分析技术邀请赛（特定视频识别）参赛奖（特定视频识别项目负责人）. 中国科学院信息工程研究所.

[7] 柳茂林，**欧新宇**，李叶，余成跃，陆竭. 2016 年第一届全国网络舆情（音视频）分析技术邀请赛（拷贝检测）参赛奖. 中国科学院信息工程研究所.