密级:_____

中国科学院大学
University of Chinese Academy of Sciences

博士学位论文

基于语义的场景图像特征表示与分类问题研究

作者姓名：魏朋旭

指导教师：焦建彬　教授

中国科学院大学

学位类别：工学博士

学科专业：计算机应用技术

培养单位：电子电气与通信工程学院

2017 年 11 月

Semantic based Feature Representation and Image Classification for Scene Images

by
Pengxu Wei

A dissertation submitted to
The University of Chinese Academy of Sciences
in partial fulfillment of the requirements
for the degree of
Doctor of Computer Application Technology

School of Electronic, Electrical and Communication Engineering

November, 2017

中国科学院大学直属院系
研究生学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行的研究工作及所取得的研究成果。尽我所知，除文中已经标注引用的内容外，本论文中不包含其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的说明或致谢。

作者签名：_____ 日期：_____

中国科学院大学直属院系
学位论文授权使用说明

本人完全了解并同意遵守中国科学院大学有关保留、使用学位论文的规定，即：中国科学院大学有权保留送交学位论文的副本，允许该论文被查阅，可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密的学位论文在解密后使用本声明。

作者签名：_____ 导师签名：_____ 日期：_____

摘　要

场景图像分类是计算机理解场景图像、识别和感知周围世界的重要途径。场景图像分类的主要任务是如何让计算机尽量按照人类认知的方式识别不同场景类别，涉及机器学习、神经心理认知和计算机视觉等多学科交叉技术，对于实现快速有效地组织、管理大规模图像数据具有重大的现实意义。由于场景所包含目标的多样性和不确定性，场景分类面临类内差异性和类间相似性两大挑战性问题。为了解决这两类问题，同时为了避免传统分类方法过多依赖图像分割、目标检测、人工标注语义等计算代价大的处理操作，本文针对语义学习以及基于语义的场景特征表示进行了深入研究，主要工作如下：

1. 提出了一种隐关联语义表示的方法，旨在去掉隐语义学习时的局部图像块之间的独立同分布假设，捕获场景语义关联的自然特性。提出 deep-BoW，避免了 BoW 传统构建方法所具有的性能低、计算复杂度高的问题；为了解决词袋特征一词多义和一义多词的语义模糊问题和捕获场景图像中广泛存在的语义关联问题，采用 logistic 正态先验分布，学习隐关联语义，并将其应用于场景分类；

2. 提出了一种关联主题向量的方法，旨在解决隐语义表示对于分类任务判别力弱的问题。基于 Fisher Kernel 理论，探索生成式模型与判别式模型的结合，提出关联主题向量，提升隐语义的判别能力。为了能让所提出的方法更适合大规模数据集，进一步给出了基于变分贝叶斯求解和吉布斯采样求解的两种关联主题向量实现策略。所提出的方法在大规模数据集上通过实验验证了其有效性，展示出其对 CNN 特征的较大性能提升，对基于深度特征的 Fisher Kernel 表现出巨大的潜力。所提出的关联主题向量与混合高斯系列的 Fisher Vector 一起，为图像语义表示构建了一个更加完备的生成式模型。

3. 提出了一种隐目标发现的方法，旨在自适应地发掘判别性、表示性的图像区域。结合最小熵准则，提出全局到局部、局部到全局的隐目标学习方法，并且结合 Fisher Vector 表示进一步提升模型性能，避免了显式的目标标注或者是依赖大量预训练的目标检测器的庞大计算。另外，所提出的隐目标与 Fisher Vector 特征编码方法融合，得到很好的性能表现。该方法也为进一步探索场景图像语义提供了比较大的潜力。另外，所提出的隐目标发现模型被扩展到另一个计算机视觉问题，弱监督目标检测任务，探究其泛化性。

关键词： 场景图像，图像分类，特征表示，隐语义，隐目标，Fisher Vector

Abstract

Scene classification is the gateway to understand scene images, recognize and perceive the surrounding world. Its primary task is how to make computers recognize scenes in the similar way to the human cognition, and this task involves interdisciplinary technologies, such as machine learning, neuropsychology, computer vision and so on. Scene classification is of great importance for effectively organizing images. Object diversities and uncertainties in scene images invite intra-class variability and inter-class similarity for scene classification. To solve these problems, we propose a series of algorithms semantic learning and semantic based image representation, without costly work of the conventional strategies on conventional strategies image segmentation, object detction and manual annotations. The contributions of this dissertation are as follows:

1. A latent correlated semantic representation is proposed, to remove the independence assumptions for latent semantic learning among local image regions and capture the natural property of semantic correlation for scenes. Deep-BoW is proposed to improve the conventional BoWs without costly computations. The latent correlated semantic representation is learned to capture the correlations among semantics for scenes by introducing the logistic normal prior distribution, and is applied for scene classification to deal with the semantic ambiguity problem.

2. Correlated Topic Vector (CTV) is proposed, to improve the discriminative ability of the latent semantic representation for scene classification. Based on Fisher Kernel theory, the combination of the generative model and discriminative model is exploited and correlated topic vector is derived to improve the discriminative ability of the latent semantic representation. To make the method suitable for the large-scale datasets, we further provide a Variational Bayesian solution and a Gibbs sampling solution. Experiments on large-scale datasets validate the effectiveness of CTV, showing its great improvement over CNN features and great potential to other Fisher Kernel based deep features. Together with Gaussian Mixture Models based Fisher Vector and Latent Dirichlet Allocation based Fisher Vector, CTV constructs a more complete generative model for image semantic representations.

3. A latent object discovery approach is proposed, to adaptively discover discrim-

inative and representative image regions. Following the min-entropy principle, latent objects are learned in a global to local and local to global manner and Fisher Vector feature encoding strategy is developed to further improve the model. The proposed approach can adaptively exploit discriminative and representative image regions without any explicit object annotation or predefined object detector. It not only unifies the explicit objects and cluster-derived regions, but also works well with a state-of-the-art feature encoding method. Experimental results validate its effectiveness to model the uncertainty of complex scene images, and its great potential to process complex spatial distributions of scene objects. Furthermore, the proposed model is extended to another computer vision task, i.e., weakly supervised object detection, exploiting its generalization.

Keywords: Scene images, Image classification, Latent semantic, Latent object, Fisher vector

目 录

图形列表

表格列表

第一章　绪论

1.1　引言

“视觉，是从图像中发现图像呈现了现实世界的什么和是哪里的过程”[1]。

这是著名神经学家和心理学家 David Marr 在 20 世纪提出的关于视觉的定义。场景理解是人类执行许多有价值行为与任务的基石，是计算机从人类场景感知的角度来感知现实世界的过程 [2]。场景理解最基本的任务是将一个自然图像分类到众多语义类别之一，即场景图像分类任务。场景图像分类任务，主要是回答图像呈现的是哪里的问题。

1.2　研究背景与意义

1.2.1　研究背景

近年来，随着电子摄像设备的普及、现代通信的发展和计算机技术的进步，数字图像作为现代信息传播的重要媒体，每天都在大规模的产生并发布。如 Google、百度等搜索引擎网站的图像库均已超过数十亿，Flickr、Facebook、Instagram、微博等网络社交服务平台每天接受着来自世界各地的人们上传的千百万幅图像。另外，视频监控、机器人、无人驾驶汽车、医学成像等应用领域也在不断产生多种多样的图像。如此海量的图像资源在丰富人们生活的同时，也令人很容易迷失在浩如烟海的数据之中而无法快速有效地获取有用信息。传统的依靠人工标注的图像分类管理方式，由于存在标注代价大和主观标注带来的片面标注问题，无法应对急速增长更新的庞大图像数据库。因此，如何利用计算机自动将图像尽量按照人类认知的方式进行不同类别语义的分类，从而快速有效地组织、管理大规模图像数据具有重大的现实意义。

图像分类是借助于计算机技术来寻找合适的图像表示方式、为图像确定所属语义类别的过程。图像的语义类别可以分为两种：第一是目标类别，即图像中“有什么”；第二是场景类别，即图像描述的是“什么地方”。图1.1中的两幅图包含了建筑、车、树、行人等目标，描述的是一个“street”的场景。本研究课题关注于第二类语义。

在计算机视觉领域，众多文献给出了关于场景的定义：

- 在高层场景感知的研究中，场景被定义为语义关联的、人类从不同视角观察到的现实世界环境，该环境中所包含的背景元素和多个离散目标以一种空间秩序排布 [3]。

图 1.1 “street”场景图像中目标标注

图 1.2 场景图像样例与场景概念云

- 场景是在一定程度上几个目标不可预测的叠放 [4]。
- 场景是一定视角下的环境呈现：所包含的目标和显露物体部分以一种有意义的方式组织在一起 [5]。
- 场景是指人类可以在其中活动的地方，或者人类可以前往的地方 [2, 6]。

尽管这些场景定义的说法不同，但是从由这些定义得到的场景概念云（图1.2）发现，他们共同概括出场景表达的是一个地方，并且包含了具有一定空间排布的实体（目标）。场景图像分类是对视觉基础层面的感知分类，它基于对场景的视觉表示，利用分类器将一幅图像分类到一个场景类别，从而为图像赋予一个语义类别/标号，如教室、办公室、高速公路、森林、羽毛球场、咖啡馆、教堂等等。这些场景语义类别，即场景标号，从人类的角度来讲，能够捕捉环境的丰富性和多样性；从机器的角度来讲，具有高度抽象性和复杂性，这就为机器识别场景带来巨大的挑战。

图像可解析为场景（scene）– 目标（object）– 部件（part）/区域（region）

图 1.3 场景理解

– 子部件（subpart）/子区域（subregion）– 边缘（edge）/纹理（texture）– 像素（pixel），形成了自上而下、自下而上的层次结构，如图1.3所示。由于场景图像中目标外观、视角的差异性以及目标分布的不确定性，层次化、复杂化的场景结构，造成了场景图像较大的类内差异性和类间相似性问题。因此，我们不能简单地构建场景图像的特征表示。不过，该场景结构从本质上揭示了场景图像丰富的语义信息。场景语义信息，是对场景自然结构的理解，刻画了场景的实体个体、实体与实体之间的语义关系和空间关系，为场景分类任务提供了重要信息。因此，基于语义的场景图像特征表示，成为场景分类的核心研究问题。对于场景分类问题，从高层语义 – 场景 – 对图像自然结构、上下文信息的高度抽象和概括，到低层语义 – 边缘/纹理、像素 – 图像原始特征的描述，两者之间存在着巨大的“语义鸿沟”（semantic gap）[7, 8]，极大地限制着场景分类性能。

1.2.2 研究意义

基于语义的场景图像特征表示与分类问题的研究，主要是针对场景图像所在的“语义鸿沟”问题，在无需显式的图像分割、人工标注或者大量的目标检测的条件下，挖掘场景语义信息，寻求表示力强的语义。更进一步地，基于学习到的语义，结合人类场景感知方式，构建判别力强的特征表示，增强场景分类模型解决场景图像类内差异性、类间相似性的能力，进而提升场景分类的准确度。场景分类技术的提升，直接或间接地影响着场景理解相关任务的进步，同时对于计算机视觉问题的深入研究以及视觉研究技术在人们生活中的应用都有着重要的意义。场景分类技术的研究意义主要体现在以下几个方面：

1. 目标检测和识别。心理物理学家 Torralba 2009 年发现：如果理解了一个场景的整体概貌，那么场景中的目标可以精确地识别出来，即使是在很低的 6*6 图像分辨率下 [9]。而当目标脱离了它所在的上下文时，目标识别性能就

会随之变差。所以场景图像分类能够为图像目标检测、定位、识别等更高层次的图像理解提供有效的上下文语义线索，推动相关视觉研究的发展。

2. 图像检索。当用户搜索图像时，场景分类是将大量的图像进行抽象语义分类，然后在同一场景类别的图像中，帮助用户更快捷地寻找包含某个目标的图像，或者满足一定语义条件的图像。基于语义或者内容图像分析和检索，可借助于场景图像分类帮助人们对图像的快速浏览、检索和理解，更高效地组织、管理海量的图像数据。

3. 视频检索。对于海量的监控视频来说，当检索或者跟踪满足出现在某个场所的目标时，人工逐帧搜索视频耗费时间长，代价大，而场景分类将会大大提高检索监控视频的效率。

4. 机器人导航。场景图像分类能够为机器人感知其当前所处的环境，为其提供重要的先验信息，快速感知周围环境中的实体目标，辅助路径规划，并且针对不同的机器人任务，指导机器人对环境的交互行为，这对机器人导航等研究与应用有着重要的意义。

5. 遥感图像应用。随着卫星、航空等技术的进步，获取到的遥感图像由于分辨率的提高而包含了越来越丰富的信息量，对其进行场景的分类可快速理解遥感图像的环境信息，为遥感图像的深入研究提供指导。

6. 无人驾驶汽车应用。近年来，无人驾驶汽车从概念已逐渐变为现实。场景图像分类能够快速识别汽车复杂多变的周围环境，为汽车速度和路径的智能化调控提供信息。

1.3 研究动机与研究内容

1.3.1 难点问题分析

场景图像表示与分类问题涉及计算机视觉、机器学习和神经心理认知科学等多学科交叉技术，如图1.4所示。MIT 从 2006 年至 2017 年组织了多届场景理解研讨会，来自神经生理学、认知神经学、视觉认知和计算机视觉等领域的顶级学者明确了场景分类的研究问题和其重要性。随着图像内容的复杂化，场景图像表示与分类问题已成为计算机视觉领域中的热门研究方向。

与目标图像分类相比，场景图像分类面临更大的挑战。从数据集图像中就可以很直观地观察到两类分类问题的差异。图1.5展示了分别来自两个数据集的图像：以目标为中心（object-centric）的目标数据集 PASCAL VOC 2007[10]、以场景为中心（scene-centric）的场景数据集 SUN 397[6]。PASCAL VOC 2007 数据集

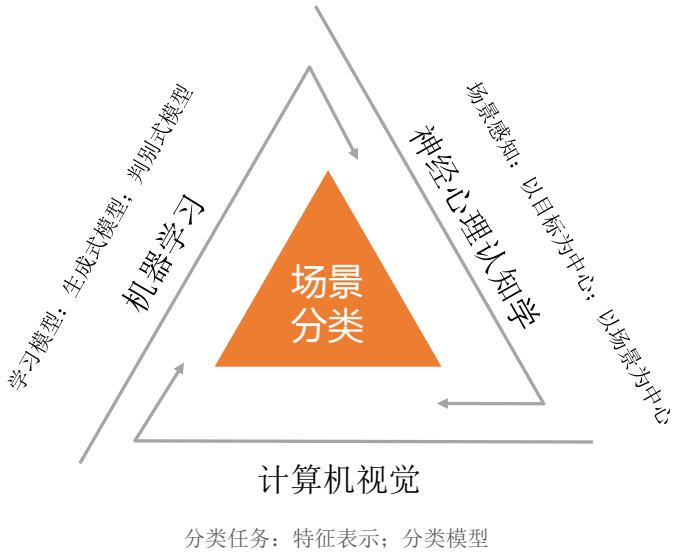


图 1.4 场景图像分类主要涉及的研究领域

图 1.5 在目标数据集与场景数据集中包含目标“car”的图像对比

表 1.1 场景数据集与目标数据集统计信息对比

类型	场景数据集	目标数据集	
数据集	SUN 397	PASCAL	ImageNet
每一幅图像中实例平均数目	16.8	1.69	1.59
每一幅图像中目标类别平均数目	9.46	1.5	1.5
每一幅图像中目标尺度平均大小	0.0863	0.241	0.358

中的图像包含的目标大多比较显著，比如，目标“car”总是位于图像中心、一幅图像中目标“car”数量通常只有一个实例。但是在场景图像中，所包含的目标在数量、尺度、遮挡、背景、光线、位置、视角等方面存在比较大的不确定性，比如，目标“car”在多个场景中出现时，表现出了较强的随机性，造成场景较大的

图 1.6 场景图像的类内差异性

变化性。表1.1给出了两类数据集的统计信息对比结果 [6]。从对比结果来看，场景数据集中出现的目标实例数目多、目标种类多样、目标尺度小，这为场景图像分类问题带来了比较大的难题。

场景图像分类面临着两大难题。（1）存在较大的类内差异性。由于自然界场景的多样化特性，即使是同一类别的场景图像，总是呈现出不同的形式，这样就导致场景图像类内的较大差异性。如图1.6所示，夏季无白雪覆盖的山脉与冬季有白雪覆盖的山脉呈现出很大的不同，不同街道场景由于视角、街道环境等的差异也呈现出很大的不同。（2）类间的相似性。同一类目标可能出现在多个场景中，造成了不同类别之间的场景图像存在相似之处，比如“天空”会存在于大部分自然场景图像中。另外复杂、多变的光照影响（不同的光照环境条件下获取的图像存在较大差异）、尺度影响（拍摄镜头的远近带来的差异性）、场景类别语义的主观性的不确定性（对同一幅场景图像，不同的观察者可能有不同的理解等），进一步增加了场景图像分类的难度。

1.3.2 研究动机

心理认知学家对于人类场景感知方式主要由两种观点：（1）以目标为中心的方式：场景可以由目标集合表示 [11]，所以我们可以通过识别场景中的目标来用于用于场景表示和分类；（2）以场景为中心的方式：人类大脑可以快速识别出场景图像 [12–14]，所以我们无需识别出目标，可以直接利用场景信息来识别场景图像。

依据心理认知学家的这两种观点，现有的场景分类技术主要分为两大类：以目标为中心的特征表示和以场景为中心的特征表示。两类研究的共同点是，大部

分相关研究工作 [15–20] 以尝试解决低层特征带来的“语义鸿沟”问题为出发点，探索表示力、判别力强的场景语义以及相关语义特征表示方法，提升分类模型区分场景图像类内差异性和类间相似性的能力。但是，两类研究采用的策略存在很大的不同，为场景分类任务带来不同的问题。

- 以目标为中心的特征表示方法

  该类表示方法广泛采用穷举场景中出现的目标的策略：（1）预定义可能出现在场景中的目标种类，（2）预训练目标检测器，（3）全图遍历检测目标，（4）用检测到的目标表示场景图像。该类特征表示方法极大地依赖于预先定义、预先训练的目标检测器，而这些目标检测器涵盖的目标种类非常有限，而且目标检测器并不总是能够鲁棒地解决场景图像中所存在的目标遮挡、尺度、光线、背景等检测难题，获得可靠的目标检测结果。另外。穷举式的策略导致其计算复杂高。

- 以场景为中心的特征表示方法

  该类表示方法采用的策略是从全图学习隐语义信息，形成对场景图像的一种全局描述。该类方法避免了第一类方法粗暴的表示方式和庞大的检测计算量，但是其判别力与表示力较大程度受限于场景类内差异性和类间相似性。

总的来讲，对于场景分类研究，无论是以目标为中心的特征表示还是以场景为中心的特征表示，它们研究的本质问题都是探索场景的语义信息，这也正是场景图像特征表示研究的核心问题。狭义地讲，语义可以是场景中出现的实体目标，广义地讲，语义可以是场景中具有一定上下文信息（context）、紧致（compact）、连续的视觉元素。因此，在本文中，我们可以把场景图像分类问题概括为两个关键问题：

(1) 如何刻画语义（semantic，也可称为主题 topic/theme）？

(2) 基于语义如何获得优异的场景特征表示？

1.4 研究内容与主要贡献

针对于场景分类任务的挑战性问题，本文研究内容以语义为中心展开，结合以场景为中心和以目标为中心的人类视觉感知方式，探究两个关键性问题：（1）如何刻画场景语义、（2）基于语义如何获得优异的场景特征表示，提出表示力、判别力强的特征表示并用于场景分类任务。如图1.7所示，本文主要从隐关联语义表示、关联主题向量和隐目标挖掘三个部分来探究两个关键性问题。表1.2给出了各个部分与两个关键性问题的关系。本文的研究内容和主要贡献概括为以下几点：

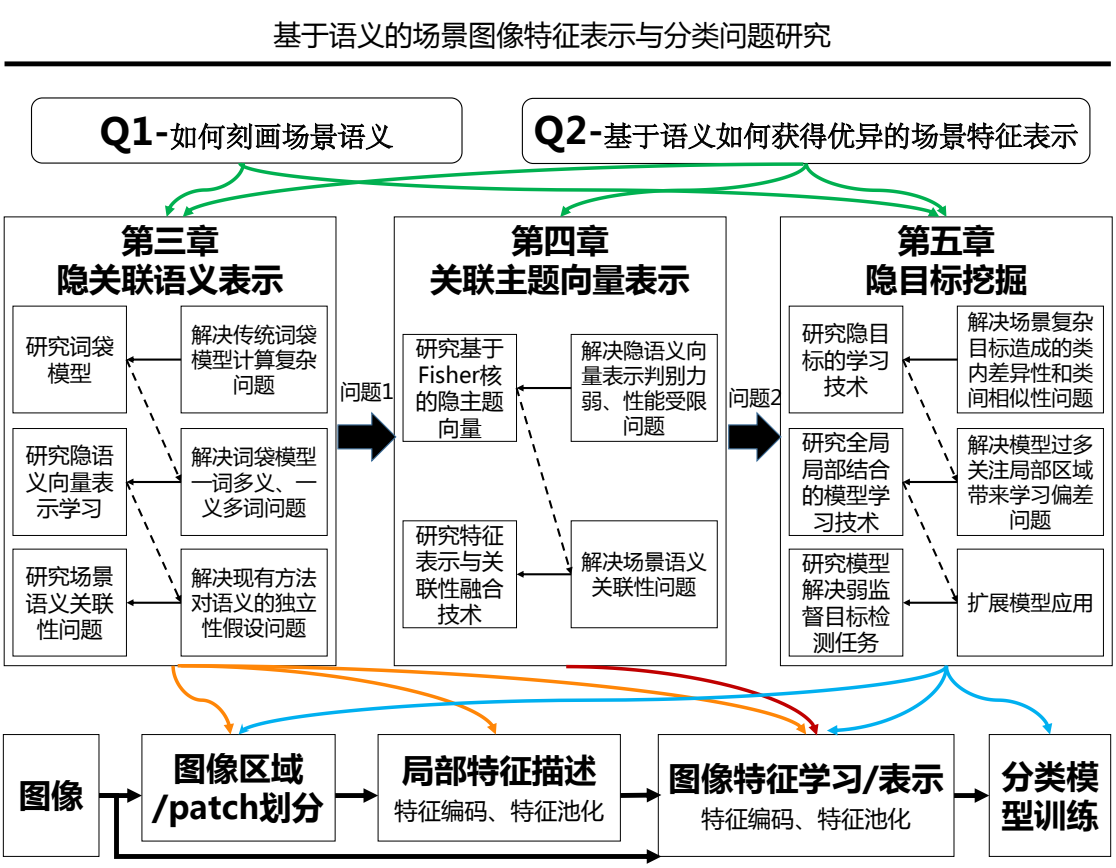


图 1.7 本文研究思路和研究内容及与场景分类框架对应关系

表 1.2 特征表示与语义

特征表示	Q1-如何刻画场景语义	Q2-基于语义如何获得优异的场景特征表示
BoW	K-means 聚类，视觉字典	硬性量化，直方图
Fisher Vector(GMM)	混合高斯概率	Fisher Kernel
CNN	神经元激活	卷积、池化
隐主题向量	隐主题/语义	隐主题向量
关联主题向量	隐关联主题/语义	关联主题向量
隐目标发现	隐目标	隐目标学习

- 基于隐关联语义表示的场景分类问题研究

  为了解决传统词袋（Bag of Words，BoW）特征在场景图像特征表示时的计算复杂度高的问题，本文研究结合了 CNN（Convolutional Neural Network，CNN）的语义特性，构建更高效的词袋特征，即 deep-BoW。deep-BoW 打

破了传统的基于 SIFT（Scale-Invariant Feature Transform，SIFT）、CNN 等局部描述子的 BoW 构建方式，比较好地利用了 CNN 的语义特性（优异的语义泛化性和强大的语义聚类效应）。另外，与传统的基于聚类构建 BoW 方式不同的是，deep-BoW 避免了那些聚类、量化等计算代价大的算法，保留了更多的信息。

进一步，为了解决 BoW 特征的一词多义、一义多词问题 [21]，本文研究利用主题模型学习场景图像的隐语义表示。隐语义表示的研究，避免了人工标注、目标检测、图像分割等额外的、计算量大的处理。但是，隐狄利克雷主题模型（Latent Dirichlet Allocation，LDA）在学习隐语义时存在一个比较严格的假设，即语义/主题之间是相互独立的。该独立性假设严重地违背了现实场景图像中所存在的语义关联特性，比如桌子、椅子、电脑总是一起出现在办公室场景中，天空、树木、道路总是一起出现在高速场景中。为了能够更接近真实的学习语义，我们去除掉独立性假设，学习隐关联语义并将其应用于场景分类任务中。主要贡献是：打破传统 BoW 构建方式，提出 deep-BoW，替代传统基于 SIFT 的 BoW；学习隐关联语义特征表示，捕获场景图像中存在的语义关联/共现现象，无需人工标注、目标检测、图像分割等复杂的辅助处理。

- 基于关联主题向量的场景分类问题研究

  为了解决基于主题语义的隐语义判别力弱的问题，我们探索从生成式模型（Generative Model）与判别式模型（Discriminative Model）相结合的角度，提升主题模型学习到的隐语义的判别力。主题概率分布表示虽一定程度上解决了场景图像类内差异性问题，但是聚类作用反而模糊了场景图像类间的区分性，用主题概率分布作为场景表示过于单一；对于场景图像分类任务，基于隐语义/主题特征表示的性能受到较大限制，并不能一直随着主题数目、字典数目的增加而增加。基于 Fisher Kernel 理论 [22]，所提出的隐主题向量和关联主题向量，在用中层关联主题语义编码弥补低层特征的表示能力的同时，尝试将生成式模型与判别式模型相结合，提出具有较强判别性和表示性的特征表示。主要贡献是：从模型角度来看，我们去掉狄利克雷先验分布对于主题的独立性假设，借助于 logistic 正态先验分布来进一步刻画主题之间的关联结构；从 Fisher Kernel 角度来看，特征向量的构建考虑了语义之间的相关性。在关联主题模型中，先验分布与后验分布的非共轭特性，使得对数似然无法解析计算，造成基于 Fisher Kernel 的关联主题向量推导的难度。为了解决该问题，我们分别完成了基于变分贝叶斯（Variational Bayesian，VB）求解和基于吉布斯采样（Gibbs Sampling, GS）求解的两种关联主题向量实现方法。

- 基于隐目标的场景分类问题研究

  场景图像包含了层次化语义，如场景类别到场景目标再到局部块。为了避免任何目标检测、人工标注等需要额外工作的方法，我们提出隐目标方法，学习场景图像中显著性的、占主导地位的视觉语义元素。为了避免模型由于过多关注局部区域所带来的学习偏差问题，我们提出了全局到局部、局部到全局的隐目标挖掘策略。该策略的合理性在于遵循了人类场景感知的自然特性[23]，从全局到局部来看：一个场景可能被分解为“目标”，激励发掘场景图像中那些表示力强的“目标”或者局部区域；从局部到全局来看：“目标”是被结构化地组织为一个场景，激励那些被发掘的“目标”更具有判别性地解释一个场景图像。隐目标挖掘主要解决现有方法的以下三个问题：模型依赖大量预先定义实体目标类别的、预先训练的目标检测器，面临庞大的检测计算量和复杂的目标检测挑战；各个目标的检测是相互独立的，即使是好的检测结果，也无法保证能够表达出丰富的场景语义；分类模型过多地关注于场景局部区域。我们主要的贡献在于：提出隐目标的概念，采用全局与局部相结合的学习方式。另外，所提出的隐目标挖掘模型被扩展应用于弱监督目标检测任务中，进一步探索模型的有效性。

1.5 本文的组织结构

第一章，绪论。论述场景图像分类的研究背景和意义，描述场景图像分类在目标检测和识别、图像检索、视频检索、机器人导航、遥感图像应用和旅游导航等方面的应用。分析了场景图像分类的难点问题，明确了本文的主要研究内容和贡献。

第二章，场景图像分类技术的发展与现状。概述场景图像分类的主要方法，并且针对于低层、中层和高层特征表示等方面进行了全面详细的分析。

第三章，基于隐关联语义表示的场景分类。打破传统 BoW 构建方式，并结合 CNN 泛化性强、聚类效应紧致的语义特性，提出 deeo-BoW 构建方法。针对 BoW 的一词多义、一义多词的语义模糊问题，基于主题模型学习隐语义表示。考虑到场景图像中所存在的语义关联特性，采用 logistic 正态先验分布（logistic normal prior distribution），去除掉局部图像块之间的独立同分布假设，更好地建模场景图像并用于特征学习。大量的实验结果验证了我们所提出的 Deep BoW 和隐关联语义的有效性。

第四章，基于关联主题向量的场景分类。针对隐语义判别力弱的问题，以 Fisher Kernel 为理论依据，提出结合生成式模型和判别式模型的特征表示构建方法，提升特征的表示力和判别力。所提出的关联主题向量，旨在利用主题之间的关联性，并将其编码于 Fisher Vector 框架中以此来提升表示的判别能力。为了能

让所提出的方法更适合大规模数据集，进一步给出了变分贝叶斯求解和吉布斯采样求解的关联主题向量实现策略。另外，在大规模数据集上的实验也验证了关联主题向量的有效性，并展示出其对 CNN 特征的较大性能提升，对基于深度特征的 Fisher Kernel 表现出巨大的潜力。与 GMM 系列的 Fisher Vector 和 LDA 系列的 Fisher Vector 一起，所提出的关联主题向量为图像语义表示构建了一个更加完备的生成式模型。

第五章，基于隐目标挖掘的场景分类。为了自适应地发掘具有判别性和表示性的图像区域、避免显式的目标标注或者大量预定义的显式目标检测器，提出了隐目标发现的方法。所提出的方法，结合最小化熵准则和 Fisher Vector 表示来提升隐目标的发掘。所学习到的隐目标不仅统一了显式的目标和聚集的区域，同时与 Fisher Vector 特征编码方法融合，得到不错的性能表现。该方法也为进一步探索场景图像语义提供了比较大的潜力。通过实验也验证了所提出方法的优异性能表现。另外，为了能够进一步扩展隐目标挖掘模型的应用，利用所提出的模型解决弱监督目标检测任务，同时通过实验验证了所提出模型的有效性和泛化性。

最后，第六章总结本文的主要工作，提出对未来工作方向的展望。

第二章　场景图像分类技术的发展与现状

场景图像分类已成为计算机视觉领域中的热门研究方向。如何寻求对场景图像的合适表示以及如何依赖机器学习算法完成分类任务，是场景图像分类的主要研究内容。场景图像存在较大类内差异性和类间相似性，同时还受气候、光照、尺度、拍摄角度等影响，这为场景图像分类任务带来了极大的挑战。近年来，场景分类研究取得了显著的进步，涌现出大量的研究工作。场景分类问题的广泛探究，推动着视觉场景理解的深入研究和相关计算机视觉任务的发展，如目标识别[24, 25]，图像检索 [26–28]，和智能机器人导航 [29, 30]。

2.1　引言

场景类别是人们理解的抽象语义概念，是赋予场景图像的高层语义信息，场景图像分类的过程就是寻找低层特征到该抽象高层语义概念的映射关系。从计算机视觉的角度来看，场景分类的解决策略主要依赖于特征提取与机器学习方法。总的来讲，一个完整的场景图像分类框架主要包括三部分：特征提取、分类器和性能评测，如图2.1所示。常用的场景分类的性能评测是采用平均准确度（mean Average Precision, mAP）；常用的分类器有支持向量机（Support Vector Machine，SVM），决策树（Decision Tree），K 近邻（K-Nearest Neighbors，KNN）等。对于特征提取，国内外大量的研究工作专注于特征表示的提取，以期获得具有优异表示力、强大判别力的特征表示。随着计算机视觉的发展，图像特征表示的维度也在与日俱增，如图2.2所示。从早些年只是简单地采用颜色、纹理、边缘等低层特征，到基于语义的中层特征表示，场景图像特征表示主要还是依赖于手工设计的方式来构建。但是，手工设计特征表示的主观性偏差、任务性偏差限制了其表现性能和广泛应用。为了解决该问题，基于语义的特征学习方法涌现出来，尤其是卷积神经网络的兴起，极大地推动着场景图像特征表示的发展。从语义表示方面出发，现有场景图像特征表示可以分为低层、中层、高层三个层次（见图2.3）。

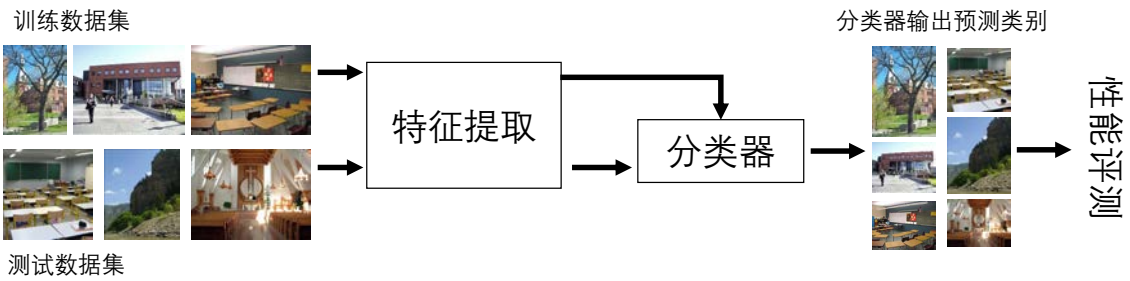


图 2.1 场景图像分类基本流程图

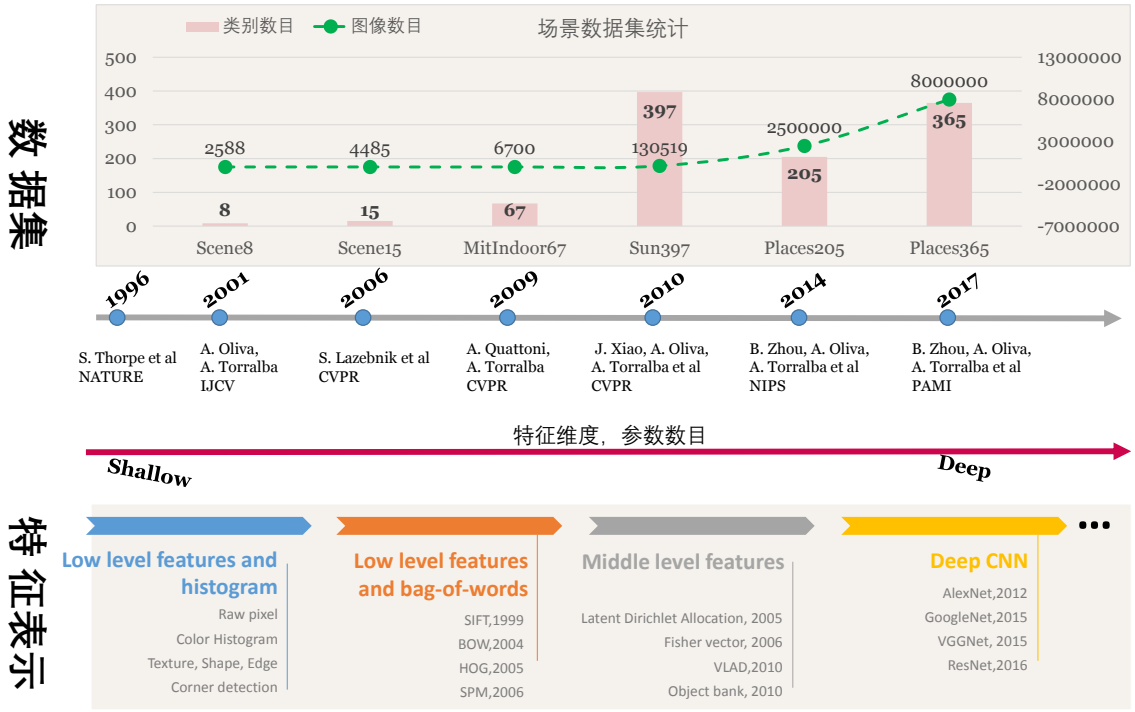


图 2.2 场景图像数据集统计与场景图像特征表示概述

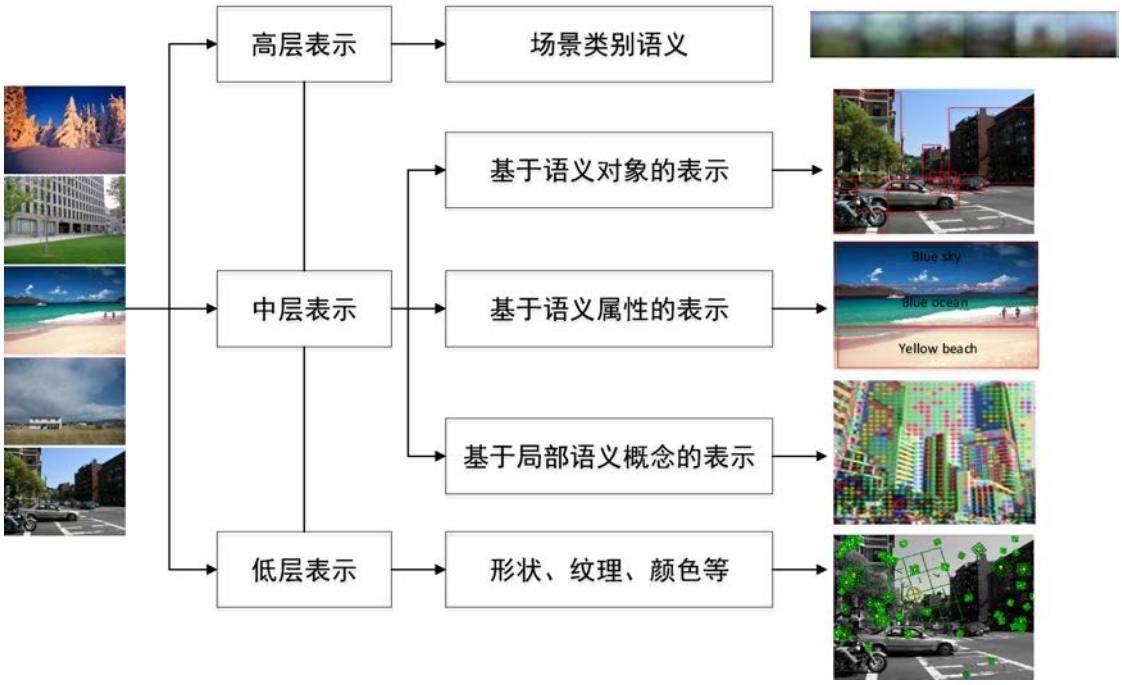


图 2.3 场景图像分类方法

2.2 场景图像的低层特征表示

早期的场景图像分类可简单看做是基于低层特征表示的二分类问题：室内（indoor）与室外（outdoor）、乡村（country）与城市（city）等等。低层特征表示一般通过直接在图像像素数据上计算得到，能够客观、直接地反映图像内容，例如颜色、纹理和形状等特征。Szummer 等 [31] 最早采用低层特征多阶段分类实现场景图像分类。在第一阶段，首先均匀地对图像进行分块，然后假设各子块相互独立，分别在各个子块中提取 MSAR 纹理特征和 Ohta 空间颜色直方图特征，利用 KNN 分类器将各字块识别为“indoor”或者“outdoor”；第二阶段，连接第一阶段各字块纹理和颜色分类结果，形成新的特征向量，借助于另一个分类器得到最终场景分类结果“indoor”或者“outdoor”。Vailaya 等 [32] 则是提取图像的特征，借助于二元贝叶斯分类器，首先将图像分类为“indoor”/“outdoor”/“other”三种场景类别，然后利用 SVM 进一步将“outdoor”场景细分为“landscape”/ “city”/“other”，再进一步将“landscape”细分为“sunset”/“forest”/“mountain”/“other”。2004 年，Lowe[33] 提出尺度不变特征变换特征 SIFT，SIFT 特征对旋转、缩放和亮度变化等均保持不变性，对视角变化、仿射变换和噪声也具有较强的稳定性，近年来它在场景分类、图像检索等计算机视觉领域的应用十分广泛。随后，ColorSIFT[34]、HOG[35]、LBP[36]、SSIM[37]、texton[38]、texton forests[39]、OTC[40] 等众多描述子被提出且用于场景分类中。

2.3 场景图像的中层特征表示

低层特征空间中的相似性不能总是直接地反映高层语义的相似性，这就导致两者之间存在很大的语义鸿沟 [7]。为了弥补低层特征与高层语义概念之间的语义鸿沟，众多的研究工作致力于建立图像的中层语义表示。与此同时，场景分类研究问题的重点由二分类问题逐渐转变为场景类别是多种具体场景类别，如海岸、街道、足球场、剧院、电影院等。中层特征表示分类通过机器学习技术建立视觉特征与语义概念之间的关系，具体可分为三种方法：基于语义对象的表示方法、基于语义属性的表示方法和基于主题模型的表示方法。

2.3.1 基于语义对象的表示方法

基于语义对象的表示方法来源于以目标为中心的场景认知（场景是由有限数目且依一定层次结构出现的目标组合而成），所以可以用出现在场景的目标分布及其相互之间的关系来表示场景。一般来说，人们借助于分割、检测或识别算法获得图像中的目标对象 (如海岸、高山和汽车等) 或者区域。常用的分割算法是

Felzenszwalb 基于图的分割算法 [41]、Jianbo Shi 基于 Normalized Cut 的图像分割 [42]，Sande 的 Selective Search[43]，常用的目标检测算法有 DPM (Deformable Part Models)[44]。Josef 等 [45] 先对图像进行区域分割，再利用分类器对分割后的区域进行目标识别，最后在局部信息的基础上进行场景的分类。Fan 等 [46] 通过标注一系列有意义的区域来训练检测器，再利用检测器寻找测试图像中的目标区域，最后利用最大后验概率识别图像的场景类别。Fredembach 等 [47] 在对图像分割之后，利用主成分分析方法获得特征区域（Eigenregion）再并分别对各局部区域进行分类，最后根据图像区域的局部分类结果完成对整幅图像的场景分类。

从本质上讲，基于语义对象的表示方法是将场景分类问题转化为了图像分割和目标识别问题。目前，图像分割算法的性能并不理想，如 Selective Search 分割算法在获得到完整目标区域的同时也产生了大量的语义对象割裂或者语义对象重复的区域；而目标识别本身仍是一个尚待解决的问题，并且目标检测需要训练众多的目标检测器，这样的话，就需要手工标记训练集图像中的目标，将导致耗费时间长，花费大。因此，基于语义对象的表示方法受到图像分割和目标识别性能的限制。

2.3.2 基于语义属性的表示方法

基于语义属性的表示方法通过人为指导将图像场景按照事先定义的语义属性来描述，从而借助于这些感官属性概念来解释图像的语义。Aude Oliva 与 Antonio Torralba 提出的 gist 特征是对场景的整体描述，规避对图像的分割或对图像目标、区域的处理，由图像的自然度、开阔度、粗糙度、方向度、崎岖度等感知属性来表示场景的空间结构 [48]。Torralba 等 [49] 将图像分为一个个小区域，再人工赋予每个小区域一个语义属性，通过统计多个图像区域的语义属性来表示图像的语义内容。Shuo Wang 等人 [50] 采用弱监督的方式，在借助于分层空间解析表示算法学习场景结构的同时，关联与或树节点和图像中出现的场景属性，训练场景中目标与属性对的模型，同时实现图像属性定位。

基于语义属性的表示方法很大程度上依赖于主观设定的图像属性特征，且需要人工标注图像中的语义属性，因此其泛化性能差，人工标注方式的代价太大，并且标注存在标注者个人偏好与理解，可能会导致标注结果存在一定的偏差。

2.3.3 基于主题模型的表示方法

2003 年，Josef Sivic 与 Andrew Zisserman[51] 借鉴文本处理领域中的 BoW(Bag Of Word) 模型，首次提出了视觉词典模型的概念，构建表示视觉图像的 BoW 特征，随后 BoW 以其显著的性能被广泛的应用于图像分类。但是传统的 BoW 特征是将低层特征向量强硬地量化到一个视觉词 word 上，不同图像的 BoW 表示

图 2.4 “coast”场景图像以及人工标注结果

的相似度判断是建立在各图像之间的 word 重复次数的基础上，因此，BoW 只是关注视觉词出现频率，所以 BoW 在丢掉了图像的空间层次信息的同时忽视属于语义层的相关，这就造成了特征表示的一词多义和一义多词的语义模糊问题 [52]。2006 年，Svetlana Lazebnik 等人提出 SPM (Spatial Pyramid Matching)[53]，为 BoW 编码空间结构信息。为了给 BoW 构建更紧致的视觉词典，Aymen Shabou 等 [54] 提出局部约束的空间正则编码 BoW。

为了消除 BoW 带来的 word 模糊问题，主题模型借助于模型隐变量挖掘场景图像中层语义，已成为目前场景图像分类的热门研究方法。主题模型最早出现在文本分析与处理领域，近年来在计算机视觉领域也得到了广泛的应用。最具代表性的语义主题模型是概率潜在语义分析 pLSA (probabilistic Latent Semantic Analysis) 模型 [55] 和潜在狄利克雷分布 LDA 模型 [56]。基于主题模型的场景图像表示与分类方法是基于两种人们对场景图像的感知。其一是场景可认为是多种目标按照一定布局组成，依据目标及其目标之间的关系实现对场景识别；其二，神经学家和视觉研究学家发现：人类可以快速并且有效地分类自然场景图像 [12–14]，不管类别是由场景中一个目标所定义还是由整个场景的特性所定义。

对场景图像分类任务来说，第一种场景感知方式一般需要借助于图像分割、目标检测算法分割出区域或检测到目标，建立场景图像到区域/目标，再到图像 patch 或者 word 之间的分层依赖关系，以及目标区域与场景类别之间的关联关系，如 Josef Sivic 等人 [45] 利用 pLSA 挖掘图像中的目标并定位，Erik B. Sudderth 等人 [57] 利用 LDA 模型建立“scene”到“object”或者“region”再到“patch”的关系。第二种场景感知方式无需识别出场景中的所有目标，只需将图像划分成规则的子区域，均匀捕捉图像局部语义信息，从整幅图像全局的角度人们就可以识别出场景类别。受此启发，许多学者尝试避开图像分割和目标识别过程，直接从图像中学习场景内容表示，比如 Feifei Li 等人 [17] 借助于 LDA 模型直接从视觉数据中学习场景的中层表示，避免了人工标注图像中的目标或者语义

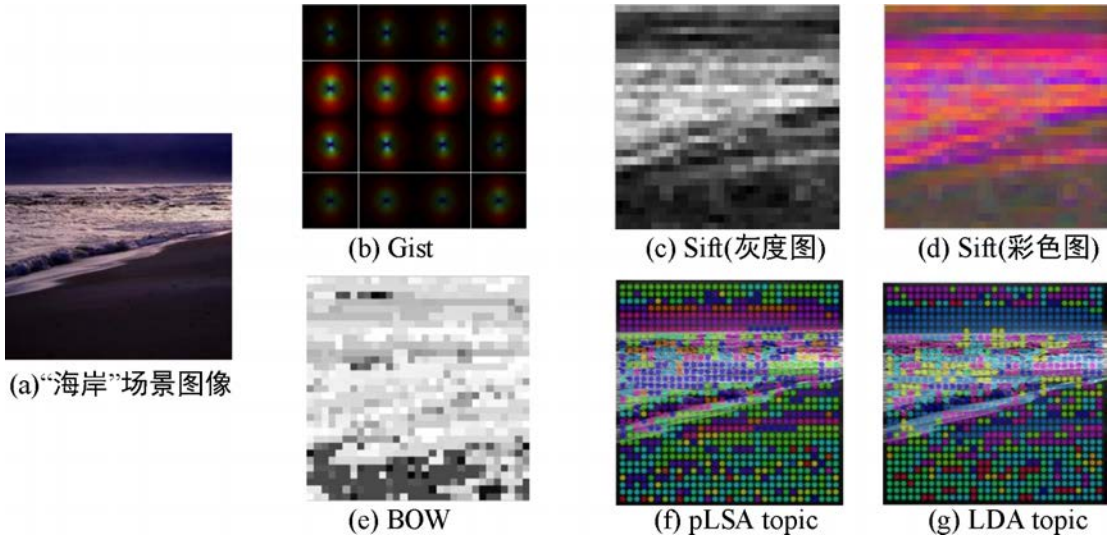


图 2.5 场景图像特征对比

区域；Anna Bosch[18, 19] 首先将 pLSA 生成模型学习到的一幅图像 topic 概率分布作为特征表示，再利用 SVM、KNN 判别模型分类场景图像。

无论是 pLSA 还是 LDA，主题模型的隐变量 topic 都是以非监督的方式进行挖掘，在模型训练阶段忽略掉图像目标的类别信息，所以描述力强大的主题模型在用于分类任务时受到限制，一般后续识别任务还需要判别分类器来实现。另一个不容忽视的问题是，主题模型处理的是已丢掉图像空间结构信息和上下文信息的 BoW 特征表示数据，所以 pLSA 和 LDA 模型的学习也是无图像空间结构信息和上下文信息。为了提高主题模型的判别力，一般研究者在主题模型中加入图像类别信息，希望学习到具有判别力的主题模型特征表示，如 Wang 借助于 sLDA[58] 建模场景类别与图像标注的关系，softmax 函数不仅将原本连续的标号转换为离散的类别输出，而且分类出场景图像所属的场景类别与图像标注；Mandar Fixit 等人 [59] 在 LDA 模型中的 topic 上面加入图像类别的监督指导作用，增强 LDA 的判别力。其他类似工作还有 cLDA[17]，Discriminative LDA[60]，MedLDA[61, 62]，cssLDA[63] 等。另一方面，在主题模型中加入空间信息，能够在实现分类性能提升的同时达到目标或者属性定位的目的，因此 Liangliang Cao[64] 提出空间相关主题模型，建立 patch 对应的 word 字典和图像子区域外观 word 字典，过分割图像，并迫使在同一个同构的区域中，像素点之间是共享相同的 topic；Zhenxing Niu[65] 等人认为一场景由位于不同空间位置的场景视觉元素组成，在 DiscLDA 模型中加入 patch 位置的可观测变量，用隐变量 topic 表示场景中的视觉元素，每一个图像 patch 对应的隐变量的值代表的是该 patch 来自于场景中哪一个视觉元素。

基于主题模型的表示分类方法在图像场景分类应用中取得了较好的分类结果，已经成为当前图像场景分类中图像表示分类的主流方法。然而对于图像数据，相

邻图像块之间存在着非常强的上下文语义共生关系，这被现有的大部分主题模型忽略。因此，如何有效利用图像区域之间的上下文语义信息，从而提高语义主题模型的分类能力是基于语义主题模型的图像场景分类算法需要进一步研究的问题。

2.4 场景图像的高层特征表示

Li Lijia[66, 67] 认为图像可以用出现在图像中的目标来表示，通过检测图像中的目标，构建 Object Bank 作为对场景图像的高层特征表示。在多个数据集上的结果表明一个带有语义意义的图像表示能够有助于减少高层视觉识别任务与低层图像表示之间的语义鸿沟。

近年来，随了神经网络的兴起，人们借助于神经网络学习、记忆多样化图像数据模式，获得判别力强大的高层特征表示。Zhen Zuo 等人 [68] 对于每一类别以数据自适应的方式激活一部分 filters，并保持不同类之间共享 filter 的特性，学习出具有很强判别力的 Filter Bank，用于编码特征。Bolei Zhou 等人 [69] 搜集了大小是 SUN 场景数据集 60 倍的 Places 数据集，并基于此数据集训练关于场景图像的卷积网络模型。训练好的网络可获得对其他场景图像数据集中图像的高层特征表示，且取得显著的分类性能。

不过，训练卷积神经网络参数需要大规模的数据，并且网络训练时间长。另外，受数据集和训练时间的限制，在利用卷积神经网络处理目标识别问题时，一般都会采用 ImageNet、Places 数据集训练好的网络模型参数，这样就会带来数据集域自适应的问题。

2.5 场景数据集发展

近年来，场景分类方法取得了很大的进步。与此同时，场景数据集的规模也有了比较大的发展。如图2.2所示，随着时间的变迁，主流的场景数据集无论是其图像数目还是其类别数目都有较大程度的增长，尤其是 2010 年之后，场景数据集的类别增长到了 397 类，图像数目增长到了 130519 幅。到了 2017 年，Places356 场景数据集 800 万的数据量，更是达到了场景数据集前所未有的数据规模。

第三章　基于隐关联语义表示的场景分类

一个场景图像通常是由若干个语义实体组成，比如，*sky*、*rock*、*street*，和*car*。这些实体总是以不可预测的层次结构组织在一起 [4, 71]，同时它们还被多个类别所共享。这就很可能为场景识别任务带来严重的类内差异性和类间相似性问题。在场景分类任务中，场景的标号，比如，*coast*、*village*、*coast*和 *inside city*，等价于对场景图像的整体认知和高层语义抽象，而这恰恰很难被低层特征捕捉到。与以目标为中心的图像分类任务相比，场景图像分类任务更加具有挑战性。本章以探究场景语义为中心，研究利用主题模型学习隐语义表示相关技术，主要回答场景图像特征表示时如何刻画场景语义的问题。考虑到场景图像中所存在的语义关联这一重要特性，建模图像中的语义关联性，提出了隐关联语义特征表示。我们期望，关联性的加入会提升所学习的生成式模型的判别力，提高场景图像的识别精度。

3.1　引言

针对于场景分类任务，为了刻画语义，首先要明确的一个前提是所要研究的语义是显式的还是隐式的。显式的语义一般指完整的实体目标，如汽车、桌子、杯子等；隐式的语义则指代的内容比较抽象，因为其主要是通过机器学习方法以非监督的方式学习得到的。不同类型的语义将会导致不同的语义获取策略，同时也会影响不同的场景特征表示的编码。

在现有工作中，常见的显式的语义是预先定义好的显式语义或者场景类别。此类语义需要用人工标注或分割区域，进而训练特定的语义分类器。语义获取的方式可以是人工标注或者视觉任务（如图像分割、语义分割、目标检测等）。一种简单的语义获取方式是场景类别就是可以被利用的语义。该策略假设一幅图像的场景类别是被该图像的所有区域块共享的 [16]。最流行的语义获取策略是依赖预先训练好的目标检测器。在给定的以目标为中心的图像数据集上预先训练一组目标检测器，利用这些训练好的目标检测器在场景图像上执行检测操作，以检测到的目标当做“真实”的目标而忽略其中的错检和漏检目标 [15]。基于显式语义的特征表示方法，本质上就是以目标为中心的场景图像表示方法。尽管该类方法获得了不错的分类性能，但是它们极大地依赖于人对于语义的预先定义（如定义可能出现在场景图像中的目标类别）。更严重的是，其极大地受限于显式方式获取到的语义结果。这实质上是把原本一个场景图像特征表示的问题转变为近似目标检测、图像分割等任务，很可能造成问题研究核心的模糊与偏差，忽视了探究场景图像的本质性上自然结构等问题。

在现有工作中，隐式的语义主要是从图像中学习获得，其策略可概括为依赖机器学习方法，以非监督或者监督的方式学习语义。因此，此类语义一般不需要显式的图像分割、人工主题标注或者大量的目标检测。基于隐式语义的特征表示，本质上就是以场景为中心（scene-centric）的特征表示 [17–20]。本章中的研究对象是隐式语义。

3.2 相关工作

一般地，以场景为中心的特征表示是基于 BoW 模型而构建的。BoW 是以无序的局部描述子的集合来编码一幅图像。它将由 k-means 等聚类方法获得的聚类中心作为语义，并且编码语义直方图作为图像的特征表示。BoW 量化局部描述子的过程是有信息损失的，这势必会造成词的语义模糊（word ambiguity）[21]：一义多词（synonymy）和一词多义（polysemy）。一词多义指的是，不同的视觉词可能表示了相同的语义；一义多词指的是，相同的视觉词在不同的上下文信息中可能表示不同的语义。如图3.1所示，即使前两幅属于“*village*”场景，BoW 特征仍旧呈现非常大的差异性。该现象表明了 BoW 对于类内变化性的处理能力比较弱。为了解决 BoW 的语义模糊问题，起源于统计文本语料的生成模型，概率化隐语义分析模型（probabilistic Latent Semantic Analysis，pLSA）[55] 被用于学习场景隐语义 [18, 19]。但是 pLSA 模型的学习受样本数量的影响，容易产生过拟合问题。隐狄利克雷分布模型 LDA[56] 通过引入先验分布的方式解决了 pLSA 模型的过拟合问题。在 2005 年，LDA 被用于处理场景图像，学习中层隐主题/语义特征，从而进一步提升了 BoW 特征性能表现 [17]。

场景图像呈现比较强的语义关联特征，而这一特性在场景特征表示与分类问题研究中总是被忽略。如图3.1的第三行所示，我们可以观察到一组主题，即 *sky-rock-house-tree*，总是在“*village*”场景中共现。很明显，一个场景图像会呈现出一种很强的语义/主题关联特性，并且更重要的是，该特性对于一个场景类别是特定的，而且是可以将该类别与其他类别区别开来的。但是糟糕的是，这样的关联性在现有的大部分工作中都被忽略，包括在 BoW 模型中。例如，LDA 在主题比例上加入狄利克雷先验分布 [56]，而这种做法比较强硬地假设主题（themes/topics）之间是相互独立的。

在本章中，我们利用关联主题模型（Correlated Topic Model，CTM）[117, 120] 来捕捉主题之间的关联性，并将其衍生的隐主题作为一种语义表示。CTM 是将经典的隐狄利克雷分布模型的狄利克雷先验分布替换为一个更加灵活的 logistic 正态分布 [116]，而该分布主要是引入了主题之间的协方差测度。然而，如果我们依旧采用传统方法提取语义，即仅仅考虑隐主题分布 [18, 19]，那么从 CTM 中获得的隐语义特征表示将无法保证一直都有好的识别性能表现。这个类似的情况在

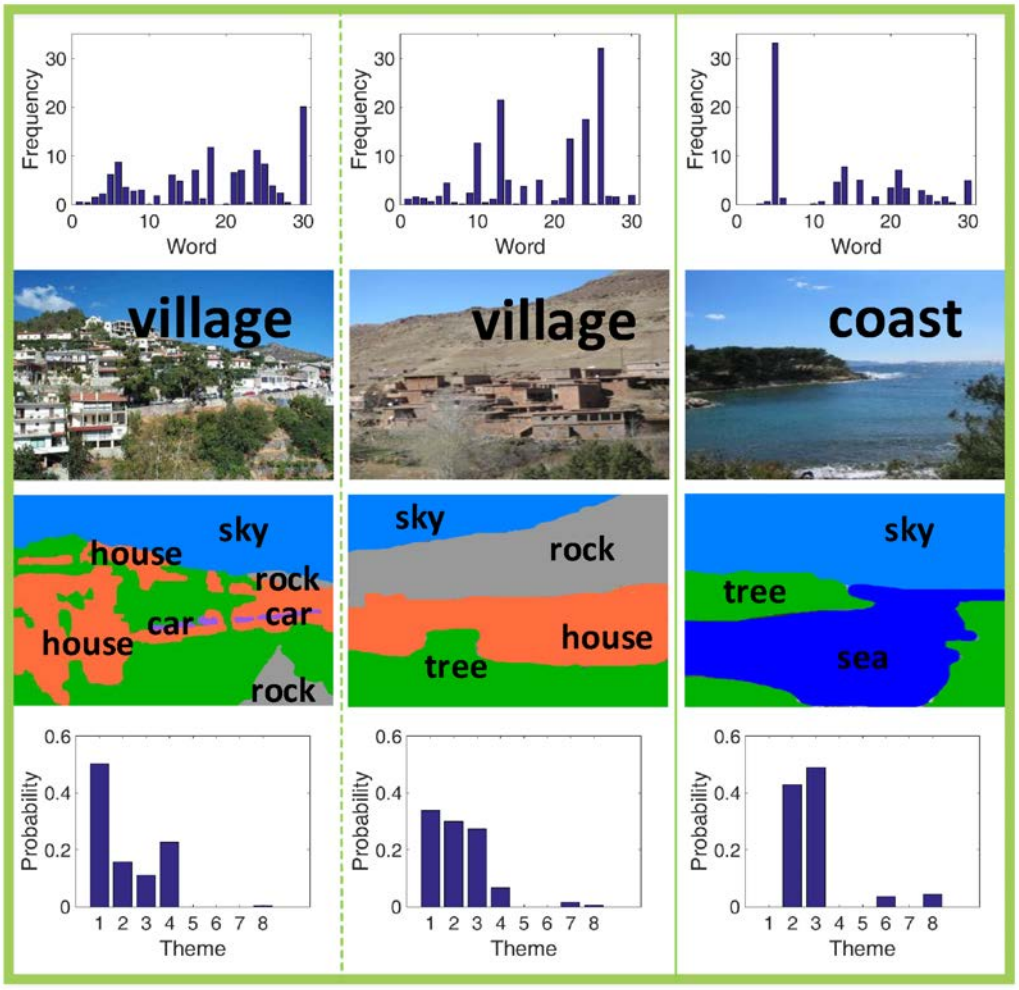


图 3.1 “village”场景和“coast”场景图像样例。第二行是场景图像样例，第一行是它们各自所对应的视觉词直方图，第三行是它们各自对应的主题。第四行给出了它们各自所对应的主题概率分布。

其他主题模型上也时有发生 [105]，例如 pLSA 和 LDA。由于它们的非监督学习特性，一般地，从这些主题模型中得到的隐主题表示总是缺少判别力的。

为了获得与时俱进的性能表现，所提出的隐关联语义表示的实现是基于卷积神经网络（Convolutional Neural Network，CNN）[89] 特征的。有文献表明，在 ImageNet[77] 数据集上训练的 CNN，其深层学习到了语义特征 [107]，尤其是其全连接层提取到的特征展示出了明显的语义聚类效果。因此，对于场景识别问题，卷积神经网络特征可以被用来作为一种特征描述，无需任何目标检测或者目标分割的额外工作。直观地，我们可以把全连接隐层的 CNN 特征当作是学习到的软赋值（soft-assignment）的词直方图，从而避免了把 CNN 作为局部描述子实施聚类、距离度量等操作来建立视觉字典。

3.3 deep-BoW

传统的 BoW 构建方式一般包括提取局部描述子、聚类获取视觉字典（visual dictionary）、度量距离、直方图计算等操作 [17, 19]，其中视觉字典包含的是视觉词（word）。在 CNN 出现之后，局部描述子由常用的 SIFT 特征替换为 CNN 来构建 BoW[16, 92, 115]。在本文中，我们将此 BoW 称之为 CNN-BoW。该 BoW 构建方式带来了复杂、代价大的处理。为了解决此问题，我们结合 CNN 优异的语义特征，打破传统的 BoW 构建方法，提出了 deep-BoW。

deep-BoW 的实现包括三个过程：局部特征提取、特征编码（encoding）和特征池化（pooling）。图3.2是 deep-BoW 的构建流程。首先，一幅图像被划分为图像块（patch），并在一个密集网格上抽样图像块。在本文中，我们选择卷积神经网络的最后一个全连接层作为局部特征。接下来，局部特征将通过维度采样的方式进行特征编码，获得局部 soft-assignment BoW。最后，采用 average pooling 的方式，来处理该图像中所有图像块的局部 BoW 向量，从而获得一个全局 soft-assignment deep-BoW。由于主题模型的输入值一般是是整型数值，所以为了方便后续主题模型的学习，我们将实数值的 soft-assignment deep-BoW 正则化为数值为整数的 deep-BoW。

在特征编码的过程中，选取卷积神经网络全连接层特征作为局部基础描述子。考虑到对于预训练好的图像分类网络，这些全连接层的神经元输出值的大小反映了神经元对图像视觉模式的响应程度，而且每个神经元倾向于对特定图像视觉模式或者目标产生响应 [143]。因此我们可以把卷积神经网络当做视觉字典，选取语义抽象性更强的全连接层的神经元作为视觉字典的视觉词，把全连接层输出作为一种软赋值的 BoW。假如采用传统的 BoW 构建方式，基于卷积神经网络深度特征的 BoW 仅仅是把深度特征替换常用的传统手工设计描述子，但是该方式仍旧无法避免聚类、距离度量等庞大计算量，尤其是聚类、距离度量高维深度特征。为了满足不同模型复杂度的需求，特征维度采样得到模型特定维度的 BoW 向量。采样策略可以是 fixed sampling，average sampling，或者 max sampling 等。在后面的实验部分，我们将给出对不同采样方式的评估结果，并验证所提出 deep-BoW 的有效性：无论采用哪种采样策略，所提出的 deep-BoW 特征分类性能表现均优于基于聚类算法（如 GMM、k-means）的 CNN-BoW。

deep-BoW 的合理性可归纳为以下几点：

- 语义特性。深度 CNN 已被验证具有显著的识别性能 [89, 107, 124]，其深层的激活特征表现出了优异的图像表示泛化性和强大的语义聚类效应 [107]。
- 软赋值特性。CNN 特征可以被认为是一种软赋值的 BoW。
- 非负性。调整线性单元变换（rectified linear unit transformation，ReLU

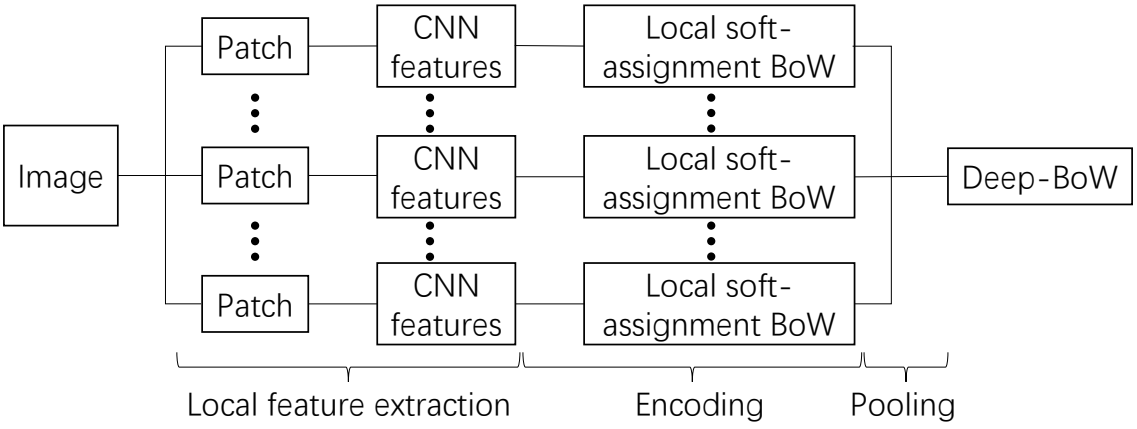


图 3.2 deep-BoW 构建流程

transformation）保证了特征值是非负的，与 BoW 本质相符合。

- 计算量小。与传统的基于聚类构建 BoW 方式不同的是，避免了那些聚类、量化等计算代价大的算法，保留了更多的信息。其构建过程主要涉及的是简单的线性代数计算操作。

3.4 隐关联语义表示

语义学习模型采用 CTM，一方面是为了去掉狄利克雷先验分布对于主题的独立性假设 [114, 115]，另一方面是借助于 logistic 正态先验分布 [116, 125] 来进一步建模主题之间的关联结构。

我们假定一个场景数据集包含 D 幅图像。给定 V 个视觉词 word 的字典，每一幅图像是由一个视觉词的集合来表示。正式地，$w_d = \{w_{d,n}, n \in 1, ..., N_d\}$ 表示视觉词在字典中的索引，每一维度分别对应了在图像 d 中采样的 N_d 个图像块，也就是，$w_{d,n}$ 是图像 d 第 n 个图像块所被赋值的视觉词。在主题模型中，一幅图像被建模为关于 K 个隐主题的混合，其中每一个主题是由关于 V 个视觉词的多项式分布来表示的。给定某个主题，每一个词采样于一个多项式分布，词的概率是由一个矩阵 $\beta = (\beta_{ij})_{K \times V}$ 参数化的。CTM 的图模型结构可见图3.3，其生成式过程可见表3.1。

语义/主题之间的关联结构由模型中的 logistic 正态分布 [116, 125] 来建模。logistic 正态分布的参数是 K 维均值向量 μ 和 $K \times K$ 的协方差矩阵 Σ。这两个参数被称之为超参。图像 d 的主题比例是 $\theta_d = [\theta_d^1, ..., \theta_d^i, ..., \theta_d^K]$，其中

$$\theta_d^i = f(\eta_d^i) = \exp \eta_d^i / \sum\nolimits_{i'} \exp \eta_d^{i'} \tag{3-1}$$

i 或者 i' 指代的是 K 主题中的第 i 或者第 i' 个主题。CTM 假设 η_d 服从一个参数为 $\mathcal{N}\{\mu, \Sigma\}$ 的正态分布。因此，函数 $f(\eta_d)$ 将 η_d 映射到它的均值参数形式：θ_d。

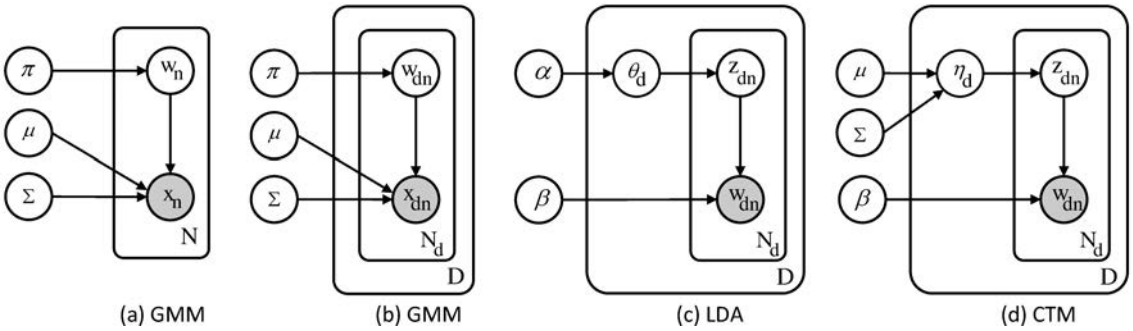
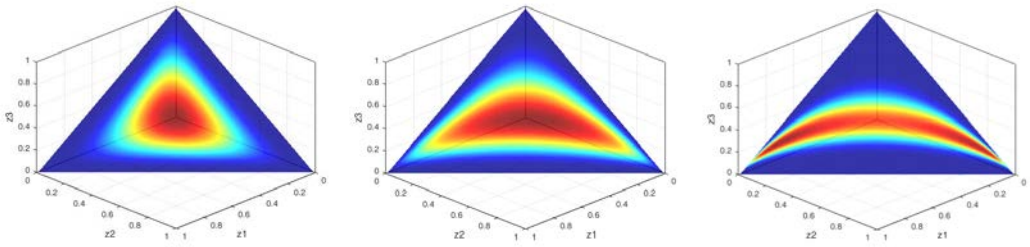


图 3.3 GMM、LDA 和 CTM 图模型。（a）GMM 图模型，（b）考虑样本数的 GMM 等价图模型，（c）LDA 图模型，（d）CTM 图模型



(a) 正相关　　(b) 独立　　(c) 负相关

图 3.4 logistic 正态分布的二维单纯形示例

表 3.1 主题模型的生成过程

1. 采样 $\eta_d|\{\mu,\Sigma\}\sim\mathcal{N}\{\mu,\Sigma\}$，其中 μ 和 Σ 是参数：均值和协方差。
2. 对于每一个词 $w_{dn}, n\in\{1,...,N_d\}$：
   (a) 从多项式分布 $Mult(f(\eta_d))$ 中，采样主题 z_n，其中 $f(\eta_d)$ 是主题比例 η_d 对于平均参数化（mean parameterization）θ_d 的自然参数形式；
   (b) 对于每一个主题，从多项式分布 $Mult(\beta_{z_n})$ 中，采样词 $w_{d,n}|\{z_n,\beta\}$。

θ_d 是位于 $K-1$ 维的主题单纯形（topic simplex）上的一个点。值得一提的是，参数 Σ 揭示了主题之间的关联关系。如图3.4所示，主题之间的正相关、相互独立和负相关分别在单纯形上有不同情况的呈现。

图像 d 的对数似然是 $L=\log p(w_d|\mu,\Sigma,\beta)$：

$$\begin{aligned}&p(w_d|\mu,\Sigma,\beta)\\&=\int p(\eta|\mu,\Sigma)(\prod_{n=1}^{N_d}\sum_{z_n}p(z_n|\eta)p(w_{d,n}|z_n,\beta))d\eta,\end{aligned}\tag{3-2}$$

其中 z_n 指代的是词 $w_{d,n}$ 的主题赋值向量（topic assignment vector），也就是词 n 在图像 d 出现的出现一次。该向量只有一项等于 1，其他项全部等于 0。

很明显，主题比例 $p(\eta|\mu,\Sigma)$ 所服从的 logistic 先验分布，与主题赋值 $p(z_n|\eta)$ 所服从的多项式后验分布，两者是非共轭的（non-conjugate）[117]。因此，公式

(3-2) 中的积分是很难解析地计算出来。因此，CTM 的求解一般采用变分贝叶斯 [119] 方法，引入变分参数来近似求解。变分贝叶斯是一个近似方法，用于优化对数似然函数（log-likelihood）的确定性目标下界 [119]。具体的是，在 mean-field 假设下 [120]，原始的图模型可以用变分参数（variational parameters）$\{\lambda, \nu^2, \phi\}$ 来简化。此时，我们可以得到 $L = L_{VB} + D_{KL}(q||p) \geq L_{VB}$，其中 D_{KL} 是分布 q 与分布 p 之间的 KL 散度（Kullback-Leibler divergence，KL divergence）。L_{VB} 是对数似然函数的下界。L 可以近似为 L_{VB}：

$$\begin{aligned} L_{VB} = E_q[\log p(\eta|\mu, \Sigma)] + \sum_{n=1}^{N_d} E_q[\log p(z_n|\eta)] + \\ \sum_{n=1}^{N_d} E_q[\log p(w_{d,n}|z_n, \beta)] + H(q), \end{aligned} \tag{3-3}$$

其中 $E_q[\cdot]$ 是关于变分分布（variational distribution）q 的期望，该变分分布的参数是 $\{\lambda, \nu^2, \phi\}$；$H(q)$ 是该变分分布的熵。变分参数 $\{\lambda, \nu^2, \phi\}$ 是 K 维的图像特定（image-specific）的向量。关于如何求解变分参数 $\{\lambda, \nu^2, \phi\}$ 和模型参数 $\{\mu, \Sigma, \beta\}$ 可以见文献 [117]。

在 CTM 模型中，尽管主题数目对于整个数据集中的图像都是 K，但是主题的比例，也就是 θ_d，在不同的图像之间是随机变化的。这主要是因为主题是从先验分布中随机采样得到的。主题比例 θ 随图变化（即图像特定化）的特点，使得每一幅图像都有自己特定的主题比例大小，解决了 BoW 词义模糊问题；同时，主题比例大小受 logistic 正态分布影响，使得每一幅图像呈现出主题关联存在的特性。因此，使用图像的主题比例 θ 作为我们想要的一种隐语义表示，是一种合理的方式。

3.5 实验验证与模型分析

在本节，在实验设置上，我们采用与文献 [83, 118, 135] 类似的实验设计：基于所提出的特征，训练一个 one-vs-all 的线性 SVM 分类器，并且采用平均分类准确度 [92] 来度量分类性能。我们的实验均是在设备参数为 2.10GHz CPU、64G RAM 和 NVIDIA Tesla K40C GPU 的计算机上运行的。

3.5.1 设置

数据集。 我们在两个公开场景数据集上进行实验验证：SCENE 8 和 MIT Indoor 67 [87]。MIT Indoor 67 数据集包含 67 个场景类别，该数据集被划分为 5360 幅训练图像和 1340 幅测试图像，也就是每一类别分别包含 80 幅训练图像和 20 幅测试图像。

实验设置。 对于图像预处理，一幅图像的大小被调整到 256×256 像素。对于图像块的大小（在文中也称之为尺度）P，在实验中我们选用了 3 个尺度，即图像块的大小分别对应 256×256、128×128 和 64×64 像素。在所有尺度上，抽样图像块的网格步长是 32 个像素。对于局部特征描述子，我们利用 Caffe[124] 分别基于 ImageNet 数据集 [77] 和 Places 数据集 [69] 上预训练的 Alex-network[89] 提取 CNN FC7 特征。对于 deep-BoW，我们选用 max sampling 策略。

3.5.2 实验结果

deep-BoW 由于图像的大小迥异，所以不同的图像会产生不同数目的图像 patch 块。考虑到 deep-BoW 依赖于一幅图像中局部 patch 的数目，我们沿用 caffe 深度学习工具中对于输入图像的处理方式：将整幅图像归一化放缩到 256×256 大小，该做法从直观上来看会相对更高效一些。为了验证这个假设，我们又依据另一个构建 deep-BoW 的策略，并做了实验。该策略是：首先裁剪原始图像，然后将裁剪成的 patch 块放缩到 256x256 大小，再一一进行 CNN 特征提取。我们将该策略生成的 deep-BoW 简称为 deep-BoW(CR)。我们对比了本章所提出的 deep-BoW(ours) 和 deep-BoW(CR)，给出了它们在不同情况下的性能表现，见图3.6和图3.7。从实验结果可以看出，deep-BoW(ours) 几乎在所有的测试情况下的性能表现均优于 deep-BoW(CR)。因此，在本文所有涉及到 deep-BoW 的实验中，均采用本章节所提的 deep-BoW(ours) 构建策略。

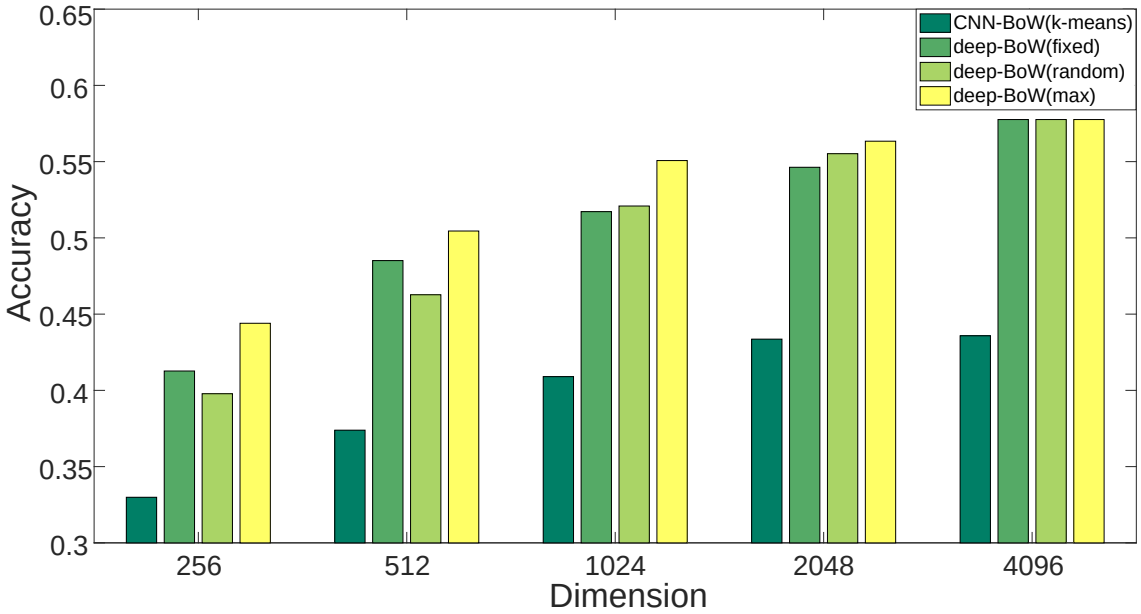


图 3.5 CNN-BoW 与 deep-BoW 在 MIT Indoor 67 数据集上的对比结果

隐关联语义表示 为了验证 CTM 的性能，我们利用 t-sne[138] 在图3.8中可视化了三种特征：BoW、基于 LDA 的隐语义表示和基于 CTM 的隐语义表示。这

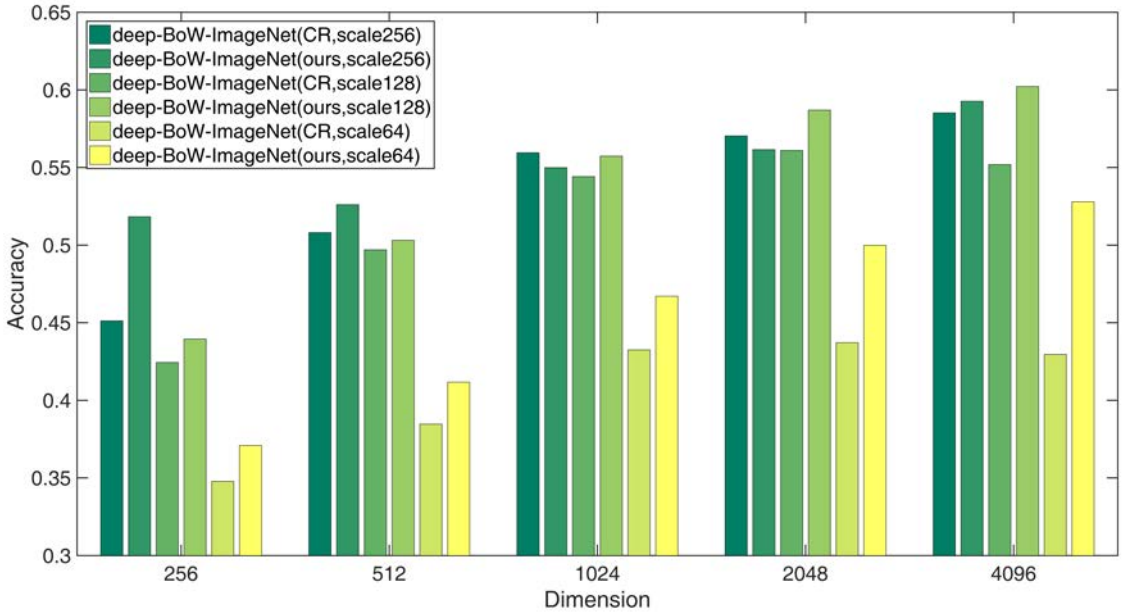


图 3.6 在不同尺度下 deep-Bow (CR) 与 deep-BoW (ours) 的性能对比

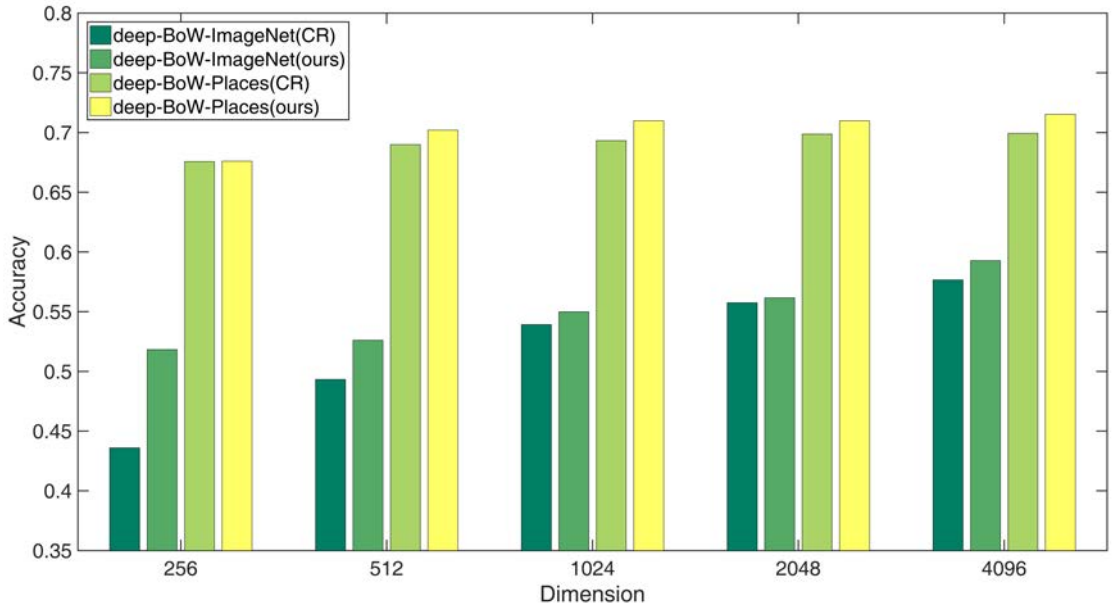


图 3.7 在 256 尺度下不同神经网络下 deep-Bow (CR) 与 deep-BoW (ours) 的性能对比

些实验选择在 SCENE 8 数据集上 [48] 进行。另外，我们从三个测度上对三种特征的性能进行了进一步的评估：聚类纯度 [139]，Dunn Validity Index (dvi) [140] 和平均分类精度 aca[92]。聚类纯度（purity）指的是聚类得到的单个 cluster 中所属占主要比例的类别的样本数与该 cluster 的大小的比值，cluster 大小等于该聚得的 cluster 中所包含的样本总数。纯度越高，聚类方法的效果就越好 [139]。dvi 衡量的是聚类得到的 clusters 的紧致程度 [140]。aca 是每一类分类正确率的平均值 [92]。从图3.8中可观测到，在隐语义特征空间中，基于 CTM 的隐语义特征都

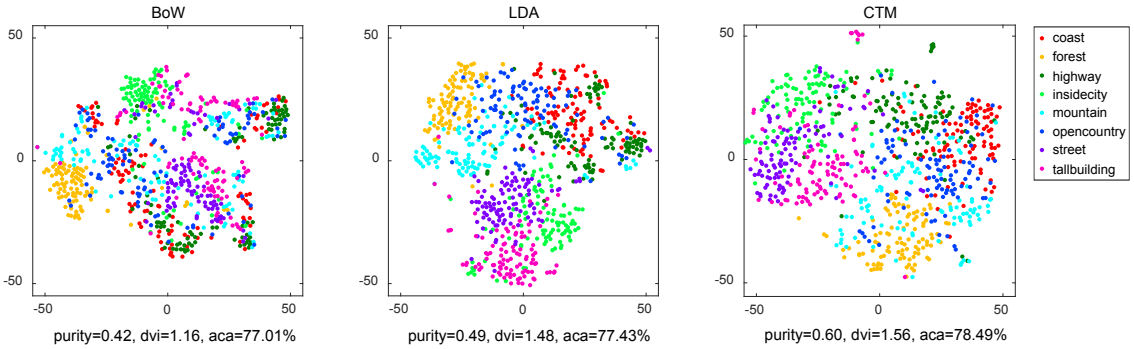


图 3.8 利用 t-sne 方法的特征可视化结果图

表现出了优越的聚类效果；同时它还获得了三个特征中的最好场景分类性能表现。

图3.9给出了隐关联语义表示在 SCENE 8 数据集上的分类表现。值得注意的一点是，隐关联语义表示并不能随着主题数目的增加而一直增加，图3.9中的实验也验证了这一情况。固定字典大小不变，对于字典大小等于 200 和 256 的两种情况，隐主题表示特征随着主题数目的增加，其分类性能也分别随之增加；但是当主题数目分别大于 50 和 60 时，其分类性能反而随主题数目的增加而下降。基于 LDA 的隐语义表示也存在类似情况，并且其情况在文献 [105] 中也得到了验证。总的来讲，隐语义表示的局限性通常不是来自于槽糕的统计估计结果，而是来自于最根本的 BoW 表示所存在的内在性的语义模糊 [20]。另一个原因是，源于词共现的隐语义特征并没有利用语义/主题与词之间的统计信息。这两个原因解释了为什么基于 CTM 的隐语义表示用于分类任务时的局限性，尽管它通过建模主题之间的关联结构更好地刻画了场景语义。

为了进一步基于语义获得表现优异的场景特征表示，我们尝试从信息几何的角度，探究底层词对于中层语义的贡献。为了实现这个目标，我们在第四章将会提出关联主题向量的解决方案；关联主题向量是基于 Fisher Kernel 架构，能够融合生成式和判别式两种方法的优势。

主题语义关联性讨论 为了进一步论证学习到的主题，我们选用 SUN 397 数据集图像中具有较全的目标分布标注信息 [6] 的图像。为了简单起见，我们将这些选出的图像称之为 SUN-anno 数据集，并且在该数据集上来评估学习到的主题以及主题之间的关联性。详细地讲，我们利用手工分割的目标区域来构建主题概率图（Topic Probability Maps，TPMs），统计每个分割出的目标区域与主题之间的对应信息。这些主题概率图 TPMs 用于展示学习到的主题与实际目标之间的对应关系，更显式地展示我们学习到的隐主题/语义是什么。为了构建 TPM，基于学习到的全局参数 β 和图像特定（每一图像都有自己所对应的）的参数 ϕ，我们对一幅图像的每一像素赋予一个 K 维的主题概率向量。其中，位于一个图像块的像素共享一个相同的主题向量，这是因为在为 CTM 模型构建 BoW 时是将图像分为图像块的，deep-BoW 的一个图像块对应一个 word。正如我们了解的，每一

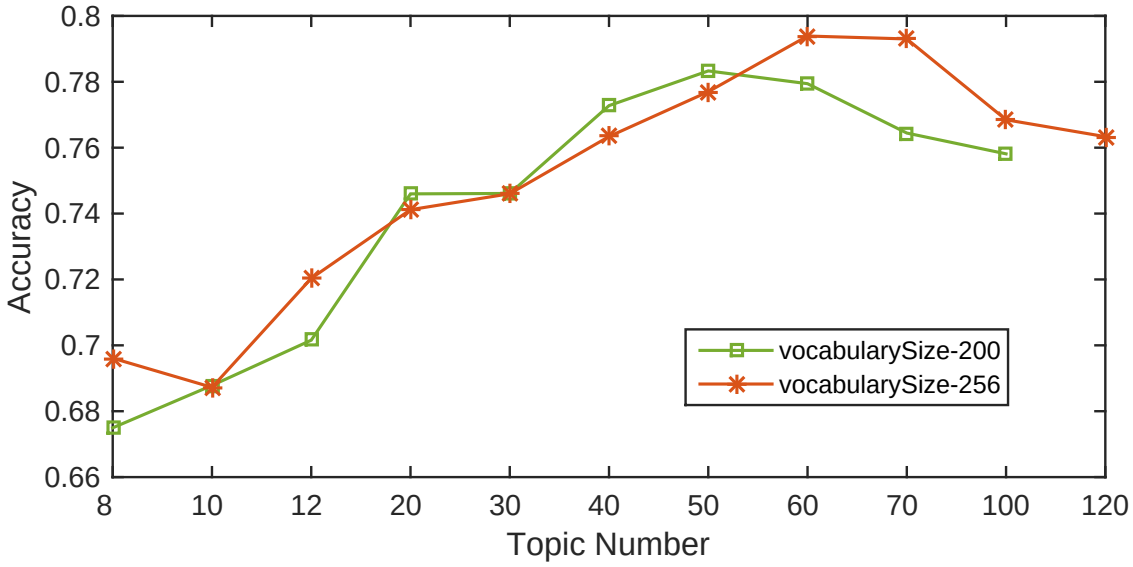


图 3.9 基于 CTM 的隐语义表示特征在 SCENE 8 数据集上的分类性能结果

个像素可能被若干个图像块共享（因为采样图像块的方式密集采样，图像块的大小和网格的步长影响了各个图像块之间的重叠情况）。基于一幅图像中的所有图像块，对每一个像素的主题概率向量分别进行加和，这样每一个像素都对应一个主题概率向量。由于每一幅图像大小不等，所以每一个像素的主题概率向量将会除以像素被图像块共享的次数，从而得到归一化的主题向量，也即得到该图像的主题概率图 TPM。更具体地讲，由于每一个像素点都对应了一个 K 维的主题概率向量，所以一幅图像的主题概率图 TPM 其实可等价于 K 个图像大小的特定主题的概率图。

图3.10首先给出了两大场景种类的全局主题概率图（global TPM）：室内场景类别和室外场景类别。全局主题概率是由加和一幅图像的 K 个特定主题的概率图而得到的，即它是一个图像大小的概率图。尽管主题的学习是非监督的，但是，从图3.10可以观察到，TPMs 中高概率的像素很好地对应了场景图像中分割出的目标。另外，有意思的是，令人感兴趣的目标被一起显现出来，例如，*toilet*目标和 *washbasin*目标一起显现在 *bathroom*场景中；*cushion*目标/*window*目标和 *bed*目标一起共现在 *bedroom*场景中；*building*目标和 *car*目标一同存在在 *street*场景中。

如图3.11所示，某些主题的 TPMs 表现出一种有意思的对应关系：模型所学习到的主题与人工目标/检测到的目标之间的对应。例如，在 *bedroom*场景中个，Topic2 的两个高概率的区域很好地对应了 *desk lamp*目标和 *chair*。进一步观察可看出，Topic3 高概率的区域对应 *table*目标和 *night table*目标，Topic4 对应 *window*，Topic5 对应 *chair*，Topic6 对应 *ceiling*，Topic7 对应 *table*，Topic8 对应 *floor*和 *bed*。在 *highway*场景图像中，Topic1 的概率高的区域对应目标 *truck*和目标 *occluded truck*，Topic3 对应目标 *field*，Topic5 对应目标 *tree*，Topic7 对应目

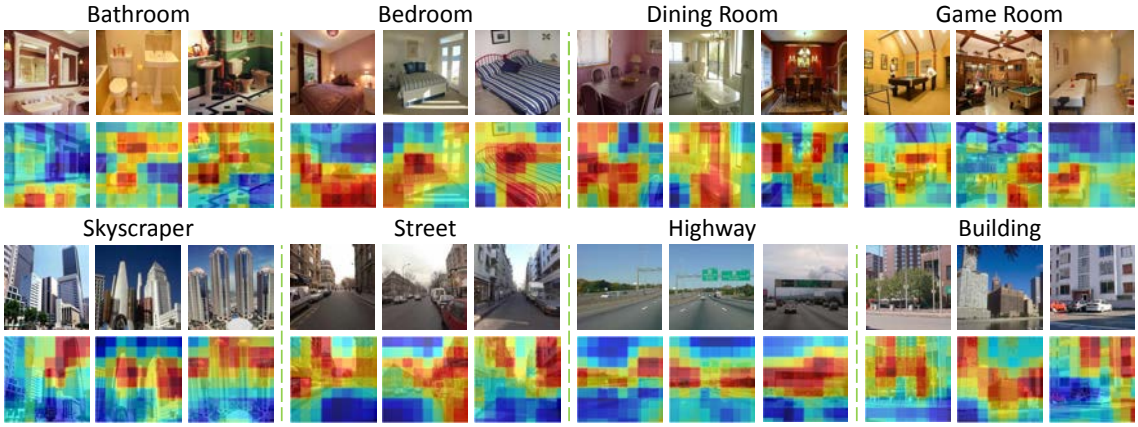


图 3.10 全局主题概率图。从图中可以发现，场景图像中概率高的主题通常对应了关联目标，例如，在 *bathroom*场景中，目标 *toilet*与目标 *washbasin*相关联地出现；在 *bedroom*场景中，目标 *cushion*与目标 *bed*相关联地出现。

标 *occluded truck*，Topic8 对应目标 *sky*。这些结果是在意料之中，也验证了隐主题语义生成的能力。值得注意的一点是，即使 K 个主题是所有图像共享的，但是由于所推断出的图像特定的参数 ϕ 的缘故，一个主题会在不同的图像中展示出不同的“目标”语义。总的来讲，在没有任何目标分割或者检测标注的情况下，所学习到的隐主题可以发掘出起主要作用的显式目标，同时在场景图像上呈现出了类别特定性（category-specific）的特质。

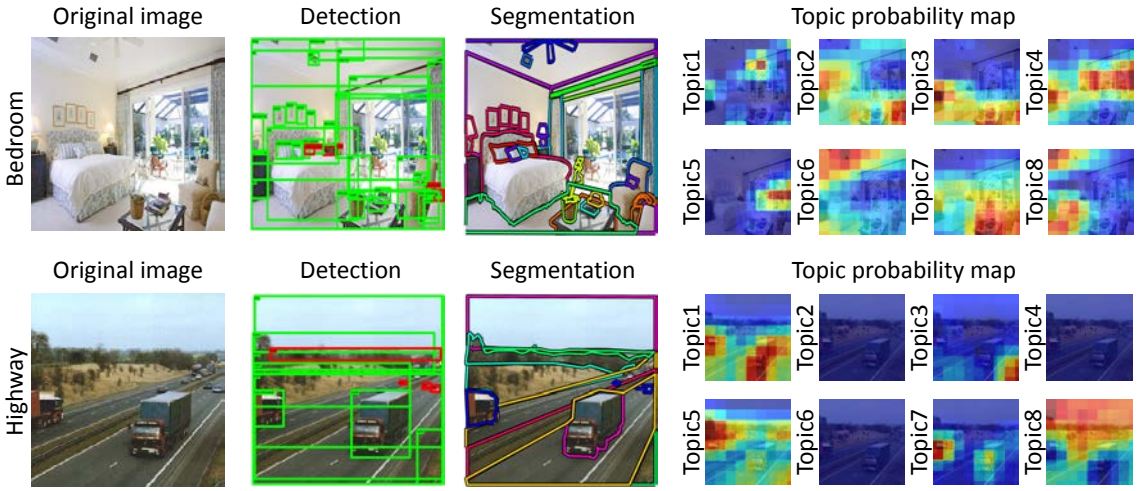


图 3.11 主题概率图以及他们与图像中检测和分割标注结果的对应。我们发现有意思的是，学习到的主题很好地对应了实体目标和图像分割区域。

为了显式地展示主题之间的关联性，我们在图3.12给出了所学习到的 8 个主题的关联矩阵，同时在图3.13中给出了针对于所有学习到的主题对应的部分目标，以及它们所对应的目标的统计结果。正如所预期的，这些结果在一定程

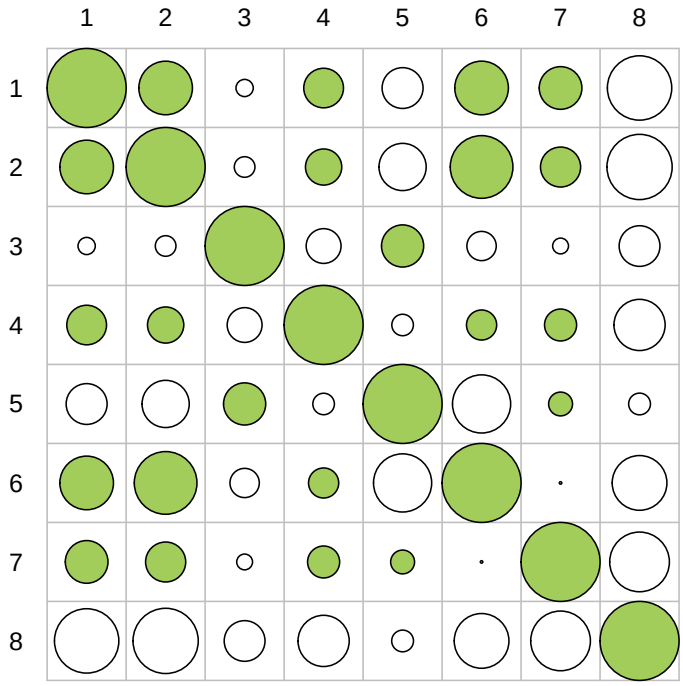


图 3.12 在 SUN-anno 数据上学习到的 8 个主题 topic 之间的关联关系。实心圆代表两个主题是正相关，空心圆表示两个主题是负相关的。圆的半径越大，正/负相关度越高。

度上解释了主题之间的正/负相关性。例如，Topic1 中的 *chair*目标和在 Topic2 中的 *table*目标总是一起出现，尤其是在室内场景图像中；而这一现象也正好与图3.12所示的 Topic1 与 Topic2 之间的正相关的结论相符合。

在众多场景中目标之间的共现现象和目标与所学到的主题之间精妙的对应关系，验证了关联问题确实在场景图像中存在。

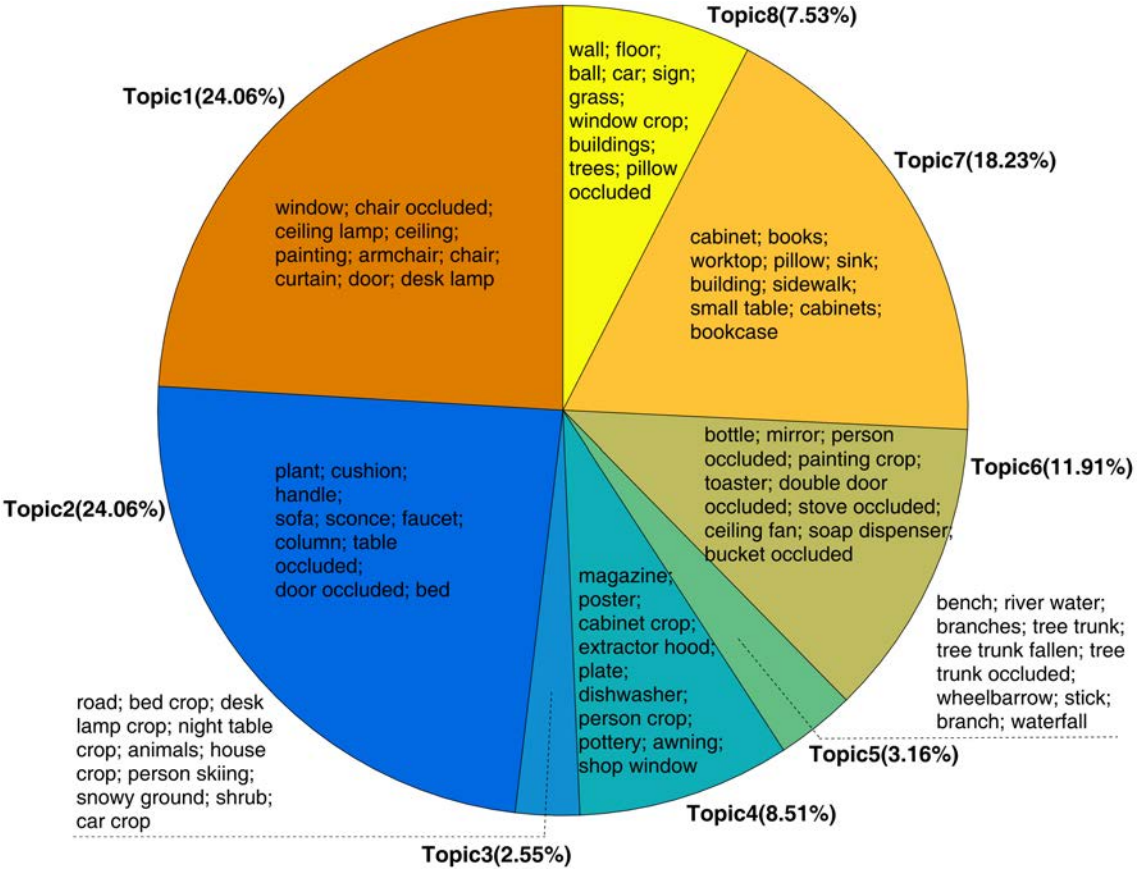


图 3.13 一个主题对应的目标比例，并且这些目标对应的最有可能的主题。图中也给出了每一个主题 topic 所对应的概率最大的 10 种类别的目标。从图中我们可以观察到，Topic1 中的目标 *chair*和 Topic2 中的目标 *table*总是一起出现，尤其是在室内场景图像中。这一现象正好与图3.12中所示的 Topic1 与 Topic2 是正相关的属性相一致。

3.6 本章小结

在本章中，我们提出了隐关联语义表示用于场景分类，旨在建模主题之间的关联性。我们去除掉局部图像块之间的独立同分布假设，采用罗基斯特正态先验分布，更好地建模场景图像并用于特征学习。另外，我们基于蕴含丰富语义信息的 CNN 特征提出了 deep-BoW 特征用于隐关联语义表示的学习，能够进一步提升隐关联语义表示的性能。

第四章　基于关联主题向量的场景分类

场景图像通常蕴含语义关联，尤其是对于大规模图像数据集。本章提出了一个生成式图像表示：隐主题向量（Latent Topic Vector，LTV）和关联主题向量（Correlated Topic Vector，CTV）。LTV 和 CTV 探索与利用了视觉词汇（visual words）对于主题（topic）的贡献，通过与 Fisher Kernel 框架的融合来指示场景间的差异性的方式，提升隐语义表示的判别力。考虑到场景图像的语义关联性总是被传统特征编码所忽略的问题，CTV 是在 LTV 的基础上去掉了语义独立性假设，加入语义关联性进一步提升特征的判别力。所提出的 CTV 还进一步与深度 CNN 特征和吉布斯采样结合，展现出了很大的潜力，尤其是处理大规模、复杂的场景图像。两个场景数据集上的实验结果验证了 CTV 的优越性。

4.1　引言

在本文中，为了提升隐语义表示对于分类任务判别力弱的问题，我们基于 Fisher Kernle 理论提出了隐主题向量 LTV 和关联主题向量 CTV。我们所提出的 LTV 和 CTV 本质上是从信息几何论的角度，进一步探索底层的视觉词对于中层主题学习的贡献与影响作用。而这是与 BoW 与隐语义表示是非常不同的，因为 BoW 依赖于视觉词出现的次数并且隐语义表示只是关注于主题的分布。对于两幅来自不同场景类别的图像，相似外观的区域总是倾向于相同的视觉词，但是隐语义由于有限的主题会导致其很难为识别任务提供较大的视觉词差异性。我们所提出的 LTV 和 CTV 基于 Fisher Kernel，结合了生成式方法和判别式方法的优势，并且将视觉词与主题的这些属性考虑了进去。考虑到场景图像中语义关联这一重要特性，CTV 是在 LTV 的基础上去掉了语义独立性假设。总之，我们所提出的 CTV 有以下贡献：

(1) 在 Fisher Kernel 空间中推导出了 CTV 的表示公式，旨在提升隐关联语义表示的判别力。

(2) 给出了高效的吉布斯采样求解方式，旨在在大规模数据集上提升 CTV 的可行性。

4.2　相关工作

受文本分类 [109] 的启发，BoW [108] 已被广泛应用于图像识别任务。它是以视觉词的共现来刻画一幅图像。然而硬性的词赋值与直方图编码方式导致了图像

空间信息的损失和每一个词的语义模糊问题，更不用说其对于场景识别任务的语义关联性这个重要属性的问题。场景图像的中层“主题”或者“语义”表示则是对 BoW 的一种扩展，它们是在尽力填补底层图像特征与高层语义概念的语义鸿沟。

直接将显示主题赋值于图像块（patch）或者区域（region）的方法，面临主题标注的庞大工作量与多样化目标的不可靠检测结果等挑战性问题。Li 等人 [15, 129] 提出了 “Object Bank” (OB) 的方法，其实质是借助于众多目标检测器在不同图像尺度上进行目标检测，以此获得每一个像素上每一个目标出现的概率。OB 方法一共需要在 12 个图像尺度、21 个空间金字塔网格上执行 177 个目标类别的检测器。如此庞大的检测计算量使其很难泛化到大规模场景图像数据集中，比如 SUN 397 [6] 或者 Places 205 [69]。除此之外，Li 等人从 1000 类目标中认真挑选了 177 类目标，以人工的方式解析这些 177 类目标之间的特性与语义关系。然而，这些关系并没有用于场景识别任务。

一些工作则是采用 Fisher Kernel 来提升 BoW[110]。Jaakkola 和 Haussler [116] 为分类任务给出了 Fisher Kernel 的一种表述方式。Perronnin 和 Dance 则推动了 Fisher Kernel 在图像分类任务中的发展。他们基于混合高斯模型推导出了更加具体的 Fisher Vector 形式，并将其引入到图像分类中 [83]。局部增加描述子向量（Vector of Locally Aggregated Descriptors，VLAD）[112] 是一种紧致的特征表示，可以看做是基于 GMM 的 Fisher Vector 的一种简化版本，一定程度上提升了 BoW 性能表现。总的来讲，Fisher Kernel 是依赖于生成模型的，目前其已经被广泛应用于图像分类中 [111]。由基于狄利克雷分布的 GMM 推导出的 Fisher Kernel [113] 则是作为特征变换的一种方式应用于图像分类任务。其假设 L_1 归一化的直方图局部描述子可以由狄利克雷分布来建模。

CNN 特征表示近年来在 ImageNet 目标识别问题上取得了令人瞩目的成果。其成功极大地激励着场景图像识别领域的研究者采用 CNN 特征嵌入用于场景分类任务，替换传统的基于 SIFT 局部描述子的 Fisher Vector 架构。例如，Gong 等人 [92] 从图像局部块提取 CNN 全连接层的特征，以此特征表示集合作为局部描述子来表示一个场景图像，并且为图像识别任务构建 VLAD 特征嵌入表示。Dixit 等人 [128] 将语义引入 Fisher Kernel 框架，提取局部图像块的 CNN 特征，并把这些特征作为语义多项式（Semantic MultiNomial，SMN）描述子。当局部语义描述子被建模为多项式分布时，在狄利克雷混合模型（Dirichlet Mixture Models，DMM）的基础下，我们就可以推导出比 GMM FV 更自然的一种特征嵌入，即 DMM FV。除此之外，Dixit 等人利用自然参数化变换来缓解 SMN 描述子的高度非欧式特性，并且在自然参数空间中采用 GMM FV 相同的做法计算出语义 FV。

基于用于图像分类任务的 Fisher Kernel 框架，相当多的工作已经取得了很大的进步；但是，一般地，他们都是假设所有图像的块是独立同分布地

（independent and identically distributed，i.i.d.）采样于相关的生成式模型中。很明显，这种独立同分布的假设违背了图像的本质特性。另外，语义关联性也很少在现有工作中被考虑到。Cinbis 等人 [114, 115] 把一幅图像看作是一个无序的区域集合，利用狄利克雷先验分布来参数化变量，而该变量在不同图像间是变化的。他们选用了那些能够捕捉到图像局部区域依赖关系的生成式模型，例如 LDA 和隐 GMM 模型。比如对于隐混合高斯模型，他们把 GMM 的参数作为隐变量，引入先验分布；从样本数据中学习先验分布参数，即模型的超参，用模型的超参控制关于隐模型参数的先验；应用 Fisher Kernel 准则，求取观测数据的对数似然函数关于模型超参的一阶导数。尽管在这些工作中隐语义被广泛地研究，但是语义关联性仍旧没有被考虑。

4.3 Fisher Kernel

定理 对于任一参数是 Θ 的概率模型 $P(X|\Theta)$，Fisher 核 $K(X_i, X_j) = U_{X_i}^T I^{-1} U_{X_j}$，其中 $U_X = \nabla_\Theta \log P(X|\Theta)$，具有以下性质 [110]：

- 它是一个有效的核函数；
- 它对于参数的任何可逆（和可微分的）变换具有不变性；
- 假设一个模型的标号是隐变量，如果一个分类器采用的是基于此模型获得的 Fisher 核，那么该分类器的结果将会渐进地表现出至少与基于同一模型的最大后验估计标号一样好。

在 Fisher 核中，U_X 被称为 Fisher Score，是对数似然函数关于模型参数的一阶导数，表示的是沿着坐标曲线、且经过参数空间内 Θ 点的速度；I 被称为 Fisher 信息矩阵（Fisher Information Matrix，FIM），是对数似然函数关于模型参数一阶导数的方差，扮演着张量度量的角色。在某正则条件下，FIM 本质是对数似然函数关于参数的二阶导数 [70]，即 $I_{[\Theta]}=E[{u_{[\Theta]}}^T u_{[\Theta]}] = -E[\partial^2 L/\partial\Theta^2]$。

Fisher 核可以看做是我们将原样本空间变换为一个距离度量为 I 的流形空间，U_X 是原样本 X 映射到此空间的一个点；在新空间中，距离度量变为 Fisher 度量 I，核的内积计算即是衡量两个样本相似性。进一步，由于 $K(X_i, X_j) = U_{X_i}^T I^{-1} U_{X_j}$，其中 $\Phi_{X_i} = I^{-1/2} U_X$，Fisher 核的内积的另一种等价计算是衡量在欧式空间中两个样本的相似性，只不过此时样本在欧式空间的特征表示变为了 Φ_X，而不是原来样本在欧式空间的语义概率的特征表示。因此，我们可以在欧式空间构建 Fisher 向量 Φ_X 作为特征表示，并将其送入分类器进行处理。

4.4 隐主题向量

隐主题向量是基于生成式模型 LDA[56] 推导出的 Fisher 向量。LDA 模型假设图像 d 的主题比例是服从参数为 α 的狄利克雷分布。图像中的词 w 出现概率服从多项式分布 $p(w_n|z,\beta)$，其中 z 是主题，β 是模型的另一个参数。对于 LDA，一个文档的似然函数是 $p(w|\alpha,\beta)=\int p(\theta|\alpha)(\prod_{n=1}^{N}\sum_{z_n}p(z_n|\theta)p(w_n|z_n,\beta))d\theta$。由于 LDA 模型中存在参数耦合情况，无法直接求解。通过引入变分参数 γ 和 ϕ，采用变分贝叶斯方法将原模型转换为一个可因式分解的近似模型，此时的似然函数是 $L(\gamma,\phi;\alpha,\beta)$。在 LDA 模型变分法求解的 E 步中，固定模型参数 α、β，通过最大化似然，得到最优的近似模型变分参数的 $\gamma^*(w)$、$\phi^*(w)$；在 M 步中，固定变分参数 $(\gamma^*(w),\phi^*(w))$，同样最大化似然，得到最优的模型参数 α、β[56]。每一幅图像都对应一个 $(\gamma^*(w),\phi^*(w))$ 对，所以参数 $(\gamma^*(w),\phi^*(w))$ 是与图像相关的，近似模型参数 $\phi^*(w)$ 是某一幅图像中出现的每一个 word 的 topic 概率分布，$\gamma^*(w)$ 是狄利克雷参数，它可以看做是在 topic 单纯形上一幅图像的一种表示。参数 β 是图像集级别的，它表示的是从整个图像数据的角度来看每一个 topic 所对应的字典中 word 的概率分布。

依据 Fisher 核理论，我们可以基于变分贝叶斯方法获得的对数似然函数推导出隐主题向量。首先，LDA 模型的对数似然函数关于模型超参数 α,β 和变分参数 ϕ,γ 的 Fisher Score 推导如下：

$$\frac{\partial L}{\partial \alpha_i}=\psi(\sum_{j=1}^{k}\alpha_j)-\psi(\alpha_i)+\psi(\gamma_i)-\psi(\sum_{j=1}^{k}\gamma_j), \tag{4-1}$$

$$\frac{\partial L}{\partial \beta_{ij}}=\sum_{n=1}^{N}\frac{\phi_{ni}w_n^j}{\beta_{ij}}, \tag{4-2}$$

$$\frac{\partial L}{\partial \phi_{ni}}=\psi(\gamma_i)-\psi(\sum_{j=1}^{k}\gamma_j)+log\beta_{iv}-\log\phi_{ni}-1+\lambda, \tag{4-3}$$

$$\frac{\partial L}{\partial \gamma_i}=\psi'(\gamma_i)(\alpha_i+\sum_{n=1}^{N}\phi_{ni}-\gamma_i)-\psi'(\sum_{j=1}^{k}\gamma_j)\sum_{j=1}^{k}(\alpha_j+\sum_{n=1}^{N}\phi_{nj}-\gamma_i). \tag{4-4}$$

其中 ψ 是对数 Γ 函数（gamma 函数）的一阶导数 [144]，λ 是模型求解中引入的一个新的变分参数 [56]。

LDA 模型的对数似然函数关于模型超参数 α,β 和变分参数 ϕ,γ 的 FIM 推导如下：

$$I_{\alpha_i}^{X}=\sum_{n=1}^{N}p(x_n|\theta)(\frac{\partial L}{\partial \alpha_i})^2=-E[\frac{\partial^2 L}{\partial {\alpha_i}^2}]=\psi'(\alpha_i)-\psi'(\sum_{j=1}^{k}\alpha_j), \tag{4-5}$$

$$I_{\beta_{ij}}^{X}=\sum_{n=1}^{N}p(w_n|\theta)(\frac{\partial L}{\partial \beta_{ij}})^2=-E[\frac{\partial^2 L}{\partial {\beta_{ij}}^2}]=\sum_{m=1}^{N}\frac{\phi_{mi}w_m^j}{{\beta_{ij}}^2}, \tag{4-6}$$

$$I_{\phi_{ij}} = \sum_{n=1}^{N} p(w_n|\theta)(\frac{\partial L}{\partial \phi_{ij}})^2 = -E[\frac{\partial^2 L}{\partial {\phi_{ij}}^2}] = 1/\phi_{ni}, \tag{4-7}$$

$$\begin{aligned} I_{\gamma_i} &= \sum_{n=1}^{N} p(x_n|\theta)(\tfrac{\partial L}{\partial \gamma_i})^2 = -E[\tfrac{\partial^2 L}{\partial {\gamma_i}^2}] \\ &= \psi'(\gamma_i) - \psi''(\gamma_i)(\alpha_i + \textstyle\sum_{n=1}^{N} \phi_{ni} - \gamma_i) + \\ &\quad \psi''(\textstyle\sum_{j=1}^{k} \gamma_j) \sum_{j=1}^{k} (\alpha_i + \textstyle\sum_{m=1}^{N} \phi_{mi} - \gamma_i) - \psi'(\textstyle\sum_{j=1}^{k} \gamma_j). \end{aligned} \tag{4-8}$$

基于推导出的 Fisher Score 和 FIM，我们根据 Fisher 向量 $\Phi_{X_i} = I^{-1/2} U_X$ 公式计算关于 $\{\alpha, \beta, \phi, \gamma\}$ 参数的四个 Fisher 向量，连接并正则化向量即得到隐主题向量 LTV。正则化可以是 power 正则化，或者 L_2 正则化 [118, 126]。

4.5 关联主题向量

基于假设前提：CTM 可以较合理地为隐语义特征建模主题之间的关系，并且 Fisher Kernel 可以进一步提升特征的判别力，那么，场景分类任务的解决方案就相当直观：首先从训练集中估计 CTM 的参数，然后在 Fisher Kernel 框架下为训练集和测试集图像各自构建关联主题向量 CTV。CTV 作为最终的特征表示，输入到一个线性 SVM 分类器中，用于识别不同场景类别。在本小节，我们将会具体讨论 CTV 的推导与求解。我们已经在第三章介绍了隐语义，尤其是隐关联语义，主要回答如何刻画/学习语义问题。接下来，我们提出的 CTV，结合生成式模型与判别式模型提升（隐）语义的判别特性，主要回答如何基于学习到的语义构建优异的特征表示。CTV 编码的基本方案具体可见图4.1。

关联主题向量是基于 CTM 生成式模型推导出的 Fisher 向量。生成式模型 CTM 的参数是 $\Theta = \{\mu, \Sigma, \beta\}$（即模型超参 $\{\mu, \Sigma\}$ 和全局参数 β），其关于图像 d 的对数似然函数如公式 (3-2) 所示。为了推导 CTV，我们需要计算对数似然函数关于模型参数 $\Theta = \{\mu, \Sigma, \beta\}$ 的 Fisher Score 和 FIM。正如第三章中所讨论的那样，CTM 中为刻画主题之间关联结构而引入的 logistic 先验分布与后验分布多项式分布是非共轭特性，和似然函数中参数耦合，造成似然函数计算、模型求解的难度，这也使我们不能直接推导出对数似然函数的梯度用于 CTV 特征计算。接下来，我们将结合 CTM 模型的两种求解策略分别给出两种 CTV 的实现。

4.5.1 基于变分贝叶斯的关联主题向量

我们首先讨论如何利用变分贝叶斯方法来推导出 CTV 特征的一个形式表达。公式 (3-2) 是图像 d 关于 CTM 的对数似然函数。我们采用变分贝叶斯方法 [119]

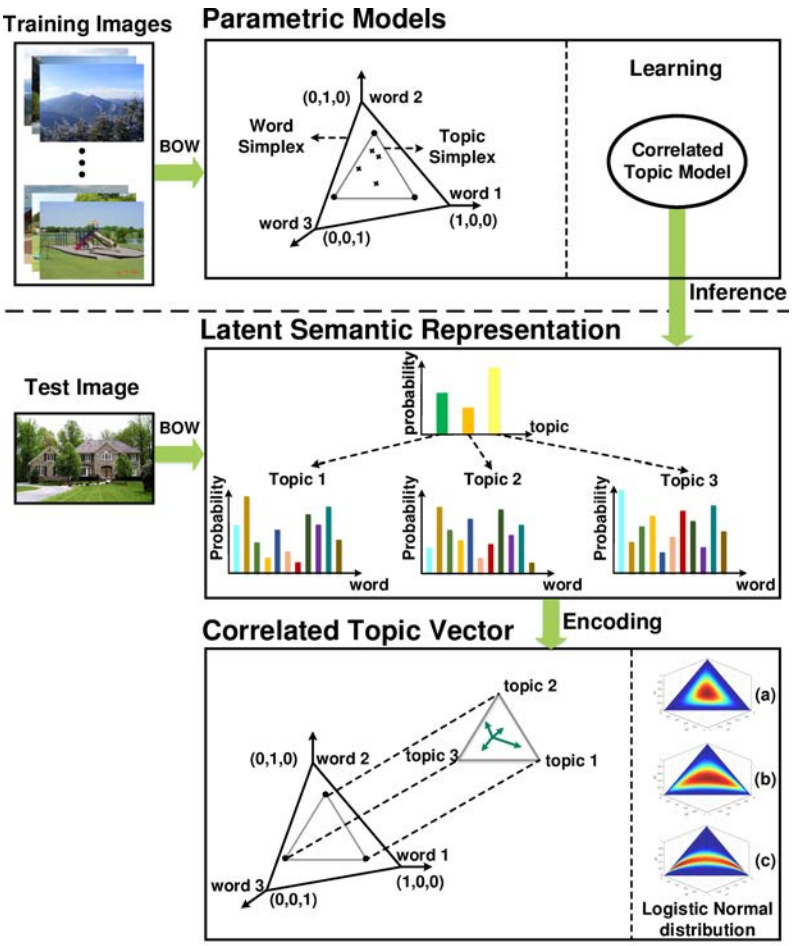


图 4.1 关联主题模型的学习以及关联主题向量的编码示意图

来推导出 CTV 的形式表达。正如第三章中所给出的，该对数似然函数被近似为公式 (3-3) 的 L_{VB}。接下来我们计算 L_{VB} 关于超参 $\{\mu, \Sigma\}$ 的 Fisher Score：

$$u_{[\mu]} = \partial L/\partial \mu = \Sigma^{-1}(\lambda_d - \mu), \tag{4-1}$$

$$u_{[\Sigma^{-1}]} = \partial L/\partial \Sigma^{-1} = 1/2(\Sigma - diag(\nu_d^2) - (\lambda_d - \mu)^T(\lambda_d - \mu)). \tag{4-2}$$

关于全局参数 β 的 Fisher Score 是 $u_{[\beta]} = (u_{[\beta_{ij}]})_{K\times V} = (\partial L/\partial \beta_{ij})_{K\times V}$，其中

$$\partial L/\partial \beta_{ij} = \sum_{n=1}^{N_d} \phi_{d,ni} w_{d,n}^j / \beta_{ij}. \tag{4-3}$$

μ 和 Σ 是真实多元高斯分布的参数，他们是从所有图像中学习到的。前面已经提到，λ_d 和 ν_d^2 是图像特定的，这主要是因为他们是从单个观测图像数据 w_d 中拟合得到的。$(\lambda_d - \mu)$ 衡量的是真实先验分布的均值与它的近似变分分布均值之间的差异。类似的，$\Sigma - diag(\nu_d^2)$ 项衡量的是两个分布方差之间的差异。$\phi_{d,ni}$ 是一个多项式分布参数，指的是在给定 topic i 的条件下，word $w_{d,n}$ 出现的可能性。$u_{[\beta]}$

可以看作是 word 出现次数的期望，其中 word 出现的概率 $\phi_{d,ni}$ 是有一个全局参数 β 加权的。为了避免矩阵乘法，我们将求解对数似然函数关于 Σ^{-1} 的偏导数，即公式 (4-2) 中 Σ 的逆，而不是求解对数似然函数关于参数 Σ 的偏导数。公式 (4-1)-(4-3) 的具体推导见附录A。

另外，关于超参 $\{\mu, \Sigma\}$ 的 Fisher 信息矩阵可以简单地表示为：

$$I_{[\mu]} = -E[\partial^2 L/\partial\mu^2] = \Sigma^{-1}, \tag{4-4}$$

$$I_{[\Sigma^{-1}]} = -E[\partial^2 L/\partial(\Sigma^{-1})^2], \tag{4-5}$$

$$\begin{aligned} I_{[\beta]} &= -E[\partial^2 L/\partial\beta_{ij}^2] \\ &= -\sum_{n=1}^{N_d} p(w_{d,n}|\theta_d)\partial^2 L/\partial\beta_{ij}^2 \\ &= -\sum_{n=1}^{N_d} p(w_{d,n}|\theta_d)\sum_{m=1}^{N_d}\phi_{d,mi}w_{d,m}^j/\beta_{ij}^2 \\ &= -\sum_{m=1}^{N_d}\phi_{d,mi}w_{d,m}^j/\beta_{ij}^2. \end{aligned} \tag{4-6}$$

有了 Fisher Score 和 FIM，我们立即就有了三个关于 $\{\mu, \Sigma, \beta\}$ 参数的近似 Fisher 向量：$\varphi_{[\mu]} = {I_{[\mu]}}^{-1/2}u_{[\mu]}$，$\varphi_{[\Sigma]} = {I_{[\Sigma]}}^{-1/2}u_{[\Sigma]}$，$\varphi_{[\beta]} = {I_{[\beta]}}^{-1/2}u_{[\beta]}$。接下来，我们连接并正则化这三个向量，就得到了想要的 CTV。正则化采用与 LTV 相同的 power 正则化，或者 L_2 正则化 [118, 126]。

4.5.2 基于吉布斯采样的关联主题向量

上一小节中，我们介绍了变分贝叶斯方法有确定性的（近似）对数似然函数 [117, 121] 表达式，可以被用于 CTV 形式表达的推导。但是，由于非共轭先验的本质问题 [121, 122]，在训练阶段，基于近似变分贝叶斯方法的 CTM 参数学习，需要耗费昂贵的计算代价。对于日益增长的数据集，大规模、复杂的数据势必会使得该问题日渐加剧。为了解决该限制问题，我们采用可扩展吉布斯采样算法 (scalable Gibbs Sampling algorithm) [121]。该算法比较容易执行，并且本质上是并行化的，适合大规模数据的学习。在接下来的讨论中个，我们将基于 VB 方法推导得到的 CTV，记作是 CTV-VB；将基于 Gibbs Sampling 的 CTV，记作是 CTV-GS。

Gibbs Sampling 避免了对数似然函数中积分项的确定性计算，是通过对一个随机抽样的隐变量采用随机变换操作而非优化对数似然函数的下界。但是，与此同时，我们很难直接推导出基于 Gibbs Sampling 的 CTV 的具体形式。解决该问题的一个策略是，近似出 Gibbs Sampling 解法的对数似然函数。这个直觉认识来

自于对于 Gibbs Sampling 与变分方法的联系的论证 [123]：L_{GS} 加上 D_{KL} 的期望等于 L_{VB}，即 $L_{GS} = L_{VB} - E_{q(z_T|w)}\{D_{KL}[q(y|z_T, x)||r(y|z_T, x)]\} \leq L_{VB}$，其中 x 是观测数据，z_T 是迭代采样的结果，$y = z_0, z_1, ..., z_{T-1}$ 是每一次迭代的状态变量序列（a series of state variables），$r(y|z_T, x)$ 是 $q(y|x, z_T)$ 的一个特定的近似分布。D_{KL} 是分布 q 与分布 r 之间的 KL 散度 [123]。L_{VB} 可以看作是 L_{GS} 的一个上界。因此，我们可以近似由 D_{KL} 的期望控制 Gibbs Sampling 的对数似然函数。对于 CTV-GS，我们尝试融合变分贝叶斯与吉布斯采样的优点来构建 CTV。具体地说，基于变分贝叶斯方法的 CTV 推导，为 CTV 特征提供一个一般式的形式化表达。对于 CTV-GS，其编码方式依旧依赖于 CTV-VB 的表达形式，其相关的参数学习将采用 Gibbs Sampling 学习。变分贝叶斯和吉布斯采样，两个方法都是对 CTM 的对数似然函数的近似，所以它们描述的是该分层概率图模型中相同的变量间的依赖关系。另外，我们将会在小节4.6中，通过实验来验证 CTV-GS 的这种近似方式的合理性。

4.6 实验验证与模型分析

4.6.1 隐主题向量

隐主题向量在场景分类任务上评估，我们选用 SIFT 特征作为局部描述子，数据集是 SCENE 8，分类模型采用的是 one-vs-all 的线性 SVM 分类器，性能评估采用平均分类准确度 [92]。

在 LTV 特征的计算中，我们增加对参数 ϕ 的对齐操作，实验结果对应表4.1中“Improved”一列。从实验结果可看出，改进后的关于参数 ϕ 分量的特征以及与其他分量特征的组合，大部分情况下都有了性能上的提高。

我们对 LTV 结果中四个参数对应的分量分别进行评估来分析各个分量对特征的重要程度，如表4.1所示。对于单个分量，关于参数 β 的 LTV 分量的分类性能最高，关于参数 α 则最低。两个分量的组合特征，性能略微超过单个分量特征。

我们通过在 LTV 中加入 FIM 的策略，对 LTV 中 FIM 进行了实验验证。不加入 FIM 的 LTV 相当于 Fisher Score，特征编码只用了一阶信息。加入 FIM 的 LTV 是尝试加入模型的二阶偏导信息并用于编码特征：最大似然估计方法在求解概率分布参数时，是令 log 似然的一阶导数等 0。而 Fisher vector 的核心 Fisher Score 也正是 log 似然关于模型参数的一阶偏导，且 Fisher Score 的期望等于 0。从目前实验结果来看，二阶 log 似然作用不大。

表 4.1 在字典大小为 64、主题数目为 8 时隐主题向量在 SCENE 8 数据集上的场景分类结果

LTV 特征				Accuracy(%) (LTV 计算中加入信息矩阵)		Accuracy(%) (LTV 计算中去掉信息矩阵)	
LTV(α)	LTV(β)	LTV(γ)	LTV(ϕ)	Plain	Improved	Plain	Improved
√	–	–	–	35.85	31.20	60.17	60.17
–	√	–	–	71.61	71.61	77.07	77.07
–	–	√	–	44.44	44.44	48.62	48.62
–	–	–	√	68.75	73.15	61.12	56.83
√	√	–	–	72.67	72.67	76.96	76.96
–	–	√	√	68.80	73.52	64.72	64.19
√	–	√	–	52.97	52.97	66.05	66.05
√	–	–	√	68.96	73.46	62.45	59.80
–	√	–	√	73.25	75.64	72.93	75.69
√	√	√	√	73.15	73.15	77.12	77.12
√	√	√	√	73.52	75.90	72.72	75.74

4.6.2 关联主题向量

在该小节，我们评估了所提出的 CTV 的性能，并且将其与多个最相关的方法进行了对比。在实验设置上，采用与文献 [83, 118, 135] 类似的实验设计：基于所提出的 CTV 特征，训练一个 one-vs-all 的线性 SVM 分类器，并且采用平均分类准确度 [92] 来度量分类性能。

4.6.2.1 设置

数据集。 我们在两个 benchmark 数据集上进行实验验证：SUN 397 [6, 88] 和 MIT Indoor 67 [87]。

实验设置。 考虑到 CNN 深层的激活特征表现出的优异的图像表示泛化性和强大的语义聚类效应 [107]，在具体实现上，我们采用 CNN 特征 [89] 作为局部描述子构建 deep-BoW，并用来评估所提出的 CTV。其中 deep-BoW 的实现过程见第三章内容。在训练阶段，deep-BoW 向量将会输入到 CTM 模型中用于学习模型参数，而 CTM 的学习方法是基于 VB 或者 GS 方法。给定了学习到的 CTM 参数，我们根据公式 (4-1 4-6)，将 deep-BoW 向量用于生成 CTV 特征，并将该特征用于测试阶段的分类任务。在实验中，我们在三个尺度上（小、中、大尺度，也就是取不同大小数值的 P）实现了 CTV-VB 和 CTV-GS（为了简化，两者统称为 CTVs）。对于多尺度 CTV（multi-scale CTV），考虑到 Fisher Vector 是高维

的，在具体实验中，我们并没有直接串联三种尺度上的原始 CTV 特征来获得多尺度 CTV。多尺度 CTV-VB/CTV-GS 是三种尺度的特征各自对应的 SVM 得分向量的串联。

由于 CTV 的实现是基于第三章的 deep-BoW，所以 CTV 实验中对于图像预处理，采用与 deep-BoW 相同的方法。我们首先把一幅图像的大小调整到 256×256 像素。对于图像块的大小，在实验中我们选用了 3 个尺度，即图像块的大小分别对应 256×256、128×128 和 64×64 像素。在所有尺度上，抽样图像块的网格步长是 32 个像素。对于局部特征描述子，我们利用 Caffe[124] 分别基于 ImageNet 数据集 [77] 和 Places 数据集 [69] 上预训练的 Alex-network[89] 提取 CNN FC7 特征。对于 deep-BoW，我们选用 max Sampling 策略。为了学习 CTV 中涉及到的模型参数，deep-BoW 会输入到基于 VB 学习的 CTM 模型中或者是基于 GS 学习的 scalable CTM 模型 [121] 中。另外，在 logistic 正态分布对应的各维度（概率）之和需满足等于 1 的约束，这意味着各维度是线性相关的，需要在 CTM 的求解中将其中一个维度的均值和方差全置为零，因此一般情况下协方差是非满秩、奇异的，这就为 Fisher 信息计算公式中计算协方差矩阵的逆带来了难度。因此，在具体的实验中，我们用 Fisher Score 来近似 CTV 特征。除之此外，文献 [22] 也所指出 FIM 是不重要的。

在接下来的讨论中，在 ImageNet 数据集 [77] 上训练的深度 CNN 网络被缩写为 CNN-I，在 Places 数据集 [69] 上训练的深度 CNN 网络被缩写为 CNN-P（PlacesNet）。基于 CNN-I 特征得到的 CTVs 被缩写为 CTV-Is (包括 CTV-VB-I 和 CTV-GS-I)，基于 CNN-P 特征得到的 CTVs 被缩写为 CTV-Ps (包括 CTV-VB-P 和 CTV-GS-P)。

4.6.2.2 主要实验结果

SUN 397 数据集。 表4.2展示了在大规模数据集 SUN 397 上的主要实验结果。我们主要从两个方面对比了所提出的 CTVs(CTV-VB and CTV-GS)：（1）与基于 Fisher Vector 框架的最相关的方法的对比，（2）与其他 state-of-the-art 方法的对比。在表4.2中给出的相关方法中，DMM FV [128] 是单尺度的，而 VLAD [92]、Semantic FV [128]、CTV-GSs 和 CTV-VBs 是多尺度的，具体地讲，是三个尺度的。

基于 ImageNet 的实验结果。 第一组对比方法采用 CNN-I 的 FC7 特征作为描述子。这些最相关的对比方法包括了 baseline CNN 和若干基于 Fisher Vector 的方法：DMM FV [128]、Semantic FV [128] 和 VLAD [92]。在表4.2中，baseline CNN-I 取得了 42.61% 的分类准确率 [69]。所提出的 CTV-VB-I 取得了 53.35% 的准确率，CTV-GS-I 取得了 53.21% 的准确率；它们分别比 baseline CNN-I 高出了 10.74% 和 10.60%。另外，对比其他基于 Fisher Vector 的方法，CTV-Is 也有

表 4.2 在 SUN 397 数据集上的实验性能对比

Methods		Accuracy(%)	Year	Description
相关方法	CNN-I[69] (baseline)	42.61	2012	在 ImageNet 数据集上训练的深度网络
	CTV-GS-I (ours)	**53.21**	–	基于 CNN-I 描述子和 Gibbs Sampling 求解的 CTV；三个尺度
	CTV-VB-I (ours)	**53.35**	–	基于 CNN-I 描述子和 Variational Bayesian 求解的 CTV；三个尺度
	CNN-P[69] (baseline)	54.32	2014	在 PlacesNet 数据集上训练的深度网络
	CTV-GS-P (ours)	**58.43**	–	基于 CNN-P 描述子和 Gibbs Sampling 求解的 CTV；三个尺度
	CTV-VB-P (ours)	**58.39**	–	基于 CNN-I 描述子和 Variational Bayesian 求解的 CTV；三个尺度
	DMM FV[128]	49.86	2015	基于 Dirichlet Mixture Model 的 Fisher Vector；单尺度
	Semantic FV[128]	51.80	2015	自然参数化的 GMM Fisher Vector；性能最好的三个尺度
	VLAD[92]	51.98	2014	VLAD；三个尺度
State-of-the-art	SPMSM[16]	28.20	2012	空间金字塔匹配（Spatial pyramid matching，SPM）；预定义好的语义主题
	Meta-classes[136]	36.80	2014	基于分类器的特征
	SUN(MKL)[88]	38.00	2010	多核学习
	DeCaF[107]	40.94	2014	Decaf；全局 CNN 特征

表 4.3 不同尺度上提取到的特征表示性能评估

Methods	**MIT Indoor 67**				**SUN 397**			
	256×256	128×128	64×64	Multi-scale	256×256	128×128	64×64	Multi-scale
VLAD[92]	53.73	65.52	62.24	68.88	39.57	45.34	40.21	51.98
Semantic FV[128]	59.50	65.10	–	68.80	43.76	48.30	–	51.80
CTV-GS-I (ours)	58.88	65.07	61.57	68.36	43.11	49.60	44.52	53.21
CTV-VB-I (ours)	59.78	65.52	62.31	68.88	44.30	50.08	47.00	53.35
VLAD (PlacesNet)* [92]	66.27	66.12	54.70	67.61	51.50	48.97	40.58	51.73
CTV-GS-P (ours)	70.82	68.88	57.61	73.51	54.95	52.46	41.49	58.43
CTV-VB-P (ours)	70.90	68.73	58.13	73.88	55.16	52.95	43.34	58.39

* 我们利用作者公开的代码来实现的。

表 4.4 在 MIT Indoor 67 数据集上的实验性能对比

Methods		Accuracy(%)	Year	Description
相关方法	CNN-I[69] (baseline)	56.79	2012	在 ImageNet 数据集上训练的深度网络
	CTV-GS-I (ours)	**68.36**	–	基于 CNN-I 描述子和 Gibbs Sampling 求解的 CTV；三个尺度
	CTV-VB-I (ours)	**68.88**	–	基于 CNN-I 描述子和 Variational Bayesian 求解的 CTV；三个尺度
	CNN-P[69] (baseline)	68.24	2014	在 PlacesNet 数据集上训练的深度网络
	CTV-GS-P (ours)	**73.51**	–	基于 CNN-P 描述子和 Gibbs Sampling 求解的 CTV；三个尺度
	CTV-VB-P (ours)	**73.88**	–	基于 CNN-I 描述子和 Variational Bayesian 求解的 CTV；三个尺度
	Latent GMM FV[115]	65.00	2015	Latent GMM based Fisher Vector；单尺度
	Spare Coding FV[134]	68.20	2014	Sparse Coding based Fisher Vector；单尺度
	DMM FV[128]	68.50	2015	Dirichlet Mixture Model based Fisher Vector；单尺度
	Semantic FV[128]	68.80	2015	自然参数化的 GMM Fisher Vector；性能最好的三个尺度
	VLAD[92]	68.88	2014	VLAD；三个尺度
State-of-the-art	Improved Object Bank[15]	46.60	2014	众多预先训练的目标检测器
	DeCaF[107]	58.40	2014	Decaf；全局 CNN 特征
	FV + Bag of parts[132]	63.18	2013	GMM Fisher Vector；判别性部件检测器；部件共现
	Mid-level elements[131]	64.03	2013	中层视觉元素发掘

图 4.2 “village”场景的识别结果。第一行的所有图像是 CTV 分类正确而 CNN 特征错分为反例的结果。第二行至第四行，给出了该幅图像被 CNN 特征错分的类别对应的场景图像。比如第一行第一列的 *castle*场景图像，被 CNN 错分为 *castle*场景类别；基于第二行至第四行给出了 *castle*类别对应的场景图像，我们可以观察：*village*和 *castle*两个类别从外观来看，存在较大的相似性。

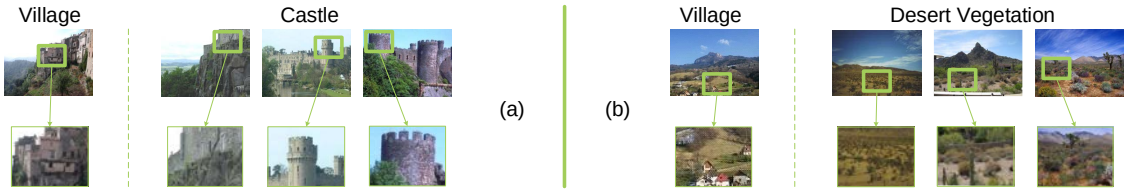


图 4.3 区域样例

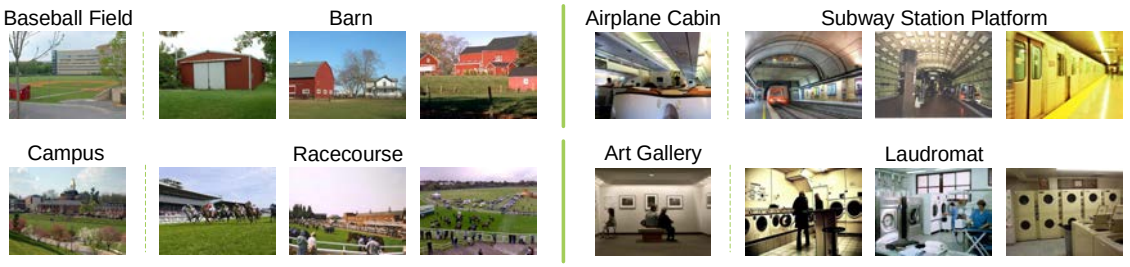


图 4.4 四个场景类别的识别结果。对于每一类，我们给出了一幅图像，且该图像是 CTV-GS 分类正确而 CNN 特征错分为反例的结果。

比较好的表现。CTV-I-VB 和 CTV-I-GS 分别比 DMM FV [128] 高出了 3.49% 和 3.35% 的分类准确度。尽管 DMM FV 和 CTVs 都是基于 Fisher Vector 的，但是 DMM FV 与 CTVs 两者之间的一个关键性的不同之处在于，DMM FV 并没有考虑主题之间的关联性（theme/topic correlations）。基于 GMM 的 Semantic FV [128] 主要是采用多项式参数向量的自然参数化（natural parameterizations）来提升 DMMFV 分类准确率。不过，它的性能依旧比 CTV-VB-I 和 CTV-GS-I 分别低了 1.55% 和 1.41%。即使是在四个尺度上串联的 Semantic FVs 的最好性能表现 53.0% [128] 也同样低于 CTV-Is。除此之外，VLAD 被认为是基于 GMM 的 Fisher Vector[128] 的一种近似。Gong 等人 [92] 在论文中指出，基于 CNN-I 的多尺度 VLAD 可以将分类性能提升到 51.98%。同样地，CTV-GS-I 和 CTV-VB-I 依旧超过了多尺度 VLAD。一般地讲，在这些最相关的方法中个，DMM FV 是建立在 DMM 之上，而 DMM 是假设主题（themes/topics）之间是相互独立的；VLAD 和 Semantic FV 是依赖于 GMM 的，而 GMM 是假设图像中的图像块是相互独立的 [115]，并没有主题关联性。总的来讲，这些与 CTV-Is 的对比结果验证了：（1）独立同分布的假设并不是总能适合于建模主题/语义，（2）主题/语义之间的关联性的引入，可以提升基于 Fisher Vector 衍生的特征对于识别任务的性能。

表4.3给出了我们所提出的 CTVs 在不同尺度上的实验结果。在每一个单尺度上，所提出的 CTV-VB-I 的分类准确度分别是：256×256 尺度上 44.30%、128×128 尺度上 50.08%、64×64 尺度上 47.00%。在每一个单尺度上的 CTV-GS-I 的性能，与 CTV-VB-I 是相差不多的。两者之间的差距最低仅仅是 0.48%，最高是 2.48%。正如上面提到的，对于两者的多尺度情况，这个差距缩小到了 0.14%。但是我们不能忽略的一点是，用于求解 CTM 的变分贝叶斯方法达到收敛时，耗费了比较长的时间，尤其是对于大规模数据集，比如 SUN 397。基于 GS 的求解方法 [121] 使 CTV-GS-I 的学习更加高效。总的来讲，在性能上，两者的分类性能是不相上下的；在计算量上，我们的 CTV-GS-I 比 CTV-VB-I 更加有效。

相比于 VLAD [92]，我们所提出的 CTV-VB-I 在 256×256 尺度上提升了 4.73% 的分类性能，在 128×128 尺度上提升了 4.74%，在 64×64 尺度上提升了 6.79%。类似地，CTV-GS-I 的分类表现超出 VLAD 平均有 4 个百分点。在 128×128 尺度上，CTV-VB-I 和 CTV-GS-I 优于 Semantic FV。在 256×256 尺度上，CTV-VB-I 表现得比 Semantic FV 好而 CTV-GS-I 取得了与 Semantic FV 可比的性能。

基于 PlacesNet 的实验结果。 为了更加充分地验证所提出的 CTVs，我们又基于以场景为中心（scene-centric）的 PlacesNet[69] 开展了实验。相比 object-centric ImageNet [77]，PlacesNet 是 scene-centric 的。正如表4.2所展示的，CTV-Ps 的性能明显比 CTV-Is 高出了大约 5 个百分点。CTV-VB-P 取得了 58.39% 的分类准

确度，CTV-GS-P 取得了 58.43% 分类准确度。与 CNN-P 相比，CTV-VB-P 和 CTV-GS-P 分别取得了 4.07% 和 4.11% 性能提升。

值得注意的一点是，基于不同的尺度，CTV-VB-P 和 CTV-GS-P 具有相同的性能表现趋势：尺度越小，性能越低，具体实验数据可见表4.3。呈现该趋势的根本原因可能是 PlacesNet 和 ImageNet 两者所学习到的语义之间的差异性。ImageNet 是 object-centric 的，因此，所学习到的 CNN 特征关注于高层 object-oriented 的语义/目标。这一点符合场景图像的一个事实：目标总是出现在场景图像中小的区域/尺度上。也正因为如此，小尺度上的 CTV-Is 的分类性能优于大尺度上的 CTV-Is，这是合理的。然而，PlacesNet 是 scene-centric 的，它所学习到的 CNN 特征是趋向于全局性的 scene-oriented 语义。尺度越小，图像块包含的场景级别的信息越少。因此，所提出的 CTV-Ps 在较大的尺度上性能更好，这也是合理的。

我们采用文献 [92] 发布的公开代码，采用 CNN-I 特征替换为 CNN-P 特征来构建多尺度 VLAD。在三个单尺度上，PlacesNet 与 ImageNet 两者学习到的语义差异性，已经严重地影响到基于 CNN-P 的 VLAD 的性能表现。从 256×256 尺度到 64×64 尺度，VLAD 性能下降的程度更大。相比于多尺度的 CTV-Ps，基于 CNN-P 的多尺度 VLAD 低了 6.7 个百分点。对比 CNN-I 情况，我们的基于 CNN-P 的 CTVs 较大程度上地超过了 VLAD。

为了进一步分析 CTVs，我们展示了在测试图像上 CTV-Is 的分类实验结果。图4.2的第一行是关于 *village*类别的识别样例，该场景图像中 *buildings*、*sky*、*trees*和 *rocks*是共现的。所提出的 CTV 利用从词共现（word co-occurrence）学习到的关联隐主题来描述该语义共现并缓解词义模糊问题。除此之外，这些图像是 CTV 正确分类的正例（true positive）而被 CNN 特征错分为反例（false negative）的结果。以图4.2中的第一行第一幅图像为例，该图像被 CTV-GS 正确地识别为 *village*场景类别，但是被 CNN 特征错误地识别为 *castle*场景类别。通过观察这幅图像所给出的 *village*类别和 *castle*类别的场景图像，*buildings*、*sky*和 *trees*在两个场景类别中总是一起出现，这使得两类场景图像外观上比较相似。该类间相似性问题为场景图像特征表示带来了巨大的挑战。对于我们提出的 CTV 特征来讲，它其中关于全局隐参数 β 的部分，从根本上推动着 visual words 影响着 latent topic。也正是 word 与 topic 之间的影响作用，有助于识别场景类别之间的差异性，这是因为一个主题是受限于一幅图像特定属性的。图4.2中用绿色框标注出来的 building，在不同的场景中变化是巨大的，比如，*castle*、*abbey*、*construction site*、*slum*和 *kasbah*。图4.3给出了图4.2中第一列和倒数第二列的两幅 *village*场景图像的区域。我们可以清晰地看到，这两幅图像中所标出的区域，存在比较大的差异。除此之外，图4.4也展示了来自其他类别的四个样例。这四个来自四个类别的图像，对于 CTV 来讲是正例，对于 CNN 特征是被错分的反例（被 CNN 特征

错误地识别为其他类别）。在图4.3（a）中，*village*场景图像被 CNN 特征错分为*castle*场景类别，但是从给出的三幅 *castle*场景图像可以看出，对于 *castle*场景中典型的 *castle*目标局部区域，他们与 *village*场景图像的典型区域 cottage 目标局部区域，存在很大的不同。同样的情况，也存在于图4.3（b）中 *village*场景类别与*desert vegetation*场景类别之间。所以我们借鉴 Fisher Kernel 的思想，采用主题与视觉词之间的关系来捕获外观相似的两个场景类别之间的差异性。也正是基于此观察，启发着我们对 CTV 的探究。

MIT Indoor 67 数据集。 表4.4展示了 CTV 在 MIT Indoor 67 数据集上的主要实验结果。与 SUN 397 类似，我们也主要从两个方面对比了所提出的 CTVs：（1）与 baseline CNN 特征和 Fisher Vector 衍生的最相关的方法的对比，（2）与其他 state-of-the-art 方法的对比。对在表4.4中所给出的相关方法中，DMM FV [128] 和 Sparse Coding FV [134] 是单尺度的，而 VLAD [92]、Semantic FV [128]、CTV-GSs 和 CTV-VBs 是多尺度的，具体地讲，是三个尺度。

基于 ImageNet 的实验结果。 除了 Sparse Coding FV [134] 采用第六层的 CNN-I 特征作为描述子，第一组对比方法采用的是 CNN-I 的 FC7 特征作为描述子。在这些相关的方法中，CNN baseline 的分类准确度是 56.79% [69]。CTV-VB-I 取得了 68.88% 的分类准确度，CTV-GS-I 获得了 68.36% 的分类表现。与 SUN 397 数据集实验结果类似，我们也在 MIT Indoor 67 数据集上得出相同的结论：CTV-VB-I 和 CTV-GS-I 的性能表现相差不大，是可比较的，并且都很大程度上地超出了 CNN-I baseline。只不过，前者识别结果更准确，而后者的时间复杂度更低。另外，所提出的 CTV-Is 与 VLAD、DMM FV、Semantic FV 和 Sparse Coding 是可比较的。Sparse Coding FV [134] 是在局部 CNN-I 特征的基础上，采用 Sparse Coding 模型来构建 Fisher Vector 的。与 Semantic FV 和 VLAD 不同的是，Latent GMM FV 方法 [115] 为 GMM 模型的混合权重（mixing weights）参数加上了一个狄利克雷先验分布。当 Latent GMM FV 采用与我们类似的密集网格采样（dense grid sampling）的图像块的方式时，Latent GMM FV 获得了 65.0% 的准确度。由于狄利克雷先验分布的缘故，Latent GMM 模型明显地指出，它其中的每一个单 Gaussian 之间是互相独立的，而这些单 Gaussian 模型可看做是一个聚类中心或者一个主题/语义。总得来讲，Latent GMM FV 并没有把这些 Gaussian 之间的关联性考虑进去。相反地是，所提出的 CTV-Is 在保证 Fisher Vector 的优良特性的同时，考虑了主题/语义之间的关联性。从实验结果可以明显看出，CTV-VB-I 超出 Latent GMM FV3.88 个百分点的分类准确度。因此，独立性的假设对于刻画场景图像的语义太过严格了。

另外，我们也评测了不同尺度上 CTVs 的场景图像识别性能表现，具体结果可见表4.3。在 256×256 尺度上，CTV-VB-I 取得了 59.78% 的分类准确度，超出了 VLAD 6.06 个百分点。CTV-GS-I 取得了 58.88% 的分类准确度，超出 VLAD

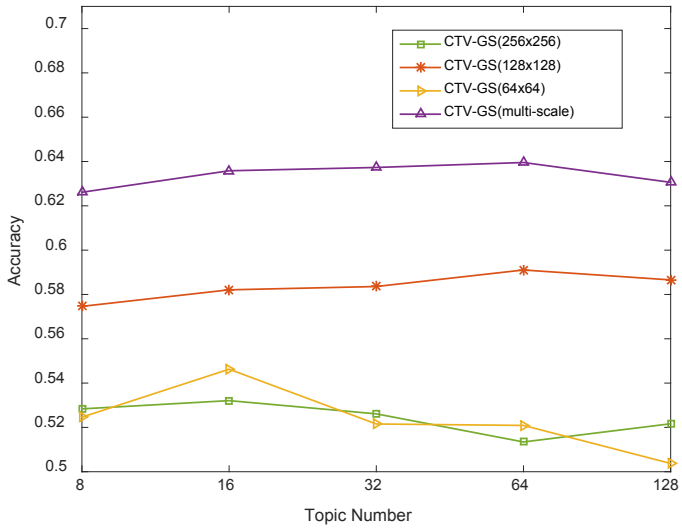


图 4.5 关于主题数目的评估

5.15 个百分点。在该尺度上，CTVs 是从全局图像上编码特征的，而不是把图像裁剪成图像块。对于室内场景图像来讲，他们总是呈现出复杂的目标组织结构，进而导致较大的类内变化性；如果采用裁剪图像块的方式来编码特征，可能会减小类内变化性，因为局部图像块一般都包含比较少数量的目标，减少了图像的复杂度。

基于 PlacesNet 的实验结果。 基于 CNN-P （PlacesNet），所提出的 CTVs 获取了高达 73.88% 的分类准确度，具体结果可见表4.3。与 SUN 397 实验结果相似，CTV-Ps 在单尺度上更倾向于在较大的尺度上有更好的性能表现。CTV-VB-P 在 256×256 尺度上取得了 70.90% 的分类准确度。相比于 128×128 尺度上的分类结果，256×256 尺度取得了 2.17% 的性能提升；相比于 64×64 尺度上的分类结果，256×256 尺度取得了 12.17% 的显著性能提升。这样的实验结果是与 SUN 397 数据集类似的，而这其中的根本原因也与上面讨论的相同。

4.6.2.3 模型评估与参数评估

学习到的主题与主题之间的关联性讨论 所学习到的主题/语义融合了类别特定性的关联关系，他们蕴含的丰富语义可以增强表示类内的相似性。再加上 Fisher Vector 方法对于类间判别能力的增强，CTV 对于场景分类任务表现出了巨大的潜力。

主题数目的评估: 主题数目的评估是在 MIT Indoor 67 数据集上进行的，具体实验结果见图4.5。当主题数目从 8 到 128 增大时，多尺度 CTV-GS 的分类性能变化得很小。这一定程度上表明了主题数目对于 CTV 并不是一个主要因素。这是因为，CTVs 源于隐关联主题/语义，而有限的主题数目已经足够在 Fisher Kernel 空间中捕获到一幅图像的主题与词之间的微小差异。

模型求解算法的评估：我们针对两种不同 CTM 求解算法所得到的 CTVs 特征进行实验分析：CTV-VB 和 CTV-GS。在表4.2、4.3和4.4中，可以观察到，无论是多尺度还是单尺度，CTV-GSs 总是与 CTV-VBs 相差不多。例如，在三个单尺度上，对于 MIT Indoor 67 数据集，CTV-VB-I 与 CTV-GS-I 之间的差距，平均只有 0.70 个百分点；对于 SUN 397 数据集，该差距平均只有 1.38 个百分点。在多尺度上，对于 MIT Indoor 67 数据集，这个差距降到了 0.52%；对于 SUN 397 数据集，这个差距更是降到了 0.06%。这些可忽略的性能差距表明了 Gibbs Sampling 对于变分贝叶斯算法是一个很好的近似。

4.7 本章小结

在本章中，为了在隐语义的基础上获得优异的场景图像，我们结合生成式模型和判别式模型，提出了隐主题向量特征表示。更进一步，为了去除局部图像块之间的独立同分布假设，采用 logistic 正态先验分布建模隐语义关联关系，并提出关联主题向量 CTV。CTV 特征表示的实现不仅结合了蕴含丰富语义信息的 CNN 特征，同时也探究了潜在的关联性语义，并将其编码于 Fisher Vector 框架中, 以此来提升表示的判别能力。为了能让所提出的方法更适合大规模数据集，我们进一步给出了变分贝叶斯求解和吉布斯采样求解的 CTV 实现。在大规模数据集上的实验验证了 CTV 的有效性，并展示出其对 CNN 特征的较大性能提升，对基于深度特征的 Fisher Kernel 表现出巨大的潜力。总之，与 GMM 系列的 Fisher Vector 和 LDA 系列的 Fisher Vector 一起，我们所提出的 CTV 为图像语义表示构建了一个更加完备的生成式模型。

第五章　基于隐目标挖掘的场景分类

场景图像的复杂性主要体现在不同种类、不同数目、不同尺度的目标分布在场景中不同的位置，造成场景图像较大的类内差异性和类间相似性问题。依赖目标检测、图像分割的方式尽量识别出场景中所有目标的策略，不仅需要处理目标自遮挡和相互遮挡、尺度多样、种类多样、实例数量多样等挑战性问题，而且这些目标的检测或者分割的一个假设提前是目标都是相互独立存在于场景中。现有目标检测器只覆盖了极有限的目标类别，无法满足对场景图像中可能存在的各类刚性或者非刚性目标的检测，也无法满足场景图像丰富特征表示的需求。在本章中，我们提出隐目标（Latent Object，LO）来表示场景的语义信息，无需任何图像分割、人工标注或大量预训练目标检测器；结合全局和局部的学习方式，提出隐目标发现（Latent Object Discovery，LOD）的方法，自适应地发掘场景图像中具有判别性、表示性的区域，并系统化地组织隐目标来构建场景特征表示。

5.1　引言

理解视觉场景图像是识别、推断我们周围世界的途径之一。一个“场景”描述的是一个充满了丰富实体（比如，天空，沙滩，树木，桌子，门等等）的一个地方。不同种类、大小、尺度和位置的实体目标造成了非常大的场景外观不确定性。考虑到场景由目标组成，许多研究工作致力于挖掘那些常见的目标。该策略得益于在现有公开数据集（如 PASCAL 和 ImageNet）上已经训练好的目标检测器。但是，这些数据集覆盖的目标类别非常有限，比如 PASCAL 目标检测数据集只包含了 20 类，ImageNet 目标图像数据集包含了 200 类。在场景图像中，如图5.1所示，比如天空、沙滩和沙漠，是没有明确的、规则的目标边界，所以现有目标检测器所发掘出的区域，并不是总是具有表示力的。场景图像中的目标存在严重的自遮挡和相互遮挡、尺度多样、种类多样、实例数量多样等挑战性问题，这就造成检测场景图像中的目标面临非常大的挑战。即使目标检测器有着优异的性能表现，场景中的目标是被独立检测到的；这些检测到的目标个体无法满足对复杂场景的描述。不同类别的场景类别很可能会出现相同类别的目标，那么目标检测器很可能在这些不同类别场景图像中输出相似区域，无法满足描述场景类别差异性的需求。

我们提出隐目标用于场景识别，无需任何标注和预训练的常见目标检测器。隐目标既可能是那些常见的实体目标又可以是实体目标混合体，或者是目标部件，或者是一些纹理化的、颜色丰富的图像区域。LO 不是仅限于这些常见的目标类别（如 PASCAL 中 20 类目标），而是为场景图像发掘显著的、主导性的图像区

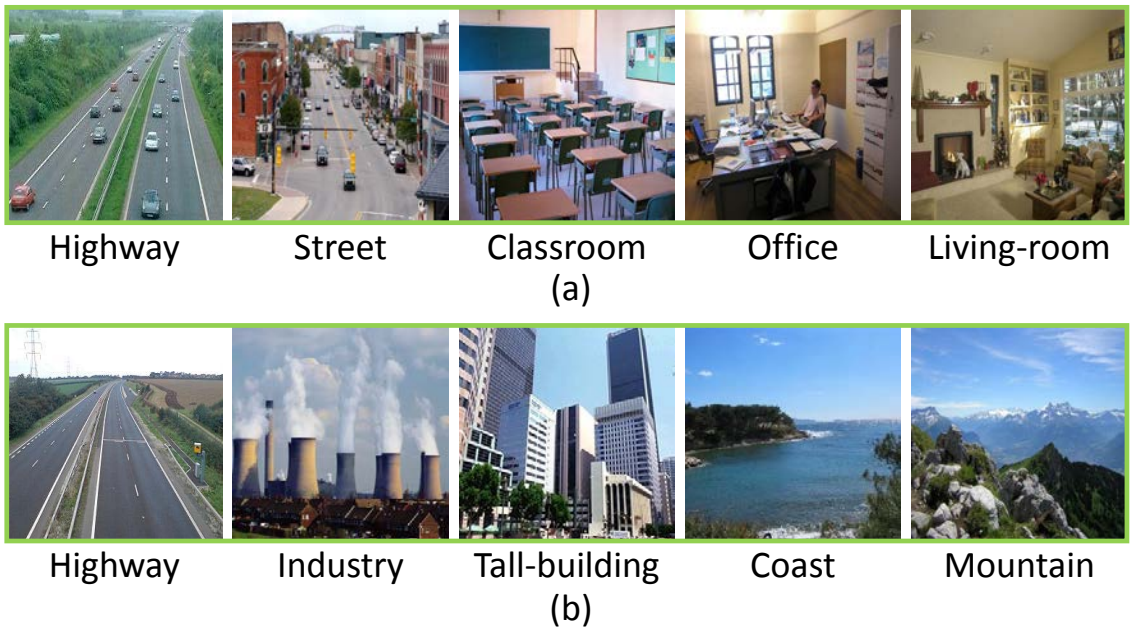


图 5.1 场景图像示例。(a) 场景图像样例 –包含了一些常见的目标，比如，汽车桌子、瓶子、沙发、人、电脑键盘、狗、书架等等，这些目标是现有的目标检测器尽力去学习和检测的对象；(b) 场景图像样例 –现有的目标检测器在这些图像中很可能失效，因为这些场景图像很难被分解为有界的、边界清晰的目标。比如，在“海岸”场景中，海水呈现的是非刚性边界的特性。对于“高速”场景，按照常识，我们可以检测车辆目标；但是“高速”场景中，没有任何车辆的情况也时有发生。因此依赖于众多预训练的显式目标检测器的场景识别策略，需要面临极其庞大的检测计算量和极其复杂的目标检测挑战（比如，目标遮挡、不同目标尺度、不同目标大小等）。

域。不同于现有的显著性研究工作，LO 趋向于那些具有判别性、表示性的区域，并且是与类别相关的。

我们将隐目标发掘建模在一个最小化熵的框架中，采用全局到局部、局部到全局的方式进行模型学习。隐目标发掘不依赖于预先定义的或者预训练的实体目标检测器，它是采用最小熵判别函数来训练一个场景特定的隐模型，用来自动地、自适应地为每一幅场景图像发掘一组语义区域。通过 LO 对图像级别分类的共线性来加权隐目标，从而实现自适应局部特征编码。隐目标挖掘的主要贡献概括如下：

1. 提出最小化熵隐目标发掘策略，用来提取具有表示性和判别性的图像块。

2. 提出全局和局部相结合的学习框架，用来学习基于最小化熵的隐目标。

5.2 相关工作

关于人类感知的研究指出，人类可以快速识别一个场景而无需首先识别出场景中的目标 [17]。基于此研究发现，通常，场景识别的常用策略是获得全局的场景表示。基于此策略的方法是从一整张图像进行学习，并且生成全局描述子 [17–20]。通常，这些方法是建立在 BoW 特征之上的。由于粗糙的局部向量量化，BoW 势必会造成一词多义、一义多词的问题，从而损害场景识别的性能。考虑到场景可以被解析为目标的组合，因此场景识别的第二种策略是以目标为中心的方法，学习场景的局部信息。相关的工作大致可分为两类：基于预训练目标检测器的方法 [74–76] 和基于局部区域学习的方法 [72, 93–97]。

基于目标检测的方法认为特殊的目标类别促进高层语义的挖掘。该方法的一个隐含假设是场景图像的所有类别共享一套预先定义好的目标类别。因此该类方法试图标注区域并且训练相应的目标检测器。这样的检测器可以采用在额外的目标图像数据集（如 PASCAL[10]，ImageNet[89]）上预先训练的 DPM（Deformable Part-based Models，DPM）[44] 或者 RCNN 系列方法 [73–76]，只不过这些额外的数据集只包含了非常有限的目标种类。基于检测到的或者标注的区域，可以为图像识别任务目标特征提取或者特征编码。在文献 [66, 67] 中，李飞飞等人提出“Object Bank”方法，采用在 12 个尺度和 21 个空间金字塔网格上检测 177 种目标，获得每一个目标出现在每一个像素处的概率。这样的以目标为中心的方法倾向于寻找经常出现的一般性目标，但是这些方法不仅检测计算代价大，而且需要解决目标训练域和目标域的差异性问题。Wu 等人 [72] 提出生成 Meta 目标和一系列共享检测器，采用基于非监督的聚类方法和弱监督的学习方法来解决这些问题。不过，这些方法都过多地寻找不同类别场景图像中都可能存在的目标，忽略了不同场景图像的差异性。

基于局部区域学习的方法致力于学习一个判别块集合，并将它们用于中层语义表示。在文献 [93] 中，作者沿用 DPM[44] 的思想，将一个场景视为一种特殊的目标，该目标部件的空间位置在学习过程中相对于真实目标框而进行调整，并且还能带来局部块的空间对齐。然而，由于极大的外观变化性和复杂的块空间分布，相对于目标来讲，与 DPM 类似的方法对于表示场景过于规则化。近年来的工作 [72, 93–97] 更倾向于弱监督学习（比如，隐支持向量机）或者非监督学习（比如，判别聚类），以此来发掘中层局部块。Singh 等人 [93] 提出采用非监督判别聚类来挖掘一个判别块集合，并且这些判别块可以用于完全非监督中层语义表示。但是该方法缺少更进一步的特征编码策略。Durand 等人 [78] 引入了最小 -最大化隐结构 SVM（min-max latent structural SVM，MANTRA），对于一个类别的场景，组合局部块的最高检测得分和最低分，获得该类别的预测得分。隐金字塔区域（Latent Pyramidal Regions，LPR）[79] 为局部区域构建空间金字塔特征

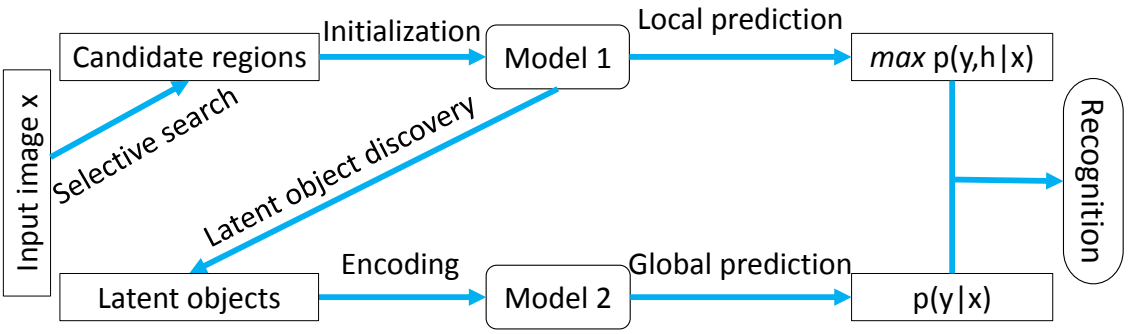


图 5.2 隐目标挖掘流程图

表示，借助于隐变量模型 Latent SVM 从众多区域中选择得分最高的区域并预测场景标号。这些方法仍旧都过多地关注图像中局部区域，忽略了全局信息以及全局与局部信息两者之间的关联关系。在本文中，我们采用隐模型学习来发掘隐目标，在最小化熵准则下结合局部和全局策略，学习场景图像的整体信息和局部隐目标。考虑到目标域场景的关系，从全局到局部的角度来看，一个场景可能被分解为“目标”，激励发掘场景图像照片中的那些具有表示力的“目标”或局部区域；从局部到全局的角度来看，“目标”被结构化地组织为一个场景，激励那些被发掘的“目标”更具有判别性地解释一个场景图像。自适应地学习和发掘的隐目标，避免了传统构建场景特征表示固定无需大量预训练目标检测器用于检测有限类别的目标。

5.3 隐目标挖掘

建模 假设训练集是 $\mathcal{D}=\{(x_1,y_1),...,(x_N,y_N)\}\in(\mathcal{X}\times\mathcal{Y})^{\mathcal{N}}$。对每一对 (x,y)，它们都有一个隐变量 h 的集合与之相关联。其中 $x\in\mathcal{X}$ 是一个场景图像，它的场景类别标号是 $y\in\mathcal{Y}$。$h\in\mathcal{H}$ 代表了隐目标可能的位置。模型输入 x 是可观测的变量，其值在训练阶段和测试阶段都是已知的；模型输出 y 是不可观测变量，其值只有在训练阶段是已知的；隐变量 h 的值在训练阶段和测试阶段都是未知的。变量 w 是待求的线性判别隐模型的参数。给定输入图像 x，输出变量和隐变量的联合条件概率被记作集合 $P_x=\{P(y,h|x;w),\forall(y,h)\in\mathcal{X}\times\mathcal{Y}\}$。给定输出标号 y，隐变量的条件概率被记作集合 $P_x^y=\{P(h|x,y;w),\forall h\in\mathcal{H}\}$。一般情况下，$Q_x^y=\{P(y,h|x;w),\forall h\in\mathcal{H}\}$ 是一个广义分布代表，本质上是分布 P_x 的一个子集，即输出 y 所对应的子集。与 P_x 不同的是，Q_x^y 不需要加和等于 1。AD 熵（Aczél and Daróczy，AD）[98] 被用来评估隐变量的不确定性。AD 熵是瑞利熵（Renyi Entropy）的泛化，它的定义是 $H_{r,s}(Q_x^y;w)=\frac{1}{1-r}log\left(\sum_h P(y,h|x;w)^{r+s-1}/\sum_h P(y,h|x;w)^s\right)$，其中 $r\geq 0,r\neq 1,s\geq 0,r+s\geq 1$。

我们采用最小化熵策略，发掘判别力强的隐目标，目的在于在降低隐变量不确定性的同时获得尽可能准确的输出标号预测。文献 [99, 145] 论证了关于 AD 熵的一个特性：关于 y 的广义分布的 AD 熵可以写成关于 y 的负对数似然与关于已知 y 时隐变量条件概率的 AD 熵之和：$H_{r,s}(Q_x^y;w) = -\log P(y|x;w)+H_{r,s}(P_x^y;w)$。从此式可看出，一方面，最小化 $H_{r,s}(Q_x^y;w)$ 意味着最大化关于 y^* 的对数似然，即令最优解 y^* 的概率最大；另一方面，最小化 $H_{r,s}(Q_x^y;w)$ 意味着最小化关于隐变量的不确定性。所以我们可以通过最小化该 AD 熵来预测输出标号，达到在挖掘确定性高、判别力高的隐目标的同时获得对于输出标号更准确的预测的目的：

$$y^* = \arg\min_y H_{r,s}(Q_x^y;w). \tag{5-1}$$

对于一个学习到的模型参数 w，给定一个场景图像 x，关于输出变量与隐变量的联合概率被定义为

$$P(y,h|x;w) = 1/Z(x;w)\cdot\exp\left(w^T\phi(x,y,h)\right) \tag{5-2}$$

其中 $Z(x;w)$ 是配分函数（partition function），保证概率之和被归一化为 1；对于图像 x 的一个可能的隐目标 h，$\phi(x,y,h)$ 是其联合特征向量。

通过分析公式 (5-1)，相比传统不含隐变量的判别模型，该基于最小化熵的隐变量模型在解决分类问题时不一定有太多优势。主要原因在于直接最小化 AD 熵的结果是在降低隐变量不确定性的同时提升预测的准确率，也就是隐变量的引入增加了模型预测的难度：预测的结果需要保证隐变量和输出标号两个量的最优。但是对于分类问题，我们主要关注的是尽可能输出最优标号预测。为了解决这一问题，我们提出全局和局部相结合的隐目标挖掘方法。从全局到局部，期望在图像全局标号（场景类别）的指导下，引入隐变量，学习模型趋向挖掘场景图像中关键性视觉元素，即显著性、判别力强的隐目标；从局部到全局，期望利用挖掘的隐目标构建更具有表示力和判别力的特征表示，进而提升分类模型的预测能力。基于此观点，我们提出全局和局部相结合的最小化熵隐模型学习方法：

$$y^* = \arg\min_y HL(Q_x^y;w) + HG(Q_x^y;w). \tag{5-3}$$

其中 $HL(Q_x^y;w) = H_{\infty,1}(Q_x^y;w) = -\log\max_h P(y,h|x;w)$，$HG(Q_x^y;w) = H_{2,0}(Q_x^y;w)$ $= -\log\sum_h P(y,h|x;w) + K$，$K$ 是一个常量。因此，我们预测输出标号 y：

$$\begin{aligned} y^* &= \arg\max_y\{\max_h p(y,h)\textstyle\sum_h p(y,h)\} \\ &= \arg\max_y\{\max_h p(y,h)p(y)\}. \end{aligned} \tag{5-4}$$

很明显，场景标号是从两个角度被预测的：(1) 局部区域的最大响应，(2) 全局响应。可预测的信息通过全局和局部方式结合在一起。

模型学习 对于公式 (5-3)，当只考虑其中的 $HL(Q_x^y;w)$ 部分时，$HL(Q_x^y;w) = -\log\max_h P(y,h|x;w)$，是最小熵（Minimum Entropy）[99]。相应地，给定模型参数 w_1 和联合特征 ϕ_1，公式 (5-1) 的预测，等价于给定输入变量 x 条件下的关于 y 的联合最大后验推断，

$$\begin{aligned}(\bar{y}(w_1),\bar{h}(w_1)) &= \arg\max\nolimits_{(y,h)\in Y\times H} P(y,h|x) \\ &= \arg\max\nolimits_{(y,h)\in Y\times H}\ w_1^T\phi_1(x,y,h).\end{aligned} \tag{5-5}$$

这样，我们就获得了关于 (y,h) 对的最优预测结果，接下来通过简单地抛掉关于 h 的部分，就得到了关于 y 的预测值。但是，事实上，这个关于 (y,h) 的最大后验预测结果 “过于乐观”，确定性地把隐变量赋值为 (y,h) 概率达到最大时 h 的状态、把输出标号赋值为 (y,h) 概率达到最大时 y 的值。该预测标号 y 的方法对于隐变量 h 的扰动非常敏感，它甚至对于简单的情况也很可能引入比较大的偏差。因此该联合最大后验预测 (y,h) 的最优结果并不一定是关于 y 的全局最优预测。这也激发着我们引入全局模型。

对于公式 (5-3)，当只考虑其中的 $HG(Q_x^y;w)$ 部分时，$HG(Q_x^y;w) = -\log\sum_h P(y,h|x;w) + K$，等价于对于隐变量的边际化 [99]。给定模型参数 w_2 和联合特征 ϕ_1，公式 (5-1) 的预测，等价于边际化最大后验预测 [99]，

$$\begin{aligned}\bar{y}(w_2) &= \arg\max_{y\in\mathcal{Y}}\sum\nolimits_h P(y,h|x) \\ &= \arg\max_{y\in\mathcal{Y}}\log\sum\nolimits_h \exp\left(w_2^T\phi_1(x,y,h)\right),\end{aligned} \tag{5-6}$$

很明显，该式子显式地考虑了隐变量的不确定性。另外，根据杰森不等式（Jensen's Inequality）和 $\log(\cdot)$ 函数的非凸性，我们推导出 $\log\sum_h \exp\left(w_2^T\phi_1(x,y,h)\right)$ 的下界，

$$\begin{aligned}\log\sum\nolimits_h \exp\left(w^T\phi(x,y,h)\right) &\geq \sum\nolimits_h w_2^T\phi_1(x,y,h) \\ &= w_2^T\sum\nolimits_h \phi_1(x,y,h) \\ &= w_2^T\phi_2(x,y,h).\end{aligned} \tag{5-7}$$

公式 (5-7) 将隐目标整合在一起，从单个局部区域特征表示 ϕ_1 获得一个全局特征表示 ϕ_2。简单来看，公式 (5-7的本质是，对于一个隐变量 SVM 模型，序列边际化隐变量等价于学习一个线性 SVM 模型，无需直接学习复杂的、涉及大量关于隐变量的优化计算和 *log-sum-exp* 操作的目标函数。

总的来讲，我们最终的目标是最大化输出变量 y 的后验概率 $P(y|x;w)$。基于 $HL(Q_x^y;w)$ 和 $HG(Q_x^y;w)$，隐变量的后验概率可表示为，

$$\begin{aligned}&\log P(h|y,x) = \log\left(P(y,h|x)/P(y|x)\right) \\ &\propto w_1^T\phi_1(x,y,h) - \log\sum\nolimits_h \exp\left(w_2^T\phi_1(x,y,h)\right) \\ &\leq w_1^T\phi_1(x,y,h) - w_2^T\phi_2(x,y),\end{aligned} \tag{5-8}$$

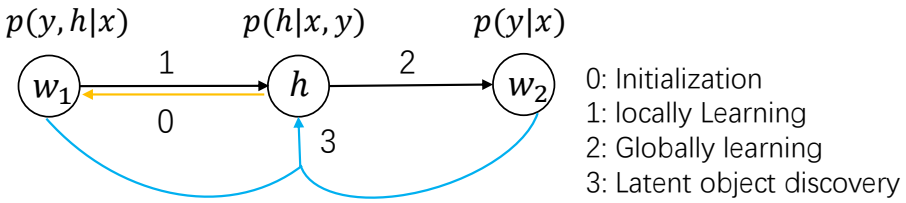


图 5.3 学习过程

$$P(h|y,x) \propto \exp\left(w_1^T \phi_1(x,y,h) / \exp w_2^T \phi_2(x,y)\right). \tag{5-9}$$

不难理解，隐变量的概率可以由一个 Softmax 函数近似获得。该 Softmax 函数依赖于全局和局部模型。局部模型关注于发现类别特定的隐目标，而全局模型关注于基于发掘的隐目标的编码特征来进行场景识别。在本章中，我们采用基于 GMM 的 Fisher Vector 特征编码方法。

在一个最大边界的建模中，模型参数 $w = (w_1, w_2)$ 的学习通过最小化下面的目标函数获得：

$$\begin{aligned} &\min_{w,\xi} \frac{1}{2}\|w\|^2 + \frac{C}{N}\sum_{i=1}^{N}\xi_i \\ &s.t.\ H(Q_{x_i}^{y};w) - H(Q_{x_i}^{y_i};w) \geq \Delta(y_i,y) - \xi_i, \\ &\quad \forall y \neq y_i, \forall (x_i,y_i) \in \mathcal{D}, \end{aligned} \tag{5-10}$$

其中 $H(Q_x^y; w) = HL(Q_x^y; w) + HG(Q_x^y; w)$；$\Delta(y_i, y)$ 是 0-1 损失，即：如果 $y_i = y$，$\Delta(y_i, y) = 0$，否则等于 1。

具体地，在参数学习阶段，首先学习参数 w_1 的局部模型，即 Model 1，主要用来发掘隐目标；接着学习参数为 w_2 的全局模型，即 Model 2，用于为图像识别任务而系统性地组织这些隐目标。同时，在全局局部模型的影响下，这些隐目标的置信度会被更新，而被用于下一次迭代学习中。参数的学习过程可参见图5.3。

模型实现 在实现过程中，采用 Selective Search[104] 策略得到局部区域，这些区域为隐目标提供初始候选集。基于 Places 205 数据集 [69] 预训练的 Alex 网络的第七层特征将作为这些区域的局部特征描述子。在训练的第一阶段阶段，Model 1 选择隐目标是依据候选区域的概率从大到小排序，按照预先设定的比例值删除掉概率低的区域，获得隐目标此阶段筛选的阈值。经过筛选，每一幅图分别获得不同数量的隐目标。不过此时的隐目标的发掘过程中并没有考虑全局信息。在第二阶段，我们采用 Fisher Vector 编码方法将图像的隐目标编码为一个全局特征向量。假设第一阶段图像 x 发掘了 T 个隐目标 $\{xx_t, t = 1, ..., T\}$，并且这些隐目标符合参数为 $\lambda = \{\alpha_i, \mu_i, \Sigma_i, i = 1...N\}$ 的 GMM 模型，其中 α_i、μ_i 和 Σ_i 分别是

GMM 中第 i 个高斯模型的权重、均值和协方差参数。GMM 的参数从所有训练图像第一阶段获得的隐目标中学习得到。由于参数 α 分量作用比较小 [118]，Fisher Vector 仅由 μ 和 Σ 对应的分量串联构成。更多细节可以参考文献 [118]。不过与文献 [118] 不同的一点是，在 LOD 中，构建 Fisher Vector 的图像块/区域数目不再是对于所有图像都相同，也就是 T 对于不同图像可以是不同的。对于包含了不同种类、不同数目、不同尺度的实体实例的复杂场景图像，不等数目的隐目标发掘的做法更加合理和灵活。经过第二阶段，隐目标将根据公式 (5-9) 进行更新并用于下一轮的模型迭代学习。

5.4 隐目标挖掘在弱监督目标检测中的应用

对于目标检测任务，监督训练检测模型需要对图像中的目标进行精确定位标定。但是由于标定工作量大，耗时耗力，尤其是对于大规模数据集，这为目标检测的实际应用带来了很大的负担。为了缓解该问题，弱监督目标检测（Weakly Supervised Object Detection，WSOD）近年来被广泛研究。WSOD 是在缺少目标在图像中精确位置、只有图像级别的目标标号的条件下训练检测模型的。由于标定数据比较弱，如何得到可靠的 WSOD 模型就更加具有挑战性。

为了训练检测模型，WSOD 需要为检测器获取可靠的目标样本。所以解决 WSOD 问题的两个关键问题是：如何收集目标样本和如何基于收集到的目标样本获得目标检测器。现有 WSOD 研究中，一般采用 Selective Search 等方法提取候选框作为目标样本候选集。但是一幅图可能会产生上千个候选框，如此大的求解空间为检测模型的求解带来庞大的计算量。对于上一节提出的 LOD 模型，目标样本的收集过程等价于隐目标的挖掘过程。所以，为了解决 WSOD 问题，利用在上一节提出的 LOD 模型，把目标样本用隐变量来表示，在学习的过程中不断地调整隐变量，逐步定位到图像中最有可能是正例的那一部分样本。不过 LOD 模型在解决 WSOD 问题时，考虑到场景分类与 WSOD 两者之间的差异性，还需要有做一点调整。表5.1对比了基于隐变量模型在解决 WSOD 与场景分类问题的不同。两者的第一个不同在于输出变量 y 的级别不同：WSOD 中 y 是目标类别，即局部标号；场景分类中 y 是场景类别标号，即全局标号。两者的第二个不同在于用于模型训练的反例来源的图像存在差异：WSOD 的反例来自于正例图像和其他类别的图像，而场景分类的反例只来自其他类别。由于 LOD 模型结合了全局和局部的学习方式，所以 LOD 模型可以直接解决第一点不同。对于第二点不同，主要来自于 WSOD 任务求取局部区域的特殊性，所以为了解决该问题，我们在 LOD 的基础上增加了关联抑制和部件抑制，去除关联目标和目标局部部件的干扰作用。本文所提出的 WSOD 学习过程是：首先对输入图像进行关联抑制，关联抑制的流程是先计算训练图像中的目标之间的关联性，然后根据关联性采样图像

表 5.1 弱监督目标检测与场景分类问题的对比

建模		弱监督目标检测任务		场景分类任务	
任务		给定目标标号，检测目标		给定场景标号，识别出图像所属的场景类别	
变量	x	图像	全局	图像	全局
	y	目标类别	局部	场景类别	全局
	h	目标可能的位置	局部	目标可能的位置	局部
训练	正例	只来自本类别的图像		只来自本类别的图像	
	反例	来自正例或其他类别图像		其他类别图像	
预测模型		$\{y^*, h^*\} = \arg\max_{y\in Y, h\in H} p(y,h\|x)$		$y^* = \arg\max_{y\in Y} p(y\|x) = \arg\max_{y\in Y} \sum_h p(y,h\|x)$	

数据集，得到去关联的图像，最后对图像提取候选框和特征，再进行 LOD 学习得到分类模型，之后进行部件抑制。为了抑制部件，需要预测出部件可能在的位置，然后使用预测到的部件学习最终的模型。

5.5 实验验证与模型分析

5.5.1 场景分类

本小节给出实验结果来评估所提出的方法的性能。评估的指标是平均分类准确度 [87]。所提出的方法在两个公共数据集上进行了实验验证：SCENE 15[53] 和 MIT Indoor 67[87]。SCENE 15 数据集有 15 个类别，在该数据集上的数据集设置与文献 [53] 相同。

5.5.1.1 主要实验结果

首先给出所提出的方法与 state-of-the-art 方法的对比结果。如表5.2所示，在 SCENE 15 数据集上，所提出的方法获得了最好的性能：94.65%。如表5.3所示，在 MITIndoor 67 数据集上，所提出的方法的性能优于大部分现有方法，只比多尺度的 MANTRA[78] 和多尺度的 MetaObject[72] 低了一点，但是依旧获得了非常有竞争力的性能。

对模型进行更深入的实验分析。实验主要在 SCENE 15 数据集上进行了验证，具体实验结果见表5.4。一方面，如果去除掉所提出模型的全局模块，所提出的模型可以简单地看作是 latent SVM，此时模型只得到了 88.16% 的识别结果。相比结合了全局和局部的 LOD，这个结果低了 6.49%。主要原因是，局部模型，

表 5.2 SCENE 15 数据集的实验对比

Method	**Accuracy(%)**
CNN-places [69]	90.19
CNN-hybrid [69]	91.59
SPM [53]	81.4
Discriminative Part Detectors [96]	86.0
Object Bank [67]	90.2
MANTRA (single-scale) [78]	80.7
MANTRA (multi-scale)[78]	93.4
Ours	94.65

也就是 latent SVM，只是单纯地试图最大化隐变量 h 和标号 y 的联合概率，也就是 $p(y,h|x;w)$，无法保证两者的联合概率最大的时候标号 y 的概率 $p(y|x;w)$ 最大。这就是为什么 latent SVM 只关注局部区域而忽略全局场景上下文信息的原因。另一方面，如果移除掉所提出模型的局部模块，模型可以简单地看作是边际化隐 SVM（Marginal latent SVM，MLSVM），此时模型只有 88.25% 的识别性能。MLSVM 等同于取消了局部区域的影响，通过对变量 h 来边际化隐变量 h 和标号 y 的联合概率；然后 MLSVM 就可以最大化关于变量 y 的概率 $p(y|x;w)$。但是 MLSVM 忽略了局部区域对场景的共享作用，它势必无法捕捉决定性作用的、判别力强的视觉细节元素。这也解释了相比于所提出的全局和局部相结合的 LOD 方法，为什么 MLSVM 下降了 6.4% 的性能。从公式 (5-7) 来看，如果令 $w_1 = w_2$，所推断出来的贝叶斯公式 (5-9)，是一个用于隐目标发现的 Softmax 函数。它获得了 93.73% 的识别性能。另外，图5.4是模型对于隐目标数目的实验评估结果。

5.5.1.2 模型分析

图5.6给出了所提出的模型在 SCENE 15 数据集上发掘的隐目标。可以观察到，所发现的隐目标具有明确的特定类别的特性，比如，“卧室”场景中的隐目标总是指包含了床、枕头和台灯的结构化区域；而“living-room”场景中的隐目标总是指那些包含了沙发、窗户和桌子的结构化区域。这两种场景呈现的特定类别属性的情况是非常不同的。所提出的模型可以学习出不规则的、灵活度高的场景视觉元素。比如，在“coast”场景中，隐目标是海水和天空的区域；在“tall building”中，隐目标是那些包含若干座高楼实例的区域。对于目标的形状，这些挖掘到的隐目标可以是具有明确边界的刚体目标，也可以是灵活度高的非刚体目标。对于目标的类别，隐目标可以是常见的目标，也可以是其他目标类别，比如

表 5.3 MITIndoor 67 数据集的实验对比

Method	**Accuracy(%)**
CNN-places [69]	68.24
CNN-hybrid [69]	70.80
SPM [53]	34.4
Discriminative Patches++ [93]	49.40
Discriminative Part Detectors [96]	51.4
FV+Bag of Parts [94]	63.18
Mid-level Elements IFV [95]	66.87
Object Bank [67]	68.2
MOP [92]	68.24
MANTRA (single-scale) [78]	56.4
MANTRA (multi-scale)[78]	76.6
MetaObject [72]	78.9
Ours	75.77

表 5.4 SCENE 15 数据集的模型对比

Method	**Accuracy(%)**
Local (LSVM)	88.16
Global (MLVM)	88.25
Lod (Softmax, ours)	93.73
Lod (FV, ours)	94.65

建筑、海水和天空等。对于目标实例的数目，隐目标可以只是一个实例，也可以是一组实例。

我们提出的方法无需大量预先定义好的、训练好的目标检测器，也不需要精确地定位目标实体，避免灾难性的检测计算量。另外，隐目标的发现可以减少来自嘈杂、语义模糊的区域对场景识别的干扰。受益于类别特定的属性，发掘隐目标有助于发掘图像中的表示力、判别力强的区域。

我们提出的方法趋向于发掘类别特定的隐目标，即对某一类上训练的模型来讲，它在场景类别相似的图像上发掘其视觉相似的区域，但在其他类别的场景图像上，它只捕获到一些杂乱的区域，得到混淆的结果。以“highway”场景为例（图5.7），在“living room”类别上训练的模型在“highway”场景上什么都没检测

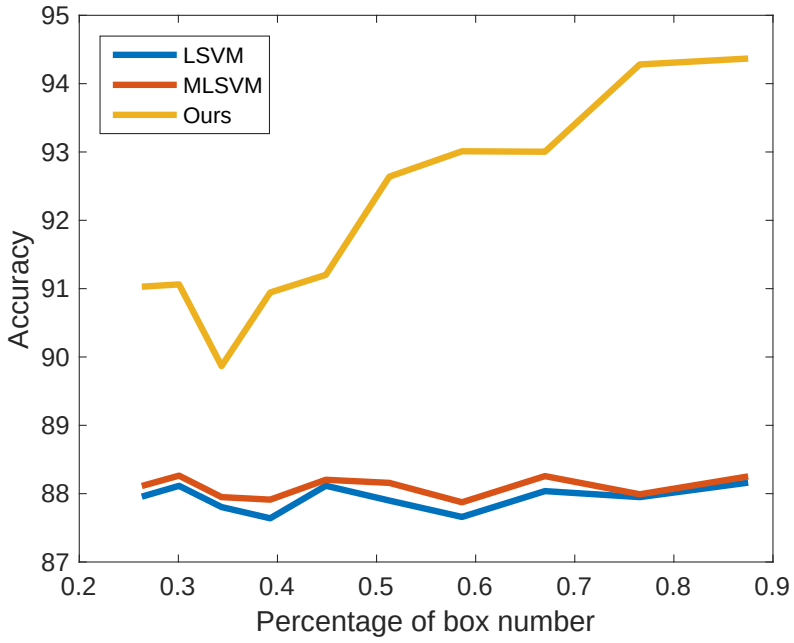


图 5.4 隐目标数目评估

到；而“tallbuilding”类别上训练的模型则是趋向于像 buildings 的区域。图5.8是从分类得分的角度印证了该结论。另外，图5.7也验证了训练的模型可以在相同类别图像上找到语义一致的区域，而在其他类别的图像上只是检测到糟乱的区域。

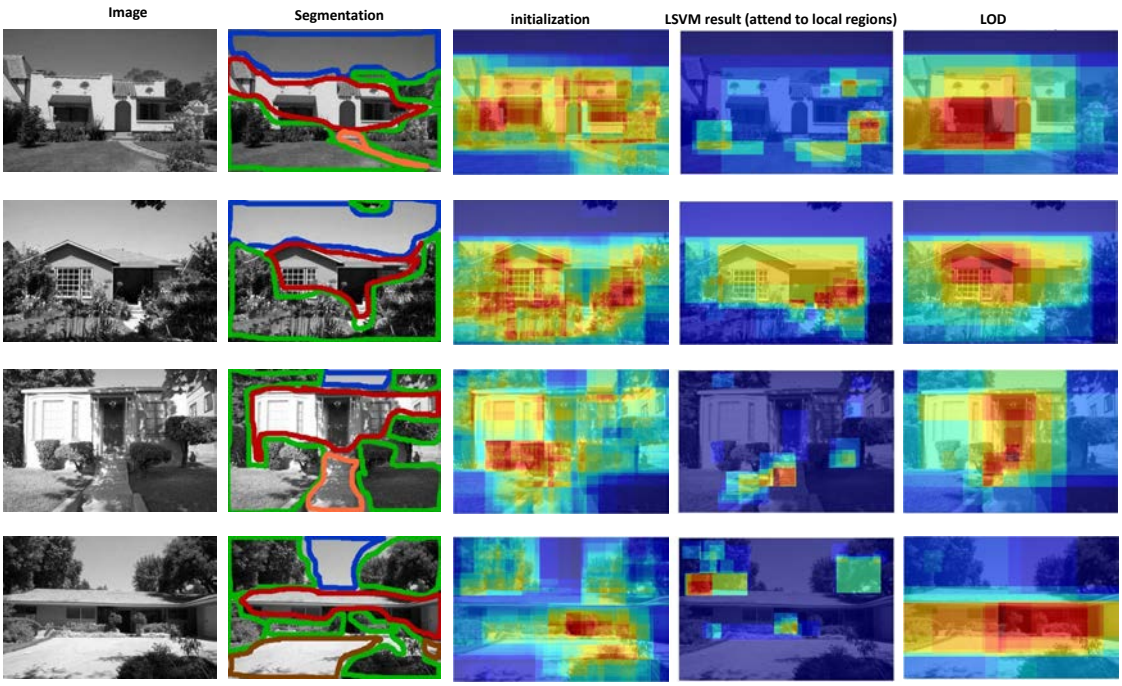


图 5.5 隐目标对比结果图

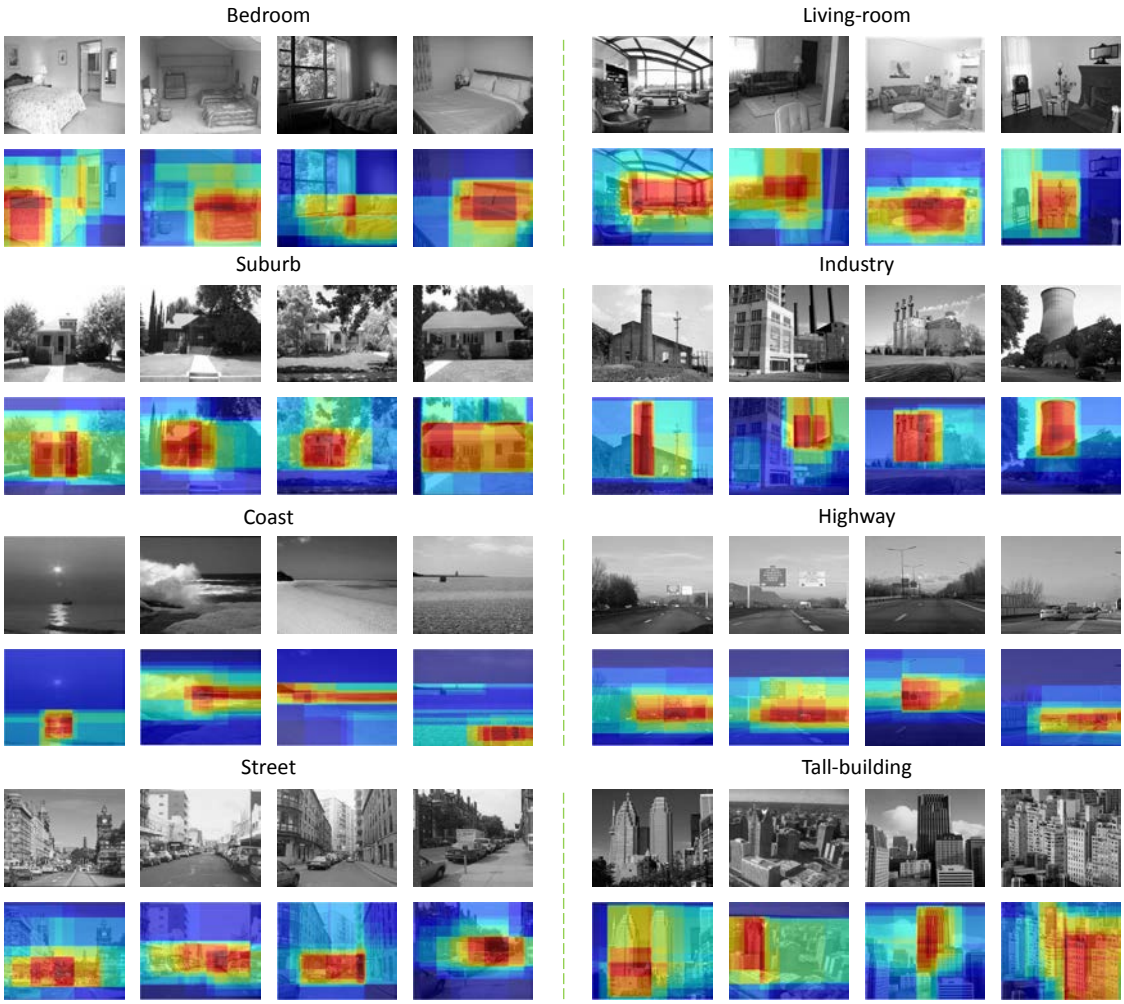


图 5.6 隐目标发现

图 5.7 不同模型的 Lod 可视化

5.5.2 弱监督目标检测

实验所用的数据集是 PASCAL VOC 2007 数据集，该数据集共有 9963 张图像，共 20 类目标。该数据集将整个图像集划分成三个部分：train、val 和 test，每个部分的图像数分别为 2501、2510 和 4952。本文按照弱监督的主流做法，将数据集分为 trainval 和 test 两个部分，分别作为训练集和测试集。本文没有使用图像中对目标的标定，取而代之的是图像的标号。模型性能采用由查全率（Recall）

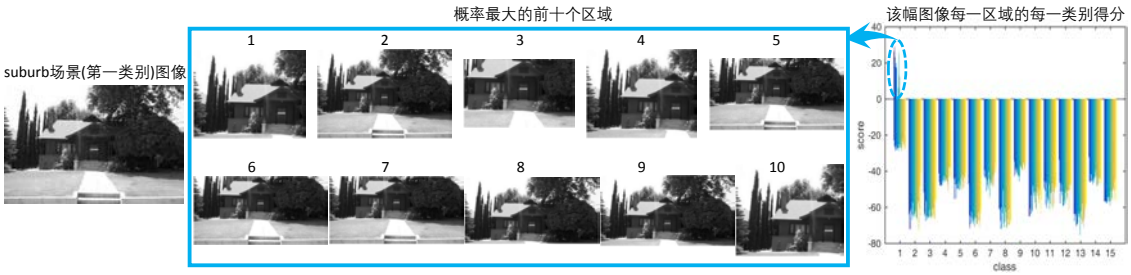


图 5.8 Class-1（“suburb”场景类别）图像的得分最高的前十个区域

和准确率（Precision）两个量计算得到的平均准确度来评估。本文在 PASCAL VOC 2007 数据集上的 20 个类的检测结果如表5.6所示。

表 5.5 PASCAL VOC 2007 数据集的各类目标的图片数

类别	图像张数		类别	图像张数	
	trainval	test		trainval	test
plane	238	204	table	200	190
bike	243	239	dog	421	418
bird	330	282	horse	287	274
boat	181	172	mtbike	245	222
bottle	244	212	person	2008	2007
bus	186	174	plant	245	224
car	713	721	sheep	96	97
cat	337	322	sofa	229	223
chair	445	417	train	261	259
cow	141	127	tv	256	229
共计	5011	4952			

实验结果与分析： 实验中的关联抑制的策略不仅让模型能学的更好，并且还减少了内存消耗，降低了模型的学习时间。在表5.6中，“LOD”、“LOD-C”、“LOD-P”和“LOD-CP”分别表示本文实验的 LOD 算法、LOD 结合关联抑制、LOD 结合部件抑制和结合关联 -部件抑制的结果。对比相关工作“SLSVM”，所提出的模型均得到了不同程度的性能提升。图5.9给出了本文提出的模型和改进前的结果对比。图中第一列是待检测图像，正例样本包括“汽车”、“瓶子”、“电视”和“狗”，而与之关联的反例目标分别是“小汽车”、“人”、“桌子”和“羊”。正例样本为图中的绿色框所示。图中第二列是“SLSVM”模型的样本得分加权累计的结果。从该结果可以看出，本文提出的方法缓解了关联目标易被误判为正例的问题。

图5.10是本文提出的模型和改进之前的模型的对比结果。图中，红色框是对比方法“SLSVM”（Soft Latent SVM，SLSVM）的结果，绿色框是本文提出方法的结果。从图中可以看出，本文方法比较好地改善了目标部件易被误判为目标的问题。图5.10中第三列是本文方法的结果。从图中可以很明显的看出，在第三列中的结果中，本文方法很好地抑制了关联的反例目标，这使得模型训练受到的干扰较小。

表 5.6 弱监督目标检测实验结果及对比

Method	**Song**[146]	**Song**[147]	**Bilen**[103]	**SLSVM**[102]	**LOD**	**LOD-C**	**LOD-P**	**LOD-CP**
plane	27.6	36.3	42.2	46.2	39.8	36.1	37.2	44.9
bike	41.9	47.6	43.9	46.9	42.0	36.4	40.2	52.2
bird	19.7	23.3	23.1	24.1	22.7	18.1	22.2	24.5
boat	9.1	12.3	9.2	16.4	9.1	10.5	16.0	14.4
bottle	10.4	11.1	12.5	12.2	12.9	10.0	4.5	11.2
bus	35.8	36.0	44.9	42.2	42.1	35.1	43.2	40.6
car	39.1	46.6	45.1	47.1	42.1	43.3	45.9	52.2
cat	33.6	25.4	24.9	35.2	23.4	31.8	28.8	35.2
chair	0.6	0.7	8.3	7.8	9.3	2.7	7.3	3.4
cow	20.9	23.5	24.0	28.3	21.3	21.7	31.3	28.9
table	10.0	12.5	13.9	12.7	8.0	10.6	11.5	3.4
dog	27.7	23.5	18.6	21.5	17.9	19.1	23.9	25.9
horse	29.4	27.9	31.6	30.1	27.6	30.9	37.3	39.4
mtbike	39.2	40.9	43.6	42.4	41.9	34.3	40.1	44.4
person	9.1	14.8	7.6	7.8	10.8	14.9	13.9	24.5
plant	19.3	19.2	20.9	20.0	18.8	14.2	19.3	17.2
sheep	20.5	24.2	26.6	26.8	19.9	17.5	27.4	19.1
sofa	17.1	17.1	20.6	20.8	18.4	8.3	16.5	18.2
train	35.6	37.7	35.9	35.8	33.5	30.1	39.6	40.7
tv	7.1	11.6	29.6	29.6	18.2	13.9	22.5	24.9
mAP	22.7	24.6	26.4	27.7	24.0	22.0	26.4	28.3

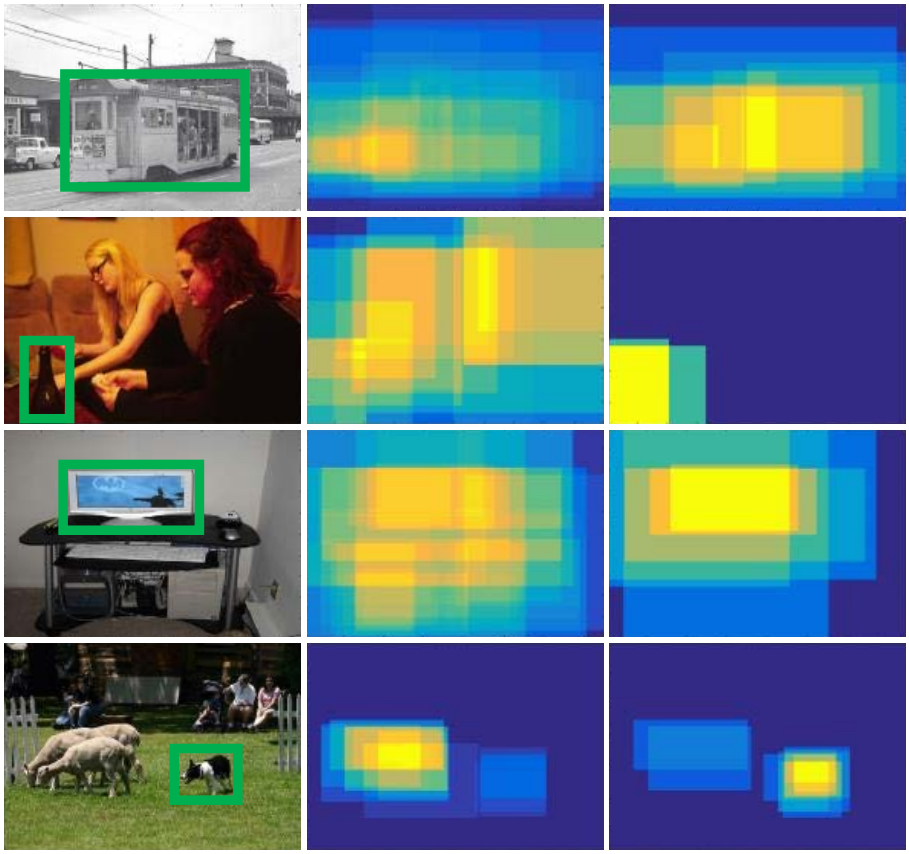


图 5.9 检测结果得分对比图。第一列为原图，图中的绿框表示正例样本，第二、三列分别为对比方法 SLSVM 和本文方法的检测结果。

图 5.10 检测结果对比图。图中绿色框为本文方法的检测结果，红色框为对比方法 SLSVM 的检测结果。

5.6 本章小结

本章提出一个隐目标发现方法，用于自适应地发掘判别性、表示性强的图像区域，避免使用目标标注或者是大量预定义的显式目标检测器。结合最小化熵准则和 Fisher Vector 表示来提升分类模型。所提出的隐目标不仅统一了显式的目标和聚集的区域，同时与 Fisher Vector 特征编码方法融合，得到不错的性能表现。该方法也为进一步探索场景图像语义提供了比较大的潜力。通过实验也验证了所提出方法的优异性能。

第六章　总结与展望

6.1　本文工作总结

本文主要针对场景图像存在较大类内差异性和类间相似性的问题，解决如何寻求对场景图像合适的特征表示以及如何依赖机器学习算法完成分类任务。研究核心是探究如何学习与挖掘场景图像中的语义特征和如何基于语义获得优异的场景特征表示两大问题。所提出的方法主要学习场景图像的隐语义表示和探究语义关联这一重要特性，并基于 Fisher Kernel 理论结合生成式模型和判别式模型来构建判别力强的场景特征表示。进一步地，为了解决特征构建时图像区域划分固定而忽略场景中显著性语义或区域的问题和特征构建与分类模型分离的问题，提出了隐目标作为语义的一种表示形式和隐目标挖掘的场景分类方法。本文的主要研究成果如下：

- 传统的 BoW 特征采用聚类和直方图统计特征的方式，分别用来表示语义和编码语义。为了避免聚类、距离度量等代价大的计算，本文借助于 CNN 的优异语义特性和聚类效应，打破传统的 BoW 构建方法，提出 deep-BoW 的 BoW 构建方法。为了解决 BoW 的一词多义、一义多词的语义模糊问题，本文基于主题模型学习场景图像的隐语义表示。考虑到场景图像中存在语义关联特性，提出隐关联语义表示，旨在通过采用 logistic 正态先验分布去除局部图像块语义之间的独立同分布假设，避免了传统场景图像分类方法依赖图像分割、目标检测、人工标注语义等干预方式带来的计算量大等问题。

- 虽然隐关联语义表示一定程度上解决了我们如何表示场景图像语义的问题，但对于场景图像分类任务来讲，其极大地受限于模型的隐语义/主题数目、视觉词典的大小、特征维度等因素，严重地影响其分类能力。因此，我们提出隐主题向量和关联主题向量，将学习到的关联语义编码于 Fisher Vector 框架中来提升表示的判别能力。为了能让所提出的方法更适合大规模数据集，进一步给出了变分贝叶斯求解和吉布斯采样求解的关联主题向量实现。在大规模数据集上的实验验证了关联主题向量的有效性，并展示出其对 CNN 特征的较大性能提升，对基于深度特征的 Fisher Kernel 表现出巨大的潜力。与 GMM 系列的 Fisher Vector 和 LDA 系列的 Fisher Vector 一起，所提出的关联主题向量为图像语义表示构建了一个更加完备的生成式模型。

- 在隐语义表示学习和基于隐语义获得的关联主题向量中，特征的学习和构建是与分类模型的训练相互分离的。另外所涉及到的局部区域是固定划分的，这也将会影响到对场景中判别力强的信息的挖掘。本文提出了一个隐目

标发现的方法，用于自适应地发掘判别性、表示性强的图像区域，避免目标标注或者大量预定义的目标检测器的使用，同时结合最小化熵准则和 Fisher Vector 表示来提升隐目标的发掘。所提出的隐目标不仅统一了显式的目标和聚集的区域，而且融合了 Fisher Vector 特征编码，得到不错的性能表现。该方法为进一步探索场景图像语义提供了比较大的潜力。另外，所提出的隐目标发现模型被扩展到计算机视觉中的弱监督目标检测，深入分析隐目标发现模型，探究其泛化性。

6.2 未来工作展望

在关联主题向量的学习中，生成式模型与判别式模型的学习各自独立，除了带来较大的计算量之外，两者参数的学习并没有考虑相互之间的影响，整个模型的学习是非端到端的，一定程度上影响了模型的效率和性能。在隐目标挖掘中，模型的学习也存在类似的问题。所以，针对模型的学习方式问题，在未来的工作中，计划从场景图像特征表示的角度出发，以 Fisher Vector 为中心，考虑深度学习架构，探索生成式模型与判别式模型的端到端学习框架。

关联主题向量关注的是隐语义之间的关联性，而现实中，隐语义概念比较模糊，对场景的可解释性较弱。另外，隐语义学习主要是基于 logistic 正态先验分布的主题模型，而该类模型复杂度高，先验分布与后验分布的非共轭特性增加了模型计算难度，近似求解较难保证模型学习的收敛性。另外，本研究的重心从场景图像隐语义的角度来探究场景图像的表示特征并应用于分类任务的问题，在接下来的工作中，可以拓展研究的重心来探究场景图像，比如场景中实体目标之间的关系、空间结构等方面，进一步从计算机视觉角度来增强对场景的理解。另外还计划扩展研究问题，从场景图像可解释性和考虑场景空间结构的角度，探究 3D 场景图像的特征表示，探究深度信息与 RGB 信息对于场景图像表示、理解的作用，并在数据集上对提出的理论和方法进行验证。

附录 A　基于变分贝叶斯的 CTV 推导过程

CTM 的参数是 $\Theta=\{\mu,\Sigma,\beta\}$。图像 d 的 log-likelihood 近似为 L_{VB}：

$$\begin{aligned}
L_{VB} &= E_q[\log p(\eta|\mu,\Sigma)] + \sum_{n=1}^{N_d} E_q[\log p(z_n|\eta)] + \\
&\quad \sum_{n=1}^{N_d} E_q[\log p(w_{d,n}|z_n,\beta)] + H(q) \\
&= 1/2\log|\Sigma^{-1}| - K/2\log 2\pi - \\
&\quad - 1/2[Tr(diag(\nu_d^2)\Sigma^{-1}) + (\lambda_d-\mu)^T\Sigma^{-1}(\lambda_d-\mu)] + \\
&\quad \sum_{n=1}^{N}\{\sum_{i=1}^{K}\lambda_{d,i}\phi_{d,ni} - \zeta^{-1}(\sum_{i=1}^{K}\exp(\lambda_{d,i}+\nu_{d,i}^2/2)) + \\
&\quad 1 - log\zeta\} + \sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{d,ni}\log\beta_{i,w_{d,n}} + \\
&\quad \sum_{i=1}^{K} 1/2(\log 2\pi + \log\nu_{d,i}^2 + 1) - \\
&\quad \sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{d,ni}\log\phi_{d,ni},
\end{aligned} \tag{A-1}$$

其中，ζ 是变分参数。L_{VB} 中涉及参数 μ 的项是：

$$L_{VB}^{[\mu]} = 1/2(\lambda_d-\mu)^T\Sigma^{-1}(\lambda_d-\mu). \tag{A-2}$$

L_{VB} 中涉及参数 Σ 的项是：

$$L_{VB}^{[\Sigma]} = 1/2(\log|\Sigma^{-1}| + Tr(diag(\nu_d^2)) + (\lambda_d-\mu)^T\Sigma^{-1}(\lambda_d-\mu)). \tag{A-3}$$

L_{VB} 中涉及参数 β 的项是：

$$L_{VB}^{[\beta]} = \sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{d,ni}\log\beta_{i,w_{d,n}}. \tag{A-4}$$

根据公式(A-2)-(A-4)，分别关于参数 μ,Σ,β 的一阶偏导即得到关于这三个参数的 Fisher Score 计算公式：公式 (4-1)-(4-3)。考虑到 FIM 可等价于对数似然函数关于参数的二阶导数的负数，根据公式(A-2)-(A-4)分别关于参数 μ,Σ,β 的二阶导数，即可得到关于这三个参数的 FIM 计算公式：公式 (4-4)-(4-6)。

参考文献

[1] MARR D. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information[M]: W.H. Freeman and Company, 1982.

[2] XIAO J. A 2D+ 3D rich data approach to scene understanding[D]: Massachusetts Institute of Technology, 2013.

[3] HENDERSON J M, HOLLINGWORTH A. High-level scene perception[J]. Annual review of psychology. 1999, 50 (1): 243–271.

[4] QUELHAS P, MONAY F, ODOBEZ J, et al. Modeling scenes with local descriptors and latent aspects[C]. Proceedings of the IEEE International Conference on Computer Vision. 2005.

[5] INTRAUB H. Visual scene perception[J]. Encyclopedia of cognitive science. 2006.

[6] XIAO J, EHINGER K A, HAYS J, et al. SUN database: Exploring a large collection of scene categories[J]. International Journal of Computer Vision. 2014, 119 (1): 3–22.

[7] SMEULDERS A W M, WORRING M, SANTINI S, et al. Content-based image retrieval at the end of the early years[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000, 22 (12): 1349–1380.

[8] CHEN Y, WANG J Z, KROVETZ R. An unsupervised learning approach to content-based image retrieval[C]. Proceedings of International Symposium on Signal Processing and Its Applications. 2003.

[9] TORRALBO A, WALTHER D B, CHAI B, et al. Good exemplars of natural scene categories elicit clearer patterns than bad exemplars but not greater bold activity[J]. PloS one. 2013, 8 (3): e58594.

[10] EVERINGHAM M, GOOL L J V, WILLIAMS C K I, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision. 2010, 88 (2): 303–338.

[11] TREISMAN A M, GELADE G. A feature-integration theory of attention[J]. Cognitive psychology. 1980, 12 (1): 97–136.

[12] BIEDERMAN I. Recognition-by-components: A theory of human image understanding[J]. Psychological review. 1987, 94 (2): 115–147.

[13] THORPE S, FIZE D, MARLOT C. Speed of processing in the human visual system[J]. nature. 1996, 381 (6582): 520–522.

[14] LI F F, VANRULLEN R, KOCH C, et al. Rapid natural scene categorization in the near absence of attention[J]. Proceedings of the National Academy of Sciences of the United States of America. 2002, 99: 9596-9601.

[15] LI L J, SU H, LIM Y, et al. Object bank: An object-level image representation for

high-level visual recognition[J]. International Journal of Computer Vision. 2014, 107 (1): 20–39.

[16] KWITT R, VASCONCELOS N, RASIWASIA N. Scene recognition on the semantic manifold[C]. Proceedings of European Conference on Computer Vision. 2012.

[17] FEI-FEI L, PERONA P. A bayesian hierarchical model for learning natural scene categories[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2005.

[18] BOSCH A, ZISSERMAN A, MUÑOZ X. Scene classification via plsa[C]. Proceedings of European Conference on Computer Vision. 2006.

[19] BOSCH A, ZISSERMAN A, MUOZ X. Scene classification using a hybrid generative/discriminative approach[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2008, 30 (4): 712–727.

[20] RASIWASIA N, VASCONCELOS N. Holistic context models for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012, 34 (5): 902–917.

[21] VAN GEMERT J C, VEENMAN C J, SMEULDERS A W M, et al. Visual word ambiguity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010, 32 (7): 1271–1283.

[22] JAAKKOLA T S, HAUSSLER D. Exploiting generative models in discriminative classifiers[C]. Advances in Neural Information Processing Systems. 1998.

[23] RENSINK R A. The dynamic representation of scenes[J]. Visual Cognition. 2000, 7: 17–42.

[24] TORRALBA A. Contextual priming for object detection[J]. International Journal of Computer Vision. 2003, 53 (2): 169–191.

[25] TORRALBA A, MURPHY K P, FREEMAN W T. Contextual models for object detection using boosted random fields[C]. Advances in Neural Information Processing Systems. 2004.

[26] WANG J Z, LI J, WIEDERHOLD G. Simplicity: Semantics-sensitive integrated matching for picture libraries[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2001, 23 (9): 947–963.

[27] CHANG E Y, GOH K, SYCHAY G, et al. CBSA: content-based soft annotation for multimodal image retrieval using bayes point machines[J]. IEEE Transactions on Circuits and Systems for Video Technology. 2003, 13 (1): 26–38.

[28] VAILAYA A, FIGUEIREDO M, JAIN A, et al. Content-based hierarchical classification of vacation images[C]. Proceedings of International Conference on Multimedia Computing and Systems. 1999.

[29] SIAGIAN C, ITTI L. Gist: A mobile robotics application of context-based vision in

outdoor environment[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop. 2005.

[30] MANDUCHI R, CASTANO A, TALUKDER A, et al. Obstacle detection and terrain classification for autonomous off-road navigation[J]. Autonomous robots. 2005, 18 (1): 81–102.

[31] SZUMMER M, PICARD R W. Indoor-outdoor image classification[C]. Content-Based Access of Image and Video Database. Proceedings, 1998 IEEE International Workshop(CAIVD). 1998.

[32] VAILAYA A, FIGUEIREDO M A T, JAIN A K, et al. Content-based hierarchical classification of vacation images[C]. Proceedings of IEEE International Conference on Multimedia Computing and Systems. 1999.

[33] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision. 2004, 60 (2): 91–110.

[34] VERMA A, BANERJI S, LIU C. A new color sift descriptor and methods for image category classification[C]. International Congress on Computer Applications and Computational Science. 2010.

[35] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[J]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2005, 1: 886-893.

[36] AHONEN T, HADID A, PIETIKÄINEN M. Face description with local binary patterns: Application to face recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2006, 28 (12): 2037–2041.

[37] SHECHTMAN E, IRANI M. Matching local self-similarities across images and videos[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2007.

[38] JULESZ B, OTHERS. Textons, the elements of texture perception, and their interactions[J]. Nature. 1981, 290 (5802): 91–97.

[39] SHOTTON J, JOHNSON M, CIPOLLA R. Semantic texton forests for image categorization and segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2008.

[40] MARGOLIN R, ZELNIK-MANOR L, TAL A. Otc: A novel local descriptor for scene classification[C]. Proceedings of European Conference on Computer Vision. 2014.

[41] FELZENSZWALB P F, HUTTENLOCHER D P. Efficient graph-based image segmentation[J]. International journal of computer vision. 2004, 59 (2): 167–181.

[42] SHI J, MALIK J. Normalized cuts and image segmentation[J]. IEEE Transactions on pattern analysis and machine intelligence. 2000, 22 (8): 888–905.

[43] VAN DE SANDE K E, UIJLINGS J R, GEVERS T, et al. Segmentation as selective

search for object recognition[C]. Proceedings of the IEEE International Conference on Computer Vision. 2011.

[44] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D A, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010, 9 (32): 1627–1645.

[45] SIVIC J, RUSSELL B C, EFROS A A, et al. Discovering objects and their location in images[C]. Proceedings of the IEEE International Conference on Computer Vision. 2005.

[46] FAN J, GAO Y, LUO H, et al. Statistical modeling and conceptualization of natural images[J]. Pattern Recognition. 2005, 38 (6): 865–885.

[47] FREDEMBACH C, SCHRODER M, SUSSTRUNK S. Eigenregions for image classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2004, 26 (12): 1645–1649.

[48] OLIVA A, TORRALBA A. Modeling the shape of the scene: A holistic representation of the spatial envelope[J]. International Journal of Computer Vision. 2001, 42 (3): 145–175.

[49] TORRALBA A, FERGUS R, FREEMAN W T. 80 million tiny images: A large data set for nonparametric object and scene recognition[J]. IEEE transactions on pattern analysis and machine intelligence. 2008, 30 (11): 1958–1970.

[50] WANG S, JOO J, WANG Y, et al. Weakly supervised learning for attribute localization in outdoor scenes[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.

[51] SIVIC J, ZISSERMAN A. Video google: A text retrieval approach to object matching in videos[C]. Proceedings of the IEEE International Conference on Computer Vision. 2003.

[52] SU Y, JURIE F. Visual word disambiguation by semantic contexts[C]. Proceedings of the IEEE International Conference on Computer Vision. 2011.

[53] LAZEBNIK S, SCHMID C, PONCE J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2006.

[54] SHABOU A, LEBORGNE H. Locality-constrained and spatially regularized coding for scene categorization[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2012.

[55] HOFMANN T. Probabilistic latent semantic indexing[C]. Proceedings of ACM SIGIR conference on Research and development in information retrieval. 1999.

[56] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research. 2003, 3: 993–1022.

[57] SUDDERTH E B, TORRALBA A, FREEMAN W T, et al. Learning hierarchical

models of scenes, objects, and parts[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2005.

[58] BLEI D M, MCAULIFFE J D. Supervised topic models[C]. Advances in neural information processing systems. 2007.

[59] DIXIT M, RASIWASIA N, VASCONCELOS N. Class-specific simplex-latent dirichlet allocation for image classification[C]. Proceedings of the IEEE International Conference on Computer Vision. 2013.

[60] LACOSTE-JULIEN S, SHA F, JORDAN M I. Disclda: Discriminative learning for dimensionality reduction and classification[C]. Advances in neural information processing systems. 2009.

[61] ZHU J, AHMED A, XING E P. Medlda: maximum margin supervised topic models for regression and classification[C]. Proceedings of the International Conference on Machine Learning. 2009.

[62] ZHU J, AHMED A, XING E P. Medlda: maximum margin supervised topic models[J]. Journal of Machine Learning Research. 2012, 13: 2237–2278.

[63] RASIWASIA N, VASCONCELOS N. Latent dirichlet allocation models for image classification[J]. IEEE transactions on pattern analysis and machine intelligence. 2013, 35 (11): 2665–2679.

[64] CAO L, FEI-FEI L. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes[C]. Proceedings of the IEEE International Conference on Computer Vision. 2007.

[65] NIU Z, HUA G, GAO X, et al. Spatial-disclda for visual recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2011.

[66] LI L, SU H, XING E P, et al. Object bank: A high-level image representation for scene classification and semantic feature sparsification[C]. Advances in Neural Information Processing Systems. 2010.

[67] LI L, SU H, LIM Y, et al. Object bank: An object-level image representation for high-level visual recognition[J]. International Journal of Computer Vision. 2014, 107 (1): 20–39.

[68] ZUO Z, WANG G, SHUAI B, et al. Learning discriminative and shareable features for scene classification[C]. Proceedings of European Conference on Computer Vision. 2014.

[69] ZHOU B, LAPEDRIZA A, XIAO J, et al. Learning deep features for scene recognition using places database[C]. Advances in Neural Information Processing Systems. 2014.

[70] RAO S. Lectures on statistical inference[M]. 2017.

[71] GALLEGUILLOS C, RABINOVICH A, BELONGIE S J. Object categorization us-

ing co-occurrence, location and appearance[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2008.

[72] WU R, WANG B, WANG W, et al. Harvesting discriminative meta objects with deep CNN features for scene classification[C]. Proceedings of the IEEE International Conference on Computer Vision. 2015.

[73] GIRSHICK R B, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[74] GIRSHICK R B. Fast R-CNN[C]. Proceedings of the IEEE International Conference on Computer Vision. 2015.

[75] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence. 2017, 39 (6): 1137–1149.

[76] REDMON J, DIVVALA S K, GIRSHICK R B, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[77] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2009.

[78] DURAND T, THOME N, CORD M. MANTRA: minimum maximum latent structural SVM for image classification and ranking[C]. Proceedings of the IEEE International Conference on Computer Vision. 2015.

[79] SADEGHI F, TAPPEN M F. Latent pyramidal regions for recognizing scenes[C]. Proceedings of European Conference on Computer Vision. 2012.

[80] BORJI A, SIHITE D N, ITTI L. What stands out in a scene? a study of human explicit saliency judgment[J]. Vision Research. 2013, 91: 62–77.

[81] RENSINK R A. Internal vs. external information in visual perception[C]. Proceedings of the International Symposium on Smart graphics. 2002.

[82] SIAGIAN C, ITTI L. Rapid biologically-inspired scene classification using features shared with visual attention[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2007, 29 (2): 300–312.

[83] PERRONNIN F, DANCE C R. Fisher kernels on visual vocabularies for image categorization[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2007.

[84] KRAPAC J, VERBEEK J J, JURIE F. Modeling spatial layout with fisher vectors for image categorization[C]. Proceedings of the IEEE International Conference on Computer Vision. 2011.

[85] ZHOU X, CUI N, LI Z, et al. Hierarchical gaussianization for image classification[C]. Proceedings of the IEEE International Conference on Computer Vision. 2009.

[86] CHANG C, LIN C. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology. 2011, 2 (3): 1–27.

[87] QUATTONI A, TORRALBA A. Recognizing indoor scenes[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2009.

[88] XIAO J, HAYS J, EHINGER K A, et al. SUN database: Large-scale scene recognition from abbey to zoo[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2010.

[89] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]. Advances in Neural Information Processing Systems. 2012.

[90] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. CoRR. 2014, abs/1409.1556.

[91] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[92] GONG Y, WANG L, GUO R, et al. Multi-scale orderless pooling of deep convolutional activation features[C]. Proceedings of European Conference on Computer Vision. 2014.

[93] SINGH S, GUPTA A, EFROS A A. Unsupervised discovery of mid-level discriminative patches[C]. Proceedings of European Conference on Computer Vision. 2012.

[94] JUNEJA M, VEDALDI A, JAWAHAR C V, et al. Blocks that shout: Distinctive parts for scene classification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.

[95] DOERSCH C, GUPTA A, EFROS A A. Mid-level visual element discovery as discriminative mode seeking[C]. Advances in Nueral Information Processing Systems. 2013.

[96] SUN J, PONCE J. Learning discriminative part detectors for image classification and cosegmentation[C]. Proceedings of the IEEE International Conference on Computer Vision. 2013.

[97] PANDEY M, LAZEBNIK S. Scene recognition and weakly supervised object localization with deformable part-based models[C]. Proceedings of the IEEE International Conference on Computer Vision. 2011.

[98] ACZÉL J, DARÓCZY Z. On measures of information and their characterizations[J]. New York. 1975.

[99] BOUCHACOURT D, NOWOZIN S, KUMAR M P. Entropy-based latent struc-

tured output prediction[C]. Proceedings of the IEEE International Conference on Computer Vision. 2015.

[100] PING W, LIU Q, IHLER A T. Marginal structured SVM with hidden variables[C]. Proceedings of the 31th International Conference on Machine Learning. 2014.

[101] SÁNCHEZ J, PERRONNIN F, MENSINK T, et al. Image classification with the fisher vector: Theory and practice[J]. International Journal of Computer Vision. 2013, 105 (3).

[102] BILEN H, PEDERSOLI M, TUYTELAARS T. Weakly supervised object detection with posterior regularization[C]. Proceedings of British Machine Vision Conference. 2014.

[103] BILEN H, PEDERSOLI M, TUYTELAARS T. Weakly supervised object detection with convex clustering[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[104] UIJLINGS J R, VAN DE SANDE K E, GEVERS T, et al. Selective search for object recognition[J]. International Journal of Computer Vision. 2013, 104 (2): 154–171.

[105] QUELHAS P, MONAY F, ODOBEZ J M, et al. A thousand words in a scene[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2007, 29 (9): 1575–1589.

[106] SYDOROV V, SAKURADA M, LAMPERT C H. Deep fisher kernels - end to end learning of the fisher kernel GMM parameters[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[107] DONAHUE J, JIA Y, VINYALS O, et al. Decaf: A deep convolutional activation feature for generic visual recognition[C]. Proceedings of the International Conference on Machine Learning. 2014.

[108] CSURKA G, DANCE C, FAN L, et al. Visual categorization with bags of keypoints[C]. Proceedings of European Conference on Computer Vision Workshop. 2004.

[109] JOACHIMS T. Text categorization with support vector machines: Learning with many relevant features[C]. Proceedings of European Conference on Machine Learning. 1998.

[110] JAAKKOLA T S, HAUSSLER D. Exploiting generative models in discriminative classifiers[C]. Advances in Neural Information Processing Systems. 1999.

[111] HOLUB A, WELLING M, PERONA P. Combining generative models and fisher kernels for object recognition[C]. Proceedings of the IEEE International Conference on Computer Vision. 2005.

[112] JÉGOU H, DOUZE M, SCHMID C, et al. Aggregating local descriptors into a compact image representation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2010.

[113] KOBAYASHI T. Dirichlet-based histogram feature transform for image classification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[114] CINBIS R G, VERBEEK J, SCHMID C. Image categorization using fisher kernels of non-iid image models[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2012.

[115] CINBIS R G, VERBEEK J J, SCHMID C. Approximate fisher kernels of non-iid image models for image categorization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016, 38 (6): 1084–1098.

[116] AITCHISON J. The statistical analysis of compositional data[J]. Journal of the Royal Statistical Society. Series B (Methodological). 1982, 44 (2): 139–177.

[117] BLEI D M, LAFFERTY J D. A correlated topic model of science[J]. The Annals of Applied Statistics. 2007, 1 (1): 17–35.

[118] SÁNCHEZ J, PERRONNIN F, MENSINK T, et al. Image classification with the fisher vector: Theory and practice[J]. International Journal of Computer Vision. 2013, 105 (3): 222–245.

[119] JORDAN M I, GHAHRAMANI Z, JAAKKOLA T S, et al. An introduction to variational methods for graphical models[J]. Machine Learning. 1999, 37 (2): 183–233.

[120] BLEI D M, LAFFERTY J D. Correlated topic models[C]. Advances in Neural Information Processing Systems. 2006.

[121] CHEN J, ZHU J, WANG Z, et al. Scalable inference for logistic-normal topic models[C]. Advances in Neural Information Processing Systems. 2013.

[122] MIMNO D, WALLACH H M, MCCALLUM A. Gibbs sampling for logistic normal topic models with graph-based priors[C]. Advances in Neural Information Processing Systems Workshop on Analyzing Graphs. 2008.

[123] SALIMANS T, KINGMA D, WELLING M. Markov chain monte carlo and variational inference: Bridging the gap[C]. Proceedings of the International Conference on Machine Learning. 2015.

[124] JIA Y, SHELHAMER E, DONAHUE J, et al. Caffe: Convolutional architecture for fast feature embedding[C]. Proceedings of ACM International Conference on Multimedia. 2014.

[125] ATCHISON J, SHEN S M. Logistic-normal distributions: Some properties and uses[J]. Biometrika. 1980, 67 (2): 261–272.

[126] PERRONNIN F, SÁNCHEZ J, MENSINK T. Improving the fisher kernel for large-scale image classification[C]. Proceedings of European Conference on Computer Vision. 2010.

[127] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption

generator[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[128] DIXIT M, CHEN S, GAO D, et al. Scene classification with semantic fisher vectors[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[129] LI L J, SU H, FEI-FEI L, et al. Object bank: A high-level image representation for scene classification & semantic feature sparsification[C]. Advances in Neural Information Processing Systems. 2010.

[130] PANDEY M, LAZEBNIK S. Scene recognition and weakly supervised object localization with deformable part-based models[C]. Proceedings of the IEEE International Conference on Computer Vision. 2011.

[131] DOERSCH C, GUPTA A, EFROS A A. Mid-level visual element discovery as discriminative mode seeking[C]. Advances in Neural Information Processing Systems. 2013.

[132] JUNEJA M, VEDALDI A, JAWAHAR C V, et al. Blocks that shout: Distinctive parts for scene classification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.

[133] ZHANG L, ZHEN X, SHAO L. Learning object-to-class kernels for scene classification[J]. IEEE Transactions on Image Processing. 2014, 23 (8): 3241–3253.

[134] LIU L, SHEN C, WANG L, et al. Encoding high dimensional local features by sparse coding based fisher vectors[C]. Advances in Neural Information Processing Systems. 2014.

[135] CHATFIELD K, LEMPITSKY V S, VEDALDI A, et al. The devil is in the details: an evaluation of recent feature encoding methods.[C]. Proceedings of British Machine Vision Conference. 2011.

[136] BERGAMO A, TORRESANI L. Classemes and other classifier-based features for efficient object categorization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014, 36 (10): 1988–2001.

[137] RAZAVIAN A S, AZIZPOUR H, SULLIVAN J, et al. Cnn features off-the-shelf: an astounding baseline for recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop. 2014.

[138] VAN DER MAATEN L, HINTON G. Visualizing data using t-sne[J]. Journal of Machine Learning Research. 2008, 9: 2579–2605.

[139] ZHAO Y, KARYPIS G. Criterion Functions for Document Clustering: Experiments and Analysis[R]. Minneapolis, MN, 2001.

[140] RASIWASIA N, VASCONCELOS N. Well-separated clusters and optimal fuzzy partitions[J]. Journal of cybernetics. 1974, 4 (1): 95–104.

[141] DURAND T, THOME N, CORD M. MANTRA: minimum maximum latent struc-

tural SVM for image classification and ranking[C]. Proceedings of the IEEE International Conference on Computer Vision. 2015.

[142] WU R, WANG B, WANG W, et al. Harvesting discriminative meta objects with deep CNN features for scene classification[C]. Proceedings of the IEEE International Conference on Computer Vision. 2015.

[143] ZHOU B, KHOSLA A, LAPEDRIZA À, et al. Object detectors emerge in deep scene cnns[C]. Proceedings of International Conference on Learning Representations. 2015.

[144] ABRAMOWITZ M, STEGUN I A. Handbook of mathematical functions[J]. American Journal of Physics. 1966, 34 (2): 177-177.

[145] MILLER K, KUMAR M P, PACKER B, et al. Max-margin min-entropy models[C]. Artificial Intelligence and Statistics. 2012.

[146] SONG H O, GIRSHICK R, JEGELKA S, et al. On learning to localize objects with minimal supervision[C]. Proceedings of International Conference on Machine Learning. 2014.

[147] SONG H O, LEE Y J, JEGELKA S, et al. Weakly-supervised discovery of visual pattern configurations[C]. Advances in Neural Information Processing Systems. 2014.

作者简历及攻读学位期间发表的学术论文与研究成果

作者基本情况

2007.092011.07	中国矿业大学（北京）	计算机科学与技术	学士
2011.092017.11	中国科学院大学	计算机应用技术	博士

攻读学位期间发表的学术论文

1. **Pengxu Wei**, Fei Qin, Fang Wan, Yi Zhu, Jianbin Jiao, Qixiang Ye. Correlated Topic Vector for Scene Classification [J], *IEEE Transactions on Image Processing (TIP)*, 2017, 27(7):3221-3234. (SCI)

2. **Pengxu Wei**, Qixiang Ye, Jianbin Jiao. Latent Object Discovery for Scene Recognition [J], *Pattern Recogition Letter (PRL)*, 2017. (SCI, 在投)

3. Xiaodan Zhang, Shengfeng He, Xinhang Song, **Pengxu Wei**, Shuqiang Jiang, Qixiang Ye, JianbinJiao, Rynson Lau. Keyword-driven Image Captioning via Context-dependent Bilateral LSTM [C]. *IEEE International Conference on Multimedia and Expo (ICME)*, 2017. (EI)

4. Fang Wan, **Pengxu Wei**, Zhenjun Han, Kun Fu, Qixiang Ye. Weakly Supervised Object Detection with Correlation and Part Suppression [C]. *IEEE International Conference on Image Processing (ICIP)*, 2016. (EI, 共同一作)

5. Wei Ke, Yao Zhang, **Pengxu Wei**, Qixiang Ye, and Jianbin Jiao. Pedestrian Detection via PCA Filters Based Convolutional Channel Features [C]. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015:1394-1398. (EI)

6. Xiaogang Chen, **Pengxu Wei**, Wei Ke, Qixiang Ye, and Jianbin Jiao. Pedestrian Detection with Deep Convolutional Neural Network [C]. *Proceedings of Asian Conference on Computer Vision (ACCV) Workshop*, 2014:354-365. (EI)

7. Jialing Zou, Xiaogang Chen, **Pengxu Wei**, Zhenjun Han, and Jianbin Jiao. A Belief Based Correlated Topic Model for Semantic Region Analysis in Far-Field Video Surveillance Systems [C]. *Proceedings of Pacific-Rim Conference on Multimedia (PCM)*, 2013:779-790. (EI)

专利

1. 一种基于少量样本的遥感目标检测方法 (已受理)

致　谢

在写这篇博士论文的过程中，我查看了研究生阶段自己按照实验室惯例每周都要写的周报，感触颇多。每一篇周报都让我回想起当时自己的样子：或斗志昂扬的、或低沉迷茫的、或纠结的、或兴奋的、或痛苦的、或愧疚的、或坚定的，或崩溃到痛哭，或兴奋到狂笑..... 时至今日，我仍清晰地记得小崔姐在我硕士二年级决定是否转博的时候跟我说：博士就是熬，熬过去就好了。整个博士学习阶段确实比较煎熬，但是我觉得一切的经历都是有意义的。串起周报中的内容，让我看到自己整个研究生在科研方面的成长，逐渐领悟读博士的价值所在。值此论文完成之际，谨在此向多年来给予我关心和帮助的老师、学长、同学、朋友和家人表示衷心的感谢！

感谢我的导师焦建彬教授。焦老师学识渊博、为人谦和、治学严谨，对我的培养倾注了很大的心血。焦老师一直以来给予我很大的信任，让我能够保持乐观积极、坚定执着的心态进行研究探索。无论是在科研上还是生活上，当我遇到问题的时候，焦老师用心的指导和帮助，让我能够明白问题所在，不断发现和完善自己的不足，更让我了解自己。我会永远铭记焦老师对我的激励、支持与指导。

感谢叶齐祥教授。叶老师一直在教我做有价值的工作，尽管我做得不好，但是这份信念潜移默化地影响着我，让我对自己有了更多的期许和追求，让我能够对科研保持一颗初心，让我能够不时地审视自己、思考自己要实现的个人价值。我的每一次成长和进步都倾注着叶老师无私的付出，我也渐渐明白叶老师对我的良苦用心。叶老师赤诚的科研之心、乐观自在的生活态度，也将会继续影响着我。谢谢叶老师。

感谢秦飞副教授。秦老师严谨的科学态度、清晰的思维逻辑和精益求精的工作作风对我影响至深。与秦老师的沟通总是让我收益颇丰，每一次我都很期待与秦老师的沟通。秦老师在逐词逐句带我讨论、改文章的过程，极大地锻炼着我发现问题和解决问题的能力。秦老师一直是我敬仰和学习的榜样。

感谢韩振军副教授。自我进实验室，韩老师就成为我学习和敬佩的榜样。无论是科研上还是生活上，韩老师做事的热情和冲劲，每一次都能给我很大的触动和激励。无论我遇到什么问题，韩老师总是能站出来支持，让我真真切切地感受到坚持下去的力量，让我能够特别特别的安心。

感谢邹佳凌师兄从我进实验室就开始给我讲读博士、做科研的意义，帮助我纠正我的毛病，影响着我成为更独立的自己。感谢柯炜给了很多有价值的意见，也让我不断地审视自己，影响着我成为更理性的自己。感谢万方容忍着我怪脾气，也正是你耐心听我讲各种奇怪想法，让我在迷乱之中理清自己的思路，影响着我成为更明朗的自己。感谢彭艺师姐、李策师姐、张晓丹师姐、崔妍婷师妹，你们

不仅是我学习的榜样，而且也是我最贴心的朋友，总是给我特别暖心的建议和帮助，用心教我做事的态度和方式，影响着我成为更好的自己。

感谢与我一起奋斗的模式识别与智能系统开发实验室的兄弟姐妹。正是你们陪伴着我度过每一天，容忍着我的倔强，忍受着我可怕的笑声，还不忘锻炼我一把拎起 20kg 重的桶装水来换水。虽然你们总是损我，但是我也没吃亏，因为我慢慢也学会了损自己和你们。无论是生活还是学习上，从你们身上我真的学到了很多，感谢你们为我带来的欢乐、鼓励和帮助。在实验室的时光里，实验室为我带来很多的东西。在写致谢时，回忆起实验室的每一个人每一件事都让我泪流满面，原来实验室早已成为我感情上难以割舍的一部分，你们每一个人都已成为我人生中非常重要的人，我会一直铭记在心里。

感谢我的父母多年以来对我的支持和鼓励。虽然我从不讲我在学校的压力，但是您们对我的一切都了如指掌。每次回家，您们都能给予我最大的努力动力。感谢哥哥嫂子、弟弟弟妹对我的支持和对家的照顾，能够让我可以安心地学习。感谢亦师亦友的徐天天一直以来对我无私的照顾，教我做事做人的道理。

最后，感谢在百忙之中抽出宝贵时间参加论文评审和答辩的各位老师！

魏朋旭

2017 年 11 月