

# 硕士学位论文

## 基于深度学习的敏感目标检索方法研究

作者姓名： 郝杰东

指导教师： 谭铁牛 研究员 中国科学院自动化研究所

学位类别： 工学硕士

学科专业： 模式识别与智能系统

培养单位： 中国科学院自动化研究所

2018 年 6 月



**Research on Sensitive Object Retrieval Based on Deep Learning**

A thesis submitted to the  
University of Chinese Academy of Sciences  
in partial fulfillment of the requirement  
for the degree of  
Master of Science in Engineering  
in Pattern Recognition and Intelligent Systems  
By  
**Jiedong Hao**  
Supervisor: Professor Tieniu Tan

**Institute of Automation, Chinese Academy of Sciences**

**June, 2018**



## 中国科学院大学 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研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何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，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日 期：

## 中国科学院大学 学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关保存和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，可以按照学术研究公开原则和保护知识产权的原则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。

涉密及延迟公开的学位论文在解密或延期后适用本声明。

作者签名：

导师签名：

日 期：

日 期：



## 摘要

基于内容的图像检索是计算视觉领域一个非常重要而且经典的研究方向，同时，相关的技术在工业界也有非常广泛的应用。近些年来，随着深度学习的兴起，由于卷积神经网络对图像特征很好的表达能力，基于深度卷积神经网络的方法在图像分类，图像检索，物体检测和语义分割等领域都取得了超越传统方法的结果。尽管图像检索技术已经被研究多年，但仍然面临很多挑战，图像中物体的尺寸，姿态以及图像光照的变化都给检索算法的性能带来严重的干扰。在本论文中，我们主要研究了基于深度卷积神经网络的图像检索方法以及该方法在敏感图像检索上的应用，论文的工作以及贡献总结如下：

### 1. 建立了一个大规模枪支图像数据库 — Firearm14k

在当前的社交网络上，充斥着各种各样的令普通用户感到不适的枪支图片，这些图片可能会激起暴力等不良后果，因此有必要对枪支图片进行适当的监管与处理。另外一方面，基于深度卷积神经网络的方法，在网络模型的训练过程中往往需要大量的训练图片，如果训练数据过少，学到的模型很容易过拟合。截至目前，学术界并没有一个大规模枪支图像数据库存在，为了方便研究者针对这个领域进行研究，我们收集了一个大规模的枪支图像数据库，包含 167 类不同类型的枪支，图片总数为 14755 张，我们将其简称为 Firearm14k。该数据库包含了真实世界拍摄的枪支图片，因此图片中物体尺寸，姿态，背景等变化很大，识别难度较高。该数据库可以用于枪支图片精细检索的研究，也可以用于枪支图片的精细分类等研究工作。

### 2. 提出了一种多尺度全卷积的图像实例检索方法

目前已有很多工作利用卷积神经网络提取图像特征进行图像检索，但是这些工作并未对影响图像特征有效性的各种因素进行详细分析，例如，图像尺寸缩放的策略，影响多尺度特征有效性的因素等，因此各种因素如何影响检索的性能仍不明确。在该工作中，我们对输入神经网络的图像尺寸缩放策略，提取图像多尺度特征的方式，以及 PCA 和白化矩阵学习这三个重要的因素进行了研究，通过实验分析了这些因素对检索结果的影响。在此基础上，我们提出了多尺度全卷积的图像特征提取方法。该方法简单而有效，我们在 Oxford5k, Paris6k, Oxford105k 以及 UKB 这四个常用数据库上进行了实验，大量的实验结果表明我们提出的方法有着良好的检索效果。

### 3. 提出了一种基于双阈值对比损失函数的精细枪支图像检索方法

在社交网络上或者是在取证领域，人们需要能够自动监管一些不适当的枪支图片或者鉴定枪支的类型等，基于图像检索的技术能够帮助人们有效解决此类问题。通过重新微调已有的神经网络模型，基于卷积神经网络的检索方法取得了很好的效果。传统的单阈值对比损失函数，由于其简单并且有效，被大量使用，但是我们发现将该损失函数用在 Firearm14k 图像库枪支检索任务上时，网络的性能并不好，原因有两点：第一，在网络训练过程中，相似与不相似样本贡献的损失不平衡；第二，Firearm14k 数据库与 ImageNet 数据库的图片风格差异巨大。我们提出了双阈值对比损失函数来解决网络训练中正负样本贡献的损失不平衡的问题；为了解决 Firearm14k 与 ImageNet 数据库的差异问题，我们使用了两步训练的策略，首先用分类任务微调网络，然后再使用检索任务微调网络。大量实验结果表明我们所提出的方法在枪支检索上的准确率超过了当前主流的方法。

**关键词：** 深度卷积神经网络，精细图像检索，多尺度特征表达，全卷积网络，双阈值对比损失函数

## Abstract

Content-based image retrieval is an important and traditional research topic in computer vision. It has also been widely applied in industrial applications. In recent years, deep learning methods have been very popular. Due to the excellent ability of the convolutional neural networks to represent an image, approaches based on it have achieved remarkable success over the traditional methods in areas such as image classification, image retrieval, object detection and semantic segmentation. Although it has been studied for a long time, image retrieval still faces a lot of challenges, for example, the large variation of object scale, pose and lighting conditions in different images will affect the performance of retrieval methods significantly. In this paper, we mainly study image retrieval based on deep convolutional neural networks and its application on sensitive image retrieval. The work and contributions of this paper are summarized below:

### 1. We build a large scale firearm image dataset — Firearm14k

The proliferation of firearm images in the social media may incite violence. So we need to properly regulate the appearance of these shocking firearm images. On the other hand, for approaches based on deep convolutional neural networks, the models are data-hungry during the training process. If there are not enough training data, the model may overfit on the dataset. Right now, no large dataset of firearm images exists in academic community. To facilitate research in this area, we have built a large scale firearm dataset — Firearm14k, which consists of 14,755 images from 167 categories of various firearm types. The dataset contains images from real world, which has large variability in object size, pose and background, etc., thus is challenging to recognize. This dataset can be used both for research on fine-grained firearm image retrieval and firearm image classification.

### 2. A novel multi-scale fully-convolutional approach for visual instance retrieval

There has been a lot of work on image retrieval based on deep convolutional neural networks. But few work has given a detailed analysis on the various factors that impact the effectiveness of image features extracted from the network. The impact of some of the factors such as image resizing strategy, multi-scale feature representation, has

not been fully explored. In this work, we studied the image resizing strategy, the way to extract multi-scale image features and the suitable way to learn PCA and whitening matrix and analyzed their impact on retrieval performance. Based on our analysis and experimental results, we propose a multi-scale fully-convolutional approach for visual instance retrieval, which is simple yet effective. We conduct experiments on four common datasets, i.e., Oxford5k, Paris6k, Oxford105k and UKB. Our method shows promising results compared to other state-of-the-art methods.

### 3. An approach for fine-grained firearm image retrieval based on double margin contrastive loss

There are great needs for automatically regulating shocking firearm images in social media or identifying firearm types in forensic science. Image retrieval techniques have a great potential to solve such problems. Recent advances in image retrieval are mainly driven by fine-tuning state-of-the-art convolutional neural networks for retrieval task. The contrastive loss, known for its simplicity and good performance, has been widely used. We find that it performs poorly for the Firearm14k dataset due to: (1) Loss contributed by similar and dis-similar image pairs during training is unbalanced. (2) A huge domain gap exists between this dataset and ImageNet. We propose to deal with the unbalanced loss by employing a double margin contrastive loss. We tackle the domain gap issue with a two-stage training strategy, where we first fine-tune the network for classification and then fine-tune it for retrieval. Extensive experiments show that our approach outperforms the state-of-the-art methods on firearm image retrieval task.

**Keywords:** deep convolutional neural network, fine-grained image retrieval, multi-scale feature representation, fully-convolutional network, double margin contrastive loss

## 目 录

<b>第1章 绪论 .....</b>	<b>1</b>
1.1 引言 .....	1
1.2 研究背景及意义 .....	2
1.3 论文的组织 .....	3
<b>第2章 图像检索方法综述 .....</b>	<b>5</b>
2.1 概述 .....	5
2.2 早期图像检索方法 .....	6
2.3 基于 SIFT 局部特征的检索方法 .....	6
2.3.1 基于视觉词袋模型（BOF）的检索方法 .....	7
2.3.2 基于 VLAD 和 Fisher Vector 的检索方法 .....	8
2.4 基于卷积神经网络的图像检索方法 .....	9
2.4.1 卷积神经网络简介 .....	9
2.4.2 基于神经网络直接提取图像特征的检索方法 .....	10
2.4.3 微调已有网络的检索方法 .....	12
2.5 本章小结 .....	13
<b>第3章 大规模枪支图片数据库的构建 .....</b>	<b>15</b>
3.1 引言 .....	15
3.2 数据库图片收集 .....	17
3.2.1 图像的收集 .....	17
3.2.2 图片的清理 .....	17
3.3 图像标注 .....	18
3.4 标注质量评估及图像筛选 .....	19
3.4.1 图像标注质量评估 .....	19
3.4.2 有效图片筛选 .....	20
3.5 数据库的划分与相关统计信息 .....	21
3.6 本章小结 .....	23
<b>第4章 基于多尺度全卷积网络的图像实例检索方法 .....</b>	<b>25</b>
4.1 引言 .....	25
4.2 基于多尺度全卷积网络的图像实例检索方法 .....	26
4.2.1 背景知识 .....	26
4.2.2 图像尺寸改变策略 .....	26

4.2.3 多尺度特征表达	27
4.2.4 特征降维与白化	28
4.3 实验	28
4.3.1 实现细节	28
4.3.2 评价指标与数据库	28
4.3.3 实验结果与分析	30
4.3.4 与其他方法的对比	33
4.4 相关工作	34
4.5 本章小结	35
<b>第5章 基于双阈值对比损失函数的枪支图像检索方法</b>	<b>37</b>
5.1 引言	37
5.2 基于双阈值对比损失函数的枪支图像检索方法	39
5.2.1 图像特征表达	39
5.2.2 双阈值对比损失函数	39
5.2.3 两步训练策略	41
5.2.4 特征降维	42
5.3 实验	42
5.3.1 评价指标	42
5.3.2 实验细节	42
5.3.3 实验结果与分析	44
5.3.4 检索效果的可视化	47
5.3.5 与其他方法的对比	48
5.4 相关工作	50
5.5 本章小结	51
<b>第6章 工作总结与展望</b>	<b>53</b>
6.1 工作总结	53
6.2 工作展望	54
<b>参考文献</b>	<b>57</b>
<b>致 谢</b>	<b>65</b>
<b>作者简历及攻读学位期间发表的学术论文与研究成果</b>	<b>67</b>

## 图形列表

2.1 视觉词袋模型的工作流程 .....	7
3.1 UKB 数据集中一些示例图片 .....	15
3.2 Facebook 上一个枪支交易群 .....	16
3.3 在线图像标注系统的前端评分界面 .....	18
3.4 枪支数据库打分数据的偏差统计信息 .....	19
3.5 Firearm14k 数据与其他数据库图片尺寸比值的概率分布 .....	21
3.6 枪支图像数据库各个类别图片数量 .....	21
3.7 来自 Firearm14k 数据库的一些示例图片 .....	23
4.1 一张图像的 3 尺度特征表示 .....	27
4.2 学习 PCA 与白化矩阵的方式对检索结果的影响 .....	33
5.1 基于双阈值对比损失方法的网络结构 .....	40
5.2 使用不同损失函数时，正负样本对对损失的贡献率 .....	45
5.3 双阈值对比损失函数中两个阈值对模型检索性能的影响 .....	46
5.4 不同模型下相似与不相似图像对特征距离的概率分布 .....	47
5.5 我们的方法的一些示例检索结果 .....	49
5.6 不同方法 Rank-k 准确率的比较 (%) .....	50
6.1 两个图片剪切篡改的实例 .....	54



## 表格列表

3.1 每个标注者有最大的个人偏差的类别数目 .....	20
3.2 Firearm14k 数据库详细统计信息 .....	23
4.1 不同图像尺寸变化策略之间的比较 .....	31
4.2 不同数据库中图像最小尺寸落在不同区间的比例 .....	31
4.3 不同设置下多尺度特征检索结果对比 .....	32
4.4 与其他方法的对比 .....	34
5.1 单阈值方法与双阈值方法的对比 .....	44
5.2 两步训练策略的结果 .....	45
5.3 与其他主流方法的结果比较 .....	48



## 第1章 绪论

### 1.1 引言

本文的研究内容为基于内容的图像检索（content-based image retrieval，CBIR）。在计算机发展早期，由于拍照设备的昂贵与匮乏，图像数量相对较少，人们通常采用一些关键词来描述图像，在需要检索图像的时候，采用文本检索相关的关键词来找到对应的图像，这种方法称为基于文本的图像检索。这种基于文本的检索方式存在明显的缺点，一方面，这种方法的准确率依靠图像标注的准确度，而且图像的标注也十分耗时耗力，不同的人之间存在一定的主观性偏差；另外，图像中颜色以及纹理等信息也很难用文字精确地描述。从 20 世纪 90 年代以来，随着互联网和个人电脑的流行，以及拍照和存储设备的大量普及，数字图像出现井喷，通过人工对图像进行标注与管理，变得更加不现实与困难。为了克服基于文本的图像搜索的问题，高效地组织和检索这些图片，人们提出了基于内容的图像检索技术。

所谓基于内容的图像检索，通俗来说就是「以图搜图」，是把计算机视觉相关技术应用到图像检索中，利用计算机来自动提取图像的颜色，形状，纹理等信息或者更高层的图像语义信息，将这些信息表示为图像的特征，计算查询图像特征与数据中的图像特征的相似度，然后从图像库返回与查询的图像相似的图片。这里的「内容」指的是图像颜色，形状，纹理等可以从图像中直接得到的信息，而不是图像的标注信息等文本信息。这种技术的优点是无需人工描述图像内容，可以应用到大规模场景，只要图像的特征提取准确，就能保证很高的检索准确率。

从 1990 年到 2000 年的这十几年间，是 CBIR 技术的发展初期，相关研究者对基于内容的图像检索技术进行了大量的研究。在这一时期，研究者们使用的特征主要是图像的颜色直方图特征，纹理以及形状等比较简单的特征，同时，由于硬件以及算法的限制，相关的图像库通常比较简单，图片数量也相对较少。学术界和工业界也开发了一些早期的图像检索的系统，例如 PhotoBook [1]，QBIC [2]，Virage [3]，PicToSeek [4] 等。

2000 年以后，随着 SIFT（scale invariant feature transform）[5]局部特征特征描述子提取算法的提出，由于 SIFT 特征对图像的旋转，缩放以及光照的变化都有很好的鲁棒性，因而被大量使用，成为最流行的图像局部特征表示方法。

之后，基于 SIFT 的视觉词袋（bag of features, BOF）模型被提出 [6]。BOF 借鉴自文本检索领域的词袋模型，很好地利用了图像的局部特征，用一个特征向量编码了图像的信息，在检索任务上取得了不错的成绩。在这之后，有大量的基于 BOF 或者 SIFT 特征的检索方法出现 [7–11]，图像检索方法研究迎来又一波热潮。随着深度学习的兴起，深度卷积神经网络在计算机视觉各个领域都取得了大幅超越其他方法的结果，基于卷积神经网络的图像检索方法也大量涌现。在工业界，也出现了一些支持以图搜图的功能的图像搜索系统，国外的此类网站如 TinEye<sup>1</sup> 和 Google 图像搜索引擎<sup>2</sup>，国内则有百度识图<sup>3</sup>以及搜狗图像搜索<sup>4</sup>等。

除了应用于通用的图像搜索引擎，基于内容的图像搜索技术在其他方面也有着非常广泛的应用：在医学领域，检索相同器官的 X 光照片；在社会公共安全及打击犯罪方面，检索嫌疑人的指纹以及面部信息等，确定数据库中是否有相关资料；在设计领域，设计师通过给出一些包含特定纹理以及颜色的照片，从数据库中或者互联网上寻找具有相似风格的照片；在电商购物网站，消费者希望通过服装或者商品的照片，从网站上找到同款或者相似的商品；在版权保护领域，版权所有人可以采用图像检索的技术查找自己的图片是否有被其他人未经授权使用。其他类型的应用场景还有很多，总的来说，基于内容的图像检索技术在实际生活中有着广泛的应用。

## 1.2 研究背景及意义

智能手机的流行给用户拍照带来了便利，拍照不再是专业摄影师的权力，随着时间的推移，用户们也累积了大量的个人照片，如何组织管理这些照片成为一个问题，基于图像检索的技术可以帮助用户对这些照片进行有效的聚类与管理。伴随着移动互联网时代的来临以及互联网服务接入的便利化，一些社交类的网站和应用，例如 Facebook, Twitter, Weibo, Wechat 等，以及图片视频分享网站和应用，如 Instagram, Flickr, YouTute, 优酷等发展迅速。随之而来的是互联网产生的大量的多媒体内容（视频，图片，音频，文本等）。一些典型的社交应用每天产生的图片量十分惊人，例如，著名的图片分享应用 Instagram 大约会产生 5200 万张图片，Facebook 约为 3 亿张，在中国非常流行的微信每天

---

<sup>1</sup><https://www.tineye.com/>

<sup>2</sup><https://images.google.com/>

<sup>3</sup><http://image.baidu.com/?fr=shitu>

<sup>4</sup><http://pic.sogou.com/>

朋友圈上传的图片数量则达到了 10 亿级别。这些海量的图片给图片的管理与检索带来了挑战，同时，在社交平台上，敏感图片的散布也会造成不良的影响甚至严重的后果。很多 Facebook 用户会在个人账户或者一些群组上传一些枪支图片 [12, 13]，微博用户也在微博上散发一些暴力图片等，这些敏感图片需要适当的监管与处理，以避免引起严重后果，这些需求都可以使用图像检索相关的技术来满足。

近些年，随着电脑硬件的不断进步，特别是图形处理器（graphics processing unit, GPU）的普及，以及诸如 ImageNet [14] 级别的大规模图像数据库的发布，使得深度神经网络的训练成为可能，基于深度学习的方法成为主流的机器学习方法。深度卷积神经网路是深度学习方法 [15] 中常用的模型之一，卷积神经网络具有强大的数据表达能力，在图像分类，物体检测，图像分割，目标跟踪等领域都取得了巨大的成功。随着它的流行 [16]，研究者们也开始尝试把卷积神经网络与图像检索的技术结合，通过神经网络来直接提取具有区分性的图像特征，或者通过度量学习的方式，更新网络参数，学习图像的特征。这一领域的进展迅速，在短短几年时间内，现有的一些方法 [17] 在图像检索常用数据集 [7, 8, 18, 19] 上的效果已经超越了传统的基于 SIFT 特征的视觉词袋模型等方法。

现有的一些基于神经网络的方法未能全面探索影响提取的特征有效性的因素，如何提升神经网络提取的图像特征的有效性仍然需要很多的研究，另外，图像检索依然面临着很多的挑战，例如算法不能有效处理图像中物体的视角和姿态的大幅度变化，复杂背景图像的干扰以及光照变化等都会造成图像检索算法准确度的下降。目前为止，这些问题都没有得到很好的解决，我们仍然需要研究更准确与更高效的算法来解决上述难题。基于上述背景，并且结合敏感图像检索的需要，我们在本文中主要研究基于深度卷积神经网络的敏感目标检索技术。

### 1.3 论文的组织

本文其它各章的组织如下：

第 2 章为图像检索相关方法的综述。该章先介绍了图像检索领域一些早期的工作，包括使用颜色，纹理特征的检索方法，然后我们介绍了基于 SIFT 局部特征描述子的方法，其中包含基于视觉词袋模型的方法以及基于 VLAD 和 Fisher Vector 的方法。最后我们介绍了卷积神经网络的一些基本概念，然后介绍

了直接从已有的网络模型提取图像特征进行检索的方法以及对神经网络进行微调进行检索的方法。

第 3 章介绍我们建立的大规模枪支图片数据库 Firearm14k。主要介绍了枪支图片的收集与清理过程，然后介绍了图片的标注，标注人员标注质量的评估以及图片的筛选过程。最后我们列出了数据库的一些特点，我们也展示了一些示例图片，介绍了数据库的划分（训练集，验证集，测试集）过程，给出了数据库的具体统计信息。

第 4 章提出了一种多尺度全卷积的图像实例检索方法。我们先介绍了已有一些利用卷积神经网络提取图像特征的方法没有详细研究的问题，然后详细介绍了影响神经网络提取的特征有效性的三个因素：输入图像尺寸的变化策略，多尺度特征表达，降维以及白化矩阵的学习。我们通过大量的实验确定了三个因素对特征有效性的影响，在此基础上我们提出了多尺度全卷积的图像实例检索方法。最后，我们在多个数据库进行了实验，实验结果表明我们提出的方法有着良好的检索性能。

第 5 章提出了一种基于双阈值对比损失函数的枪支图像检索方法。我们发现，将传统的基于单阈值对比损失函数的方法应用到枪支图像检索问题，不能取得很好的检索精度，一方面，这是由于在训练过程中，正负样本对产生的损失不平衡，另外一方面，这是由于 Firearm14k 数据库与 ImageNet 存在了很大差异。为了解决这些问题，我们提出使用双阈值对比损失函数，并结合两步训练的策略来微调神经网络模型。在该章，我们首先介绍了提出的方法，然后给出了一些对比实验的结果以及可视化的结果，最后，我们把提出的方法和当前主流方法进行了对比，实验结果表明我们的方法在不同的特征维度下都取得了很好的检索精度。

第 6 章对本文的研究工作做一总结，并对未来的研方向进行了展望。

## 第2章 图像检索方法综述

### 2.1 概述

图像检索是一个经典的研究问题，研究者们围绕这个问题提出了很多的解决方法。图像检索问题的核心是如何表示图像的内容，也就是说如何用数字化的方式来表达图像中的物体和场景，这种数字化的表达可以称之为图像的特征，另外一个重要问题是衡量图像特征之间的相关性（relevance）或者相似性（similarity），大量的研究都围绕这些问题展开。

早期的图像检索方法试图从图像中提取一些局部或者全局的颜色以及纹理信息作为图像的特征，然后使用不同的图像相关性计算方法来度量两个图像之间的相似性，由于当时计算资源等的限制，这些方法使用的特征大都比较简单，并且使用的图像数据库规模并不大，多在几百到一千的范围内。

到了 2000 年以后，随着 SIFT [20] 提出，由于 SIFT 特征描述子对图像旋转，尺度缩放，光照变化等因素有一定的鲁棒性，因而成为主流的图像局部特征描述方法。此后的图像检索方法大都使用基于 SIFT 局部特征描述子的方法来提取图像特征。基于 SIFT 特征的视觉词袋的方法（bag of features，BOF）是一种常用的图像特征表示方法，研究者在该方法基础上进行了大量的改进与创新，使得该方法成为一种检索效果出色的方法。除了 BOF 方法，也有研究者研究基于 SIFT 特征的局部描述子聚合向量方法（vector of locally aggregated descriptor，VLAD）或者 Fisher 特征向量方法等。

随着深度学习方法的兴起，卷积神经网络在图像分类 [14, 21–23]，物体检测 [24–27]，图像语义分割 [28–30] 等任务上取得了超越传统方法的结果，图像检索领域的研究者们将研究方向转向了基于卷积神经网络的图像检索，提出了一系列的方法，例如，直接提取图像的全连接层或者卷积层的输出，作为图像的特征，或者对已有的神经网络模型进行微调，从而改进得到的特征在检索任务上的效果等。

在本章接下来的部分，在 2.2 节我们首先介绍一些早期的研究图像检索方法的探索性工作，接着在 2.3 节我们将介绍基于 SIFT 特征描述子的图像检索方法。然后，在 2.4 节，我们首先对卷积神经网络的相关发展及必要的概念进行简要回顾与介绍，然后介绍基于卷积神经网络的图像检索方法。最后在 2.5 节，我们对本章的内容进行总结。

## 2.2 早期图像检索方法

颜色是图像的一种重要特征，Swain 和 Ballard [31] 首先提出使用彩色图像的多维颜色直方图作为图像的特征，他们使用直方图交叉（histogram intersection）方法确定两个图像之间的相关性，他们发现颜色直方图特征能够在一定程度上对物体的在图像上的平移，遮挡以及视角变化等有一定鲁棒性。然而，Swain 和 Ballard 的方法对图像的光照有一定的要求，各个图像的光照条件不能差异太大，如果光照差异太大，该方法需要首先对图像进行预处理 [32] 来消除光照变化影响，为了克服这个问题，Funt 和 Finlayson 提出了光照无关的图像颜色直方图特征提取方法 [33]。Deng 等人 [34] 则提出使用聚类方法对图像局部颜色聚类，得到具有代表性的颜色，然后分别使用这些有代表性的颜色来匹配包含这些颜色的图像，最后来自不同代表性颜色的匹配被融合在一起，得到最终的检索结果。也有一些研究者尝试采用纹理以及形状等信息作为图像特征，Manjunath 和 Ma [35] 提出，可以设计 Gabor 小波的滤波器，利用 Gabor 小波来分析和代表图像的纹理信息。Niblack 等人 [2] 在 IBM 的 QBIC (query by image content) 项目中则使用了多种特征来表示图像，包括颜色，纹理，形状以及速写特征等，其中颜色特征采用了直方图特征，纹理特征则结合粗糙度，对比度以及方向特征等，形状特征则结合二值画图像中的形状区域面积、圆度、奇异度以及代数矩不变量等信息，速写特征则是基于降低分辨率以后的图像边缘图信息。Jain 和 Vailaya [36] 提出使用颜色以及形状信息进行图像检索，他们使用的形状特征是由图像的边缘信息得来的（使用 Canny 边缘提取算法提取图像的边缘信息）。更多的关于早期图像检索的方法，可以参考 Smeulders 等人 [37] 的综述。

## 2.3 基于 SIFT 局部特征的检索方法

SIFT 是由 Lowe [5] 提出的一种图像局部特征描述子，具有良好的旋转以及尺度不变性，同时对图像的仿射变换以及光照变化和噪声等都有一定的鲁棒性，因而被广泛用于描述图像的特征。传统的图像检索中的很多方法都是基于 SIFT 描述子来得到图像的特征，然后再进行图像检索的后续过程。基于 SIFT 描述子的方法，常用的方法，一类是基于 BOF 方法，另外一类是基于 VLAD 以及 Fisher 向量的方法。

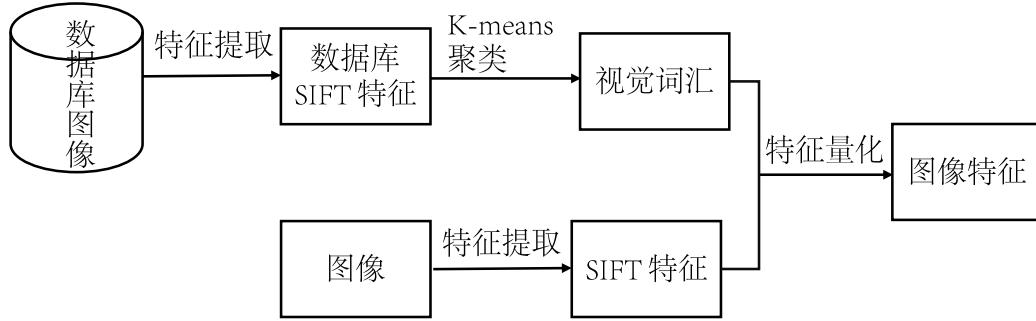


图 2.1 视觉词袋模型的工作流程

### 2.3.1 基于视觉词袋模型（BOF）的检索方法

视觉词袋模型（BOF）[\[38–40\]](#)来源于信息检索（information retrieval, IR）领域非常流行的词袋（bag of words, BOW）方法[\[41, 42\]](#)，该方法将图像看成是一系列「视觉词汇」的集合，通过统计图像中各个视觉词汇的频率作为图像的正特征，该方法被广泛应用到图像检索领域，也常用于图像分类等任务。图 2.1 展示了 BOF 方法的工作流程，具体的工作原理如下：

步骤一：数据库图像 SIFT 特征提取。在这个阶段，通常使用密集采样[\[43, 44\]](#)的方式把图像划分成规则的小块，然后提取局部特征，或者采用一些兴趣点检测算法[\[45\]](#)检测到一些尺度及仿射不变的兴趣点，然后使用 SIFT 局部特征描述子提取算法提取数据库中每张图像的 SIFT 特征。

步骤二：视觉词汇的计算。使用 K-means 等聚类算法，对数据库中所有图片得到的 SIFT 特征进行聚类，得到  $K$  个聚类中心，该聚类中心称为视觉词汇（visual words），其中  $K$  是一个可以调节的超参数。

步骤三：量化图像局部特征。对于每一张图片，把从该图像提取的 SIFT 特征描述子分配给距离最近的视觉词汇，类似投票的过程。

步骤四：计算图像特征。得到量化的图像特征以后，计算视觉词汇出现的频率，经过归一化等操作，该特征就是该图像的特征，用于后续的分类或者检索等任务。

Sivic 和 Zisserman[\[6\]](#)于 2003 年第一次提出使用 BOF 方法来进行图像检索，在该工作中，作者借鉴使用了逆文档频率（inverse document frequency, IDF），停用词（stop words）等文档检索中的常用技术。Sivic 与 Zisserman[\[6\]](#)文章中用到的数据库图片样本比较少，如果数据库样本比较大，从所有图片中提取的 SIFT 描述子的数目将非常庞大，传统的 K-means 算法无法有效聚类大量的 SIFT 特征描述子，从而生成视觉词汇，因此 Philbin 等人[\[8\]](#)对比了两种 K-means 的改

进方法：近似 K-means（approximate K-means, AKM）以及层次 K-means 方法（hierarchical K-means, HKM）。作者使用了不同的视觉词汇大小来测试两种聚类方法的性能，发现 AKM 方法比 HKM 方法能够得到更好的检索效果。传统的 BOF 方法，得到视觉词汇以后，对于一张图片中的 SIFT 描述子，其权重会完全分配给距离该描述子最近的一个视觉词汇，如果一个 SIFT 描述子同时距离两个视觉词汇都非常相近，以这样的分配方式来计算词频向量显然是有问题的，会损失图像原有的信息。从直觉上来说，如果一个描述子同时距离几个视觉词汇都很近，那么显然该描述子可以按照距离远近给这几个视觉词汇同时赋予一定的权重，Philbin 等人 [7] 即是按照这个思路对原有的 BOF 方法进行了改进，作者称之为软分配（soft assignment）。另外一种对原有 BOF 方法的改进来自 Jégou 等人 [19]，该文章提出汉明空间嵌入方法（Hamming embedding, HE），对于查询图像与数据库图像，如果分别来自这两张图像的两个描述子对应于同一个视觉词汇，那么这两个描述子对图像相似度得分不一定有贡献，我们需要把原有的视觉词汇所在的空间进行进一步细分，得到每个描述子在这个细分的空间中的一个二进制编码，只有当这两个描述子编码的距离小于一个阈值，才对两幅图像之间的相似度有贡献。传统的 BOF 方法得到特征通常维度相当高，不适用于大规模图像检索场景，因此 Jégou [46] 提出近似方法来代表 BOF 特征，压缩图像特征维度，他们提出的方法的检索速度要比传统的 BOF 快一个数量级。

查询扩展（query expansion, QE）也是信息检索领域常用的一种方法，QE 就是把检索返回的结果与初次的查询目标进行融合，试图弥补原有方法对初次查询信息表达的不足，希望能够得到原有查询的更加全面的表达，Chum 等人 [11] 对这个方法在图像检索领域的使用进行了全面的研究，对不同的查询扩展的方式进行了试验，试图确定最佳的查询扩展方式。也有论文从其他方面进行了探索，Mikulík 等人 [9] 针对原有的 BOF 方法计算图像相似度的不足，提出了一种使用概率方法估计查询图像和数据库图像相似度的方法，该相似度通过非监督方式学习得到，效果要优于 HE 以及 soft assignment。Arandjelović 和 Zisserman [10] 则试图对原有的 SIFT 特征描述子进行改进，提出使用 RootSIFT 来代替原有的 SIFT 特征，发现能够取得更好的效果。

### 2.3.2 基于 VLAD 和 Fisher Vector 的检索方法

BOF 方法把图像看成是没有空间关系的一系列小块的组合，Fisher Vector 方法 [47] 把一张图片上所有 SIFT 特征描述子看成是对一个高斯混合模型（Gaussian mixture model, GMM）采样得到的样本点，在此基础上，可以得出高

斯混合模型相对于每个高斯分量的平均值的梯度，然后把对于所有分量的梯度拼接起来，作为图像的特征表达。原始的 Fisher Vector 维度很高，并且是非稀疏的，Perronnin 等人 [48] 提出一种简单的压缩办法来给 Fisher Vector 降维，发现这种方法效果要优于哈希的方法。Douze 等人 [49] 则试图把属性特征与 Fisher Vector 融合起来作为图像特征来进行图像检索。

Fisher Vector 特征计算复杂，计算量大，受 Fisher Vector 启发，VLAD 简化了 Fisher Vector 的操作，可以认为是 Fisher Vector 的简化版本。VLAD 最早由 Jégou 等人于 2010 年提出 [50]，VLAD 方法同样需要计算 BOF 中用到的视觉词汇，不同于 BOF 的是，VLAD 方法把图像中的局部特征描述子分配给视觉词汇以后，求取分配给某个视觉词汇的特征描述子与该视觉词汇的残差和（residual sum），作为图像特征表达的一个分量，最后将对应于每一个视觉词汇的残差和向量拼接起来，作为图像的特征表达，因此 VLAD 特征维度要高于对应的 BOF 特征<sup>1</sup>。Arandjelović 和 Zisserman [51] 对原始 VLAD 操作的一些步骤进行了改进，提出对图像 VLAD 特征的每一个残差和分量分别归一化，而不是整体拼接再进行归一化，作者发现这样的归一化操作可以减少图像的突变特征 [52]，有助于提升最终检索的精度，同时作者也提出使用 Multi-VLAD 来替代原始的单一尺度下的 VLAD，可以认为 Multi-VLAD 是普通 VLAD 方法的多尺度版本。

## 2.4 基于卷积神经网络的图像检索方法

### 2.4.1 卷积神经网络简介

卷积神经网络（convolutional neural network, CNN）由 Le Cun 等人提出 [53, 54]，最初用于邮政编码图像中手写数字的识别问题。不同于传统的多层感知机神经网络（multi-layer perceptron, MLP），CNN 专门为处理图像内容所设计，因此包含一些 MLP 没有的结构，典型的 CNN 包含卷积层（convolutional layer），池化层（pooling layer），非线性激活函数（non-linear activation）以及全连接层（fully-connected layer）等结构。其中，卷积层包含一系列的滤波器，每个滤波器都包含多个滤波核（filter kernels），通过训练，这些滤波核可以学习到图像中不同的模式 [16, 55]，卷积层的权重被设计为共享方式，通过滑动窗口（sliding window）的方式提取上一层的输出信息，相比于传统的 MLP，大大降低了参数的数量，卷积层的设计也使得网络对图像中物体的平移有一定的鲁棒性；卷积层后面通常是池化层，通过池化操作，CNN 对物体的扭曲以及变化

<sup>1</sup>具体的计算，可以参考 <http://www.vlfeat.org/api/vlad-fundamentals.html>

有一定的鲁棒性；非线性激活函数可以有多种选择，如 Tanh 函数，ReLU [56]，PReLU [57] 和 Leaky ReLU [58] 等，非线性激活函数使神经网络不仅仅能够表达简单的线性关系，也能够表达复杂的非线性关系；全连接层则相当于分类器，对图像进行分类。在神经网络的结构中，卷积层，池化层，非线性激活函数通常构成一个完整模块，该模块被重复多次，形成多层神经网络，通过这种由低层到高层的结构，网络能够在低层学习图像的一些低级特征，如线条，形状等特征，在高层，则能够学习到更加高级与抽象的特征，如代表图像类别的语义信息 [16]，这种层级结构增强了卷积神经网络的特征表达能力。

虽然卷积神经网络在 90 年代初已经被提出，但是并未引起很大的轰动，一方面因为当时的硬件无法支持大规模的训练，另外一方面，当时也不存在大规模的图像数据库。即使到了 2000 年以后，人们还在使用受限玻尔兹曼机（restricted Boltzmann machine，RBM）[59, 60] 来学习图像的特征，深度 RBM 训练十分复杂，需要首先进行逐层的无监督的预训练，最后才能进行监督式的训练。到了 2012 年，随着 GPU 等硬件的成熟以及 ImageNet [14] 大规模数据库的提出，Krizhevsky 等人首次 [16] 展示，通过使用大量的训练数据直接对深度神经网络进行监督式训练的方式，完全可以在图像分类任务上取得大幅度超越前人的结果。在这之后，研究者对深度神经网络的研究掀起了热潮，神经网络的结构以及深度也在不断进化 [21–23, 55, 61]；另一方面，研究者也将基于深度卷积神经网络的方法应用到各个领域，如物体检测 [24–27]，图像语义分割 [28–30]，图像风格转换 [62, 63]，生成对抗网络 [64, 65]，深度强化学习 [66, 67] 等等。

当然，也有研究者将研究转向基于深度卷积神经网络的检索方法，基于深度学习的方法，按照是否需要进行训练，可以分为两类方法，第一类是基于已有网络模型，直接提取图像特征（off-the-shelf）的方法，第二类则是在已有模型基础上，对模型的结构进行修改，然后对模型参数进行微调（fine-tuning）的检索方法。我们之前的工作 [68]（见第 4 章）属于利用已有的网络提取特征的方法，在第 5 章，我们提出的基于双阈值对比损失函数的方法则是属于微调神经网络的方法。

#### 2.4.2 基于神经网络直接提取图像特征的检索方法

基于已有神经网络模型提取图像特征的方法，通常采用在 ImageNet 1000 类分类数据库上训练的模型，如常见的 AlexNet [16]，VGGNet [21]，ResNet [23] 等。该方法在输入图像以后，提取 CNN 的全连接层或者是卷积层输出，进行特

征的聚合或者是后处理操作，把得到的特征作为图像的特征，这种方法的优点是速度快，不需要对网络再进行有监督的训练，适用于数据库样本量很小无法进行训练的情况。Krizhevsky 等人 [16] 在原始的 AlexNet 文章中就指出，可以用 CNN 全连接层的输出作为图像特征，进行粗略的检索任务，如果两幅图像的特征之间的欧式距离很小，那么说明两幅图像是相似的。Razavian 等人 [69] 直接使用全连接层的特征作为图像的特征表达，在进行图像检索时，他们提取了查询图像与数据库图像在不同尺度的小块，用滑动窗口的方式确定每个尺度下查询图像与数据库图像的相似度，然后取各个尺度下的最大相似度的平均值作为查询图像与数据库图像的相似度，该方法效果不错，但是缺点是十分耗时，提取不同尺度下的图像小块特征需要消耗大量的时间。Gong 等人 [70] 提出了一种被称为 MOP (multiple orderless pooling) 的图像特征表示方法，该方法使用 AlexNet 全连接层输出作为一张图像的原始特征，在三个尺度上提取图像特征，在尺度 2 和 3 上，对图像块特征使用 VLAD 方法进行特征聚合并且使用 PCA 对特征降维。最后将三个尺度的特征拼接起来，作为图像最终的特征表达，显而易见，该方法同样非常耗时。Ng 等人 [71] 从卷积神经网络的不同层提取特征，并且结合 VLAD 编码方法，得到图像的特征。与此同时，Babenko 等人 [72] 则研究了使用 CNN 哪一层的特征更有效的问题，他们的研究发现，CNN 全连接层的输出通常代表图像整体上的信息，更适合分类任务，对于图像检索任务，这种特征会缺乏一些细节信息，而中间的卷积层特征，有更丰富的图像的局部信息，因而检索效果更好。

Babenko 和 Lempitsky [73] 尝试使用 CNN 的卷积层输出的特征图 (feature map) 作为图像的特征表达，由于卷积层输出的特征图是二维的，作者提出 SPoC (sum pooling of convolutions) 方法来聚合特征，该方法对特征图的元素进行加权求和，作者还假设物体一般位于图像中间，因此特征图中间的元素被赋予更大的权重。Tolias 等人 [74] 也选择使用卷积层的输出作为图像的特征，但是与 Babenko 和 Lempitsky [73] 的做法不同，该论文对特征图使用最大池化方法，为了增强图像特征有效性，作者提出了 R-MAC (regional maximum activation of convolutions) 方法，该方法为多尺度特征表达方法，在求图像区域特征时，并不直接把图像区域重新送入 CNN 网络提取特征，而是仿照 ROIpooling 的想法 [27]，假定图像区域和特征图上的区域存在线性映射，直接在特征图上求图像区域特征，最后对不同尺度的特征进行融合。Seddati 等人 [75] 则提出了对 R-MAC 方法的一些改进。Zhou 等人 [76] 提出结合经典的 SIFT 特征与神经网络

提取的图像特征，共同编码图像，也取得了良好的效果。

#### 2.4.3 微调已有网络的检索方法

直接从已有网络提取图像特征的方法有一定的局限性，因为这些模型都是针对图像分类任务训练，并不是特别针对图像检索的任务，并且这些模型都是在 ImageNet 数据库上训练，并不能很好地应对数据库风格发生巨大变化的情况<sup>2</sup>。因此，针对特定的数据，对网络进行微调，能够进一步提升图像检索的准确率。

Babenko 等人 [72] 针对检索任务，重新收集了一个较大的 Landmark 数据库<sup>3</sup>，在这个数据库，使用分类损失函数，对 AlexNet 网络的参数值进行了微调，作者发现这种微调对于提升在某些数据库上的检索准确率有一定帮助，因为这些数据库与微调使用的数据库图片比较类似。值得一提的是，该文章直接使用的是全连接层的输出，并未使用多尺度的方法，另外，虽然他们对网络进行了微调，但是使用的是还是分类的损失函数，并未针对检索任务设计其他损失函数。在检索任务中，通常希望相似或相关的图像与查询图像在特征空间的距离要小于不相似图像与查询图像的特征在特征空间的距离，为了这个目标，研究者也提出了各种基于度量学习的方法。三元组网络（triplet network）[77, 78] 在度量学习中应用广泛，Wang 等人 [77] 较早使用 triplet network 进行图像检索的工作，他们收集了一个大规模的图像库，作者在文中提出使用多尺度的图像特征，并且提出了采样训练所用的 (anchor, positive, negative) 三元组的算法。Arandjelović 等人 [79] 将弱监督形式下的三元组损失（triplet loss）应用到地点检索任务上，并结合传统的 VLAD 方法，提出了 NetVLAD 层，该方法能够有效进行地点检索，学习的特征区分度高，能够有效忽略一些背景的干扰。现有图像检索任务，图片中物体并未对齐，因此物体在图片中位置以及大小都不确定，这给提取特征带来了巨大挑战，Gordo 等人 [17] 提出使用 region proposal network (RPN)，找出图像中可能的物体区域，从而提取更加有效的特征，同时作者提出了一种自动清洗噪声数据的方法，并且采用算法自动估计图片中物体的位置，避免了耗时费力的人工标注。Siamese 网络 [80, 81] 也是一种常见的度量学习网络，Radenovic 等人 [82] 使用了两个分支的 Siamese 网络，利用单阈值对比损失函数来监督网络的训练，该论文使用了 structure from motion 技术，构建三维模

<sup>2</sup>在第 5 章，我们将会展示，直接利用已有模型提取特征的方法在我们建立的 Firearm14k 数据集上效果十分不理想

<sup>3</sup>该数据库获取地址：<http://sites.skoltech.ru/compvision/projects/neuralcodes/>

型来挑选训练图像对，这是一种全自动的方法，不需要人工的干预即可选出训练样本。以上提到的这些方法都是学习到了图像的全局特征，全局特征对图像检索任务有时并不有效，因此 Noh 等人 [83] 提出利用分类任务并结合显著性检测方法学习深度局部特征（deep local features），然后使用传统的图像匹配的算法来计算图像相似度，在他们收集的超过 100 万的数据库上效果超过了现有的方法<sup>4</sup>。

## 2.5 本章小结

本章中，我们简要回顾了基于颜色、形状等特征的传统图像检索方法，然后我们介绍了基于 SIFT 特征的方法：分别是基于 BOF 的图像检索方法以及基于 VLAD 和 Fisher Vector 的图像检索方法。在 2.4 节，我们介绍了两种常见的基于深度卷积神经网络的图像检索方法，第一种是基于已有神经网络直接提取图像特征的方法，该方法简单且无需额外训练数据，但是精度不高且有一定限制，第二种方法是微调已有模型的方法，这种方式针对检索任务，采用端到端训练的方式优化模型参数，在检索任务上能够取得更好的效果。

---

<sup>4</sup>该数据集获取地址：<https://www.kaggle.com/c/landmark-retrieval-challenge>



## 第3章 大规模枪支图片数据库的构建

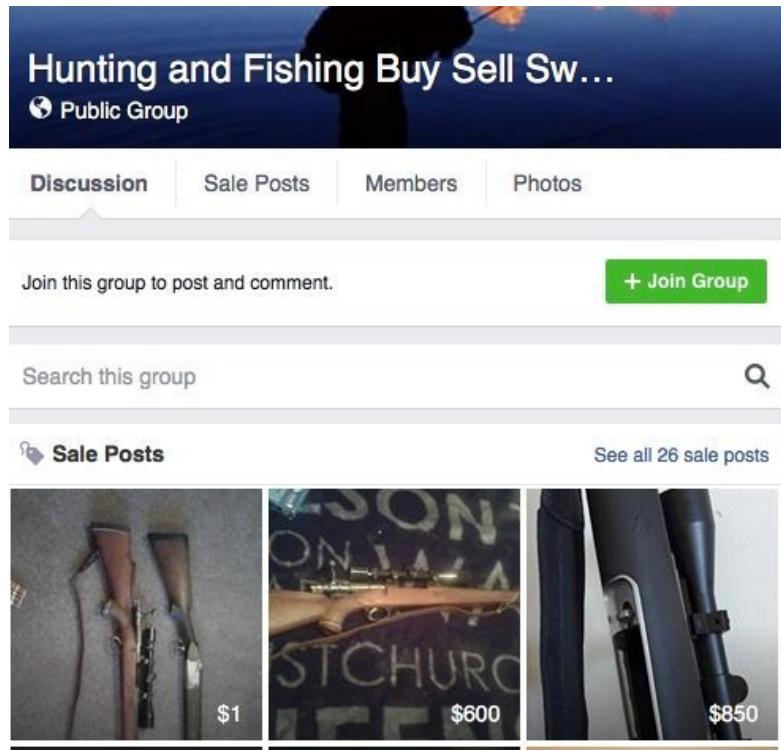
### 3.1 引言

在计算机视觉领域的每个研究方向，都存在着一些为大家所熟知的数据库，例如，在物体检测与物体分割领域，常用的有 PASCAL VOC 数据库 [84] 以及 Microsoft COCO 数据库 [85]；在人脸识别领域，则有 LFW (labeled faces in the wild) [86]，YTF (YouTube Faces) [87] 以及 MegaFace Challenge 数据库 [88]。一个图片数量丰富，标注准确的数据库，对于推动一个研究领域的发展有着重要的作用。正是由于 PASCAL VOC 数据库和 ImageNet 数据库 [14] 的提出，研究者可以在同一数据库用相同的标准比较自己的算法，才推动了研究者对图像分类方法的研究，从而使得图像分类的准确率不断提高，直至超过人类的水平，一个好的数据库对相关研究的推动可见一斑。



图 3.1 UKB 数据集中一些示例图片

图像检索领域，也有一些通用的数据库，最常用的数据库有牛津大学 Visual Geometry Group (VGG) 建立的 Oxford building 数据集 [8]（简称 Oxford5k），Paris building 数据集 [7]（简称 Paris6k），法国 INRIA 建立的 Holidays 数据集 [19] 以及 Nistér 和 Stewénius 建立的 UKBench 数据集 [18]（简称 UKB）。这些数据库的提出时间都较早，建立的时间都在基于深度学习的图像检索方法兴起之前，因此这些数据库的规模都较小，例如 Holiday 数据集只有 1491 张照片，Paris6k 和 Oxford5k 的规模在 5000–6000 张；另外，有的数据库的图片比较简

图 3.2 Facebook 上一个枪支交易群<sup>1</sup>

单，例如 UKB 数据集，如图 3.2 所示，相似图片均是同一物体或场景在不同角度下的照片，因此识别起来相对简单。

在深度学习兴起以后，基于卷积神经网络的图像检索方法也开始流行，众所周知，基于深度学习的方法，在网络的训练过程中往往需要大量的训练图片，否则学到的模型很容易过拟合，即使我们是在已有的网络的基础上进行微调 (fine-tuning)，也需要一定量的图片，才能取得不错的检索精度。另一方面，在目前的社交网络上，充斥着各种各样的枪支图片，例如许多 Facebook 的用户出于不同的目的在自己的 Facebook 账户上传枪支照片 [12, 13]，这些枪支图片可能会激起暴力，也可能会助长枪支交易的泛滥（参见图 3.1），因此有必要对枪支图片进行必要的监管与处理。基于这两点考虑，为了方便研究者针对这方面的应用进行研究，我们收集了一个大规模的枪支图片数据库，该数据库包含 167 类不同类型的枪支，总图片数为 14755 张，我们将其简称为 Firearm14k。

本章的结构组织如下：3.2 节介绍数据库图像的收集与清理过程，3.3 节介绍枪支图像的标注过程，然后 3.4 节介绍我们如何对各个标注人员的评分质量进行评估，从而筛选出有效的图片。3.5 节介绍枪支数据库的划分（训练集，验证集，测试集），给出各个集合的具体信息，介绍数据库的一些特点，并且给出一些数据库中的示例图片。

## 3.2 数据库图片收集

### 3.2.1 图像的收集

为了构建我们的枪支数据库，首先我们需要找到一些关键词，我们从一些论坛以及网站搜集了一批枪支名称的关键词。由于一个枪支可能会有不同的型号以及很多仿照的型号，例如，根据维基百科，*M16*<sup>2</sup>这个枪支就有至少三种不同的型号，分别为 *M16A2*, *M16A3*, *M16A4*，同时也有一些仿照的枪支，例如 *XM177*。因此有必要处理收集到的关键词列表，去掉重复的关键词。我们参考维基百科对各个枪支类型的介绍，认真审查了得到的关键词列表，确保一个枪支和它的不同型号以及仿照枪支不同时出现。经过人工的过滤以后，我们总共得到了 167 个关键词，这些关键词对应着不同的枪支类型，如「来复枪」，「手枪」，「散弹枪」等。

为了抓取图片，我们编写了一个简单的图片爬虫，用来下载图片。在得到这些枪支名称关键词以后，我们使用谷歌图像搜索引擎<sup>3</sup>，把这些关键词作为查询，提交到搜索引擎。对于每一个关键词，由于对应的枪支类型流行程度不同，因此返回的结果数目也不同，我们使用图片爬虫下载其中 300 – 500 张图像作为初步的结果。

### 3.2.2 图片的清理

完成图片的收集以后，我们对图片进行了清理。我们主要去掉了以下几类图片：

- 损坏的图片（由于下载时候出现的错误，某些图片未能完整下载）
- 灰度图片（grayscale images）
- 最小边小于 128 的图片
- 非 JPEG 格式的图片（包括 GIF, TIFF, PNG 等格式）

经过图片的清理过程，我们总共得到了 62642 张图片，在后续过程中，我们将对这些图片进行人工标注，然后进行评分与过滤，最后得到最终的枪支图片数据库。

---

<sup>1</sup>图片来源: <https://nyti.ms/2JksjoL>

<sup>2</sup>[https://en.wikipedia.org/wiki/M16\\_rifle](https://en.wikipedia.org/wiki/M16_rifle)

<sup>3</sup><https://images.google.com/>



图 3.3 在线图像标注系统的前端评分界面

### 3.3 图像标注

尽管根据我们提供的查询关键词，谷歌图像搜索引擎搜索到的图片质量已经相对较高，但是不可避免地，利用检索词搜索到的图片还存在一定的噪声（也就是说，根据某个检索词抓取的图片仍有一部分并不是对应于该检索词），因此我们必须要对每一类图片<sup>4</sup>进行人工的标注，找出真正属于该类的图片。为了达到这个目的，我们专门设计了一个简单的图片在线标注系统，标注人员可以通过网页化的界面 (Web Interface) 对图像进行打分，从而帮助我们找出真正有效的图片。我们的图片在线标注系统的前端界面如图 3.3 所示。

对于每一类图片，我们事先人工挑选出五张真正属于该类的示例图片，然后我们对标注者进行了培训，指导他们标注应该如何进行。在标注过程中，标注者会根据当前要标注的图片（图 3.3 右上部分）与该类的示例图片（图 3.3 左边）的相似程度，对要标注的图片打分，分数范围为 [0,9]（0 分代表当前图片与示例图片毫不相关，9 分代表当前图片与示例图片最相似）。我们总共招募了 5 名标注者参与标注工作，每名标注者独立对所有的图片进行打分。因此，在标注人员完成图片标注任务后，对于每一类图片下的每张图片，我们都有 5 个评分。

<sup>4</sup> 我们把通过同一个查询关键词抓取的图片当作一类

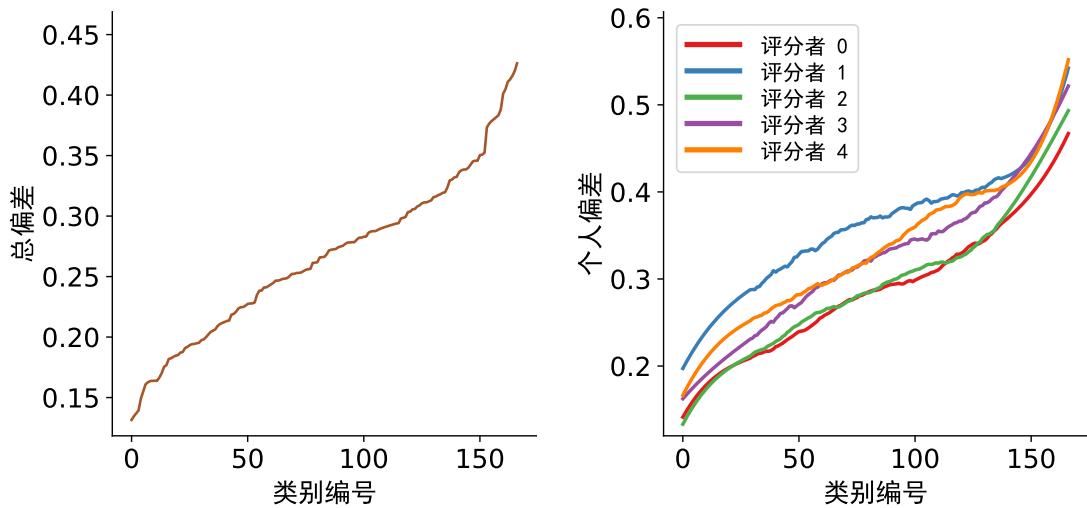


图 3.4 枪支数据库打分数据的偏差统计信息。左：总体的偏差信息；右：个人的偏差信息

### 3.4 标注质量评估及图像筛选

#### 3.4.1 图像标注质量评估

不可避免地，不同的标注者的标注质量并不是相同的，标注质量的高低取决于很多因素，例如，标注者在标注过程中是否认真，以及标注者对于相似这个概念的认识也可能有一定的差异。为了给每一类下的每一张图片一个可靠的评分，我们必须科学地对标注者的标注质量进行评估。

首先，对于每一类图像，我们计算出不同打分者对该类图片评分之间的相关系数，我们用一个矩阵  $C$  来表示打分相关性信息。矩阵  $C$  的大小为  $5 \times 5$ ，每个元素  $C[i, j]$  代表第  $i$  个标注者与第  $j$  个标注者的评分之间的相关性，显然矩阵  $C$  是一个对称矩阵。在理想情况下，两个标注者对某一类下面的图片评分肯定会有所不同，但是他们的打分相关性应该足够高，因此矩阵  $C$  的每个元素都应该接近 1。如果一个标注者标注的质量不高，那么他/她的评分和其他人的评分之间的相关关系就会比较弱，我们可以据此定义每一类的总偏差（overall deviation, OD）和个人偏差（individual deviation, ID）。对于某一类图片，总偏差可以定义为：

$$OD = \frac{15 - \sum_{i=0}^4 \sum_{j=i}^4 C[i, j]}{15}, \quad (3.1)$$

对于标注者  $i$ ，个人偏差可以定义为：

$$ID[i] = \frac{5 - \sum_{j=0}^4 C[i, j]}{5}. \quad (3.2)$$

表 3.1 每个标注者有最大的个人偏差的类别数目

标注者编号	0	1	2	3	4
数目	4	93	4	33	33

在以上两个公式中， $C$  是对应某类的相关性矩阵。

在图 3.4 中，我们展示了偏差信息的总体分布情况与个人分布情况：左图展示的总体的偏差分布情况，我们把偏差按照从小到大进行了排序；右图展示的是每个人的偏差分布情况<sup>5</sup>，和总体偏差采用了相同的类别编号顺序。一个标注者在某类上偏差越大，那么他在该类上的评分就越不可靠。从图 3.4 中可以看出，标注者 1 在整体上有着最大的个人偏差（这意味着比较差的标注质量），标注者 3 和 4 相对来说，个人偏差较小，因此他们的标注质量较好，标注者 0 和 2 的个人偏差最小，因此他们的标注质量是最好的。

然后，我们统计了每个标注者在多少个类别上具有最大的个人偏差，这个指标可以用来衡量每个人的评分质量。在表 3.1 中，我们列出了统计数据。从中可以看出，在所有的 167 类中，标注者 1 在其中 93 类有最大的偏差值，这表明标注者 1 总体的标注质量很差；标注 3 和 4 在其中 33 类上有最大偏差值，这表明他们的标注质量中等，对于标注者 0 和 2，他们分别只在 167 类中的 4 类有最大的偏差值，这表明他们的标注质量是极其优秀的。

### 3.4.2 有效图片筛选

为了选出每一类的真正有效的图片，我们需要选出和给出的示例图片相似性得分高的图片。对于每一类别下的每张图片，我们使用五个标注者打分的加权平均分作为每张的图片的相似性得分。由于标注者 1 的评分质量太低，我们把标注者 1 的权重设为了 0，对于其他标注者，权重反比于该标注者在多少个图像类别上有最大的个人偏差，这些权重的和为 1，我们最终使用的权重为

$$w = [0.445, 0, 0.445, 0.055, 0.055]$$

对于每类下的每张图片，其最终的相似性得分  $s$  为

$$s = \sum_{i=0}^4 w_i s_i, \quad (3.3)$$

<sup>5</sup>为了滤除一些噪声，显示出趋势，我们对曲线进行了平滑处理

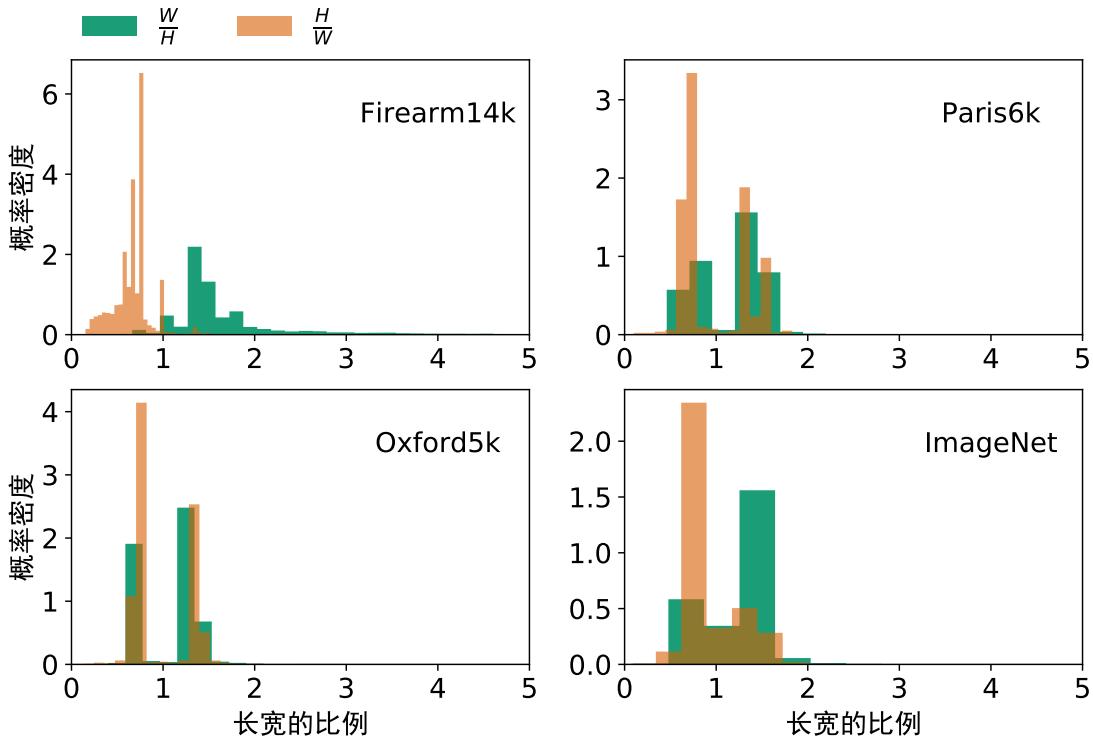


图 3.5 Firearm14k 数据与其他数据库图片尺寸比值的概率分布

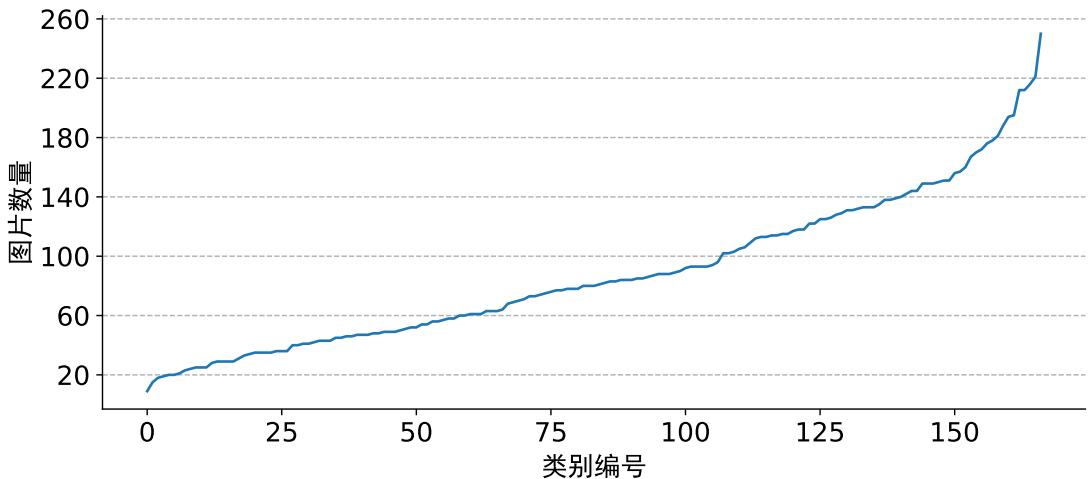


图 3.6 枪支图像数据库各个类别图片数量

其中  $s_i$  表示第  $i$  个标注者对图片的评分。然后我们选择相似性得分在 6 分以上（含 6 分）的图片作为每类的有效图片，这些图片构成了我们的枪支数据库。

### 3.5 数据库的划分与相关信息

最终，我们得到的枪支数据库含有 14755 张图片，这些图片来自 167 类不同类型的枪支，我们将该数据库命名为 Firearm14k。该数据库有几个值得注意

的特点：

1. 该数据库具有高度不平衡 (imbalanced) 的特点 (图 3.6 展示了数据库中各类的图片数目, 按照图片数目从小到大排序), 有的类别的图像包含的图片数目超过 200, 还有一些类别图片数量大约只有 20 左右。
2. 由于该数据库包含很多长枪的图片, 因此数据库中很多图片的长宽比都很大, 在图 3.5 中, 我们分别画出了 Firearm14k, Paris6k, Oxford5k 以及 ImageNet 数据库中图片尺寸的比例关系, 对于每一个数据库, 我们分别画出了宽和长的比 ( $W/H$ ) 以及长和宽的比例 ( $H/W$ )。Firearm14k 数据库由于长枪居多, 所以宽长比和长宽比的分布呈现明显的两极化特点, 两个分布交叉很少 (见图 3.5, 左上), 另外三个数据库由于长宽比不那么极端, 两个分布呈现交错的状态。对于长宽比较极端的数据库, 如果在训练神经网络时把图片缩放为正方形, 将会更严重地破坏图像的内容, 保持图像长宽比显然是更好的选择。在第 5 章, 我们通过实验也发现, 保持图像的长宽比来训练网络, 比固定图像长宽的方式得到的检索结果至少要高 2%。
3. 该数据库中的图片是从真实世界中得到, 图片中的枪支在尺度, 拍照的视角, 姿态, 光照等因素有很大的变化, 背景也十分复杂, 因而识别的难度很大。

在图 3.7 中, 我们展示了一些随机选取自枪支数据库的示例图片, 从图中可以看出, 这些枪支图片中枪支的尺寸, 拍摄的视角等变化千差万别, 这给枪支图片的识别带来了很大的困难, 也是我们试图解决的问题。

我们把数据库划分为训练集, 验证集以及测试集, 每个集合分别占总的图片量的大约 65%, 10% 和 25%, 为了模拟真实的检索场景, 三个集合的图片类别互不交叉 (具体做法是把 167 类图片随机划分为 3 个互不相交的类的集合, 使得三个集合的图片数目占总图片量的比例接近预定的比例)。对于验证集和测试集, 我们从每一类图片中随机选择了两张作为查询图片<sup>6</sup>, 我们把剩余的所有图片作为分别作为验证集和测试集的图像库。在表 3.2 中, 我们列出了 Firearm14k 数据库的详细统计信息, 表 3.2 第 2 行与第 3 行列出了三个集合的图片类别数目, 图像数目。第 4 行和第 5 行列出了验证集以及测试集的查询图片数量以及对应的数据库图片数量。

---

<sup>6</sup>验证集的查询图像中, 有一张标注有错误, 我们从查询图片中去掉了该图片

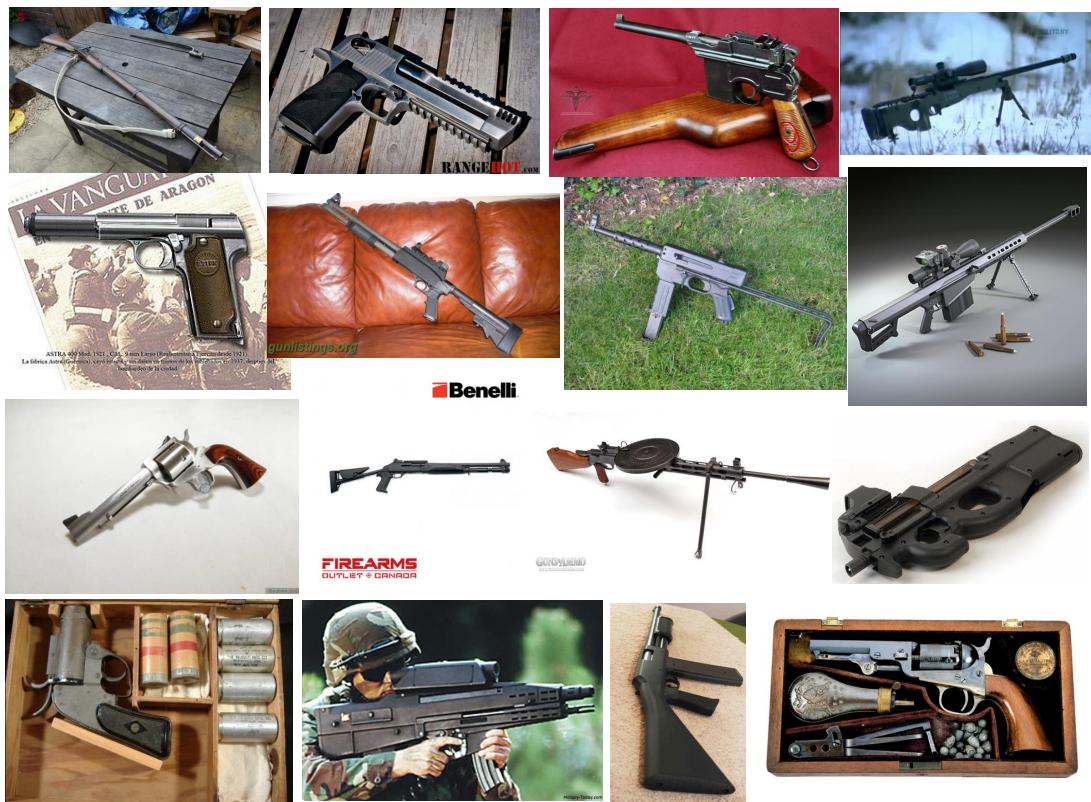


图 3.7 来自 Firearm14k 数据库的一些示例图片

表 3.2 Firearm14k 数据库详细统计信息

	训练集	验证集	测试集	总数
类别数目	107	20	40	167
图片数目	9628	1478	3649	14755
查询图片数目	-	39	80	-
图片库图片数目	-	1438	3569	-

### 3.6 本章小结

本章主要介绍了我们的枪支图片数据库的构建过程。首先，我们介绍了枪支图片的收集与清理，然后我们介绍了图像的标注过程，以及以相关性系数矩阵为基础，确定标注者的标注质量，进而得到图片的相似度分数，筛选出每一类真正有效的图片。最后，我们介绍了数据库的划分，列出了数据库的几个特点，并给出了数据库的详细统计信息。



## 第4章 基于多尺度全卷积网络的图像实例检索方法

### 4.1 引言

图像检索是学术界以及工业界都非常关注的一个问题，按照检索的精细程度来分类，我们可以把图像检索方法分为两大类：第一类是基于类别的检索，在这种设定下，只要返回的某个搜索结果与查询图片属于同一大类（如「车」，「猫」等），就认为该结果和查询图像是相关的；另外一类方法是精细检索，在本章中，我们研究的是图像实例级别的检索，在这种情况下，只有当返回的图片包含和查询图片相同的物体实例或者场景的情况下，我们认为该图片与查询图片相关。尽管图像的实例检索是一个经典问题 [6, 8, 18, 69, 72–74]，前人的研究已经提出了很多的方法，但是由于图像中物体尺寸，方向，位置以及光照等条件，在不同图像上变化很大，图像实例检索方法仍不能很好处理这些问题。

图像检索领域的经典方法最常使用的是基于 SIFT 局部特征描述子 [5] 的 bag of features (BOF) 方法。通常为了提高检索的效果，研究者经常使用一些后处理技术，例如查询扩展 [11] 以及空间验证 [8]。近些年来，由于卷积神经网络的流行，研究者也开始研究基于卷积神经网络的图像检索方法。他们的实验显示卷积神经网络在图像检索领域的有效性 [69, 72, 74]，他们发现，如果使用全局图像特征并且不使用后处理方法，基于卷积神经网络的方法和传统的 BOF 和局部聚合向量方法 (VLAD) [50] 性能相似甚至超过这些方法。尽管使用卷积神经网络来表达图像特征取得了这些进步，一些影响提取的图像特征有效性的潜在因素并未被详细研究，例如，图像缩放的策略，影响多尺度特征有效性的因素等。研究清楚这些问题将帮助我们构建更加鲁棒与准确的检索系统。

我们详细研究了影响图像特征有效性的一些因素，做出了一定的创新。不同于其他研究，我们使用了全卷积神经网络，研究了图像尺寸对检索结果的影响，然后，我们研究了提取图像多尺度特征的不同设置对检索结果的影响，最后我们实验了不同的学习 PCA 与白化矩阵的策略。结合我们的实验发现，我们提出了一种新的多尺度图像特征表达，该特征简洁而有效。我们在四个数据库上测试了提出的方法，大量的实验结果证明了我们提出的方法的有效性。

本章的结构组织如下：4.2 节介绍我们提出的多尺度全卷积方法的几个构成要素，包括图像尺寸缩放方法，多尺度特征表达以及 PCA 与白化矩阵的使用。

4.3 节给出各种因素对检索结果影响的实验结果以及我们的方法和其他同类方法的对比结果。4.4 节介绍与本章的工作直接相关的一些工作。最后在 4.5 节，我们对本章的内容进行总结。

## 4.2 基于多尺度全卷积网络的图像实例检索方法

### 4.2.1 背景知识

在本方法中，我们关注的是如何使用已有的神经网络模型，从中提出维度较低并且具有区分性的特征。对于图像  $I$ ，我们分别减去 RGB 三个通道的均值，然后将图像送入神经网络，图像在神经网络中经过一系列卷积，池化以及非线性操作。神经网络某一层输出的特征图可以认为是图像的原始特征，基于此，我们可以进一步构建更复杂的图像特征。特征图构成了大小为  $C \times H \times W$  大小的张量，其中  $C$  是特征通道的数目， $H$  与  $W$  是每个特征图的高和宽。我们用以下的公式代表特征图：

$$F = \{F_i\}, i = 1, 2, \dots, C, \quad (4.1)$$

上式中， $F_i$  代表第  $i$  个特征图，最简单的图像特征可以用下面的公式表示：

$$f = [f_1, f_2, \dots, f_i, \dots, f_C]^T, \quad (4.2)$$

其中， $f_i$  是通过对特征图  $F_i$  使用最大池化得到的。我们使用的特征图都经过 ReLU 操作，因此特征图每个元素都是非负的<sup>1</sup>。在得到图像特征以后，我们对特征进行 PCA 以及白化处理。

### 4.2.2 图像尺寸改变策略

计算机视觉各个领域的研究者为了自己的研究目的，通常使用一些在 ImageNet [14] 上训练的模型 [16, 21–23]，然后对这些模型进行一些改造以适应自己的研究。这些网络通常会对输入网络的图像尺寸有要求，为了满足网络对图像尺寸的要求，一些检索方面的工作 [70, 73] 通常对输入图像进行变化，使得图像的长宽变成固定值。图像的缩放操作可能会导致图像中物体信息的丢失或者扭曲，因此从网络中提取的图像特征的区分性不强，检索效果变差。对于图像检索任务，我们认为最好保持图像原有的大小，直接把图像输入神经网络。我们

<sup>1</sup>我们也实验了没有使用 ReLU 操作的特征图，发现效果并不好

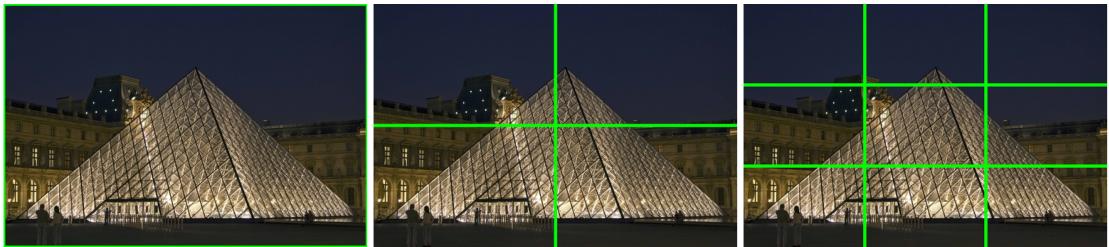


图 4.1 一张图像的 3 尺度特征表示

研究了三种不同的改变图像大小的策略：

- 图像的长和宽都被设定为固定值，这种策略用 *two-fixed* 表示。
- 图像的最小边为固定值，保持图像的长宽比不变，这种策略用 *one-fixed* 表示。
- 保持图像原尺寸，这种策略用 *free* 来表示。

#### 4.2.3 多尺度特征表达

不同于 SIFT [5] 等局部特征描述子，从神经网络提取的特征向量是全局描述子，编码的是图像的全局信息，因此得到的图像特征和图像的语义上的类别关系紧密但是缺乏一些局部以及细节信息，这些细节信息正是评价图像相似性所需要的。受一些多尺度图像特征表达工作 [89, 90] 的影响，我们试图把这种有效的特征表达方式与合适的图像尺寸改变策略结合在一起，得到具有区分性的特征。在我们的方法中，一张图像由  $L$  个不同尺度的图像特征金字塔表示，在每个尺度，图像被分为多个互相重合或者不重合的区域，图 4.1 展示了一张图像 3 尺度表示的示意图。然后我们计算每个小区域的特征表达，并把来自同一尺度的局部特征表达结合起来形成每个尺度的特征。最后我们把每个尺度的特征结合起来， $l_2$  归一化作为图像最后的特征。

我们试验了三种不同的用来结合局部特征以及不同尺度特征的方法。第一种方法和 He 等提出的空间金字塔 [90] 相似，来自不同尺度的所有区域特征被  $l_2$  归一化，然后被拼接成一个特征向量，因此该向量的维度很高，为  $N \times C$ ，其中  $N$  是所有来自不同尺度的小区域的个数， $C$  是特征通道的个数。第二种方法中，计算区域特征的和并归一化以后形成尺度级别的特征，然后拼接尺度级别的特征形成图像特征，因此这个特征的维度为  $L \times C$ ，其中  $L$  代表尺度的个数。在第三种方法中，区域级别的特征被加起来，然后归一化形成尺度特征，然后尺度特征随即被求和、归一化形成图像的特征表达，图像特征维度为  $C$ （特征通道的个数）。

如何有效计算图像各区域的特征也是需要考虑的问题，如果我们将每个图像小块分别送入神经网络来提取特征，计算一张图像特征需要消耗大量的时间，这对于图像检索应用是不现实的。受 ROIpooling [27] 以及 R-MAC (regional maximum pooling of activations) 方法 [74] 启发，如果我们假设图像区域与特征图对应区域是线性映射关系，那么我们就可以有效地计算图像区域特征，而不用把图像小块重新输入网络。在实验部分，我们实验了不同的提取多尺度特征的设置，报告不同设置得到的特征的检索性能并给出我们的分析。

#### 4.2.4 特征降维与白化

主成份分析法 (principal component analysis, PCA) 是一种常用的有效的特征降维方法，该方法可以使特征的各个分量之间相互独立，白化 (whitening) 则是使得每个维度特征均值为 0，方差为 1。之前的工作 [72] 已经表明经过 PCA 和白化以后，特征的检索效果将会提高。在本章中，我们研究了学习 PCA 与白化矩阵的不同方式对图像检索结果的影响，并给出一些发现。

### 4.3 实验

#### 4.3.1 实现细节

我们使用开源深度学习框架 Caffe [91] 来进行我们所有的实验，所使用的 GPU 为英伟达 Tesla K20m，显存容量为 4G，所使用的 CPU 为两个英特尔 Xeon 8 核心处理器。我们使用的模型是非常流行的 VGGNet 模型 [21]，该模型是在 ImageNet [14] 分类数据上训练得到的。在实验中，我们使用的特征图来自于网络的最后一层卷积层的输出。

由于原有的 VGG 模型只能接受固定大小的图像输入，这种方式对于需要提取图像特征的检索任务来说，并不是最好的方式。为了使网络能够处理各种大小与长宽比的图像并且测试我们提出的图像尺寸改变的策略，我们把网络变成了全卷积的网络，使得网络能够接受任何大小与长宽比的图像。

#### 4.3.2 评价指标与数据库

##### (1) 评价指标

图像检索领域两个最基本的指标是准确率 (precision) 与召回率 (recall)，准确率指的是返回的结果中，有效的样本占所有返回结果的比例，随着返回结果的增多，准确率通常会降低，常用的还有前  $K$  个返回结果的准确率，用  $\text{precision}@k$  表示；召回率则指的是返回的结果中，有效的样本占所有有效样本

的比例，随着返回结果的增多，召回率通常会上升，同样地，研究者也常使用前 K 个返回结果的召回率，用  $\text{recall}@k$  表示。准确率与召回率是考虑系统性能的两个重要指标，一般来说，准确率高的时候，召回率会比较低，而召回率高的时候，准确率通常又比较低（极端情况就是返回所有的样本，召回率 100%，但是准确率很低），因此一个好的检索方法必须同时在准确率与召回率上都取得较好的结果，在两个指标上取得平衡。

对于单个查询来说，为了综合反映检索算法的性能，通常并不单独使用准确率和召回率，二是使用综合性的指标，最常使用的一个指标是平均准确率（average precision，AP），该指标同时考虑准确率与召回率，实际上计算的是准确率-召回率曲线下面的面积，用公式表示该指标如下：

$$AP = \sum_{k=1}^n P(k) \Delta r(k) \quad (4.3)$$

公式 4.3 中， $P(k)$  代表  $\text{precision}@k$ ， $\Delta r(k)$  代表从  $k - 1$  到  $k$ ，召回率的变化。

评价检索算法在某个数据库上的性能，常用的评价指标是 mean average precision（简称 mAP），该指标指的是一个数据库上所有查询图像对应的平均准确率（AP）的均值，用公式表示如下：

$$\text{mAP} = \frac{\sum_{q=1}^Q AP(q)}{Q} \quad (4.4)$$

公式 4.4 中， $AP(q)$  代表某个查询  $q$  对应的平均准确率， $Q$  代表所有查询图像的数目。

## （2）数据库

在本章的实验中，我们总共使用了四个数据库，现对这四个数据库以及对应的评价指标做一简要介绍：

1. **Oxford5k** 数据库 [8] 包含了 5062 张从 Flickr<sup>2</sup> 上抓取的 Oxford 建筑图片。该数据库总共包含 55 个查询图像，每个查询图像包含有具体查询区域的坐标，在实验中，对于每一幅查询图像，我们使用了两种查询，分别是 *full-query*(使用整个查询图像) 和 *cropped-query*(使用坐标指定区域内的图像块)。在该数据库上使用的评价指标为 mAP。

<sup>2</sup><https://www.flickr.com/>

2. **Paris6k** 数据库 [7] 包含了 6412 张来自巴黎的建筑图片<sup>3</sup>。与 Oxford5k 数据库类似，该数据库也包含了 55 个查询图像，同时附带每个图像对应的感兴趣区域的坐标。在这个数据库上的使用的评价指标是 mAP。

3. **Oxford105k** 数据库包含了的是 Oxford5k 数据库的图片以及额外 100000 张来自 Flickr 的图片<sup>4</sup> [8]。这 100000 张图片与 Oxford5k 没有相同的图片，被用来作为干扰图片，主要是为了验证数据库规模扩大时的算法的性能。在这个数据库上使用的评价指标与 Oxford5k 完全一致。

4. **UKB** 数据库 [18] 总包含了 2550 个物体的 10200 张照片，每个物体 4 张照片。实验时，每张图片被用作查询图片去查询其他图片（包含该图片本身），在这个数据上的评价指标比较特殊，用返回的前四张图片与查询图片相似的平均数目来衡量，是区间  $[0, 4]$  上的一个数，也就是  $4 \times recall@4$ 。

### 4.3.3 实验结果与分析

在本部分，我们对不同因素对检索结果的影响进行了详细的实验，并给出相应的分析。

#### (1) 图像尺寸变化对检索的影响

我们对 4.2.2 节给出的三种图像尺寸变化的策略进行了实验，对于 *two-fixed* 和 *one-fixed* 策略，由于图像的尺寸可以选择多种，我们使用网格搜索来找到性能最好时对应的图像尺寸。我们发现，总体来说，增加图像的尺寸，对于 *full-query* 或者 *cropped-query* 这两种情况，都会提升检索的效果，试验结果如表 4.1 所示，表中括号后面的数字表示取得最好的 mAP 时使用的图像尺寸。从表上可以看出，对于 *cropped-query* 这种情况，*free* 策略能够有效地提升检索的性能，对于 *full-query*，*free* 与 *one-fixed* 策略则比较接近，但检索的效果都要好于 *two-fixed* 策略。我们可以看出 *one-fixed* 策略此时要稍高于 *free* 策略，但是在使用 *one-fixed* 策略的情况下，图像最小的边长度被固定在 800，因为数据库中大部分图像最小尺寸都要小于 800（在表 4.2 中，我们列出不同数据库中的图像最小边尺寸的统计信息），这种设置将大大增加提取图像特征所需要的计算时间。采用 *free* 方式，不仅减少了特征提取时间，并且能够取得很好的检索结果。

实验表明，改变图像的长宽比（对应 *two-fixed* 策略）对图像信息的干扰最大，因此检索的精度有明显的下降。*one-fixed* 策略虽然没有改变图像的长宽比，但

<sup>3</sup>按照惯例，该数据有 20 张被损坏的图片被去掉，因此实际图片数量是 6392。

<sup>4</sup>有一张损坏图片 `oxc1_100k\portrait\portrait_000801.jpg` 被移除

**表 4.1 不同图像尺寸变化策略之间的比较**

方法	full-query	cropped-query
<i>two-fixed</i>	55.5 (864)	38.7 (896)
<i>one-fixed</i>	59.0 (800)	39.3 (736)
<i>free</i>	58.0	52.6

**表 4.2 不同数据库中图像最小尺寸落在各个区间的比例 ( $s$  代表最小尺寸)**

Dataset	$s \leq 500$	$500 < s \leq 800$	$s > 800$
Oxford5k	0.87	96.86	2.27
Paris6k	1.30	95.59	3.11
Flickr100k	1.49	93.91	4.60
UKB	100	0	0

是由于图像尺寸的变化，仍然有一部分信息的丢失。*free* 方式能够编码图像未经扭曲与破坏的信息，因而得到的特征效果更好，能够取得更好的检索结果。

## (2) 图像的多尺度特征表达

首先我们通过实验比较了 4.2.3 节提出的区域特征以及尺度特征融合策略，实验结果表明前两种特征融合策略效果不如第三种策略。如果使用第一种策略，检索效果比第三种策略要低 41%。同时，使用前两种方法得到的图像特征维度（维度至少为 1500 维）也要高于第三种策略（特征维度为 512 维）。高维度特征也会使得检索时间变长，综合考虑这些，我们采用第三种方式来融合不同的特征。

我们进行了大量实验来确定多尺度方法最好的配置，实验结果列于表 4.3 中。在表中，「overlap」用来指示每个尺度的各个图像区域是否有重叠部分，「s2」和「s3」表示重叠发生在尺度 2 和尺度 3。「weighing」表示各个尺度特征融合的时候，是否采用了不同的权重或者相同的权重。「version」则表示每个尺度使用不同数目的图像块。

首先，我们研究了尺度的个数对检索精度的影响。当使用的尺度数目分别为 2 和 3 时，每个尺度的区域数目分别为  $\{1 \times 1, 2 \times 2\}$  和  $\{1 \times 1, 2 \times 2, 3 \times 3\}$ 。对于尺度为 4 的情况，我们使用三种不同版本的区域数目，用「v1」，「v2」，「v3」表示，所使用的图像区域数目分别为  $\{1 \times 1, 2 \times 2, 3 \times 3, 4 \times 4\}$ ， $\{1 \times 1, 2 \times 2, 3 \times 3, 5 \times 5\}$  和  $\{1 \times 1, 2 \times 2, 3 \times 3, 6 \times 6\}$ 。表 4.3 中 (a1)、(b1)、(c6) 三行列出了分别使用 2, 3, 4 个尺度表示数据图像特征得到的检索结果。很显然，尺度数目的增加也使得检索结果性能变好，特别是对于 *cropped-query*，检索的结果提高了 3.9%。

表 4.3 不同设置下多尺度特征检索结果对比

	scale	overlap	weighing	version	full	cropped
(a1)	2	×	×	-	63.5	59.0
(a2)	2	×	✓	-	63.9	61.0
(b1)	3	×	×	-	64.2	60.9
(b2)	3	×	✓	-	62.6	61.0
(b3)	3	s2	×	-	64.8	60.8
(c1)	4	s3	×	v1	65.1	61.4
(c2)	4	s3	✓	v1	64.8	60.7
(c3)	4	s2,s3	×	v1	65.5	60.8
(c4)	4	s2,s3	×	v2	65.9	61.5
(c5)	4	s2,s3	✓	v2	65.4	61.2
(c6)	4	×	×	v3	64.5	61.3
(c7)	4	s3	×	v3	65.8	62.2
(c8)	4	s2,s3	×	v3	<b>66.3</b>	<b>62.6</b>

接着，我们研究对不同的尺度的特征采用加权融合的方式是否会提升检索的效果。我们使用的不同的尺度的特征加权融合方式类似空间金字塔匹配方法 [89]，也就是说，给来自粗糙尺度上特征更少的权重，给来自精细尺度的特征更大的权重。假设当前总共有  $L$  个尺度，对应的特征为  $f^1, f^2, \dots, f^L$ ，则图像特征  $f$  计算方式为：

$$f = \frac{1}{2^{L-1}} f^1 + \sum_{i=2}^L \frac{1}{2^{L-i+1}} f^i. \quad (4.5)$$

关于空间金字塔匹配方法更多的细节，可以参见 Labebnik 等的论文 [89]。比较 (a1) 和 (a2) 行的结果，似乎加权的方式可以提升检索的性能。我们做了更多的实验，发现随着尺度数目的提高，利用加权方式来求图像特征并不能提高检索的性能，反而得到了更差的结果，例如，比较 (b1)(b2) 两行的结果，或者比较 (c1)(c2) 两行。这里的结果表明，深度学习提取的特征和传统的特征描述子不同，因此当我们使用对传统的 SIFT 特征有效的方式时，应该有所警惕 [73]。基于这里的实验结果，我们在计算图像特征时，不使用权重不同的策略。

最后，我们研究了同一尺度下不同图像区域重叠（overlapping）是否会提高检索的结果。对于不同尺度数目，我们在一个或者两个尺度下使用重叠的策略。对于行 (b1)(b3) 和 (c1)(c3)，我们观察到使用重叠的方式提高了 *full-query* 情况下的结果，但是却降低了 *cropped-query* 情况下的检索结果。但是对于尺度数目为 4，版本是「v3」的情况（比较 (c7)(c8)），我们可以观察到对于 *full-query* 和

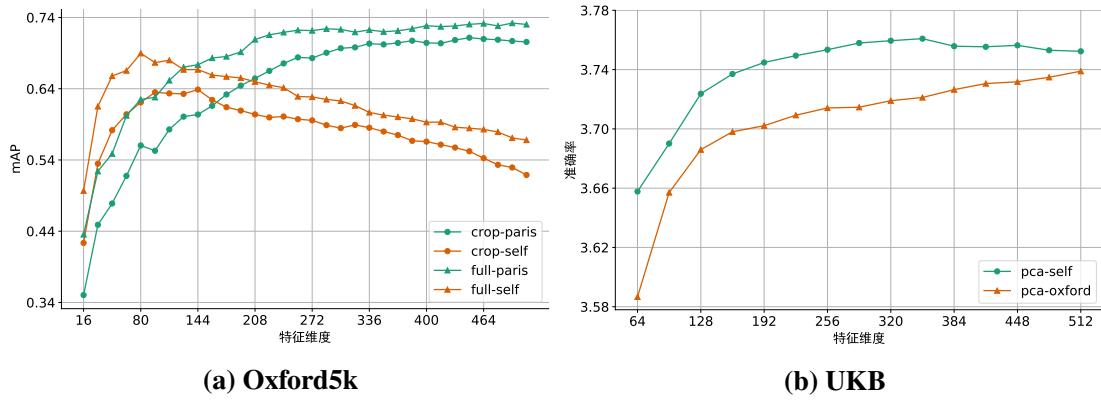


图 4.2 学习 PCA 与白化矩阵的方式对检索结果的影响

*cropped-query*, 检索结果都有所提高。因此计算最终的图像特征时, 我们在尺度 2 和尺度 3 上使用了重叠的策略。

### (3) 主成份分析与白化方法

我们在 Oxford5k 以及 UKB 数据库上进行了 PCA 与白化对检索影响的实验。对于 Oxford5k 数据库, 我们分别从 Oxford5k 数据库本身 (分为「full-self」以及「crop-self」两种情况) 或者从 Paris6k 数据库 (分为「full-paris」以及「crop-paris」两种情况) 学习 PCA 与白化矩阵; 对于 UKB 数据集, PCA 与白化矩阵则从 UKB 本身 (称为「pca-self」) 或者从 Oxford5k 上 (称为「pca-oxford」) 学习得到。我们发现, 对于 Oxford5k 数据库, 如果从 Paris6k 上学习 PCA 与白化矩阵, 无论是对于 *full-query* 还是 *cropped-query* 这种情况, 检索结果都会提高。但是 UKB 数据库, 由于它和 Oxford5k 数据库和 Paris6k 数据库风格不同, 从 Oxford5k 数据库学习到的 PCA 与白化矩阵反而对检索结果有害。图 4.2 清楚地展示了这种不同。同时, 对于 Paris6k 数据库, 我们也有同样的结果, 当使用从 Oxford5k 数据库学习到的 PCA 与白化矩阵时, 检索的结果也会有所提升。

#### 4.3.4 与其他方法的对比

基于前面的实验结果以及我们的对不同影响因素的分析, 我们提出了 **MFC**: 一种多尺度全卷积的图像特征表达方法。对于一张图像, 我们不对它进行缩放 (*free* 的方式), 直接输入到卷积神经网络, 然后我们从卷积神经网络的最后一层卷积层的输出的特征图的基础上, 提取 4 尺度图像特征表达。在多尺度特征表达中, 我们使用最大池化处理特征图, 并且使用了区域重叠的策略, 我们把区域级别的特征加起来形成尺度特征并归一化, 然后把各个尺度的特征求和得到图像特征, 再一次归一化。之后, 我们学习 PCA 与白化矩阵, 对图像特征进行降维。对于不同待测数据库, 我们用来学习 PCA 和白化矩阵的数据库也不相

表 4.4 与其他方法的对比

方法	D	Oxford5k		Paris6k		Oxford105k		UKB
		full	cropped	full	cropped	full	cropped	
Razavian et al. [69]	256	58.9	-	57.8	-	-	-	3.65
SPoC [73]	256	58.9	53.1	-	-	57.8	50.1	3.65
Neural codes[72]	512	55.7	-	-	-	52.2	-	3.56
R-MAC [74]	512	-	66.8	-	83.0	-	61.6	-
Ours	256	72.2	68.4	82.5	83.4	68.0	62.9	3.75
Ours	512	73.0	70.6	82.0	83.3	68.9	65.3	3.75

同：对于 Oxford5k 及 Oxford105k，我们使用的是 Paris6k；对于 Paris6k 和 UKB，我们使用数据库分别是 Oxford5k 以及 UKB。经过 PCA与白化处理的特征，被再一次归一化，得到最终的特征，这个特征用来报告各种方法的性能。

我们把提出的方法与一些主流的方法进行了比较，这些方法都使用了较低维度的特征，并且没有使用复杂的后处理方法，如几何重排序 [8] 以及查询扩展 [11]。我们使用了不同的特征维度，实验结果见表 4.4，「D」代表特征的维度，「full」与「cropped」分别代表 *full-query* 以及 *cropped-query*。和其他方法相比，在不同的特征维度下，我们的方法都取得了不错的结果。

#### 4.4 相关工作

Labebnik 等人 [89] 提出使用空间金字塔匹配（spatial pyramid matching, SPM）来编码图像的空间信息，在该方法中，可以用多个尺度的金字塔来表示一张图像。来自不同的尺度的特征被结合在一起形成图像的特征，在特征融合的时候，来自精细尺度的特征得到了更多的权重，来自粗糙尺度的特征得到了更少的权重。He 等人 [90] 提出空间金字塔池化方法（spatial pyramid pooling, SPP），该方法把神经网络最后一层卷积层输出的特征图划分为不同尺度的金字塔，每个尺度下的区域特征，以及各个尺度的特征被拼接在一起，形成了图像的特征。在我们的工作中，我们也探索了这两种方式是否可行，我们发现使用更简单的特征聚合方式反而取得了更好的结果。

本章采用的是直接提取图像特征不经过网络微调的方式来得到图像特征，在这方面有一些相关的工作。Tolias 等人 [74] 以及 Babenko 等人 [73] 提出使用神经网络的最后一层卷积层的输出作为图像特征，他们分别采用了不同的特征池化方式。Tolias 等人还提出使用一种多尺度的特征表达方式，被称为 R-MAC (regional maximum activation of convolutions)，特征图被分为 3 个不同的尺度，

每个尺度有一些相互重叠的区域构成，每个尺度的区域个数由原图的长宽比决定。在我们的方法里，我们使用了更简单的策略，每个尺度下区域的个数是固定的，区域的大小根据图像长宽比不同而变化。另外这些工作也没有讨论图像尺寸对检索结果的影响以及该如何学习 PCA 和白化矩阵，才能提升检索的性能。

#### 4.5 本章小结

在本章中，我们介绍了所提出多尺度全卷积的图像特征提取方法。前人工作没有详细地探讨各种因素对检索结果的影响，在本工作中，我们对图像尺寸缩放的策略，多尺度提取特征的方式，以及 PCA 和白化矩阵学习这三个重要的因素进行了研究，并进行了大量的实验，分析三个因素对检索的影响，在此基础上我们提出了多尺度全卷积的图像特征提取方法。我们在四个数据库上进行了实验，实验结果也表明我们提出的方法有着良好的检索效果。



## 第5章 基于双阈值对比损失函数的枪支图像检索方法

### 5.1 引言

在社交媒体上，经常会有用户把枪支等敏感图片上传到自己的主页，这些图片可能助长非法的枪支交易 [12, 13]，也可能引发其他用户的不适，因此某些类型的枪支图片（如机关枪，自动步枪等）需要适当的监管与控制 [92]，如果普通用户在社交媒体上展示不能随意购买的枪支，这可能预示着非法的交易。另外，在取证科学上，取证人员有时也需要根据枪支的图片确定枪支的具体品牌与型号，由于枪支类型的繁杂，除非取证人员经过专门的训练，否则很难从庞大的枪支数据库准确找到具体的枪支类型。基于图像检索的任务能够有效地帮助解决以上的需求，我们只需要使用已经训练好的模型提取图像的特征，然后计算图像特征之间的距离，采用最近邻方法就可以确定具体的枪支类型。

在本章中，我们研究枪支图像的精细（fine-grained）检索问题，也就是说，给定一个检索的枪支图像，我们的方法需要从数据库找到和给定图像属于同样精细类别的图片 [93, 94]。图像精细检索任务相当困难，因为不同于行人再识别（ReID） [95, 96]与人脸识别 [97, 98]等任务，图像精细检索任务中使用的图片都是非对齐的（un-aligned）图片，也就是说，在训练和测试过程中使用的图片，其中的物体可能存在不同的大小，可能位于图像不同的位置，并不像人脸或者行人图片一样已经经过检测算法的检测和裁剪，这些都给图像检索任务带来了相当大的挑战。

随着基于卷积神经网络的方法于 2012 年在 ImageNet [14] 图像分类竞赛中取得压倒性的成功 [16]，研究者们发现对已有的网络模型进行端到端的微调能够大幅度刷新卷积神经网络在不同任务上的性能，他们将基于卷积神经网络的方法应用到了物体检测 [24–27]，图像语义分割 [28–30] 等领域。与此同时，图像检索领域的研究者也开始使用基于卷积神经网络的方法，希望通过网络的微调提高模型在检索任务上的准确率。其中，Radenovic 等人 [82] 使用 Siamese 架构的网络结合对比损失（contrastive loss）函数 [81, 99, 100] 来训练整个模型的结构，Gordo 等人 [17] 则使用了三通道的网络结构结合常用的三元组损失（triplet loss）函数 [77, 78, 101, 102] 来学习图像的特征。无论是使用两个分支的网络的结构还是使用具有三个分支的网络结构，他们的目的都是希望通过端到端的学习，得到好的网络参数，使得相似图像在特征空间的距离更近，不相似的图

片在特征空间距离更远。当然他们得到的实验结果也表明，端到端的训练能够显著提升模型在常用数据库上的检索效果。

基于 Siamese 架构的方法 [80, 82, 99] 是一种常用的学习特征距离度量的方法，该方法通过输入图像对 (image pair) 来进行训练，目的在于通过训练，学习到一个好的图像特征嵌入空间 (embedding space)，在这个空间中，相似图像之间的特征距离较小，不相似图像之间的特征距离较大。然而，该方法有一个比较明显的缺陷，其使用的对比损失函数是非平衡的：在网络的训练过程中，即使相似图像之间的距离已经非常小，使用该损失函数仍然会产生损失。这个损失函数就像加在相似与不相似图像对上的无形的作用力，对于相似图像对，通过训练，把他们在特征空间的距离拉近，对于不相似图像对，通过训练，把他们的特征距离在特征空间推远。但是传统的对比损失函数对相似图像对的「牵引力」太强，导致网络在训练的时候过分注重相似图像对，因此网络的检索性能并不是很高。另外已有的方法的另外一个问题是，他们使用的检索数据库 [7, 8] 和 ImageNet 数据库在风格上相似，因而在 ImageNet 数据库训练的分类模型能够容易地适应新的数据库。事实上，这些在 ImageNet 上训练得到的模型，如有名的 AlexNet [16], VGGNet [21], GoogleNet [22], ResNet [23] 等，即使不经过微调，直接使用从这些模型的全连接层或者卷积层提取的特征，也能够在上述的数据库上取得非常不错的效果 [72–74]，我们在第 4 章的实验也说明了这一点。但是对于枪支数据库，由于该数据库上的图片与 ImageNet 图片风格差异巨大，直接使用检索任务训练这些在 ImageNet 上训练得到的模型，效果并不理想，在 5.3 节实验部分，我们将展示一些实验结果：即使这些分类模型经过检索任务的微调，在枪支数据库上的效果仍然不佳。

为了解决网络训练过程中相似图像对 (similar image pair) 与不相似图像对 (dis-similar image pair) 贡献的损失不平衡的问题，我们提出了基于双阈值对比损失函数的方法。在该方法中，我们分别给相似与不相似图像对设定阈值，因此在模型训练过程中，正负图像对贡献的损失能够更加平衡。在实验过程中，我们通过采样相似与不相似图像对图像特征距离，得到距离的概率分布，然后在此基础上通过实验选出相似与不相似图像对对应的最佳阈值。实验表明，我们的提出的双阈值方法在相同的情况下检索性能比传统的单阈值方法高出 34.2%（使用的基础模型为 VGGNet [21]）。为了解决原始的 ImageNet 数据集与 Firearm14k 数据库之间差异过大的问题，我们提出了两步训练的策略。具体来说，我们首先在 Firearm14k 上用分类任务来微调原始的模型，然后在此基础上，

使用双阈值误差来微调得到的模型。这种策略非常有效，进一步提升了模型的检索性能。

本章的结构组织如下：5.2 节简要回顾单阈值对比损失，然后介绍我们提出的基于双阈值对比损失的方法，5.3 节介绍具体的实验结果，包括双阈值与单阈值对比损失方法的结果比较，两步训练的效果，阈值的选择对实验结果的影响，最后比较了我们的方法和其他的方法在不同维度下的检索效果。5.4 节介绍了一些与本章的方法直接相关的工作，5.5 节总结本章的内容。

## 5.2 基于双阈值对比损失函数的枪支图像检索方法

我们的方法使用了卷积神经网络来得到图像在欧式空间的特征，我们使用了 Siamese 网络架构，结合双阈值对比损失函数来学习有区分性的图像特征表达。在训练的时候，我们采用如图 5.1 上部分所示的结构，两个网络的参数是共享的，其中 MAC (maximum activation of convolutions) 是由 Tolias 等人 [74] 提出的图像特征表达方法， $l_2$  代表特征的  $l_2$  归一化。在测试的时候，图片通过测试网络，经过  $l_2$  归一化，得到最终的特征，要计算两个图像之间的相似度，可以直接使用点积得到。

### 5.2.1 图像特征表达

我们使用 MAC [74] 方法来得到图像的特征表达。对于一个图像  $I$ ，我们使用神经网络的最后一层卷积层输出的特征图，该特征图需经过 ReLU 操作，然后对该特征图使用全局最大池化 (global max pooling) 操作，得到图像的特征。然后，我们对特征进行  $l_2$  归一化，得到最终的图像特征，该特征维度为  $C$ ， $C$  是网络最后一层卷积层输出的通道个数。

这种特征表示方式的优点在于该网络是全卷积形式的，因此可以接受不同大小的图像作为网络的输入，能够保持图像的长宽比。He 等人 [90] 以及 Hao 等人 [68] 的文章都指出，保持图像长宽比，能够取得更好的结果，我们在实验中也发现，在训练时保持图片长宽比，测试集上得到的结果比训练时使用固定长宽的方式要高 2%。

### 5.2.2 双阈值对比损失函数

首先我们简要回顾传统的单阈值对比损失函数。给定图像对  $(I_p, I_q)$ ，以及对应的相似性标签  $y$ （如果两张图片是相似的， $y = 1$ ，否则  $y = 0$ ）。假如我们用  $f(I)$  表示图片  $I$  经过神经网络得到的特征，那么单阈值对比损失可以用如下

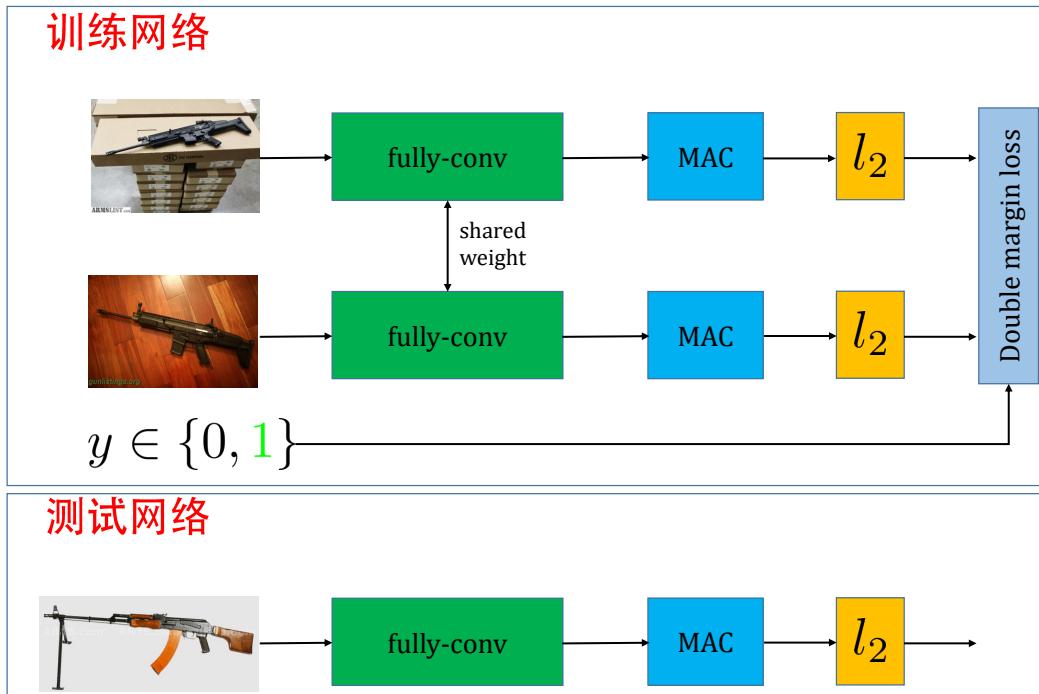


图 5.1 基于双阈值对比损失方法的网络结构，上面为训练网络，下面为测试网络

的公式来表示

$$L(I_p, I_q) = \frac{1}{2} [y d^2 + (1 - y) \max(\alpha - d, 0)^2], \quad (5.1)$$

上式中， $d$  是两个图像  $I_2$  归一化特征之间的欧式距离 ( $d = \|f(I_p) - f(I_q)\|_2$ )， $\alpha$  是对不相似样本设定的一个阈值。单阈值对比损失函数试图把相似图像尽可能拉近，同时把不相似图像之间的距离推的超过设定的阈值  $\alpha$ 。

单阈值方法的一个明显缺点是网络在训练过程中会偏向相似图片对，因为相似图片对总是会贡献损失，除非相似图片之间的距离为 0。另外一个问题是在实际训练过程中，很难为不相似图像对设定一个合适的阈值  $\alpha$ ，使得正负图片对贡献的损失是平衡的。总的来说，这两个问题都导致在实际中很难平衡正负样本贡献的损失，因而不能学习到很好的特征距离度量，导致学习到的模型在 Firearm14k 数据库上效果不佳。

为了缓解不平衡的损失问题，我们提出使用双阈值对比损失函数来优化网络模型。该损失函数可以用如下的公式表示

$$L(I_p, I_q) = \frac{1}{2} [y \max(d - \alpha_1, 0)^2 + (1 - y) \max(\alpha_2 - d, 0)^2], \quad (5.2)$$

公式 5.2 中,  $\alpha_1$  和  $\alpha_2$  是对应于相似与不相似图像对的阈值。很显然, 为了使得损失函数有效且有意义, 阈值不应该超过两个归一化的特征之间的距离的最大值 (为  $\sqrt{2}$ ), 因此下面的不等式成立

$$0 \leq \alpha_1 \leq \alpha_2 \leq \sqrt{2}. \quad (5.3)$$

比较公式 5.1 与 5.2, 我们可以看出, 单阈值与双阈值方法之间的差异主要在于相似图像对的损失如何计算:

- 单阈值方法: 相似图像对总是会贡献损失
- 双阈值方法: 只有两张相似图像的特征距离大于阈值  $\alpha_1$  的图像对才会贡献损失

当我们需要对模型的参数进行小幅度更新以取得更好的结果时, 基于双阈值损失函数的方法尤其有效, 我们将在实验部分对这种情况下的结果进行说明。我们在实验中发现, 由于这种对相似样本更加「缓和」的损失, 模型才能在分类任务的基础上继续提高检索的效果, 具体实验结果见 5.3.3 节。

双阈值对比损失函数 (公式 5.2) 中的两个阈值是该损失函数的核心, 我们根据相似与不相似图像对的特征距离的分布来选择这两个阈值。由于数据库中图片数量庞大, 可以组成的相似与不相似图像对的数目也十分庞大, 因此不可能穷举所有组合, 实际上, 我们的做法是随机从训练集中采样大约 400000 对图像 (相似图像对与不相似图像对的数目相等), 并计算图像对特征距离, 分别得到相似与不相似图像对特征距离的分布。然后我们把两个分布的均值作为选取两个阈值  $\alpha_1$  和  $\alpha_2$  的起点。

### 5.2.3 两步训练策略

我们在训练中使用的基础模型是从 ImageNet 数据集训练得到的分类模型, 由于 ImageNet 数据库和我们的 Firearm14k 数据库差异非常大, Firearm14k 图像的特征不能很好地被基础模型所表达。为了解决这个问题, 我们首先使用分类任务微调基础模型, 经过分类任务微调以后, 模型已经能够比较好地表达 Firearm14k 数据库中的图像特征, 此时的检索性能已经较好。然后, 我们使用分类微调后的模型, 进一步使用我们提出的双阈值损失函数进行微调, 实验表明, 检索的结果会得到进一步的提升。

### 5.2.4 特征降维

当训练过程完成以后，我们进一步使用主成份分析法（PCA）对特征进行降维，因为 PCA 能够进一步提升特征的效果 [72–74, 82]。在我们的实验中，我们使用了 PCA 方法，但是并未使用白化 (whitening)，因为我们发现白化特征会引起检索效果的大幅度下降。

## 5.3 实验

### 5.3.1 评价指标

在实验中，我们在训练集上训练模型，在验证集上采用 mAP 来选择合适的模型。我们使用两种评价指标来报告模型在测试集上的结果。第一种评价指标是 mean average precision (mAP)，mAP 常用来衡量检索系统的总体性能。第二个评价指标是 Rank-k 准确率 [93, 103]，Rank-k 准确率计算的是在所有的查询图片中，其  $K$ -邻近图片中存在至少一张同类图片的查询图片的比例。Rank-k 准确率注重于检索方法返回的前面的结果，在实际场景下，如图像搜索引擎及电商网站在线购物，用户更关心前面返回  $K$  个结果，因此 Rank-k 准确率这个指标更加能够反映实际需求。

### 5.3.2 实验细节

我们使用了开源框架 PyTorch<sup>1</sup> 来实现我们的整个方法，我们使用的服务器搭载了英伟达 TITAN X (Pascal) GPU，显存容量为 12G，同时搭载两颗英特尔 Xeon 12 核 CPU。在实验中，我们所用的基础模型为 VGG16 模型 [21]，该模型是在 ImageNet 分类数据库上训练得到的。针对我们的任务，我们对模型进行了一些修改，我们把模型的全连接部分去掉，只使用了该模型的全卷积部分。

我们使用了两步训练的办法对模型进行了训练，下面分别介绍这两步训练的过程。

#### 1. 使用分类损失微调网络

在这个步骤中，我们使用了 Firearm14k 数据库中的训练集和验证集的所有 127 类图片，我们把这些图片分为训练和验证数据，比例为 70% 和 30%。我们在第 3 章 3.5 节已经提到过，该数据库类别高度不平衡，因此我们使用了加权的交叉熵损失函数来训练网络，对于某个样本来说，该损失函数形式如下：

<sup>1</sup><http://pytorch.org/>

$$l = -w_k \log \frac{\exp(x_k)}{\sum_{i=0}^{126} \exp(x_i)} \quad (5.4)$$

上式中， $k$  是该样本的真实类别，神经网络的输出为 127 个节点， $x_i$  是神经网络在第  $i$  个节点的输出值， $w_k$  是第  $k$  类对应的权重。我们使用的批 (batch) 大小为 128，训练过程 (epoch) 数量为 50，初始学习率 (learning rate) 为 0.001，冲量 (momentum) 为 0.9，权值衰减 (weight decay) 系数为 0.0005，学习率每过 30 个训练过程，减为之前的十分之一。

为了防止模型过拟合，我们也使用了数据增广 [16] (data augmentation)，由于我们使用的网络可以接受不同大小的图像，因此我们在实验中保持图像长宽比不变，把图像的最长边缩放到 [256, 384] 这个范围 (步长为 8)，同时我们也使用了随机的水平翻转，随机的旋转以及随机的颜色变换 (亮度，对比度等变化)。基础网络的分类微调过程，大约耗时 2 个小时。最后，我们选择在验证数据上分类准确率最高的模型，然后在这个模型基础上再进行第二步的训练。

## 2. 使用双阈值对比损失函数微调网络

正如 5.2 节介绍的，使用 Siamese 结构的网络训练图像时，需要输入一对图像以及相应的标签，因此我们需要生成训练数据。对于 Firearm14k 训练集中的每一类图像，我们随机生成 180 对相似图像对以及 180 对不相似图像对。我们总共生成了 38520 对图像来训练第一步得到的网络模型。当然，对于神经网络来说，这样的训练数据量并不够，为了增加图像的多样性，增强模型的泛化能力，每 5 个训练过程，我们会重新生成一次训练图像对。在训练过程中，我们同样使用了数据增广方法，和分类训练数据增广方法一致，这里不再赘述。

由于输入网络的图像大小以及长宽比不同，我们无法使用传统的批处理 (batch processing) 的方法来输入图像，在训练过程中，我们每次输入一个图像对，然后反传误差，等输入网络的图像对达到 64 对时<sup>2</sup>，我们对网络的参数进行一次更新。在验证集上，我们使用 mAP 衡量网络的性能，从而选择最佳的网络模型。我们使用随机梯度下降方法优化网络的参数，实验中使用的参数设置为：初始学习率为 0.001 (每过 10 个训练过程变为原来的十分之一)，冲量为 0.9，权值衰减为 0.0005。网络的最大训练过程为 30，整个训练大约需要 14 个小时的时间。

---

<sup>2</sup>所以可以认为我们使用的批大小是 64

**表 5.1 单阈值方法与双阈值方法的对比**

方法	mAP(%)	Rank-k 准确率(%)			
		k=1	k=2	k=4	k=8
(a) VGG	31.1	83.75	91.25	95.0	96.25
(b1) retr-s	35.1	80.0	86.25	92.5	96.25
(b2) retr-d	47.1	87.5	88.75	95.0	97.5

### 5.3.3 实验结果与分析

本节介绍具体的实验，主要有单阈值与双阈值对比损失函数的比较与结果分析，两步训练的结果以及分析，以及双阈值函数中的两个阈值选取对检索性能的影响分析。

#### (1) 单阈值与双阈值对比损失函数的比较

在本部分，我们评估了所提出的双阈值对比损失函数的有效性，和单阈值对比损失函数进行了比较。在我们的比较中，双阈值与单阈值方法使用了相同的基础模型（VGG16 模型）并且使用了相同的设置。对于两种方法，我们都进行了多轮实验，然后我们从多轮实验中选择效果最好的模型来比较这两个方法。我们在表 5.1 列出了两种方法在测试集上的实验结果，对于 Rank-k 准确率，我们都使用了  $k = [1, 2, 4, 8]$ 。在表 5.1 中，第一行「VGG」是使用原始 VGG 网络得到的结果，第二行「retr-s」代表使用单阈值对比损失函数得到的检索结果，第三行「retr-d」代表使用双阈值损失函数得到的结果。从这些结果，我们可以看出：

- 单阈值与双阈值方法都可以提高基础的 VGG 模型的 mAP。
- 原始的 VGG 的模型已经在 Rank-k 准确率这个指标上取得了不错的效果，因而 Rank-k 准确率这个指标更加具有挑战性。基于双阈值损失函数的方法在这个指标上仍然有所提高，但是基于单阈值损失函数的方法在这个指标上反而比原始的 VGG 模型还要差。

总的来说，基于双阈值对比损失函数的方法在两种评价指标下的效果都优于传统的单阈值损失函数，双阈值损失函数在 mAP 指标上取得了 47.1% 的结果，而单阈值方法只取得了 35.1% 的结果，双阈值方法相对于单阈值方法整整提高了 34.2% 的 mAP，而且没有额外的时间和内存等开销，这说明了双阈值损失函数的有效性。

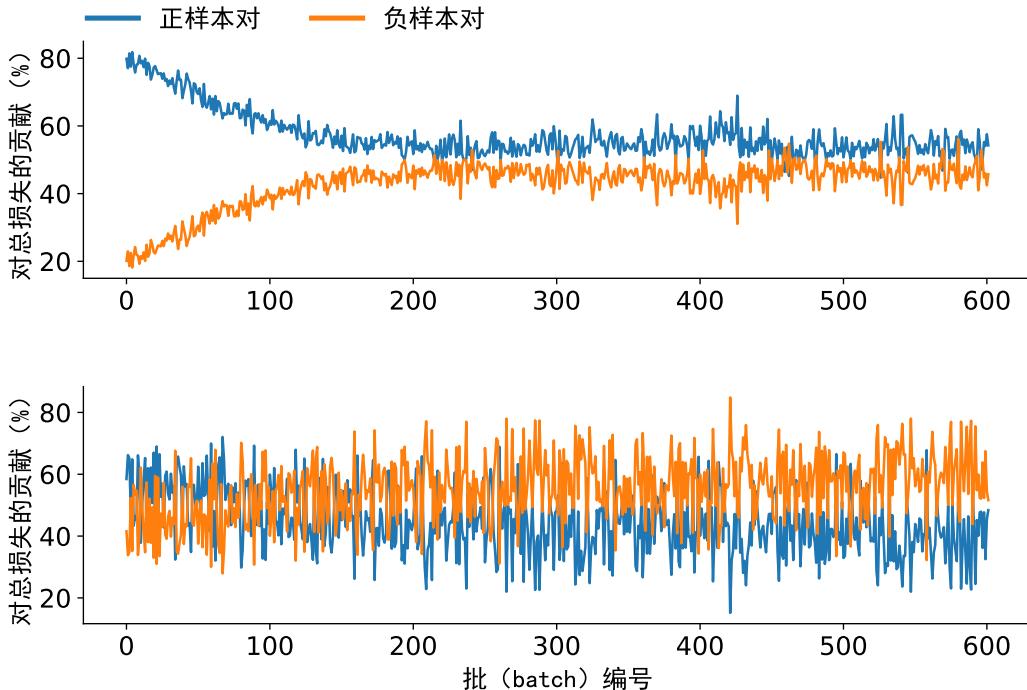


图 5.2 使用不同损失函数（上图：单阈值对比损失函数，下图：双阈值对比损失函数），一个训练过程中不同批次数据，相似样本对与不相似样本对对总体损失的贡献率

表 5.2 两步训练策略的结果

方法	mAP(%)	Rank-k 准确率(%)			
		k=1	k=2	k=4	k=8
(a) VGG	31.1	83.75	91.25	95.0	96.25
(b1) Cls	65.4	92.5	96.25	97.5	98.75
(b2) Cls + retr-s	-	-	-	-	-
(b3) Cls + retr-d	68.4	95.0	98.75	98.75	100.0

我们也从直观上研究了为什么基于双阈值损失函数的方法更加有效。我们分别画出了基于双阈值函数的方法以及基于单阈值函数的方法，在一个训练过程中不同批数据中，相似样本对以及不相似样本对对整个损失的贡献率，所得得到的结果如图 5.2 所示。图 5.2 上部分展示的是使用单阈值对比损失函数的结果，下部分展示的使用双阈值对比损失函数的结果。可以看出，使用单阈值损失函数时，刚开始训练时的一些批次数据，正样本对总的损失的贡献明显大于负样本对，这样的话，网络的训练可能会被引入错误的方向，导致后续模型优化出现方向错误，因而检索效果很差；使用双阈值损失函数，从训练一开始，正负样本贡献的损失与负样本贡献的损失相对平衡，因而网络优化方向更好，检索的效果也更好。

## (2) 两步训练策略的效果

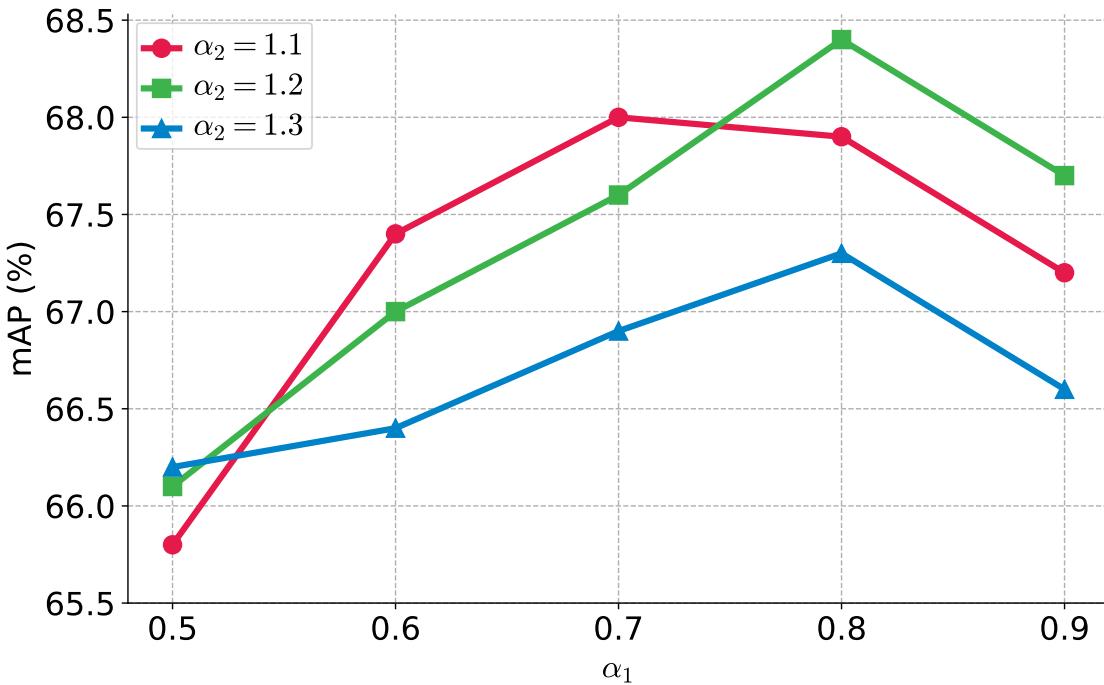


图 5.3 双阈值对比损失函数中两个阈值对模型检索性能的影响

我们进一步研究了两步训练策略的有效性。首先我们使用分类任务（采用加权的交叉熵损失）来微调网络的参数，然后我们在分类微调得到的模型基础上，再使用双阈值对比损失函数进行优化。我们将实验的结果列在表 5.2 中，「Cls」表示使用分类任务训练得到的结果，「Cls + retr-s」代表先使用分类任务后使用单阈值损失函数得到的结果，「Cls + retr-d」代表先使用分类任务后使用双阈值损失函数得到的结果。

从实验结果，我们可以看出，仅仅通过在分类任务上训练模型，在 mAP 这个指标上，我们得到的模型相对于基础的 VGG 模型已经取得了 110.2% 的效果增强。我们发现，在分类微调的模型基础上，在使用单阈值对比损失函数无法进一步提高模型的检索效果，但是使用双阈值对比损失函数微调模型，无论在 mAP 还是 Rank-k 准确率的指标上都可以进一步提高网络的检索效果。实际上，我们发现，如果使用单阈值方法，无论我们设定阈值为多少，随着训练过程的进行，模型的性能会持续下降（因此在表 5.2 中，我们没有列出「Cls + retr-s」的结果）。这种情况是可以预见的：因为模型经过分类任务的微调，当我们再使用检索任务微调模型的参数时，必须要小心，使用单阈值的方法将会使得模型优化向错误的方向前进（因为单阈值方法中，相似图像对的损失占上风）。正是由于使用双阈值损失函数，正负样本贡献的损失更加平衡，所以模型的检索效果得到进一步提升。

### (3) 阈值选取及对检索结果的影响

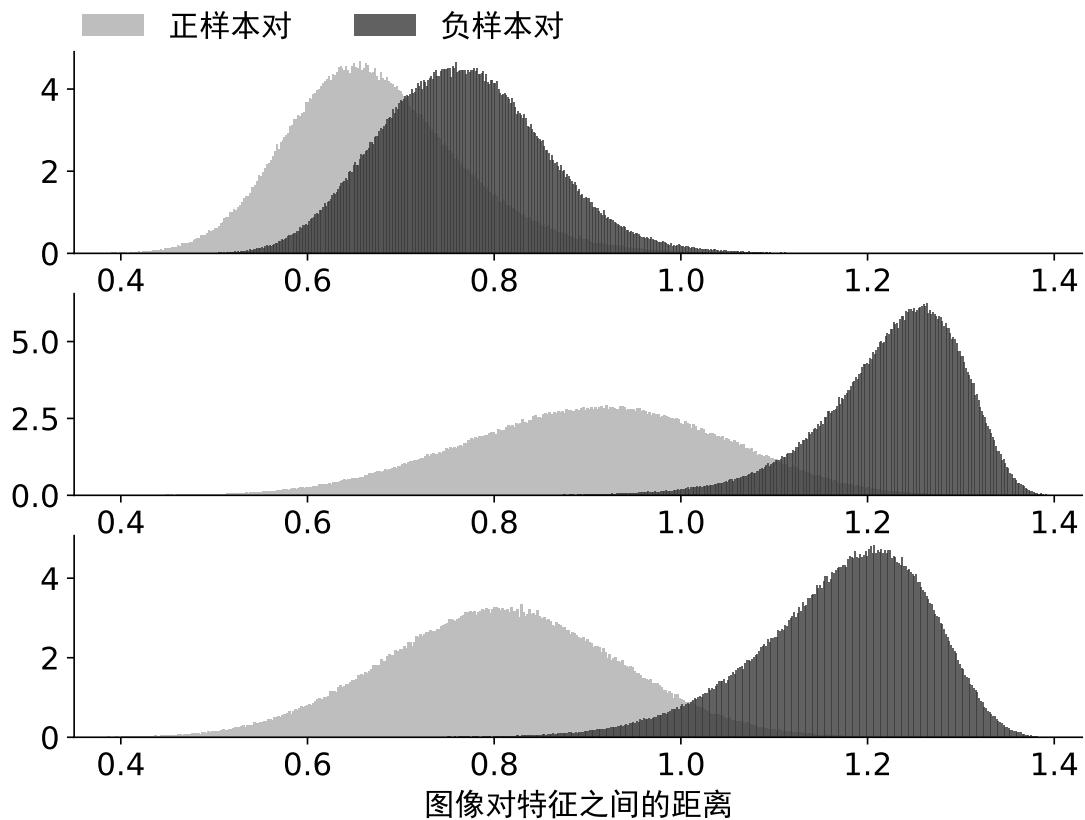


图 5.4 不同模型下相似与不相似图像对特征距离的概率分布

在本部分，我们将讨论双阈值损失函数中的两个阈值  $\alpha_1$ ,  $\alpha_2$  对模型检索效果的影响。正如在 5.2 节讨论的那样，我们使用在分类任务上微调好的模型（因为检索的模型是在分类模型的基础上进行优化的），从训练数据集采样图像对，然后计算图像对中两幅图像的特征之间的距离，然后画出相似图像对与不相似图像对特征距离的分布曲线。这两个分布曲线近似为正太分布，具体例子可以参见图 5.4 中的某一张图。我们把两个分布的均值作为阈值选择的起点，对于相似和不相似图像对，两个分布均值分别为 0.9 以及 1.2。图 5.3 展示了两个阈值对模型检索性能的影响（以 mAP 指标来衡量）。从图上可以看出，两个阈值的选取应该在分布的均值附近选择，才能取得比较高的检索结果，另外，两个阈值之间的间隔非常重要，这个间隔不能太「松弛」（也就是说  $\alpha_2 - \alpha_1$  太小），也不能太「紧」（ $\alpha_2 - \alpha_1$  太大），无论间隔太大或者太小，都会影响模型的性能。最终，我们选择设定  $\alpha_1 = 0.8$  以及  $\alpha_2 = 1.2$  来优化我们的模型。

### 5.3.4 检索效果的可视化

在本部分，我们将采用可视化的方式来帮助我们理解为什么所提出的方法能够取得更好的效果。我们比较了三种不同的模型，第一种是原始的 VGG16

表 5.3 与其他主流方法的结果比较

方法	特征维度					
	D=512	D=256	D=128	D=64	D=32	D=16
Neural codes [72]	15.9	15.8	15.7	14.6	13.7	10.5
SPoC [73]	23.3	23.4	23.1	22.4	20.6	18.4
MFC [68]	30.0	29.8	29.5	28.0	24.5	20.9
MAC [74]	36.1	36.4	36.5	35.4	32.3	27.8
Siamese-MAC [82]	35.6	36.2	37.2	37.3	35.2	31.4
TripletNet <sup>†</sup> [17]	67.98	68.57	69.60	69.97	68.1	60.67
Ours (retr-d)	45.79	46.17	46.61	46.75	45.36	42.03
Ours (cls)	66.57	67.2	68.53	68.67	67.47	59.99
Ours best (cls + retr-d)	<b>68.63</b>	<b>69.15</b>	<b>70.14</b>	<b>70.07</b>	<b>68.48</b>	<b>61.59</b>

<sup>†</sup> TripletNet 的权重是也由 Firearm14k 分类微调的模型初始化的。

模型，第二种是在 Firearm14k 数据库使用分类任务微调以后的模型，第三种是在分类模型基础上继续使用双阈值损失函数微调的模型。我们随机采样了大约 500000 相似与不相似图像对，然后我们分别使用了不同模型计算相似与不相似图像对特征距离。图 5.4 展示了不同模型下，正负样本对特征距离的分布曲线。对于原来的 VGG 模型（图 5.4，上图），相似图像对与不相似图像对的图像特征距离分布重合非常大，因此模型输出的特征对于判断两个图像是否是相似有很大的不确定性，所以检索结果较差（31.1% mAP）。经过在分类任务上的微调，我们得到了第二个模型（图 5.4，中图），相似与不相似图像对的特征距离分布之间的重叠大幅减少，因此模型的检索精度也有了很大的提升（检索的 mAP 为 65.4%）。当我们进一步在分类模型基础上使用双阈值损失函数对模型进行微调后，我们得到了最终的模型（图 5.4，下图），我们可以看到特别是相似图像对的特征距离分布发生了明显的变化：分布曲线的形状发生了变化，相对于中图的「矮胖」，变得更加「高瘦」。分布的变化也带来了检索结果的进一步提升（检索 mAP 68.4%）。在图 5.5 中，我们也展示了一些采用我们的方法得到的检索结果，从这些结果我们可以看出，当枪支图片只占据图像的一小部分时候，我们的方法可以准确找到相似的图片（第 2 个和第 7 个检索结果），另外我们的方法对于枪支的姿态角度变化也具有较强的鲁棒性。

### 5.3.5 与其他方法的对比

在本节，我们比较了所提出的双阈值方法与其他主流方法的检索结果，这些方法中，有的方法使用的是未经过微调的模型（称为 off-the-shelf 模型），另



图 5.5 我们的方法的一些示例检索结果

外一些方法使用的是经过微调的模型。为了保证对比的公平性，所有的方法都由我们使用 VGG16 模型实现。对于那些基于 off-the-shelf 模型的方法 [68, 72–74]，我们根据论文中所给的设置来实现对应的方法，在这些方法中，MFC 是我们在第 4 章提出的多尺度全卷积的方法。对于那些使用模型微调的方法 [17, 82]，我们使用了和我们的方法尽可能一致的设置，然后在 Firearm14k 数据库上训练模型。按照惯例 [73, 82]，从这些方法得到的图像特征都经过  $l_2$  归一化，然后经过了 PCA 变换，最后再一次经过  $l_2$  归一化处理。

首先，我们报告不同的方法在 mAP 指标上的结果，所有的实验结果参见表 5.3。这里的结果表明，off-the-shelf 模型（表 5.3，上部）的 mAP 分数都相对较低，这是由于 Firearm14k 数据库和 ImageNet 数据库的巨大差异，这些预训练的模型无法生成具有区分性的特征，因而检索效果并不理想。使用单阈值方法微调的模型 [82] 表现也不理想，使用三元组损失的 TripletNet [17] 的表现相对单阈

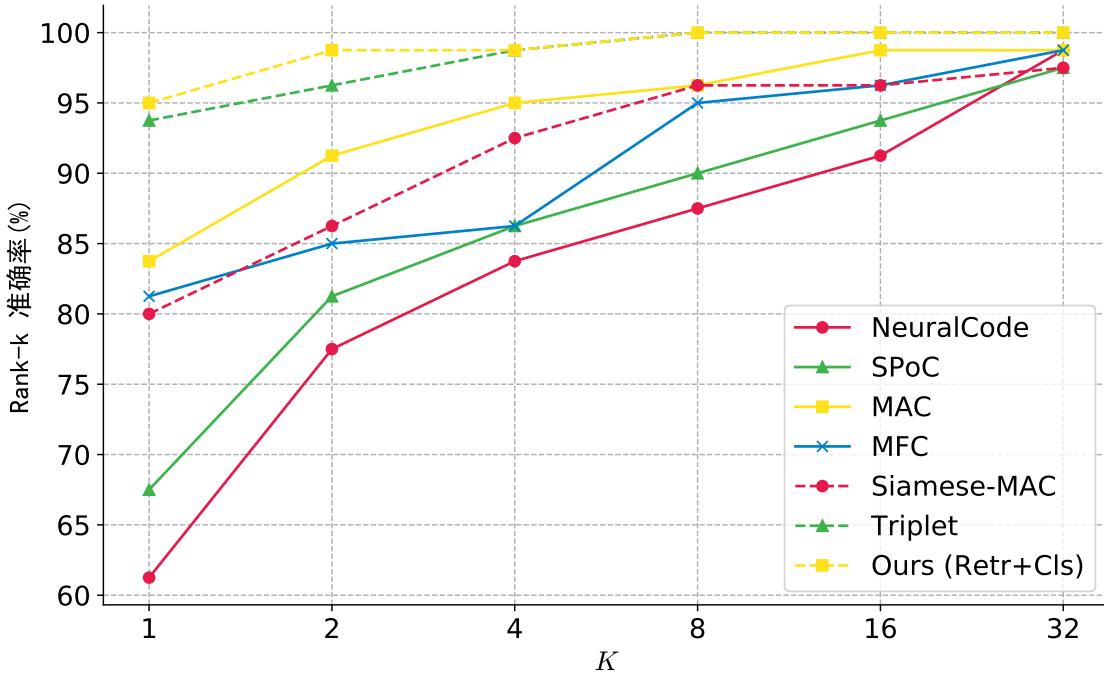


图 5.6 不同方法 Rank-k 准确率的比较 (%)  
Figure 5.6 Rank-k accuracy of different approaches (%)

值方法提高了很多。最后，我们的提出的方法在不同的特征维度下都超过了其他方法。值得注意的是，我们的方法在特征维度被压缩到极低的情况下仍然表现良好，例如当特征维度为 16 时，检索的 mAP 仍然达到了 61.59%。

我们也计算了不同方法在 Firearm14k 测试集上的 Rank-k 准确率，在这个实验中，我们使用的是 512 维特征，得到的实验结果如图 5.6 所示。从图中可以看出，当  $k$  比较大时 ( $k = 16$  或  $k = 32$  时)，各种方法的 Rank-k 准确率都较高，这说明 Rank-k 指标是比较容易的指标；当  $k$  比较小，SPoC [73]，NeuralCode [72] 和 Siamese-MAC [82] 等方法，表现较差，TripletNet 方法 [17] 和我们的方法此时表现较好，特别是我们的方法，在  $k = 1$  或  $k = 2$  时，表现要优于 TripletNet 方法。

#### 5.4 相关工作

枪支图像的检索问题之前很少有人研究，Wen 和 Yao [104] 提出用枪支的轮廓到中心的距离分布来检索相似图片，分布之间的不同用来衡量两个枪支图片的相似度。他们用到的图片背景都比较简单，并且数据库规模很小，大约为 300 张图片。

Babenko 等人 [72] 提出使用分类任务微调神经网络，并观察到检索的结果有所提高，他们使用的是交叉熵损失，这种损失并不直接针对检索任务，对未

知类别泛化能力不强。Arandjelovic 等人 [79] 将传统的 VLAD 方法用神经网络来实现，使用三元组损失来优化模型，将模型用在了地点检索任务上。Gordo 等人 [17] 提出使用三元组损失，并学习针对检索任务的图像的特征嵌入方法，他们同时结合了物体检测中的 ROIpooling 方法 [27] 来提升模型对物体的定位能力，提高模型的性能。Radenovic 等人 [82] 则关注于采用 structure from motion 技术，自动从大量图像中产生合适的训练样本对，他们采用了和我们的方法一样的双分支的 Siamese 网络来学习图像特征，发现使用双分支的网络结构能够取得比三分支结构更好的结果。Cao 等人 [105] 使用了和我们的双阈值损失函数相似的方法来处理过图像实例检索问题，他们在实验中采用了更加复杂的网络结构。

## 5.5 本章小结

在本章中，我们详细介绍了提出的双阈值对比损失函数，我们提出该方法的动机主要有两个，一是目前单阈值对比损失函数无法保证正负图像对在训练时贡献的损失的平衡，二是我们提出的 Firearm14k 数据库与 ImageNet 数据库之间的巨大差异导致直接微调模型效果欠佳。我们提出使用双阈值对比损失函数来解决模型训练过程中正负图像对贡献的损失不平衡的问题，同时提出使用两步训练的策略来解决数据库之间图像差异大的问题。在实验部分，我们把提出的方法与当前的主流方法进行了比较，并给出可视化的分析，证明我们的方法为什么有效，实验结果表明我们的方法在不同的特征维度下检索效果都超过了当前的主流方法。



## 第6章 工作总结与展望

从卷积神经网络在 2012 年的兴起 [16] 到图像检索领域的研究者将卷积神经网络与图像检索的技术结合起来 [70, 72]，再到最近基于卷积神经网络的方法在常用的一些数据库上效果已经超过了传统的方法 [17, 83]，短短四五年时间，基于卷积神经网络的方法在图像检索方向就取得了巨大的进步。本文的工作也是围绕卷积神经网络在图像检索领域的应用展开的，同时结合了对与敏感枪支图像检索的需求，下面对本文的工作做一总结，并对未来工作进行展望。

### 6.1 工作总结

本文围绕卷积神经网络在图像检索领域的应用，分析了目前方法的优缺点，并且结合对敏感图像识别的需求，开展了相关的工作，主要工作总结如下：

#### 1. 我们建立了一个大规模的枪支图像数据库

在社交网络上，大量的枪支图片会引起用户不适，需要适当的控制与处理。目前的基于深度卷积神经网络的检索方法，在模型的训练过程中都需要大量的训练数据，否则模型将过拟合。这些都要求一个数据量丰富的数据库，从而方便研究者开展该方面的研究，但是目前并没有一个大规模的枪支图片数据库存在。为了这样的需求，我们构建了一个大规模的枪支图像数据库，该数据包含 14755 张来自 167 类不同类型枪支的图片，可以为枪支图片分类与检索研究提供所需的数据基础。

#### 2. 我们提出了一种多尺度全卷积的图像实例检索方法

在工作中，我们发现目前基于卷积神经网络的图像检索方法在提取图像特征时，并未详细分析影响提取的特征有效性的因素，采用的设置都是一些比较随意的选择，因此我们通过实验详细分析了三个重要因素对提取的特征有效性的影响，这三个因素分别是：输入神经网络的图像尺寸，多尺度特征表达以及 PCA 和白化矩阵的学习方式。结合我们的实验和分析，我们提出了一种多尺度全卷积的图像实例检索方法，我们的方法在多个数据库上都取得了良好的效果。

#### 3. 我们提出了一种基于双阈值对比损失函数的枪支图像检索方法

在当前的社交网络上，大量的枪支图片的出现，要求社交媒体的监管者对



图 6.1 两个图片剪切篡改的实例（左：donor 图片，右：剪切拼接后图片），请放大查看

这些图片进行适当的处理，在取证中，也有大量需要鉴定枪支图片类型的需求，这些问题都可以通过图像检索相关的技术得到有效解决。在本工作中，我们研究了微调卷积神经网络进行枪支图片检索的可行性，我们发现，传统的单阈值对比损失函数对于枪支图像检索，存在一定的缺点：第一，使用该函数，在模型的训练过程中，正负样本贡献的损失不平衡，模型的训练偏向负样本对；第二，由于枪支图片与训练这些基础模型的 ImageNet 存在巨大的域差异，直接使用这些模型针对检索任务微调，并不能取得良好的效果。我们提出使用双阈值对比损失函数解决网络训练中正负样本贡献的损失不平衡的问题。我们采用两步训练的策略缓和域差异带来的性能下降问题，第一步，我们在 Firearm14k 上针对分类任务微调网络，第二步再针对检索任务（使用双阈值对比损失函数）微调模型。在 Firearm14k 测试集上，我们把提出的方法与其他主流方法进行了对比，实验结果表明，我们的方法得到的检索结果在不同特征维度上都要优于其他主流方法。

## 6.2 工作展望

虽然我们的方法在图像检索任务上取得了不错的成绩，但是该领域仍有许

多没有完全解决的问题，未来的工作将关注以下几个问题：

#### 1. 利用显著性分析或注意力机制解决图像中小物体检索的问题

在我们的工作中，我们发现基于卷积神经网络的方法，在图像中的物体尺寸非常小的情况下，效果并不理想，因为卷积神经网络提取的是图像的全局特征，并不像 SIFT 特征描述子一样提取的是局部特征。包含小物体的图像，如果使用卷积神经网络提取特征，特征中可能包含了大量周边环境的特征，对小物体的检索造成了干扰，我们可以结合一些显著性分析或者注意力机制的方法 [106]，自动检测图像中可能的物体，然后提取显著性区域的特征，忽略周边环境的干扰，从而增强小物体图像检索的有效性。

#### 2. 研究基于深度学习的局部特征描述特征

深度学习特征本质上是图像全局信息的表达，特别是网络经过多层卷积与池化以后，特征图中一个元素在原图上对应着很大一片区域（感受野比较大），无法像传统的 SIFT 特征一样，进行空间几何校验，进行检索结果重排序，提高检索的准确率。目前这方面的工作还比较少，Noh 等 [83] 提出的方法就是试图学习局部深度特征，但是他们使用的训练方法并未针对检索的任务，局部特征有效性有待提高，因此这方面的研究也是未来的一个方向。

#### 3. 研究基于特征量化和哈希的精细图像检索问题

在本文中所使用的特征都是实值特征，并不是二值化的特征，在检索数据库规模小的情况下，这些方法是可行的，但是当数据库规模巨大（亿级别或者十亿级别）时，这样的方法弊端是明显的，一方面特征存储需要消耗大量空间，另外计算特征之间的相似度也会消耗大量时间，达不到实时性的要求。另外目前哈希方法使用的数据库都是一些粗糙类别层次的数据库，如 CIFAR10<sup>1</sup>, NUSWIDE [107]，并不是精细类别的数据库，因此基于特征量化或者深度学习的图像特征哈希编码学习也是一个非常重要的研究方向。

#### 4. 利用图像检索的方法解决图像取证的问题

随着数字图像技术的发展，图像伪造技术也不断进步，网络上以及媒体上出现了各种伪造的照片。在图像取证中，有一类图片伪造方法是剪切拼接 (splicing) [108]，也就是把一张图片（称为 donor 图片）一部分剪切拼接到另一张图片上（称为 base 图片），达到欺骗目的，图 6.1 展示了两例剪切拼接的实例（绿色框中是对应的相同物体）。如果能通过搜索技术找到剪切后的图像中

<sup>1</sup><https://www.cs.toronto.edu/~kriz/cifar.html>

物体的原图，那么就可以很容易判断图像是被篡改过的，因此基于图像检索的技术也能在剪切检测中发挥一定的作用。

## 参考文献

- [1] PENTLAND A, PICARD R W, SCLAROFF S. Photobook: Content-based manipulation of image databases[J]. International Journal of Computer Vision, 1996, 18(3): 233-254.
- [2] NIBLACK W, BARBER R, EQUITZ W, et al. The QBIC project: Querying images by content, using color, texture, and shape[C]//Storage and Retrieval for Image and Video Databases. Washington: SPIE, 1993: 173-187.
- [3] BACH J R, FULLER C, GUPTA A, et al. Virage image search engine: An open framework for image management[C]//Storage and Retrieval for Image and Video Databases. Washington: SPIE, 1996.
- [4] GEVERS T, SMEULDERS A W M. PicToSeek: combining color and shape invariant features for image retrieval[J]. IEEE Transactions on Image Processing, 2000, 9(1): 102-119.
- [5] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60: 91-110.
- [6] SIVIC J, ZISSERMAN A. Video Google: A text retrieval approach to object matching in videos[C]//Proceedings Ninth IEEE International Conference on Computer Vision. Washington: IEEE Computer Society, 2003: 1470-1477.
- [7] PHILBIN J, CHUM O, ISARD M, et al. Lost in quantization: Improving particular object retrieval in large scale image databases[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2008: 1-8.
- [8] PHILBIN J, CHUM O, ISARD M, et al. Object retrieval with large vocabularies and fast spatial matching[C]//2007 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2007: 1-8.
- [9] MIKULÍK A, PERDOCH M, CHUM O, et al. Learning a fine vocabulary[C]//Computer Vision – ECCV 2010. Berlin/Heidelberg: Springer, 2010: 1-14.
- [10] ARANDJELOVIĆ R, ZISSERMAN A. Three things everyone should know to improve object retrieval[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2012: 2911-2918.
- [11] CHUM O, PHILBIN J, SIVIC J, et al. Total recall: Automatic query expansion with a generative feature model for object retrieval[C]//2007 IEEE 11th International Conference on Computer Vision. Washington: IEEE Computer Society, 2007: 1-8.
- [12] DRANGE M. Why is this canadian hacker better than facebook at detecting gun photos? [EB/OL]. (2016-03-31)[2018-03-10]. [https://www.forbes.com/sites/mattdrange/2016/03/31/facebook-guns-beet\\_farmer-image-recognition/#5fae664024f2](https://www.forbes.com/sites/mattdrange/2016/03/31/facebook-guns-beet_farmer-image-recognition/#5fae664024f2).
- [13] MELE C. Facebook banned gun sales. so why is it still 'full of them'?

- [EB/OL]. (2016-07-21)[2018-04-05]. <https://www.nytimes.com/2016/07/22/technology/facebook-banned-gun-sales-so-why-is-it-still-full-of-them.html>.
- [14] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115(3): 211-252.
- [15] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.
- [16] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. New York: Curran Associates, 2012: 1097-1105.
- [17] GORDO A, ALMAZÁN J, REVAUD J, et al. Deep image retrieval: Learning global representations for image search[C]//Computer Vision – ECCV 2016. Berlin/Heidelberg: Springer, 2016: 241-257.
- [18] NISTÉR D, STEWÉNIUS H. Scalable recognition with a vocabulary tree[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Washington: IEEE Computer Society, 2006: 2161-2168.
- [19] JÉGOU H, DOUZE M, SCHMID C. Hamming embedding and weak geometric consistency for large scale image search[C]//Computer Vision – ECCV 2010. Berlin/Heidelberg: Springer, 2008: 304-317.
- [20] LOWE D G. Object recognition from local scale-invariant features[C]//Proceedings of the Seventh IEEE International Conference on Computer Vision. Washington: IEEE Computer Society, 1999: 1150-1157.
- [21] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. CoRR, 2014, abs/1409.1556.
- [22] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington: IEEE Computer Society, 2015: 1-9.
- [23] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington: IEEE Computer Society, 2016: 770-778.
- [24] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]//Computer Vision – ECCV 2016. Berlin/Heidelberg: Springer, 2016: 21-37.
- [25] REDMON J, DIVVALA S K, GIRSHICK R B, et al. You only look once: Unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington: IEEE Computer Society, 2016: 779-788.
- [26] LIN T Y, GOYAL P, GIRSHICK R B, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV). Washington: IEEE Computer Society, 2017: 2999-3007.

- [27] REN S, HE K, GIRSHICK R B, 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.
- [28] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.
- [29] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.
- [30] NOH H, HONG S, HAN B. Learning deconvolution network for semantic segmentation [C]//2015 IEEE International Conference on Computer Vision (ICCV). Washington: IEEE Computer Society, 2015: 1520-1528.
- [31] SWAIN M J, BALLARD D H. Color indexing[J]. International Journal of Computer Vision, 1991, 7(1): 11-32.
- [32] FORSYTH D A. A novel algorithm for color constancy[J]. International Journal of Computer Vision, 1990, 5(1): 5-35.
- [33] FUNT B V, FINLAYSON G D. Color constant color indexing[J]. IEEE Transantions on Pattern Analysis and Machine Intelligence, 1995, 17(5): 522-529.
- [34] DENG Y, MANJUNATH B S, KENNEY C S, et al. An efficient color representation for image retrieval[J]. IEEE Transactions on Image Processing, 2001, 10(1): 140-147.
- [35] MANJUNATH B S, MA W Y. Texture features for browsing and retrieval of image data[J]. IEEE Transantions on Pattern Analysis and Machine Intelligence, 1996, 18(8): 837-842.
- [36] JAIN A K, VAILAYA A. Image retrieval using color and shape[J]. Pattern Recognition, 1996, 29(8): 1233-1244.
- [37] SMEULDERS A W M, WORRING M, SANTINI S. Content-based image retrieval at the end of the early years[J]. IEEE Transanctions on Pattern Analysis and Machine Intelligence, 2000, 22(12): 1349-1380.
- [38] O'HARA S, DRAPER B A. Introduction to the bag of features paradigm for image classification and retrieval[J]. CoRR, 2011, abs/1101.3354.
- [39] NOWAKE, JURIE F, TRIGGS B. Sampling strategies for bag-of-features image classification [C]//Computer Vision – ECCV 2006. Berlin/Heidelberg: Springer, 2006: 490-503.
- [40] YANG J, JIANG Y G, HAUPTMANN A G, et al. Evaluating bag-of-visual-words representations in scene classification[C]//Proceedings of the International Workshop on Multimedia Information Retrieval. New York: ACM, 2007: 197-206.
- [41] MANNING C D, RAGHAVAN P, SCHÜTZ H. Introduction to information retrieval[M]. Cambridge: Cambridge University Press, 2008.
- [42] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

- [43] VOGEL J, SCHIELE B. Natural scene retrieval based on a semantic modeling step[C]// International Conference on Image and Video Retrieval. Berlin/Heidelberg: Springer, 2004: 207-215.
- [44] LI F, PERONA P. A bayesian hierarchical model for learning natural scene categories[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Washington: IEEE Computer Society, 2005: 524-531.
- [45] MIKOLAJCZYK K, SCHMID C. Scale & affine invariant interest point detectors[J]. International Journal of Computer Vision, 2004, 60(1): 63-86.
- [46] JÉGOU H, DOUZE M, SCHMID C. Packing bag-of-features[C]//2009 IEEE 12th International Conference on Computer Vision. Washington: IEEE Computer Society, 2009: 2357-2364.
- [47] PERONNIN F, DANCE C R. Fisher kernels on visual vocabularies for image categorization [C]//2007 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2007: 1-8.
- [48] PERONNIN F, LIU Y, SÁNCHEZ J, et al. Large-scale image retrieval with compressed fisher vectors[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2010: 3384-3391.
- [49] DOUZE M, RAMISA A, SCHMID C. Combining attributes and fisher vectors for efficient image retrieval[C]//2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2011: 745-752.
- [50] JÉGOU H, DOUZE M, SCHMID C, et al. Aggregating local descriptors into a compact image representation[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2010: 3304-3311.
- [51] ARANDELOVIĆ R, ZISSERMAN A. All about VLAD[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2013: 1578-1585.
- [52] JÉGOU H, DOUZE M, SCHMID C. On the burstiness of visual elements[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2009: 1169-1176.
- [53] CUN Y L, MATAN O, BOSEN B, et al. Handwritten zip code recognition with multi-layer networks[C]//Proceedings of the 10th International Conference on Pattern Recognition. Washington: IEEE Computer Society, 1990: 35-40.
- [54] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
- [55] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]// Computer Vision – ECCV 2014. Berlin/Heidelberg: Springer, 2014: 818-833.
- [56] GLOROT X, BORDES A, BENGIO Y. Deep sparse rectifier neural networks[C]//Proceedings

- of the Fourteenth International Conference on Artificial Intelligence and Statistics. Fort Lauderdale: PMLR, 2011: 315-323.
- [57] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//2015 IEEE International Conference on Computer Vision (ICCV). Washington: IEEE Computer Society, 2015: 1026-1034.
- [58] MAAS A L, HANNUN A Y, NG A Y. Rectifier nonlinearities improve neural network acoustic models[C]//Proceedings of the 30th International Conference on Machine Learning. Fort Lauderdale: PMLR, 2013: 3-8.
- [59] SALAKHUTDINOV R, HINTON G E. Deep boltzmann machines[C]//Proceedings of the 12th International Conference on Artificial Intelligence and Statistics: volume 5. Fort Lauderdale: PMLR, 2009: 448-455.
- [60] SALAKHUTDINOV R, HINTON G E. An efficient learning procedure for deep boltzmann machines[J]. Neural computation, 2012, 24(8): 1967-2006.
- [61] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington: IEEE Computer Society, 2017: 2261-2269.
- [62] GATYS L A, ECKER A S, BETHGE M. Image style transfer using convolutional neural networks[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington: IEEE Computer Society, 2016: 2414-2423.
- [63] JOHNSON J, ALAHI A, FEI-FEI L. Perceptual losses for real-time style transfer and super-resolution[C]//Computer Vision – ECCV 2016. Berlin/Heidelberg: Springer, 2016: 694-711.
- [64] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// Advances in Neural Information Processing Systems 27. New York: Curran Associates, 2014: 2672-2680.
- [65] MIRZA M, OSINDERO S. Conditional generative adversarial nets[J]. CoRR, 2014, abs/1411.1784.
- [66] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518: 529-533.
- [67] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of go with deep neural networks and tree search[J]. Nature, 2016, 529: 484-489.
- [68] HAO J, WANG W, DONG J, et al. MFC: A multi-scale fully convolutional approach for visual instance retrieval[C]//2017 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). Washington: IEEE Computer Society, 2017: 513-518.
- [69] RAZAVIAN A S, AZIZPOUR H, SULLIVAN J, et al. CNN features off-the-shelf: An astounding baseline for recognition[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Washington: IEEE Computer Society, 2014: 512-519.
- [70] GONG Y, WANG L, GUO R, et al. Multi-scale orderless pooling of deep convolutional

- activation features[C]//Computer Vision – ECCV 2014. Berlin/Heidelberg: Springer, 2014: 392-407.
- [71] NG J Y H, YANG F, DAVIS L S. Exploiting local features from deep networks for image retrieval[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Washington: IEEE Computer Society, 2015: 53-61.
- [72] BABENKO A, SLESAREV A, CHIGORIN A, et al. Neural codes for image retrieval[C]// Computer Vision – ECCV 2014. Berlin/Heidelberg: Springer, 2014: 584-599.
- [73] BABENKO A, LEMPITSKY V S. Aggregating local deep features for image retrieval [C]//2015 IEEE International Conference on Computer Vision (ICCV). Washington: IEEE Computer Society, 2015: 1269-1277.
- [74] TOLIAS G, SICRE R, JÉGOU H. Particular object retrieval with integral max-pooling of CNN activations[J]. CoRR, 2015, abs/1511.05879.
- [75] SEDDATI O, DUPONT S, MAHMOUDI S, et al. Towards good practices for image retrieval based on cnn features[C]//2017 IEEE International Conference on Computer Vision Workshops (ICCVW). Washington: IEEE Computer Society, 2017: 1246-1255.
- [76] ZHOU W, LI H, SUN J. Collaborative index embedding for image retrieval.[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99): 1-1.
- [77] WANG J, SONG Y, LEUNG T, et al. Learning fine-grained image similarity with deep ranking [C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2014: 1386-1393.
- [78] SCHROFF F, KALENICHENKO D, PHILBIN J. FaceNet: A unified embedding for face recognition and clustering[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington: IEEE Computer Society, 2015: 815-823.
- [79] ARANDJELOVIĆ R, GRONĀT P, TORII A, et al. NetVLAD: CNN architecture for weakly supervised place recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington: IEEE Computer Society, 2016: 5297-5307.
- [80] BELL S, BALA K. Learning visual similarity for product design with convolutional neural networks[J]. ACM Transanctions on Graphics (SIGGRAPH), 2015, 34(4): 98:1-98:10.
- [81] HADSELL R, CHOPRA S, LECUN Y. Dimensionality reduction by learning an invariant mapping[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06): volume 2. Washington: IEEE Computer Society, 2006: 1735-1742.
- [82] RADENOVIC F, TOLIAS G, CHUM O. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples[C]//Computer Vision – ECCV 2016. Berlin/Heidelberg: Springer, 2016: 3-20.
- [83] NOH H, ARAUJO A, SIM J, et al. Large-scale image retrieval with attentive deep local features[C]//2017 IEEE International Conference on Computer Vision (ICCV). Washington: IEEE Computer Society, 2017: 3476-3485.

- [84] EVERINGHAM M, ESLAMI S M A, GOOL L V, et al. The pascal visual object classes challenge: A retrospective[J]. International Journal of Computer Vision, 2014, 111(1): 98-136.
- [85] LIN T Y, MAIRE M, BELONGIE S J, et al. Microsoft COCO: Common objects in context [C]//Computer Vision – ECCV 2014. Berlin/Heidelberg: Springer, 2014: 740-755.
- [86] HUANG G B, RAMESH M, BERG T, et al. Labeled faces in the wild: A database for studying face recognition in unconstrained environments[R]. Amherst: University of Massachusetts, 2007.
- [87] WOLF L, HASSNER T, MAOZ I. Face recognition in unconstrained videos with matched background similarity[C]//2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington: IEEE Computer Society, 2011: 529-534.
- [88] KEMELMACHER-SHLIZERMAN I, SEITZ S M, MILLER D, et al. The megaface benchmark: 1 million faces for recognition at scale[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2016: 4873-4882.
- [89] LAZEBNIK S, SCHMID C, PONCE J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Washington: IEEE Computer Society, 2006: 2169-2178.
- [90] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.
- [91] JIA Y, SHELHAMER E, DONAHUE J, et al. Caffe: Convolutional architecture for fast feature embedding[C]//Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM, 2014: 675-678.
- [92] HSU T. Bumble dating app bans gun images after mass shootings[EB/OL]. (2018-03-05)[2018-03-12]. <https://www.nytimes.com/2018/03/05/business/bumble-dating-app-gun-images.html>.
- [93] SONG H O, XIANG Y, JEGELKA S, et al. Deep metric learning via lifted structured feature embedding[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington: IEEE Computer Society, 2016: 4004-4012.
- [94] WANG J, ZHOU F, WEN S, et al. Deep metric learning with angular loss[C]//2017 IEEE International Conference on Computer Vision (ICCV). Washington: IEEE Computer Society, 2017: 2612-2620.
- [95] ZHAO R, OUYANG W, WANG X. Unsupervised salience learning for person re-identification [C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2013: 3586-3593.

- [96] LI W, ZHAO R, XIAO T, et al. DeepReID: Deep filter pairing neural network for person re-identification[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2014: 152-159.
- [97] WEN Y, ZHANG K, LI Z, et al. A discriminative feature learning approach for deep face recognition[C]//Computer Vision – ECCV 2016. Berlin/Heidelberg: Springer, 2016: 499-515.
- [98] TAIGMAN Y, YANG M, RANZATO M, et al. DeepFace: Closing the gap to human-level performance in face verification[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2014: 1701-1708.
- [99] CHOPRA S, HADSELL R, LECUN Y. Learning a similarity metric discriminatively, with application to face verification[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Washington: IEEE Computer Society, 2005: 539-546.
- [100] HAN X, LEUNG T, JIA Y, et al. MatchNet: Unifying feature and metric learning for patch-based matching[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington: IEEE Computer Society, 2015: 3279-3286.
- [101] G V K B, CARNEIRO G, REID I D. Learning local image descriptors with deep siamese and triplet convolutional networks by minimizing global loss functions[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington: IEEE Computer Society, 2016: 5385-5394.
- [102] WEINBERGER K Q, BLITZER J, SAUL L K. Distance metric learning for large margin nearest neighbor classification[C]//Advances in Neural Information Processing Systems 18. New York: Curran Associates, 2006: 1473-1480.
- [103] JÉGOU H, DOUZE M, SCHMID C. Product quantization for nearest neighbor search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(1): 117-128.
- [104] WEN C Y, YAO J Y. Pistol image retrieval by shape representation[J]. Forensic science international, 2005, 155: 35-50.
- [105] CAO J, HUANG Z, WANG P, et al. Quartet-net learning for visual instance retrieval[C]//Proceedings of the 2016 ACM on Multimedia Conference. New York: ACM, 2016: 456-460.
- [106] SONG J, YU Q, SONG Y Z, et al. Deep spatial-semantic attention for fine-grained sketch-based image retrieval[C]//2017 IEEE International Conference on Computer Vision (ICCV). Washington: IEEE Computer Society, 2017: 5552-5561.
- [107] CHUA T S, TANG J, HONG R, et al. NUS-WIDE: a real-world web image database from national university of singapore[C]//Proceedings of the ACM International Conference on Image and Video Retrieval. New York: ACM, 2009.
- [108] FARID H. Image forgery detection[J]. IEEE Signal Processing Magazine, 2009, 26: 16-25.

## 致 谢

在此论文完成之际，感谢中科院自动化所智能感知与计算研究中心为我的学习科研提供的软硬件环境，感谢导师的指导，感谢实验室的各位老师，同学以及工作人员等对我的帮助！

另外，感谢我的朋友们对我的支持，感谢我的父母对我的支持。



## 作者简历及攻读学位期间发表的学术论文与研究成果

### 作者简历:

2010 年 9 月 – 2014 年 6 月，在中南大学信息科学与工程学院，获得学士学位。

2014 年 9 月 – 2018 年 7 月，在中国科学院自动化研究所攻读硕士学位。

### 已发表(或正式接受)的学术论文:

1. **HAO J, WANG W, DONG J, et al.** MFC: A multi-scale fully convolutional approach for visual instance retrieval[C]//2017 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2017: 513 – 518
2. **HAO J, DONG J, WANG W, et al.** DeepFirearm: Learning Discriminative Feature Representation for Fine-grained Firearm Retrieval[C]//The 24th International Conference on Pattern Recognition (ICPR 2018), 2018

### 申请或已获得的专利:

1. 谭铁牛, 王伟, 董晶, 郝杰东.基于目标体图像的目标体检索方法、系统及装置。申请号: 201810460265.8

### 参加的研究项目:

1. 国家自然科学基金：“基于迁移学习的图像隐写分析新方法研究”，项目编号：U1536120
2. 国家自然科学基金重点项目：“应对大数据分析的情景融合个性化隐写理论与方法”，项目编号：U1636201
3. 国家自然科学基金青年科学基金项目：“基于成像环境约束的低质量图像篡改取证研究”，项目编号：61502496
4. 国家重点研发计划项目：“群体视觉大数据的透彻感知关键技术”，项目编号：2016YFB1001003
5. 北京市自然科学基金项目：“基于深度学习的数字图像取证研究”，项目编号：4164102

6. 国家自然科学基金重点项目：“具有强泛化能力的通用图像隐写检测技术研究”，项目编号：U1736119
7. 国家自然科学基金面上项目：“基于深度对抗学习的隐蔽通信新方法研究”，项目编号：61772529