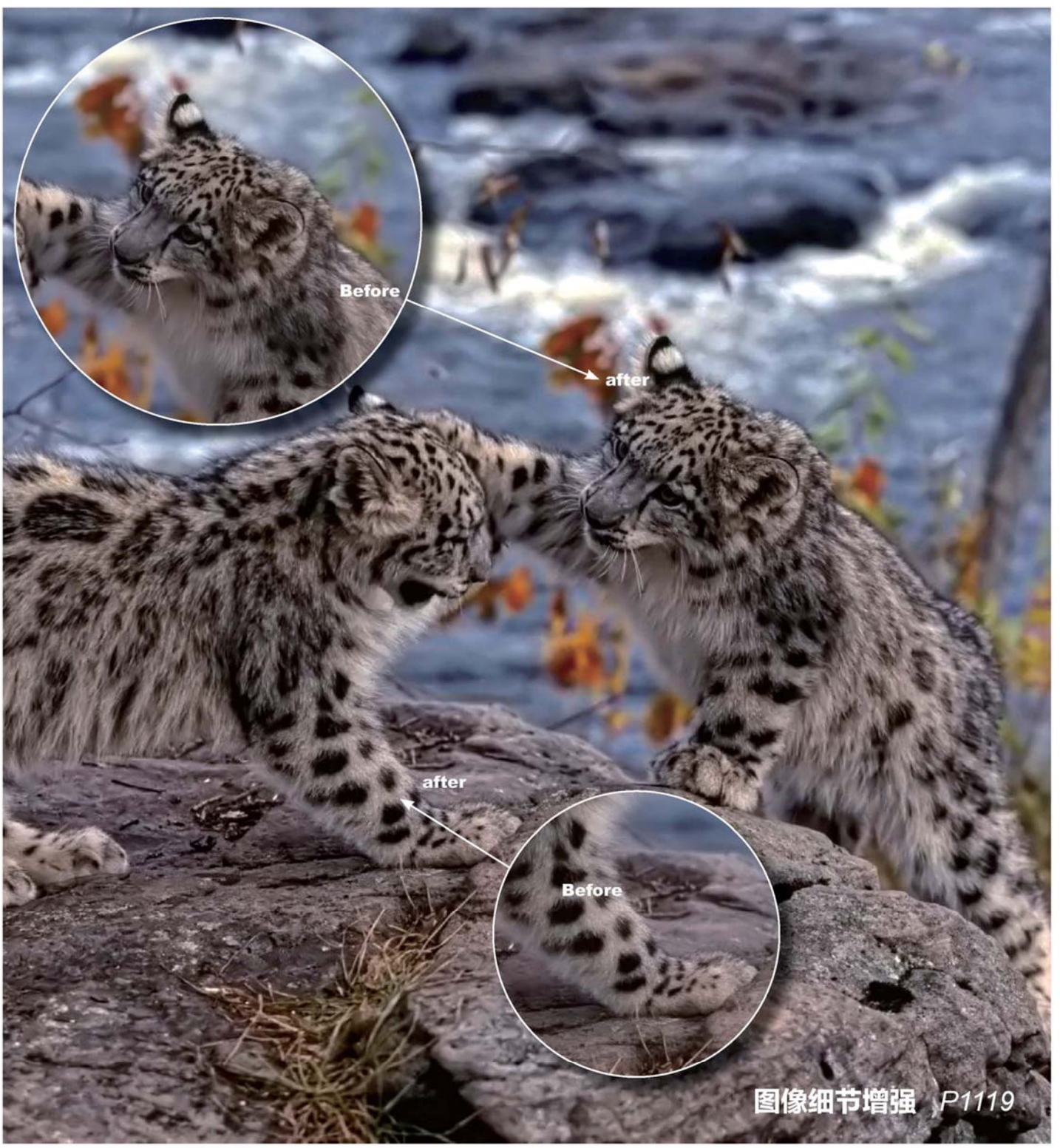
JIG
JOURNAL OF IMAGE AND GRAPHICS
主办：中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2016
09
VOL.21
ISSN1006-8961
CN11-3758/TB



图像细节增强 P1119

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）



第21卷第9期（总第245期）
2016年9月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 顾行发
编辑出版 《中国图象图形学报》编辑出版委员会
邮政信箱 北京9718信箱
邮　　编 100101
电子信箱 jig@radi.ac.cn
电　　话 010-64807995
网　　址 www.cjig.cn

广告经营许可证 京朝工商广字第0361号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics
Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Institute of Remote Sensing and Digital Earth, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Gu Xingfa
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
P.O.Box 9718, Beijing, P.R.China
Zip code 100101
E-mail jig@radi.ac.cn
Telephone 010-64807995
Website www.cjig.cn

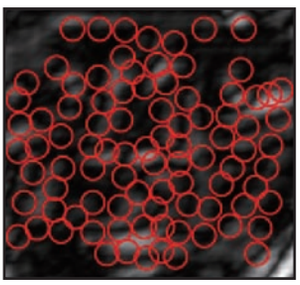
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

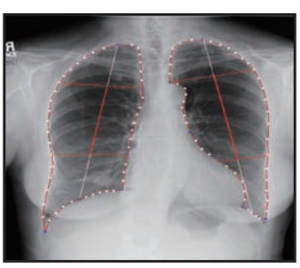
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2016年9月 第21卷 第9期（总第245期）
中国图象图形学报 Zhongguo Tuxiang Tuxing Xuebao

目次



基于频率簇模型的人脸识别(第1166页)

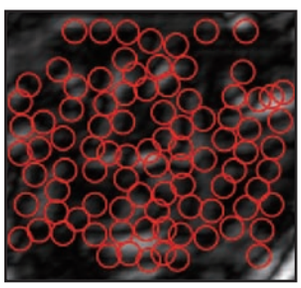


胸部解剖结构回归模型的虚拟双能量X线减影方法(第1247页)

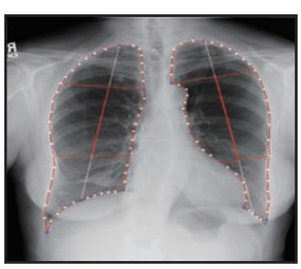
参考1维光谱差异的区域生长种子点选取方法(第1256页)

Volume 21, Number 9
Published September 16, 2016

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Face recognition method based on frequency cluster (P1166)



Virtual dual-energy subtraction method for X-ray radiographs by using regression model based on chest anatomical structure(P1247)

Seed extraction method for seeded region growing based on one-dimensional spectral differences(P1256)

中图法分类号：TP391　　**文献标识码**：A　　**文章编号**：1006-8961(2016)09-1147-08

论文引用格式：Feng Y S, Wang Z L. Fine-grained image categorization with segmentation based on top-down attention map[J]. Journal of Image and Graphics, 2016, 21(9): 1147-1154.[冯语姗, 王子磊. 自上而下注意图分割的细粒度图像分类[J]. 中国图象图形学报, 2016, 21(9): 1147-1154.][DOI: 10.11834/jig.20160904]

自上而下注意图分割的细粒度图像分类

冯语姗，王子磊

中国科学技术大学自动化系，合肥　230027

摘　要：**目的**　针对细粒度图像分类中的背景干扰问题，提出一种利用自上而下注意图分割的分类模型。**方法**　首先，利用卷积神经网络对细粒度图像库进行初分类，得到基本网络模型。再对网络模型进行可视化分析，发现仅有部分图像区域对目标类别有贡献，利用学习好的基本网络计算图像像素对相关类别的空间支持度，生成自上而下注意图，检测图像中的关键区域。再用注意图初始化 GraphCut 算法，分割出关键的目标区域，从而提高图像的判别性。最后，对分割图像提取 CNN 特征实现细粒度分类。**结果**　该模型仅使用图像的类别标注信息，在公开的细粒度图像库 Cars196 和 Aircrafts100 上进行实验验证，最后得到的平均分类正确率分别为 86.74% 和 84.70%。这一结果表明，在 GoogLeNet 模型基础上引入注意信息能够进一步提高细粒度图像分类的正确率。**结论**　基于自上而下注意图的语义分割策略，提高了细粒度图像的分类性能。由于不需要目标窗口和部位的标注信息，所以该模型具有通用性和鲁棒性，适用于显著性目标检测、前景分割和细粒度图像分类应用。

关键词：细粒度图像分类；卷积神经网络；自上而下注意图；GraphCut；GoogLeNet

Fine-grained image categorization with segmentation based on top-down attention map

Feng Yushan, Wang Zilei

Department of Automation, University of Science and Technology of China, Hefei　230027, China

Abstract: **Objective** This paper addresses the problem of fine-grained visual categorization requiring domain-specific and expert knowledge. This task is challenging because of all the objects in a database belonging to the same basic level category, with very fine differences between classes. These subtle differences are easily overcome by image background information that is seldom discriminative and often serves as a distractor, which reduces recognition performance in fine-grained image categorization. Therefore, segmenting the background regions and localizing discriminative image parts are important precursors to fine-grained image categorization. To this end, a new segmentation algorithm based on top-down attention maps is proposed to detect discriminative objects and discount the influence of background. Once the objects are localized, they are described by convolutional neural network (CNN) features to predict the corresponding class. **Method** The proposed method first recognizes a dataset through the CNN model to obtain a ConvNet basis with GoogLeNet structure. The

收稿日期：2016-03-15；**修回日期**：2016-05-12

基金项目：国家自然科学基金项目(61233003)；安徽省自然科学基金项目(1408085MF112)；中央高校研究基金科研业务费专项资金(WK3500000002)

第一作者简介：冯语姗(1990—)，女，中国科学技术大学信息科学技术学院模式识别与智能系统专业在读硕士研究生，主要研究方向为计算机视觉、深度学习、图像分类。E-mail: fyushan@mail.ustc.edu.cn

Supported by: National Natural Science Foundation of China(61233003)

ConvNet basis is visualized to build reliable intuitions of the visual information contained in the trained CNN representations, and the visualization result shows that the saliency parts in the images correspond to the object regions while the activations of background regions are very small. Then, according to the learned ConvNet, we predict the top-1 class label of a given image and determine the spatial support of the predicted class among the image pixels. Spatial support is rearranged to produce a top-down attention map, which can effectively locate the informative image object regions. Next, given an image and its corresponding image-specific class attention map, we compute the object segmentation mask with the GraphCut segmentation algorithm. The high-quality foreground segmentations are then used to encode the image appearance into a highly discriminative visual representation by finetuning the ConvNet basis to learn a new segmentation ConvNet. Finally, the ConvNet basis and segmentation ConvNet are combined to conduct fine-grained image recognition. We also use the original images to finetune the segmentation ConvNet for improved accuracy. **Result** The proposed model was tested on two new benchmark datasets available for fine-grained image categorization. The two databases were Cars196 and Aircrafts100, which were designed for fine-grained image recognition with public annotations, including class labels, object bounding boxes, and part locations. Only the class label annotation was used in our evaluation, and the final average accuracy rates of Cars196 and Aircrafts100 databases were 86.74% and 84.70%, respectively. These results show that adding visual attention information is more accurate than the GoogLeNet model alone. **Conclusion** A semantic segmentation strategy based on top-down attention map was proposed to improve the accuracy of fine-grained image categorization. Our method did not need any bounding box or part annotations, making it very robust and applicable to a variety of datasets. The experimental results show that the attention information was very useful for fine-grained image recognition. The proposed novel model proved capable of application to salient object detection, foreground segmentation, and fine-grained image categorization.

Key words: fine-grained visual categorization; convolutional neural network (CNN); top-down attention map; GraphCut; GoogLeNet

0 引 言

图像分类的目标是从固定的类别标签集合中，预测出给定图像对应的类别。在机器视觉研究中，图像分类任务主要包括粗粒度图像分类和细粒度图像分类两种。其中，粗粒度图像分类的对象属性差异较大，例如汽车、人、树等；而细粒度图像分类的对象通常属于同一个大类，例如细粒度图像库CUB200[1]中的200种鸟类和Flower102[2]中的102种花类等。由于细粒度类别属于同一个大类，所以各类别之间的差距很小，这些细微的差距容易被光照、颜色、背景、形状和位置等变化因素覆盖，导致细粒度图像分类相对困难。

最近几年，随着大型数据库ImageNet[3]、图形处理器(GPU)以及训练技巧(dropout正则化等)的发展，卷积神经网络(CNN)模型在图像分类中得到了突破进展，分类效果显著提升。对于粗粒度图像分类，Krizhevsky[4]使用CNN模型，在ILSVRC2012分类挑战赛中，将top-5平均分类错误率从传统FV(fisher vector)特征模型的26.1%直接降到了16.4%；Zeiler[5]对Caltech101图像数据库采用传统特征分类模型，获得的平均分类正确率为84.3%，采用CNN特征则为86.5%，Caltech256数据库的正确率也从55.2%上升到了74.2%。然而，尽管CNN模型在粗粒度图像分类中的效果很好，但是Russakovsky[6]却表明CNN不适合直接应用于细粒度图像分类。

目前，针对细粒度图像分类问题，大部分研究工作分两步进行，首先进行目标和目标各部位的检测，再对检测区域进行特征表示与识别。例如，Zhang[7]利用选择性搜索算法提取大量的图像区域，再根据目标和部位之间的几何约束来实现图像的pose-normalized[8]表示，最后进行分类，该模型在目标和部位检测期间的计算量较大。Goring[9]提出一个标签转移算法，给定测试图像，从训练集中找出k邻近图像，再把邻近图像的部位标注信息转移到测试图像，从而实现部位检测，最后实现分类，该模型在寻找邻近图像阶段的复杂度较高。文献[10-11]则首先利用梯度方向直方图(HOG)和DPM(deformable parts model)[12]以及Poselet[13]方法检测目标和各个部位，再对检测区域提取CNN特征进行分类，然而基于HOG特征的目标检测器鲁棒性不强。此外，上述方法在训练或测试阶段需要目标和关键部位的位置

标注信息，获取这些信息需要大量的人工标注。

Angelova[14]表明细粒度图像分类的关键是找到图像的显著性区域。Chai[15]认为细粒度图像分类的重点是判别性区域的定位以及背景区域的剔除。文献[16-18]等也突出了判别性区域检测对细粒度图像分类的重要性。

将细粒度图像的分类过程划分为两个步骤：首先提取图像的判别性区域；再对判别性区域进行 CNN 特征表示和学习，从而提高分类正确率。但是，不同于上述方法，本文方法具有以下几个特点：1）不需要目标和部位的标注信息，仅使用类别标签信息，增加模型的通用性；2）采用卷积神经网络模型生成自上而下的注意图来分割目标，增强模型的鲁棒性；3）所提模型能够同时实现图像的显著性检测、目标分割与分类，具有广泛的应用。最后，采用自上而下注意图分割的判别性学习方法在公开的细粒度图像库 Cars196[19]和 Aircraft100[20]中进行实验分析，结果表明，所提方法能够有效提升分类性能。

1　模型概述

具体的模型框架如图 1（a）所示，首先用 CNN 模型对图像进行初分类，得到基本网络模型为 BaseNet；然后使用梯度可视化方法对 BaseNet 进行可视化分析，发现仅有部分图像区域对目标类别有贡献，给定一幅图像和感兴趣的类别，在 BaseNet 上使用 BP（back propagation）算法进行梯度反传，生成该图像针对特定类别的自上而下注意图；再利用生成的注意图初始化 GraphCut[21]算法，对图像进行语义目标分割；最后对分割图像进行 CNN 特征学习得到分割网络模型 SegNet，迫使卷积神经网络集中关注于图像的关键区域，从而提高细粒度图像分类的性能。

卷积神经网络是一个特征学习模型，能够自动学习数据集的样本特征，不需要人为参与和设计，具有很强的鲁棒性，分类性能远远超过 HOG 和 SIFT（scale invariant feature transform）等传统特征。常用的 CNN 模型主要包括 AlexNet[4]，GoogLeNet[22]和 VGG[23]等网络结构，本文选择分类性能最好的 GoogLeNet 作为基本网络，如图 1（b）所示，该网络具有 22 层结构，由于网络层数较深，参数太多，直接使用较小的数据库学习所有参数会导致过拟合和梯度弥散等问题，所以本文使用 ImageNet 对网络进行预训练，再用目标数据库微调得到初始分类模型。其中“2-inception module”表示 2 个连续的 inception[22]结构。

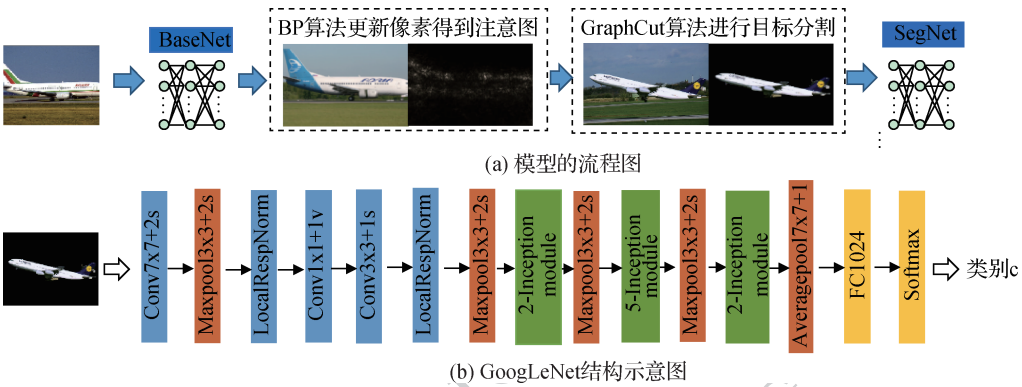


图 1　模型框架与 GoogLeNet 结构示意图

Fig. 1　Overview of proposed model and GoogLeNet structure ((a) flow chart of model; (b) GoogLeNet structure)

2　自上而下注意图的生成

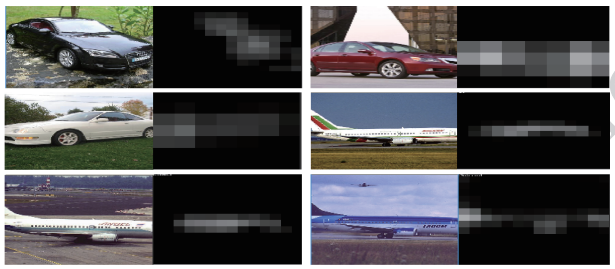
为了研究基本网络模型学到的图像表示，首先对 BaseNet 进行可视化，分析图像各个区域对目标识别的影响，并对网络的特征图响应进行可视化和分析，考察目标神经元学习到的图像模式。假设 BaseNet 的权值为 $\boldsymbol{W}$，图像 $\boldsymbol{I}$ 在类别 c 上的分数为 $S_c(\boldsymbol{I})$，采用有监督的可视化学习方法[24]进行可视化分析，保持权值 $\boldsymbol{W}$ 不变，随机初始化输入图像，在像素空间利用梯度下降法更新像素值，找到一幅图像 $\boldsymbol{I}'$ 使得感兴趣的神经元 c_0 最大化，即

$$\boldsymbol{I}' = \arg\max_{I} S_{c_0}(\boldsymbol{I}) + \lambda \|\boldsymbol{I}\|_2^2 \qquad (1)$$

式中，得到的 $\boldsymbol{I}'$ 代表网络模型 BaseNet 在类别 c_0 上

的一个典型输入模式，使用 BP 算法求解式(1)，最后的可视化结果如图 2 所示，可以发现，网络模型对目标区域的响应较大，对背景区域的响应却很小，所以背景信息基本没有判别性。

(a) 原图与可视化图像



(b) 原图与其在网络高层中响应较大的一个特征图

图 2　基本网络的可视化结果

Fig. 2　Visualization of BaseNet ((a) original images and visualization images; (b) original images and the corresponding feature maps with great response in high layer)

为了得到具有判别性的目标区域，根据 BaseNet 模型计算图像像素对目标类别的贡献度，再生成自上而下的注意图来提取图像中的重要区域。在 BaseNet 中输入图像 $\boldsymbol{I}$，根据 $S_c(\boldsymbol{I})$ 预测 $\boldsymbol{I}$ 的类别为

$$c_0 = \arg\max_{c=0,1,\cdots,L-1} S_c(\boldsymbol{I}) \tag{2}$$

式中，L 表示目标数据库的类别数，对于 CNN 模型，$S_c(\boldsymbol{I})$ 是高度非线性的，所以在图像 $\boldsymbol{I}_0$ 处对 $S_c(\boldsymbol{I})$ 进行一阶的泰勒级数展开得到

$$S_c(\boldsymbol{I}) \approx \boldsymbol{W}^{\mathrm{T}}\boldsymbol{I} + b \tag{3}$$

$$\boldsymbol{W} = \left.\frac{\partial S}{\partial \boldsymbol{I}}\right|_{\boldsymbol{I}_0} \tag{4}$$

式中，b 表示偏差，$\boldsymbol{W}$ 为权重，权重的幅值则表示像素对类别 c 的贡献度，幅值越大表明该像素对类别 c 的支持度越大。

在实验中，首先利用式(2)预测图像 $\boldsymbol{I}_0$ 的 top-1 类别 c_0，再根据式(4)计算权值，将得到的权值 $\boldsymbol{W}$ 根据对应的像素 $\boldsymbol{I}_0$ 进行重新排列，生成图像 $\boldsymbol{I}_0$ 在类别 c_0 上的注意图。结果如图 3 所示，可以发现，图像背景对应的注意值很小，注意值较大的区域对应于图像的目标区域，因此，自上而下注意图给出了图像目标的位置，利用注意图对图像进行前景分割，突出具有判别性的目标区域，删除背景区域对分类性能的干扰，可以进一步提高细粒度图像的分类性能。

图 3　原图与其在 top-1 预测类别上的注意图

Fig. 3　Original images and image-specific class attention maps for the top-1 predicted class in BaseNet

3　基于注意图的目标分割

细粒度图像分类的难点在于各类别之间差异较小，这些差异容易被复杂的背景信息覆盖，使得网络模型难以学到真正的差异性特征。为了解决细粒度图像分类中的背景干扰问题，增加图像目标的判别性，迫使模型学习类别之间真正的差异性特征，利用上文得到的自上而下注意图对图像目标进行语义分割。注意图给出了图像中的关键区域，图像像素对应的注意值则表明该像素对目标类别的空间支持度，注意值越大表明该像素对相应类别的贡献越大。因此，在注意图上选择合适的阈值，进行合理的分割初始化，可以实现图像的语义分割。

假设图像 $\boldsymbol{I} = (I_1, \cdots, I_n, \cdots, I_N) \in \mathbf{R}^{w\times h\times 3}$ 的注意图为 $\boldsymbol{S} = (S_1, \cdots, S_n, \cdots, S_N) \in \mathbf{R}^{w\times h\times 3}$，$w$、$h$ 为图像的宽高，选择合适的阈值 T_1、T_2，对 GraphCut 初始化为

$$I_n = \begin{cases} \text{前景} & S_n > T_1 \\ \text{不确定} & \text{其他} \\ \text{背景} & S_n < T_2 \end{cases} \tag{5}$$

式中，T_1 为 $\boldsymbol{S}$ 的均值，T_2 为均值的 0.4 倍。分别对前景和背景区域建立混合高斯模型，利用 GraphCut 算法进行目标分割，结果如图 4 所示，可以发现，利用自上而下注意图能够很好地分割出图像的目标。

图 4　分割结果

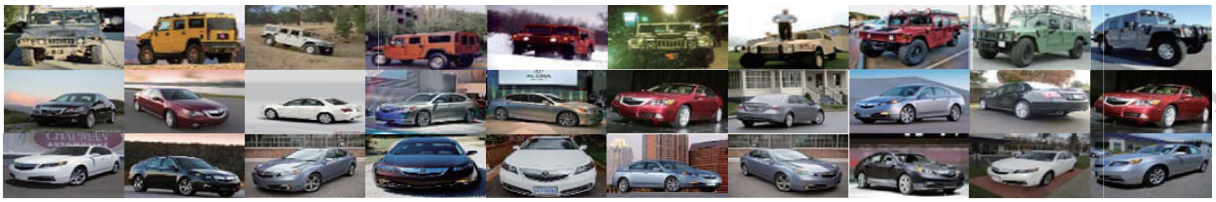
Fig. 4　Results of object segmentation

4　实验与结果

4.1　数据库

在两个较新的细粒度图像库上进行分类研究，分别为 Cars196 [19] 和 Aircrafts100 [20] 数据库。其中 Cars196 图像库包含 196 类汽车，每类包含 48 ～136 幅图像，总共 16 185 幅图像。该数据库按照汽车的制造、模型和年代进行分类，每类包含 24 ～68 幅训练图像，剩下的图像作为测试集。图 5(a)列出 3 种汽车类别的若干样本，可以发现，该数据库的图像背景复杂；同一类别内，目标的颜色、光照变化大；不同类别之间，目标的形状、颜色非常相似，所以该数据库适用于细粒度图像分类研究。

Aircrafts100 图像库则是由 Maji[20] 提出的较新的细粒度图像库，包含 100 种飞机模型，每类包含 100 幅图像，总共 10 000 幅图像，其中，每类训练集、验证集和测试集分别为 33、34 幅图像。作为 FG-Comp2013 挑战赛的一部分，该数据库图像的分辨率比较大。如图 5(b)所示，数据库类内的差距大，类别之间的差距很小，例如类别 Boeing737-300 与 Boeing737-400 在形状、颜色等方面非常相似。



(a) 每一行分别为汽车类别AMM General Hummmer SUV 2000,Acura RL Seddan 2012, Acuraa TL Sedan 20112

(b) 每一行分别为飞机类别Booeing737-300,BBoeing737-4000,Boeing737-700

图 5　目标数据库的图像示例

Fig. 5　Images from target databases ((a) each row is a car class named AM General Hummer SUV 2000, Acura RL Sedan 2012 or Acura TL Sedan 2012; (b) each row is an aircraft class named Boeing737-300, Boeing737-400 or Boeing737-700)

4.2　实验结果

4.2.1　分类结果

首先，用目标数据库对 GoogLeNet 进行微调得到基本分类模型 BaseNet。在实验中，将图像缩放到 256 ×256 ×3 大小，减去图像均值，再裁剪出 224 × 224 ×3 大小的输入图像，并利用图像翻转来扩增训练数据。在微调网络之前，把分类器层的 1 000 个神经元替换为目标数据库的类别数，对于 Cars196

数据库替换为196个神经元,Aircrafts100则替换为100,对最后一层的参数进行随机初始化,将对应的学习速率扩大10倍。由于使用较小的目标数据库对预训练网络进行微调,所以整个网络的初始学习速率较小,设置为0.001,使用随机梯度下降法训练网络,当验证集的损失不再下降时将学习速率下降10倍。

得到基本网络模型之后,首先使用BaseNet对图像库进行初分类,结果如表1所示,196类汽车测试集的平均分类正确率为85.50%,100类飞机测试集的平均分类正确率为83.01%,分别比传统FV-SIFT特征高出26.3%和22.01%,比VGG网络模型高出5.7%和8.91%;然后根据BaseNet生成图像针对top-1预测类别的注意图;再利用注意图进行前景分割;最后,针对分割图像,采用以下两种方式进行细粒度图像分类研究,并在表1给出最终结果。

表1 Cars196和Aircrafts100数据库的分类结果

Table 1 Classification results on Cars196 and Aircrafts100 database

方法	Cars196分类结果/%	Aircrafts100分类结果/%
FV-SIFT[13]	59.2	61.0
SIFT ensemble[25]	82.7	80.7
symbiotic segmentation(BBox)[23]	78.0	72.5
VGG[23]	79.8	74.1
FV-CNN[26]	85.7	80.1
B-CNN[23]	83.9	84.1
BaseNet	85.50	83.01
BaseNet + SegNet	86.74	83.43
分部训练模型	86.70	84.70

1)利用分割图像微调BaseNet得到分割网络SegNet,再综合BaseNet和SegNet得到多模型结果。在SegNet中,196类汽车测试集的平均分类正确率为86.06%,100类飞机测试集的平均分类正确率则为82.83%。将两个网络的分数相加得到综合模型,最后汽车和飞机数据库的平均分类正确率分别为86.74%和83.43%,汽车图像库的正确率在GoogLeNet模型基础上提高了1.24%,飞机图像库的正确率也有提高。

2)采用分部训练的方法来提高分类正确率。分析第1种方法,可以发现,由于少部分样本的分割效果较差,丢失部分目标区域的信息,如图4最后一列所示,导致SegNet在汽车测试集上的正确率仅提高0.5%,在飞机测试集上却没有提高,所以分割网络的分类性能没有得到显著提升,但是分割图像给出了图像的目标区域,所以使用分割图像学习得到的分割网络主要关注于图像的关键区域,从而得到较优的初始化权值,因此,本文将SegNet作为初始化模型,利用原始图像进行再次微调得到分部训练模型。由于分割网络已经充分挖掘分割图像的判别性信息,所以再次微调时,初始学习速率较小,设置为0.000 1,保证分部训练网络的权值在一定范围内变化,获得稳定的性能,并使得分部训练模型在挖掘关键目标区域信息的同时,又能补充分割网络缺失的互补信息,从而提高分类正确率,最后196类汽车在分部训练模型上的测试结果为86.70%,100类飞机的测试结果为84.70%,分别比GoogLeNet基本模型提高了1.2%和1.69%。

表1中,FV-SIFT[13]字典大小为256,空间金字塔只有1层。而FV-SIFT[25]使用的是多尺度SIFT特征,字典大小为1 024,空间金字塔有2层,为1×1,3×1区域,结果表明,使用传统FV特征的分类正确率远远低于本文模型。FV-CNN[26]方法对CNN特征进行FV编码,这个过程需要通过聚类来学习码本,并且FV编码后的向量通常不是稀疏的,所以该方法对内存的要求很高,而且分类正确率也不高。BCNN[23]模型需要同时学习两个深度网络,参数个数扩大两倍,所以训练过程缓慢,BCNN给出的结果是原始图像和翻转图像的平均结果,使得测试时间加倍。Symbiotic Segmentation[23]方法要求知道目标的矩形框标注信息,并且分类正确率比本文模型低8个以上百分点。可以发现,本文方法仅使用图像的类别标签,分类性能也优于其他优秀的算法。

4.2.2 时间复杂度分析

将本文方法分别与传统模型、其他基于CNN的模型进行测试时间复杂度的分析和运行时间比较,发现本文方法降低了时间复杂度。

传统特征在线性SVM分类器上获得良好的分类性能[13,23,25]。线性SVM的测试时间复杂度为$O(dn)$,d为特征维数,n为支持向量个数。假设底层特征为D维,FV编码中混合高斯模型包含K_g个高斯分布,空间金字塔包含P个子区域,考虑均值和标准差的梯度,最后FV特征包含$2K_gDP$维。假

设底层特征为 128 维的 SIFT，那么文献[13,25]中每张图像的特征维数分别为 65 536 和 1 048 576 维，symbiotic segmentation[23]中的特征为 40 992 维，文献[26]用 PCA 对 SIFT 降维，最后将每幅图像表示为 256 000 维的向量。而本文模型中，每张图像的特征长度为 1 024 维，远远低于传统特征的维数。基于 CNN 的分类模型中，VGG 网络包含 138 M 参数，BCNN 模型包含两个网络结构，选择 VGG 结构时包含 276 M 参数，选择 GoogLeNet 则包含 10 M 参数，而本文模型仅包含 5 M 参数，远远少于其他模型。在 Tesla K40 GPU 中对图像进行特征提取时间测试，得到的结果如表 2 所示，可以发现，本文模型的运行时间最少。

表 2　特征提取时间比较

Table 2　Feature extraction speeds

方法	时间/(帧/s)
FV-SIFT[13]	10
FV-VGG[26]	8
FV-M-Net[26]	23
VGG[23]	43
BCNN[23]	10
RCNN[7]	1/47
本文	59

5　结　论

图像分类是机器视觉领域的一项基本任务，针对细粒度图像分类中的背景干扰问题，提出一种基于自上而下注意图分割的分类模型，利用卷积神经网络对图像进行分类，根据分类结果生成自上而下注意图来识别和分割目标区域，再对目标区域进行表示与识别，提高分类性能。在公开的细粒度图像库 Cars196 和 Aircrafts100 上进行分类研究，结果表明，在 CNN 模型基础上引入注意信息能够进一步提高细粒度图像分类的正确率。相对于传统特征，使用 CNN 特征增强了模型的鲁棒性；整个模型仅使用图像的类别标签信息，使其适用于其他的图像库，增加了模型的通用性；同时，也通过实验验证了图像中的大部分内容对目标识别没有贡献，判别性信息存在于关键的部分区域中，与人脑在视觉上的认知过程相符合；此外，本文模型能够同时实现图像的显著性检测、目标分割和分类，具有广阔的应用前景。

参考文献(References)

[1] Wah C, Branson S, Welinder P, et al. The Caltech-UCSD birds-200-2011 dataset: Computation & Neural Systems Technical Report, CNS-TR-2011-001[R]. San Diego, USA: California Institute of Technology, 2011.

[2] Nilsback M E, Zisserman A. Automated flower classification over a large number of classes[C]//Proceedings of the 6th Indian Conference on Computer Vision, Graphics & Image Processing. Bhubaneswar, India: IEEE, 2008: 722-729. [DOI: 10.1109/ICVGIP.2008.47]

[3] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009: 248-255. [DOI: 10.1109/CVPR.2009.5206848]

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//Proceedings of Advances in Neural Information Processing Systems 25, NIPS. Lake Tahoe, Nevada, USA: MIT press, 2012, 25(2): 1097-1105.

[5] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//Proceedings of the 13th European Conference on Computer Vision-ECCV 2014. Zurich, Switzerland: Springer International Publishing, 2014: 818-833. [DOI: 10.1007/978-3-319-10590-1_53]

[6] Russakovsky O, Deng J, Huang Z H, et al. Detecting avocados to zucchinis: what have we done, and where are we going?[C]//Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, NSW: IEEE, 2013: 2064-2071. [DOI: 10.1109/ICCV.2013.258]

[7] Zhang N, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection[C]//Proceedings of the 13th European Conference on Computer Vision-ECCV 2014. Zurich, Switzerland: Springer International Publishing, 2014: 834-849. [DOI: 10.1007/978-3-319-10590-1_54]

[8] Farrell R, Oza O, Zhang N, et al. Birdlets: subordinate categorization using volumetric primitives and pose-normalized appearance[C]//Proceedings of the 2011 International Conference on Computer Vision. Barcelona, Spain: IEEE, 2011: 161-168. [DOI: 10.1109/ICCV.2011.6126238]

[9] Göering C, Rodner E, Freytag A, et al. Nonparametric part transfer for fine-grained recognition[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014: 2489-2496. [DOI: 10.1109/CVPR.2014.319]

[10] Zhang N, Farrell R, Iandola F, et al. Deformable part descriptors for fine-grained recognition and attribute prediction[C]//Proceedings of the 2013 IEEE International Conference on Computer

Vision. Sydney, NSW, Australia: IEEE, 2013: 729-736. [DOI: 10.1109/ICCV.2013.96]

[11] Zhang N, Paluri M, Ranzato M A, et al. PANDA: pose aligned networks for deep attribute modeling[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014: 1637-1644. [DOI: 10.1109/CVPR.2014.212]

[12] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645. [DOI: 10.1109/TPAMI.2009.167]

[13] Bourdev L, Malik J. Poselets: body part detectors trained using 3D human pose annotations[C]//Proceedings of the 12th International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009: 1365-1372. [DOI: 10.1109/ICCV.2009.5459303]

[14] Angelova A, Zhu S H. Efficient object detection and segmentation for fine-grained recognition[C]//Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA: IEEE, 2013: 811-818. [DOI: 10.1109/CVPR.2013.110]

[15] Chai Y N, Lempitsky V, Zisserman A. Symbiotic segmentation and part localization for fine-grained categorization[C]//Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, NSW, Australia: IEEE, 2013: 321-328. [DOI: 10.1109/ICCV.2013.47]

[16] Krause J, Jin H L, Yang J C, et al. Fine-grained recognition without part annotations[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015: 5546-5555. [DOI: 10.1109/CVPR.2015.7299194]

[17] Yang S L, Bo L F, Wang J, et al. Unsupervised template learning for fine-grained object recognition[C]//Proceedings of Advances in Neural Information Processing Systems 25. NIPS. Lake Tahoe, Nevada, USA: MIT press, 2012: 3122-3130.

[18] Zhang N, Farrell R, Darrell T. Pose pooling kernels for sub-category recognition[C]//Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012: 3665-3672. [DOI: 10.1109/CVPR.2012.6248364]

[19] Krause J, Stark M, Deng J, et al. 3D object representations for fine-grained categorization[C]//Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops. Sydney, NSW, Australia: IEEE, 2013: 554-561. [DOI: 10.1109/ICCVW.2013.77]

[20] Maji S, Rahtu E, Kannala J, et al. Fine-grained visual classification of aircraft[J]. CoRR, 2013, abs/1306.5151.

[21] Boykov Y Y, Jolly M P. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images[C]//Proceedings of the 8th IEEE International Conference on Computer Vision. Vancouver, BC, Canada: IEEE, 2001, 1: 105-112. [DOI: 10.1109/ICCV.2001.937505]

[22] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015: 1-9. [DOI: 10.1109/CVPR.2015.7298594]

[23] Lin T Y, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 1449-1457. [DOI: 10.1109/ICCV.2015.170]

[24] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps[J]. CoRR, 2013, abs/1312.6034.

[25] Gosselin P H, Murray N, Jégou H, et al. Revisiting the fisher vector for fine-grained classification[J]. Pattern Recognition Letters, 2014, 49: 92-98. [DOI: 10.1016/j.patrec.2014.06.011]

[26] Cimpoi M, Maji S, Vedaldi A. Deep filter banks for texture recognition and segmentation[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015: 3828-3836. [DOI: 10.1109/CVPR.2015.7299007]