# 联合语义部件的鸟类图像细粒度识别

赵毅力[1,2], 徐　丹[1)*]

[1)] (云南大学信息学院　昆明　650091)
[2)] (西南林业大学大数据与智能工程学院　昆明　650224)
(danxu@ynu.edu.cn)

**摘　要:** 由于子类别的高度相似性引起的类间微小差异, 以及姿态、尺度和旋转方面的类内变化, 使得细粒度图像识别成为一个具有挑战性的计算机视觉问题. 为了对鸟类图像进行细粒度识别, 提出一种联合语义部件的深度卷积神经网络模型. 该模型由 2 个子网络组成: 一个是语义部件检测子网, 使用深度残差网络对鸟类图像语义部件进行精确定位; 另一个是分类子网, 使用三路深度残差网络对检测子网检测到的语义部件进行联合分类. 收集了一个新的鸟类图像数据集 YUB-200-2017, 用于鸟类图像细粒度识别实验. 结果表明, 在 YUB-200-2017 和 CUB-200-2011 数据集上, 文中方法具有较高的语义部件检测精度和识别准确率.

**关键词:** 细粒度图像识别; 语义部件检测; 深度学习; 卷积神经网络
**中图法分类号:** TP391.41　　**DOI:** 10.3724/SP.J.1089.2018.16781

# Joint Semantic Parts for Fine-Grained Bird Images Recognition

Zhao Yili[1,2] and Xu Dan[1)*]

[1)] (*School of Information, Yunnan University*, *Kunming*　650091)
[2)] (*School of Big Data and Intelligent Engineering, Southwest Forestry University*, *Kunming*　650224)

**Abstract:** Fine-grained image recognition is a challenging computer vision problem, due to small inter-class variations caused by highly similar subordinate categories, and the large intra-class variations in poses, scales and rotations. In order to perform fine-grained recognition on bird images, this paper proposes a deep convolution neural networks model collaborated with semantic parts detection. The model consists of two modules, one module is a parts detector network, and another module is a three-stream classification network based on deep residual network. In the meantime, a new bird images dataset was collected and labeled to facility the research of fine-grained bird images recognition. Experiment results on YUB-200-2017 and CUB-200-2011 illustrate the proposed model has higher part detection and image classification accuracy comparing with state-of-the-arts fine-grained bird image recognition approaches.

**Key words:** fine-grained image recognition; semantic parts detection; deep learning; convolutional neural networks

## 1　相关工作

细粒度识别任务, 如不同动物或植物的识别, 逐渐成为计算机视觉中的一个重要的研究领域. 因为同属于某一个大类别中的子类具有较高的相似度, 所以需要通过子类别之间存在的细小差异来进行区分, 这就使得细粒度识别成为一个比

收稿日期: 2017-07-10; 修回日期: 2018-02-23. 基金项目: 国家自然科学基金(61662072, 61540062); 云南省教育厅基金(2015Y285, 2016CYH03); 云南省应用基础研究项目(2014FA021). 赵毅力(1978—), 男, 博士研究生, 副教授, CCF 会员, 主要研究方向为计算机视觉、深度学习; 徐　丹(1968—), 女, 博士, 教授, 博士生导师, 论文通讯作者, 主要研究方向为计算机图形学.

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

粗粒度识别更具有挑战性的问题. 为了正确地对细粒度图像进行识别，通常需要观察者具备一定的领域知识. 如图 1 所示，对于同属于鹟科姬鹟属的鸟类黄眉姬鹟和白眉姬鹟，两者之间用于辨识的特征主要在于眉纹颜色的不同，而这种差异是比较细微的.

a. 黄眉姬鹟　　b. 白眉姬鹟

图 1　黄眉姬鹟和白眉姬鹟的识别

由于子类对象之间的差异主要体现在局部细节上，因此如何有效地对识别目标进行检测，并从中发现重要的局部区域信息就成为细粒度图像识别算法的关键. 当前，对于细粒度图像识别方面的研究，可以按照其使用监督信息的多少，分为基于强监督学习的分类模型和基于弱监督学习的分类模型两大类.

### 1.1　基于强监督学习的细粒度图像分类模型

Zhang 等[1]提出的 Part R-CNN 模型首先利用选择性搜索算法[2]在细粒度图像中产生物体可能出现的候选框；然后使用 R-CNN 算法[3]根据训练图像提供的目标和部件包围盒标注信息进行训练，得到 3 个部件检测模型分别用于对鸟类的整体、头部和躯干部位进行检测，并使用几何约束对检测到的包围盒进行限制；接着将检测得到的图像块作为输入，分别对预训练的卷积神经网络(convolutional neural networks, CNN)进行微调；最后使用支持向量机(support vector machine, SVM)对提取到的特征进行分类.

Branson 等[4]提出的姿态规范化细粒度识别模型首先使用可变形部件模型(deformable part model, DPM)通过语义部件的特征点来计算物体级别和部件级别的包围盒，并对语义部位图像块进行姿态对齐；然后针对细粒度图像不同级别的图像块分别提取 CNN 中不同层的卷积特征；最后将不同级别特征级联作为整幅图像的表示，再使用 SVM 进行分类.

Lin 等[5]提出的 Deep LAC 模型包含语义部件定位、对齐和分类 3 个组件. 其中，语义部件定位是通过直接对包围盒进行回归完成的；然后使用模板匹配对语义部件进行对齐，并通过值链接函数将定位、对齐和分类连接起来进行端到端训练.

Zhang 等[6]提出的 SPDA-CNN 模型由卷积共享的语义部件检测子网和部件抽象与分类子网组成. 为了能够对更多的小尺度语义部件进行检测，SPDA-CNN 使用基于 $k$ 最近邻的候选框建议生成方法生成多个语义部件检测的候选框. 给定基于 $k$ 最近邻生成的部件候选框，SPDA-CNN 的检测网络使用 Fast R-CNN[7]来生成最终的语义部件包围盒；同时，SPDA-CNN 的部件抽象和分类子网在传统的 CNN 架构上引入语义部件池化层、基于部件的全连接层和特征聚合全连接层，使其能够进行端到端的训练.

Huang 等[8]提出的 PS-CNN 模型使用全卷积网络(fully convolutional networks, FCN)来对图像中的语义部件关键点进行定位，然后使用两路网络来分别对目标的整体和语义部件进行处理，最后使用 3 个全连接层组成的子网络作为特征分类器. PS-CNN 更注重模型的可解释性：(1) 对于某个特定的类别，把这个类别和其他类别区分开来最具判别性的部件是哪一个；(2) 对于高度相似的类别，把这些类别区分开来最具判别性的部件是哪一个.

Wei 等[9]提出的 Mask-CNN 模型亦分为 2 个模块：一是部件定位模块，二是全局和局部图像块的特征学习模块. Mask-CNN 使用 FCN 学习语义部件的分割模型. 在得到语义部件的蒙版图像后，可以通过图像裁剪操作来获得对应的小图像块，并将 2 个语义部件的蒙版图像组合起来构成完整的目标整体蒙版图像；然后基于物体和部件图像块分别训练 3 个子网络用于分类，在每个子网络中，使用语义部件的蒙版图像对 CNN 的关键卷积特征进行筛选；最后对保留下来的特征进行全局平均和最大池化操作，并将 3 个子网特征再次级联作为整幅图像的特征表示.

可以看出，基于强监督学习的分类模型注重利用待识别目标的语义部件特征来进行细粒度分类，本文方法也属于这一类别.

### 1.2　基于弱监督学习的细粒度图像分类模型

基于强监督学习的分类模型能够取得较满意的分类精度，但是在训练的过程中需要提供语义部件的标注信息，如部件的包围盒坐标或关键点坐标. 与之相比，基于弱监督学习的细粒度图像分类模型在模型训练时仅使用图像级的标注信息，而不再使用语义部件标注信息.

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Lin 等[10]提出的 Bilinear-CNN 模型是一个四元组 $B=(f_A,f_B,P,C)$；其中，$f_A$ 和 $f_B$ 是特征提取函数，$P$ 是池化函数，$C$ 是分类函数. 特征提取函数是函数映射 $f:L\times I\to\mathbb{R}^{c\times D}$，即将输入图像 $I$ 和位置区域 $L$ 映射为一个 $c\times D$ 维的特征向量. 这 2 个特征提取函数的输出可以通过双线性操作进行聚合得到最终的特征描述；最后使用 SVM 对双线性特征进行分类.

Xiao 等[11]提出的细粒度分类模型主要关注识别目标物体级别和部件级别 2 个层次的特征，并通过注意力机制和视觉焦点来完成物体和局部区域的检测. 该模型分为 3 个阶段:

Step1. 从输入图像中产生大量的候选区域，然后对这些区域进行过滤并保留包含前景物体的候选区域.

Step2. 训练一个 CNN 实现对物体级图像进行分类.

Step3. 根据物体级模型训练的网络来对候选区域提取特征，并对这些特征进行谱聚类，得到 $k$ 个不同的聚类簇，每一簇可视为代表一类局部信息，从而达到对测试样本局部区域检测的目的.

Simon 等[12]提出的星座细粒度分类模型是对 CNN 提取到的卷积特征进行可视化，发现一些响应比较强烈的区域恰好对应于输入图像中一些潜在的局部区域关键点；然后根据这些关键点来提取局部区域信息. 由于特征输出的分辨率与输入图像相差较大，使得对输入图像中的区域进行精确定位比较困难，因此需要通过计算梯度图来产生区域位置. 具体而言，卷积特征的输出是一个 $W\times H\times P$ 维的张量，其中，$P$ 表示通道数目，每一个通道可以表示成一个 $W\times H$ 维的矩阵. 通过计算每一个通道对每一个输入像素的平均梯度值，可以得到与原输入图像大小相同的特征梯度图. 在特征梯度图中响应比较强烈的区域即代表输入图像中的一个局部区域，而每一张梯度图中响应最强烈的位置即作为原图中的关键点. 卷积层的输出共有 $P$ 维通道，分别对应于 $P$ 个关键点位置，通过排序选择出最重要的前 $M$ 个. 最后对关键点提取特征以完成细粒度识别.

Zhang 等[13]提出的 SWFV-CNN 模型通过对 CNN 中的滤波器进行挑选以构建复杂特征. SWFV-CNN 模型分为 2 个步骤:

Step1. 利用 CNN 的选择性来挖掘对于某些模式敏感的滤波器，从而得到弱监督的语义部件检测器，进而将该部件检测器作为初始值来训练一个更具判别性的部件检测模型.

Step2. 对 Step1 检测到的语义部件图像块提取深度卷积特征，并利用空间加权 Fisher 向量对特征进行池化，得到最后的特征描述.

与基于强监督学习的细粒度识别模型相比，目前最好的基于弱监督学习的细粒度识别模型在分类准确率方面仍然存在一定的差距. 但是由于基于弱监督学习的细粒度识别模型不需要部件的标注信息，所以在网络训练上具有独特的优势.

为了能够对不同的鸟类图像进行细粒度识别，本文提出一种联合语义部件的鸟类图细粒度识别模型. 本文还收集了一个高质量的细粒度鸟类图像数据集 YUB-200-2017 用于对该模型进行验证.

## 2 本文方法

本文提出的鸟类图细粒度识别模型由语义部件检测网络和分类网络 2 个模块组成. 通过将部件检测视为 3 个类别的目标检测任务，语义部件检测网络利用基于深度残差网络(deep residual network, ResNet)的目标检测算法对鸟类图像的语义部件进行检测. 基于检测得到的语义部件信息构建一个三路基于 ResNet 的分类模型同时对对象级和部件级的特征进行聚合和分类. 该模型架构如图 2 所示. 在分类模块的每一路网络中，使用 ResNet 作为

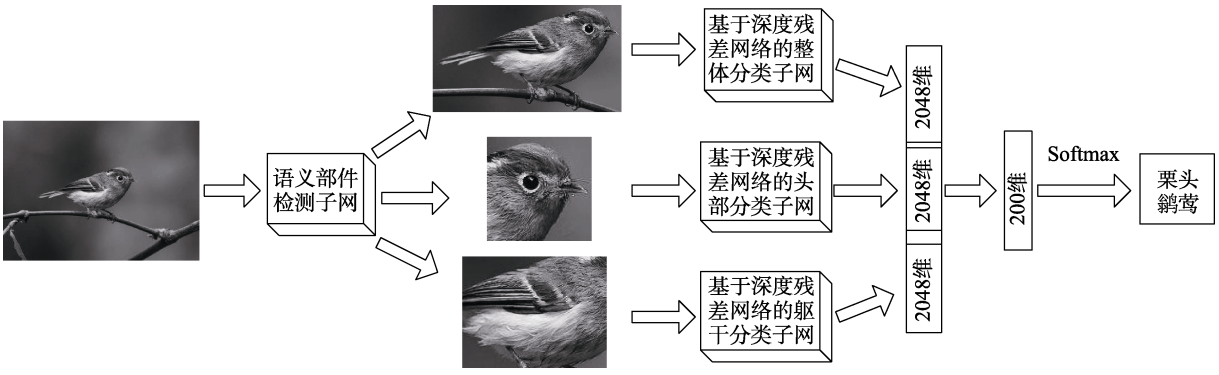


图 2 本文模型架构

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

CNN架构，并对选取的特征描述符通过全局平均池化输出为 2048 维的特征向量；然后将这三路的特征向量连接为一个 6144 维的特征向量；最后通过添加一个 200 维的分类层用于端到端的联合训练.

### 2.1　细粒度鸟类图像数据集

相对于普通分类任务的数据集，细粒度图像数据集的获取难度更大，需要一定的专业领域知识才能完成数据的采集与标注. 目前，在细粒度图像识别研究中被广泛使用的鸟类图像数据集是 CUB-200-2011，该数据集包含 200 种不同类别的北美鸟类，共 11 788 幅鸟类图像；其中，训练集包括 5 594 幅图像，测试集包括 5 794 幅图像；该数据集为每幅图像提供了 15 个部件关键点坐标以及 1 个整体包围盒坐标.

中国有着丰富的野生鸟类资源，尤其是云南省被称为“植物王国”和“动物王国”. 目前已知的野生鸟类有 1 300 多种，仅云南省就有 800 多种，占全国野生鸟类总量的 60%. 本文作者从云南省野生鸟类中收集了 200 种鸟类，每类 60 幅高质量的图像，总共 12 000 幅图像，用于鸟类图像的细粒度分类与识别研究，这个数据集被命名为 YUB-200-2017；该数据集包含的鸟类主要以鸟纲中雀形目的鸟类为主，这也是鸟纲中最丰富的一个类别，另外，部分鸟类雄性和雌性具有不同的识别特征. CUB-200-2011 只包含雄性鸟类的样本，而 YUB-200-2017 同时对雄性和雌性的鸟类样本进行了收集.

为了和 CUB-200-2011 进行比较，本文对 YUB-200-2017 进行相同的划分，即从每个类别中随机选择 30 幅图像作为训练图像，剩下的 30 幅图像作为测试图像. 因此，YUB-200-2017 的训练集和测试集各包含 6 000 幅鸟类图像，其中每一幅图像都提供了鸟类的整体、头部和躯干 3 个部件的包围盒标注信息，以及对应的鸟类类别标签. 从 YUB-200-2017 中的 200 种鸟类中随机选取一幅图像，合并后的图像如图 3 所示.

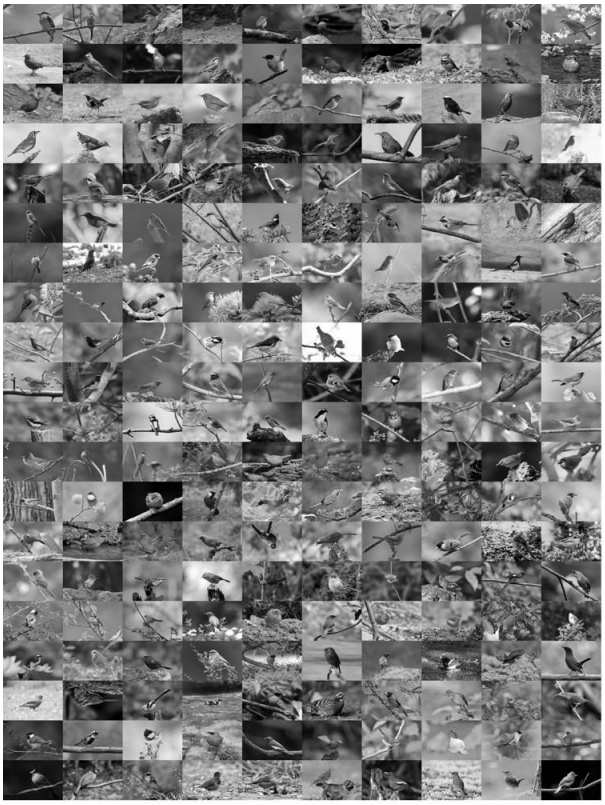

图 3　YUB-200-2017 细粒度鸟类图像数据集

### 2.2　CNN 基准架构选择

不同的 CNN 架构具有不同的分类准确率，因此选择一个合适的 CNN 架构对于鸟类图像语义部件检测和细粒度识别非常重要. 相对于 ImageNet 数据集，细粒度鸟类图像数据集 YUB-200-2017 和 CUB-200-2011 都属于小型数据集. 由于其样本数量过少，因此并不足以对一个 CNN 进行充分训练，容易导致模型对数据集的过拟合. 因为基于 ImageNet 预训练的模型具有较好的泛化性，所以不同的计算机视觉任务往往使用预训练模型对数据集进行微调. 具体地说，微调的方法是将原来模型最后的分类层替换为和新数据集类别相关的分类层，然后在新的数据集上对模型进行重新训练. 通过模型微调，既可以在较短的时间内完成模型训练，又可以避免过拟合. 本文对目前主流的 CNN 架构 AlexNet[14]，VGGNet[15]，Inception V3[16] 以 ResNet[17]在 YUB-200-2017 和 CUB-200-2011 上进行微调，其分类准确率如表 1 所示. 可以看出，对于图像分类任务，ResNet 有较高的分类准确率. 另外，使用相同的 CNN 架构在 YUB-200-2017 上的分类准确率比在 CUB-200-2011 上的准确率平均高 10%左右.

表 2 所示为不同架构的 CNN 的加载时间和单幅图像的推导时间. 可以看出，AlexNet 由于层数较少，因此具有较低的加载时间和推导时间. 相对于 VGGNet, ResNet 的加载时间和推导时间都要更少. 综合模型的分类准确率和运行时间，本文选择 ResNet-50 作为鸟类图像语义部件检测和细粒度图像分类的基准架构.

### 2.3　鸟类图像判别性语义部件选取

由于不同鸟类之间的差异主要在局部部件区域的细节，因此对识别目标关键部件的选取对于

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 1 不同架构的 CNN 分类准确率比较 %

| CNN 架构 | CUB-200-2011 | YUB-200-2017 |
|---|---|---|
| AlexNet | 55.4 | 76.7 |
| VGG-11 | 74.8 | 86.3 |
| VGG-13 | 75.1 | 87.8 |
| VGG-16 | 76.2 | 88.5 |
| VGG-19 | 75.6 | 89.1 |
| Inception V3 | 77.6 | 90.6 |
| ResNet-18 | 77.3 | 90.1 |
| ResNet-34 | 78.1 | 90.8 |
| ResNet-50 | 79.1 | 91.4 |
| ResNet-101 | 81.3 | 92.3 |
| ResNet-152 | 82.4 | 92.8 |

表 2 不同架构 CNN 的运行时间比较 s

| CNN 架构 | 模型加载时间 | 模型推导时间 |
|---|---|---|
| AlexNet | 0.63 | 0.03 |
| VGG-11 | 7.25 | 0.20 |
| VGG-13 | 7.28 | 0.29 |
| VGG-16 | 7.56 | 0.38 |
| VGG-19 | 7.86 | 0.46 |
| Inception V3 | 0.68 | 0.48 |
| ResNet-18 | 0.66 | 0.07 |
| ResNet-34 | 1.25 | 0.12 |
| ResNet-50 | 1.41 | 0.15 |
| ResNet-101 | 2.54 | 0.27 |
| ResNet-152 | 3.46 | 0.38 |

细粒度识别来说尤为重要. 虽然 CNN 在对鸟类图像进行分类时, 通过卷积层提取到的特征在经过全连接层后就丢失了位置信息. 但是 CNN 内部的注意力机制能够隐式的对鸟类图像判别性区域进行选取. Zhou 等[18]提出一种基于弱监督学习的判别性区域定位方法, 能够在只使用图像类别标签的前提下对图像的判别性区域进行定位. 对于一幅给定的图像, 令 $f_k(x,y)$ 表示CNN最后一个卷积层在空间位置 $(x,y)$ 的第 $k$ 个单元的激活值. 对于单元 $k$, 全局平均池化层的输出 $F_k=\sum_{x,y} f_k(x,y)$. 对于给定的类别 $c$, softmax 分类器的输入 $S_c=\sum_k w_k^c F_k$, 其中, $w_k^c$ 是对应于单元 $k$ 的类别 $c$ 的权重. 可以把 $w_k^c$ 视为 $F_k$ 对于类别 $c$ 的重要性权重系数. 因此, 类别 $c$ 的 softmax 分类器的输入

$$S_c=\sum_k w_k^c \sum_{x,y} f_k(x,y)=\sum_{x,y}\sum_k w_k^c f_k(x,y) \tag{1}$$

定义类别 $c$ 的类激活映射图

$$M_c=\sum_k w_k^c f_k(x,y) \tag{2}$$

因此有

$$S_c=\sum_{x,y} M_c(x,y) \tag{3}$$

即 $M_c(x,y)$ 表示空间网格 $(x,y)$ 位置处的激活值对于 CNN 将图像分为 $c$ 类的重要性.

为了生成不同类别的类激活映射图, 对于 AlexNet 和 VGGNet, 需要删除网络最后输出层之前的全连接层, 并用全局平均池化层替代. 但是, 去除全连接层会对网络的分类性能有一定的影响. 第 2.2 节通过对不同 CNN 在数据集 YUB-200-2017 和 CUB-200-2011 上的分类准确率进行比较, 选择 ResNet-50 作为鸟类图像细粒度识别的基准架构, 而 ResNet 本身就使用了全局平均池化层, 因此可以在不影响分类精度的前提下生成不同类别的类激活映射图. 基于 ResNet, 本文使用类激活映射图的方法对 YUB-200-2017 中的 200 个类别的鸟类图像判别性区域进行可视化, 图 4~图 6 所示为其中 3 个类别及其对应的判别性区域热度图. 可以看出, 当使用 CNN 在对鸟类图像进行识别时, 最具判别性的语义部件是鸟类的头部和躯干 2 个部位. 因此本文使用这 2 个语义部件再加上鸟类物体本身作为鸟类图像细粒度识别的语义部件.

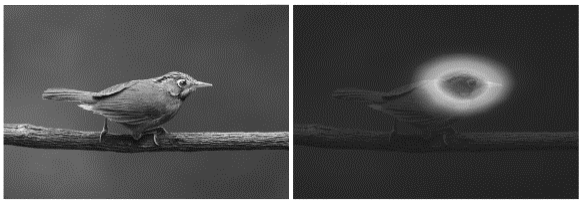

图 4 黑头穗鹛及其判别性区域热度图

图 5 蓝眉林鸲及其判别性区域热度图

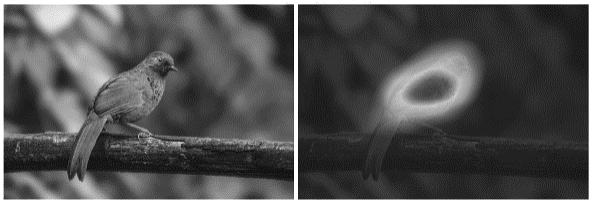

图 6 红头噪鹛及其判别性区域热度图

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

### 2.4　鸟类图像判别性语义部件检测

由于 YUB-200-2017 中的每一幅图像都提供鸟类物体、头部和躯干 3 个语义部件的包围盒标注信息，因此可以将鸟类图像语义部件检测视为目标检测任务. CUB-200-2011 并没有直接提供语义部件的包围盒标注信息，而是提供了不同语义部件关键点的标注信息，基于这些语义部件关键点的标注信息可以间接得到包围盒的标注信息.

目前，基于深度学习的目标检测主要分为 2 大类：一类是基于候选区域的深度学习目标检测算法，其包含候选区域生成和候选区域分类 2 个阶段，如 R-CNN, Fast R-CNN, Faster R-CNN[19] 和 R-FCN[20]；另一类是基于回归的深度学习目标检测算法，其不包括候选区域生成阶段，而是直接生成目标的包围盒，如 YOLO[21]和 SSD[22]. 为了选择一个合适的目标检测算法作为鸟类图像判别性语义部件检测算法，本文使用基于 IoU(intersection over union)的 PCP(percentage of correctly estimated body parts)指标，对目前主流的基于深度学习的目标检测算法在 YUB-200-2017 和 CUB-200-2011 上进行定量评测，结果如表 3~表 10 所示. 可以看出，当 IoU=0.5 时，4 种目标检测算法具有相似的 PCP 精度；当 IoU 阈值从 0.6 变化到 0.8 时，Faster R-CNN 和 R-FCN 具有相似的检测精度，而 SSD 有更高的检测精度，这对于 YUB-200-2017 和 CUB-200-2011 都是一致的.

**表 3　Faster R-CNN 在 YUB-200-2017 上的 PCP**　%

| IoU 阈值 | 整体 | 头部 | 躯干 |
|---|---|---|---|
| 0.5 | 99.9 | 98.8 | 98.9 |
| 0.6 | 99.6 | 97.5 | 97.9 |
| 0.7 | 98.6 | 91.8 | 94.7 |
| 0.8 | 93.4 | 65.4 | 78.6 |

**表 4　R-FCN 在 YUB-200-2017 上的 PCP**　%

| IoU 阈值 | 整体 | 头部 | 躯干 |
|---|---|---|---|
| 0.5 | 99.8 | 98.9 | 98.7 |
| 0.6 | 99.8 | 97.9 | 97.7 |
| 0.7 | 98.6 | 94.2 | 94.7 |
| 0.8 | 91.5 | 72.6 | 78.9 |

**表 5　YOLO 在 YUB-200-2017 上的 PCP**　%

| IoU 阈值 | 整体 | 头部 | 躯干 |
|---|---|---|---|
| 0.5 | 99.6 | 97.7 | 98.2 |
| 0.6 | 98.6 | 93.2 | 95.7 |
| 0.7 | 93.2 | 79.4 | 89.5 |
| 0.8 | 67.8 | 44.4 | 63.3 |

**表 6　SSD 在 YUB-200-2017 上的 PCP**　%

| IoU 阈值 | 整体 | 头部 | 躯干 |
|---|---|---|---|
| 0.5 | 99.8 | 99.3 | 99.5 |
| 0.6 | 99.8 | 98.2 | 98.7 |
| 0.7 | 98.2 | 93.4 | 96.1 |
| 0.8 | 95.6 | 75.5 | 76.3 |

**表 7　Faster R-CNN 在 CUB-200-2011 上的 PCP**　%

| IoU 阈值 | 整体 | 头部 | 躯干 |
|---|---|---|---|
| 0.5 | 99.0 | 93.5 | 94.2 |
| 0.6 | 98.0 | 88.0 | 90.3 |
| 0.7 | 95.0 | 71.2 | 79.8 |
| 0.8 | 82.2 | 54.6 | 63.5 |

**表 8　R-FCN 在 CUB-200-2011 上的 PCP**　%

| IoU 阈值 | 整体 | 头部 | 躯干 |
|---|---|---|---|
| 0.5 | 99.4 | 95.3 | 96.1 |
| 0.6 | 98.5 | 90.0 | 89.9 |
| 0.7 | 95.2 | 77.4 | 80.9 |
| 0.8 | 80.1 | 55.4 | 62.4 |

**表 9　YOLO 在 CUB-200-2011 上的 PCP**　%

| IoU 阈值 | 整体 | 头部 | 躯干 |
|---|---|---|---|
| 0.5 | 97.7 | 90.7 | 90.5 |
| 0.6 | 92.3 | 80.6 | 79.8 |
| 0.7 | 76.3 | 59.2 | 60.1 |
| 0.8 | 43.6 | 24.1 | 30.3 |

**表 10　SSD 在 CUB-200-2011 上的 PCP**　%

| IoU 阈值 | 整体 | 头部 | 躯干 |
|---|---|---|---|
| 0.5 | 99.2 | 92.4 | 95.5 |
| 0.6 | 98.8 | 87.3 | 90.8 |
| 0.7 | 96.7 | 80.9 | 82.9 |
| 0.8 | 82.6 | 57.3 | 65.3 |

表 11 所示为 Faster R-CNN, R-FCN, YOLO 和 SSD 对于单幅图像的语义部件检测时间比较结果. 可以看出, SSD 和 YOLO 具有较快的检测时间.

**表 11　不同方法的语义部件检测时间**　ms

| 语义部件检测方法 | 检测时间 |
|---|---|
| Faster R-CNN | 111 |
| R-FCN | 82 |
| YOLO | 20 |
| SSD | 21 |

综合语义部件的检测精度和检测时间，本文选择基于 ResNet-50 的 SSD 作为鸟类图像语义部件检测算法.

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

### 2.5 基于 ResNet 的分类子网

基于检测子网检测得到的语义部件信息，通过构建一个三路基于 ResNet 的分类模型同时对物体级和部件级的特征进行聚合，这样得到的特征既有整体的语义信息，又包含局部的语义部件信息；然后对每一路输出的特征进行全局平均池化；最后将这三路的特征向量连接起来，通过添加一个 200 维的分类层用于端对端的联合训练.

在训练时采用随机梯度下降算法，每次迭代的采样数目为 64 幅图像，总共训练 30 轮. 因为三路分类网络是基于在 ImageNet 上预训练的 ResNet-50 模型，所以网络学习率的初始值为 0.001，并且在第 20 轮的时候将学习率降低为 0.000 1.

## 3 实验结果与分析

本文实验使用的显卡是 NVIDIA Quadro GP100，显存为 16 GB，使用 CUDA 8.0；操作系统是 Ubuntu 16.04，深度学习框架使用 Caffe.

### 3.1 语义部件检测比较

图 7 所示为本文基于 SSD 的语义部件检测的一些实例，每幅图像中的 3 个矩形分别表示检测到的鸟类整体包围盒、鸟类头部包围盒和鸟类躯干包围盒. 可以看出，对于不同类别和不同姿态下的鸟类图像，都能够准确地对其语义部件进行检测.

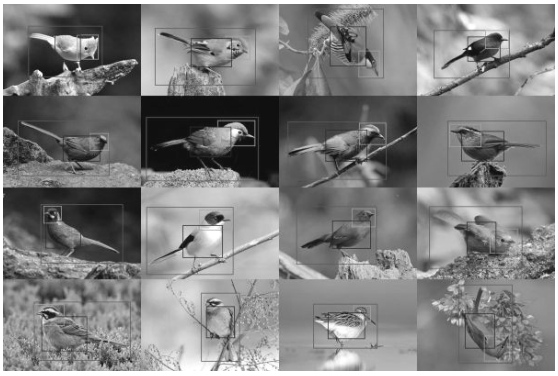

图 7 鸟类图像语义部件检测示例

若鸟类图像成像质量较低或有遮挡，有时会出现语义部件检测失败的情况，这时系统会回退到使用单个模型 ResNet-50 对鸟类图像进行分类.

对不同方法在 CUB-200-2011 上的语义部件定位精度进行比较，当 IoU 阈值为 0.5 时 PCP 结果如表 12 所示. 可以看出，本文基于 SSD 的语义部件检测模型对于鸟类头部的检测精度为 92.4%，对于鸟类躯干的检测精度为 95.5%.

表 12 不同方法语义部件检测的 PCP 比较 %

| 语义部件检测方法 | 头部 | 躯干 |
|---|---|---|
| Part R-CNN[1] | 61.4 | 70.7 |
| Deep LAC[5] | 74.0 | 96.0 |
| SPDA-CNN[6] | 93.4 | 94.9 |
| MASK-CNN[9] | 84.6 | 89.8 |
| 本文 | 92.4 | 95.5 |

### 3.2 鸟类图像分类比较

在数据集 YUB-200-2017 和 CUB-200-2011 上，对不同方法的分类准确率比较结果如表 13 所示. 可以看出，由于语义部件的精确定位以及 ResNet 优异的分类性能，再加上对不同语义部件的特征进行聚合，本文方法在这 2 个数据集上的分类精度分别达到 96.2%和 87.4%.

表 13 各方法在 2 个数据集上分类准确率比较 %

| 鸟类图像分类方法 | YUB-200-2017 | CUB-200-2011 |
|---|---|---|
| SWFV-CNN[13] | 80.3 | 75.2 |
| Part R-CNN[1] | 75.8 | 70.7 |
| Deep LAC[5] | 86.6 | 80.3 |
| SPDA-CNN[6] | 92.7 | 85.1 |
| PS-CNN[8] | 81.2 | 76.2 |
| MASK-CNN[9] | 93.4 | 85.2 |
| Attention-CNN[11] | 83.6 | 77.9 |
| Bilinear-CNN[10] | 91.3 | 84.1 |
| Constellation-CNN[12] | 89.7 | 81.0 |
| 本文 | 96.2 | 87.4 |

## 4 结 语

本文提出一种联合语义部件进行鸟类图像细粒度识别的模型. 在使用基于 SSD 的检测网络对鸟类图像的语义部件进行定位后，使用三路深度残差网络模型对待识别鸟类的语义部件信息进行编码，再进行分类；并收集了一个高质量的细粒度鸟类图像数据集，用于对鸟类图像细粒度识别进行研究. 本文对不同 CNN 架构和不同目标检测算法在 YUB-200-2017 和 CUB-200-2011 上的分类准确率以及语义部件检测精度进行了定量比较. 实验结果表明，本文模型在鸟类图像语义部件定位和细粒度分类 2 个方面都具有较高的准确率.

## 参考文献(References):

[1] Zhang N, Donahue J, Girshick R, *et al*. Part-based R-CNNs for

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

fine-grained category detection[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2014, 8689: 834-849

[2] Uijlings J R, Sande K E, Gevers T, *et al*. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154-171

[3] Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 580-587

[4] Branson S, Van Horn G, Belongie S, *et al*. Bird species categorization using pose normalized deep convolutional nets[C] //Proceedings of British Machine Vision Conference. Nottingham: BMVA Press, 2014: 1–14

[5] Lin D, Shen X Y, Lu C W, *et al*. Deep LAC: deep localization, alignment and classification for fine-grained recognition[C] //Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 1666-1674

[6] Zhang H, Xu T, Elhoseiny M, *et al*. SPDA-CNN: unifying semantic part detection and abstraction for fine-grained recognition[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 1143-1152

[7] Girshick R. Fast R-CNN[C] //Proceedings of International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 1440-1448

[8] Huang S L, Xu Z, Tao D C, *et al*. Part-stacked CNN for fine-grained visual categorization[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 1173-1182

[9] Wei X S, Xie C W, Wu J X, *et al*. Mask-CNN: localizing parts and selecting descriptors for fine-grained bird species categorization[J]. Pattern Recognition, 2018, 76(4): 704-714

[10] Lin T Y, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition[C] //Proceedings of IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 1449-1457

[11] Xiao T J, Xu Y C, Yang K Y, *et al*. The application of two-level attention models in deep convolutional neural network for fine-grained image classification[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 842-850

[12] Simon M, Rodner E. Neural activation constellations: unsupervised part model discovery with convolutional networks[C] //Proceedings of IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 1143-1151

[13] Zhang X P, Xiong H K, Zhou W G, *et al*. Picking deep filter responses for fine-grained image recognition[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 1134-1142

[14] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90

[15] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. [2017-07-10]. https://arxiv.org/abs/1409.1556

[16] Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2818-2826

[17] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778

[18] Zhou B L, Aditya K, Agata L, *et al*. Learning deep features for discriminative localization[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2922-2929

[19] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149

[20] Dai J F, Li Y, He K M, *et al*. R-FCN: object detection via region-based fully convolutional networks[OL]. [2017-07-10]. https://arxiv.org/abs/1605.06409

[21] Redmon J, Divvala S, Girshick R, *et al*. You only look once: unified, real-time object detection[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 779-788

[22] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot multibox detector[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2016, 9905: 21-37

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net