基于语义信息跨层特征融合的细粒度鸟类识别

李国瑞 何小海* 吴晓红 卿粼波 滕奇志
(四川大学 四川 成都 610065)

摘 要 有效识别各种鸟类目标具有重要的生态环境保护意义。针对不同种类鸟类之间差别细微、识别难度大等问题,提出一种基于语义信息跨层特征融合的细粒度鸟类识别模型。该模型由区域定位网络、特征提取网络和一种跨层特征融合网络(Cross-layer Feature Fusion Network,CFF-Net)组成。区域定位网络在没有局部语义标注的情况下,自动定位出局部有效信息区域;特征提取网络提取局部区域图像特征和全局图像特征;CFF-Net 对多个局部和全局特征进行融合,提高最终分类性能。结果表明,该方法在 Caltech-UCSD Birds200-2011(CUB200-2011)鸟类公共数据集上,取得了 87.8% 的分类准确率,高于目前主流的细粒度鸟类识别算法,表现出优异的分类性能。

关键词 鸟类识别 细粒度识别 区域定位 特征提取 特征融合

中图分类号 TP391.41 **文献标志码** A DOI: 10.3969/j.issn.1000-386x.2020.04.022

FINE-GRAINED BIRD RECOGNITION BASED ON SEMANTIC INFORMATION CROSS-LAYER FEATURES FUSION

Li Guorui He Xiaohai* Wu Xiaohong Qing Linbo Teng Qizhi
(*Sichuan University*, *Chengdu 610065*, *Sichuan*, *China*)

Abstract In view of the subtle differences between different bird species and the difficulty of recognition, we propose a fine-grained bird recognition model based on cross-layer feature fusion of semantic information. It consists of regional location network, feature extraction network and cross-layer feature fusion network (CFF-Net). The regional location network automatically located the local effective information region without local semantic annotation; feature extraction network extracted local and global image features; CFF-Net combined multiple local and global features to improve the final classification performance. The results show that the classification accuracy is 87.8% on Caltech-UCSD Birds200-2011 (CUB200-2011) dataset, which is higher than the current mainstream fine-grained bird recognition algorithm. And it shows excellent classification performance.

Keywords Bird recognition Fine-grained recognition Regional location Feature extraction Feature fusion

0 引 言

细粒度图像识别是深度学习领域的重要研究方向,其目的是对属于同一基础类别的图像进行更加细致的从属类别划分,由于从属类别内部之间差别细微,细粒度图像识别任务相较于传统通用图像识别任务难度更高。近年来,随着我国生态保护事业的蓬勃发展,物种监控图像视频数量剧增,生物种类识别的需求也剧增。细粒度鸟类识别成为其中重要的任务之一,其识别结果可以帮助生物学家有效监控鸟类种群分布及生态环境的变迁。目前,针对细粒度图像识别任务,大多数研究都以卷积神经网络(Convolutional Neural networks,CNN)为基础,主要分为基于强监督学习的细粒度图像识别和基于弱监督学习的细粒度图像识别两大类[1]。

收稿日期:2019-06-17。国家自然科学基金项目(61871278);四川省科技计划项目(2018HH0143);四川省教育厅项目(18ZB0355)。李国瑞,硕士生,主研领域:图像处理。何小海,教授。吴晓红,副教授。卿粼波,副教授。滕奇志,教授。

基于强监督学习的细粒度图像识别，除了使用图像真实类别标签以外，还使用了目标标注框坐标等局部语义标注信息。Wei 等[2]提出的 Mask-CNN，是首个端到端地将深度卷积特征运用到物体检测的细粒度图像识别模型。基于强监督学习的细粒度图像识别方法使用了局部语义标注信息，相较于传统 CNN 方法，检测精度和模型泛化性能均有明显提升。但由于人工标注成本昂贵，且不能确保局部语义的有效性，此类算法在实际应用中受到限制。因此，目前主流的研究方法基于弱监督学习的思想，其优点在于，模型仅使用图像真实类别标签，不再使用局部语义标注，也能准确定位到局部关键区域，得到与基于强监督学习相当的准确率。Yu 等[3]提出 HBP 模型，开发了一种简单有效的跨层双线性池化技术，以一种相互增强的方式学习图像的细粒度表征。Yang 等[4]首次提出一种新颖的自监督机制网络 NTS-Net，可以有效定位出关键区域而无需局部语义标注信息，在广泛的基准数据集上实现了最先进的性能。

尽管对细粒度图像识别的研究已经取得了不少成果，但仍有诸多问题亟待解决。本文工作基于弱监督学习的思想，主要解决细粒度鸟类识别的两大难点：第一是在没有局部语义标注的情况下，自动定位到具有有效信息的关键区域；第二是提出一种有效的特征融合方式以提高最终分类性能。

1　相关基础网络

1.1　残差网络

残差网络(ResNet)[5]首次提出残差块结构，其基本结构如图 1 所示。该结构在增加网络深度的同时，能有效减少网络参数量，防止过拟合现象发生，一定程度上避免网络性能随深度增加而降低。

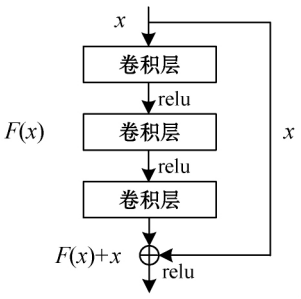

图 1　ResNet 网络残差块结构

ResNet 有不同的网络层数，其中 ResNet-50 共有 50 层网络层，分为 Conv_1 层、Conv_2 层、Conv_3 层、Conv_4 层和 Conv_5 层，Conv_1 层为 1 个单独卷积层，Conv_2 层到 Conv_5 层分别包含 3、4、6、3 个残差块结构，Conv_5 层后为全局均值池化层(Global Average Pooling，GAP)，GAP 层后为全连接层(Fully Connected Layers，FC)。

1.2　区域建议网络

Ren 等[6]在多目标检测中提出区域建议网络(Region Proposal Networks，RPN)，利用 CNN 卷积操作后的特征图谱生成具有有效信息的区域，代替了选择性搜索等方法，在检测速度上提升明显。

RPN 是一种全卷积神经网络，整个网络没有全连接层，所以该网络能接受任意尺寸的图像输入，输出一系列图像局部矩形区域坐标及每个区域是目标和背景的概率得分，原理如图 2 所示。锚点是特征图谱上的一个像素映射到原图像上的像素位置，对应于一组预先设定的 k_1 个不同尺度和 k_2 个宽高，以相应锚点为中心，生成 $k=k_1\times k_2$ 锚点框。对于 $m\times m$ 大小的特征图谱，采用 $3\times3\times256$ 卷积核进行卷积，得到 $m\times m\times256$ 维向量，再用大小为 $1\times1\times2k$ 卷积核对每个 256 维向量进行卷积，从而得到对应的 $2k$ 个置信度，代表特征图谱上相应像素对应原图像上的锚点对应的锚点框前景和背景的概率。同时，用 $1\times1\times4$ 大小卷积核对 256 维向量进行卷积操作，得到 $4k$ 个锚点框的位置信息，每个锚点框的位置信息由矩形框的左上角点的横坐标偏移量、纵坐标偏移量和矩形框的长度偏移量、宽度偏移量共 4 个数据组成。

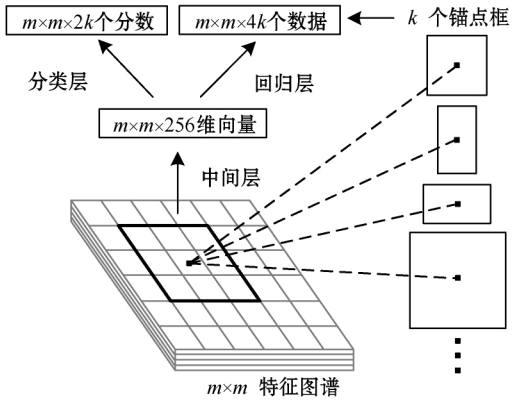

图 2　RPN 核心原理图

RPN 网络会产生大量的候选框坐标，因此存在对同一目标产生多个近似的候选框。Neubeck 等[7]提出的非极大值抑制(Non-MaximumSuppression，NMS)算法，能有效去除同一目标冗余的候选框，保留信息含量最丰富的候选框。首先，对同一类别的候选框按 RPN 网络得分高低排序，选出得分最高的候选框；其次，遍历剩余得分候选框，计算与当前最高得分候选框的重叠面积(Intersection over Union，IoU)，设为 S，如果 S 大于一定阈值，将该遍历的候选框删除；重复前面步骤，直到所有的剩余候选框都和得分最高的候选框比较过，留下非冗余候选框。S 的计算方法如下：

$$S_{A,B}=\frac{A\cap B}{A\cup B} \tag{1}$$

式中:A 代表得分最高的候选框区域,B 代表每次遍历的候选框区域,$A\cap B$ 表示区域 A、B 交集部分的面积,$A\cup B$ 表示区域 A、B 并集部分的面积。

2 细粒度鸟类识别网络模型

针对细粒度鸟类识别的两大难点,本文提出如图 3 所示的细粒度鸟类识别网络模型。该模型由 3 种网络组成,分为区域定位网络、特征提取网络和特征融合网络 CFF-Net。其中,特征提取网络用于提取全局和局部图像特征,且所有特征提取网络共享网络训练参数;区域定位网络用于在没有局部语义标注信息的情况下,自动定位出图像中信息含量最丰富的 Top-n 个局部区域,且对每一个区域信息量含量打分,按大小分别排序为:I_0,I_1,I_2,…,I_n;CFF-Net 网络对多个局部和全局特征进行融合,有效结合局部与全局特征信息,提高最终分类性能。

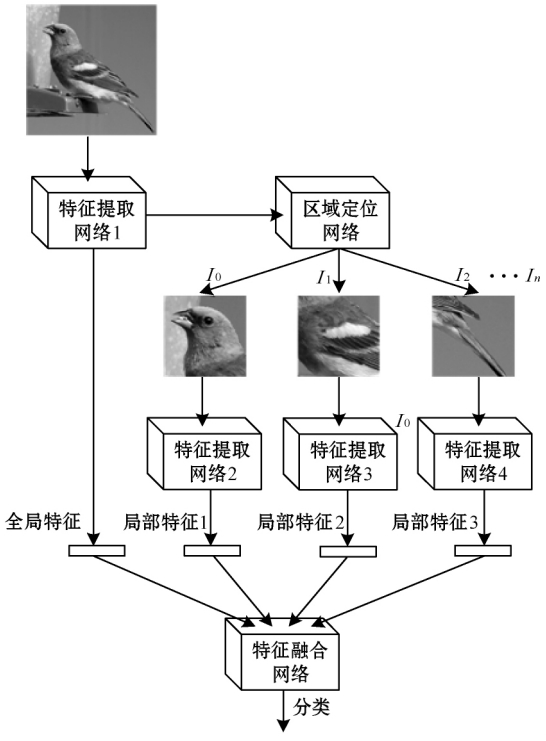

图 3 细粒度鸟类识别网络模型

2.1 特征提取网络

基本特征提取网络既要保证能提取有效图像特征及较快的损失收敛速度,还要避免过拟合现象的发生。本文采用 ResNet-50 作为特征提取网络,由于训练集图片与测试集图片数量相当,在 ResNet-50 的 GAP 层后增加一个 dropout 层,dropout radio 设为 0.5,防止过拟合现象发生。同时修改 FC 层输出参数,使其最终输出为 200 维,满足数据集类别总数。特征提取网络在训练时加载在 ImageNet 图像库预训练好的模型参数,对网络进行微调。

2.2 区域定位网络

基于弱监督学习的细粒度图像识别,其难点是在没有局部语义标注信息的情况下,自动定位出具有有效信息的关键区域。由于图 2 中原始 RPN 网络需要局部语义标注信息为监督对位置进行精细修正,因此对原始 RPN 网络做如下精简:

(1) 删除图 2 中 RPN 网络回归层,直接得出对应于原图的局部区域坐标,不做位置的精细修正;

(2) 修改图 2 中分类层卷积核大小为 $1\times1\times256\times k$,只得到 k 个置信度,将其定义为每个锚点框内所含有效信息的丰富程度,而不再代表每个锚点框内目标和背景的概率大小。

本文引入精简后的 RPN,改进提出图 4 所示的区域定位网络。

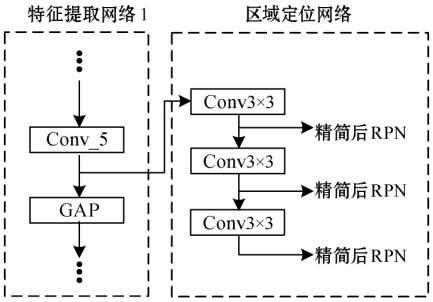

图 4 区域定位网络原理图

区域定位网络共享特征提取网络的前 Conv_5 层,然后引入连续的 3 个卷积网络,在每层卷积网络输出的特征图谱上使用精简后的 RPN 网络。原始 RPN 网络进行目标检测时,以最后一层卷积作为输入,该方法在大目标定位时性能优异,但对于小目标,卷积池化到最后一层时,语义信息已经基本消失,不能很好地定位出小目标。因此,借鉴 FPN [8] 网络的思想,针对不同的特征图谱层使用精简后的 RPN 网络,较浅的特征图谱层用于定位更小的目标,而较深的特征图谱层用于定位更大区域的目标。

由于输入图片大小为 448×448,对每一层特征图谱选定的基准锚点框大小分别设为 64×64、128×128 和 256×256,单独设置每一层锚点框的面积比和宽高比,共产生 1 614 个候选框与对应的信息量得分。经 NMS 算法后,选取 Top-n 个得分最高的候选区域,作为局部区域,将其上采样到 224×224 大小,送入后续的特征提取网络,其中,NMS 算法中 IoU 取 0.3。

区域定位网络作为选取局部有效信息区域的重要手段,在没有局部语义标注信息的前提下,如何保证得

分越高的区域，其所含有效信息越丰富，成为影响定位网络性能的核心因素。NTS-Net 提出了一种新颖的自监督机制，将整幅图像的类别标签作为局部区域图像的类别标签，局部区域图像经特征提取网络后有图像类别标签作为监督，得到置信度，即为标定图像(ground-truth)的概率大小。置信度越高的局部区域图像，其所含有效信息越丰富，对应的通过定位网络的打分应该更高。基于该思想，定义经过区域定位网络得到的前 M 个得分最高的区域为:

$$R=\{R_1,R_2,R_3,\cdots,R_M\} \tag{2}$$

其得分为:

$$I=\mathcal{I}(R)=\{I_1,I_2,I_3,\cdots,I_M\} \tag{3}$$

对应的置信度为:

$$C=\{C_1,C_2,C_3,\cdots,C_M\} \tag{4}$$

由此可知，若 $C_1>C_2>C_3>\cdots>C_M$，则 $I_1>I_2>I_3>\cdots>I_M$。其损失函数定义为:

$$L_{\mathcal{I}}(I,C)=\sum_{(i,s):C_i<C_s}\mathcal{F}(I_s-I_i) \tag{5}$$

上述公式中，函数应为非增函数，用来确保当 $C_s>C_i$ 时，$I_s>I_i$。该损失函数能有效监督区域定位网络的打分性能，定位出局部有效信息区域，结合全局图像信息，使最终分类性能提升。

2.3　特征融合网络 CFF-Net

局部区域有助于更细致地表征对象，因此融合局部区域特征和全局图像的特征将获得更好的分类性能。目前，局部图像和全局图像特征融合方式大多采用简单的级联，不能充分利用局部区域所表达的更细微的特征信息。本文提出了一种基于跨层技术的特征融合网络 CFF-Net，其网络结构如图 5 所示。

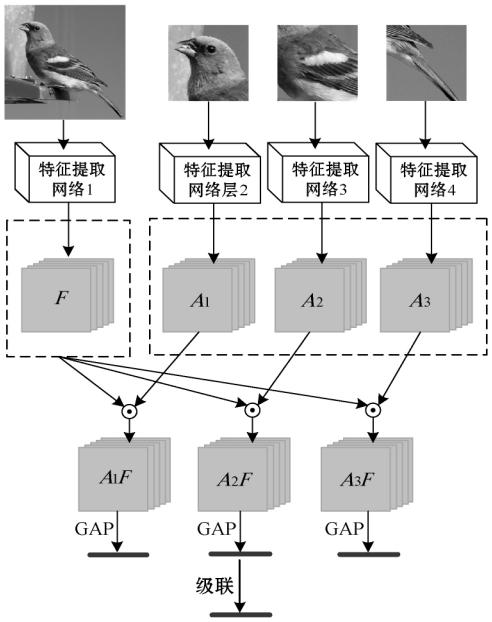

图 5　CFF-Net 网络结构示意图

区域定位网络定位出前 M 个局部区域，经特征提取网络后，形成局部特征图谱 A_1、A_2、A_3、…，与全局特征图谱 F 分别进行点乘操作。该过程可用以下公式表示:

$$F_k=A_k\odot F\quad k=1,2,\cdots,M \tag{6}$$

点乘操作得到的特征图谱，经 GAP 层后得到特征向量，该过程可定义为:

$$f_k=\mathcal{H}(F_k) \tag{7}$$

CFF-Net 整个特征融合过程可用如下公式表示:

$$P=\mathcal{R}(A,F)=\begin{pmatrix}\mathcal{H}(A_1\odot F)\\ \mathcal{H}(A_2\odot F)\\ \vdots\\ \mathcal{H}(A_M\odot F)\end{pmatrix}=\begin{pmatrix}f_1\\ f_2\\ \vdots\\ f_M\end{pmatrix} \tag{8}$$

3　实　验

3.1　实验数据集

本文所用数据集 CUB-200-2011，是一个鸟类公共图片数据库，包含 200 种鸟类，共计 11 788 幅图片。其中 5 994 幅为训练集，5 794 幅为测试集。图 6 展示了数据集中两种不同种类的鸟类，如果观察者没有相关领域的专业知识，很难区分两种鸟类，而该数据集包含很多类似细微差别的种类。

Black_Billed_Cuckoo　　Mangrove_Cuckoo

图 6　数据集中两种鸟类对比

3.2　模型训练

本文实验使用 GPU 为 Nvidia GTX1080Ti，显存 11 GB，CPU 为 Inter Core i5 7500，内存为 8 GB，使用 CUDA 8.0，操作系统为 Ubuntu 16.04，深度学习框架使用 Pytorch-0.4.1。

本文所提细粒度鸟类识别网络能实现端到端地的训练。实现细节中，batch_size 设为 10，初始学习率设为 0.001，每过 60 次 epoch 学习率变为原来十分之一，使用 SGD 优化器。图 7 展示了训练时总损失的收敛情况，可以看出，在 40 次 epoch 后总损失不再明显下降，50 次 epoch 后总损失几乎不再变化。其原因为: 学习率设置过大，导致梯度来回震荡。因此 60 次 epoch 后将学习率降低，可见总损失继续收敛。

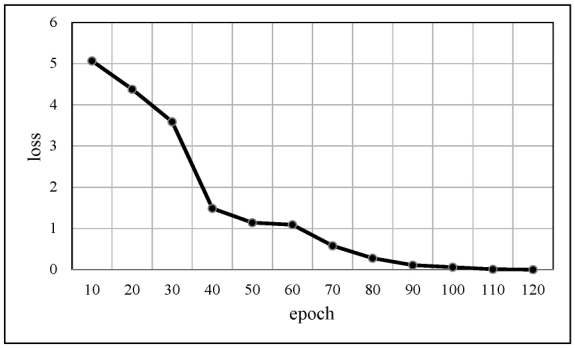

图 7 训练损失迭代

3.3 区域定位网络性能测试

图 8 展示了本文改进的区域定位网络在数据集上定位的前 4 个得分最高的局部区域。可见，定位的局部信息区域基本包括了鸟类的头部、躯干及翅膀等能明显表征鸟类特点的部位，显示了区域定位网络在没有局部语义标注信息下优异的自动定位能力。

图 8 区域定位网络定位的局部区域

3.4 特征融合网络 CFF-Net 性能测试

为了验证 CFF-Net 的性能，本文在实验时对改进后的 ResNet 不同特征图谱或特征向量分别做特征融合操作。融合方式如表 1 所示。

表 1 特征融合组合方式

融合方式	特征提取 1	特征提取 2－4	特征大小
方式 1	Conv_4	Conv_3	1 048 × 28 × 28
方式 2	Conv_5	Conv_4	2 048 × 14 × 14
方式 3	GAP	GAP	2 048 × 1 × 1
方式 4	FC	FC	200 × 1 × 1

局部信息区域在特征提取前上采样到 224 × 224 大小，全局图像大小为 448 × 448，因此，对于同样的卷积层输出，特征提取层 2、3、4 总是比特征提取层 1 得到的特征图谱小。方式 1 将全局图像输入的 Conv_4 层特征图谱和局部区域输入的 Conv_3 层特征图谱做特征融合操作，得到的特征图谱大小为 1 048 × 28 × 28；方式 3 和方式 4 直接对特征向量做融合操作，得到的也是特征向量。由于 NTS-Net 中局部区域个数 M 取 4 时表现出最优异的分类性能，表 2 展示了在 $M=4$ 时，不同特征融合方式的最终分类准确率对比实验结果。

表 2 不同特征融合方式准确率

融合方式	准确率/%
方式 1	85.3
方式 2	87.1
方式 3	87.6
方式 4	**87.8**

由表 2 可以看出，特征融合方式 1 分类性能较低，原因在于方式 1 将融合后大小为 28 × 28 的特征图谱直接进行了全局均值池化，导致语义特征信息的丢失。方式 2 和方式 3 使用更深层卷积的特征图谱进行融合，丢失的语义信息变少，分类性能明显提高，方式 4 准确率最高为 87.8%，高于目前主流的细粒度鸟类识别算法，验证了本文所提 CFF-Net 网络的有效性。

为了进一步验证不同局部区域个数对最终分类性能的影响，本文测试时选取分类性能最优异的特征融合方式 4，加入不同局部区域个数 M 做对比实验，其中 $M=0,1,2,3,4,5$。结果如表 3 所示，可以看出，局部信息区域的加入对最终分类结果提升了 2.7%，随着加入区域个数不断增多，识别准确率不断增加，当 $M=4$ 时，准确率最高为 87.8%；当 $M=5$ 时，识别率开始下降，表明后续局部区域所含有效信息已经不能更好地表征图像全局信息。该结果验证了局部有效信息区域的加入对分类性能的提升，同时也进一步验证了 M 取 4 的有效性。

表 3 不同 M 下准确率对比

M 取值	准确率/%
0	85.1
1	86.4
2	87.3
3	87.6
4	**87.8**
5	87.2

3.5 不同算法分类性能对比

表 4 展示了不同方法在 CUB-200-2011 数据集上分类准确率的比较。由于区域定位网络优异的定位性能及 CFF-Net 有效的特征融合能力，本文方法在 CUB-200-2011 数据集上取得了 87.8% 的分类准确率，高于目前主流的细粒度鸟类识别方法，表现出优异的分类性能。

（下转第 191 页）

tice Hall，2007.

[7] Bartoli A，Sturm P. The 3D Line Motion Matrix and Alignment of Line Reconstructions [J]. International Journal of Computer Vision，2004，57(3)：159－178.

[8] Solà J，Vidal-Calleja T，Civera J，et al. Impact of landmark parametrization on monocular EKF-SLAM with points and lines [J]. International Journal of Computer Vision，2012，97(3)：339－368.

[9] 张鸿燕，耿征. Levenberg-Marquardt 算法的一种新解释 [J]. 计算机工程与应用，2009，45(19)，5－8.

[10] Ma L，Kerl C，Stuckler J，et al. CPA-SLAM: Consistent plane-model alignment for direct RGB-D SLAM [C]//2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE，2016.

[11] Zhang Z，Forster C，Scaramuzza D. Active exposure control for robust visual odometry in HDR environments [C]//IEEE International Conference on Robotics & Automation. IEEE，2017.

[12] 戴立根. 欠点特征环境的视觉同步定位与制图研究 [D]. 哈尔滨: 哈尔滨工业大学，2016.

[13] 毛星云，冷雪飞. Opencv3 编程入门 [M]. 北京: 电子工业出版社，2015.

[14] 魏玉慧，王永军，王国东，等. 点线特征融合的误匹配剔除算法 [J]. 计算机科学，2019，46(2)：286－293.

[15] 贾哲，冷建伟. 线特征融合光流的单目 SLAM 算法 [J]. 计算机工程与科学，2018，40(12)：2198－2204.

(上接第 136 页)

表 4　不同算法分类性能对比

鸟类识别方法	CUB-200-2011/%
MASK-CNN [2]	85.2
RA-CNN [3]	85.3
DT-RAM [10]	86.0
MA-CNN [11]	86.5
MMC-CNN [12]	86.5
HBP [3]	87.1
NTS-Net [4]	87.5
本文方法	**87.8**

4　结　语

本文针对鸟类图像的细粒度识别，提出了一种基于语义信息跨层特征融合的识别网络。区域定位网络在没有局部语义标注信息的情况下，自动定位出局部有效信息区域; 特征提取网络提取全局图像与局部区域图像的有效特征; CFF-Net 对多个特征图谱或向量融合。本文方法在 CUB-200-201 数据集上表现出良好的自动局部区域定位能力，取得了 87.8% 的分类准确率，高于目前主流的细粒度鸟类识别算法。

参 考 文 献

[1] 罗建豪，吴建鑫. 基于深度卷积特征的细粒度图像分类研究综述 [J]. 自动化学报，2017，43(8)：1306－1318.

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

[3] Yu C，Zhao X，Zheng Q，et al. Hierarchical bilinear pooling for fine-grained visual recognition [C]//Proceedings of the European Conference on Computer Vision. 2018: 574－589.

[4] Yang Z，Luo T，Wang D，et al. Learning to navigate for fine-grained classification [C]//Proceedings of the European Conference on Computer Vision. 2018: 420－435.

[5] He K，Zhang X，Ren S，et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770－778.

[6] Ren S，He K，Girshick R，et al. Faster r-cnn: Towards real-time object detection with region proposal networks [C]//Advances in neural information processing systems. 2015: 91－99.

[7] Neubeck A，Van Gool L. Efficient non-maximum suppression [C]//18th International Conference on Pattern Recognition(ICPR'06). IEEE，2006，3: 850－855.

[8] Lin T Y，Dollár P，Girshick R，et al. Feature pyramid networks for object detection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117－2125.

[9] Fu J，Zheng H，Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4438－4446.

[10] Li Z，Yang Y，Liu X，et al. Dynamic computational time for visual attention [C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 1199－1209.

[11] Zheng H，Fu J，Mei T，et al. Learning multi-attention convolutional neural network for fine-grained image recognition [C]//Proceedings of the IEEE international conference on computer vision. 2017: 5209－5217.

[12] Sun M，Yuan Y，Zhou F，et al. Multi-attention multi-class constraint for fine-grained image recognition [C]//Proceedings of the European Conference on Computer Vision. 2018: 805－821.