MVA Microsoft Virtual Academy

DevX

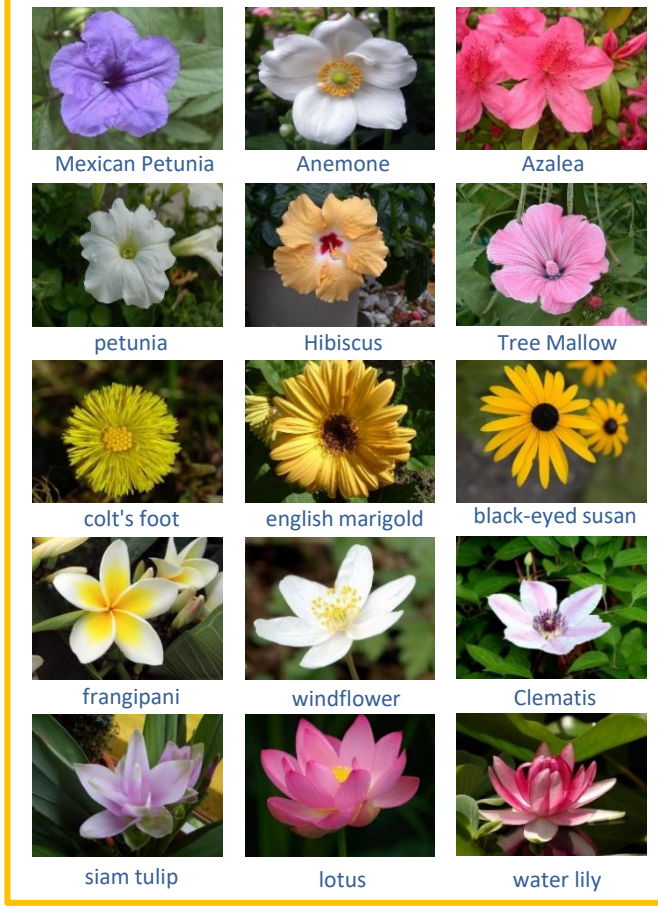
精细化物体识别

Microsoft

傅建龙 | 微软亚洲研究院 多媒体搜索与挖掘 副研究员
jianf@microsoft.com

Fine-grained Recognition [MSRA, CVPR'17, ICCV'17]

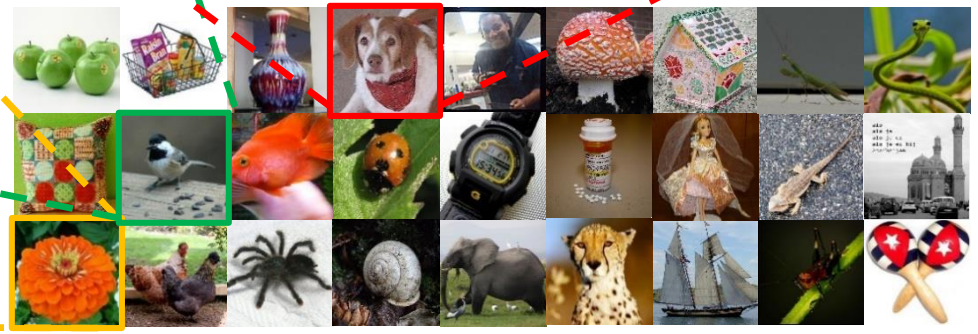
Fine-grained flowers (~250k species in the world)



Fine-grained birds (~10k species in the world)

Fine-grained dogs (~340 breeds in the world)

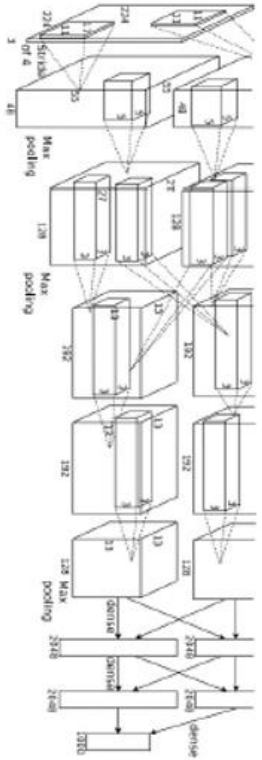
ImageNet (includes ~20 flower species, ~30 bird species and ~80 dog breeds)



General recognition networks

Year 2012

AlexNet



[Krizhevsky NIPS 2012]

Year 2014

GoogLeNet

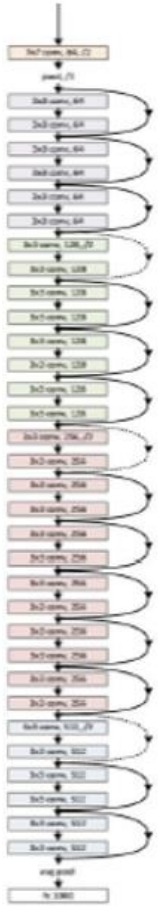
[Szegedy CVPR 2015]

VGG

[Simonyan ICLR 2015]

Year 2015

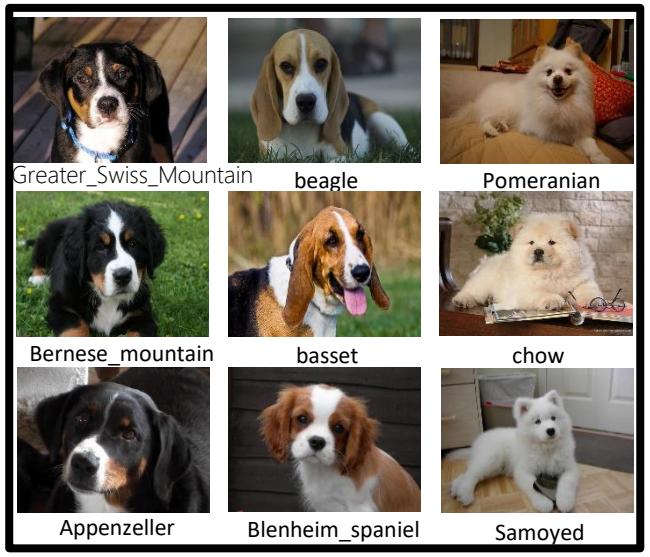
ResNet-152



[He CVPR2016]

Challenges on fine-grained categories

CUB-200-2011 [P. Welinder et.al. 2010]



Stanford-Dogs [Fei-fei Li et.al. 2011]

Challenges:

- Discriminative region localization
  - Localizing the very marginal visual differences from highly-localized regions
- Fine-grained feature learning
  - Describing the subtle visual differences by representative visual features

The state-of-the-art general recognition network (Resnet-152)

Dataset	CUB-Birds (200 categories)	Stanford-Cars (196 categories)	Stanford-Dogs (120 categories)
Accuracy	77.3%	87.5%	87.3%

Our fine-grained image recognition network (CVPR 2017)

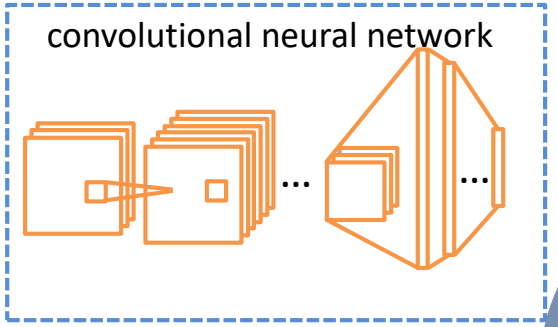
Dataset	CUB-Birds (200 categories)	Stanford-Cars (196 categories)	Stanford-Dogs (120 categories)
Accuracy	**86.5%** ↑ 9.2%	**93.8%** ↑ 6.3%	**89.3%** ↑ 2.0%

Technique overview

Artic Tern

Fine-grained recognition

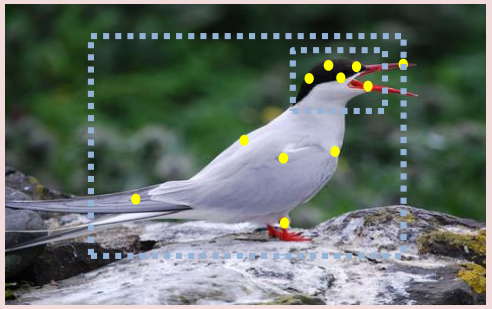
=



General recognition net

+

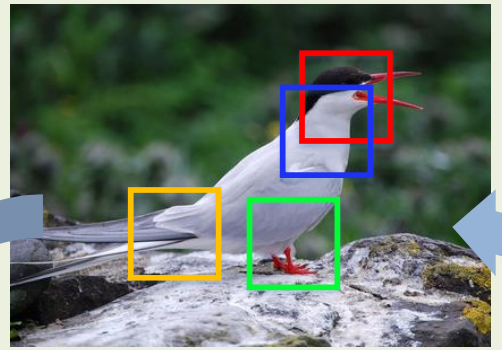
earlier research focus (from 2009-2014)



X = human-annotated part key points

Problem: heavily depends on human involvement, which is hardly extend to large-scale data.

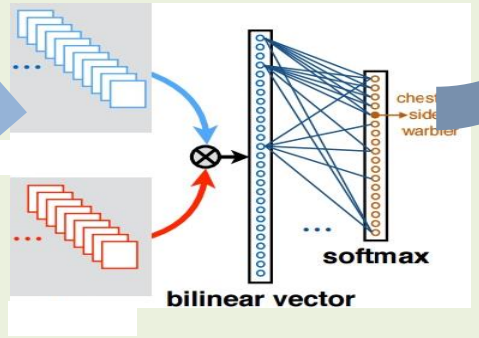
recent research focus (from 2015)



X = weakly-supervised part learning

Problem: the learned part regions may not be optimal, due to the less representative visual features.

Recurrent learning in a reinforced way

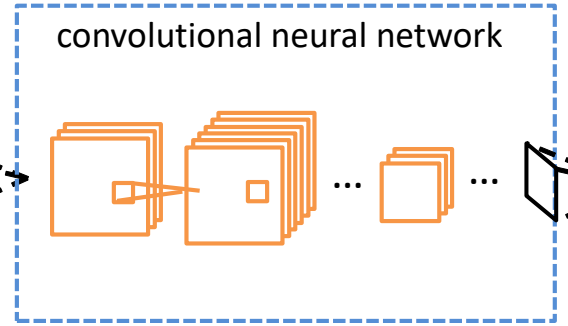


= fine-grained feature learning

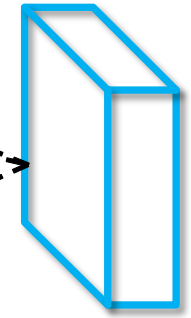
Problem: more fine-grained feature representations on parts cannot be further learned.

Part learning from feature maps

Bohemian Waxwing



convolution
+pooling
+activation



Conv5_4
(512 channels)

12th from 512 channels

128th from 512 channels

267th from 512 channels

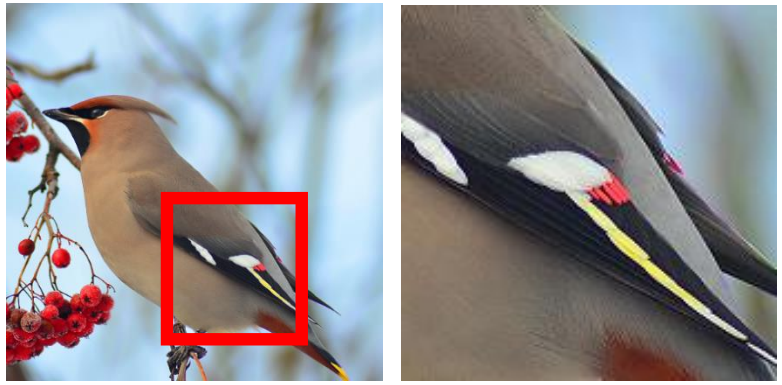
507th from 512 channels

Different feature maps from conv5_4 of a VGG-19 net
(trained on CUB-200-2011 Birds)

Observation: deep feature maps (i.e., feature learning) can help localize semantic parts.

Fine-grained recognition: challenges

original amplified



Bohemian Waxwing (太平鸟)

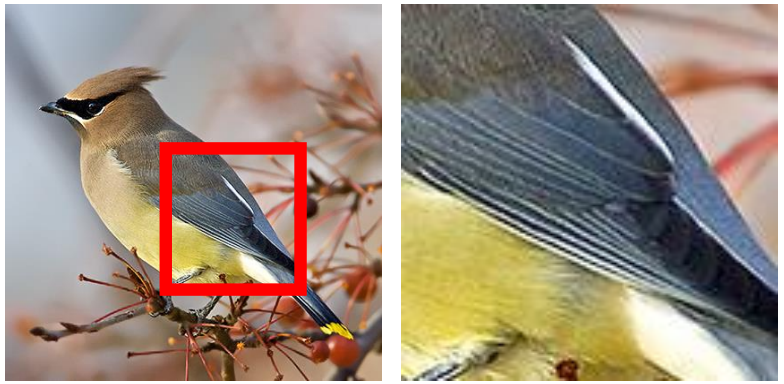
original amplified

Elegant Tern

original amplified

Nashville Warbler

original amplified



Cedar Waxwing (雪松太平鸟)

original amplified

Caspian Tern

original amplified

Orange Crowned Warbler

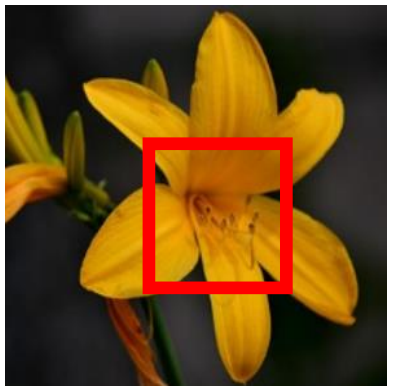
Observation: the subtle visual differences can be clearly represented from amplified parts.

Fine-grained recognition: challenges

original amplified

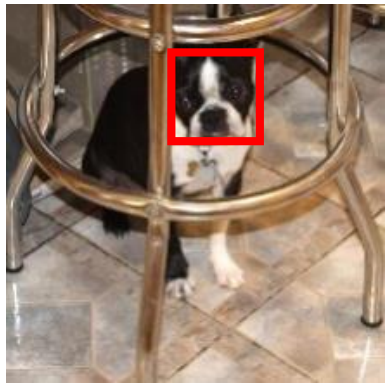

Lilium Ochraceum Fr.

original amplified



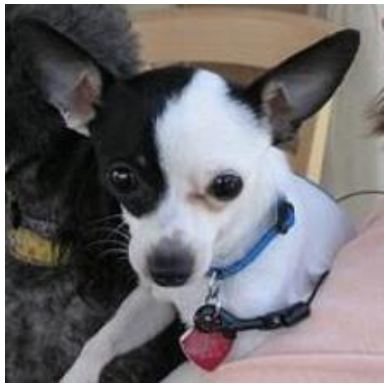

Hemerocallis Fulva (L.) L.

original amplified



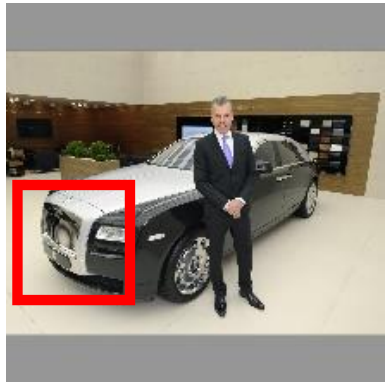

Boston Bull

original amplified

Chihuahua

original amplified




Rolls Royce

original amplified


Buick

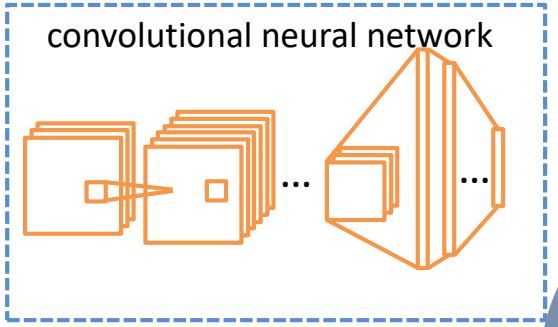
Observation: the subtle visual differences can be clearly represented from amplified parts.

Technique overview

Artic Tern

Fine-grained recognition

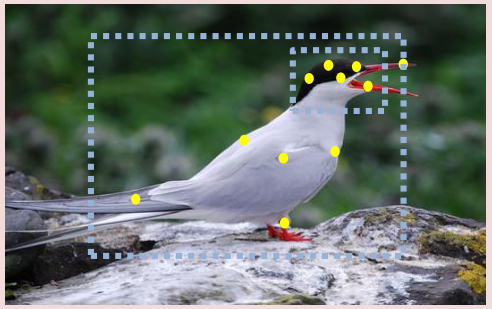
=



General recognition net

+

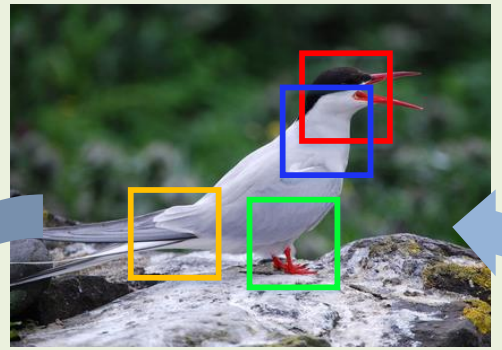
earlier research focus (from 2009-2014)



X = human-annotated part key points

Problem: heavily depends on human involvement, which is hardly extend to new fine-grained domains.

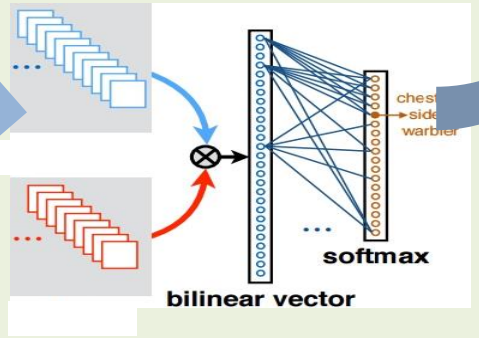
recent research focus (from 2015)



X = weakly-supervised part learning

Problem: the learned part regions may not be optimal, due to the less representative visual features.

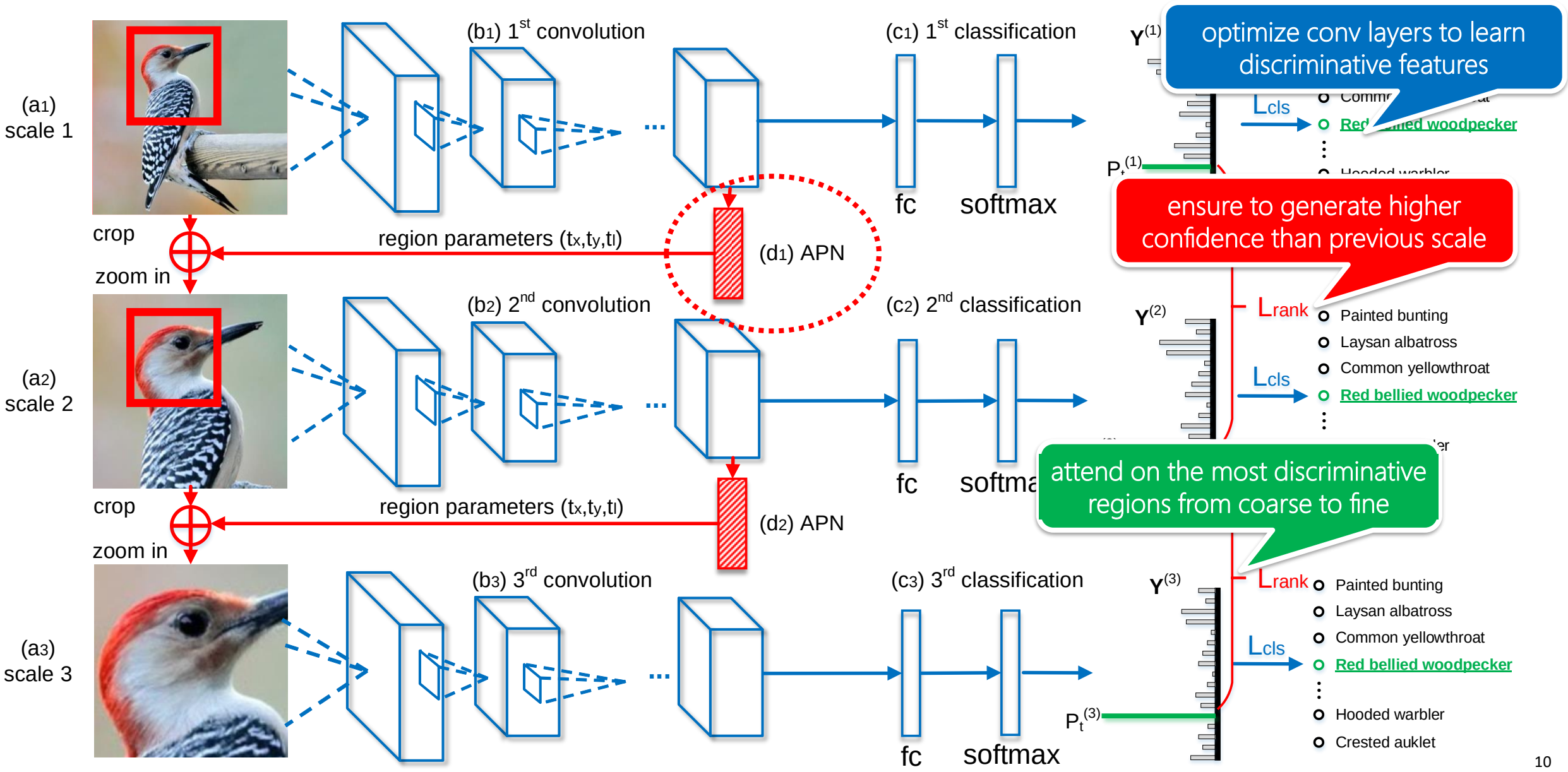
Recurrent learning in a reinforced way



= fine-grained feature learning

Problem: more fine-grained feature representations on parts cannot be further learned.

Recurrent-Attention Network (Fu. CVPR 2017 oral)



Multi-task Formulation

Classification Net

$$\mathbf{p}(\mathbf{X}) = f(\mathbf{W}_c * \mathbf{X})$$

Attention Proposal Net

$$[t_x, t_y, t_l] = g(\mathbf{W}_c * \mathbf{X})$$

Input for the Next Scale

$$\mathbf{X}^{att} = \mathbf{X} \odot \mathbf{M}(t_x, t_y, t_l)$$

$$\mathbf{X}^{amp}_{(i,j)} = \sum_{\alpha,\beta=0}^{1} |1 - \alpha - \{i/\lambda\}||1 - \beta - \{j/\lambda\}|\mathbf{X}^{att}_{(m,n)}$$

where M indicates an attention mask.

Optimization

$$L(\mathbf{X}) = \sum_{s=1}^{3}\{L_{cls}(\mathbf{Y}^{(s)}, \mathbf{Y}^*)\} + \sum_{s=1}^{2}\{L_{rank}(p_t^{(s)}, p_t^{(s+1)})\}$$

where L_{cls} indicates classification loss in each scale, L_{rank} denotes pair-wise ranking loss.

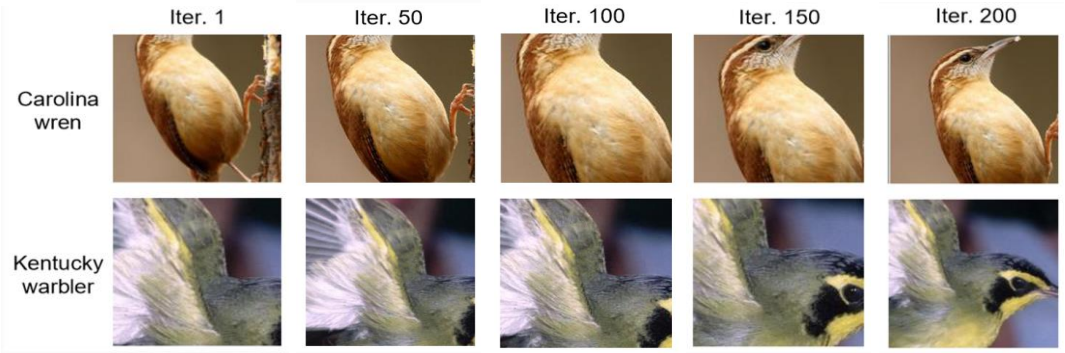


Figure. An illustration of the learning process for region attention.

Experiment Settings

Evaluation	Domain	# category	# Training	# Testing
1. CUB-200-2011		200	5,994	5,794
2. Stanford-Dogs		120	12,000	8,580
3. Stanford-Cars		196	8,144	8,041
4. FGVC-Aircraft		100	6,667	3,333
5. VIREO Food		172	66,144	33,072

Experiment Results

Table 1: Results on CUB-200-2011 dataset

Approach	Train Anno.	Accuracy
DeepLAC [34]	✓	80.3
Part-RCNN [33]	✓	81.6
PA-CNN [14]	✓	82.8
MG-CNN [28]	✓	83.0
FCAN [20]	✓	84.3
B-CNN (250k-dims) [19]	✓	85.1
SPDA-CNN [32]	✓	85.1
PN-CNN [2]	✓	85.4
VGG-19 [27]		77.8
TLAN [31]		77.9
DVAN [35]		79.0
NAC [26]		81.0
MG-CNN [28]		81.7
FCAN [20]		82.0
PDFR [34]		82.6
B-CNN (250k-dims) [19]		84.1
ST-CNN (Inception net) [11]		84.1
RA-CNN (scale 2)		82.4
RA-CNN (scale 3)		81.2
RA-CNN (scale 1+2)		84.7
RA-CNN (scale 1+2+3)		**85.3**

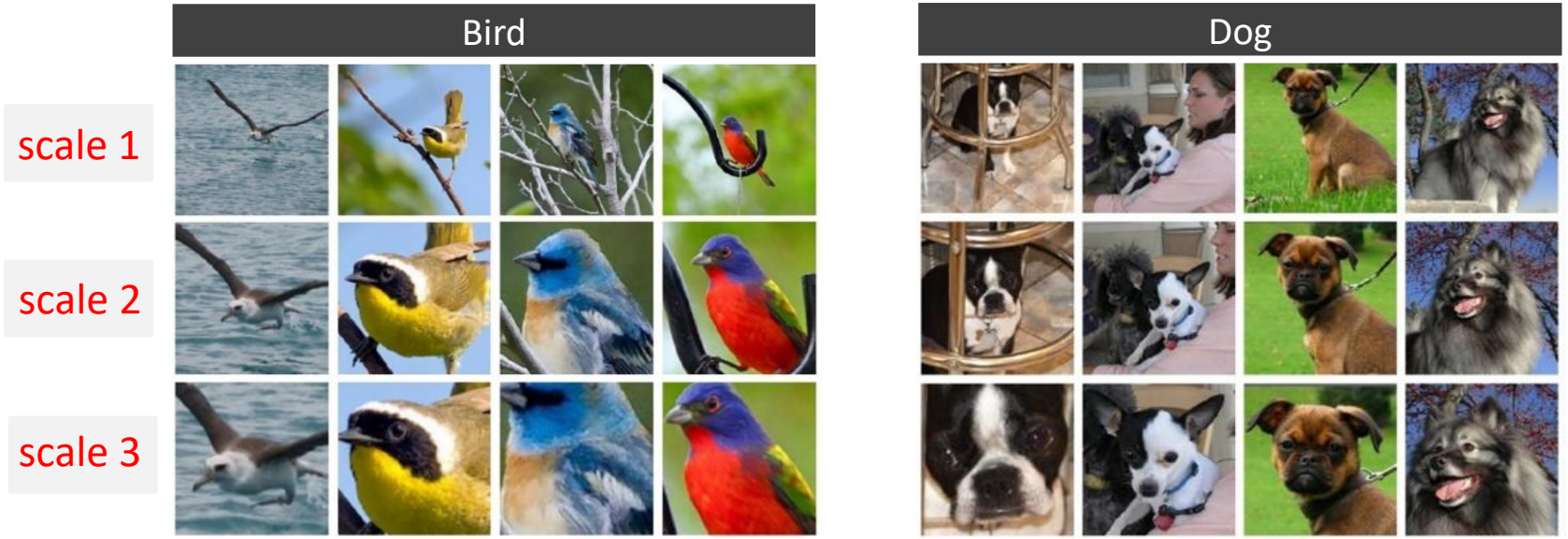
Table 2: Results on Stanford-Dog dataset

Approach	Accuracy
NAC (AlexNet) [26]	68.6
PDFR (AlexNet) [34]	71.9
VGG-16 [27]	76.7
DVAN [35]	81.5
FCAN [20]	84.2
RA-CNN (scale 2)	85.9
RA-CNN (scale 3)	85.0
RA-CNN (scale 1+2)	86.7
RA-CNN (scale 1+2+3)	**87.3**

Table 3: Results on Stanford-Car dataset

Approach	Train Anno.	Accuracy
R-CNN [7]	✓	88.4
FCAN [20]	✓	91.3
PA-CNN [14]	✓	92.8
VGG-19 [27]		84.9
DVAN [35]		87.1
FCAN [20]		89.1
B-CNN (250k-dims) [19]		91.3
RA-CNN (scale 2)		90.0
RA-CNN (scale 3)		89.2
RA-CNN (scale 1+2)		91.8
RA-CNN (scale 1+2+3)		**92.5**

Experiment Results



Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-grained Image Recognition

Jianlong Fu[1], Heliang Zheng[2], Tao Mei[1]
[1]Microsoft Research, Beijing, China
[2]University of Science and Technology of China, Hefei, China
[1]{jianf, tmei}@microsoft.com, [2]zhenghl@mail.ustc.edu.cn

Abstract

Recognizing fine-grained categories (e.g., bird species) is difficult due to the challenges of discriminative region localization and fine-grained feature learning. Existing approaches predominantly solve these challenges independently, while neglecting the fact that region detection and fine-grained feature learning are mutually correlated and thus can reinforce each other. In this paper, we propose a novel recurrent attention convolutional neural network (RA-CNN) which recursively learns discriminative region attention and region-based feature representation at multi-

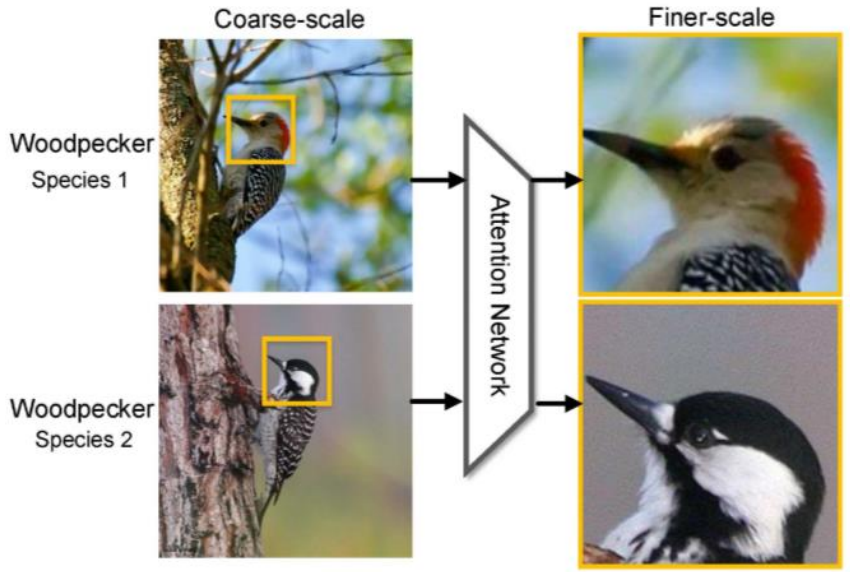
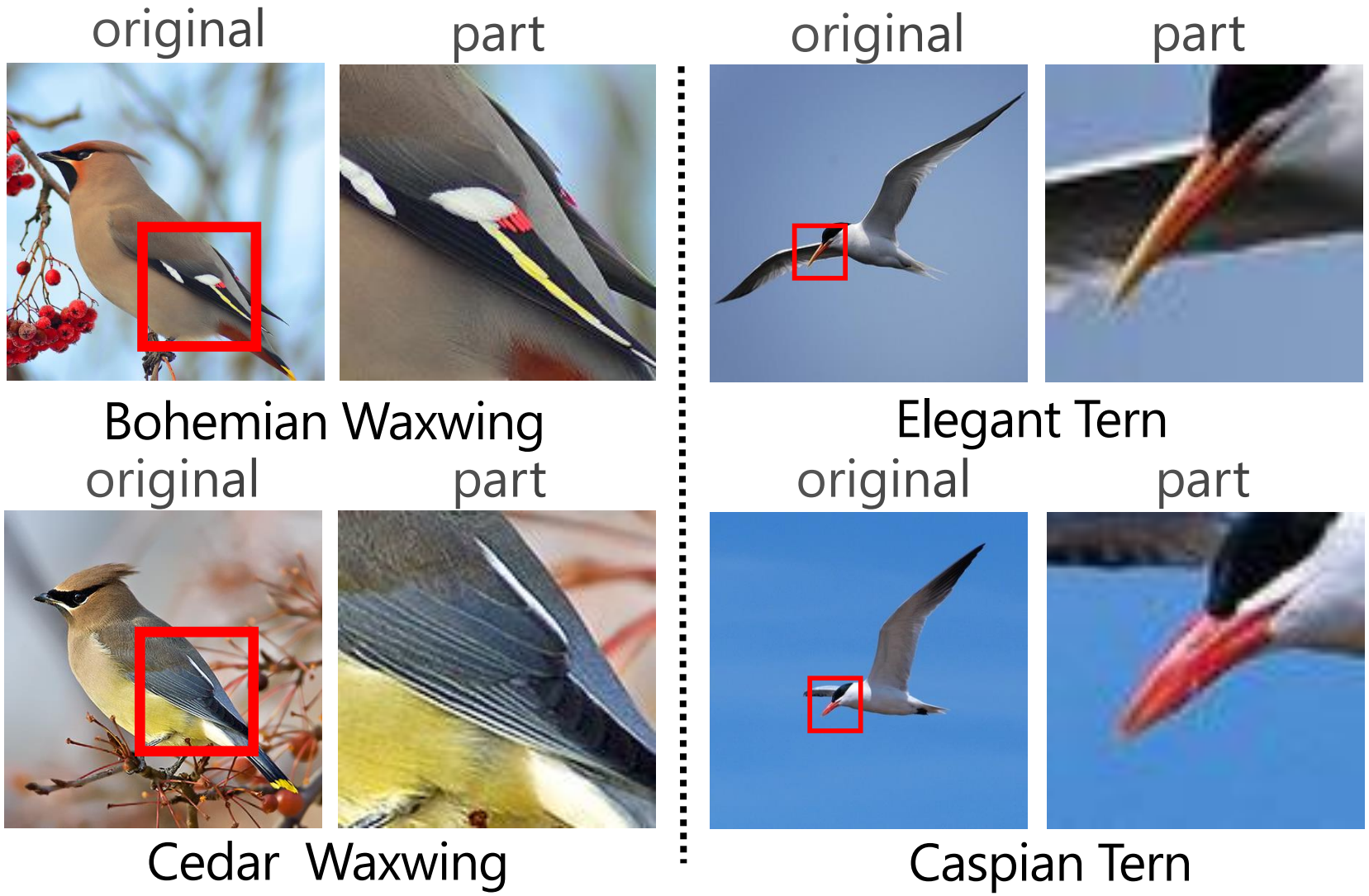


Figure 1. Two bird species of woodpecker. We can observe the

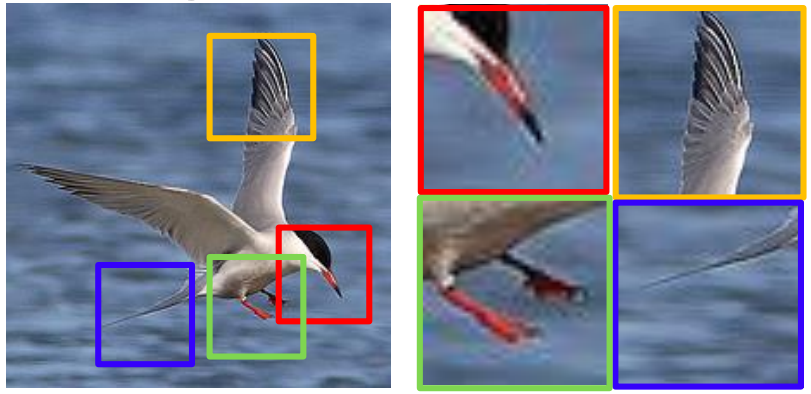
Feature learning from parts (going forward)



Observation: the subtle visual differences can be clearly represented from amplified parts.

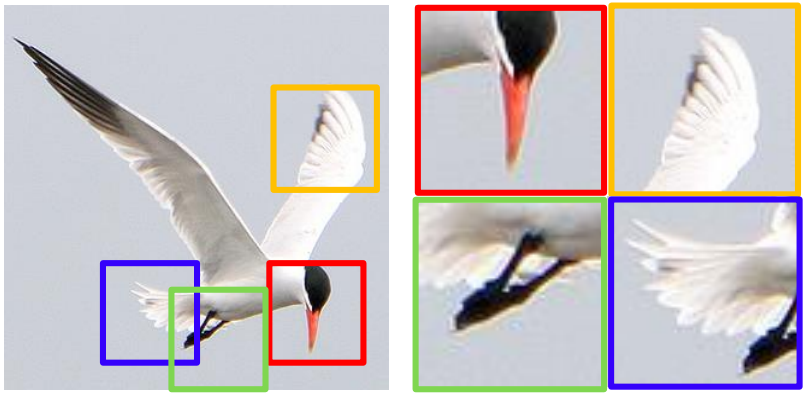
Subtle difference in multiple parts

original 4 parts



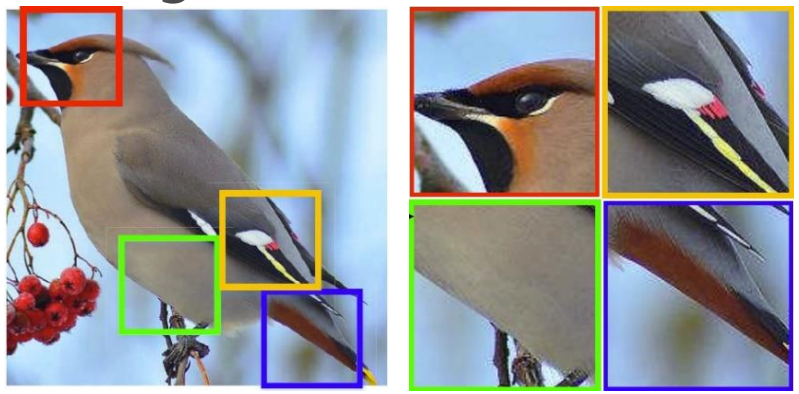
Common Tern

original 4 parts



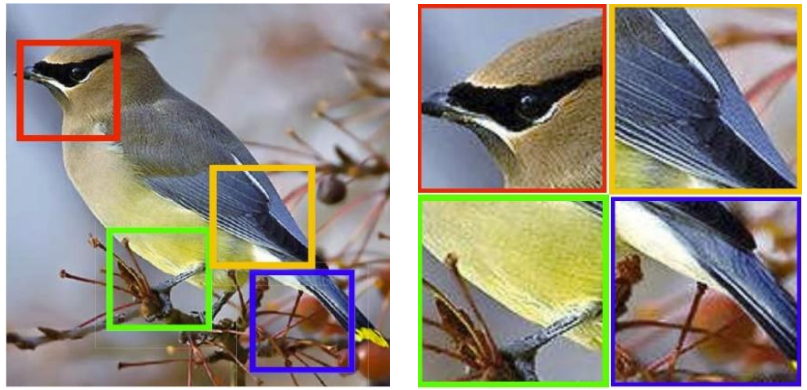
Caspian Tern

original 4 parts



Bohemian Waxwing

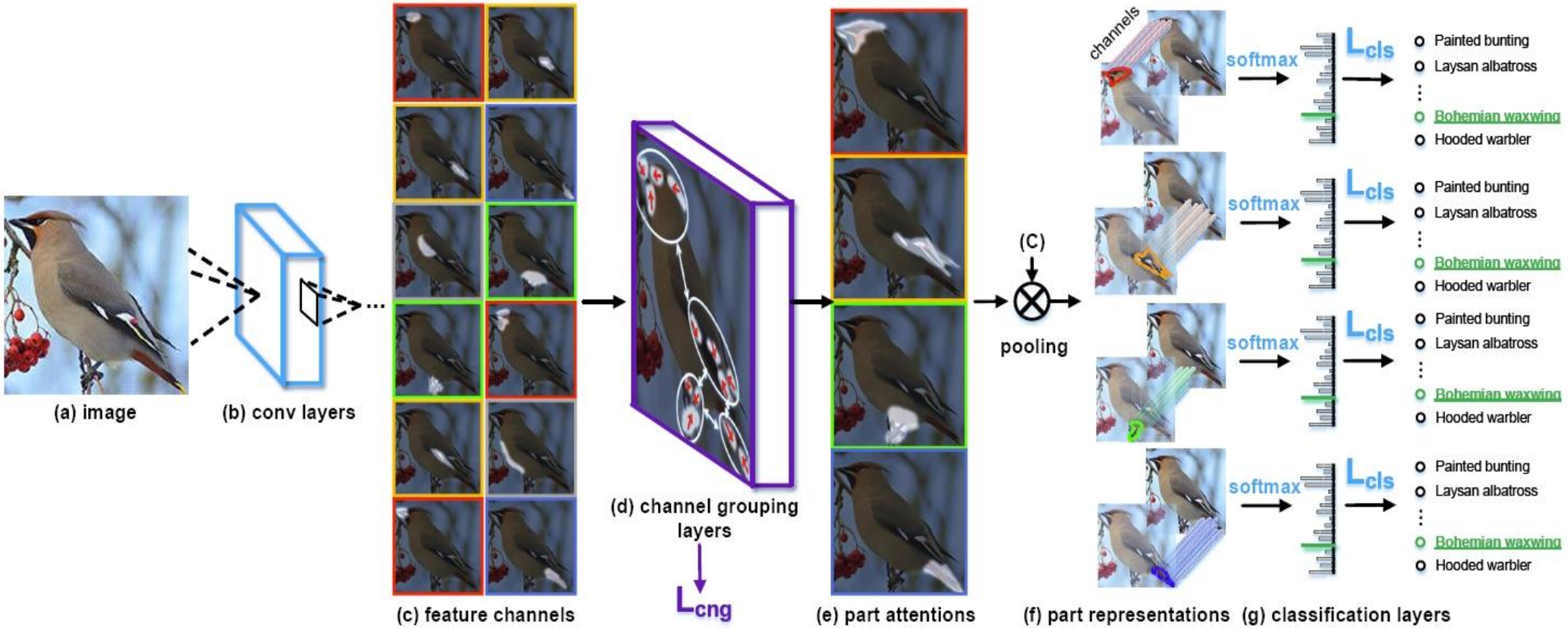
original 4 parts



Cedar Waxwing

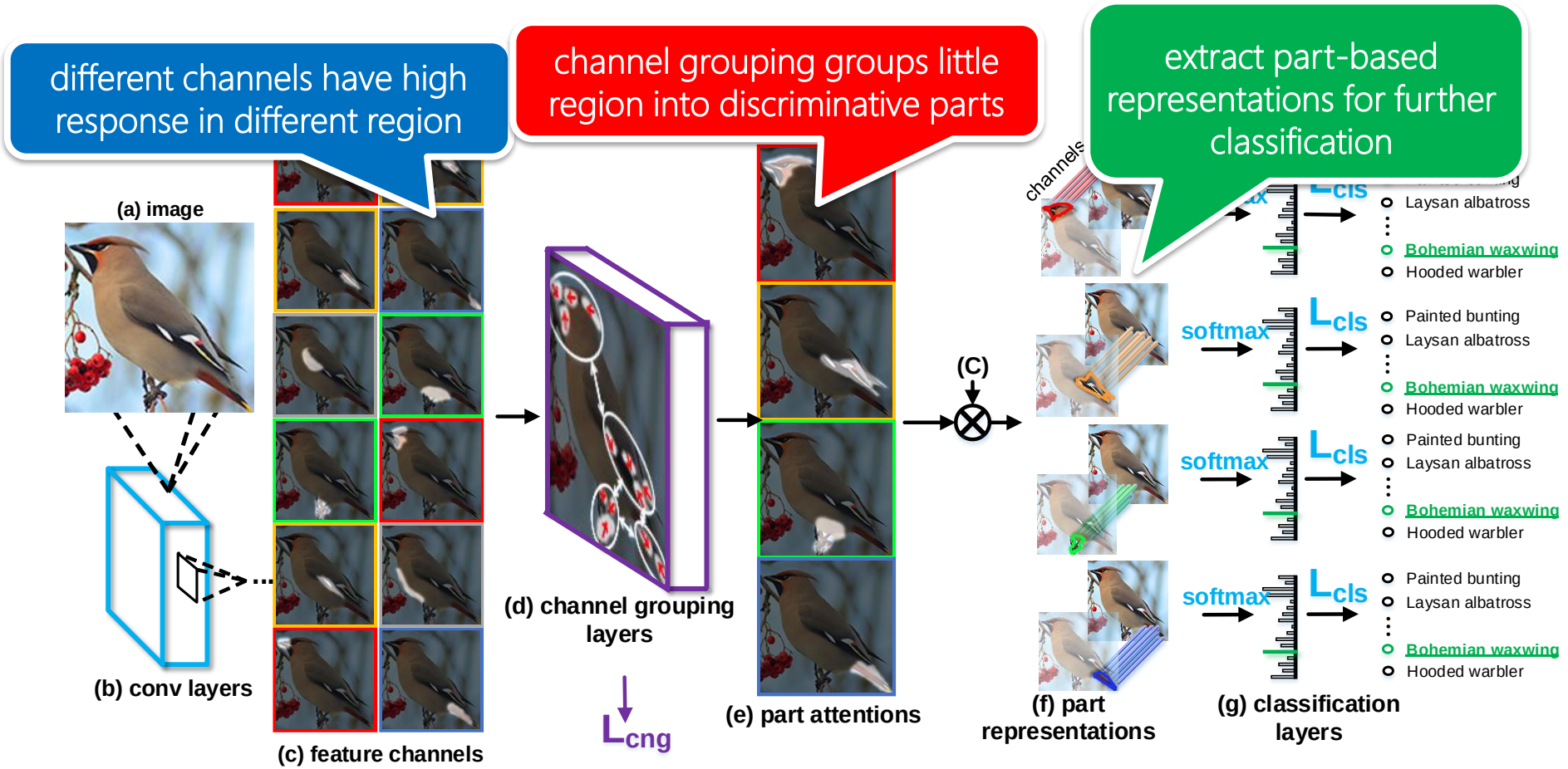
Observation: the subtle visual differences can be clearly represented from multiple parts.

Multi-Attention Network (Fu. ICCV 2017 Oral)



- Multiple attention for each image
- Joint learning of features and parts

Multi-Attention Network (Fu. ICCV 2017 Oral)



$$Dis(M_i) = \sum_{(x,y \in M_i)} m_i(x,y)[\|x - t_x\|^2 + \|y - t_y\|^2]$$

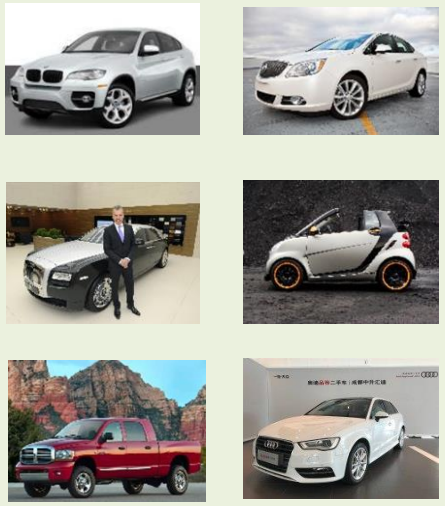
$$Div(M_i) = \sum_{(x,y \in M_i)} m_i(x,y)[max_{k \neq i} m_k(x,y) - mrg]$$

Evaluation

CUB-200-2011

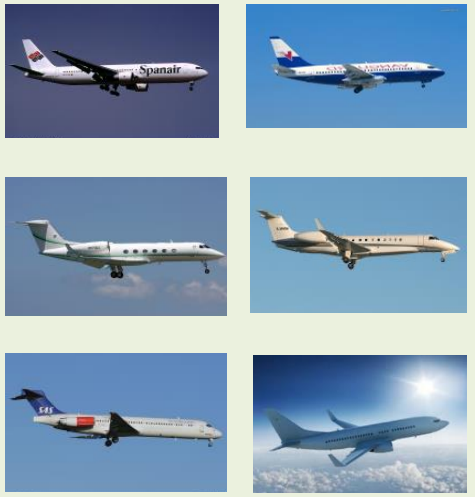
# Category	200
# Training	5,994
# Testing	5,794

Stanford-Cars



# Category	196
# Training	8,144
# Testing	8,041

FGVC-Aircraft



# Category	100
# Training	6,667
# Testing	3,333

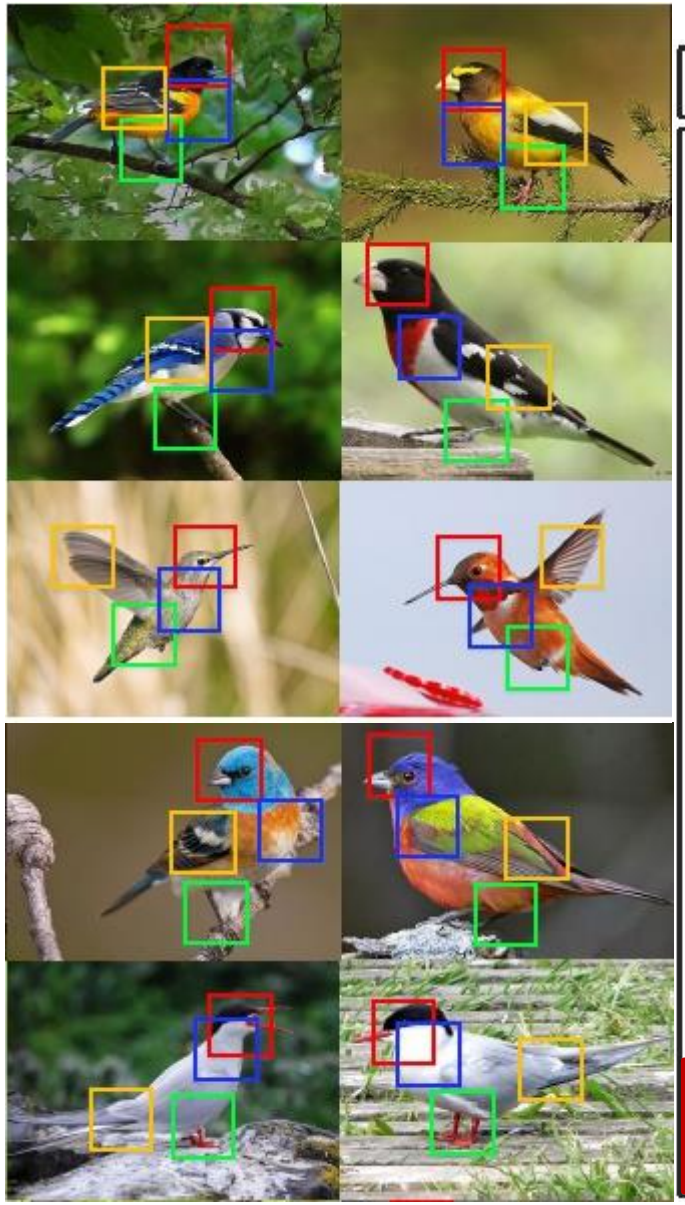
Evaluation

(a) image

(b) part attention by channel grouping

(c) part attention by joint learning

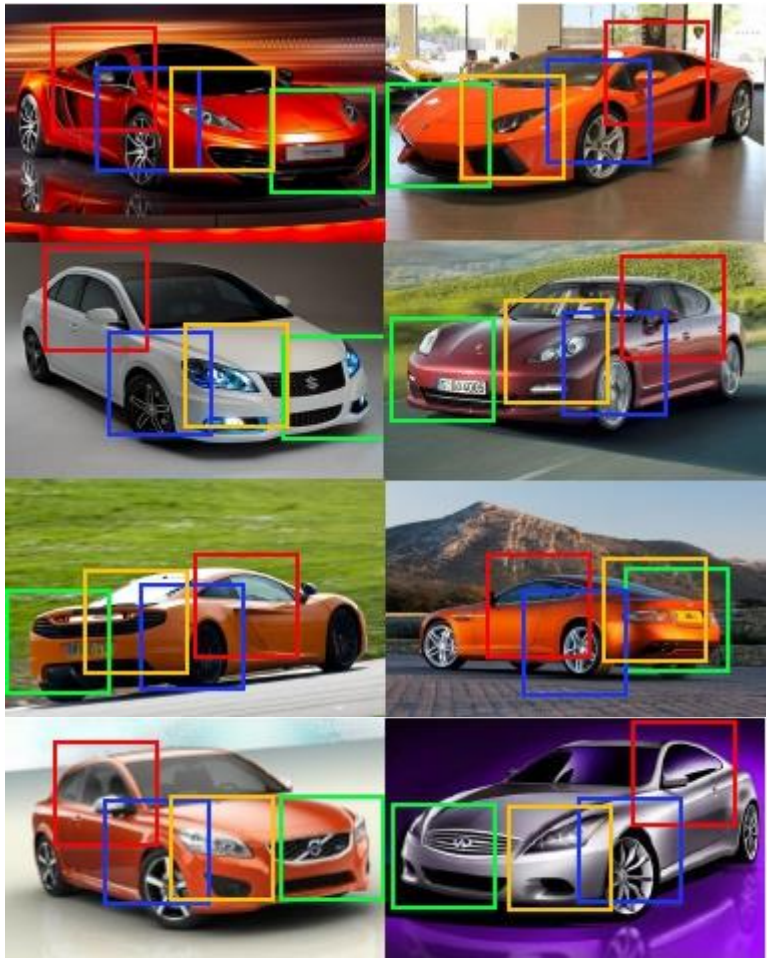
Evaluation



Comparison results on CUB-200-2011 dataset.

Approach	Train Anno.	Accuracy
PN-CNN(AlexNet) [1]	✓	75.7
Part-RCNN(AlexNet) [34]	✓	76.4
PA-CNN [14]	✓	82.8
MG-CNN [27]	✓	83.0
FCAN [18]	✓	84.3
B-CNN (250k-dims) [17]	✓	85.1
Mask-CNN [29]	✓	85.4
TLAN(AlexNet) [31]		77.9
MG-CNN [27]		81.7
FCAN [18]		82.0
B-CNN (250k-dims) [17]		84.1
ST-CNN (Inception net) [10]		84.1
PDFR [35]		84.5
RA-CNN [5]		85.3
MA-CNN (2 parts + object)		85.4
MA-CNN (4 parts + object)		**86.5**

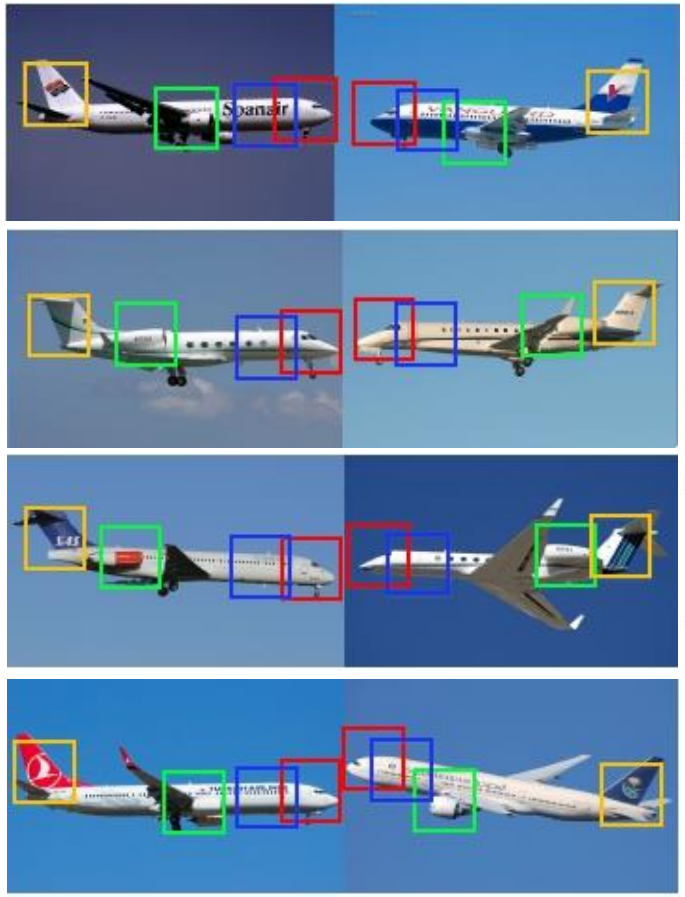
Evaluation



Comparison results on Stanford Cars dataset.

Approach	Train Anno.	Accuracy
R-CNN [6]	✓	88.4
FCAN [18]	✓	91.3
MDTP [28]	✓	92.5
PA-CNN [14]	✓	92.8
FCAN [18]		89.1
B-CNN (250k-dims) [17]		91.3
RA-CNN [5]		92.5
MA-CNN (2 parts + object)		91.7
MA-CNN (4 parts + object)		**92.8**

Evaluation



Comparison results on FGVC-Aircraft dataset.

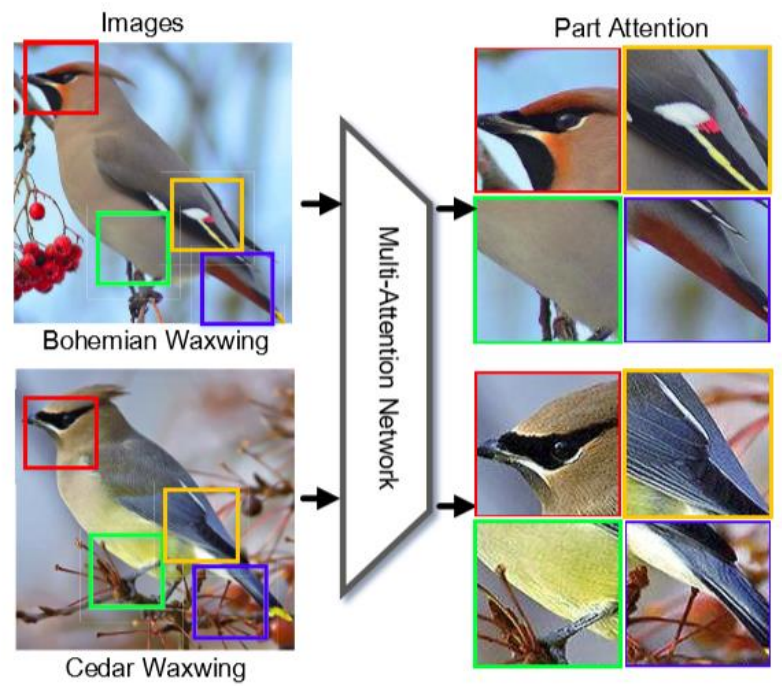
Approach	Train Anno.	Accuracy
MG-CNN [27]	✓	86.6
MDTP [28]	✓	88.4
FV-CNN [7]		81.5
B-CNN (250k-dims)[17]		84.1
RA-CNN [5]		88.2
MA-CNN (2 parts + object)		88.4
MA-CNN (4 parts + object)		**89.9**

Learning Multi-Attention Convolutional Neural Network for Fine-Grained Image Recognition

Heliang Zheng[1*], Jianlong Fu[2], Tao Mei[2], Jiebo Luo[3]
[1]University of Science and Technology of China, Hefei, China
[2]Microsoft Research, Beijing, China
[3]University of Rochester, Rochester, NY
[1]zhenghl@mail.ustc.edu.cn, [2]{jianf, tmei}@microsoft.com, [3]jluo@cs.rochester.edu

Abstract

Recognizing fine-grained categories (e.g., bird species) highly relies on discriminative part localization and part-based fine-grained feature learning. Existing approaches predominantly solve these challenges independently, while neglecting the fact that part localization (e.g., head of a bird) and fine-grained feature learning (e.g., head shape) are mutually correlated. In this paper, we propose a novel part learning approach by a multi-attention convolutional neural network (MA-CNN), where part generation and feature learning can reinforce each other. MA-CNN consists of convolution, channel grouping and part classification sub-networks. The channel grouping network takes as input feature channels from convolutional layers, and



MVA Microsoft Virtual Academy

微软识花产品介绍

Microsoft

微软亚洲研究院最新款智能识别应用

目前已知的花卉种类超过25万种

先进的花卉识别技术

植物专家

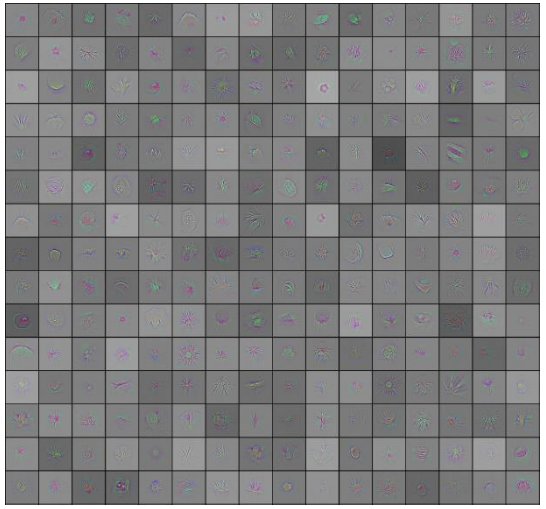
植物爱好者

徒步旅行者

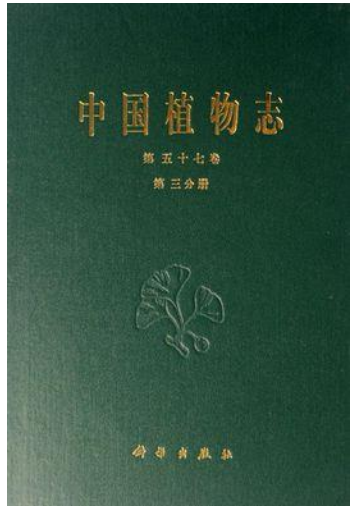
科普工作者及儿童

一键变身"识花达人"

中国科学院植物所
海量精准的花卉图像



微软亚洲研究院先进
的图像表示和分类技术



科学出版社专业
详实的植物介绍

微软识花

app功能

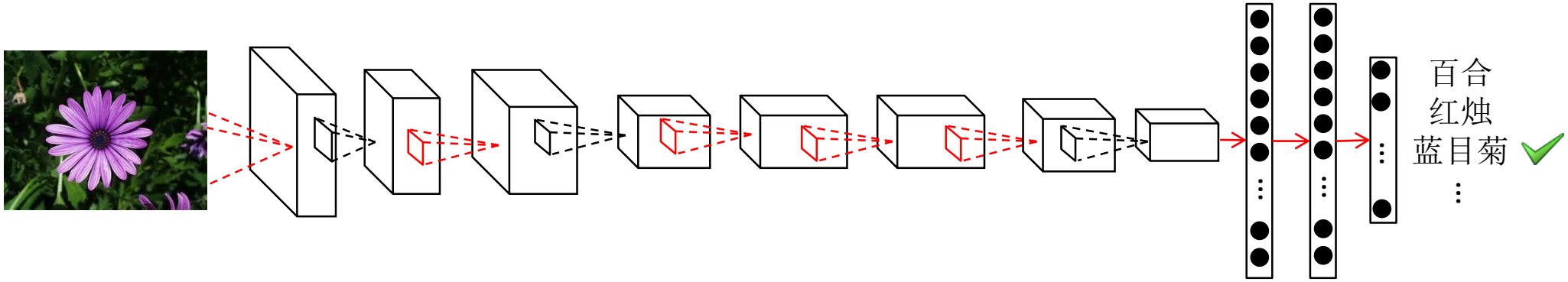
- 拍照识别：可以通过打开本地照片或手机拍照的方式对花卉进行快速识别。目前，可辨识400余种中国常见花卉。
- 离线应用：用户可在郊游、登山等有可能无法连接互联网的情况下使用。
- 寓教于乐：除了识别结果，应用还提供包括专业的植物学特征、富含情感的花语、和生动有趣的散文诗句等不同层级的描述帮助用户了解花卉知识。

产品特点

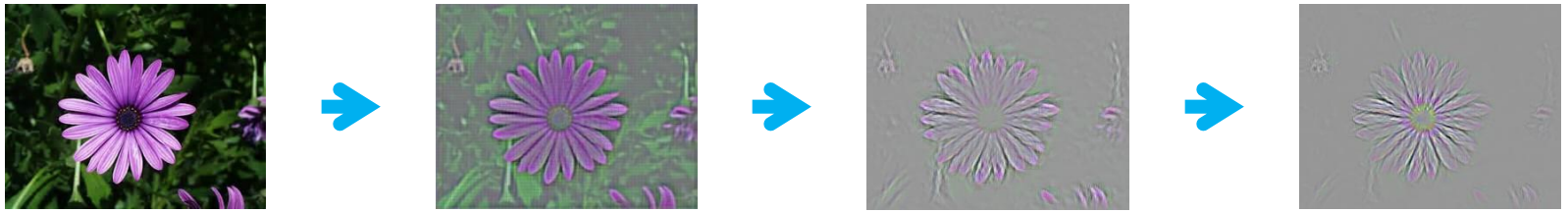
- 精准： 依托于微软亚洲研究院先进的物体识别技术和中国科学院植物研究所海量准确的植物数据，识别准确率可达92%
- 离线：目前手机移动市场上唯一的离线识别应用，使用范围广，计算速度快，节省网络流量
- 专业：与科学出版社合作，物种分布及介绍均来自《中国植物志》，可提供专业的植物学描述
- 友好：用户界面友好， 方便用户比对识别结果，学习植物学知识

深度学习

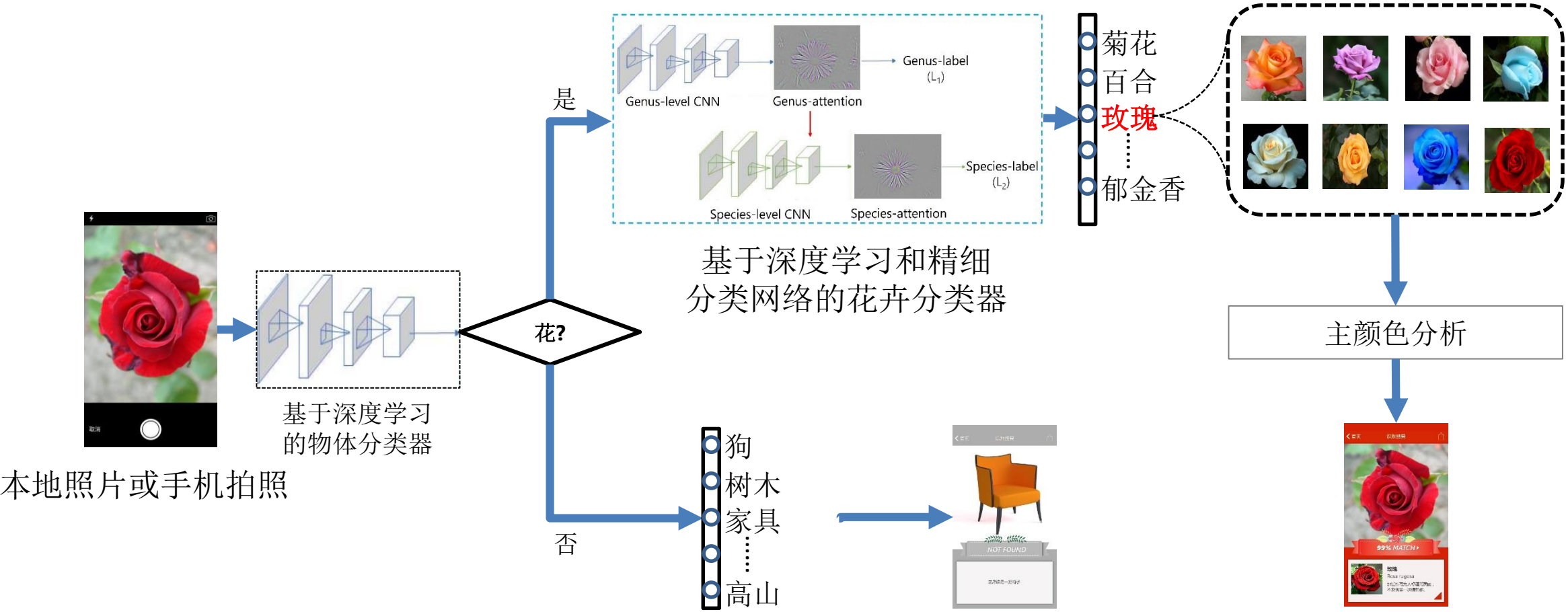
识别技术：卷积神经网络



图像信息提取示意图



识别框架


本地照片或手机拍照
基于深度学习
的物体分类器
花?
是
否
Genus-level CNN
Genus-attention
Genus-label
(L1)
Species-level CNN
Species-attention
Species-label
(L2)
基于深度学习和精细
分类网络的花卉分类器
菊花
百合
玫瑰
郁金香
狗
树木
家具
高山
NOT FOUND
主颜色分析
99% MATCH

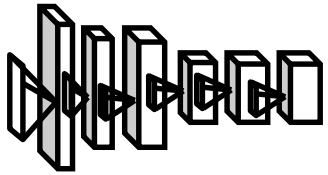
产品特点

- 精准： 依托于微软亚洲研究院先进的物体识别技术和中国科学院植物研究所海量准确的植物数据，识别准确率可达92%
- 离线：目前手机移动市场上唯一的离线识别应用，使用范围广，计算速度快，节省网络流量
- 专业：与科学出版社合作，物种分布及介绍均来自《中国植物志》，可提供专业的植物学描述
- 友好：用户界面友好， 方便用户比对识别结果，学习植物学知识

离线使用

花卉图片

深度学习网络



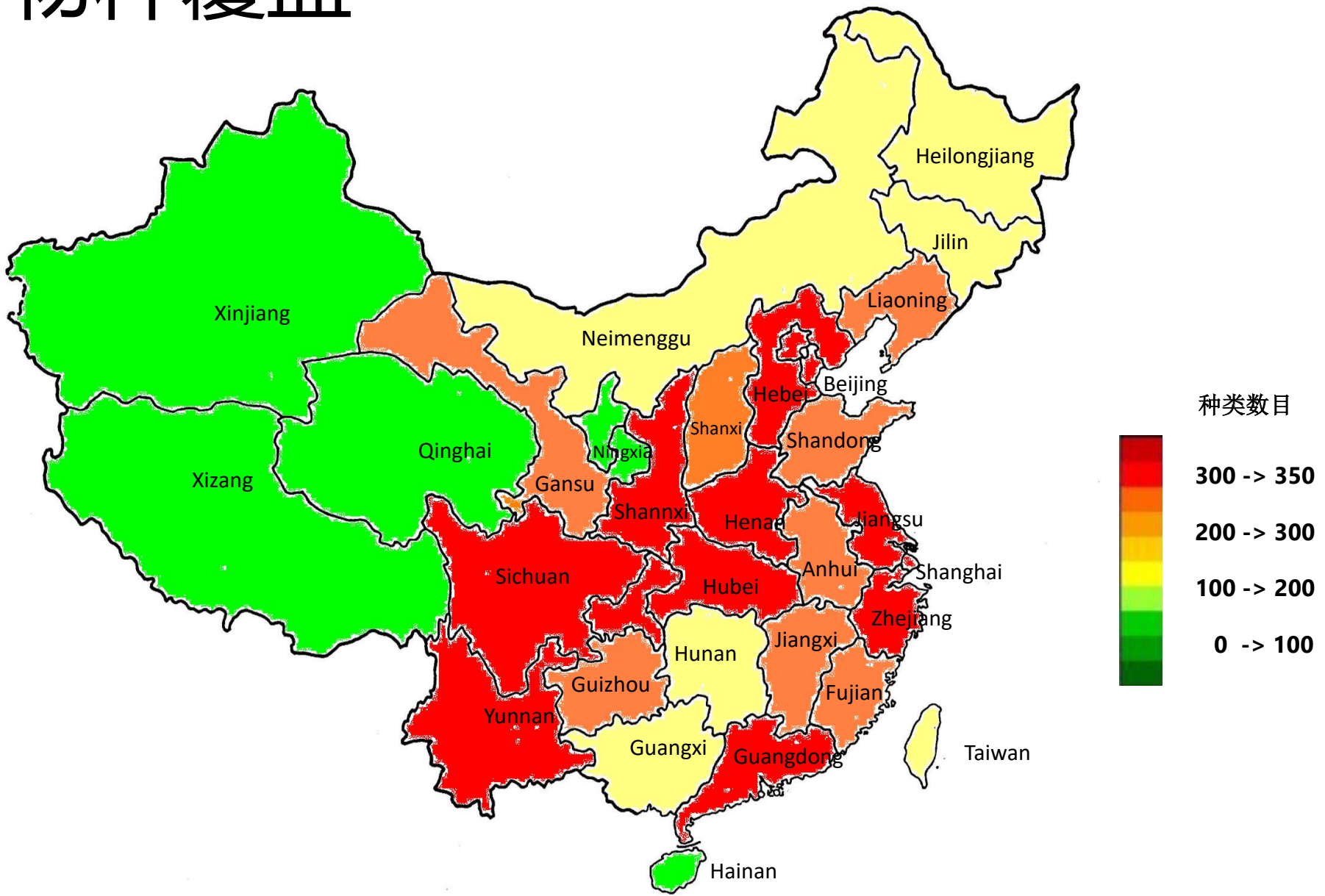
植物学描述

应用大小：100 兆

产品特点

- 精准： 依托于微软亚洲研究院先进的物体识别技术和中国科学院植物研究所海量准确的植物数据，识别准确率可达92%
- 离线：目前手机移动市场上唯一的离线识别应用，使用范围广，计算速度快，节省网络流量
- 专业：与科学出版社合作，物种分布及介绍均来自《中国植物志》，可提供专业的植物学描述
- 友好：用户界面友好， 方便用户比对识别结果，学习植物学知识

花卉物种覆盖


Heilongjiang
Jilin
Liaoning
Xinjiang
Neimenggu
Beijing
Hebei
Shanxi
Shandong
Qinghai
Ningxia
Xizang
Gansu
Shannxi
Henan
Jiangsu
Anhui
Shanghai
Sichuan
Hubei
Zhejiang
Hunan
Jiangxi
Guizhou
Fujian
Yunnan
Guangxi
Guangdong
Taiwan
Hainan
种类数目
300 -> 350
200 -> 300
100 -> 200
0 -> 100

专业植物学描述

牡丹

Moutan Peony

"想和你相约在金玉盘下，良辰美景都有了，我微笑只等你来。"

芍药属 毛茛科

花语--圆满团聚

植物学特征--落叶灌木。茎高达2米；分枝短而粗。叶通常为二回三出复叶，偶尔近枝顶的叶为3小叶；顶生小叶宽卵形，长7-8厘米，宽5.5-7厘米，3裂至中部，裂片不裂或2-3浅裂，表面绿色，无毛，背面淡绿色，有时具白粉，沿叶脉疏生短柔毛或近无毛，小叶柄长1.2-3厘米；侧生小叶狭卵形或长圆状卵形，长4.5-6.5厘米，宽2.5-4厘米，不等2裂至3浅裂或不裂，近无柄；叶柄长5-11厘米，和叶轴均无毛。花单生枝顶，直径10-17厘米；花梗长4-6厘米；苞片5，长椭圆形，大小不等；萼片5，绿色，宽卵形，大小不等；花瓣5，或为重瓣，玫瑰色、红紫色、粉红色至白色，通常变异很大，倒卵形，长5-8厘米，宽4.2-6厘米，顶端呈不规则的波状；雄蕊长1-1.7厘米，花丝紫红色、粉红色，上部白色，长约1.3厘米，花药长圆形，长4毫米；花盘革质，杯状，紫红色，顶端有数个锐齿或裂片，完全包住心皮，在心皮成熟时开裂；心皮5，稀更多，密生柔毛。蓇葖长圆形，密生黄褐色硬毛。花期5月；果期6月。

杜鹃

Rhododendron Simsii

"我为你奔赴一场风花雪月，结局任你书写，我永远属于你。"

杜鹃属 杜鹃花科

花语--永远属于你

植物学特征--落叶灌木，高2（-5）米；分枝多而纤细，密被亮棕褐色扁平糙伏毛。叶革质，常集生枝端，卵形、椭圆状卵形或倒卵形或倒卵形至倒披针形，长1.5-5厘米，宽0.5-3厘米，先端短渐尖，基部楔形或宽楔形，边缘微反卷，具细齿，上面深绿色，疏被糙伏毛，下面淡白色，密被褐色糙伏毛，中脉在上面凹陷，下面凸出；叶柄长2-6毫米，密被亮棕褐色扁平糙伏毛。花芽卵球形，鳞片外面中部以上被糙伏毛，边缘具睫毛。花2-3（-6）朵簇生枝顶；花梗长8毫来，密被亮棕褐色糙伏毛；花萼5深裂，裂片三角状长卵形，长5毫米，被糙伏毛，边缘具睫毛；花冠阔漏斗形，玫瑰色、鲜红色或暗红色，长3.5-4厘米，宽1.5-2厘米，裂片5，倒卵形，长2.5-3厘米，上部裂片具深红色斑点；雄蕊10，长约与花冠相等，花丝线状，中部以下被微柔毛；子房卵球形，10室，密被亮棕褐色糙伏毛，花柱伸出花冠外，无毛。蒴果卵球形，长达1厘米，密被糙伏毛；花萼宿存。花期4-5月，果期6-8月。

海棠花

Asiatic Apple

"到不了的都叫做远方，回不去的都叫做家乡。"

苹果属 蔷薇科

花语--游子思乡

植物学特征--乔木，高可达8米；小枝粗壮，圆柱形，幼时具短柔毛，逐渐脱落，老时红褐色或紫褐色，无毛；冬芽卵形，先端渐尖，微被柔毛，紫褐色，有数枚外露鳞片。叶片椭圆形至长椭圆形，长5-8厘米，宽2-3厘米，先端短渐尖或圆钝，基部宽楔形或近圆形，边缘有紧贴细锯齿，有时部分近于全缘，幼嫩时上下两面具稀疏短柔毛，以后脱落，老叶无毛；叶柄长1.5-2厘米，具短柔毛；托叶膜质，窄披针形，先端渐尖，全缘，内面具长柔毛。花序近伞形，有花4-6朵，花梗长2-3厘米，具柔毛；苞片膜质，披针形，早落；花直径4-5厘米；萼筒外面无毛或有白色绒毛；萼片三角卵形，先端急尖，全缘，外面无毛或偶有稀疏绒毛，内面密被白色绒毛，萼片比萼筒稍短；花瓣卵形，长2-2.5厘米，宽1.5-2厘米，基部有短爪，白色，在芽中呈粉红色；雄蕊20-25，花丝长短不等，长约花瓣之半；花柱5，稀4，基部有白色绒色，比雄蕊稍长。果实近球形，直径2厘米，黄色，萼片宿存，基部不下陷，梗洼隆起；果梗细长，先端肥厚，长3-4厘米。花期4-5月，果期8-9月。

产品特点

- 精准： 依托于微软亚洲研究院先进的物体识别技术和中国科学院植物研究所海量准确的植物数据，识别准确率可达92%
- 离线：目前手机移动市场上唯一的离线识别应用，使用范围广，计算速度快，节省网络流量
- 专业：与科学出版社合作，物种分布及介绍均来自《中国植物志》，可提供专业的植物学描述
- 友好：用户界面友好， 方便用户比对识别结果，学习植物学知识

用户界面

拍照识花

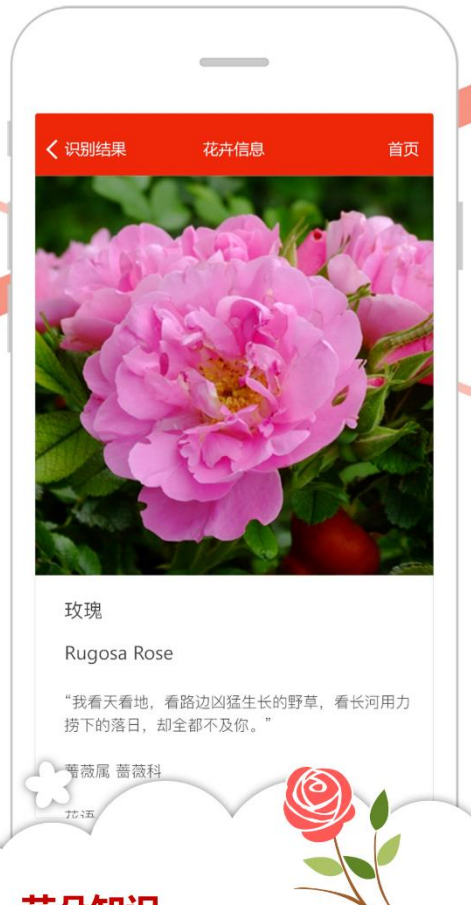
一键拍照识花 • 超方便

手动缩放

手动缩放花朵 • 更精准

匹配结果

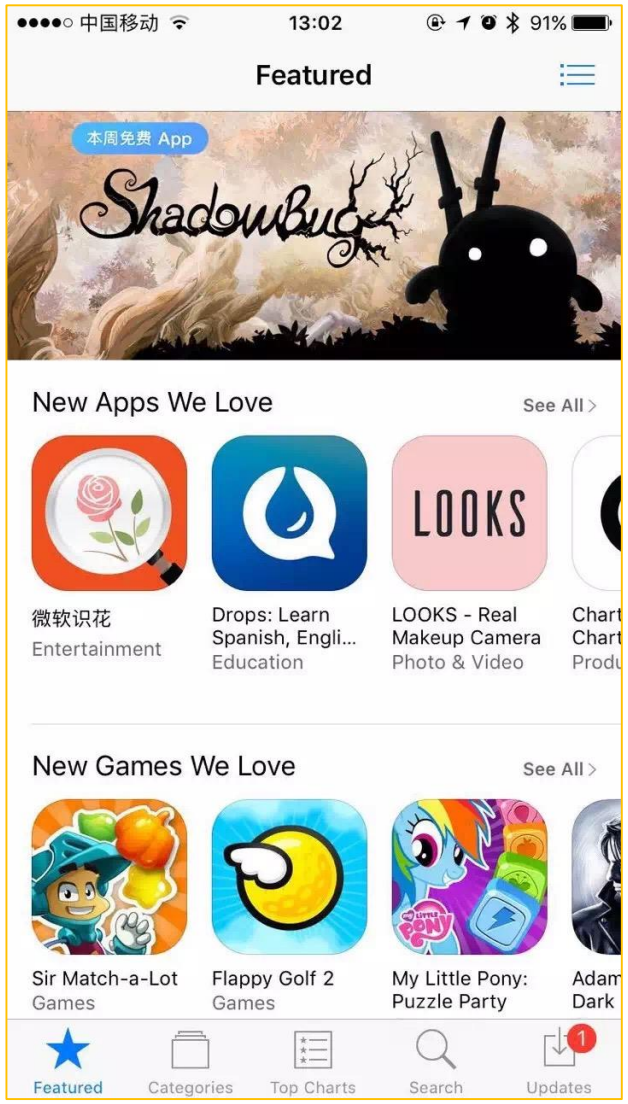
百分制匹配结果 • 最专业



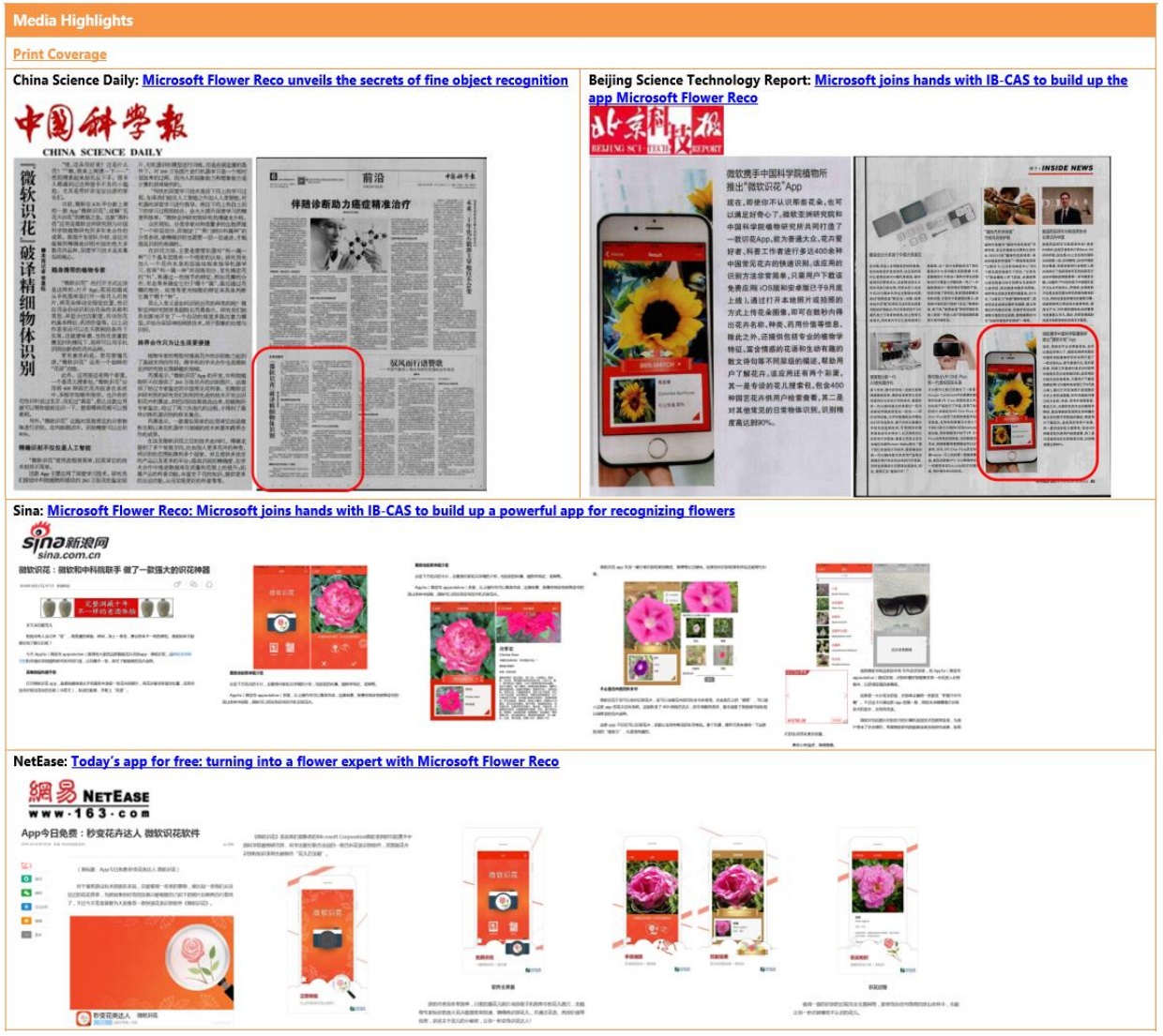
花朵知识

有趣的花朵介绍 • 涨知识

影响力



Apple store front page



Media highlights

Weibo Comments

jianf@microsoft.com

Image creation & poem generation

Aesthetic Layout [ACM TOMM17 Best Paper]

The First Poetry Book Created by AI