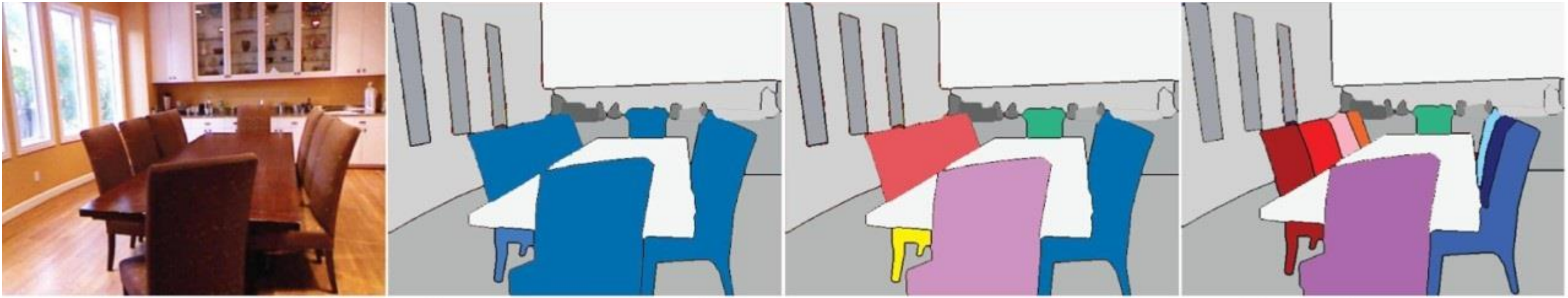


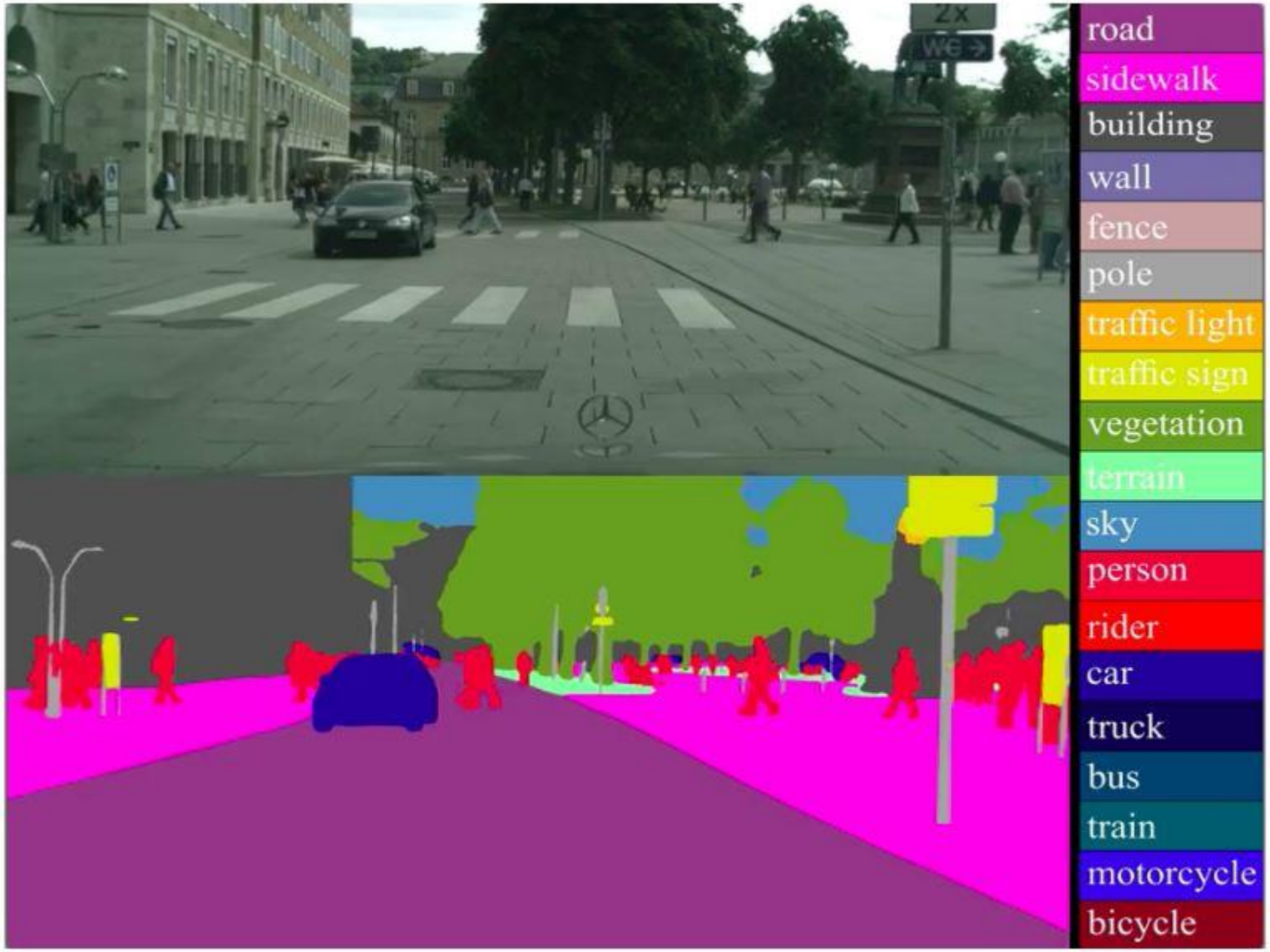
深度学习图像分割综述

邢琛聪

2019/1/14

图像分割示例

可以将图像分割看作是对图像每一个像素的分类问题



目录

- 特征提取网络
- 图像分割方法
- 评价体系
- 数据集汇总
- 发展趋势

监督学习方法

- Patch Based classification
- FCN
  - Encoder-Decoder
  - Dilated Convolution
  - Feature Ensampling
  - Large Kernel

无/弱监督学习方法

- Weakly- and Semi-Supervised Learning

目录

- **特征提取网络**
- 图像分割方法
- 评价体系
- 数据集汇总
- 发展趋势

特征提取网络

Backbone networks

- AlexNet[1]
- VGGNet[2]
- GoogleNet[3]
- ResNet[4]
- DenseNet[5]
- XceptionNet[6]

---

[1]: Krizhevsky A et al “Imagenet classification with deep convolutional neural networks”, NIPS, 2012.

[2]: Simonyan K, et al “A. Very deep convolutional networks for large-scale image recognition” .ICLR, 2015

[3]: Szegedy C, et al. “Going deeper with convolutions”. CVPR 2015.

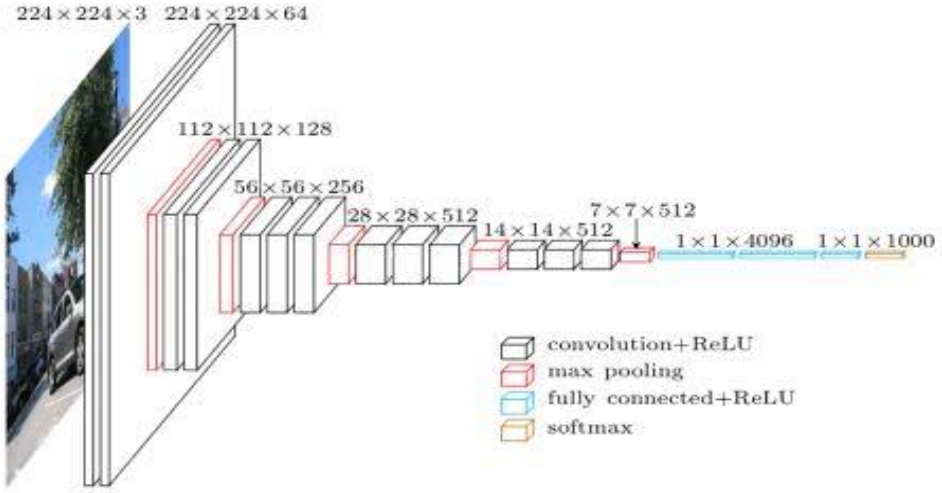
[4]: He K, et al. “Deep residual learning for image recognition” CVPR, 2016.

[5]: Huang G, et al. “Densely connected convolutional networks” CVPR 2017

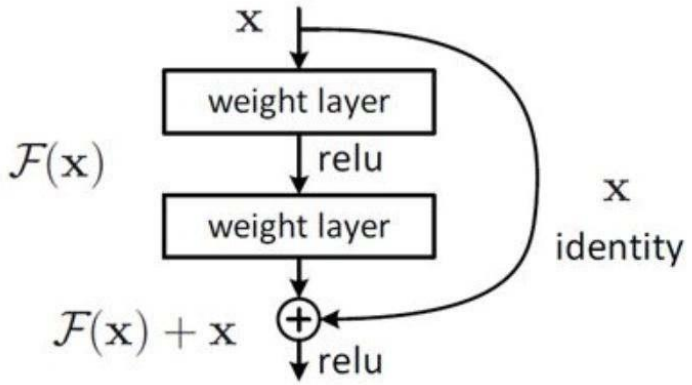
[6]: Chollet, Francois. ”Xception: Deep Learning With Depthwise Separable Convolutions.” CVPR, 2017

特征提取网络

VGGNet



ResNet



- Use skip connection to avoid gradient vanishing
- Learn residuals through residual blocks

目录

- 特征提取网络
- **图像分割方法** ⟶
- 评价体系
- 数据集汇总
- 发展趋势

监督学习方法

- Patch Based classification
- FCN
  - Encoder-Decoder
  - Dilated Convolution
  - Feature Ensampling
  - Large Kernel

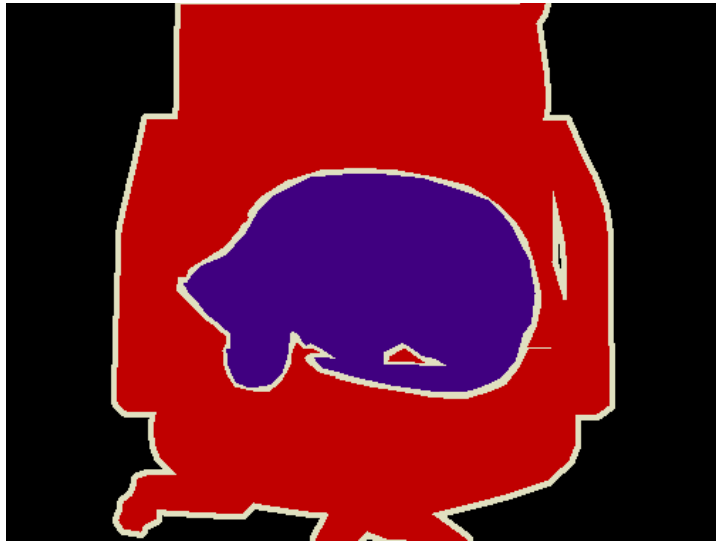
无/弱监督学习方法

- Weakly- and Semi-Supervised Learning

监督学习方法

监督学习方法需要大量逐像素标注图片作为训练集，通常这些标注图片是十分昂贵的，以下是常见的基于监督学习的图像分割模型

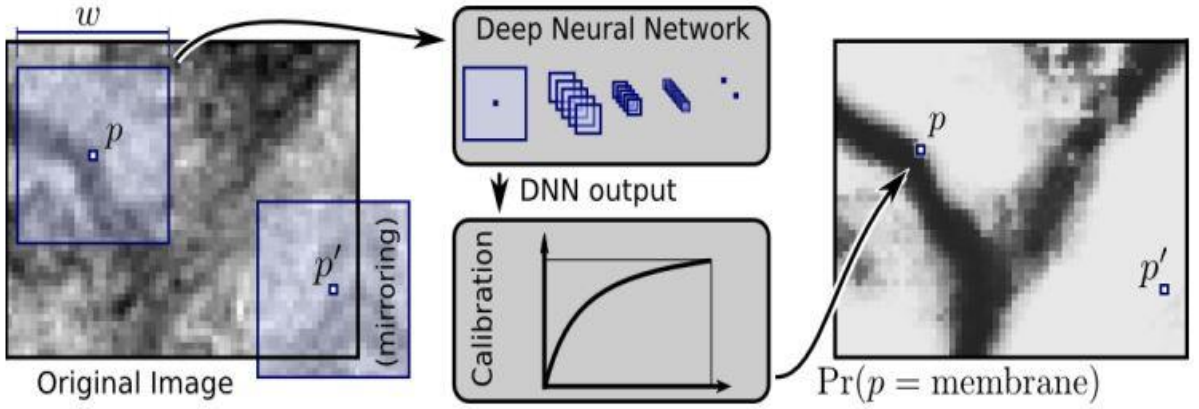
- Patch Based Classification
- FCN
  - Encoder-Decoder
  - Dilated Convolution
  - Feature Ensampling
  - Large Kernel



Patch Based Classification

将图像分割问题化为逐像素的图像分类问题

对于输入图像的每一个像素p，使用神经网络分类器预测其属于前景的概率。网络的输入是以p为中心，窗口宽度为w的正方形图像块。当p靠近边界时，采用镜像扩展的方式填补边缘



优点:

可以在小数据集上训练

模型与图像分辨率无关

缺点:

计算过程包含大量冗余

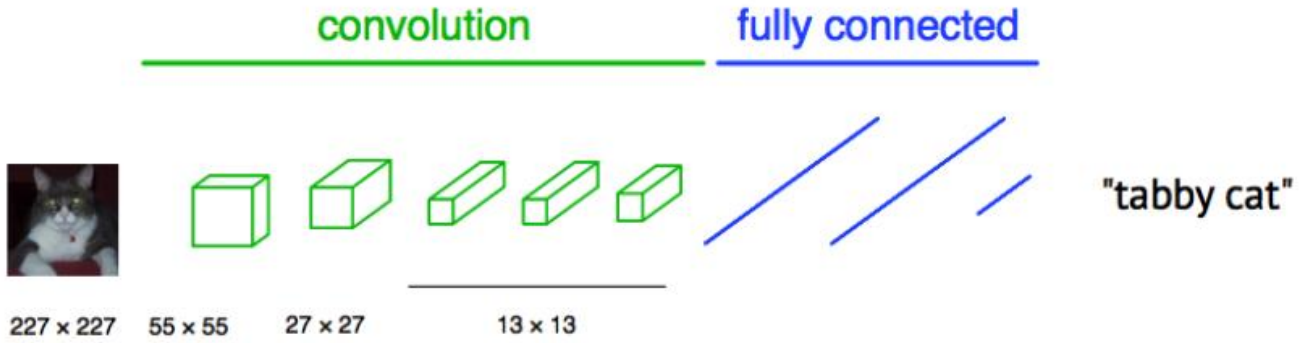
窗口宽度w的选择会影响模型性能

Layer	Type	Maps and neurons	Kernel size
0	input	1 map of 95x95 neurons	
1	convolutional	48 maps of 92x92 neurons	4x4
2	max pooling	48 maps of 46x46 neurons	2x2
3	convolutional	48 maps of 42x42 neurons	5x5
4	max pooling	48 maps of 21x21 neurons	2x2
5	convolutional	48 maps of 18x18 neurons	4x4
6	max pooling	48 maps of 9x9 neurons	2x2
7	convolutional	48 maps of 6x6 neurons	4x4
8	max pooling	48 maps of 3x3 neurons	2x2
9	fully connected	200 neurons	1x1
10	fully connected	2 neurons	1x1

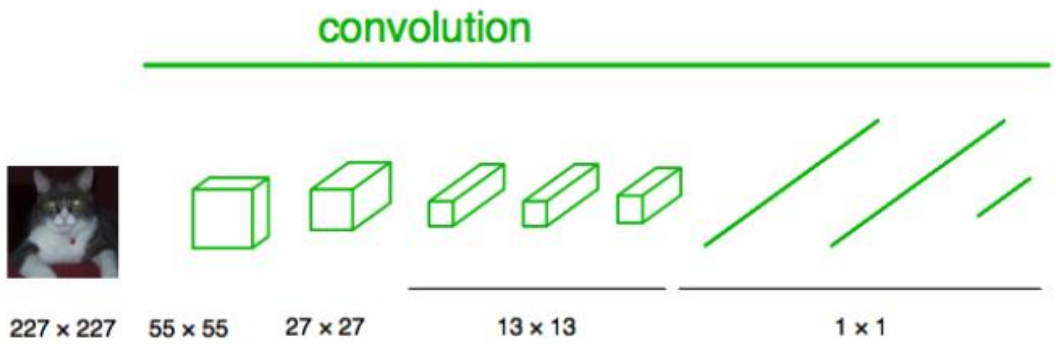
Ciresan, Dan C., et al. “Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images.” *NIPS. 2012*

从图像分类到图像分割

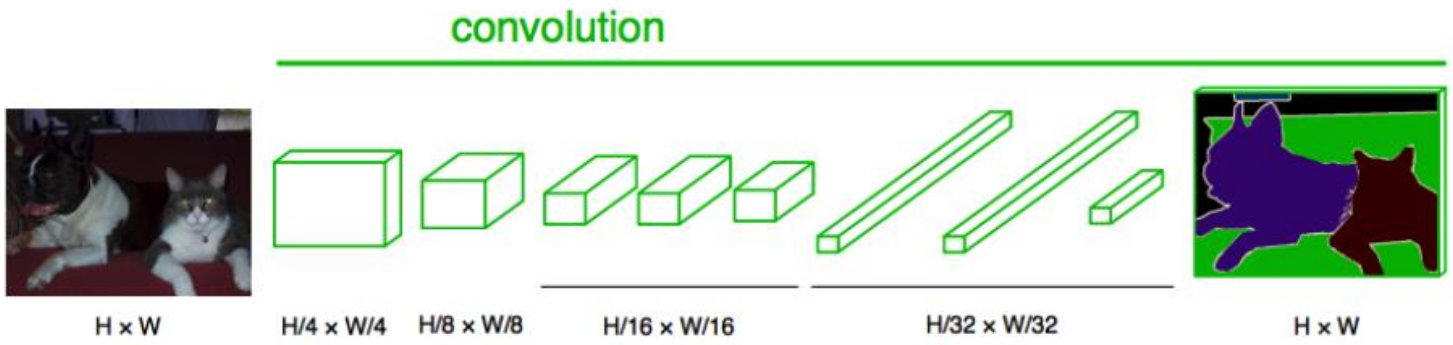
CNN模型可以自动提取特征，但会丢失物体位置、轮廓细节，无法做到精细分割



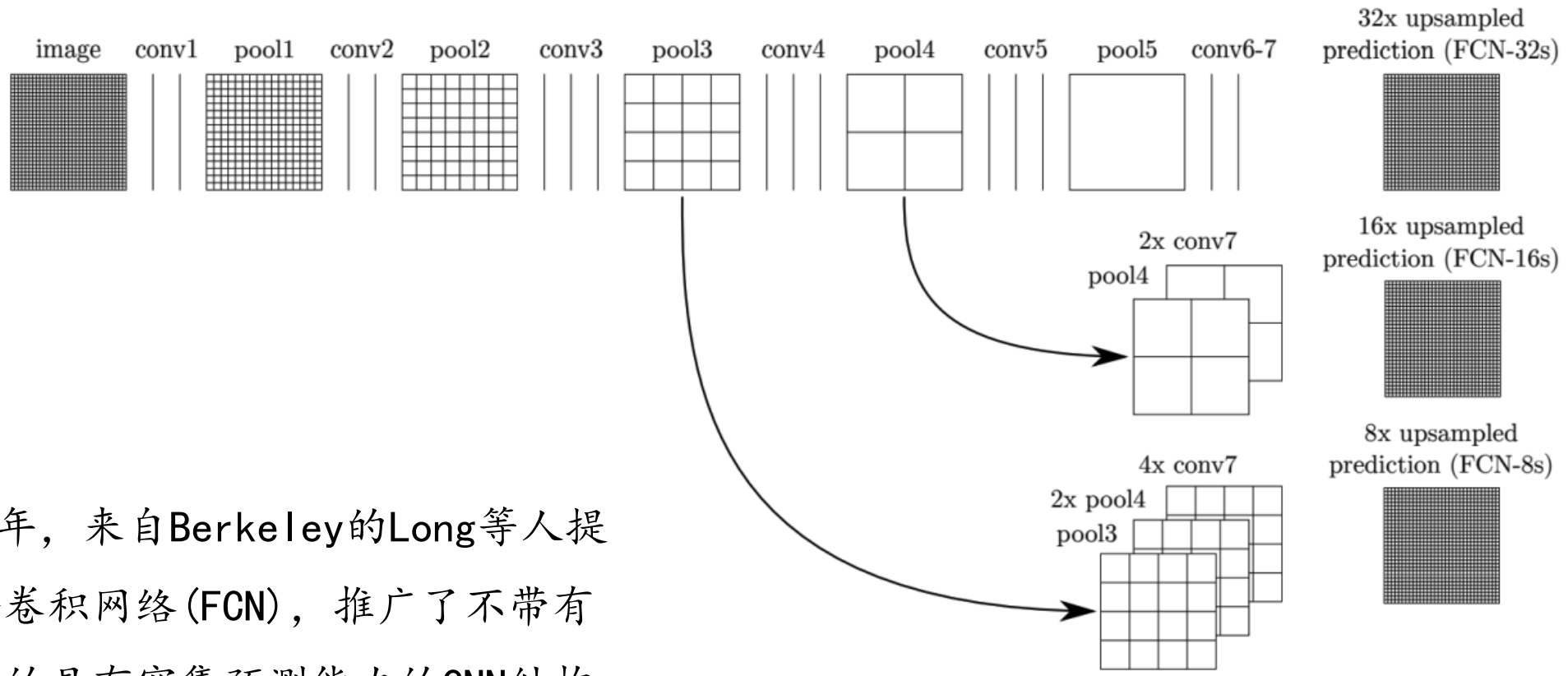
使用卷积层替代全连接层



配合上采样方法，可以得到与原图相同大小的分割结果



Fully Convolutional Network



2014年，来自Berkeley的Long等人提出了完全卷积网络(FCN)，推广了不带有全连接层的具有密集预测能力的CNN结构。

Long, Jonathan, et al. “Fully Convolutional Networks for Semantic Segmentation.” *CVPR 2015.*

Fully Convolutional Network

贡献:

实现了端到端、逐像素分割框架

能够处理不同大小的输入图片

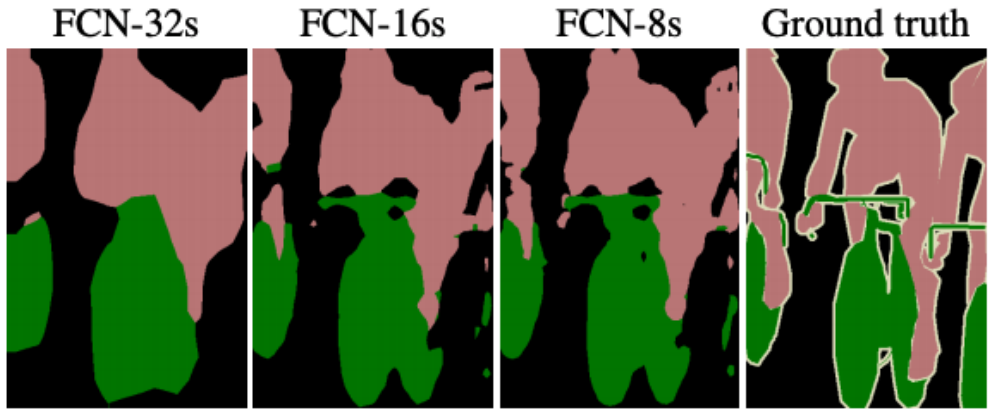
融合不同尺度的特征方法

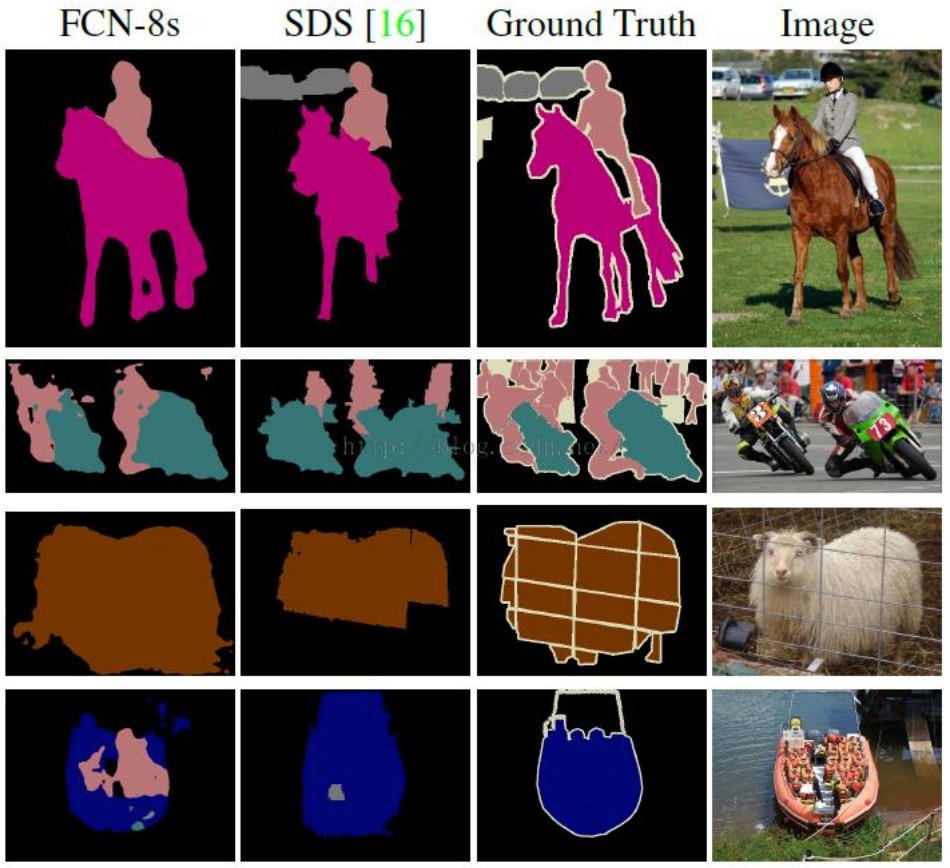
缺陷:

忽略了池化层造成的空间信息丢失

上采样的结果不够精细，对细节不敏感

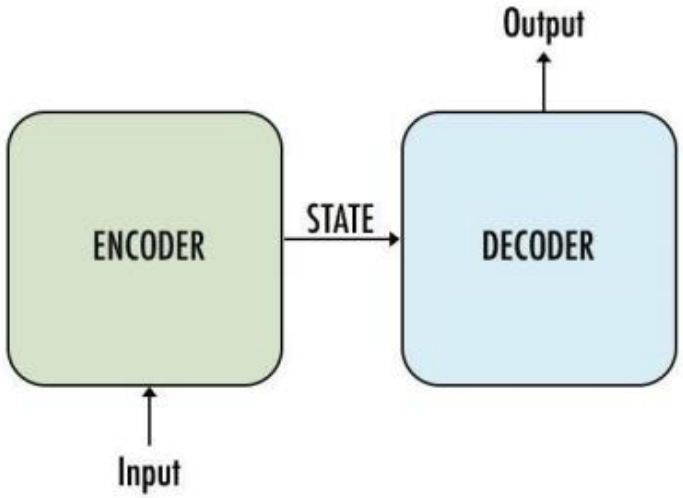
忽略了像素间空间关系，缺乏空间一致性





Encoder-Decoder Methods

Encoder-Decoder 结构



典型模型

- UNet[1]
- SegNet[2]
- DeconvNet[3]

[1]: Ronneberger O, et al. “U-net: Convolutional networks for biomedical image segmentation”, MICCAI, 2015.

[2]: V Badrinarayanan, et al. “Segnet: A deep convolutional encoder-decoder architecture ....”. CVPR. 2015.

[3]: Noh H, et al. ”Learning deconvolution network for semantic segmentation”, ICCV. 2015

U-Net

核心思想

左侧Contraction path抽取特征

增加类别信息

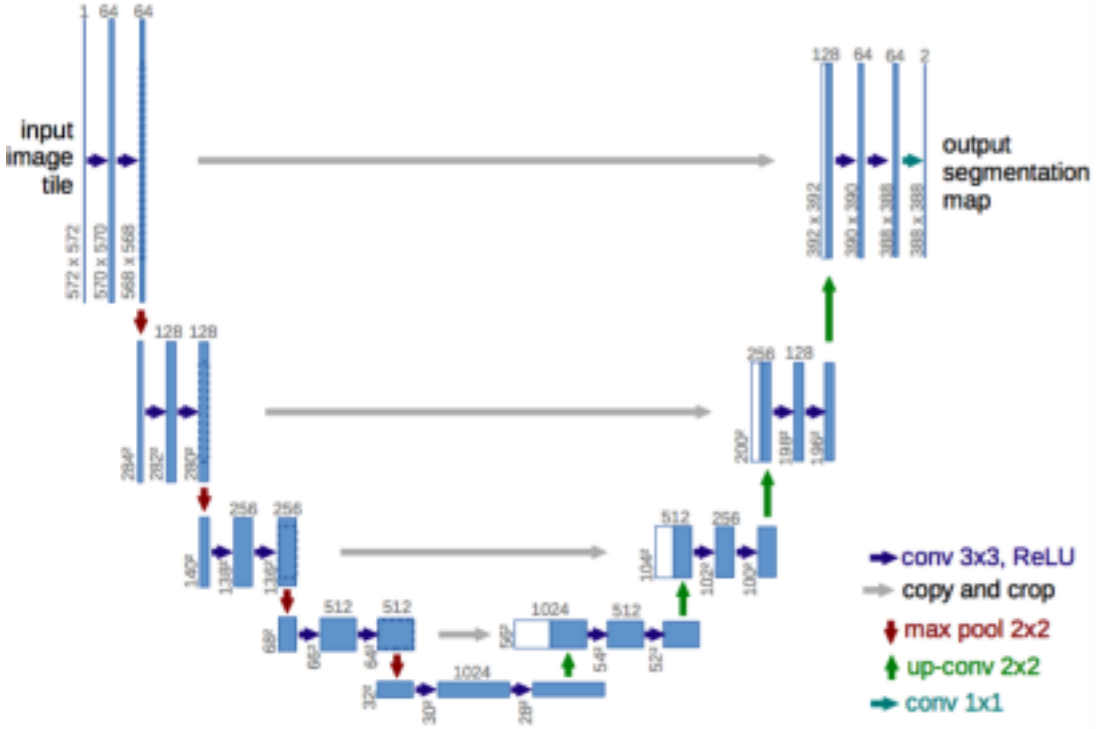
丢失空间位置信息

右侧对称的expansion path实现精准定位

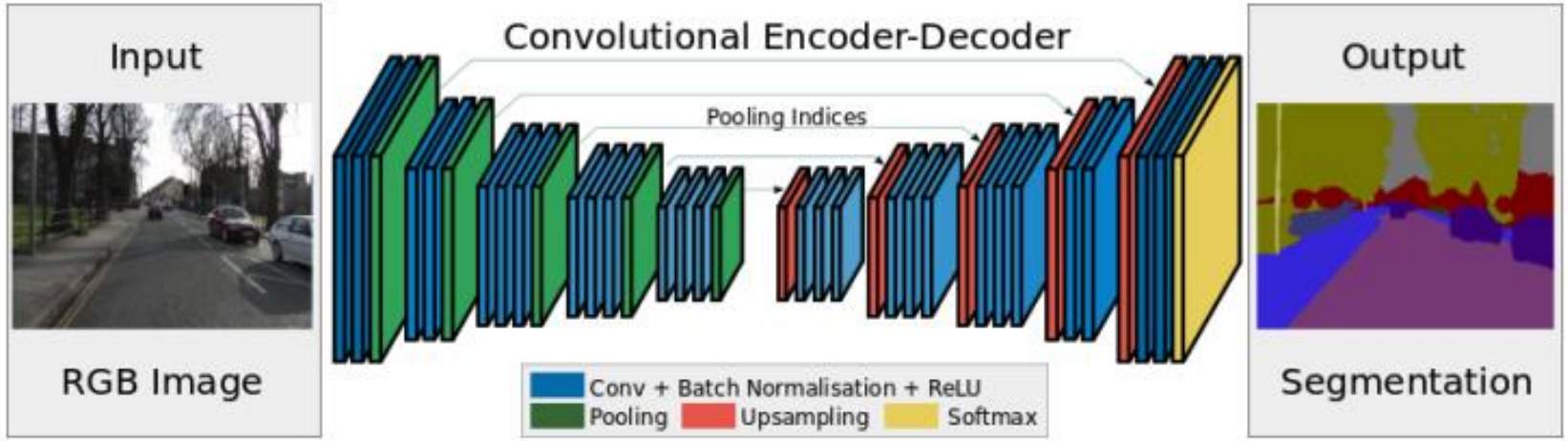
高低层次特征融合

得到高分辨率分割图

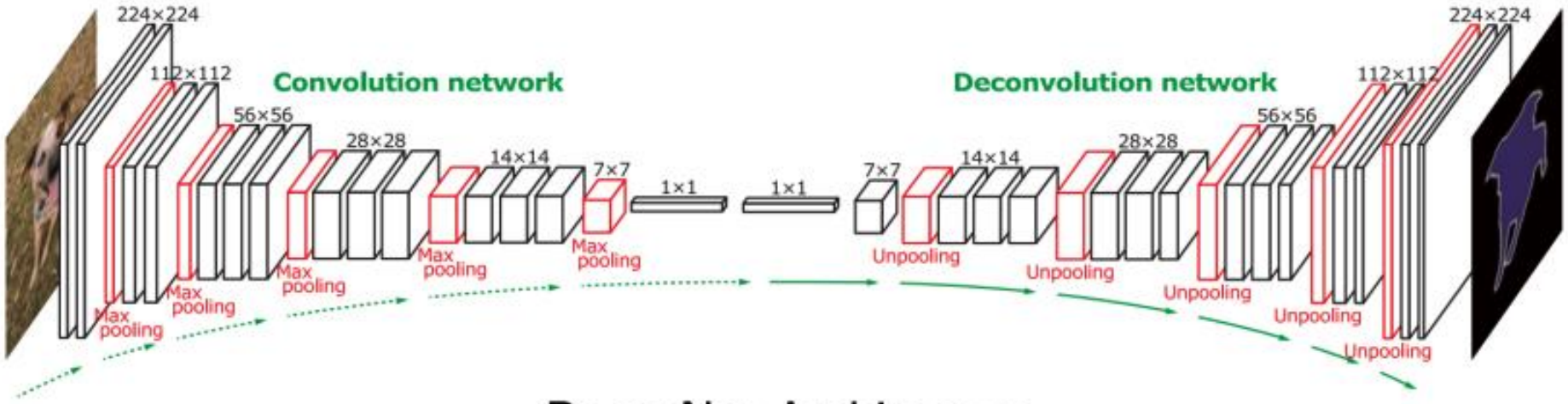
适用于医学图片的增强技巧



SegNet & DeconvNet

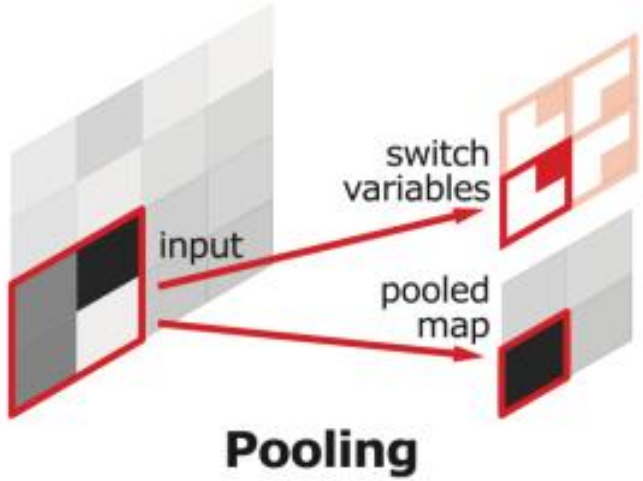


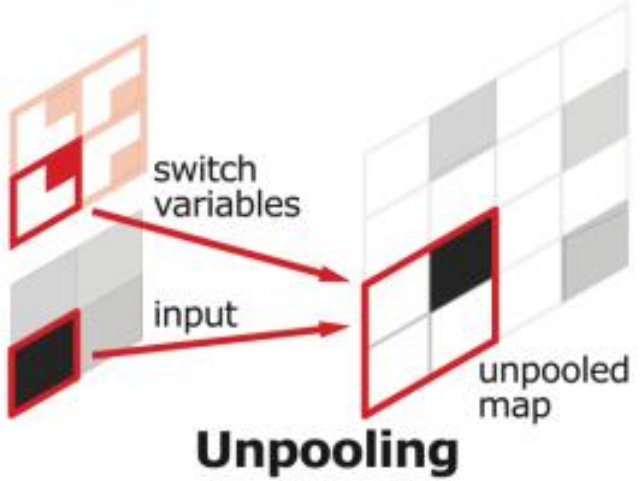
SegNet Architecture

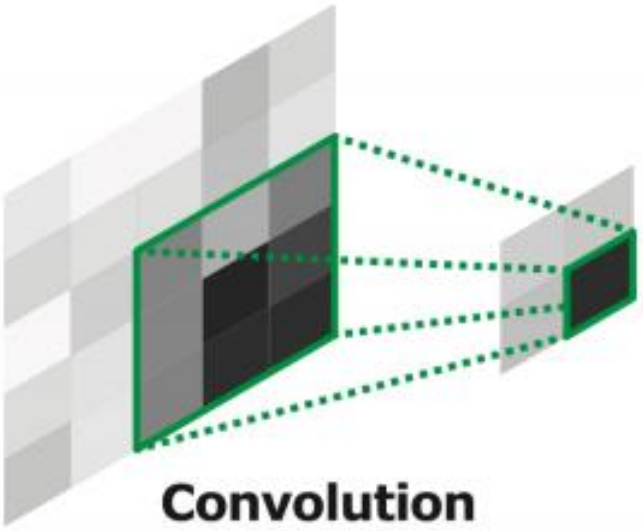


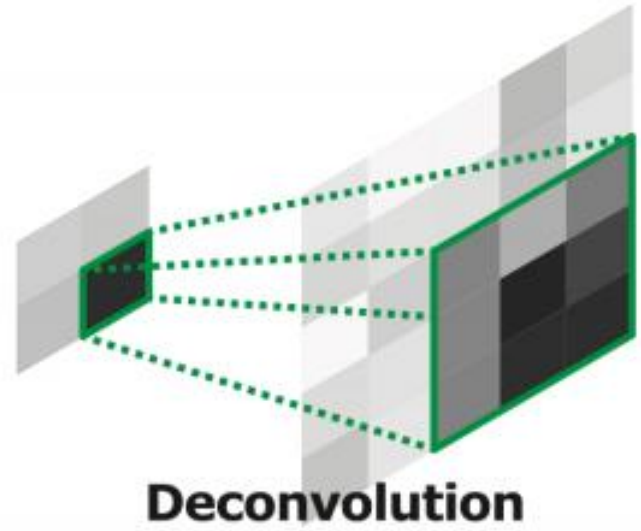
DeconNet Architecture

Unpooling & Deconvolution



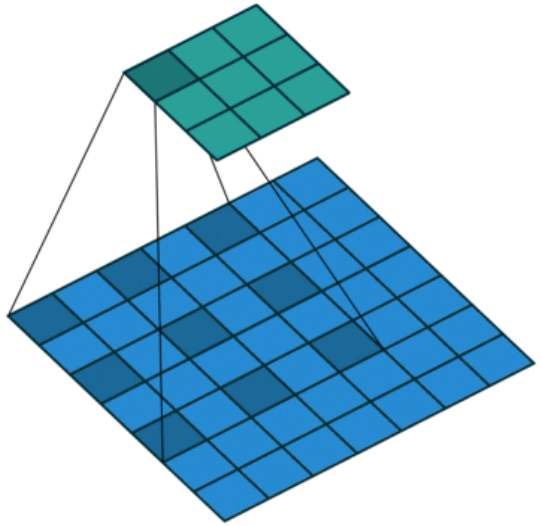






Dilated Convolution

膨胀卷积



典型模型

- DeepLab v2[1]
- DeepLab v3+[2]
- DUC[3]

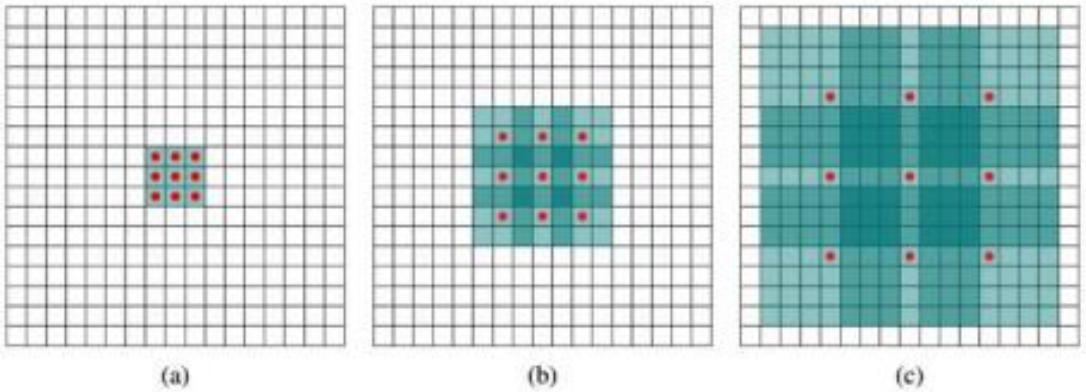
---

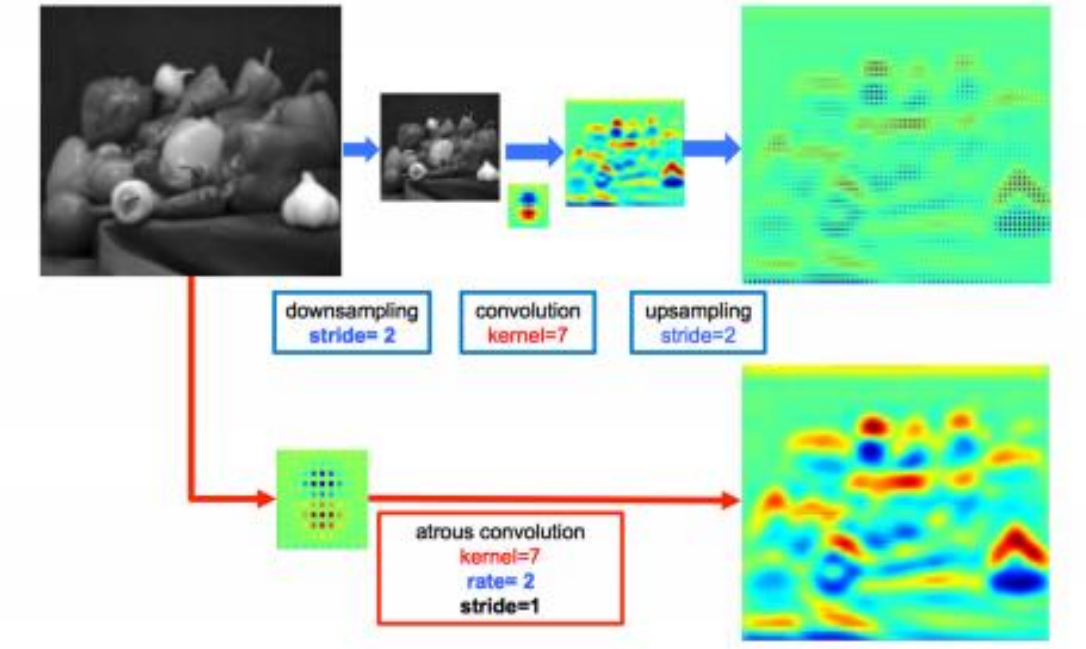
[1]: Liang-Chieh, et al. ”DeepLab: Semantic Image Segmentation with Deep Convolutional ...”, TPAMI, 2017.

[2]: Liang-Chieh, et al. ”Encoder-Decoder with Atrous Separable Convolution for ...”, ECCV,2018.

[3]: dPanqu Wang, et al, ”Understanding Convolution for Semantic Segmentation”, WACV, 2018.

Dilated Convolution





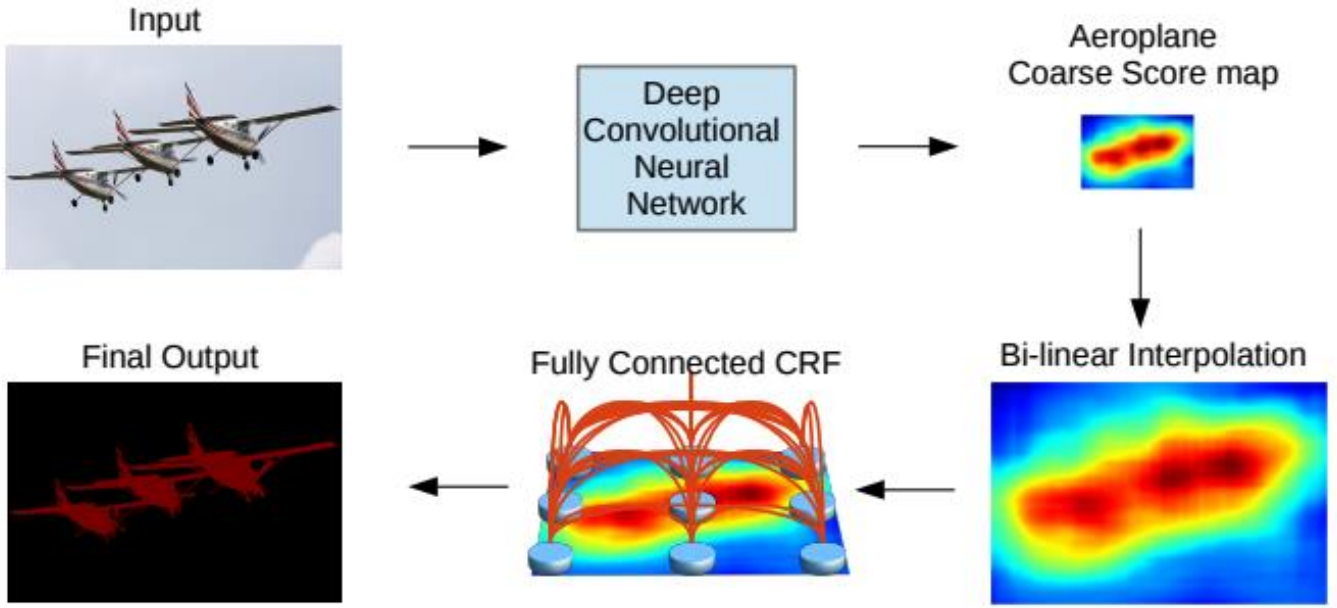
膨胀卷积优势

可以替代池化层，增大卷积器的感受野

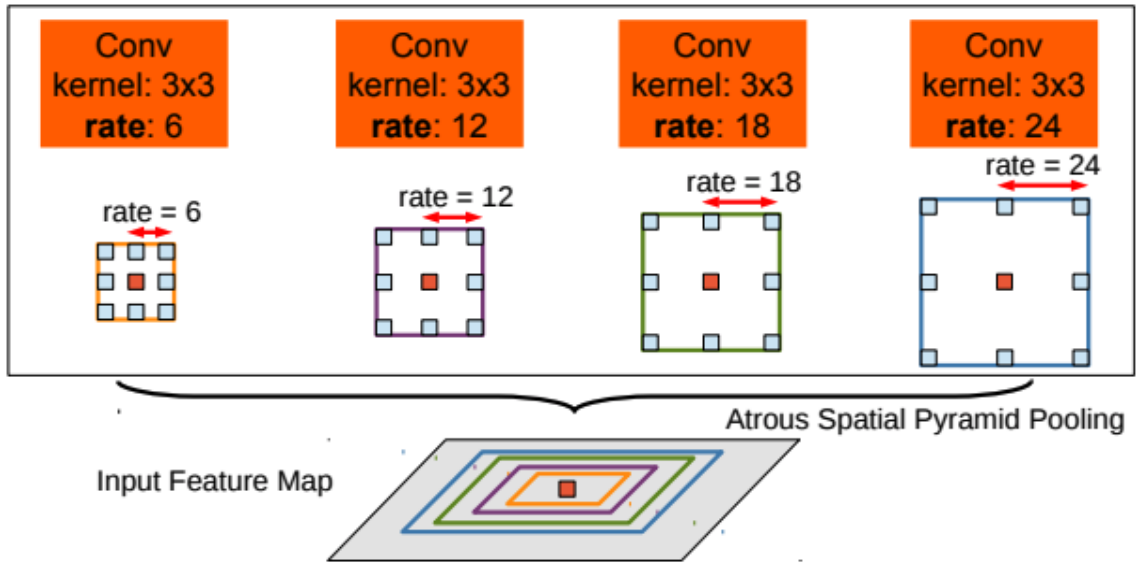
能够输出更加稠密的预测结果

DeepLab v1 & v2

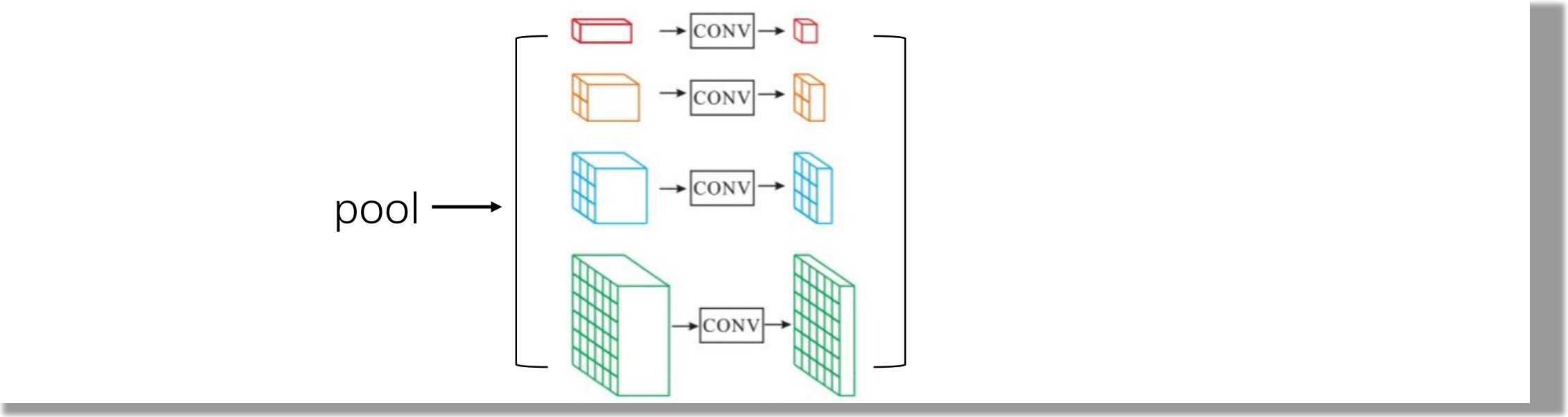
DeepLab v1



DeepLab v2



Feature Ensampling Based



典型模型

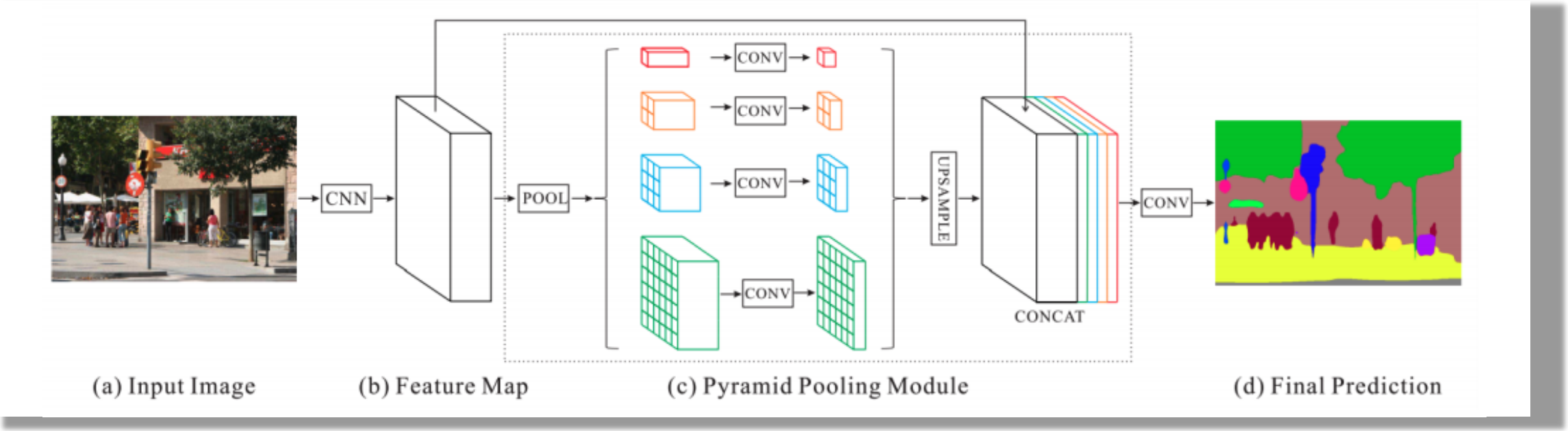
- PspNet[1]
- RefineNet[2]

---

[1]: Zhao, Hengshuang, et al. ”Pyramid scene parsing network.” CVPR. 2017.

[2]: Lin G, et al. ”Refinenet: Multi-path refinement networks for high-resolution semantic ....” CVPR, 2017.

PSPNet

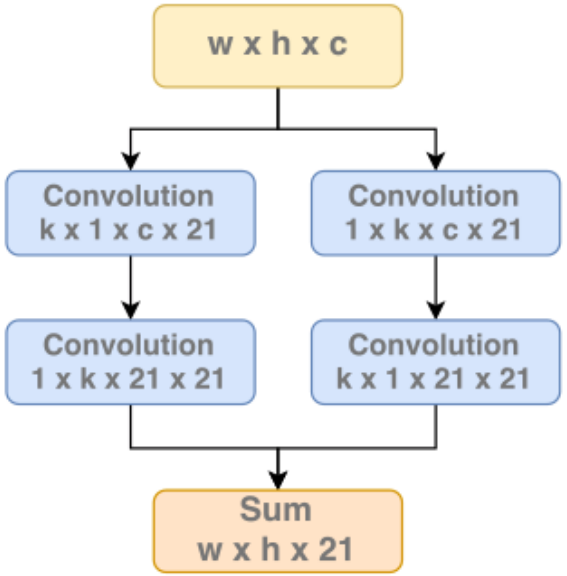


特点

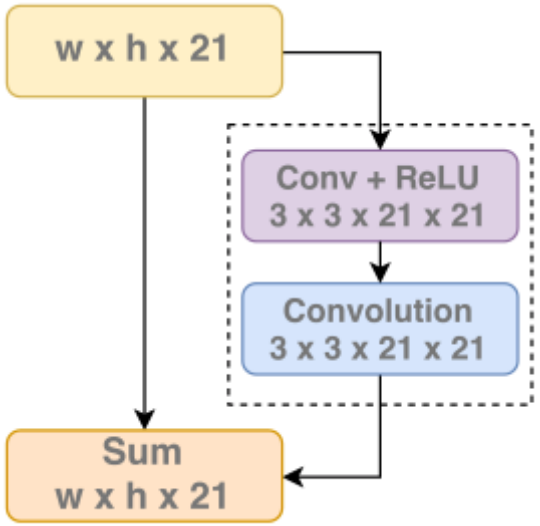
- 抽取特征时使用膨胀卷积
- 池化金字塔中使用1x1 卷积压缩特征通道数
- 整合不同尺度信息来获取全局特征表达

Large Kernel Matters

大尺度卷积



边缘锐化



典型模型

- ExFuse[1]

---

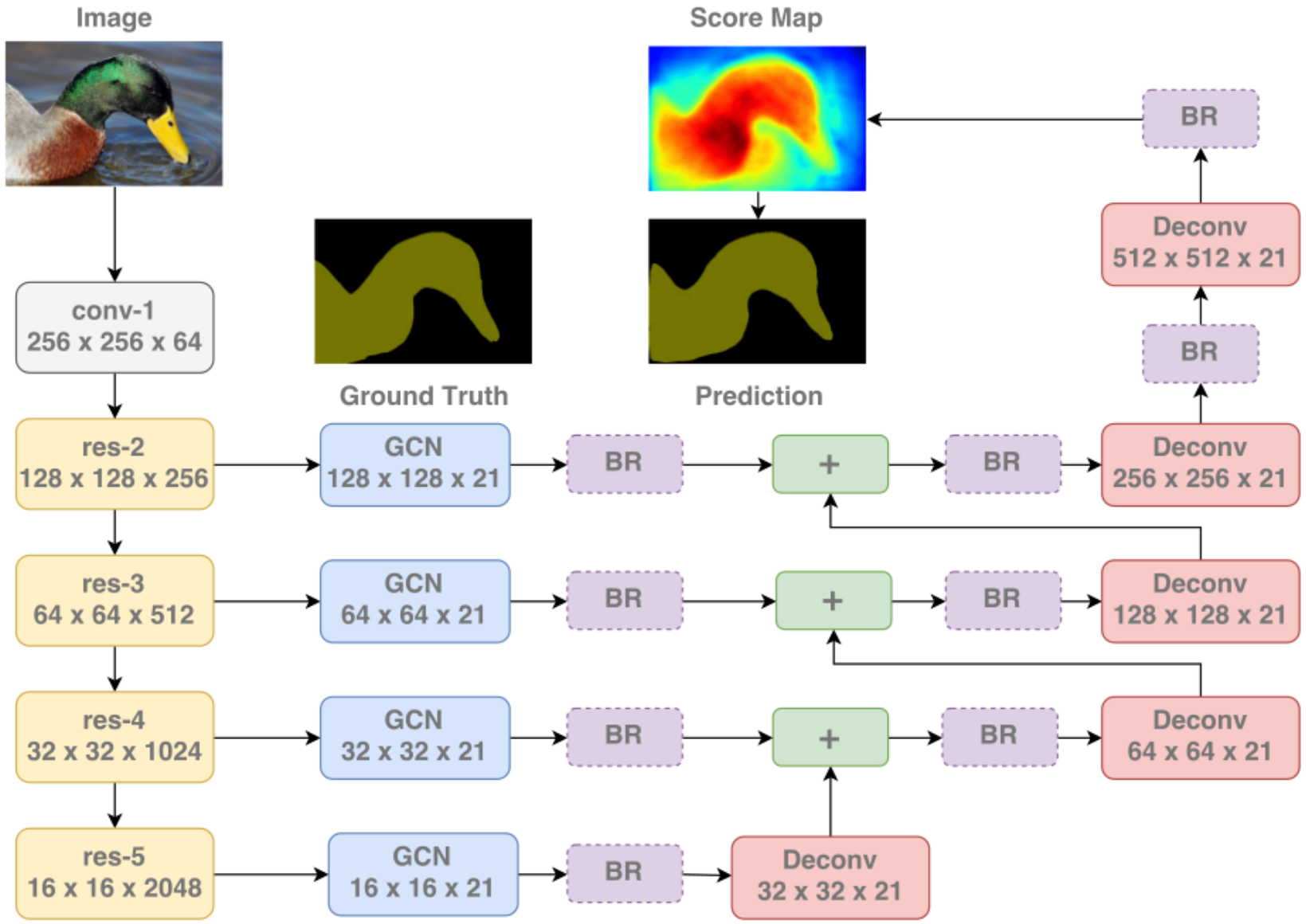
[1]: cZhenli, Zhang, et al. ”ExFuse: Enhancing Feature Fusion for Semantic Segmentation”, ECCV,2018.

Large Kernel Matters

贡献

随着GCN中k值的增加，网络的性能越来越好，而普通卷积与堆叠3x3卷积在k>5之后性能均会下降

GCN的确提升了物体内部的分割精度但对边缘精度没有什么影响，而BR的确提升了物体边缘的分割精度



无/半监督学习方法

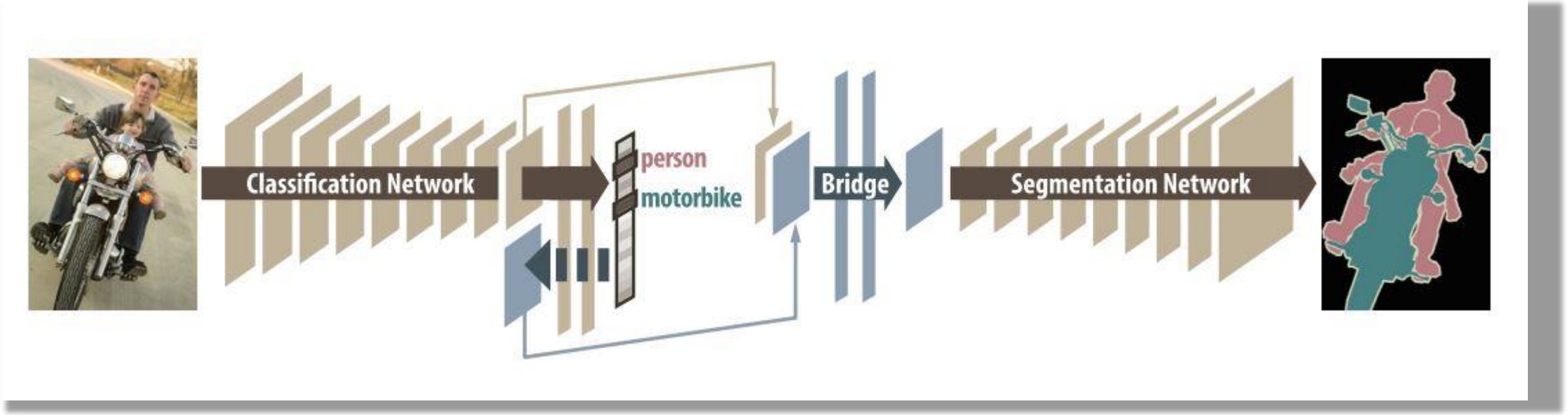
有些时候无法获取逐像素标注或者没有足够的标注数据时可以采用无监督或半监督方法

- Weakly- and Semi-Supervised Learning

标注目标类别

标注检测框

DecoupledNet



先用大量的image-level 标签训练前面的分类网络，然后用少量的pixel-level标签来进行后续网络的训练

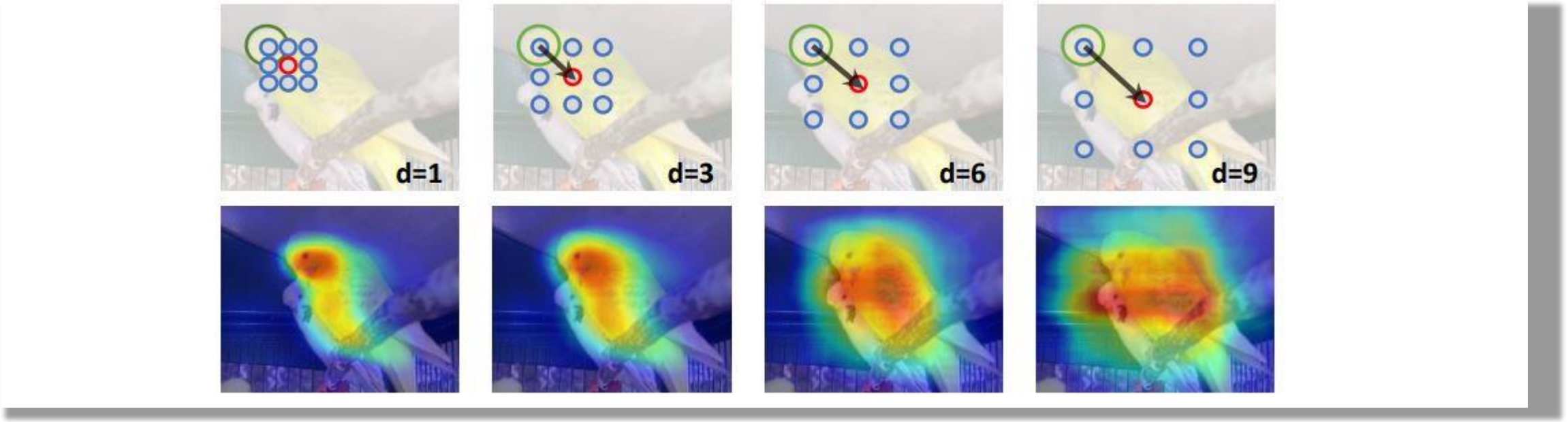
特点

提出了一个将分类和分割网络结合的半监督网络

引入了bridgelayer连接两个网络，主要是提取每一类的activation map然后进行前景背景的分割

Hong, Seunghoon, et al. “Decoupled Deep Neural Network for Semi-Supervised Semantic Segmentation.” *NIPS*, 2015

Revisiting-Dilated-Convolution



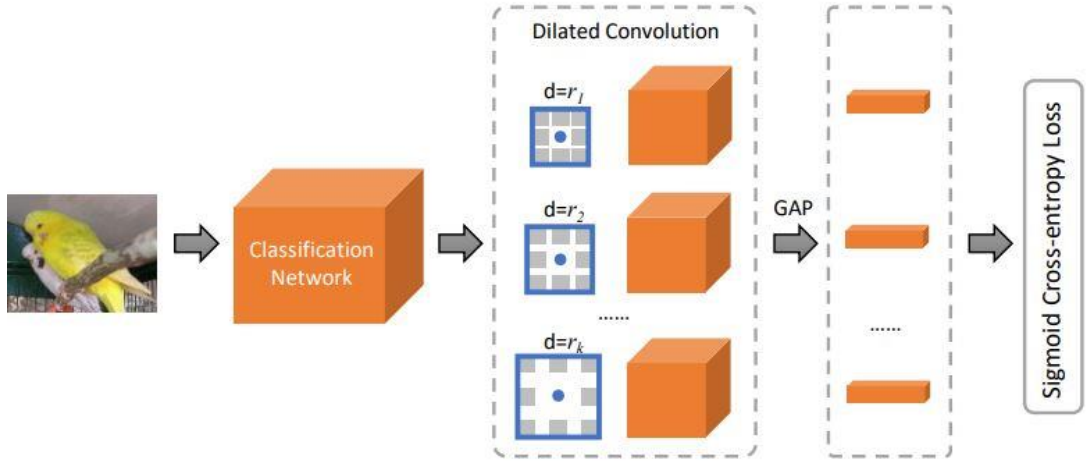
贡献

利用扩张卷积通过传递鉴别分割信息来密集定位对象

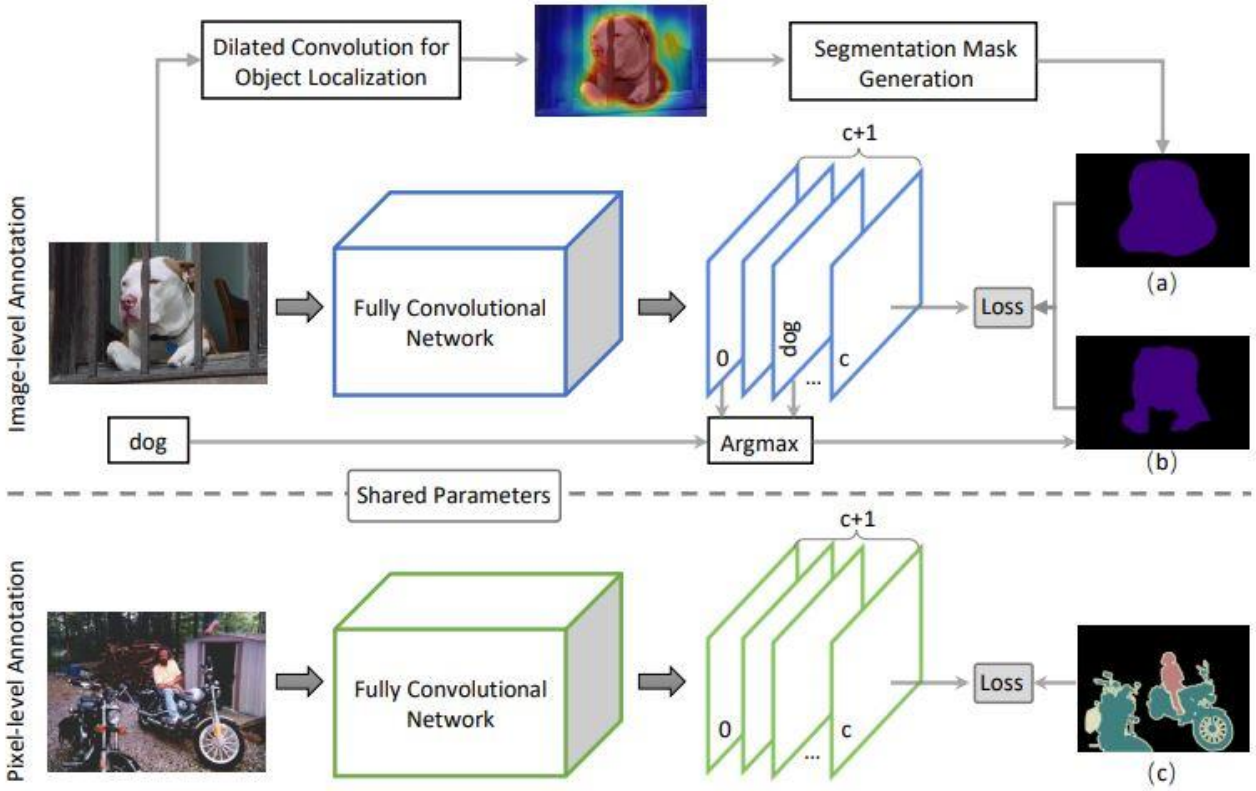
提出的方法对于以弱和半监督方式学习语义分割网络是通用的

Wei, Yunchao, et al. “Revisiting Dilated Convolution: A Simple Approach for Weakly- and Semi-Supervised Semantic Segmentation.” *CVPR*, 2018

Revisiting-Dilated-Convolution



从分类网络训练，输出密集标签



通用的半监督学习框架

目录

- 特征提取网络
- 图像分割方法
- **评价体系**
- 数据集汇总
- 发展趋势

评价体系

t_i	类别i的像素总数
k	分割目标总数
n_{ij}	属于第i类，被标记成第j类的像素总数

评价方法：

Global Accuracy

$$\frac{\sum_{i=1}^{k} n_{ii}}{\sum_{i=1}^{k} t_i}$$

mIoU(mean Intersection over Union)

mean accuracy

frequency weighted IoU

评价体系

t_i	类别i的像素总数
k	分割目标总数
n_{ij}	属于第i类，被标记成第j类的像素总数

评价方法：

Global Accuracy

mIoU(mean Intersection over Union)

$$\frac{1}{k}\sum_{i=1}^{k}\frac{n_{ii}}{t_i - n_{ii} + \sum_{j=1}^{k} n_{ji}}$$

mean accuracy

frequency weighted IoU

评价体系

t_i	类别i的像素总数
k	分割目标总数
n_{ij}	属于第i类，被标记成第j类的像素总数

评价方法：

Global Accuracy

mIoU(mean Intersection over Union)

mean accuracy

$$\frac{1}{k}\sum_{i=1}^{k}\frac{n_{ii}}{t_i}$$

frequency weighted IoU

评价体系

t_i	类别i的像素总数
k	分割目标总数
n_{ij}	属于第i类，被标记成第j类的像素总数

评价方法：

Global Accuracy

mIoU(mean Intersection over Union)

mean accuracy

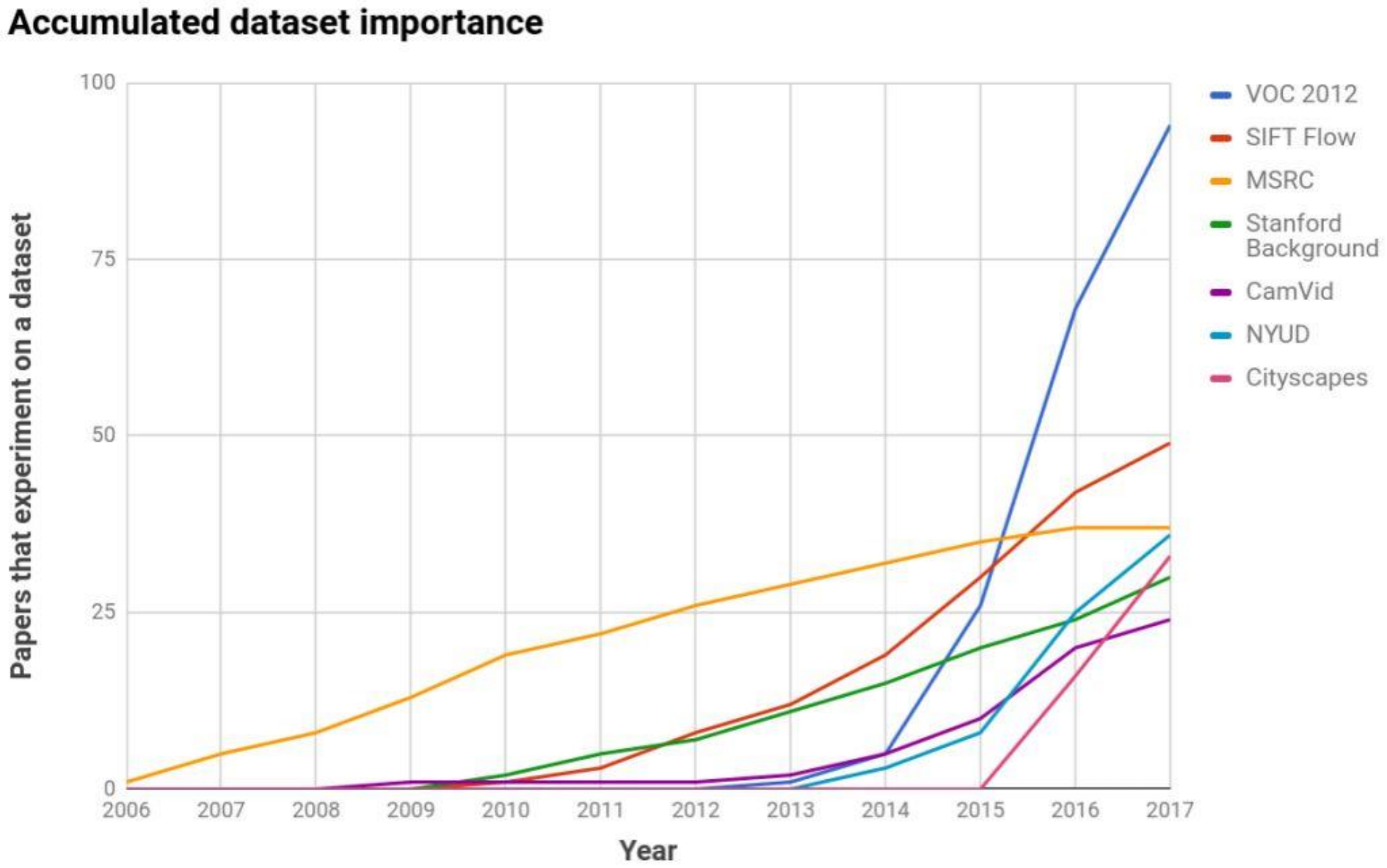
frequency weighted IoU

$$(\sum_{i=1}^{k} t_i)^{-1} \frac{1}{k} \sum_{i=1}^{k} \frac{n_{ii}}{t_i - n_{ii} + \sum_{j=1}^{k} n_{ji}}$$

目录

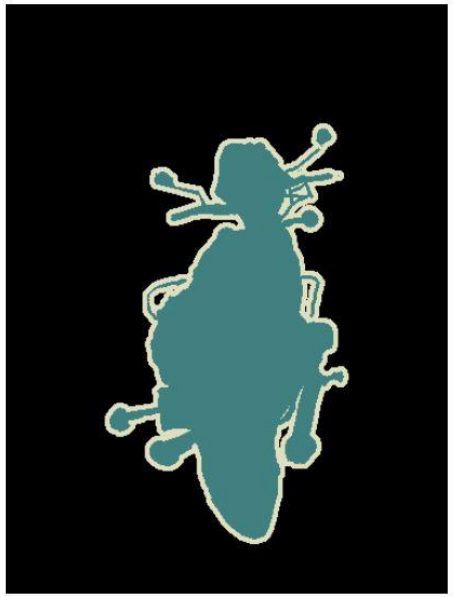
- 特征提取网络
- 图像分割方法
- 评价体系
- **数据集汇总**
- 发展趋势

数据集汇总


Accumulated dataset importance
Papers that experiment on a dataset
100
75
50
25
0
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Year
VOC 2012
SIFT Flow
MSRC
Stanford Background
CamVid
NYUD
Cityscapes

数据集汇总

Original image



Ground truth

VOC[1] 2007/2012: 20 classes

CitySpaces[2]: 30 classes

[1]: Everingham, Mark, et al. ”The pascal visual object classes challenge: A retrospective.” IJCV, 2015

[2]: Cordts, Marius, et al. ”The cityscapes dataset for semantic urban scene understanding.”CVPR. 2016.

数据集汇总

Original image

Ground truth

ADE20K[1]

Microsoft COCO[2]

[1]: Zhou, Bolei, et al. ”Scene parsing through ade20k dataset.” Proc. CVPR. 2017.

[2]: Lin, Tsung-Yi, et al. ”Microsoft coco: Common objects in context.” ECCV, 2014.

数据集汇总

Name and Reference	Purpose	Year	Classes	Data	Resolution	Sequence	Synthetic/Real	Samples (training)	Samples (validation)	Samples (test)
PASCAL VOC 2012 Segmentation [27]	Generic	2012	21	2D	Variable	✗	R	1464	1449	Private
PASCAL-Context [28]	Generic	2014	540 (59)	2D	Variable	✗	R	10103	N/A	9637
PASCAL-Part [29]	Generic-Part	2014	20	2D	Variable	✗	R	10103	N/A	9637
SBD [30]	Generic	2011	21	2D	Variable	✗	R	8498	2857	N/A
Microsoft COCO [31]	Generic	2014	+80	2D	Variable	✗	R	82783	40504	81434
SYNTHIA [32]	Urban (Driving)	2016	11	2D	960 × 720	✗	S	13407	N/A	N/A
Cityscapes (fine) [33]	Urban	2015	30 (8)	2D	2048 × 1024	✓	R	2975	500	1525
Cityscapes (coarse) [33]	Urban	2015	30 (8)	2D	2048 × 1024	✓	R	22973	500	N/A
CamVid [34]	Urban (Driving)	2009	32	2D	960 × 720	✓	R	701	N/A	N/A
CamVid-Sturgess [35]	Urban (Driving)	2009	11	2D	960 × 720	✓	R	367	100	233
KITTI-Layout [36] [37]	Urban/Driving	2012	3	2D	Variable	✗	R	323	N/A	N/A
KITTI-Ros [38]	Urban/Driving	2015	11	2D	Variable	✗	R	170	N/A	46
KITTI-Zhang [39]	Urban/Driving	2015	10	2D/3D	1226 × 370	✗	R	140	N/A	112
Stanford background [40]	Outdoor	2009	8	2D	320 × 240	✗	R	725	N/A	N/A
SiftFlow [41]	Outdoor	2011	33	2D	256 × 256	✗	R	2688	N/A	N/A
Youtube-Objects-Jain [42]	Objects	2014	10	2D	480 × 360	✓	R	10167	N/A	N/A
Adobe's Portrait Segmentation [26]	Portrait	2016	2	2D	600 × 800	✗	R	1500	300	N/A
MINC [43]	Materials	2015	23	2D	Variable	✗	R	7061	2500	5000
DAVIS [44] [45]	Generic	2016	4	2D	480p	✓	R	4219	2023	2180
NYUDv2 [46]	Indoor	2012	40	2.5D	480 × 640	✗	R	795	654	N/A
SUN3D [47]	Indoor	2013	–	2.5D	640 × 480	✓	R	19640	N/A	N/A
SUNRGBD [48]	Indoor	2015	37	2.5D	Variable	✗	R	2666	2619	5050
RGB-D Object Dataset [49]	Household objects	2011	51	2.5D	640 × 480	✓	R	207920	N/A	N/A
ShapeNet Part [50]	Object/Part	2016	16/50	3D	N/A	✗	S	31,963	N/A	N/A
Stanford 2D-3D-S [51]	Indoor	2017	13	2D/2.5D/3D	1080 × 1080	✓	R	70469	N/A	N/A
3D Mesh [52]	Object/Part	2009	19	3D	N/A	✗	S	380	N/A	N/A
Sydney Urban Objects Dataset [53]	Urban (Objects)	2013	26	3D	N/A	✗	R	41	N/A	N/A
Large-Scale Point Cloud Classification Benchmark [54]	Urban/Nature	2016	8	3D	N/A	✗	R	15	N/A	15

目录

- 特征提取网络
- 图像分割方法
- 评价体系
- **数据集汇总**
- **发展趋势**

发展趋势

Challenges

1. Memory issues for segmentation
2. Utilize location information
3. Annotations is expensive
4. Category-specific
5. Boundary is not fine
6. ResNet is good choice?
7. Why such prediction output?
8. Not robust to different data

Future direction

1. More efficient methods(video)
2. Small object is challenging now
3. Weakly-supervised method
4. Meta-learning for segmentation
5. Incorporate low-level cues
6. Segmentation-specific network
7. Interpretability(medical image)
8. Robust model to different data.

Sonagyang Zhang et al. “A Survey on DNN based Semantic Segmentation Focus on Image Semantic Segmentation”

FINISH