最全综述 | 图像分割算法

mp.weixin.qq.com/s

点击上方“计算机视觉life”，选择“星标”

快速获得最新干货

> 图像分割是计算机视觉研究中的一个经典难题，已经成为图像理解领域关注的一个热点，图像分割是图像分析的第一步，是计算机视觉的基础，是图像理解的重要组成部分，同时也是图像处理中最困难的问题之一。所谓图像分割是指根据灰度、彩色、空间纹理、几何形状等特征把图像划分成若干个互不相交的区域，使得这些特征在同一区域内表现出一致性或相似性，而在不同区域间表现出明显的不同。简单的说就是在一副图像中，把目标从背景中分离出来。对于灰度图像来说，区域内部的像素一般具有灰度相似性，而在区域的边界上一般具有灰度不连续性。 关于图像分割技术，由于问题本身的重要性和困难性，从20世纪70年代起图像分割问题就吸引了很多研究人员为之付出了巨大的努力。虽然到目前为止，还不存在一个通用的完美的图像分割的方法，但是对于图像分割的一般性规律则基本上已经达成的共识，已经产生了相当多的研究成果和方法。
>
> 本文对于目前正在使用的各种图像分割方法进行了一定的归纳总结，由于笔者对于图像分割的了解也是初窥门径，所以难免会有一些错误，还望各位读者多多指正，共同学习进步。

传统分割方法

这一大部分我们将要介绍的是深度学习大火之前人们利用数字图像处理、拓扑学、数学等方面的只是来进行图像分割的方法。当然现在随着算力的增加以及深度学习的不断发展，一些传统的分割方法在效果上已经不能与基于深度学习的分割方法相比较了，但是有些天才的思想还是非常值得我们去学习的。

1.基于阈值的分割方法

阈值法的基本思想是基于图像的灰度特征来计算一个或多个灰度阈值，并将图像中每个像素的灰度值与阈值作比较，最后将像素根据比较结果分到合适的类别中。因此，该方法最为关键的一步就是按照某个准则函数来求解最佳灰度阈值。

阈值法特别适用于目标和背景占据不同灰度级范围的图。

图像若只有目标和背景两大类，那么只需要选取一个阈值进行分割，此方法成为单阈值分割；但是如果图像中有多个目标需要提取，单一阈值的分割就会出现作物，在这种情况下就需要选取多个阈值将每个目标分隔开，这种分割方法相应的成为多阈值分割。

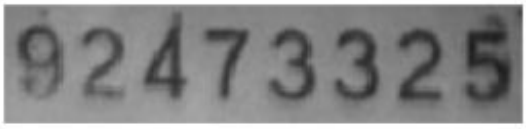


图 4. 原始图像

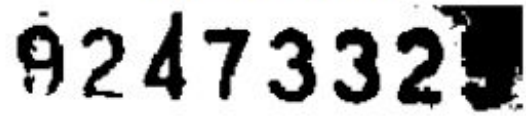


图 5. 阈值低，对亮区效果好，则暗区差

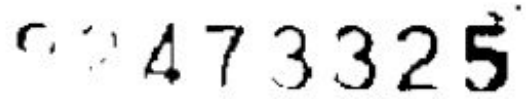


图 6. 阈值高，对暗区效果好，则亮区差

如图所示即为对数字的一种阈值分割方法。

阀值分割方法的优缺点：

- 计算简单，效率较高；
- 只考虑像素点灰度值本身的特征，一般不考虑空间特征，因此对噪声比较敏感，鲁棒性不高。

从前面的介绍里我们可以看出，阈值分割方法的最关键就在于阈值的选择。若将智能遗传算法应用在阀值筛选上，选取能最优分割图像的阀值，这可能是基于阀值分割的图像分割法的发展趋势。

2.基于区域的图像分割方法

基于区域的分割方法是以直接寻找区域为基础的分割技术，基于区域提取方法有两种基本形式：一种是区域生长，从单个像素出发，逐步合并以形成所需要的分割区域；另一种是从全局出发，逐步切割至所需的分割区域。

区域生长

区域生长是从一组代表不同生长区域的种子像素开始，接下来将种子像素邻域里符合条件的像素合并到种子像素所代表的生长区域中，并将新添加的像素作为新的种子像素继续合并过程，知道找不到符合条件的新像素为止（小编研一第一学期的机器学习期末考试就是手写该算法T.T），该方法的关键是选择合适的初始种子像素以及合理的生长准则。

区域生长算法需要解决的三个问题：

（1）选择或确定一组能正确代表所需区域的种子像素；

（2）确定在生长过程中能将相邻像素包括进来的准则；

（3）指定让生长过程停止的条件或规则。

区域分裂合并

区域生长是从某个或者某些像素点出发，最终得到整个区域，进而实现目标的提取。而分裂合并可以说是区域生长的逆过程，从整幅图像出发，不断的分裂得到各个子区域，然后再把前景区域合并，得到需要分割的前景目标，进而实现目标的提取。其实如果理解了上面的区域生长算法这个区域分裂合并算法就比较好理解啦。

四叉树分解法就是一种典型的区域分裂合并法，基本算法如下：

（1）对于任一区域，如果H(Ri)=FALSE就将其分裂成不重叠的四等分；

（2）对相邻的两个区域Ri和Rj，它们也可以大小不同（即不在同一层），如果条件H(RiURj)=TURE满足，就将它们合并起来；

（3）如果进一步的分裂或合并都不可能，则结束。

其中R代表整个正方形图像区域，P代表逻辑词。

区域分裂合并算法优缺点：

（1）对复杂图像分割效果好；

（2）算法复杂，计算量大；

（3）分裂有可能破怪区域的边界。

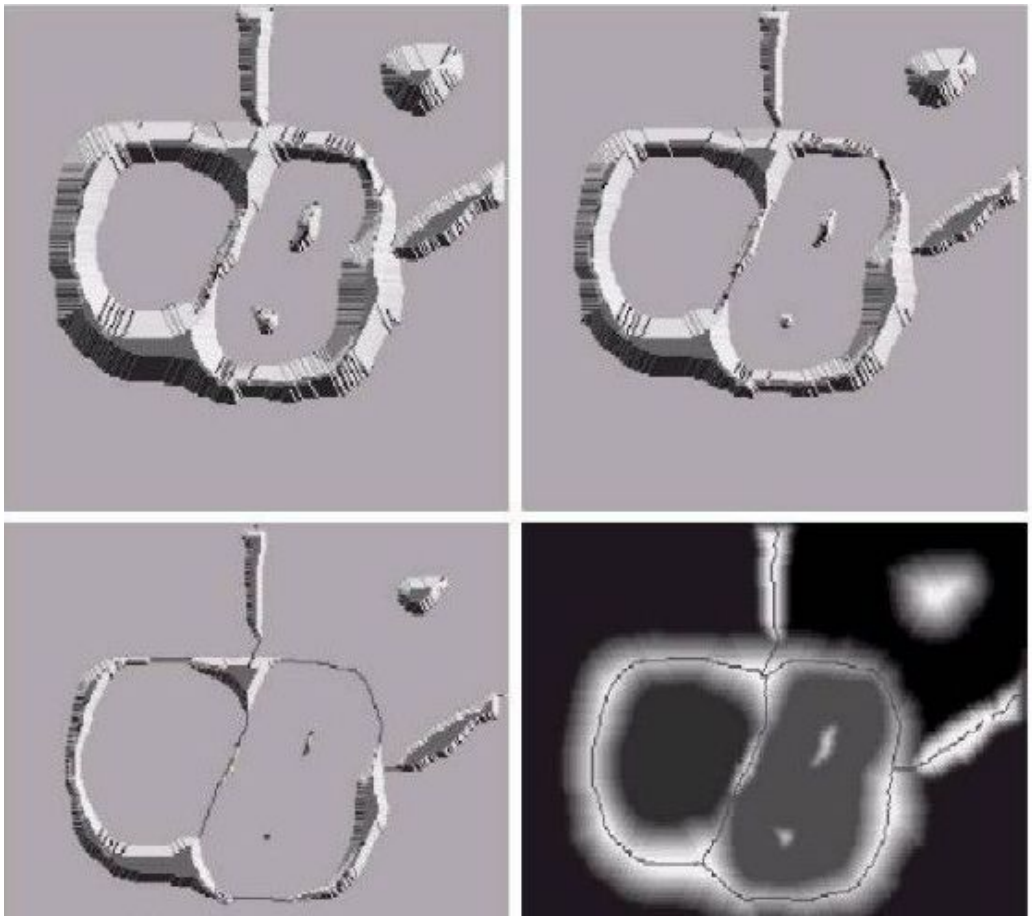
在实际应用当中通常将区域生长算法和区域分裂合并算法结合使用，该类算法对某些复杂物体定义的复杂场景的分割或者对某些自然景物的分割等类似先验知识不足的图像分割效果较为理想。

分水岭算法

分水岭算法是一个非常好理解的算法，它根据分水岭的构成来考虑图像的分割，现实中我们可以想象成有山和湖的景象，那么一定是如下图的，水绕山山围水的景象。

分水岭分割方法，是一种基于拓扑理论的数学形态学的分割方法，其基本思想是把图像看作是测地学上的拓扑地貌，图像中每一点像素的灰度值表示该点的海拔高度，每一个局部极小值及其影响区域称为集水盆，而集水盆的边界则形成分水岭。分水岭的概念和形成可以通过模拟浸入过程来说明。在每一个局部极小值表面，刺穿一个小孔，然后把整个模型慢慢浸入水中，随着浸入的加深，每一个局部极小值的影响域慢慢向外扩展，在两个集水盆汇合处构筑大坝，即形成分水岭。

分水岭对微弱边缘具有良好的响应，图像中的噪声、物体表面细微的灰度变化都有可能产生过度分割的现象，但是这也同时能够保证得到封闭连续边缘。同时，分水岭算法得到的封闭的集水盆也为分析图像的区域特征提供了可能。



e f
g h

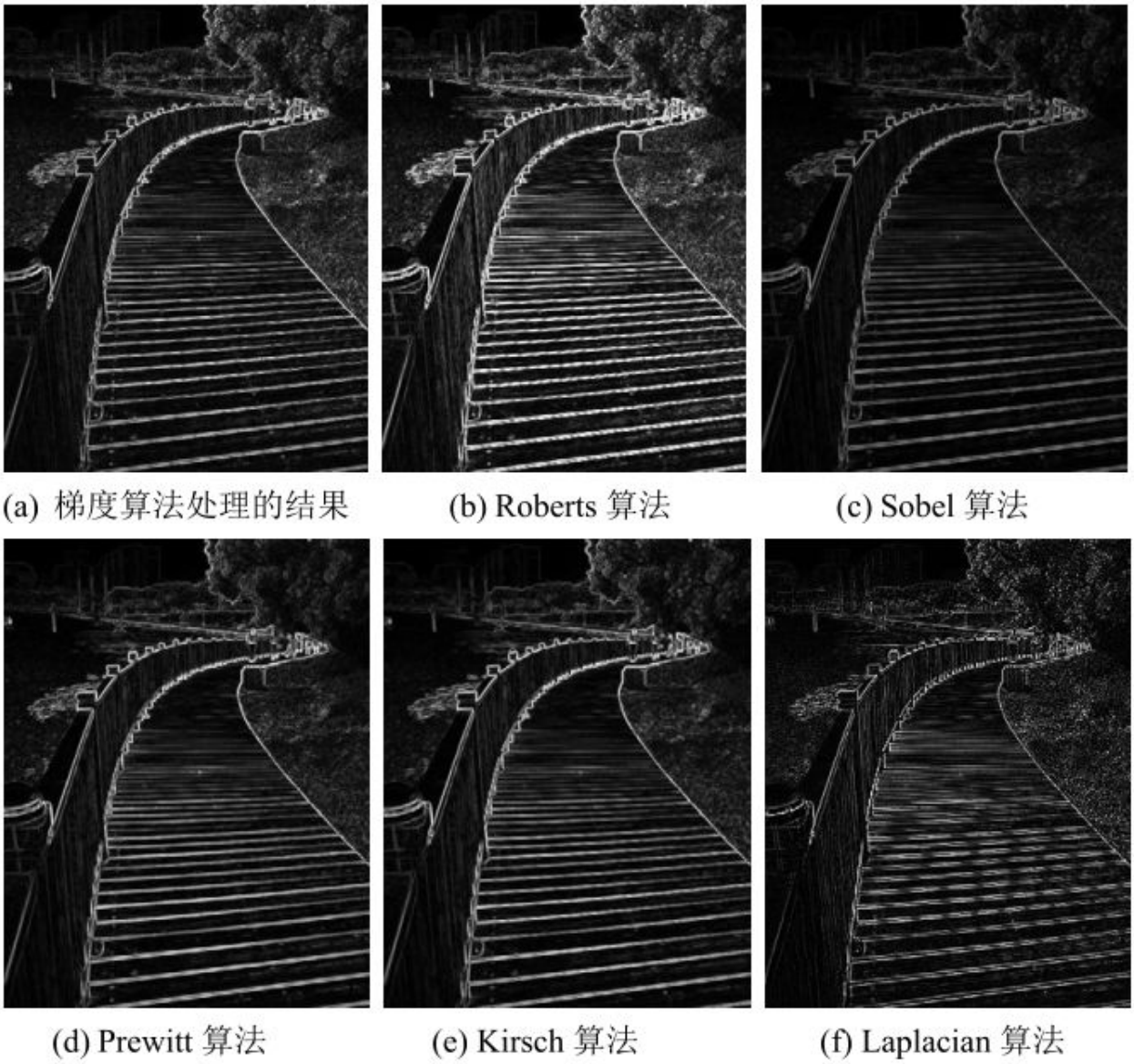
FIGURE 10.44
(Continued)
(e) Result of further flooding. (f) Beginning of merging of water from two catchment basins (a short dam was built between them). (g) Longer dams. (h) Final watershed (segmentation) lines. (Courtesy of Dr. S. Beucher, CMM/Ecole des Mines de Paris.)

3.基于边缘检测的分割方法

基于边缘检测的图像分割算法试图通过检测包含不同区域的边缘来解决分割问题。它可以说是人们最先想到也是研究最多的方法之一。通常不同区域的边界上像素的灰度值变化比较剧烈，如果将图片从空间域通过傅里叶变换到频率域，边缘就对应着高频部分，这是一种非常简单的边缘检测算法。

边缘检测技术通常可以按照处理的技术分为串行边缘检测和并行边缘检测。串行边缘检测是要想确定当前像素点是否属于检测边缘上的一点，取决于先前像素的验证结果。并行边缘检测是一个像素点是否属于检测边缘高尚的一点取决于当前正在检测的像素点以及与该像素点的一些临近像素点。

最简单的边缘检测方法是并行微分算子法，它利用相邻区域的像素值不连续的性质，采用一阶或者二阶导数来检测边缘点。近年来还提出了基于曲面拟合的方法、基于边界曲线拟合的方法、基于反应-扩散方程的方法、串行边界查找、基于变形模型的方法。



(a) 梯度算法处理的结果　(b) Roberts 算法　(c) Sobel 算法

(d) Prewitt 算法　(e) Kirsch 算法　(f) Laplacian 算法

边缘检测的优缺点：

（1）边缘定位准确；

（2）速度快；

（3）不能保证边缘的连续性和封闭性；

（4）在高细节区域存在大量的碎边缘，难以形成一个大区域，但是又不宜将高细节区域分成小碎片；

由于上述的（3）（4）两个难点，边缘检测只能产生边缘点，而非完整意义上的图像分割过程。这也就是说，在边缘点信息获取到之后还需要后续的处理或者其他相关算法相结合才能完成分割任务。

在以后的研究当中，用于提取初始边缘点的自适应阈值选取、用于图像的层次分割的更大区域的选取以及如何确认重要边缘以去除假边缘将变得非常重要。

结合特定工具的图像分割算法

基于小波分析和小波变换的图像分割方法

小波变换是近年来得到的广泛应用的数学工具，也是现在数字图像处理必学部分，它在时间域和频率域上都有量高的局部化性质，能将时域和频域统一于一体来研究信号。而且小波变换具有多尺度特性，能够在不同尺度上对信号进行分析，因此在图像分割方面的得到了应用，

二进小波变换具有检测二元函数的局部突变能力，因此可作为图像边缘检测工具。图像的边缘出现在图像局部灰度不连续处，对应于二进小波变换的模极大值点。通过检测小波变换模极大值点可以确定图像的边缘小波变换位于各个尺度上，而每个尺度上的小波变换都能提供一定的边缘信息，因此可进行多尺度边缘检测来得到比较理想的图像边缘。

上图左图是传统的阈值分割方法，右边的图像就是利用小波变换的图像分割。可以看出右图分割得到的边缘更加准确和清晰

另外，将小波和其他方法结合起来处理图像分割的问题也得到了广泛研究，比如一种局部自适应阈值法就是将Hilbert图像扫描和小波相结合，从而获得了连续光滑的阈值曲线。

基于遗传算法的图像分割

遗传算法（Genetic Algorithms，简称GA）是1973年由美国教授Holland提出的，是一种借鉴生物界自然选择和自然遗传机制的随机化搜索算法。是仿生学在数学领域的应用。其基本思想是，模拟由一些基因串控制的生物群体的进化过程，把该过程的原理应用到搜索算法中，以提高寻优的速度和质量。此算法的搜索过程不直接作用在变量上，而是在参数集进行了编码的个体，这使得遗传算法可直接对结构对象（图像）进行操作。整个搜索过程是从一组解迭代到另一组解，采用同时处理群体中多个个体的方法，降低了陷入局部最优解的可能性，并易于并行化。搜索过程采用概率的变迁规则来指导搜索方向，而不采用确定性搜索规则，而且对搜索空间没有任何特殊要求（如连通性、凸性等），只利用适应性信息，不需要导数等其他辅助信息，适应范围广。

遗传算法擅长于全局搜索，但局部搜索能力不足，所以常把遗传算法和其他算法结合起来应用。将遗传算法运用到图像处理主要是考虑到遗传算法具有与问题领域无关且快速随机的搜索能力。其搜索从群体出发，具有潜在的并行性，可以进行多个个体的同时比较，能有效的加

快图像处理的速度。但是遗传算法也有其缺点：搜索所使用的评价函数的设计、初始种群的选择有一定的依赖性等。要是能够结合一些启发算法进行改进且遗传算法的并行机制的潜力得到充分的利用，这是当前遗传算法在图像处理中的一个研究热点。

基于主动轮廓模型的分割方法

主动轮廓模型（active contours）是图像分割的一种重要方法，具有统一的开放式的描述形式，为图像分割技术的研究和创新提供了理想的框架。在实现主动轮廓模型时，可以灵活的选择约束力、初始轮廓和作用域等，以得到更佳的分割效果，所以主动轮廓模型方法受到越来越多的关注。

该方法是在给定图像中利用曲线演化来检测目标的一类方法，基于此可以得到精确的边缘信息。其基本思想是，先定义初始曲线C，然后根据图像数据得到能量函数，通过最小化能量函数来引发曲线变化，使其向目标边缘逐渐逼近，最终找到目标边缘。这种动态逼近方法所求得的边缘曲线具有封闭、光滑等优点。

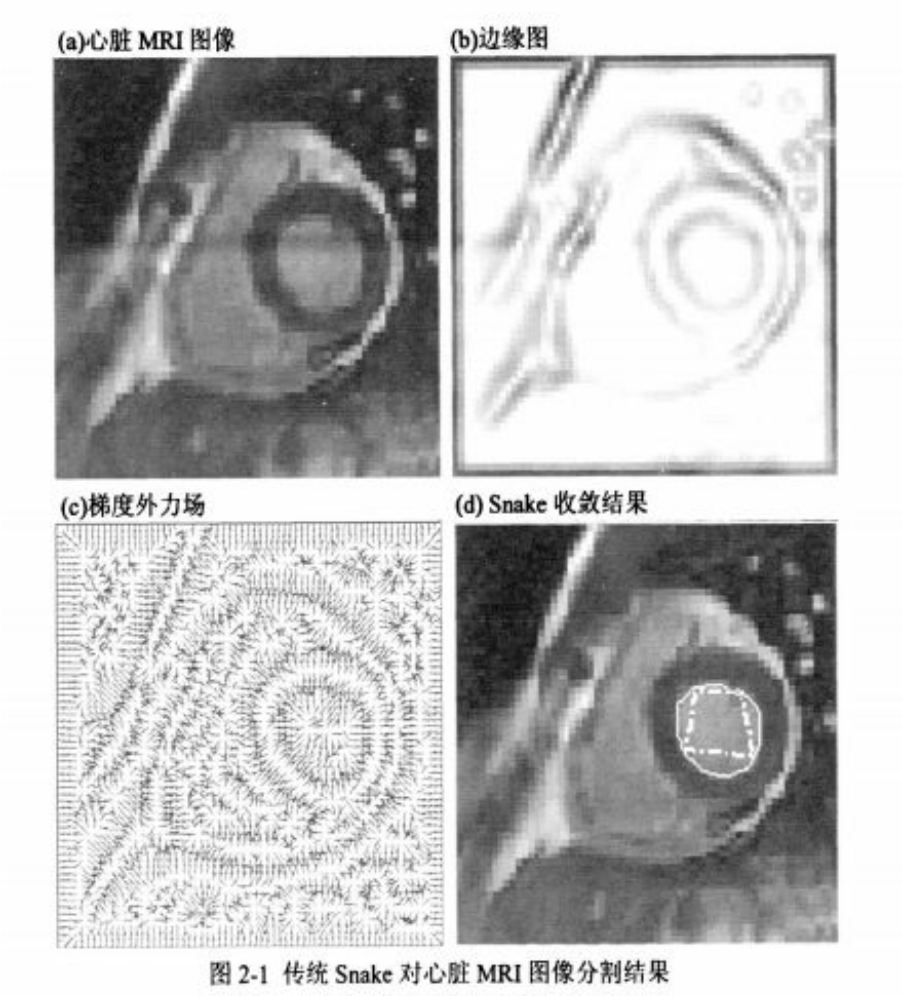


图 2-1 传统 Snake 对心脏 MRI 图像分割结果

传统的主动轮廓模型大致分为参数主动轮廓模型和几何主动轮廓模型。参数主动轮廓模型将曲线或曲面的形变以参数化形式表达，Kass等人提出了经典的参数活动轮廓模型即“Snake”模型，其中Snake定义为能量极小化的样条曲线，它在来自曲线自身的内力和来自图像数据的外力的共同作用下移动到感兴趣的边缘，内力用于约束曲线形状，而外力则引导曲线到特征此边缘。参数主动轮廓模型的特点是将初始曲线置于目标区域附近，无需人为设定曲线

的的演化是收缩或膨胀，其优点是能够与模型直接进行交互，且模型表达紧凑，实现速度快；其缺点是难以处理模型拓扑结构的变化。比如曲线的合并或分裂等。而使用水平集（level set）的几何活动轮廓方法恰好解决了这一问题。

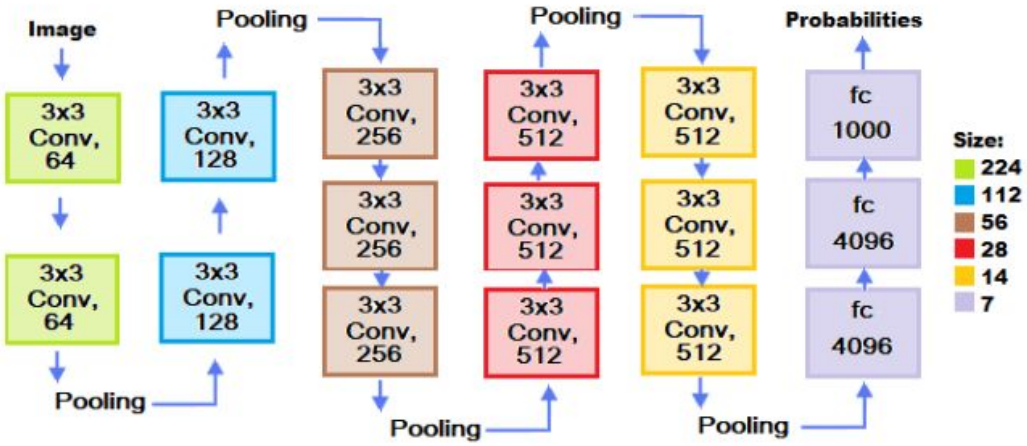
基于深度学习的分割

1.基于特征编码（feature encoder based）

在特征提取领域中VGGnet和ResNet是两个非常有统治力的方法，接下来的一些篇幅会对这两个方法进行简短的介绍

a.VGGNet

由牛津大学计算机视觉组合和Google DeepMind公司研究员一起研发的深度卷积神经网络。它探索了卷积神经网络的深度和其性能之间的关系，通过反复的堆叠*33的小型卷积核和*22的最大池化层，成功的构建了16~19层深的卷积神经网络。VGGNet获得了ILSVRC 2014年比赛的亚军和定位项目的冠军，在top5上的错误率为7.5%。目前为止，VGGNet依然被用来提取图像的特征。



VGGNet的优缺点

1. 由于参数量主要集中在最后的三个FC当中，所以网络加深并不会带来参数爆炸的问题；
2. 多个小核卷积层的感受野等同于一个大核卷积层（三个3x3等同于一个7x7）但是参数量远少于大核卷积层而且非线性操作也多于后者，使得其学习能力较强
3. VGG由于层数多而且最后的三个全连接层参数众多，导致其占用了更多的内存（140M）

b.ResNet

随着深度学习的应用，各种深度学习模型随之出现，虽然在每年都会出现性能更好的新模型，但是对于前人工作的提升却不是那么明显，其中有重要问题就是深度学习网络在堆叠到一定深度的时候会出现梯度消失的现象，导致误差升高效果变差，后向传播时无法将梯度反馈到

前面的网络层，使得前方的网络层的参数难以更新，训练效果变差。这个时候ResNet恰好站出来，成为深度学习发展历程中一个重要的转折点。

ResNet是由微软研究院的Kaiming He等四名华人提出，他们通过自己提出的ResNet Unit成功训练出来152层的神经网络并在ILSVRC2015比赛中斩获冠军。ResNet语义分割领域最受欢迎且最广泛运用的神经网络.ResNet的核心思想就是在网络中引入恒等映射，允许原始输入信息直接传到后面的层中，在学习过程中可以只学习上一个网络输出的残差（F(x)），因此ResNet又叫做残差网络。、

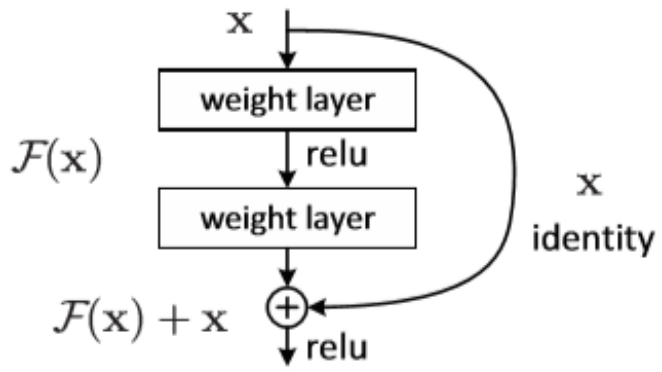


Figure 2. Residual learning: a building block.

使用到ResNet的分割模型：

- Efficient Neural Network（ENet）：该网络类似于ResNet的bottleNeck方法；
- ResNet-38：该网络在训练or测试阶段增加并移除了一些层，是一种浅层网络，它的结构是ResNet+FCN；
- full-resolution residual network(FRRN)：FRRN网络具有和ResNet相同优越的训练特性，它由残差流和池化流两个处理流组成；

- AdapNey：根据ResNet-50的网络进行改进，让原本的ResNet网络能够在更短的时间内学习到更多高分辨率的特征；

  ……

  ResNet的优缺点：

  1）引入了全新的网络结构（残差学习模块），形成了新的网络结构，可以使网络尽可能地加深；

  2）使得前馈/反馈传播算法能够顺利进行，结构更加简单；

  3）恒等映射地增加基本上不会降低网络的性能；

  4）建设性地解决了网络训练的越深，误差升高，梯度消失越明显的问题；

  5）由于ResNet搭建的层数众多，所以需要的训练时间也比平常网络要长。

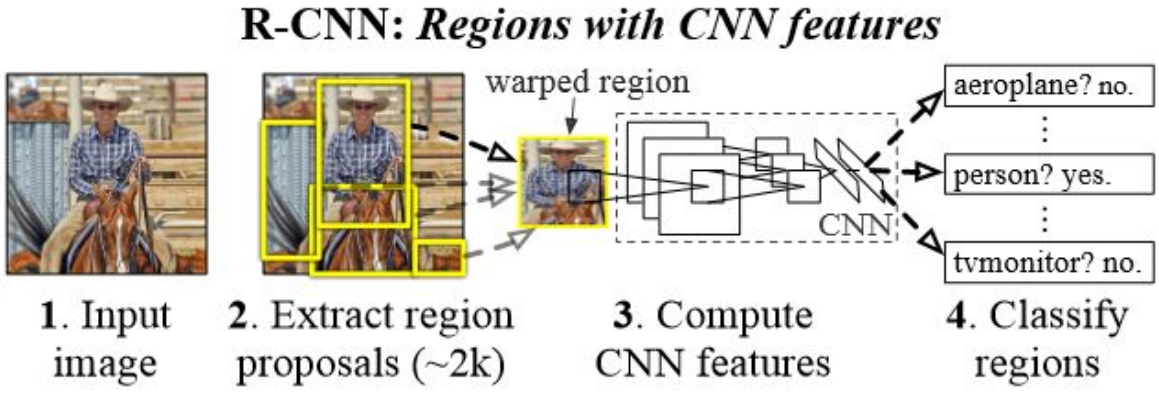
2.基于区域选择（regional proposal based）

Regional proposal 在计算机视觉领域是一个非常常用的算法，尤其是在目标检测领域。其核心思想就是检测颜色空间和相似矩阵，根据这些来检测待检测的区域。然后根据检测结果可以进行分类预测。

在语义分割领域，基于区域选择的几个算法主要是由前人的有关于目标检测的工作渐渐延伸到语义分割的领域的，接下来小编将逐步介绍其个中关系。

Stage I： R-CNN

伯克利大学的Girshick教授等人共同提出了首个在目标检测方向应用的深度学习模型：Region-based Convolutional Neural Network（R-CNN）。该网络模型如下图所示，其主要流程为：先使用selective search算法提取2000个候选框，然后通过卷积网络对候选框进行串行的特征提取，再根据提取的特征使用SVM对候选框进行分类预测，最后使用回归方法对区域框进行修正。

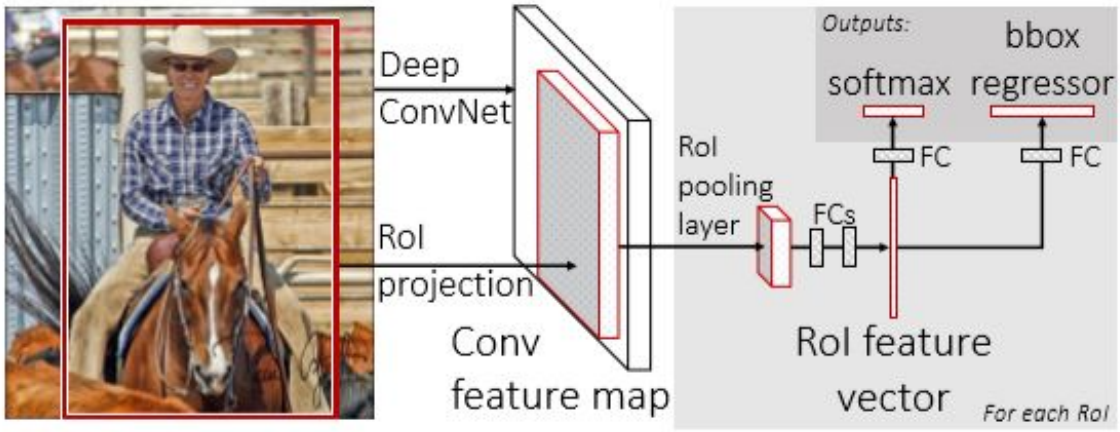


R-CNN的优缺点：

- 是首个开创性地将深度神经网络应用到目标检测的算法；
- 使用Bounding Box Regression对目标检测的框进行调整；
- 由于进行特征提取时是串行，处理耗时过长；
- Selective search算法在提取每一个region时需要2s的时间，浪费大量时间

Stage Ⅱ：Fast R-CNN

由于R-CNN的效率太低，2015年由Ross等学者提出了它的改进版本：Fast R-CNN。其网络结构图如下图所示（从提取特征开始，略掉了region的选择）Fast R-CNN在传统的R-CNN模型上有所改进的地方是它是直接使用一个神经网络对整个图像进行特征提取，就省去了串行提取特征的时间；接着使用一个RoI Pooling Layer在全图的特征图上摘取每一个RoI对应的特征，再通过FC进行分类和包围框的修正。

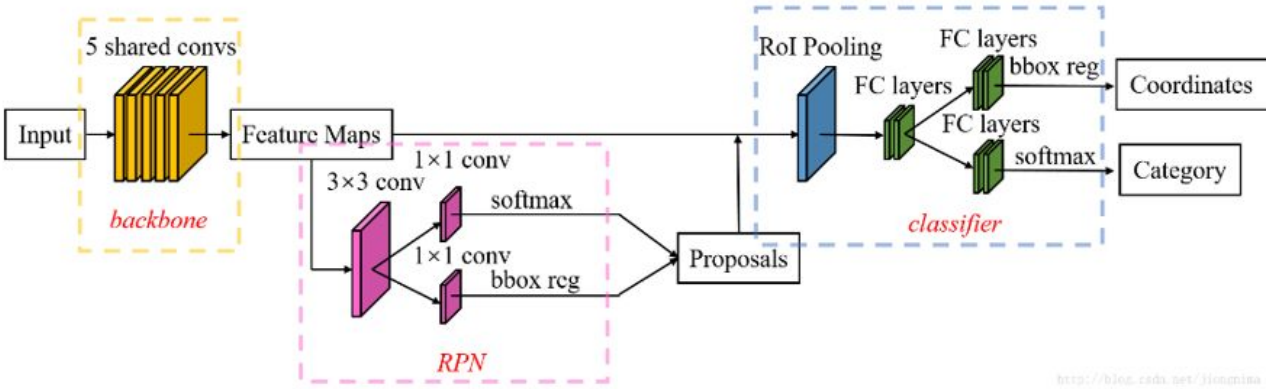


Fast R-CNN的优缺点

- 节省了串行提取特征的时间；
- 除了selective search以外的其它所有模块都可以合在一起训练；
- 最耗时间的selective search算法依然存在。

Stage Ⅲ：Faster R-CNN

2016年提出的Faster R-CNN可以说有了突破性的进展（虽然还是目标检测哈哈哈），因为它改变了它的前辈们最耗时最致命的部位：selective search算法。它将selective search算法替换成为RPN，使用RPN网络进行region的选取，将2s的时间降低到10ms，其网络结构如下图所示：



Faster R-CNN优缺点：

- 使用RPN替换了耗时的selective search算法，对整个网络结构有了突破性的优化；
- Faster R-CNN中使用的RPN和selective search比起来虽然速度更快，但是精度和selective search相比稍有不及，如果更注重速度而不是精度的话完全可以只使用RPN；

Stage Ⅳ：Mask R-CNN

Mask R-CNN（终于到分割了！）是何恺明大神团队提出的一个基于Faster R-CNN模型的一种新型的分割模型，此论文斩获ICCV 2017的最佳论文，在Mask R-CNN的工作中，它主要完成了三件事情：目标检测，目标分类，像素级分割。

恺明大神是在Faster R-CNN的结构基础上加上了Mask预测分支，并且改良了ROI Pooling，提出了ROI Align。其网络结构真容就如下图所示啦：

Mask R-CNN的优缺点：

- 引入了预测用的Mask-Head，以像素到像素的方式来预测分割掩膜，并且效果很好；
- 用ROI Align替代了ROI Pooling，去除了RoI Pooling的粗量化，使得提取的特征与输入良好对齐；
- 分类框与预测掩膜共享评价函数，虽然大多数时间影响不大，但是有的时候会对分割结果有所干扰。

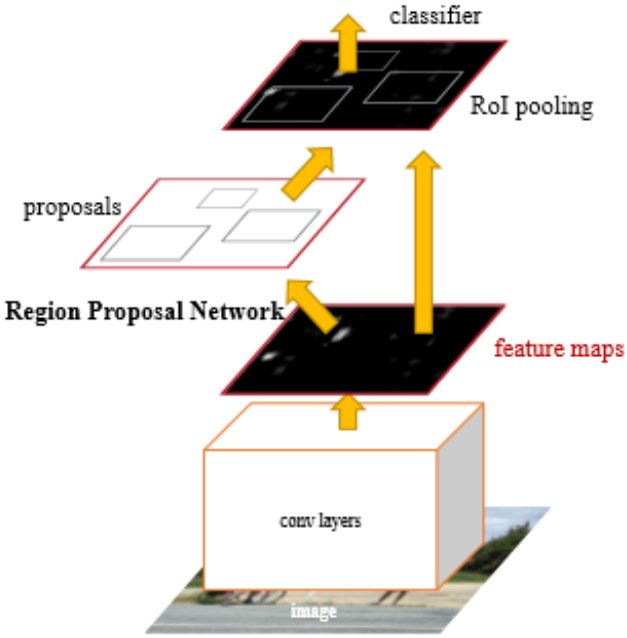


Figure 2: Faster R-CNN is a single, unified network for object detection. The RPN module serves as the 'attention' of this unified network.

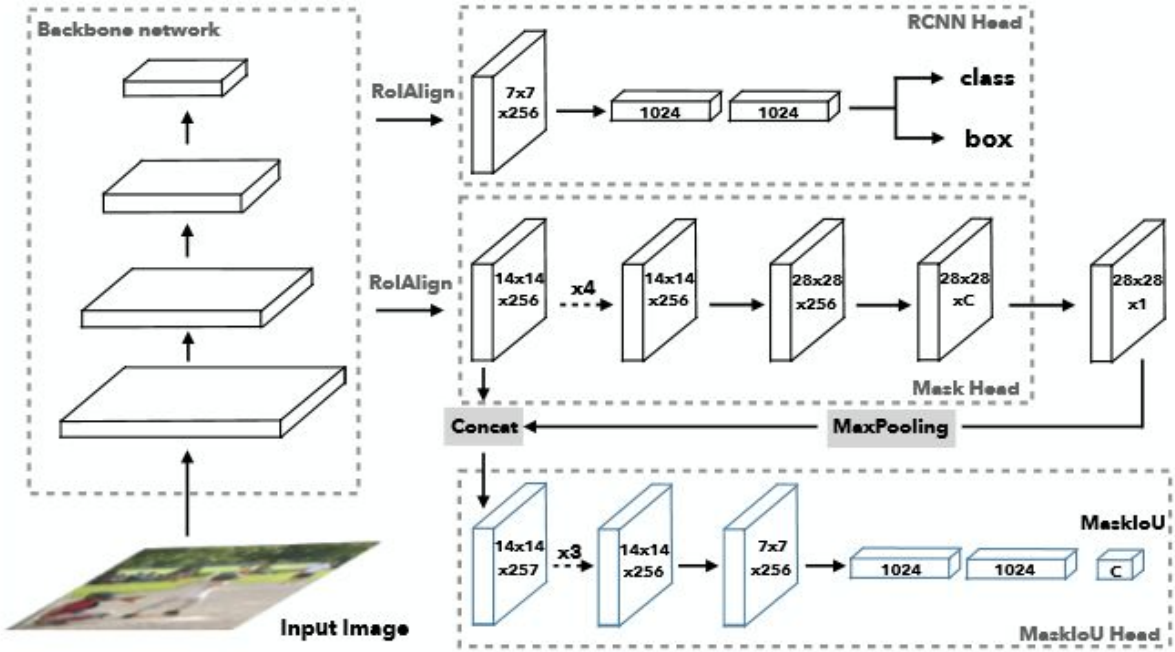
Stage Ⅴ：Mask Scoring R-CNN

最后要提出的是2019年CVPR的oral，来自华中科技大学的研究生黄钊金同学提出的

MS R-CNN，这篇文章的提出主要是对上文所说的Mask R-CNN的一点点缺点进行了修正。他的网络结构也是在Mask R-

CNN的网络基础上做了一点小小的改进，添加了Mask-IoU。

黄同学在文章中提到：恺明大神的Mask R-CNN已经很好啦！但是有个小毛病，就是评价函数只对目标检测的候选框进行打分，而不是分割模板（就是上文提到的优缺点中最后一点），所以会出现分割模板效果很差但是打分很高的情况。所以黄同学增加了对模板进行打分的MaskIoU Head，并且最终的分割结果在COCO数据集上超越了恺明大神，下面就是MS R-CNN的网络结构啦~



MS R-CNN的优缺点：

- 优化了Mask R-CNN中的信息传播，提高了生成预测模板的质量；
- 未经大批量训练的情况下，就拿下了COCO 2017挑战赛实例分割任务冠军；
- 要说缺点的话。。应该就是整个网络有些庞大，一方面需要ResNet当作主干网络，另一方面需要其它各种Head共同承担各种任务。

3.基于RNN的图像分割

Recurrent neural networks（RNNs）除了在手写和语音识别上表现出色外，在解决计算机视觉的任务上也表现不俗，在本篇文章中我们就将要介绍RNN在2D图像处理上的一些应用，其中也包括介绍使用到它的结构或者思想的一些模型。

RNN是由Long-Short-Term Memory（LSTM）块组成的网络，RNN来自序列数据的长期学习的能力以及随着序列保存记忆的能力使其在许多计算机视觉的任务中游刃有余，其中也包括语义分割以及数据标注的任务。接下来的部分我们将介绍几个使用到RNN结构的用于分割的网络结构模型：

1.ReSeg模型

ReSeg可能不被许多人所熟知，在百度上搜索出的相关说明与解析也不多，但是这是一个很有效的语义分割方法。众所周知，FCN可谓是图像分割领域的开山作，而RegNet的作者则在自己的文章中大胆的提出了FCN的不足：没有考虑到局部或者全局的上下文依赖关系，而在语义分割中这种依赖关系是非常有用的。所以在ReSeg中作者使用RNN去检索上下文信息，以此作为分割的一部分依据。

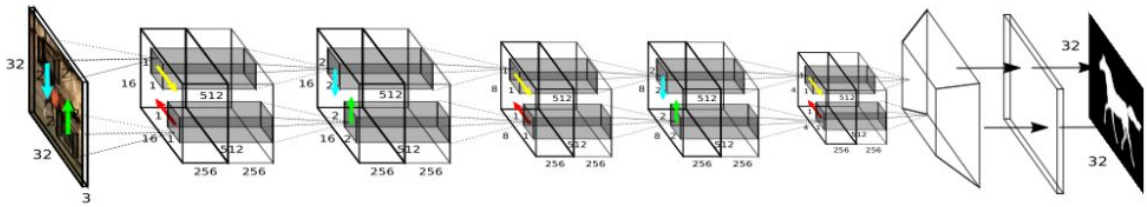


Figure 2. The ReSeg network. For space reasons we do not represent the pretrained VGG-16 convolutional layers that we use to preprocess the input to ReSeg. The first 2 RNNs (blue and green) are applied on 2x2x3 patches of the image, their 16x16x256 feature maps are concatenated and fed as input to the next two RNNs (red and yellow) which read 1x1x512 patches and emit the output of the first ReNet layer. Two similar ReNet layers are stacked, followed by an upsampling layer and a softmax nonlinearity.

该结构的核心就是Recurrent Layer，它由多个RNN组合在一起，捕获输入数据的局部和全局空间结构。

优缺点：

- 充分考虑了上下文信息关系；
- 使用了中值频率平衡，它通过类的中位数(在训练集上计算)和每个类的频率之间的比值来重新加权类的预测。这就增加了低频率类的分数，这是一个更有噪声的分割掩码的代价，因为被低估的类的概率被高估了，并且可能导致在输出分割掩码中错误分类的像素增加。

2.MDRNNs（Multi-Dimensional Recurrent Neural Networks）模型

传统的RNN在一维序列学习问题上有着很好的表现，比如演讲（speech）和在线手写识别。但是 在多为问题中应用却并不到位。MDRNNs在一定程度上将RNN拓展到多维空间领域，使之在图像处理、视频处理等领域上也能有所表现。

该论文的基本思想是：将单个递归连接替换为多个递归连接，相应可以在一定程度上解决时间随数据样本的增加呈指数增长的问题。以下就是该论文提出的两个前向反馈和反向反馈的算法。

for $x_1 = 0$ to $X_1 - 1$ **do**
  for $x_2 = 0$ to $X_2 - 1$ **do**
    ...
    for $x_n = 0$ to $X_n - 1$ **do**
      initialize $a \leftarrow \sum_j in_j^{\mathbf{x}} w_{kj}$
      for $i = 1$ to n **do**
        if $x_i > 0$ **then**
          $a \leftarrow a + \sum_j h_j^{(x_1,\ldots,x_i-1,\ldots,x_n)} w_{kj}$
      $h_k^{\mathbf{x}} \leftarrow tanh(a)$

Algorithm 1: MDRNN Forward Pass

Defining $\hat{o}_j^{\mathbf{x}}$ and $\hat{h}_k^{\mathbf{x}}$ respectively as the derivatives of the objective function with respect to the activations of the j^{th} output unit and the k^{th} hidden unit at point $\mathbf{x}$, the backward pass is:

for $x_1 = X_1 - 1$ to 0 **do**
  for $x_2 = X_2 - 1$ to 0 **do**
    ...
    for $x_n = X_n - 1$ to 0 **do**
      initialize $e \leftarrow \sum_j \hat{o}_j^{\mathbf{x}} w_{jk}$
      for $i = 1$ to n **do**
        if $x_i < X_i - 1$ **then**
          $e \leftarrow e + \sum_j \hat{h}_j^{(x_1,\ldots,x_i+1,\ldots,x_n)} w_{jk}$
      $\hat{h}_k^{\mathbf{x}} \leftarrow tanh'(e)$

Algorithm 2: MDRNN Backward Pass

4.基于上采样/反卷积的分割方法

卷积神经网络在进行采样的时候会丢失部分细节信息，这样的目的是得到更具特征的价值。但是这个过程是不可逆的，有的时候会导致后面进行操作的时候图像的分辨率太低，出现细节丢失等问题。因此我们通过上采样在一定程度上可以不全一些丢失的信息，从而得到更加准确的分割边界。

接下来介绍几个非常著名的分割模型：

a.FCN(Fully Convolutional Network)

是的！讲来讲去终于讲到这位大佬了，FCN！在图像分割领域已然成为一个业界标杆，大多数的分割方法多多少少都会利用到FCN或者其中的一部分，比如前面我们讲过的Mask R-CNN。

在FCN当中的反卷积-升采样结构中，图片会先进性上采样（扩大像素）；再进行卷积——通过学习获得权值。FCN的网络结构如下图所示：

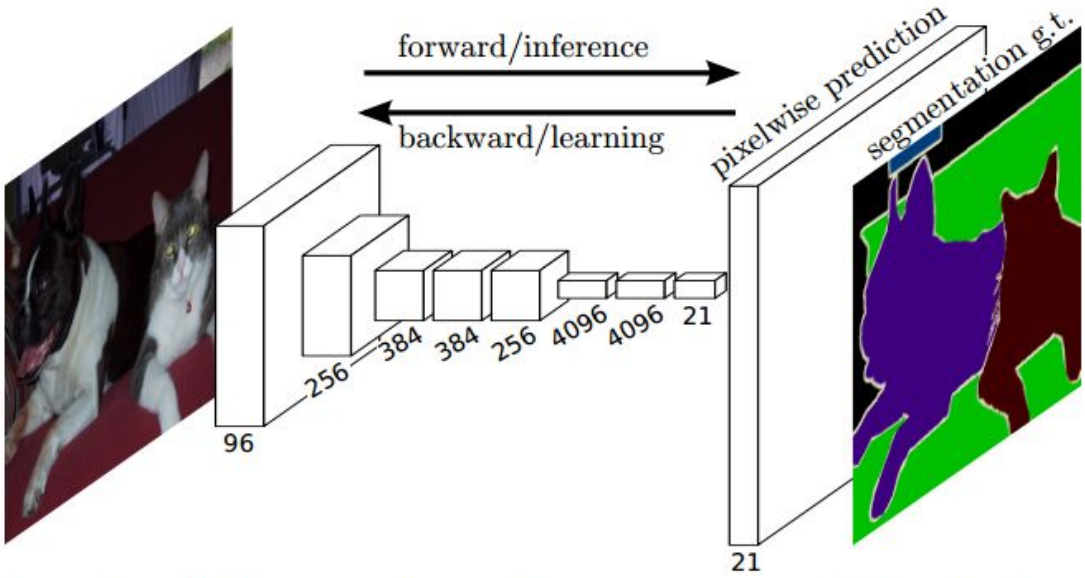


Figure 1. Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation.

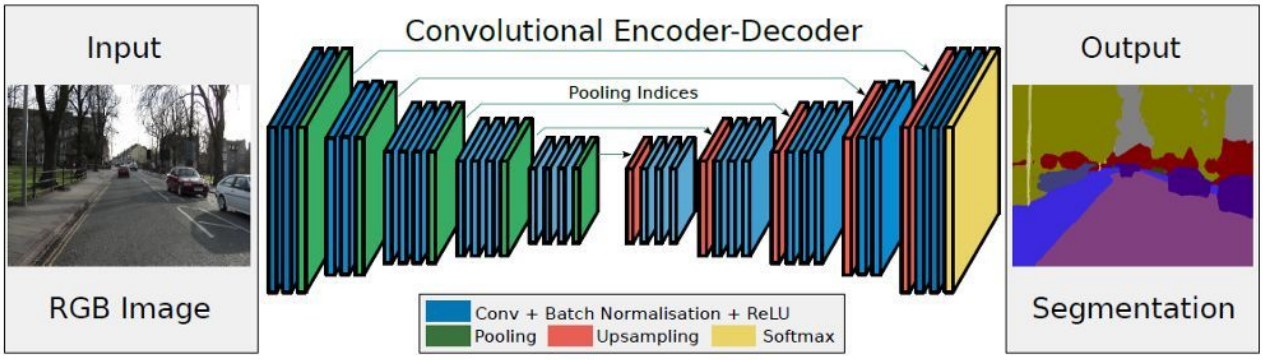
当然最后我们还是需要分析一下FCN，不能无脑吹啦~

优缺点：

- FCN对图像进行了像素级的分类，从而解决了语义级别的图像分割问题；
- FCN可以接受任意尺寸的输入图像，可以保留下原始输入图像中的空间信息；
- 得到的结果由于上采样的原因比较模糊和平滑，对图像中的细节不敏感；
- 对各个像素分别进行分类，没有充分考虑像素与像素的关系，缺乏空间一致性。

2.SetNet

SegNet是剑桥提出的旨在解决自动驾驶或者智能机器人的图像语义分割深度网络，SegNet基于FCN，与FCN的思路十分相似，只是其编码-解码器和FCN的稍有不同，其解码器中使用去池化对特征图进行上采样，并在分各种保持高频细节的完整性；而编码器不使用全连接层，因此是拥有较少参数的轻量级网络：



SetNet的优缺点：

- 保存了高频部分的完整性；
- 网络不笨重，参数少，较为轻便；
- 对于分类的边界位置置信度较低；
- 对于难以分辨的类别，例如人与自行车，两者如果有相互重叠，不确定性会增加。

以上两种网络结构就是基于反卷积/上采样的分割方法，当然其中最最最重要的就是FCN了，哪怕是后面大名鼎鼎的SegNet也是基于FCN架构的，而且FCN可谓是语义分割领域中开创级别的网络结构，所以虽然这个部分虽然只有两个网络结构，但是这两位可都是重量级嘉宾，希望各位能够深刻理解~

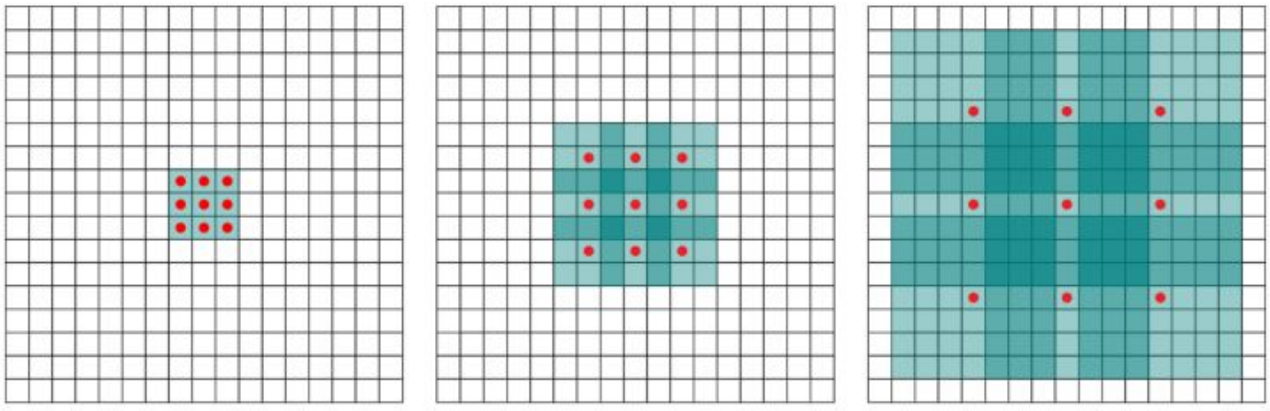
5.基于提高特征分辨率的分割方法

在这一个模块中我们主要给大家介绍一下基于提升特征分辨率的图像分割的方法。换一种说法其实可以说是恢复在深度卷积神经网络中下降的分辨率，从而获取更多的上下文信息。这一系列我将给大家介绍的是Google提出的DeepLab 。

DeepLab是结合了深度卷积神经网络和概率图模型的方法，应用在语义分割的任务上，目的是做逐像素分类，其先进性体现在DenseCRFs（概率图模型）和DCNN的结合。是将每个像素视为CRF节点，利用远程依赖关系并使用CRF推理直接优化DCNN的损失函数。

在图像分割领域，FCN的一个众所周知的操作就是平滑以后再填充，就是先进行卷积再进行pooling,这样在降低图像尺寸的同时增大感受野，但是在先减小图片尺寸（卷积）再增大尺寸（上采样）的过程中一定有一些信息损失掉了，所以这里就有可以提高的空间。

接下来我要介绍的是DeepLab网络的一大亮点：Dilated/Atrous Convolution，它使用的采样方式是带有空洞的采样。在VGG16中使用不同采样率的空洞卷积，可以明确控制网络的感受野。



图a对应3x3的1-dilated conv，它和普通的卷积操作是相同的；图b对应3x3的2-dilated conv，事迹卷积核的尺寸还是3x3（红点），但是空洞为1，其感受野能够达到7x7；图c对应3x3的4-dilated conv，其感受野已经达到了15x15.写到这里相信大家已经明白，在使用空洞卷积的情况下，加大了感受野，使每个卷积输出都包含了较大范围的信息。

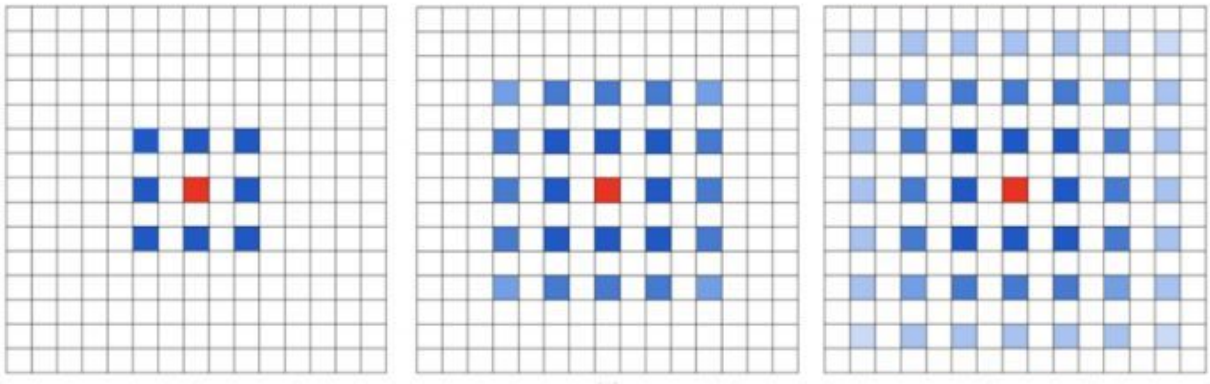
这样就解决了DCNN的几个关于分辨率的问题：

1）内部数据结构丢失；空间曾计划信息丢失；

2）小物体信息无法重建；

当然空洞卷积也存在一定的问题，它的问题主要体现在以下两方面： 1）网格效应

加入我们仅仅多次叠加dilation rate 2的 3x3 的卷积核则会出现以下问题



我们发现卷积核并不连续，也就是说并不是所有的像素都用来计算了，这样会丧失信息的连续性；

2）小物体信息处理不当

我们从空洞卷积的设计背景来看可以推测出它是设计来获取long-ranged information。然而空洞步频选取得大获取只有利于大物体得分割，而对于小物体的分割可能并没有好处。所以如何处理好不同大小物体之间的关系也是设计好空洞卷积网络的关键。

6.基于特征增强的分割方法

基于特征增强的分割方法包括：提取多尺度特征或者从一系列嵌套的区域中提取特征。在图像分割的深度网络中，CNN经常应用在图像的小方块上，通常称为以每个像素为中心的固定大小的卷积核，通过观察其周围的小区域来标记每个像素的分类。在图像分割领域，能够覆盖到更大部分的上下文信息的深度网络通常在分割的结果上更加出色，当然这也伴随着更高的计算代价。多尺度特征提取的方法就由此引进。

在这一模块中我先给大家介绍一个叫做SLIC，全称为simple linear iterative cluster的生成超像素的算法。

首先我们要明确一个概念：啥是超像素？其实这个比较容易理解，就像上面说的“小方块”一样，我们平常处理图像的最小单位就是像素了，这就是像素级（pixel-level）；而把像素级的图像划分成为区域级（district-level）的图像，把区域当成是最基本的处理单元，这就是超像

素啦。

算法大致思想是这样的，将图像从RGB颜色空间转换到CIE-Lab颜色空间，对应每个像素的（L，a，b）颜色值和（x，y）坐标组成一个5维向量V[l, a, b, x, y],两个像素的相似性即可由它们的向量距离来度量，距离越大，相似性越小。

算法首先生成K个种子点，然后在每个种子点的周围空间里搜索距离该种子点最近的若干像素，将他们归为与该种子点一类，直到所有像素点都归类完毕。然后计算这K个超像素里所有像素点的平均向量值，重新得到K个聚类中心，然后再以这K个中心去搜索其周围与其最为相似的若干像素，所有像素都归类完后重新得到K个超像素，更新聚类中心，再次迭代，如此反复直到收敛。

有点像聚类的K-Means算法，最终会得到K个超像素。

Mostahabi等人提出的一种前向传播的分类方法叫做Zoom-Out就使用了SLIC的算法，它从多个不同的级别提取特征：局部级别：超像素本身；远距离级别：能够包好整个目标的区域；全局级别：整个场景。这样综合考虑多尺度的特征对于像素或者超像素的分类以及分割来说都是很有意义的。

接下来的部分我将给大家介绍另一种完整的分割网络：**PSPNet：Pyramid Scene Parsing Network**

论文提出在场景分割是，大多数的模型会使用FCN的架构，但是FCN在场景之间的关系和全局信息的处理能力存在问题，其典型问题有：1.上下文推断能力不强；2.标签之间的关系处理不好；3.模型可能会忽略小的东西。

本文提出了一个具有层次全局优先级，包含不同子区域时间的不同尺度的信息，称之为金字塔池化模块。

该模块融合了4种不同金字塔尺度的特征，第一行红色是最粗糙的特征–全局池化生成单个bin输出，后面三行是不同尺度的池化特征。为了保证全局特征的权重，如果金字塔共有N个级别，则在每个级别后使用1×1 1×11×1的卷积将对于级别通道降为原本的1/N。再通过双线性插值获得未池化前的大小，最终concat到一起。其结构如下图：

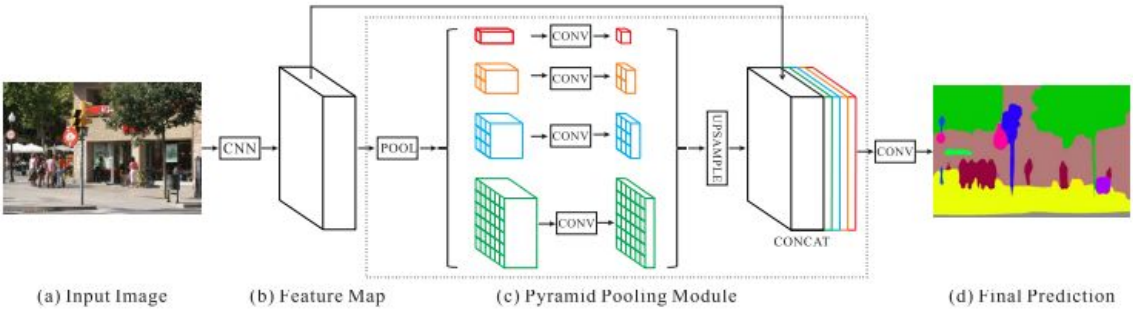


Figure 3. Overview of our proposed PSPNet. Given an input image (a), we first use CNN to get the feature map of the last convolutional layer (b), then a pyramid parsing module is applied to harvest different sub-region representations, followed by upsampling and concatenation layers to form the final feature representation, which carries both local and global context information in (c). Finally, the representation is fed into a convolution layer to get the final per-pixel prediction (d).

最终结果就是，在融合不同尺度的feature后，达到了语义和细节的融合，模型的性能表现提升很大，作者在很多数据集上都做过训练，最终结果是在MS-COCO数据集上预训练过的效果最好。

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU
FCN [26]	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
Zoom-out [28]	85.6	37.3	83.2	62.5	66.0	85.1	80.7	84.9	27.2	73.2	57.5	78.1	79.2	81.1	77.1	53.6	74.0	49.2	71.7	63.3	69.6
DeepLab [3]	84.4	54.5	81.5	63.6	65.9	85.1	79.1	83.4	30.7	74.1	59.8	79.0	76.1	83.2	80.8	59.7	82.2	50.4	73.1	63.7	71.6
CRF-RNN [41]	87.5	39.0	79.7	64.2	68.3	87.6	80.8	84.4	30.4	78.2	60.4	80.5	77.8	83.1	80.6	59.5	82.8	47.8	78.3	67.1	72.0
DeconvNet [30]	89.9	39.3	79.7	63.9	68.2	87.4	81.2	86.1	28.5	77.0	62.0	79.0	80.3	83.6	80.2	58.8	83.4	54.3	80.7	65.0	72.5
GCRF [36]	85.2	43.9	83.3	65.2	68.3	89.0	82.7	85.3	31.1	79.5	63.3	80.5	79.3	85.5	81.0	60.5	85.5	52.0	77.3	65.1	73.2
DPN [25]	87.7	59.4	78.4	64.9	70.3	89.3	83.5	86.1	31.7	79.9	62.6	81.9	80.0	83.5	82.3	60.5	83.2	53.4	77.9	65.0	74.1
Piecewise [20]	90.6	37.6	80.0	67.8	74.4	92.0	85.2	86.2	39.1	81.2	58.9	83.8	83.9	84.3	84.8	62.1	83.2	58.2	80.8	72.3	75.3
PSPNet	**91.8**	**71.9**	**94.7**	**71.2**	**75.8**	**95.2**	**89.9**	**95.9**	**39.3**	**90.7**	**71.7**	**90.5**	**94.5**	**88.8**	**89.6**	**72.8**	**89.6**	**64.0**	**85.1**	**76.3**	**82.6**
CRF-RNN† [41]	90.4	55.3	88.7	68.4	69.8	88.3	82.4	85.1	32.6	78.5	64.4	79.6	81.9	86.4	81.8	58.6	82.4	53.5	77.4	70.1	74.7
BoxSup† [7]	89.8	38.0	89.2	68.9	68.0	89.6	83.0	87.7	34.4	83.6	67.1	81.5	83.7	85.2	83.5	58.6	84.9	55.8	81.2	70.7	75.2
Dilation8† [40]	91.7	39.6	87.8	63.1	71.8	89.7	82.9	89.8	37.2	84.0	63.0	83.3	89.0	83.8	85.1	56.8	87.6	56.0	80.2	64.7	75.3
DPN† [25]	89.0	61.6	87.7	66.8	74.7	91.2	84.3	87.6	36.5	86.3	66.1	84.4	87.8	85.6	85.4	63.6	87.3	61.3	79.4	66.4	77.5
Piecewise† [20]	94.1	40.7	84.1	67.8	75.9	93.4	84.3	88.4	42.5	86.4	64.7	85.4	89.0	85.8	86.0	67.5	90.2	63.8	80.9	73.0	78.0
FCRNs† [38]	91.9	48.1	93.4	69.3	75.5	94.2	87.5	92.8	36.7	86.9	65.2	89.1	90.2	86.5	87.2	64.6	90.1	59.7	85.5	72.7	79.1
LRR† [9]	92.4	45.1	94.6	65.2	75.8	**95.1**	89.1	92.3	39.0	85.7	70.4	88.6	89.4	88.6	86.6	65.8	86.2	57.4	85.7	77.3	79.3
DeepLab† [4]	92.6	60.4	91.6	63.4	76.3	95.0	88.4	92.6	32.7	88.5	67.6	89.6	92.1	87.0	87.4	63.3	88.3	60.0	86.8	74.5	79.7
PSPNet†	**95.8**	**72.7**	**95.0**	**78.9**	**84.4**	94.7	**92.0**	**95.7**	**43.1**	**91.0**	**80.3**	**91.3**	**96.3**	**92.3**	**90.1**	**71.5**	**94.4**	**66.9**	**88.8**	**82.0**	**85.4**

为了捕捉多尺度特征，高层特征包含了更多的语义和更少的位置信息。结合多分辨率图像和多尺度特征描述符的优点，在不丢失分辨率的情况下提取图像中的全局和局部信息，这样就能在一定程度上提升网络的性能。

7.使用CRF/MRF的方法

首先让我们熟悉熟悉到底啥是MRF的CRF的。

MRF全称是Marcov Random Field，马尔可夫随机场，其实说起来笔者在刚读硕士的时候有一次就有同学在汇报中提到了隐马尔可夫、马尔可夫链啥的，当时还啥都不懂，小白一枚（现在是准小白hiahia），觉得马尔可夫这个名字贼帅，后来才慢慢了解什么马尔科夫链呀，马尔可夫随机场，并且在接触到图像分割了以后就对马尔科夫随机场有了更多的了解。

MRF其实是一种基于统计的图像分割算法，马尔可夫模型是指一组事件的集合，在这个集合中，事件逐个发生，并且下一刻事件的发生只由当前发生的事件决定，而与再之前的状态没有关系。而马尔可夫随机场，就是具有马尔可夫模型特性的随机场，就是场中任何区域都只与其临近区域相关，与其他地方的区域无关，那么这些区域里元素（图像中可以是像素）的集合就是一个马尔可夫随机场。

CRF的全称是Conditional Random Field，条件随机场其实是一种特殊的马尔可夫随机场，只不过是它是一种给定了一组输入随机变量X的条件下另一组输出随机变量Y的马尔可夫随机场，它的特点是埃及设输出随机变量构成马尔可夫随机场，可以看作是最大熵马尔可夫模型在标注问题上的推广。

在图像分割领域，运用CRF比较出名的一个模型就是全连接条件随机场（DenseCRF），接下来我们将花费一些篇幅来简单介绍一下。

CRF在运行中会有一个问题就是它只对相邻节点进行操作，这样会损失一些上下文信息，而全连接条件随机场是对所有节点进行操作，这样就能获取尽可能多的临近点信息，从而获得更加精准的分割结果。

在Fully connected CRF中，吉布斯能量可以写作：

我们重点关注二元部分：

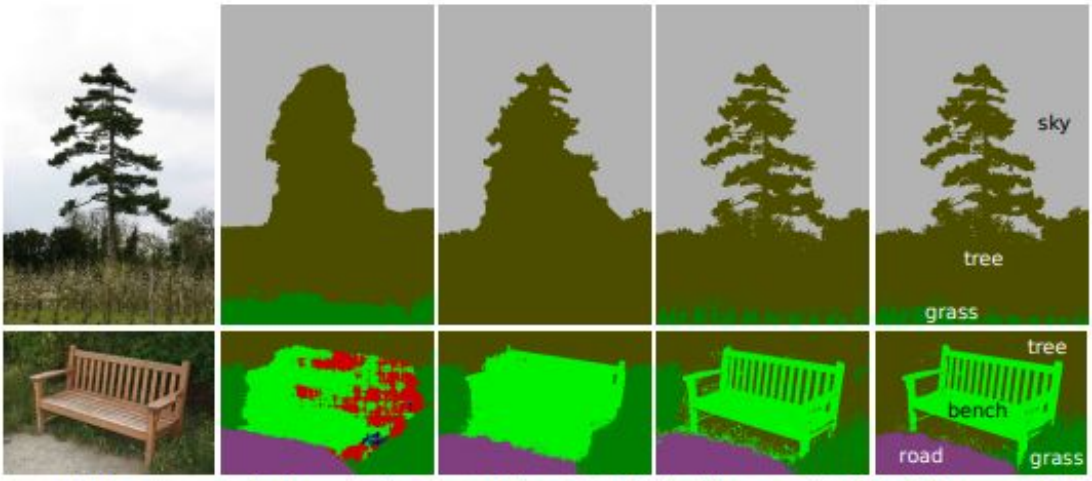
$$E(x) = \sum_i \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, y_j)$$

其中k(m)为高斯核，写作：

$$k^m(f_i, f_j) = \omega_1 \exp\left(-\frac{\|p_i - p_j\|^2}{2\sigma_\alpha^2} - \frac{\|I_i - I_j\|^2}{2\sigma_\beta^2}\right) + \omega_2 \exp\left(-\frac{\|p_i - p_j\|^2}{2\sigma_\gamma^2}\right)$$

该模型的一元势能包含了图像的形状，纹理，颜色和位置，二元势能使用了对比度敏感的的双核势能，CRF的二元势函数一般是描述像素点与像素点之间的关系，鼓励相似像素分配相同的标签，而相差较大的像素分配不同标签，而这个"距离"的定义与颜色值和实际相对距离有关，这样CRF能够使图像尽量在边界处分割。全连接CRF模型的不同就在于其二元势函数描述的是每一个像素与其他所有像素的关系，使用该模型在图像中的所有像素对上建立点对势能从而实现极大地细化和分割。

在分割结果上我们可以看看如下的结果图：



(a) Image (b) Unary classifiers (c) Robust P^n CRF (d) Fully connected CRF, MCMC inference, 36 hrs (e) Fully connected CRF, our approach, 0.2 seconds

Figure 1: Pixel-level classification with a fully connected CRF. (a) Input image from the MSRC-21 dataset. (b) The response of unary classifiers used by our models. (c) Classification produced by the Robust P^n CRF [9]. (d) Classification produced by MCMC inference [17] in a fully connected pixel-level CRF model; the algorithm was run for 36 hours and only partially converged for the bottom image. (e) Classification produced by our inference algorithm in the fully connected model in 0.2 seconds.

可以看到它在精细边缘的分割比平常的分割方法要出色得多，而且文章中使用了另一种优化算法，使得本来需要及其大量运算的全连接条件随机场也能在很短的时间里给出不错的分割结果。

至于其优缺点，我觉得可以总结为以下几方面：

- 在精细部位的分割非常优秀；
- 充分考虑了像素点或者图片区域之间的上下文关系；

- 在粗略的分割中可能会消耗不必要的算力；
- 可以用来恢复细致的局部结构，但是相应的需要较高的代价。

OK，那么本次的推送就到这里结束啦，本文的主要内容是对图像分割的算法进行一个简单的分类和介绍。综述对于各位想要深入研究的看官是非常非常重要的资源：大佬们经常看综述一方面可以了解算法的不足并在此基础上做出改进；萌新们可以通过阅读一篇好的综述入门某一个学科。

学术交流群

欢迎加入公众号读者群一起和同行交流，目前有**SLAM、算法竞赛、图像检测分割、人脸人体、医学影像、自动驾驶、综合**等微信群（以后会逐渐细分），请扫描下面微信号加群，备注：”昵称+学校/公司+研究方向“，例如：”张三 + 上海交大 + 视觉SLAM“。**请按照格式备注，否则不予通过**。添加成功后会根据研究方向邀请进入相关微信群。请勿在群内发送**广告**，否则会请出群，谢谢理解~

推荐阅读

计算机视觉方向简介 | 从全景图恢复三维结构

计算机视觉方向简介 | 阵列相机立体全景拼接

计算机视觉方向简介 | 单目微运动生成深度图

计算机视觉方向简介 | 深度相机室内实时稠密三维重建

计算机视觉方向简介 | 深度图补全

计算机视觉方向简介 | 人体骨骼关键点检测综述

计算机视觉方向简介 | 人脸识别中的活体检测算法综述

计算机视觉方向简介 | 目标检测最新进展总结与展望

计算机视觉方向简介 | 唇语识别技术

计算机视觉方向简介 | 三维深度学习中的目标分类与语义分割

计算机视觉方向简介 | 基于单目视觉的三维重建算法

计算机视觉方向简介 | 用深度学习进行表格提取

计算机视觉方向简介 | 立体匹配技术简介

计算机视觉方向简介 | 人脸表情识别

计算机视觉方向简介 | 人脸颜值打分

计算机视觉方向简介 | 深度学习自动构图

计算机视觉方向简介 | 基于RGB-D的3D目标检测

计算机视觉方向简介 | 人体姿态估计

计算机视觉方向简介 | 三维重建技术概述

计算机视觉方向简介 | 视觉惯性里程计(VIO)

目标检测技术二十年综述

最全综述 | 医学图像处理