模式识别文献综述

基于深度学习的图像语义分割算法

姓名：罗雁天

院系：清华大学电子系

学号：2018310742

更新：March 9, 2019

目录

摘要

在计算机视觉领域，图像分割指的是将数字图像细分为多个图像子区域 (像素的集合)(也被称作超像素) 的过程。图像分割的目的是简化或改变图像的表示形式，使得图像更容易理解和分析。图像分割通常用于定位图像中的物体和边界 (线，曲线等)。更精确的，图像分割是对图像中的每个像素加标签的一个过程，这一过程使得具有相同标签的像素具有某种共同视觉特性。

简单来说，图像分割可以看做是像素级别的分类，其在医疗领域、自动驾驶等方面有着重要的应用，在目前的算法研究中，图像分类可以分为语义分割和实例分割。

本文主要从图像语义分割的方向进行调研，介绍了近几年来基于深度学习的图像语义分割算法，并且对这些算法进行对比与总结，最后提出了对图像分割未来方向的展望。

关键字：图像语义分割; 深度学习; 全卷积网络; 条件随机场

Abstract

In the field of computer vision, image segmentation refers to the process of subdividing a digital image into multiple image sub-regions (a collection of pixels) (also referred to as superpixels). The purpose of image segmentation is to simplify or change the representation of the image, making the image easier to understand and analyze. Image segmentation is often used to locate objects and boundaries (lines, curves, etc.) in an image. More precisely, image segmentation is a process of tagging each pixel in an image, which results in pixels with the same tag having some common visual characteristics.

In short, image segmentation can be regarded as a pixel-level classification, which has important applications in the medical field and automatic driving. In the current algorithm research, image classification can be divided into semantic segmentation and instance segmentation.

This paper mainly investigates the direction of image semantic segmentation, introduces the image semantic segmentation algorithm based on deep learning in recent years, and compares and summarizes these algorithms. Finally, it puts forward the prospect of the future direction of image segmentation.

Key Words: Image Semantic Segmentation; Deep Learning; Fully Convolutional Networks; Conditional Random Field

第 1 章 引言

图像分割 (Segmentation) 指的是将数字图像细分为多个图像子区域 (像素的集合)(也被称作超像素) 的过程。图像分割的目的是简化或改变图像的表示形式，使得图像更容易理解和分析。图像分割通常用于定位图像中的物体和边界 (线，曲线等)。更精确的，图像分割是对图像中的每个像素加标签的一个过程，这一过程使得具有相同标签的像素具有某种共同视觉特性。

图像分割又可以分为图像语义分割 (Semantic Segmentation) 与实例分割 (Instance Segmentation)。语义分割是在像素级别上的分类，属于同一类的像素都要被归为一类，因此语义分割是从像素级别来理解图像的。而实例分割不但要进行像素级别的分类，还需在具体的类别基础上区别开不同的实例。比如说图像有多个人甲、乙、丙，那边他们的语义分割结果都是人，而实例分割结果却是不同的对象。

图像分割在实际应用中非常广泛，在医学影像中来用来进行肿瘤和其他病理的定位、组织体积的测量、计算机引导的手术等，在卫星图像中定位物体，用于人脸识别、指纹识别，近几年在自动驾驶领域图像分割也起着至关重要的作用。

1.1 评价指标

在图像分割领域主要有如下 4 个评价指标：

- pixel accuracy (Acc): 像素准确率；
- mean pixel accuracy of different categories (mAcc): 类平均像素准确率；
- mean Intersection-over-Union of different categories (mIoU): 类平均识别准确度；
- frequency-weighted IoU (fwIoU): 频率加权的识别准确度。

其中 IoU(Intersection-over-Union) 表示预测位置与真实位置之间的重叠程度，IoU 越高，预测的位置越准确。图1.1表示了 IoU 的几何意义。

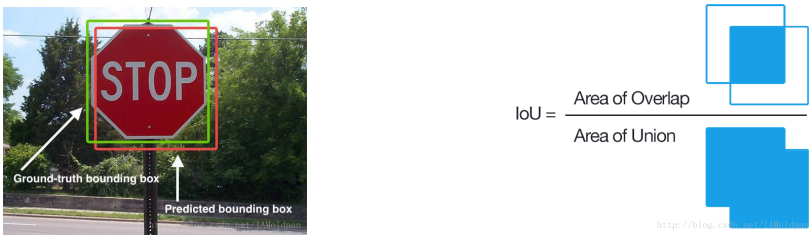

图 1.1: IoU 表示含义示意图

四个指标的计算方式如式1.1所示：

定义 1.1: 计算方式

$$
\begin{aligned}
Acc &= \sum_i \frac{n_{ii}}{s} \\
mAcc &= \frac{1}{n_C} \sum_i \frac{n_{ii}}{s_i} \\
mIoU &= \frac{1}{n_C} \sum_i \frac{n_{ii}}{s_i + \sum_j n_{ji} - n_{ii}} \\
fwIoU &= \frac{1}{s} \sum_i s_i \frac{n_{ii}}{s_i + \sum_j n_{ji} - n_{ii}}
\end{aligned}
\tag{1.1}
$$

♣

1.2 数据集

图像分割常用的数据集包括 PASCAL VOC[1]、MS COCO[2] 等，包含深度信息图像的数据集有 NYUv2[3] 等。

- PASCAL VOC(The PASCAL Visual Object Classification) 是目标检测、分类、图像分割领域一个有名的数据集。从 2005 年到 2012 年，共举办了 8 个不同的挑战赛。用于图像分割的 VOC2012 数据集提供原图以及图像语义分割和图像实例分割两种 png 图 (如图1.2所示)，共分为 20 类，包括背景为 21 类，分别如下：
  - Person: person;
  - Animal: bird, cat, cow, dog, horse, sheep;
  - Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train;
  - Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor.

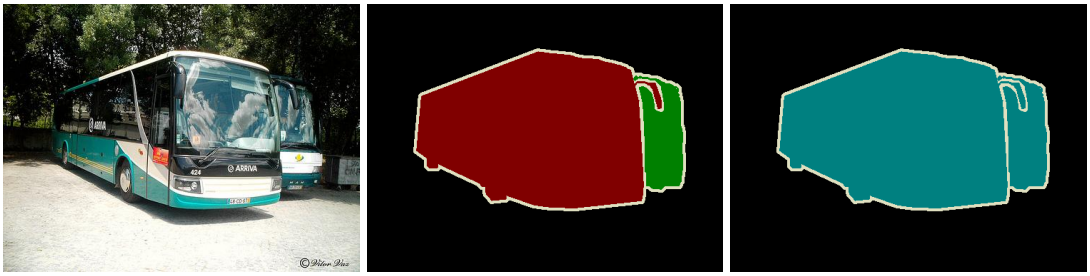

图 1.2: PASCAL VOC 图像分割数据集示例 (左图：原始图像；中图：实例分割的标签图；右图：语义分割的标签图)

- MS COCO(Common Objects in COntext) 是微软建立的数据集。这个数据集也用于多种竞赛：图像标题生成、目标检测、关键点检测和图像分割。图像包括 91 类目标，328000 影像和 2500000 个 label。图1.3展示了 MS COCO 举办的各个竞赛中数据集的示例。

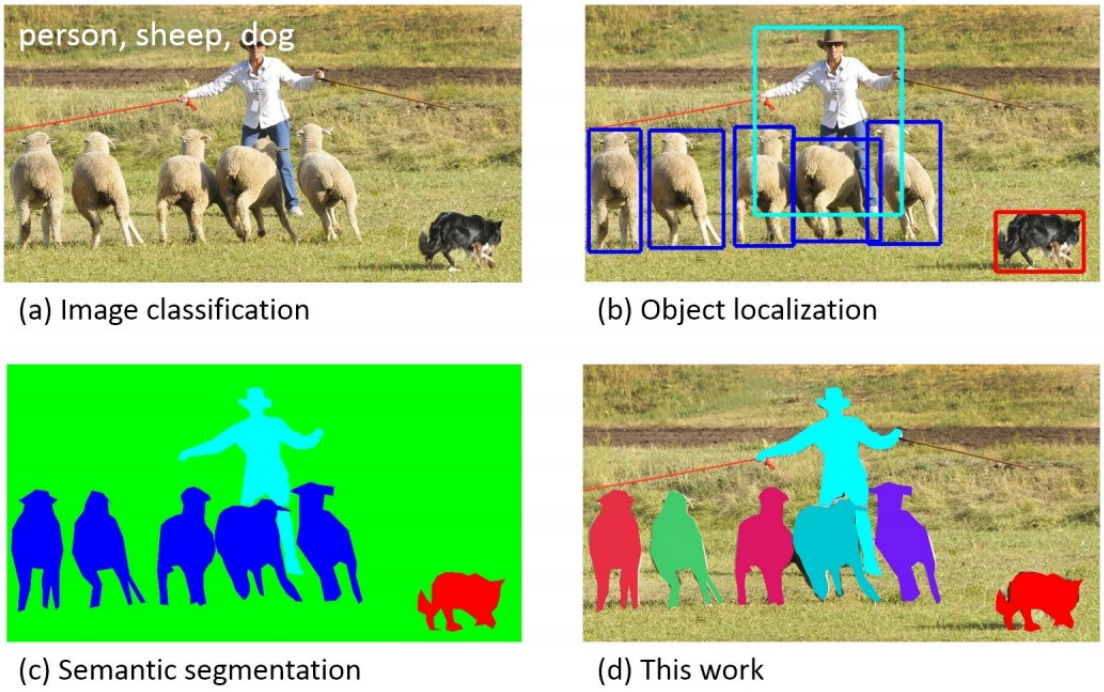

图 1.3: MS COCO 数据集多种图像任务示例

- NYUv2 数据集是使用 Kinect 采集的一系列包含深度信息的图像，包含如下几个部分：
  - 有标签的：视频数据的一个子集，伴随着密集多标签。此数据已被预处理以填补缺少的深度标签；
  - 原始数据集：利用 Kinect 测得的原始的 RGB、Depth、加速度数据；
  - 工具箱：用于操作数据和标签的有用的工具；
  - 用于评估的训练和测试部分。

  有标签的数据如图1.4所示。

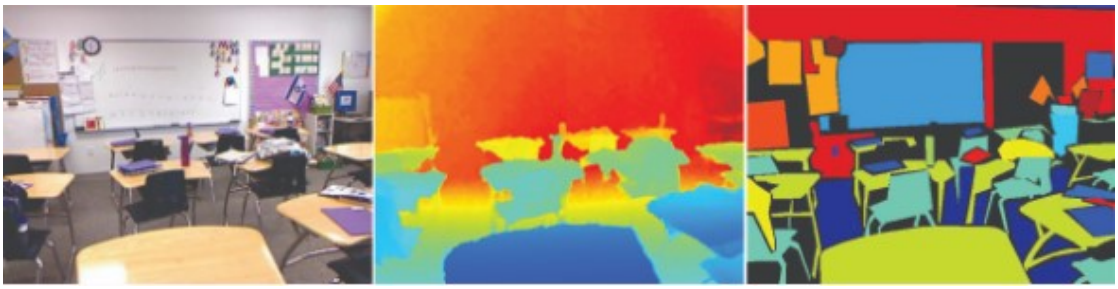

图 1.4: NYUv2 有标签的数据集示例。(左图：Kinect 相机输出的图像；中图：预处理深度信息；右图：添加一系列标签图)

第 2 章 图像语义分割算法描述

在深度学习方法流行之前，TextonForest 和基于随机森林分类器等语义分割方法是用得比较多的方法。不过在深度卷积网络流行之后，深度学习方法比传统方法提升了很多，所以这里就不详细讲传统方法了。最初的学习方法应用于图像分割就是 Patch classification。Patch classification 方法，顾名思义，图像是切成块喂给深度模型的，然后对像素进行分类。使用图像块的主要原因是因为全连接层需要固定大小的图像。由于在全卷积网络出现之后对语义分割的效果提升了很多，因此在此也不再详述 Patch Classification 的方法。

2.1 全卷积网络 (FCN)

全卷积网络 (Fully convolutional networks, FCN[4]) 于 2015 年被首次提出，并且获得了当年 CVPR 的 best paper。相比于之前使用带全连接层的卷积神经网络进行图像分割，FCN 主要涉及以下三个技术：

- 卷积化 (Convolutionalization);
- 上采样 (Upsampling)，也叫反卷积 (Deconvoltion);
- 跳跃结构 (Skip Architecture).

2.1.1 卷积化

FCN 将传统 CNN 中的全连接层转化为一个个的卷积层。如图2.1所示，上图为传统的 CNN 网络结构 (AlexNet[5])，前 5 层是卷积层，第 6 层和第 7 层分别是一个长度为 4096 的全连接层，第 8 层是一个长度为 1000 的全连接层。在 FCN 中，将最后的三层全连接层全都替换为卷积层，卷积核大小分别为 (4096,1,1), (4096,1,1), (1000,1,1)。

网络结构如图2.2所示，虚线上半部分为全卷积网络 (蓝：卷积层，绿：max pooling)，输入可为任意尺寸的图像，下半部分为反卷积 (上采样) 结构，最后输出与原图像大小相同，通道数为 21(PASCAL VOC 数据集 20 类物体类别 +1 类背景).

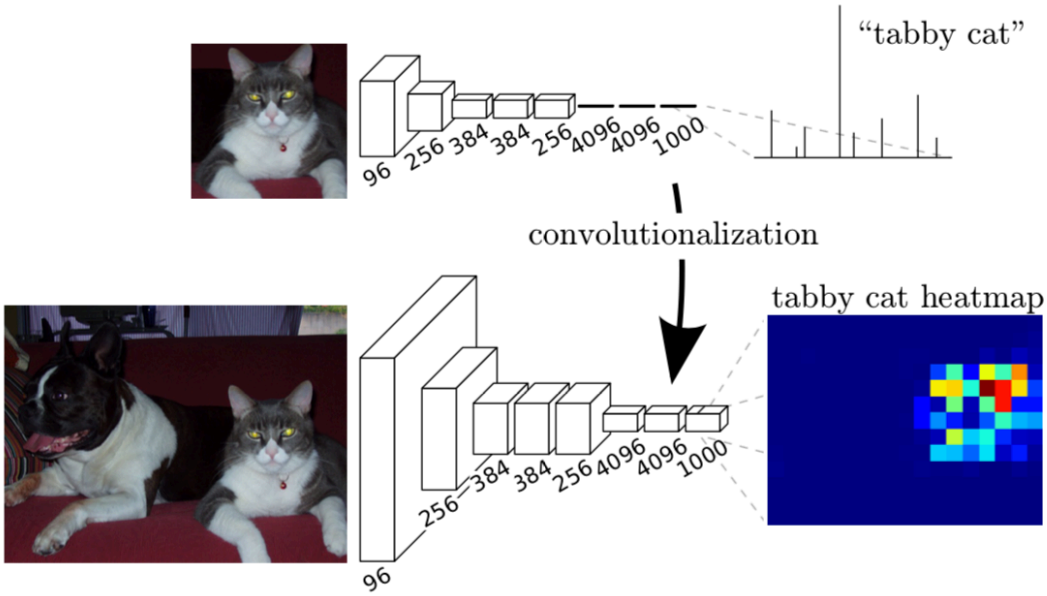

图 2.1: FCN 与传统 CNN 对比图

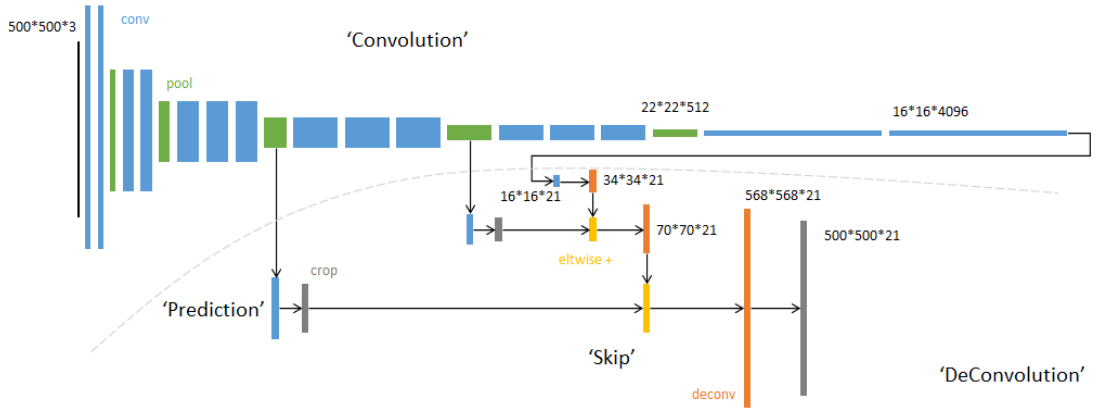

图 2.2: FCN 网络结构示意图

2.1.2 反卷积

经过全卷积网络之后得到的 feature map 相比于原图像要小，为了得到和原图像一样的 feature map 以便进行像素级的分类，FCN 采用反卷积的方式将最后一层的 feature map 进行放大 (图2.2中虚线以下橙色的部分)。

图2.3展示了正常卷积与反卷积的对比图，左图展示的正常的 no padding no strides 情况下的卷积操作，可以看出，卷积之后 feature map 会变小，右图展示的是 no padding no strides 情况下的反卷积操作，可以看出反卷积之后 feature map 会增大。简单来看，反卷积其实可以看做先对 feature map 进行上采样增加像素，然后再进行卷积的过程，卷积的参数值通过训练得到。

https://github.com/vdumoulin/conv_arithmetic给出了各种类型卷积的动态图 (正常卷积、反卷积以及之后要涉及的空洞卷积)，可以通过动态图进一步认识各

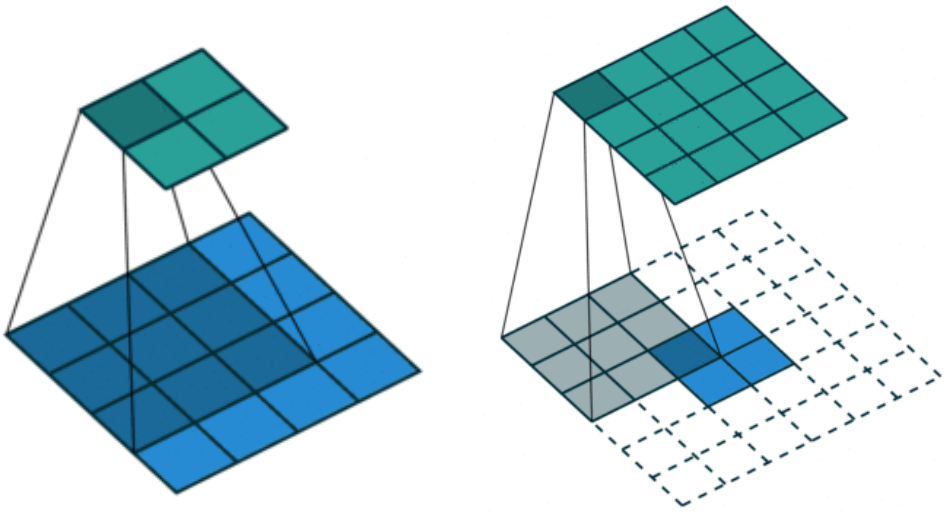

图 2.3: 正常卷积与反卷积对比图

种卷积操作。论文 [6] 给出了详细的数学公式以及卷积前后 feature map 大小的变化公式，可以更深一步的认识卷积。

2.1.3 跳跃结构

如图2.4所示展示了 FCN 中跳跃结构示意图。从图中可以看出，对原图进行卷积 conv1、pool1 后图像缩小为 1/2；对图像进行第二次卷积 conv2、pool2 后图像缩小为 1/4；对图像进行第三次卷积 conv3、pool3 后图像缩小为 1/8，此时保留 pool3 的 featuremap；对图像进行第四次卷积 conv4、pool4 后图像缩小为 1/16，此时保留 pool4 的 featuremap；对图像进行第五次卷积 conv5、pool5 后图像缩小为 1/32，然后把原来 CNN 操作过程中的全连接编程卷积操作的 conv6、conv7，图像的 featuremap 的大小依然为原图的 1/32, 此时图像不再叫 featuremap 而是叫 heatmap。其实直接使用前两种结构就已经可以得到结果了，这个上采样是通过反卷积 (deconvolution) 实现的，对第五层的输出 (32 倍放大) 反卷积到原图大小。但是得到的结果还上不不够精确，一些细节无法恢复。于是将第四层的输出和第三层的输出也依次反卷积，分别需要 16 倍和 8 倍上采样，结果过也更精细一些了。这种做法的好处是兼顾了 local 和 global 信息。

2.1.4 实验结果

在此篇论文中，作者分别使用了一次反卷积、两次反卷积以及三次反卷积操作进行实验，并且使用修改过的 VGG 网络结构进行训练，最后在 PASCAL VOC 数据集上的性能指标如图2.5所示。由结果图中可以直观的看出，FCN 的图像分割结果和 ground truth 相比已经有了较好的效果，但是边缘部分差距还是较大，在之后的几种算法中会有进一步的提升。

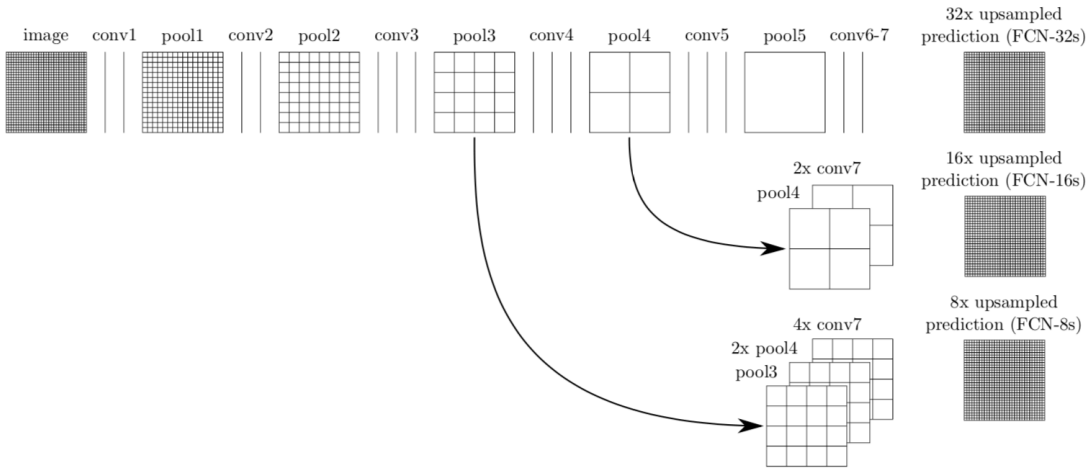

图 2.4: 跳跃结构示意图

	pixel acc.	mean acc.	mean IU	f.w. IU
FCN-32s-fixed	83.0	59.7	45.4	72.0
FCN-32s	89.1	73.3	59.4	81.4
FCN-16s	90.0	75.7	62.4	83.0
FCN-8s	**90.3**	**75.9**	**62.7**	**83.2**

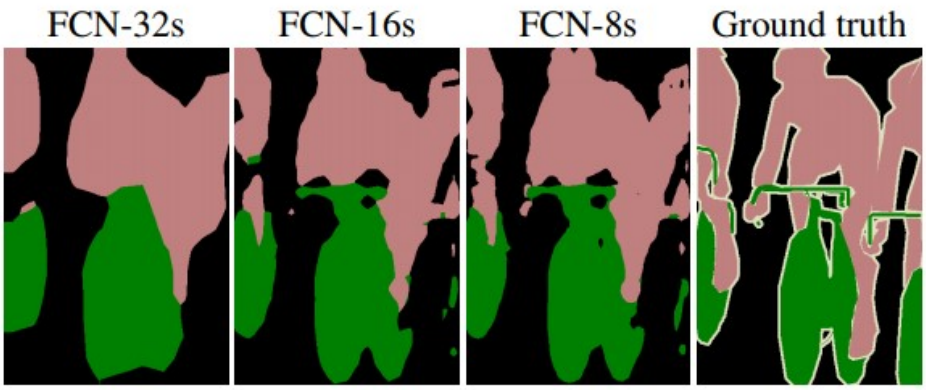

图 2.5: 实验结果性能指标与 PASCAL VOC 数据集上的实验效果图

2.2 DeepLab

从 FCN 的实验结果图中可以看出来，虽然分割的整体区域比较正确了，但是分割效果还是比较粗糙，细节不明显。DeepLab v1(文章 [7]) 使用 Hole 算法 (Atrous Algorithm) 和条件随机场 (CRF) 来进一步的提升分割效果，其算法结构图如图2.6所示。DeepLab v1 收录于 ICLR 2015，是 DeepLab 系列的第一篇文章，之后我们还会介绍该系列的后续文章。

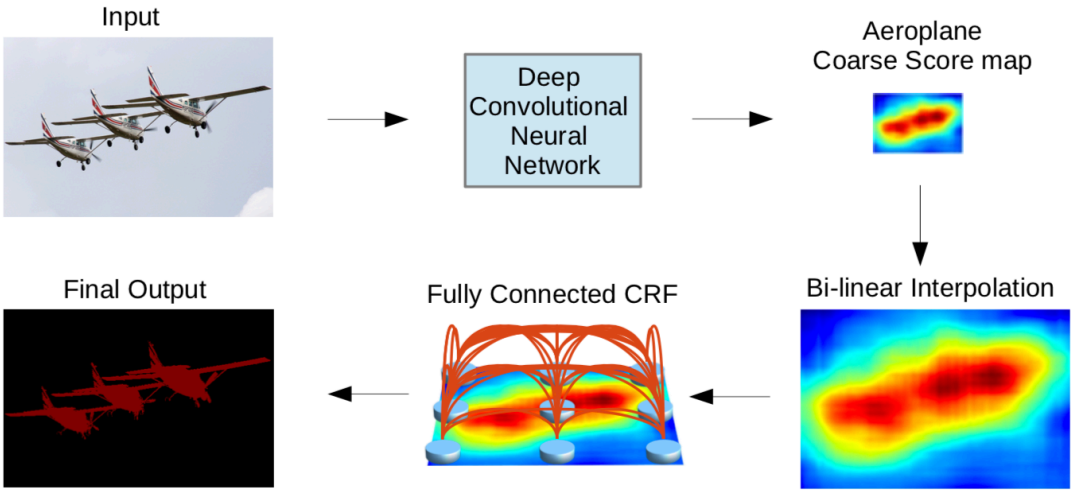

图 2.6: DeepLab 算法结构示意图

2.2.1 Hole 算法

由于普通的卷积感受野较小，需要增加池化层来增加感受野，但是池化层又会损失信息，所以使用空洞卷积在不损失信息的情况下增加感受野的范围。

Hole 算法又可以看做带孔 (Hole) 卷积，传统的卷积或者 pooling 中，一个 filter 中相邻的权重作用在 feature map 上的位置都是物理上连续的。而在 Hole 算法中，一个 filter 中相邻的权重不一定作用在 feature map 上的位置都是物理上连续的，而是跟 hole size 相关的。如图2.7所示表示的是卷积核大小 kernel_size=3，输入步长 input_stride(也就是 hole size)=2，输出步长 output_stride=1 的一维带孔卷积示意图。可以看出卷积核作用在输入 feature map 上的位置不是连续的。后续章节会有关于 Hole 算法的更详细的介绍，在此不再赘述。

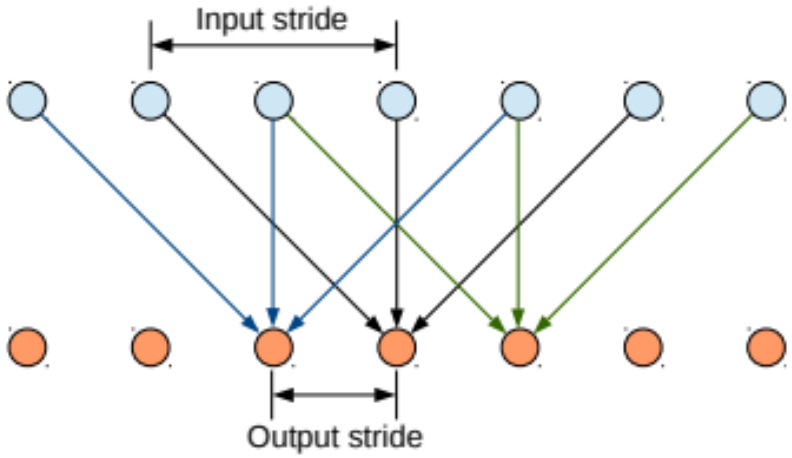

图 2.7: Hole 算法示意图

2.2.2 条件随机场

只使用全卷积网络能够预测到目标的大概位置但是位置比较模糊，论文 [8] 中提出的全连接条件随机场尝试找到图像像素之间的关系：相近且相似的像素大概率为同一标签，考虑像素的概率分配标签，通过迭代来细化分割的结果。

条件随机场服从吉布斯分布，如式 (2.1) 所示，其中 $E(X)$ 是 x 取某个值的能量，$Z(I)$ 是归一化的函数。

$$P(X|I) = \frac{1}{Z(I)} \exp(-E(X|I)) \tag{2.1}$$

为了做图像分割，只需要后验概率最大，因此只需能量函数最小即可，因此条件随机场优化的目标函数便是能量函数 $E(X)$(式 (2.2)).

$$E(x) = \sum_i \psi_i(x_i) + \sum_{i,j} \psi_{i,j}(x_i, x_j) \tag{2.2}$$

能量方程的第一项 $\psi_i(x_i)$(式 (2.3)) 称为一元势函数，用于衡量当像素点 i 的颜色值为 y_i 时，该像素点属于类别标签 x_i 的概率。在 DeepLab 中，此概率是通过 CNN 的输出得到的。

$$\psi_i(x_i) = -\log(P(x_i)) \tag{2.3}$$

能量方程的第二项 $\psi_{i,j}(x_i, x_j)$ 称之为成对势函数 (pairwise)，用于衡量两事件同时发生的概率 $p(x_i, x_j)$，我们希望两个相邻的像素点，如果颜色值 y_i, y_j 非常接近，那么这两个像素点 x_i, x_j 属于同一个类别的概率应该比较大才对；反之如果颜色差异比较大，那么我们分割的结果从这两个像素点裂开的概率应该比较大才对。这一能量项正是为了让我们的分割结果尽量从图像边缘的地方裂开，也就是为了弥补之前 FCN 边缘的地方分割的不足，我们可以采用式 (2.4) 来计算。

$$\psi_{i,j}(x_i, x_j) = u(x_i, x_j) \sum_{m=1}^{M} w^m K_G^m(f_i, f_j) \tag{2.4}$$

其中 K_G 是一个高斯核，用于度量像素点 i 和 j 的特征向量相似度的一个高斯权重项。特征向量 f_i 我们可以用 (x, y, R, G, B) 表示，也就是以像素点的像素值和坐标位置作为特征向量。然后 $u(x_i, x_j)$ 表示两个标签之间的一个兼容性度量。通过最小化式 (2.2) 的能量函数，我们就可以实现 CRF 隐变量 X 的推理。

图2.8显示了 DeepLab 中使用 CRF 迭代来细化分割结果的示意图，从图中可以看出，使用 CRF 迭代可以使得分割的边缘效果逐渐增强。

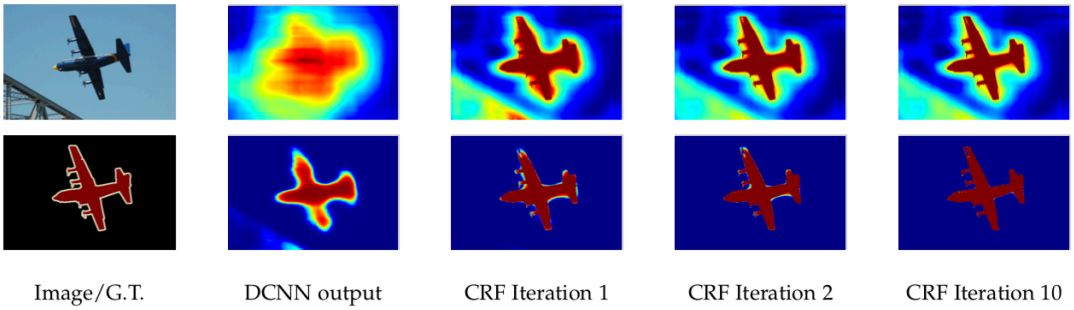

图 2.8: 使用 CRF 细化分割效果。可以看出，随着迭代次数的增加，图像分割的效果逐渐增强

2.2.3 网络结构

2.3 DilatedConv

2.4 DeepLab v2

2.5 PSPNet

2.6 DeepLab v3

第 3 章 方法结果对比

第 4 章 总结

参考文献

[1] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. *arXiv e-prints*, page arXiv: 1405.0312, May 2014.

[3] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.

[4] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[6] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *ArXiv e-prints*, mar 2016.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[8] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.