《Fully Convolutional Networks for Semantic Segmentation》文献阅读

# 研究意义

图像语义分割是计算机视觉领域的一个重要任务，旨在为图像中的每个像素分配一个预先定义好的表示其语义类别的标签[1]。语义分割最终的结果是每个像素的类别图，像素的分类任务完成，自然将图像分割为不同的类别区域。语义分割技术在自动驾驶、无人机、智慧安防等领域应用广泛。

图像分类、语义分割、目标探测和实例分割的区别可用图 1表示，分类任务只需要给出分类的最终决策结果，无需逐像素识别。而语义分割需要对每个像素进行分类，若需要区分同一类下的不同实体，则称为实例分割，目标探测只需要检测出目标的具体位置，无需对全图进行分类。

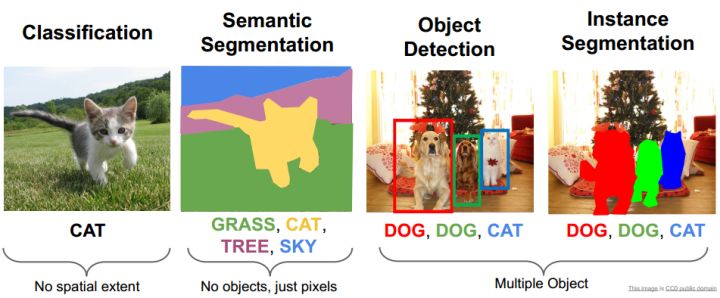


图 1 图像分类、语义分割、目标探测和实例分割的区别图[2]

随着深度学习技术的发展，卷积神经网络（Convolutional Neural Networks ,CNN）在图像语义分割任务中取得了显著的成果。语义分割的全卷积网络（Fully Convolutional Networks for Semantic Segmentation ,FCN）由Long等人在2015 IEEE Conference on Computer Vision and Pattern Recognition上提出。该网络由传统的分类网络做卷积化产生，无需复杂的预处理或后处理步骤，能够对任意输入大小的图像进行语义分割，实现了端到端、像素到像素的分割。

# 问题描述

## 卷积神经网络的架构（以VGG16为例）

CNN的架构主要包括卷积层、激活函数、池化层、全连接层、归一化层和丢弃层。其中，卷积层可以提取图像的细节特征，激活函数负责为网络引入非线性特征，池化层用于降低特征映射的维度，从而减少计算量和参数数量，全连接层主要将输入的高维特征映射到一维，负责最终的分类。

VGG16在图像分类任务中取得了优异的成绩，该网络的架构图如图 2所示，网络结构如表 1（D列）所示。VGG16共有16个权重层：卷积层13层，全连接3层，池化层不涉及权重。

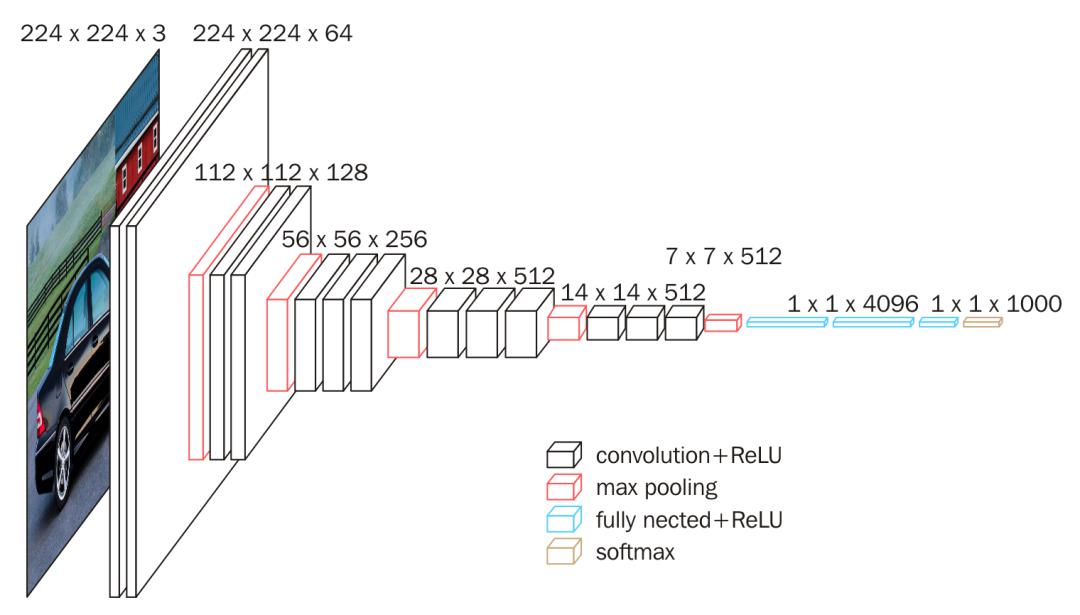


图 2 VGG16网络架构图[3]

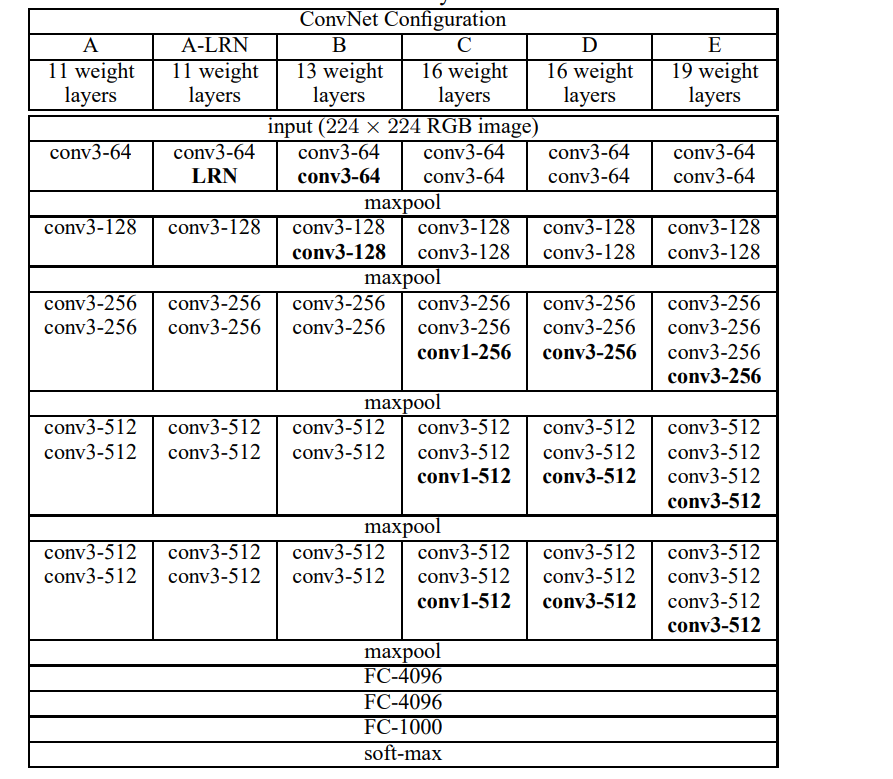


表 1 VGG网络构架[4]

VGG16网络结构简单明了，卷积核尺寸大小相对固定，没有复杂的层级，网络深度较深，能较好的提取图像的深层特征，实现分类任务。

## 分类网络语义分割的可行性

尽管VGG16能较好的完成图像分类任务，但是直接将其投入到语义分割任务，仍存在如下问题：

1. VGG16网络的输出为的向量，表示输入图像属于1000类中某一类的概率，实质上最终结果为一个标签值，与语义分割要求的像素级分类相差较远。
2. VGG16网络中的全连接层输入固定，意味着整个网络的输入图像的尺寸需要固定，否则无法继续推理，对于使用场景广泛的语义分割任务，VGG16对尺寸的要求是严苛的。
3. VGG16网络中，多次使用了最大池化层，虽然减少了后续的计算量，造成了图像分辨率下降，低层空间信息没有办法在高维特征中表达，故对于像素级分割的语义分割任务，可能效果不佳。

# 解决方案

## 全卷积网络

针对这一问题，Long等人通过替换掉传统分类网络的全连接层，丢弃分类层，使网络只包含卷积层，称为全卷积网络（FCN），这样可以使输入图像的尺寸不固定，实现任意尺寸的输入。该过程如图 3所示。

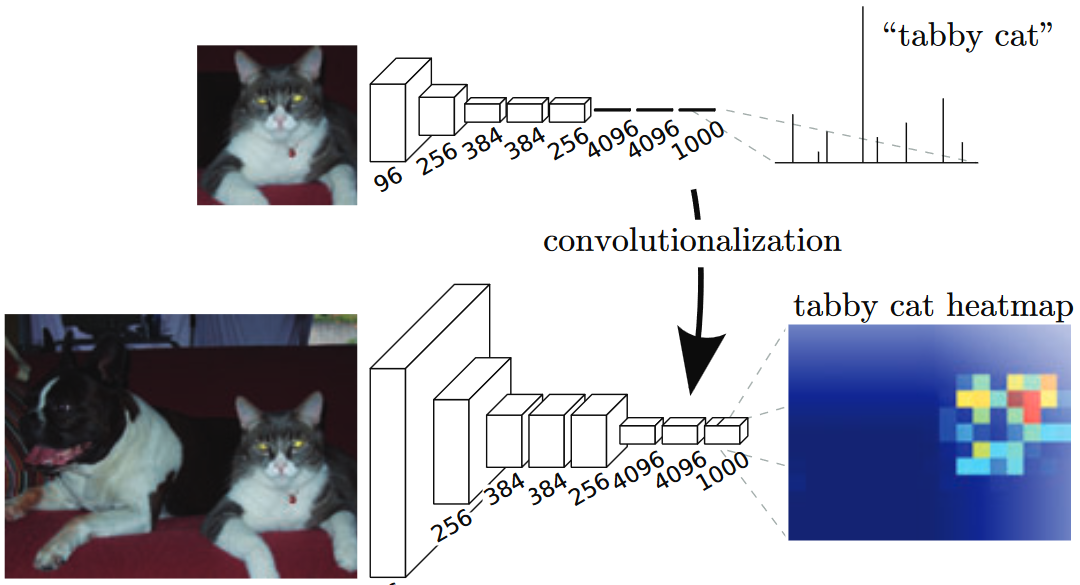


图 3 传统分类网络卷积化

这样处理的好处在于(1并行计算：可以对整幅图像进行并行计算，而不需要逐个处理图像的小patch（2参数共享：使得像素之间共用卷积核减少了计算量（3密集预测：这样的网络输出形式使其自然地适合于图像分割这种密集预测任务。

然而由于传统分类网络池化层的存在，图像被不断下采样，使得预测结果与输入尺寸不一致。故需要一定的“还原”的方法，即上采样技术。

## 上采样技术

1. Shift-and-stitch策略

Long等人介绍了Shift-and-stitch策略，为了使得网络输出为原始图像尺寸大小，经过网络后，图像被下采样了倍，即需要将原始图像每次向左，向上平移个像素，得到个图像，将这些图像都经过一次网络，得到一系列降采样图像，最后将其交错拼接成原始图像大小，如图 4所示，使用2\*2的Maxpooling层，stride=2比例因子为。得到四个输入，使用四个相同的滤波器（红，黄，绿，蓝所示）对其分别进行降采样，得到四个结果，最终交错拼接成与原始输入（黑色方块）大小相同的输出。

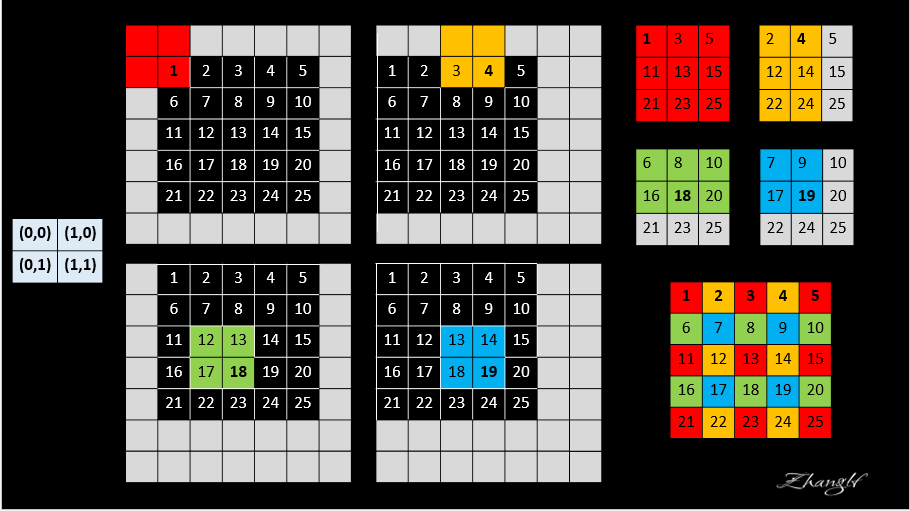


图 4 Shift-and-stitch[5]

笔者认为这种策略客观上难以理解。Long等人指出这种策略输入以倍增长，效率较低，实验中并没有采取这种策略。

1. 反卷积

反卷积又称转置卷积，逆卷积，能够将卷积结果恢复为原始图像的尺寸大小，而其值不完全一致。这种方法可以将任意的步长正向卷积结果翻转为原尺寸大小。

对于图像与卷积核在空域中卷积，可知矩阵形式为图像对应的矩阵拉伸的列向量和具有稀疏性质的卷积核矩阵的矩阵相乘，得到卷积结果同样可以表述为矩阵拉伸的列向量。有，故可以还原为原始图像大小以实现上采样。这种方式可以使得卷积核矩阵的初值不固定，在网络的训练中使得其可以自己学习参数，甚至可以实现非线性的上采样。

## 跳跃连接

当网络全卷积化后并使用上采样策略，理论上并不能很好的实现很好的分割结果。当网络的深层输入维数较高时，其高维特征得到很好的表达，但是其空间细节信息在下采样时已被损失，单纯依靠较大的上采样率实现原始图像尺寸，效果必定不够好。故Long等人提出了skip architecture。以VGG16为例，其3个版本的跳跃连接方式如图 5-图 7所示。蓝色方块为卷积层，红色为池化层，是图像尺寸下降的原因，黄色方框为全连接层的卷积化，黑色方框为1\*1的卷积层，目的是将降低维数得到分类数目维的三维张量，绿色方框为反卷积层，为实现密集预测的关键步骤。FCN-32S,FCN-16S与FCN-8S表示采样的大小分别为32，



图 5 FCN-32S



图 6 FCN-16S



图 7 FCN-8S

16与8。如图中的跳跃连接操作使得高级特征与低级特征融合，实现更准确的密集预测。

# 实验及分析

## 分类网络卷积化的分割性能对比

Long等人使用了ILSVRC14中性能极好的分类网络AlexNet，VGG16和GoogLeNet，丢弃了分类网络的分类层，将全连接层变为卷积层，同时使用1\*1的卷积核，输出21类，并采用反卷积层使网络输出大小为输入大小，得到FCN-AlexNet，FCN-VGG16(FCN-32S)，FCN-GoogLeNet。卷积化过程如图 8所示。

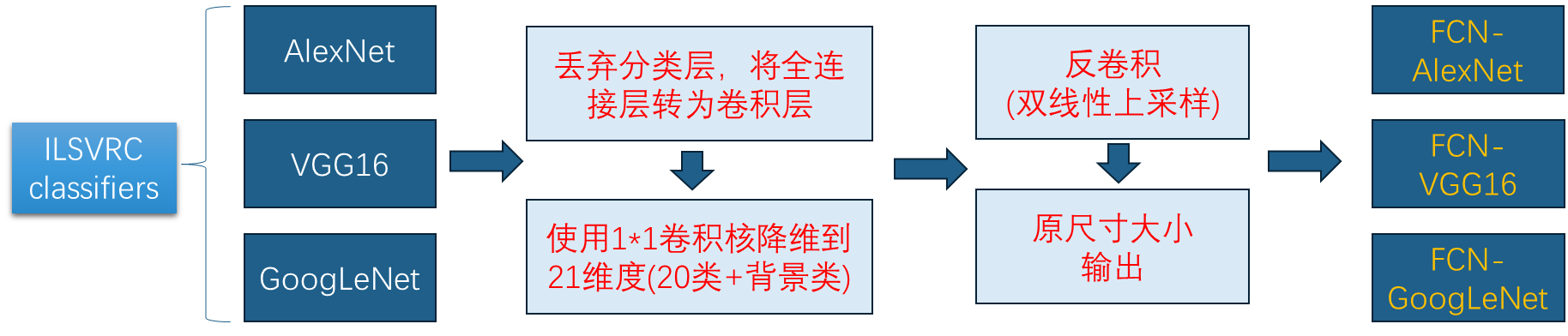


图 8 分类网络卷积化

PASCAL VOC 2011数据集上，固定学习率训练至少175个epochs的情况下，三个FCN的分割效果评价指标mean IU，推理时间forward time，卷积层数con.layers，参数量parameters，输出单元的感受野大小rf size和最大步长max strid如表 2所示。

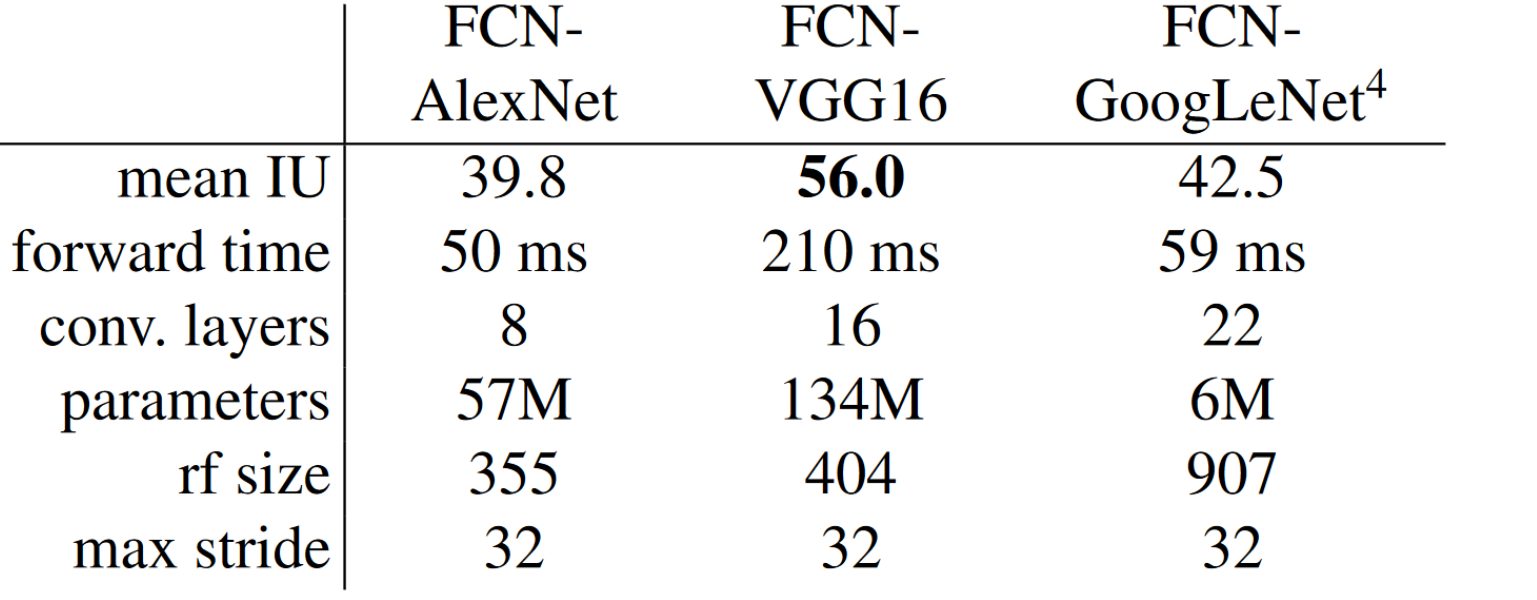


表 2 分类网络卷积化性能

可以看出，FCN-VGG16在验证集上的mean IU最大，表示分割效果最好，与下文SDS方法在验证集上52.6的mean IU相比提升比较大。说明了分类网络全卷积化的可行性，且在32倍采样的情况下，其分割效果也是优越的。

## FCN-32S，FCN-16S，FCN-8S性能对比

尽管FCN-32S已经取得了不错的分割效果，其分割的视觉效果与FCN-16S,FCN-8S仍有较大差异，如图 9所示。可以看出，随着所融合的特征来自的层级越低，其分割结果越好，即细层和粗层的特征结合，能使模型在预测局部特征时考虑全局特征，从而得到更好的分割结果。

三个网络在VOC 2011 segval数据子集上的分割性能如表 3所示，可以看出FCN-8S的分割效果最好，这与图 9的可视化结果是相符的。同时值得注意的是，FCN-32S-fixed是在

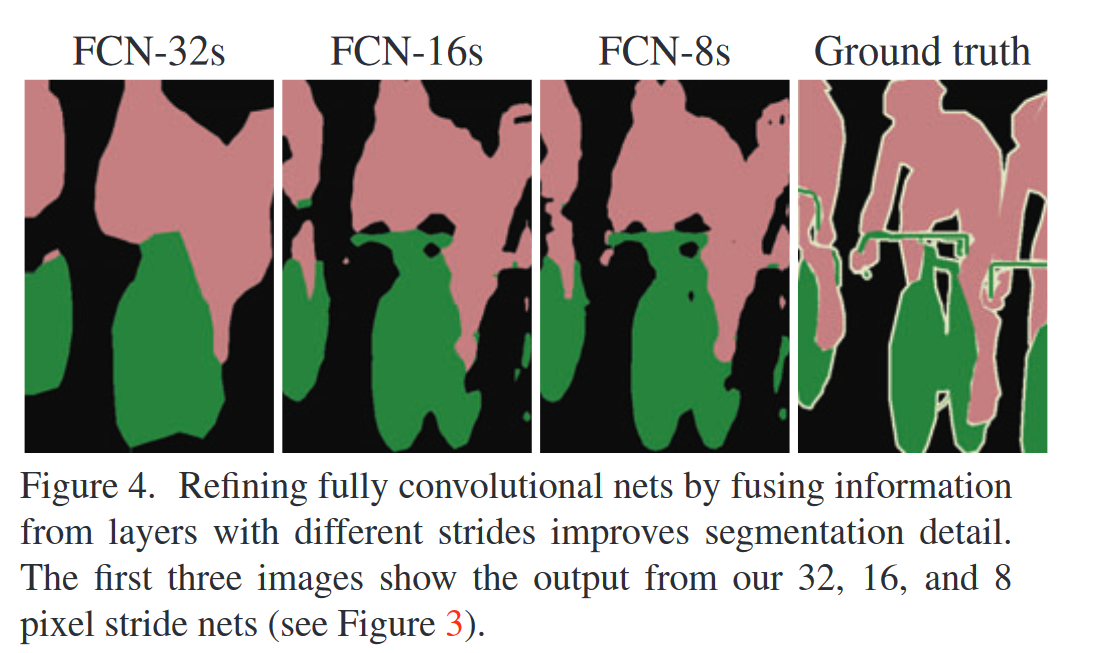


图 9 FCN-32S，FCN-16S，FCN-8S分割结果图

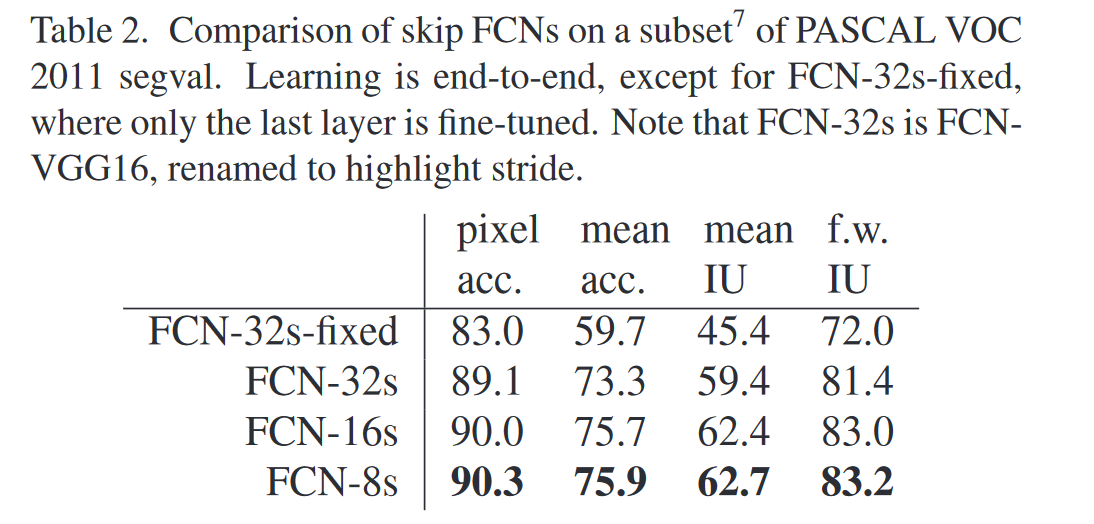


表 3 FCN-32S,FCN-16S,FCN-8S在VOC 2011 segval数据子集上的分割性能

网络训练时，只训练最后一层的结果，即其他来自分类网络的参数没有微调，可见其效果不如FCN-32S。Long等人指出，此时FCN-8S的跳跃连接技术在客观指标和视觉指标上都取得了较大收益，继续下去会收益递减，故没有实施更底层的融合。

## PASCAL VOC语义分割性能对比

在PASCAL VOC 2011 and 2012 test数据集上，FCN8-S，R-CNN和SDS三种方法的结果如表 4所示。可以看出，FCN-8S在VOC2011 test数据集和VOC2012 test数据集上mean IU相较于SDS提升了19%和21%，推理时间缩短了286倍。三种方法分割的结果如图 10

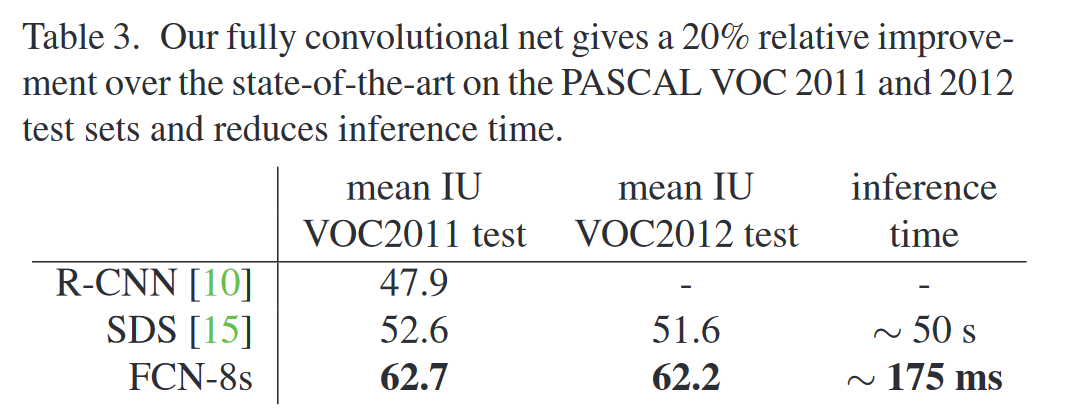


表 4 PASCAL VOC 2011 and 2012测试集上，FCN8-S，R-CNN和SDS方法的分割性能

所示，对于第一行而言，FCN-8S能较好的分割出人物和马而云朵丢失，但SDS没有完全分割出人物，马匹半个身子缺失。对于第二行，即使相交的人物和摩托车。FCN-8S仍能较好的区分，SDS不但没有将人与车分割出来，还存在缺失的现象。对于第三行，FCN-8S没有受到阻塞的影响，将铁丝网后的羊完全识别出来，而SDS受到干扰，产生了截断。当然FCN-8S也存在误检的现象，将船上的救生衣误识别为了人。

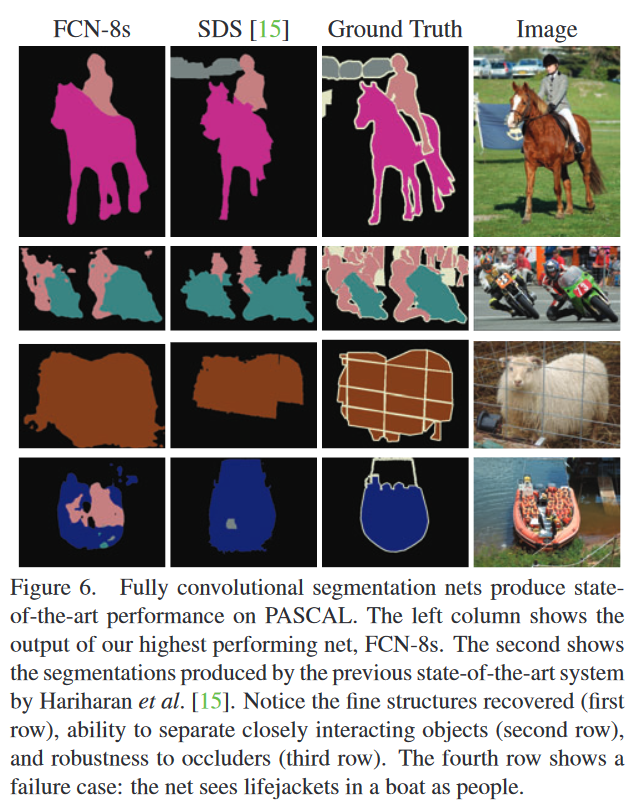


图 10 SDS,R-CNN,FCN-8S在PASCAL数据集上分割结果

## Patch Sampling 训练策略

Long等人使用整张图像进行训练而不是与分类相似使用patch进行训练。Patch training重复图块多，效率低，二者相比LOSS都收敛但是整张图像的训练速度更快。

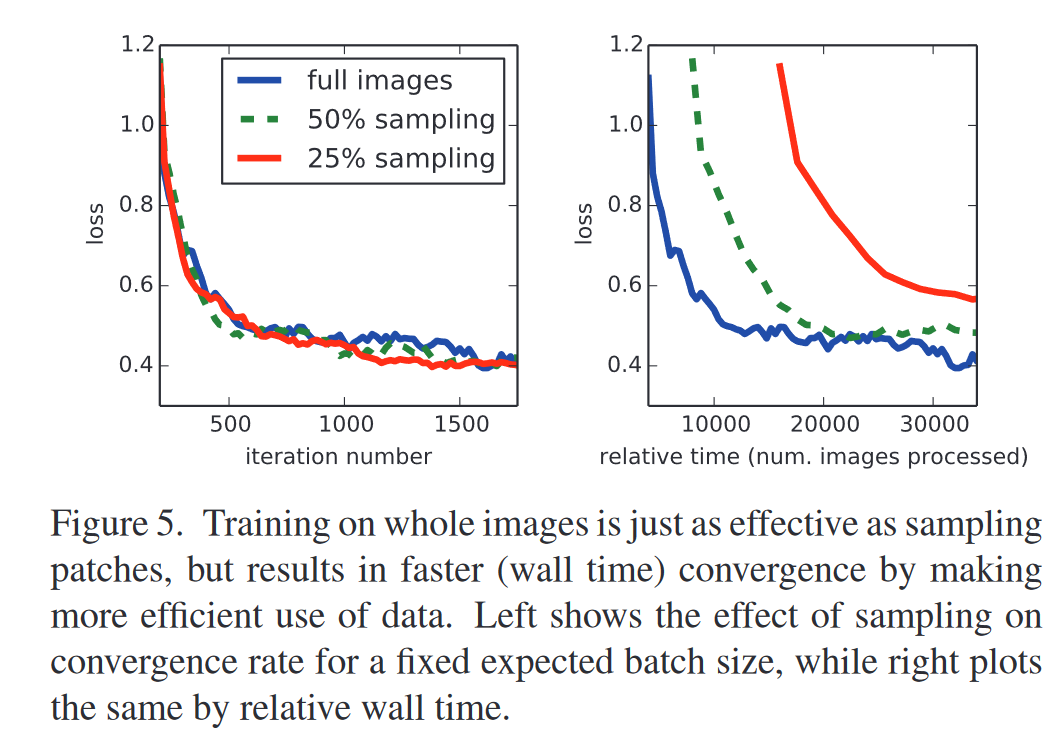


图 11 full image和Patch Sampling训练策略

如图 11所示，采用Patch Sampling策略与整幅图像进行训练的策略Loss下降速度相同，但是Patch Sampling策略耗时更长，导致不如使用Full images训练的速度快，原因如3.1中所述。

参考文献

1. Csurka G, Perronnin F.An efficient approach to semantic segmentation.Int’l Journal of Computer Vision, 2011, 95 (2) :198-212.[doi:10.1007/s11263-010-0344-8]
2. https://zhuanlan.zhihu.com/p/109999530
3. https://www.sohu.com/a/387083227\_663516
4. Very Deep Convolutional NetWorks for Large-Scale Image Recognition
5. https://blog.csdn.net/qq\_36336522/article/details/102599526