Segnet: 一种用于图像分割的深度卷积编码－解码架构

摘要

我们展示了一种新奇的有实践意义的深度全卷积神经网络结构，用于逐个像素的语义分割，并命名为 SegNet。核心的可训练的分割引擎包含一个编码网络，和一个对应的解码网络，并跟随着一个像素级别的分类层。编码器网络的架构在拓扑上与 VGG16 网络中的 13 个卷积层相同。解码网络的角色是映射低分辨率的编码后的特征图到输入分辨率的特征图。具体地，解码器使用在相应编码器的最大合并步骤中计算的池化索引来执行非线性上采样。这消除了上采样的学习需要。上采样后的图是稀疏的，然后与可训练的滤波器卷积以产生密集的特征图。我们把我们提出的架构和广泛采用的 FCN 架构和众所周知的 DeepLab-LargeFOV、DeconvNet 架构做了比较，这种比较揭示了实现良好分割性能的内存与准确性的权衡。

SegNet 的主要动机是场景理解应用。因此，它在设计的时候保证在预测期间，内存和计算时间上保证效率。在可训练参数的数量上和其他计算架构相比也显得更小，并且可以使用随机梯度下降进行端到端的训练。我们还在道路场景和 SUN RGB-D 室内场景分割任务中执行了 SegNet 和其他架构的受控基准测试。这些定量的评估表明，SegNet 在和其他架构的比较上，提供了有竞争力的推断时间和最高效的推理内存。我们也提供了一个 Caffe 实现和一个 web 样例 http://mi。eng。cam。ac。uk/projects/segnet/。

Index Terms—Deep Convolutional Neural Networks, Semantic Pixel-Wise Segmentation, Indoor Scenes, Road Scenes, Encoder,Decoder, Pooling, Upsampling。

1.介绍

语义分割具有广泛的应用范围，从场景理解，推断对象之间的支持关系到自主驾驶。依靠低级别视觉线索的早期方法已经被流行的机器学习算法所取代。特别的，深度学习后来在手写数字识别、语音、整图分类以及图片中的检测上都取得了成功[VGG][GoogLeNet]。现在图像分割领域也对这个方法很感兴趣[crfasrnn][parsent]等。然而，近来的很多方法的都尽力直接采用设计来图像分类的方法进行语义分割。结果虽然令人鼓舞，但是比较粗糙[deeplab]。这主要是因为 max-

pooling 和 sub-sampling 减少了特征图的分辨率。我们设计 SegNet 的动机就是来自于对于为了语义分割而从低分辨率的特征图到输入分辨率映射的需要。这种映射也必须产生一些特征用于精确地边界定位。

我们的架构，SegNet，设计的目的是作为一种高效的语义分割架构。它主要是由道路现场理解应用的动机，需要建模外观（道路，建筑物），形状（汽车，行人）的能力，并了解不同类别（如道路和侧面行走）之间的空间关系（上下文）。在典型的道路场景中，大多数像素属于大型类，如道路，建筑物，因此网络必须产生平滑的分段。引擎还必须具有根据其形状来描绘对象的能力，尽管它们的尺寸很小。因此，在提取的图像表示中保留边界信息是重要的。从计算的角度来看，在推理过程中，网络需要保证在内存和计算时间两方面都是高效的。进行端到端的训练为了使用诸如随机梯度下降（SGD）之类的有效的权重更新技术来联合优化网络中所有权重的能力是一个额外的好处，因为它更容易重复。SegNet 的设计源于需要符合这些标准。

SegNet 中的编码网络和 VGG16 的卷积层是拓扑上相同的。我们移除了全连接层，这样可以使 SegNet 比其他许多近来的结构[FCN][DeconvNet][ParseNet][Decoupled]显著的小并且训练起来更容易。SegNet 的关键部件是解码器网络，由一个对应于每个编码器的解码器层次组成。其中，解码器使用从相应的编码器接受的 max-pooling indices 来进行输入特征图的非线性 upsampling。这个想法来自设计用于无监督功能学习的架构。在解码网络中重用 max-pooling indics 有多个实践好处：（1）它改进了边界划分（2）减少了实现端到端训练的参数数量（3）这种 upsampling 的形式可以仅需要少量的修改而合并到任何编码－解码形式的架构[FCN][crfasrnn]。

这篇论文的一个主要贡献是，我们对 Segnet 解码技术和广泛使用的 FCN 的分析。这是为了传达在设计分割架构中的实际权衡。近来许多分割的深度架构使用相同的编码网络，例如 VGG16，但是在解码网络的形式、训练和推理上是不同的。另一个常见的特点是，这些网络通常有亿级别的训练参数，从而导致端到端的训练很困难[DeconvNet]。训练困难导致了多阶段的训练[FCN]，或者添加一个与训练的网络结构如 FCN[crfasrnn]，或者用辅助支持，例如在推理阶段使用区域 proposals[DeconvNet]，或者使用分类和分割网络的不相交训练[Decoupled]，或者用额外的数据进行与训练[Parsenet]或者全训练[crfasrnn]。另外，性能提升后处理技术也受到欢迎。尽管这些因素都很好的提高了在 voc 上的性能，但是他们的定量结果难以解决实现良好性能所必需的关键设计因素。因此我们分析了被用在这些方法[FCN][DeconvNet]中的解码过程，并揭示了他们的优点和缺陷。

我们评估了 SegNet 在两种场景分割任务中的性能，分别是 CamVid 道路场景分割和 SUN RGB-D 室内场景分割。VOC12 在过去很多年都有分割的基准挑战。但是，这个任务的大部分都有一个或两个由高度多样的背景包围的前景类。这隐含地有利于用于检测的技术，如最近关于解耦分类分割网络的工作所示[Decoupled]，其中分类网络可以用大量弱标签数据进行训练，并且独立分割网络性能得到改善。[deeplab]的方法还使用分类网络的特征图和独立的 CRF 后处理技术来执行分割。性能也可以通过额外的推理辅助来增强，例如区域 proposals[DeconvNet][Edge Boxes]。因此，因此，它与场景理解不同之处在于，其目的是利用对象的共同出现以及其他空间上下文来执行可靠的分割。为了证明 SegNet 的高效性，我们展示了一个实时的道路场景分割的在线 demo，来分割 11 类的自主驾驶兴趣类(如图 1 所示)。图 1 中展示了从 Google 中找的一些随机道路图片和 SUNRGB-D 中产生的一些随机室内测试场景图片的分割结果。

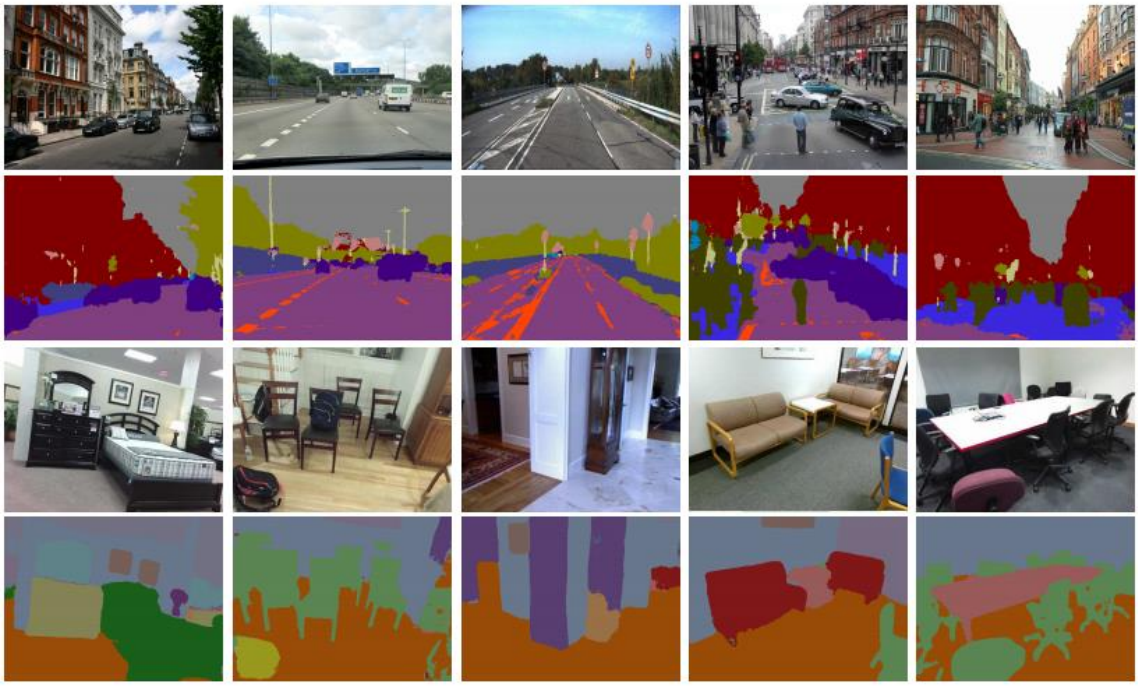

Fig. 1. SegNet predictions on road scenes and indoor scenes. To try our system yourself, please see our online web demo at http://mi.eng.cam.ac.uk/projects/segnet/.

论文的剩余部分组织如下。在 Section 2 我们回顾了近期的相关文献。在 Section 3 我们描述了 SegNet 架构和对它的分析。在 Section 4 我们评估了 SegNet 在室外和室内数据集上的性能。接下来是 Section 5 关于我们的方法的一般性讨论，指出未来的工作。Section 6 是结论。

2.文献回顾

语义分割是一个十分活跃的研究课题，其中很大一部分作用是因为作为挑战的数据集[PASCALVOC][SUN RGB-D][KITTI]。在深度学习到来之前，性能最好的方法大部分依赖于手工设计的特征来独立地分类像素。通常，将一块区域送入一个分类器例如 Random Forest 或者 Boosting 来预测中心像素的类概率。基于外观的特征或者 SfM 已经被发明用来进行 CamVid 道路场景理解的测试。后通过使用成对或更高阶的 CRF 来平滑来自分类器的每像素噪声预测（通常称为一元项）来提高精确度。最近的方法旨在通过尝试预测块中所有像素的标签，而不仅仅是中心像素来产生高质量的一元项。这改进了随机森林一元项的结果，但是稀疏的结构化类被分类不佳。从 CamVid 视频中计算出的稠密深度图也被用于使用随机森林[32]进行分类的输入。另一种方法主张结合流行的手工设计特征和时空超像素化来获得更高的准确率。CamVid 测试中性能最好的技术通过将对象检测输出与 CRF 框架中的分类器预测相结合来解决标签频率之间的不平衡。所有这些技术的结果表明需要改进的分类的特征。

自从 NYU 数据集的发布以来，室内 RGBD 像素级语义分割也得到欢迎。该数据集显示了深度通道改善分割的有用性。他们的方法使用诸如 RGB-SIFT，depth-SIFT 和像素位置的特征作为神经网络分类器的输入来预测像素一元项。然后使用 CRF 来光滑这个有噪音的一元项。使用更丰富的特征集进行改进，包括 LBP 和区域分割，以获得更高的准确性，然后是 CRF。还要一些其他的方法，所有这些方法的共同属性是使用手工设计的特征来分类 RGB 或 RGBD 图像。

深层卷积神经网络对物体分类的成功最近引导研究人员利用其特征学习能力进行结构化预测问题，如分割。还尝试将设计用于对象分类的网络应用于分割，特别是通过在块中复制最深层特征以匹配图像尺寸。然而，所得到的分类是块状的。另一种方法是使用循环神经网络[RNN]合并了几个低分辨率预测来创建输入图像分辨率预测。这些技术已经是手工设计特征的改进，但是它们划定边界的能力差。

更新的深度结构[FCN][DeconvNet][CRFASRNN][Decoupled]特别设计用于分割，通过学习解码或将低分辨率图像表示映射到像素点预测，提升了最先进的技术水平。上边几个网络的编码网络是用来产生低分辨率便是，都是使用的 VGG16 分类网络结构(13 个卷基层和 3 个全连接层)。这些编码网络的权重在 ImageNet 上进行了特殊的预训练。解码器网络在这些架构之间不同，并且是负责为每个像素生成多维特征以进行分类的部分。

全卷积网络（FCN）架构中的每个解码器都学习对其输入特征图进行上采样，并将其与相应的编码器特征图组合，以产生到下一个解码器的输入。它是一种在编码器网络中具有大量可训练参数的架构（参数个数 134M），但是非常小的解码器网络（参数个数 0。5M）。该网络的整体

大小使得难以在相关任务上端到端地进行训练。因此，作者使用了阶段性的训练过程。这里，解码器网络中的每个解码器逐渐添加到现有的训练好的网络中。网络生长直到没有观察到进一步的性能提高。这种增长在三个解码器之后停止，因此忽略高分辨率特征图肯定会导致边缘信息的丢失[DeconvNet]。除了训练的相关问题之外，解码器中重用编码器特征图的需求使其在测试时间内内存集约。我们更深入地研究这个网络，因为它是其他最新架构的核心[CRFASRNN][ParseNet]。

通过使用循环神经网络（RNN）附加到 FCN[CRFASRNN]并对其在大的数据集上[VOC][COCO]进行微调，FCN 的预测性能进一步得到改善。在使用 FCN 的特征表征能力的同时，RNN 层模仿 CRF 的尖锐边界划分能力。它们比 FCN-8 显示出显着的改进，但也表明当使用更多训练数据训练 FCN-8 时，这种差异减小。当与基于 FCN-8 的架构联合训练时，CRF-RNN 的主要优点被揭示出来。联合训练有助于其他最近的结果。有趣的是，反卷积网络[DeconvNet]的性能明显优于 FCN，但是以更复杂的训练和推理为代价。这提出了一个问题，即随着核心前馈分割引擎的改进，CRF-RNN 的感知优势是否会减少。无论如何，CRF-RNN 网络可以附加到任何深度分段架构，包括 SegNet。

多尺度的深层架构也被广泛采用。它们有两种风格，（i）使用几个尺度的输入图像和相应的深度特征提取网络，以及（ii）组合来自单个深层结构的不同层的特征图[ParseNet]。通常的想法是使用多尺度提取的特征来提供局部和全局上下文[zoom-out]，并且早期编码层的使用特征图保留更高频率的细节，从而导致更尖锐的类边界。其中一些架构由于参数大小而难以训练。因此，与数据增加一起使用多阶段训练过程。推论过程由于特征提取的多个卷积路径使用也是复杂度比较高的。其他在他们的多尺度网络上附加了一个 CRF，并共同训练他们。然而，这些在测试时间不是前馈的，需要优化才能确定 MAP 标签。

最近提出的几种最新的分割结构在推理时间上不是前馈的[DeconvNet][deeplab][Decoupled]。它们需要通过 CRF 的 MAP 推理或推荐的区域 proposals[DeconvNet]等辅助工具。我们认为通过使用 CRF 获得的感知性能提升是由于在其核心前馈分割引擎中缺乏良好的解码技术。另一方面，SegNet 使用解码器获得准确的像素级别分类效果。

最近提出的反卷积网络[DeconvNet]及其半监督变体解耦网络[Decoupled]使用编码器特征图的最大位置（pooling index）在解码器网络中执行非线性上采样。这些架构的作者独立于 SegNet（首次提交给 CVPR 2015），提出了解码网络中的解码思想。然而，它们的编码器网络由 VGG-16 网络包括全连接，其包括其整个网络的约 90％的参数。这使得他们的网络训练非常困难，因此需要更多的辅助工具，例如使用区域 proposals 来实施培训。此外，在推理阶段这些 proposals 被使用，

这显着增加了推理时间。从基准的角度来看，这也使得在没有其他辅助帮助下难以评估其架构（编码器 -解码器网络）的性能。在这项工作中，我们丢弃 VGG16 编码器网络的全连接层，使我们能够使用 SGD 优化使用相关的训练集训练网络。另一种最近的方法[deeplab]显示了在不牺牲性能，显着减少参数数量的好处是能够减少内存消耗和改进推理时间。

我们的工作是由 Ranzato 等人提出的无监督特征学习架构的启发。这种架构用于无监督的预训练进行分类。关键的学习模块是编码器-解码器网络。编码器由滤波器组卷积、tanh 非线性、max-pooling 和 sub-samplingh 组成，以获得特征图。对于每个示例，池期间计算的最大位置的索引将被存储并传递给解码器。编码 upsampling 使用存储的 pooled indices 映射进行采样。它使用一个可训练的解码器滤波器组对这个上采样映射进行卷积，以重建输入图像。该体系结构用于无监督的分类预处理。一种类似的解码技术用于对训练好的卷积网络进行可视化分类，接受完整的图像大小作为输入，学习分层编码器。然而，这种方法没有尝试使用深度编码器-解码器网络进行无监督的特征训练，因为它们在每个编码器训练之后丢弃解码器。在这里，SegNet 与这些架构不同，因为深度编码器 - 解码器网络被联合训练用于监督学习任务，因此解码器是测试时间中网络的组成部分。

使用深度网络进行像素点预测的其他应用包括图像超分辨率和来自单个图像的深度地图预测。[50]的作者讨论了学习从低分辨率的特征映射上采样的必要性，这是本文的中心主题。

3.架构

架构如图 2 所示。

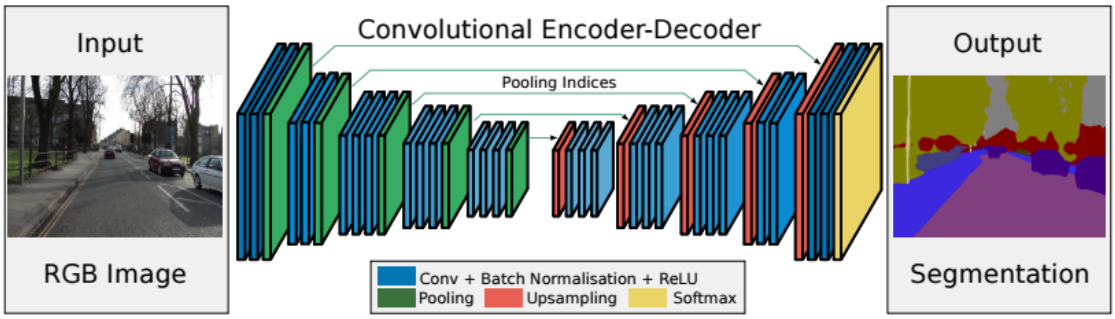

Fig. 2. An illustration of the SegNet architecture. There are no fully connected layers and hence it is only convolutional. A decoder upsamples its input using the transferred pool indices from its encoder to produce a sparse feature map(s). It then performs convolution with a trainable filter bank to densify the feature map. The final decoder output feature maps are fed to a soft-max classifier for pixel-wise classification.

SegNet 有一个编码器网络和相应的解码器网络，然后是最后的像素级分类层。编码器部分使用的是 VGG16 的前 13 层卷积网络，可以尝试使用 Imagenet 上的预训练。我们还可以丢弃完全连接的层，有利于在最深的编码器输出处保留较高分辨率的特征图。与其他最近的架构[FCN][DeconvNet]相比，这也减少了 SegNet 编码器网络中的参数数量（从 134M 到 14。7M）。如表 6 所示。每个编码器层具有对应的解码器层，因此解码器网络具有 13 层。最终解码器输出被馈送到多级 soft-max 分类器以独立地为每个像素产生类概率。

				Median frequency balancing							Natural frequency balancing						
		Storage	Infer	Test				Train			Test				Train		
Variant	Params (M)	multiplier	time (ms)	G	C	mIoU	BF	G	C	mIoU	G	C	mIoU	BF	G	C	mIoU
Fixed upsampling																	
Bilinear-Interpolation	0.625	0	24.2	77.9	61.1	43.3	20.83	89.1	90.2	82.7	82.7	52.5	43.8	23.08	93.5	74.1	59.9
Upsampling using max-pooling indices																	
SegNet-Basic	1.425	1	52.6	82.7	62.0	47.7	35.78	94.7	96. 2	92.7	84.0	54.6	46.3	36.67	96.1	83.9	73.3
SegNet-Basic-EncoderAddition	1.425	64	53.0	83.4	**63.6**	48.5	35.92	94.3	95.8	92.0	**84.2**	56.5	**47.7**	36.27	95.3	80.9	68.9
SegNet-Basic-SingleChannelDecoder	0.625	1	33.1	81.2	60.7	46.1	31.62	93.2	94.8	90.3	83.5	53.9	45.2	32.45	92.6	68.4	52.8
Learning to upsample (bilinear initialisation)																	
FCN-Basic	0.65	11	24.2	81.7	62.4	47.3	**38.11**	92.8	93.6	88.1	83.9	55.6	45.0	**37.33**	92.0	66.8	50.7
FCN-Basic-NoAddition	0.65	n/a	23.8	80.5	58.6	44.1	31.96	92.5	93.0	87.2	82.3	53.9	44.2	29.43	93.1	72.8	57.6
FCN-Basic-NoDimReduction	1.625	64	44.8	**84.1**	63.4	**50.1**	37.37	**95.1**	**96.5**	**93.2**	83.5	**57.3**	47.0	37.13	**97.2**	**91.7**	**84.8**
FCN-Basic-NoAddition-NoDimReduction	1.625	0	43.9	80.5	61.6	45.9	30.47	92.5	94.6	89.9	83.7	54.8	45.5	33.17	95.0	80.2	67.8

编码器网络中的每个编码器与滤波器组进行卷积，生成一组特征映射。每个编码器由卷积层、批归一化层、RELU 组成，之后，执行具有 2×2 窗口和步幅 2（非重叠窗口）的最大池化，输出结果相当于系数为 2 的下采样。最大池化用于实现输入图像中小空间位移的平移不变性，子采样导致特征图中每个像素的大输入图像上下文（空间窗口）。由于最大池化和子采样的叠加，导致边界细节损失增大，因此必须在编码特征图中在 sub-sampling 之前捕获和储存边界信息。为了高效，我们只储存了 max-pooling indices，原则上，对于每个 2×2 池化窗口，这可以使用 2 位来完成，因此与浮动精度的记忆特征图相比，存储效率更高。正如我们在本文稍后展示的那样，这种较低的内存存储会导致精确度的轻微损失，但仍然适用于实际应用。

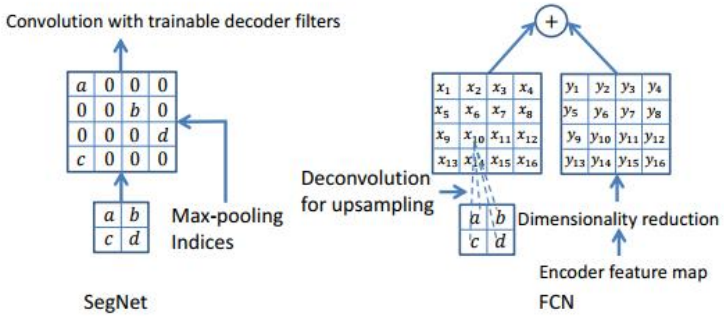

Fig. 3. An illustration of SegNet and FCN [2] decoders. a, b, c, d correspond to values in a feature map. SegNet uses the max pooling indices to upsample (without learning) the feature map(s) and convolves with a trainable decoder filter bank. FCN upsamples by learning to deconvolve the input feature map and adds the corresponding encoder feature map to produce the decoder output. This feature map is the output of the max-pooling layer (includes sub-sampling) in the corresponding encoder. Note that there are no trainable decoder filters in FCN.

SegNet 的解码技术如图 3 所示。

解码器网络中的解码器使用来自对应的编码器特征图的存储的最大池化索引来上采样至其输入特征图。此步骤产生稀疏特征图。然后将这些特征图与可训练的解码器滤波器组卷积以产生密集的特征图。然后是 BN。注意，最后一个解码器产生一个多通道的特征图，而不是 3 通道的(RGB)。然后输入给一个 softmax 分类器。这个 soft-max 独立地分类每个像素，soft-max 分类器的输出是 K 通道图像的概率，其中 K 是类的数量。预测的分割对应于在每个像素处具有最大概率的类。

我们在这里添加了另外两个架构，DeconvNe]和 U-Net 与 SegNet 共享类似的架构，但有一些不同。DeconvNet 的参数化要大得多，需要更多的计算资源，而且端到端培训也比较困难(表 6)，主要由于使用了完全连接的层(尽管是以卷积的方式)，我们报告了 DeconvNet 与进行的几个比较在论文的第 4 节中。

与 SegNet 相比，U-Net（提出用于医学影像社区）不重复使用池化指标，而是将整个特征图（以更多内存为代价）传输到相应的解码器，并将其连接上采样（通过反卷积）解码器特征图。在网络架构中，U-Net 中没有 conv5 和 max-pool 5。另一方面，SegNet 使用来自 VGG 网络的所有预先训练的卷积层权重作为预训练权重。

3.1 解码器变种

许多分段架构[FCN][deeplab][DeconvNet]共享相同的编码器网络，它们只是以其解码器网络的形式而变化。其中我们选择比较 SegNet 解码技术与广泛使用的完全卷积网络（FCN）解码技术[FCN][CRFASRNN]。

为了分析 SegNet 并将其性能与 FCN（解码器变体）进行比较，我们使用较小版本的 SegNet，称为 SegNet-Basic，它具有 4 个编码器和 4 个解码器。所有在 SegNet-Basic 中的编码器使用 max-

pooling、sub-sampling 和相应的解码器使用接收到的最大池索引对其输入进行采样。在编码器和解码器网络中，每个卷积层之后都进行正规化。在卷积之后没有使用偏差，并且在解码器网络中不存在 ReLU 非线性。选择所有编码器和解码器层的 7×7 的恒定核大小以提供用于平滑标记的宽上下文，即最深层特征图（层 4）中的像素可以追溯到上下文窗口 106×106 像素的输入图像。这种小尺寸的 SegNet-Basic 使我们能够在合理的时间内探索许多不同的变体（解码器）并进行训练。类似地，我们创建了 FCN-Basic，一个可比较的 FCN 版本，用于我们的分析，它与 SegNet-Basic 共享相同的编码器网络，但与所有解码器中使用的 FCN 解码技术（见图 3）相同。较小的变体是解码器滤波器是单通道的变体，即它们仅仅卷积它们相应的上采样特征图。该变体（SegNet-Basic-SingleChannelDecoder）显着减少了可训练参数的数量和推理时间。

FCN 模型的重要设计元素是编码器特征图的降维步骤。这压缩了编码器特征图，然后在相应的解码器中使用。使用双线性插值权重初始化上采样内核。我们还可以创建 FCN-Basic 模型的变体，该模型丢弃编码器特征映射添加步骤，并且仅学习上采样内核（FCN-Basic-NoAddition）。 除了上述变体之外，我们研究使用固定双线性插值权重的上采样，不需要上采样学习（双线性插值）。另一方面，我们可以在 SeqNet 解码器的每一层添加 64 个编码器特征映射到 SegNet 解码器的相应输出特征图，以创建更多内存扩大型 SegNet（SegNet-Basic-EncoderAddition）。这里使用上采样的 max-pooling indices，随后进行卷积步骤以使其稀疏输入变得更加密集。然后将其逐个添加到相应的编码器特征图，以产生解码器输出。

另一种和更多的内存密集型 FCN-Basic 变体（FCN-Basic-NoDimReduction）是对编码器特征映射没有进行维度降低的地方。这意味着与 FCN-Basic 不同，最终的编码器特征图在将其传送到解码器网络之前不会压缩到 K 个通道。因此，每个解码器结尾处的信道数量与相应的编码器相同(即 64)。

我们还尝试了其他通用变体，其中功能图只是通过复制进行上采样，或者通过使用固定（和稀疏）索引数组进行上采样。与上述变体相比，这些表现相当差。在编码器网络（解码器是冗余的）中没有最大池和子采样的变体消耗更多的存储器，需要更长的时间来收敛和执行等不好。最后，请注意，为了鼓励复制我们的结果，我们发布了 Caffe 执行所有变体。

3.2 训练

我们使用 CamVid 路景数据集来对基于解码器变体的性能进行基准测试。该数据集很小，由 360×480 分辨率的 367 次训练和 233 次测试 RGB 图像（白天和黄昏场景）组成。挑战是划分道路，建筑，汽车，行人，标志，极点，侧路等 11 类。我们对 RGB 输入进行局部对比度归一化。

编码器和解码器权重都使用 He 等人的方法。为了训练所有的变体，我们使用固定学习率 0。1 和动量 0。9 的随机梯度下降（SGD），使用我们的 Caffe 实现 SegNet-Basic。在每轮之前，训练集被洗牌，然后按顺序挑选每个小批量（12 张图像），从而确保每个图像在一个时代只被使用一次。我们选择在验证数据集上执行最高的模型。

我们使用交叉熵损失作为训练网络的目标函数。损失在一个小批量的所有像素上求和得到。当训练集中的每个类别（例如，道路，天空和建筑像素占主导地位的 CamVid 数据集）中像素数量的变化很大时，则需要根据真实类别不同地加权。这被称为 class balancing。我们使用 median frequency balancing，其中分配给损失函数中的类的权重是在整个训练集上计算的类频率的中值除以类频率的比率(?)。这意味着训练集中的较大类的权重小于 1，最小类的权重最高。我们还尝试了不同类型的训练，无需类平衡，也可以等效地使用 natural frequency balancing。

3.3 分析

为了定量分析不同的解码器变体。使用如下的测量:G 值是 global accuracy，测量数据集中所有像素正确分类的百分比。C 值 class average accuracy，所有类的预测准确度的平均值。还有就是在 Pascal VOC12 挑战中使用的所有类的 mIoU。mIoU 度量是比类平均精度更严格的度量，因为它惩罚了假阳性预测。然而，mIoU 度量不是通过类平衡交叉熵损失直接优化的。

mIoU 指标也被称为“雅克指数”，最常用于基准测试。然而，Csurka 等人注意到，这个度量并不总是符合人类对质量好的细分的定性判断（等级）。他们以示例的形式表明，mIoU 有利于区域平滑度，并且不评估边界准确性，FCN 作者最近也提到了这一点。因此，他们建议通过基于通常用于评估无监督图像分割质量的伯克利轮廓匹配得分的边界测量来补充 mIoU 度量。Csurka 等人简单地将其扩展到语义分割，并且表明与 mIoU 度量结合使用的语义轮廓精度的度量与分割输出的人类排序一致。

计算语义轮廓得分的关键思想是评估 F1 测量，涉及在给定一个像素公差距离的情况下计算预测和 ground truth 类边界的精确度和回调值。我们使用图像对角线的 0。75％的值作为公差距离。将存在于地面真实测试图像中的每个类的 F1 测量值进行平均以产生图像 F1 度量。BF 作为整个测试集的 F1 度量。

在对 CamVid 验证集进行 1000 次优化之后，我们对每个体系结构变体进行测试，直到训练损失收敛为止。培训 mini-batch 大小 12 这对应测试大约每 33 时代通过训练集(传球)。我们选择迭代

在全球中精度最高的评估验证集。我们报告所有三种措施的性能在这一点上了 CamVid 测试集。虽然我们在训练变体时使用类平衡，但仍然重要的是要实现高全局准确度，从而实现整体平滑分割。另一个原因是，细分对自动驾驶的贡献主要在于划分道路、建筑、人行道、天空等类别。这些类控制了图像中的大部分像素，高全局精度对应于这些重要类的良好分割我们还观察到，当等级平均值最高时报告数值性能通常可以对应于表示感知噪声分割输出的低全局精度。

在表 1 中，我们报告了分析的数值结果。我们还展示了可训练参数的大小和最高分辨率的特征映射或池索引存储内存，最大池和子采样后的第一层特征映射。我们展示的平均时间为一个与我们的咖啡实现前进传球,平均使用 360×50 多个测量 480 在 NVIDIA Titan GPU 上输入，带有 cuDNN v3 加速。我们注意到 SegNet 变体中的上采样层没有使用 cuDNN 加速进行优化。我们展示了所选迭代中所有变量的测试和培训结果。在没有类平衡的情况下，结果也被制成表格(自然频率)用于训练和测试的准确性。下面我们用类平衡来分析结果。

从表 1 中我们可以看到，基于上采样的双线性插值在没有任何学习的情况下，在所有的精度测量中表现最差。所有其他的方法，无论是使用学习的上采样(FCN-Basic 和变体)或学习解码器滤波器后的上采样(SegNet-Basic 和它的变体)表现得明显更好。这强调了需要学习解码的分割。其他作者在比较 FCN 和 FCN 时收集的实验证据也支持了这一点分段式译码技术[4]。

当我们比较 SegNet-Basic 和 FCN-Basic 时，我们发现两者在所有的准确性测试中都表现得一样好不同的是，SegNet 在推理期间使用的内存更少，因为它只存储最大池索引。另一方面，FCN-Basic 完全存储编码器功能映射，这将消耗更多的内存(11 倍多)。basic 有一个解码器，每个解码器层有 64 个特征映射。相比使用降维的 FCN-Basic 在每个解码器层中具有较少的(11)个特征映射。这减少了解码器网络中的卷积次数，因此 FCN-Basic 在推理过程中(向前传递)更快。从另一个角度来看，SegNet-Basic 中的解码器网络总体上比 FCN-Basic 的解码器网络更大。这赋予了它更大的灵活性，从而在相同的迭代次数下，达到了比 FCN-Basic 更高的训练精度。

SegNet-basic 在解码器上与 FCN-Basic-NoAdiiition 最相似，虽然 SegNet 的解码器更大。两者都学习生成密集的特征图，或者直接学习像 FCN-Basic-NoAdiiition 那样执行反褶积，或者先向上采样，然后与训练过的解码器滤波器进行卷积。SegNet-Basic 的性能优越，部分原因是它的解码器更大。与 FCN-Basic 相比，FCN-Basic-NoAdiiition 的准确度也较低。这表明，为了获得更好的性能，捕获编码器特性映射中呈现的信息是至关重要的。特别要注意的是 BF 在这两个变量之间进行度量。这也表明了 SegNet-basic 优于 FCN-Basic-NoAddition 部分原因。

fcn-basic-noadd-nodimreduce 模型的大小略大于 SegNet-Basic 模型，因为最终的编码器 feature map 没有被压缩以匹配类的数量这就使得模型的大小进行了公平的比较。这种 FCN 变体的性能比在测试中采用分段基本法，但在相同训练次数下训练精度较低。这表明，使用较大的解码器是不够的，但捕获编码器 feature map 信息，特别是细粒度轮廓信息(请注意 BF 度量的下降)也很重要。有趣的是，与 fcn-basic-nodimreduce 等大型模型相比，SegNet-Basic 具有竞争性的训练精度。

fcn-basic-noA 和 segnet-basic-single-channeldecoder 之间的另一个有趣的比较表明，使用 max-pooling index 进行上采样和使用一个更大的解码器可以获得更好的性能。这也为。提供了证据 SegNet 是一个很好的分割体系结构，特别是在需要在存储成本、准确性和推理时间之间找到折衷的时候。在最好的情况下，当内存和推理时间都不受限制时，大型模型如 fcn - basic - nodimreduce 和 segnet - encoder 加法都比其他变体更准确。特别是，在 FCN-Basic 模型中丢弃降维处理，在 BF 评分较高的 FCN-Basic 变量中，其性能最好。这再次强调了在分割架构中内存和准确性之间的权衡。

在最好的情况下，当内存和推理时间都不受约束时，诸如 FCN-Basic-NoDimReduction 和 SegNet-EncoderAddition 之类的较大型号比其他变体更准确。特别地，在 FCN-Basic 模型中丢弃维数降低导致具有高 BF 分数的 FCN Basic 变体中的最佳性能。这再次强调了分割架构中存储器与精度之间的权衡。

我们现在可以总结上述分析，具有以下一般要点:

1）编码器特征图全部存储时，性能最好。 这最明显地反映在语义轮廓描绘度量（BF）中。

2）当限制推理中的存储器时，可以使用适当的解码器（例如 SegNet 类型）来存储和使用编码器特征图（维数降低，最大聚集索引）的压缩形式来提高性能。

3）更大的解码器提高了给定编码器网络的性能。

4. 基准测试

我们使用 Caffe 实现 3 在两个场景分割基准上量化 SegNet 的性能。第一个任务是道路场景分割，这对于当前各种与自主驾驶相关的问题具有实际意义。第二个任务是室内场景分割，这是目前几个增强现实(AR)应用的直接利益。这两个任务的输入 RGB 图像是 360×480。

我们将 SegNet 与其他几个很好采用的深度架构进行了基准测试，例如 FCN，DeepLab-LargFOV[3]和 DeconvNet。我们的目标是在相同数据集上端到端地训练时理解这些体系结构的性能。

为了实现端到端训练，我们在每个卷积层之后添加了批量标准化层。对于 DeepLab-LargeFOV，我们将 max pooling 3 stride 更改为 1 以达到 45×60 的最终预测分辨率。我们将 DeconvNet 完全连接的图层中的特征尺寸限制为 1024，以便能够以与其他型号。请注意，DeepLab-LargeFOV 的作者也报告说，通过减小完全连接层的大小，性能几乎没有损失。

为了执行受控基准，我们使用相同的 SGD 求解器[17]，固定学习率为 10-3，动量为 0.9。通过数据集对超过 100 个时期进行优化，直到观察不到进一步的性能增加。在所有模型中，在 0.5 更深的卷积层的末尾添加了 0.5 的丢失，以防止过度拟合（参见 http://mi.eng.cam.ac.uk/projects/segnet/tutorial.html，例如 caffe prototxt）。对于有 11 个班级的道路场景，我们使用了 5 个小批量，对于 37 个班级的室内场景，我们使用了 4 个小批量。

4.1 道路场景分割

许多道路场景数据集可用于语义分析。如果我们选择使用 CamVid 数据集对 SegNet 进行基准测试，因为它包含视频序列。这使我们能够将我们提出的架构与使用运动和结构和视频片段的架构进行比较。我们还将[22]，[26]，[60]，[61]结合起来形成了 3433 个图像的集合，以训练 SegNet 获得额外的基准。对于道路场景分割的网络演示（见脚注 3），我们包括 CamVid 测试设置为这个更大的数据集。在这里，我们要注意的是，已经为 SegNet 和本文中使用的其他竞争架构执行了另一个近期和独立的道路场景分割基准。但是，基准测试没有得到控制，这意味着每个架构都采用不同输入分辨率的单独配方进行训练，有时还包含验证集。因此，我们认为可以使用更加可控的基准来补充他们的工作。

SegNet 预测与其他深层架构的定性比较可以在图 4 中看到。定性结果表明所提出的架构能够在道路场景中分割较小的类别，同时产生整个场景的平滑分割。事实上，在受控制下 基准设置，与一些较大的模型相比，SegNet 显示出卓越的性能.DeepLab-LargeFOV 是最有效的模型，CRF 后处理可以产生有竞争力的结果，尽管较小的类丢失。具有学习反卷积的 FCN 明显优于固定的 双线性上采样.DeconvNet 是最大的模型，也是训练效率最低的。它的预测不会保留小班。

我们还使用这个基准来首先将 SegNet 与几种非深度学习方法进行比较，包括随机森林，Boosting 结合基于 CRF 的方法。 这样做是为了向用户提供与基于经典特征工程的技术相比使用深度网络实现的精度改进的视角。

表 2 中的结果显示 SegNet-Basic，SegNet 与使用 CRF 的方法相比获得了有竞争力的结果。这表明深层架构从输入图像中提取有意义的特征并将其映射到准确和平滑的类段标签的能力。这里最

有趣的结果是，当通过组合[22]，[26]，[60]，[61]获得的大型训练数据集用于训练 SegNet 时，在类平均值和 mIOU 指标上获得了巨大的性能提升。相应地，SegNet 的定性结果（见图 4）明显优于其他方法。它能够很好地分割小类和大类。我们在这里说我们在训练 SegNet 中使用中值频率类平衡-Basic 和 SegNet。此外，整体平滑的分割质量很像 CRF 后处理通常获得的。虽然结果随着更大的训练集而改善的事实并不令人惊讶，但是改进的百分比得到了改进。经过预先训练的编码器网络和该训练集表明该架构可以部署用于实际应用。我们对来自互联网的城市和高速公路图像进行随机测试（见图 1）表明 SegNet 可以吸收大量训练集并进行推广很好地看不见图像。它还表明，当提供足够数量的训练数据时，可以减少先前（CRF）的贡献。

在表 3 中，我们将 SegNet 的性能与现在广泛采用的用于分段的完全卷积体系结构进行比较。与表 2 中的实验相比，我们没有使用任何类策略来训练包括 SegNet 在内的任何深层体系结构。这是因为我们发现很难训练更大的模型，比如使用中值频率平衡的 DeconvNet。我们基于 40K，80K 和大于 80K 迭代的基准性能，给出迷你批量大小和训练集大小大约相当于 50,100 和大于 100 个纪元。对于最后一个测试点我们也报告最大迭代次数（此处至少 150 个时期），超过此时我们观察到没有准确性改进或过度拟合设置。我们在训练阶段的三个阶段报告指标，以揭示指标如何随训练时间而变化，特别是更大的网络。这一点很重要，了解在准确度提高时，额外的培训时间是否合理。另请注意，对于每次评估，我们都要进行完成数据集的运行以获取批处理规范统计信息，然后使用此统计信息评估测试模型（有关代码，请参阅 http://mi.eng.cam.ac.uk/projects/segnet/tutorial.html。）。这些评估在大型训练集上执行是昂贵的，因此我们仅在训练阶段的三个时间点报告指标。

Method	Building	Tree	Sky	Car	Sign-Symbol	Road	Pedestrian	Fence	Column-Pole	Side-walk	Bicyclist	Class avg.	Global avg.	mIoU	BF
SfM+Appearance [28]	46.2	61.9	89.7	68.6	42.9	89.5	53.6	46.6	0.7	60.5	22.5	53.0	69.1	n/a*	
Boosting [29]	61.9	67.3	91.1	71.1	58.5	92.9	49.5	37.6	25.8	77.8	24.7	59.8	76.4	n/a*	
Dense Depth Maps [32]	85.3	57.3	95.4	69.2	46.5	**98.5**	23.8	44.3	22.0	38.1	28.7	55.4	82.1	n/a*	
Structured Random Forests [31]	n/a											51.4	72.5	n/a*	
Neural Decision Forests [64]	n/a											56.1	82.1	n/a*	
Local Label Descriptors [65]	80.7	61.5	88.8	16.4	n/a	98.0	1.09	0.05	4.13	12.4	0.07	36.3	73.6	n/a*	
Super Parsing [33]	87.0	67.1	96.9	62.7	30.1	95.9	14.7	17.9	1.7	70.0	19.4	51.2	83.3	n/a*	
SegNet (3.5K dataset training - 140K)	**89.6**	**83.4**	96.1	**87.7**	52.7	96.4	**62.2**	**53.45**	**32.1**	**93.3**	**36.5**	**71.20**	**90.40**	60.10	46.84
CRF based approaches															
Boosting + pairwise CRF [29]	70.7	70.8	94.7	74.4	55.9	94.1	45.7	37.2	13.0	79.3	23.1	59.9	79.8	n/a*	
Boosting+Higher order [29]	84.5	72.6	**97.5**	72.7	34.1	95.3	34.2	45.7	8.1	77.6	28.5	59.2	83.8	n/a*	
Boosting+Detectors+CRF [30]	81.5	76.6	96.2	78.7	40.2	93.9	43.0	47.6	14.3	81.5	33.9	62.5	83.8	n/a*	

TABLE 2
Quantitative comparisons of SegNet with traditional methods on the CamVid 11 road class segmentation problem [22]. SegNet outperforms all the other methods, including those using depth, video and/or CRF's on the majority of classes. In comparison with the CRF based methods SegNet predictions are more accurate in 8 out of the 11 classes. It also shows a good $\approx 10\%$ improvement in class average accuracy when trained on a large dataset of 3.5K images. Particularly noteworthy are the significant improvements in accuracy for the smaller/thinner classes. * Note that we could not access predictions for older methods for computing the mIoU, BF metrics.

从表 3 中我们立即看到，与其他模型相比，SegNet，DeconvNet 在所有指标中获得最高分。DeconvNet 具有更高的边界描绘精度，但与 DeconvNet 相比，SegNet 的效率更高。 这可以从计算统计如表 6 FCN 可以看出，具有完全连接层（变成卷积层）的 DeconvNet 训练速度要慢得多，并且参考 SegNet 具有相当或更高的前后通过时间。 在这里我们还注意到过度拟合不是训练这些较大模型的问题，因为在与 SegNet 的可比迭代中，它们的度量显示出增加的趋势。

对于 FCN 模型，学习反卷积层而不是用双线性插值权重来修复它们可以提高性能，特别是 BF 得分。它还可以在更短的时间内实现更高的指标。这个事实与我们早先在 3.3 节中的分析一致。

令人惊讶的是，DeepLab-LargeFOV 被训练用于以 45×60 的分辨率预测标签，因为它是参数化方面最小的模型，并且具有最快的训练时间，如表 6 所示。但是，边界精度是更穷，这是其他架构共享的。在很长一段时间内训练后，DeconvNet 的 BF 分数高于其他网络。鉴于我们在 Sec 3.3 中的分析。以及它共享 SegNet 类型架构的事实。在 DeepLab-LargeFOV-denseCRF 的最后一个时间点可以看到密集 CRF [63]后处理的影响。全球和 mIoU 都有所改善，但是班级平均水平有所下降。然而，BF 分数获得了很大的改进。请注意，密集 CRF 超参数是通过对训练集的子集进行昂贵的网格搜索过程获得的，因为没有可用的验证集。

Network/Iterations	40K				80K				>80K				Max iter
	G	C	mIoU	BF	G	C	mIoU	BF	G	C	mIoU	BF	
SegNet	88.81	59.93	50.02	35.78	89.68	69.82	57.18	42.08	90.40	71.20	60.10	46.84	140K
DeepLab-LargeFOV [3]	85.95	60.41	50.18	26.25	87.76	62.57	53.34	32.04	88.20	62.53	53.88	32.77	140K
DeepLab-LargeFOV-denseCRF [3]	not computed								89.71	60.67	54.74	40.79	140K
FCN	81.97	54.38	46.59	22.86	82.71	56.22	47.95	24.76	83.27	59.56	49.83	27.99	200K
FCN (learnt deconv) [2]	83.21	56.05	48.68	27.40	83.71	59.64	50.80	31.01	83.14	64.21	51.96	33.18	160K
DeconvNet [4]	85.26	46.40	39.69	27.36	85.19	54.08	43.74	29.33	89.58	70.24	59.77	52.23	260K

TABLE 3

Quantitative comparison of deep networks for semantic segmentation on the CamVid test set when trained on a corpus of 3433 road scenes *without class balancing*. When end-to-end training is performed with the same and fixed learning rate, smaller networks like SegNet learn to perform better in a shorter time. The BF score which measures the accuracy of inter-class boundary delineation is significantly higher for SegNet, DeconvNet as compared to other competing models. DeconvNet matches the metrics for SegNet but at a much larger computational cost. Also see Table 2 for individual class accuracies for SegNet.

Network/Iterations	80K				140K				>140K				Max iter
	G	C	mIoU	BF	G	C	mIoU	BF	G	C	mIoU	BF	
SegNet	70.73	30.82	22.52	9.16	71.66	37.60	27.46	11.33	72.63	44.76	31.84	12.66	240K
DeepLab-LargeFOV [3]	70.70	41.75	30.67	7.28	71.16	42.71	31.29	7.57	71.90	42.21	32.08	8.26	240K
DeepLab-LargeFOV-denseCRF [3]	not computed								66.96	33.06	24.13	9.41	240K
FCN (learnt deconv) [2]	67.31	34.32	24.05	7.88	68.04	37.2	26.33	9.0	68.18	38.41	27.39	9.68	200K
DeconvNet [4]	59.62	12.93	8.35	6.50	63.28	22.53	15.14	7.86	66.13	32.28	22.57	10.47	380K

TABLE 4

Quantitative comparison of deep architectures on the SUNRGB-D dataset when trained on a corpus of 5250 indoor scenes. Note that only the RGB modality was used in these experiments. In this complex task with 37 classes all the architectures perform poorly, particularly because of the smaller sized classes and skew in the class distribution. DeepLab-Large FOV, the smallest and most efficient model has a slightly higher mIoU but SegNet has a better G,C,BF score. Also note that when SegNet was trained with *median frequency class balancing* it obtained 71.75, 44.85, 32.08, 14.06 (180K) as the metrics.

4.2 SUN RGB-D 室内场景

SUN RGB-D 是一个非常具有挑战性的大型室内场景数据集，具有 5285 次训练和 5050 次测试图像。图像由不同的传感器捕获，因此具有各种分辨率。任务是分割 37 个室内场景类，包括墙壁，地板，天花板，桌子，椅子，沙发等。由于对象类具有各种形状，大小和不同姿势，因此很难完成这项任务。由于在每个测试图像中通常存在许多不同的类，因此存在频繁的部分遮挡。这些因素使其成为最难分割的挑战之一。我们只使用 RGB 模态进行训练和测试。使用深度模态将需要进行架构修改/重新设计。此外，来自当前相机的深度图像的质量需要仔细的后处理以填充缺失的测量值。它们还可能需要使用许多帧的融合来鲁棒地提取用于分割的特征。因此，我们认为使用深度进行分割需要单独的工作，这不属于本文的范围。我们还注意到，早期的基准数据集 NYUv2 作为该数据集的一部分包含在内。

道路场景图像在感兴趣的类别和它们的空间布置方面具有有限的变化。当从移动的车辆捕获时，其中摄像机位置几乎总是平行于路面，限制了视点的可变性。这使得深度网络更容易学会将它们彻底分割。相比之下，室内场景的图像更复杂，因为视点可以变化很多，并且场景中存在的类的数量和它们的空间排列的规律性较小。另一个困难是由场景中对象类的大小变化引起的。来自最近的 SUN RGB-D 数据集的一些测试样本如图 5 所示。我们观察到一些具有较少大类的场景和一些具有密集杂波的场景（底行和右）。外观（纹理和形状）在室内场景中也可以广泛变化。因此，我们认为这是计算机视觉中分段架构和方法面临的最大挑战。其他挑战，如 Pascal VOC12 显着对象分割已经占据了研究人员，但我们认为室内场景分割更具挑战性，并且具有更多当前的实际应用，例如 AR 和机器人技术。为了鼓励在这个方向上进行更多研究，我们在大型 SUN RGB-D 数据集上比较了众所周知的深层架构。

SegNet 对不同类型的室内场景样本（如卧室，客厅，实验室，会议室，浴室）的定性结果如图 5 所示。我们看到，当不同类别的大小不同时，SegNet 获得了合理的预测。 这特别有趣，因为输入模态只是 RGB。 RGB 图像也可用于分割较薄的结构，例如椅子和桌子的腿，使用当前可用传感器的深度图像难以实现的灯。 这可以从图 5 中的 SegNet，DeconvNet 的结果中看出。对于 AR 任务，将装饰性物体（例如墙壁上的绘画）分割也是有用的。 然而，与室外场景相比，分割质量显然更加嘈杂。 当杂波增加时质量显着下降（参见中间栏中的结果样本）。

表 4 中的定量结果表明，所有深层体系结构都具有较低的 mIoU 和边界度量。 全球和阶级平均值（与 mIou 相关）也很小。 SegNet 在 G，C，BF 指标方面优于所有其他方法，并且具有比 DeepLab-LargeFOV 略低的 mIoU。 作为一个独立的实验，我们训练 SegNet 使用中值频率等级平衡

[67]并且指标更高（见表 4），这与我们在 Sec 中的分析一致。3.3。 有趣的是，除了 DeepLab-LargeFOV-denseCRF 的 BF 得分度量之外，使用基于网格搜索的密集 CRF 的最佳超参数变得更糟。可能会找到更优化的设置，但鉴于密集 CRF 的推理时间很长，网格搜索过程太昂贵了。

整体性能不佳的一个原因是该分割任务中的大量类，其中许多占据图像的一小部分并且很少出现。表 5 中报告的准确度清楚地表明，较大的类具有合理的准确性，较小的类具有较低的准确性。这可以通过更大尺寸的数据集和类分布感知训练技术来改进。性能不佳的另一个原因可能在于这些深层架构（所有这些都基于 VGG 架构）无法在室内场景中实现大的可变性。我们这猜想是基于这样一个事实，即最小模型 DeepLab-LargeFOV 在 mIoU 方面产生最佳精度，相比之下，DeconvNet 中的更大参数化，即使经过更长时间的训练，FCN 也没有改善性能（DeconvNet）。这表明可能存在所有架构性能不佳的常见原因。需要更多受控数据集、来验证这一假设。

Wall	Floor	Cabinet	Bed	Chair	Sofa	Table	Door	Window	Bookshelf	Picture	Counter	Blinds
83.42	93.43	63.37	73.18	75.92	59.57	64.18	52.50	57.51	42.05	56.17	37.66	40.29
Desk	Shelves	Curtain	Dresser	Pillow	Mirror	Floor mat	Clothes	Ceiling	Books	Fridge	TV	Paper
11.92	11.45	66.56	52.73	43.80	26.30	0.00	34.31	74.11	53.77	29.85	33.76	22.73
Towel	Shower curtain	Box	Whiteboard	Person	Night stand	Toilet	Sink	Lamp	Bathtub	Bag		
19.83	0.03	23.14	60.25	27.27	29.88	76.00	58.10	35.27	48.86	16.76		

TABLE 5
Class average accuracies of SegNet predictions for the 37 indoor scene classes in the SUN RGB-D benchmark dataset. The performance correlates well with size of the classes in indoor scenes. Note that class average accuracy has a strong correlation with mIoU metric.

Network	Forward pass(ms)	Backward pass(ms)	GPU training memory (MB)	GPU inference memory (MB)	Model size (MB)
SegNet	422.50	488.71	6803	**1052**	117
DeepLab-LargeFOV [3]	**110.06**	**160.73**	**5618**	1993	83
FCN (learnt deconv) [2]	317.09	484.11	9735	1806	539
DeconvNet [4]	474.65	602.15	9731	1872	877

TABLE 6
A comparison of computational time and hardware resources required for various deep architectures. The caffe time command was used to compute time requirement averaged over 10 iterations with mini batch size 1 and an image of 360×480 resolution We used nvidia-smi unix command to compute memory consumption. For training memory computation we used a mini-batch of size 4 and for inference memory the batch size was 1. Model size was the size of the caffe models on disk. SegNet is most memory efficient during inference model.

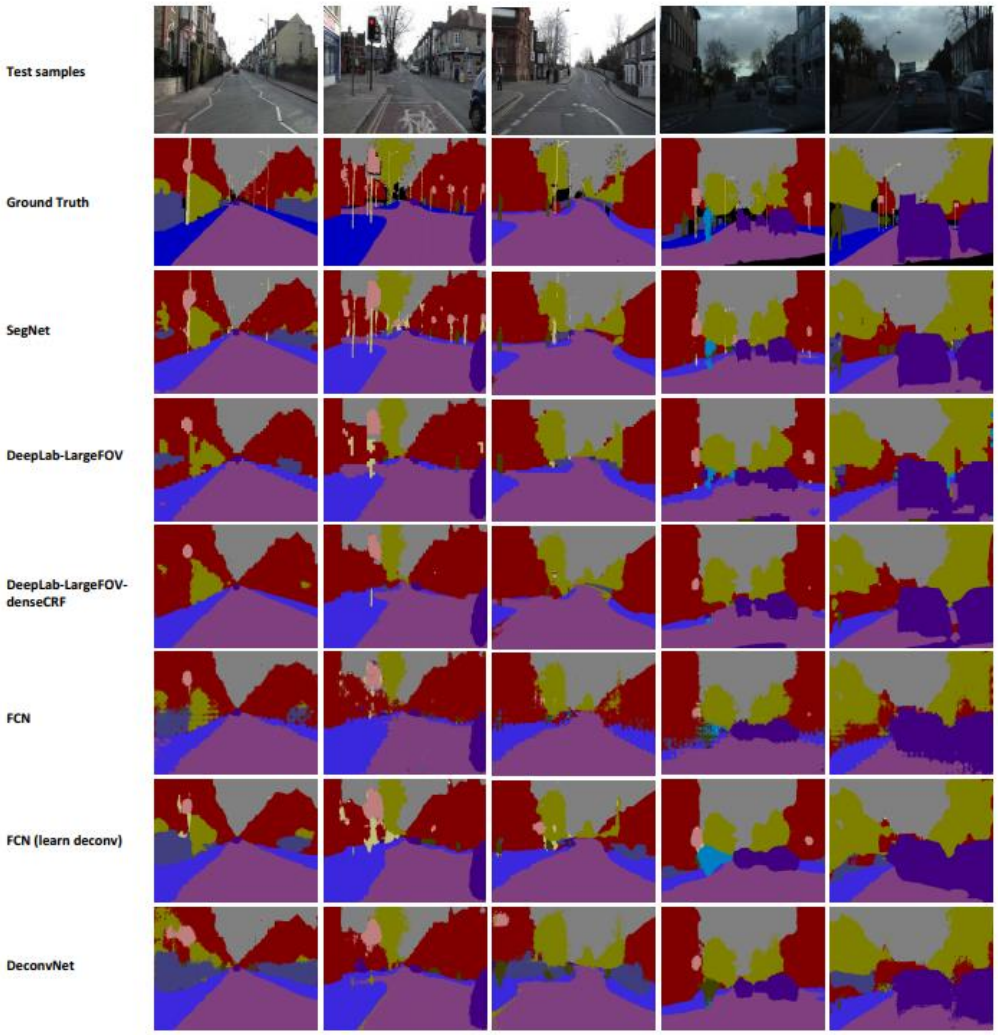

Fig. 4. Results on CamVid day and dusk test samples. SegNet shows superior performance, particularly with its ability to delineate boundaries, as compared to some of the larger models when all are trained in a controlled setting. DeepLab-LargeFOV is the most efficient model and with CRF post-processing can produce competitive results although smaller classes are lost. FCN with learnt deconvolution is clearly better. DeconvNet is the largest model with the longest training time, but its predictions loose small classes. Note that these results correspond to the model corresponding to the highest mIoU accuracy in Table 3.

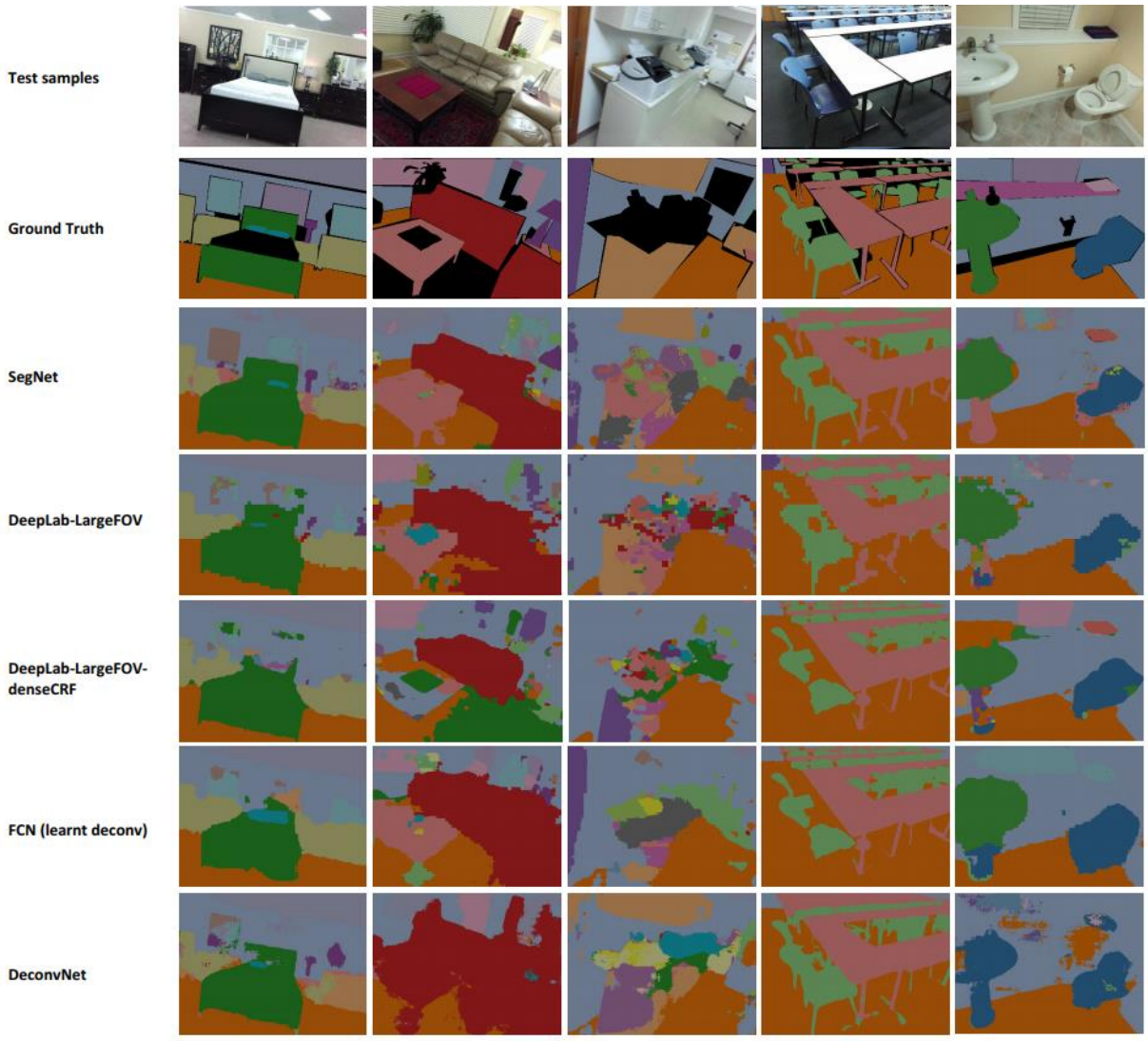

Fig. 5. Qualitative assessment of SegNet predictions on RGB indoor test scenes from the recently released SUN RGB-D dataset [23]. In this hard challenge, SegNet predictions delineate inter class boundaries well for object classes in a variety of scenes and their view-points. Overall rhe segmentation quality is better when object classes are reasonably sized but is very noisy when the scene is more cluttered. Note that often parts of an image of a scene do not have ground truth labels and these are shown in black colour. These parts are not masked in the corresponding deep model predictions that are shown. Note that these results correspond to the model corresponding to the highest mIoU accuracy in Table 4.

5. 讨论和未来工作

由于大量数据集的可用性和扩展的模型深度和参数化，深度学习模型往往取得了更大的成功。然而，在实践中，训练和测试期间的记忆和计算时间等因素是从大型模型库中选择模型时考虑的重要因素。训练时间成为一个重要的考虑因素，特别是当我们的实验显示，性能增益与增加的训练时间不相称时。测试时间记忆和计算负荷对于在专用嵌入式设备上部署模型（例如 AR 应用程序）很重要。从总体效率的角度来看，我们对于更小更多的内存，对于实时应用的时间效率模型

（如道路现场理解和 AR）的关注较少。这是 SegNet 提案的主要动机，它比其他竞争的架构明显更小，更快，但是我们已经表现出对于道路现场理解等任务的效率。

诸如 Pascal 和 MS-COCO 之类的分割挑战是对象分割挑战，其中几个类别存在于任何测试图像中。场景分割更具挑战性，因为室内场景的高度变化，同时需要分割更多的类。户外和室内场景分割的任务在现代应用中也更为实用，如自主驾驶，机器人和 AR。

我们选择了对各种深层分割架构（如边界 F1 测量（BF））进行基准测量的指标，以补充更偏向于区域精度的现有指标。从我们的实验和其他独立的基准可以看出，从移动的汽车捕获的室外场景图像更容易分割，深层结构能够很好地运行。我们希望我们的实验将鼓励研究人员注意更具挑战性的室内场景分割任务。

在对不同参数化的不同深层架构进行基准测试时，我们必须做出的一个重要选择是训练他们的方式。许多这些架构已经使用了许多支持技术和多阶段训练配方来达到数据集的高准确度，但是这使得很难在时间和内存限制下收集关于其真实性能的证据。相反，我们选择执行受控的基准测试，我们使用批处理标准化，使用相同的求解器（SGD）实现端对端训练。然而，我们注意到，这种方法不能完全解开模型与求解器（优化）在实现特定结果时的影响。这主要是由于训练这些网络涉及梯度反向传播，这是不完美的，优化是非常大的非凸的问题。承认这些缺点，我们希望这种受控分析补充了其他基准，并揭示了涉及不同知名架构的实际权衡。

对于未来，我们希望利用我们对从分析中收集到的分段架构的理解，为实时应用设计更有效的架构。我们也有兴趣从深度分段架构中估计预测的模型不确定性。

6. 结论

我们提出了 SegNet，一种用于语义分割的深度卷积网络架构。SegNet 背后的主要动机是需要设计一种有效的道路和室内场景理解架构，这在存储和计算时间方面都是有效的。我们分析了 SegNet，并将其与其他重要变体进行了比较，以揭示涉及设计分段架构的实际权衡，特别是训练时间，内存与精度。存储编码器网络特征的那些架构完整性能最好，但在推理时间消耗更多的内存。另一方面，SegNet 更有效率，因为它仅存储特征映射的最大池索引，并将其用于解码器网络以实现良好的性能。在大型和众所周知的数据集中，SegNet 具有竞争力，实现道路现场理解的高分。深层分割架构的端到端学习是一个更难的挑战，我们希望更多地关注这一重要问题。