9102年了，语义分割的入坑指南和最新进展都是什么样的

选自Medium

作者：Derrick Mwiti

机器之心编译

参与：Nurhachu Null，Geek AI

语义分割指的是将图像中的每一个像素关联到一个类别标签上的过程，这些标签可能包括一个人、一辆车、一朵花、一件家具等等。在这篇文章中，作者介绍了近来优秀的语义分割思想与解决方案，它可以称得上是 2019 语义分割指南了。

我们可以认为语义分割是像素级别的图像分类。例如，在一幅有很多辆车的图像中，分割模型将会把所有的物体（车）标记为车辆。但是，另一种被称为实例分割的模型能够将出现在图像中的独立物体标记为独立的实例。这种分割在被用在统计物体数量的应用中是很有用的（例如，统计商城中的客流量）。

语义分割的一些主要应用是自动驾驶、人机交互、机器人以及照片编辑/创作型工具。例如，语义分割在自动驾驶和机器人领域是十分关键的技术，因为对于这些领域的模型来说，理解它们操作环境的上下文是非常重要的。

图片来源:
http://www.cs.toronto.edu/~tingwuwang/semantic_segmentation.pdf

接下来，我们将会回顾一些构建语义分割模型的最先进的方法的研究论文，它们分别是：

1. Weakly- and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation
2. Fully Convolutional Networks for Semantic Segmentation

3. U-Net: Convolutional Networks for Biomedical Image Segmentation
4. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation
5. Multi-Scale Context Aggregation by Dilated Convolutions
6. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs
7. Rethinking Atrous Convolution for Semantic Image Segmentation
8. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation
9. FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation
10. Improving Semantic Segmentation via Video Propagation and Label Relaxation
11. Gated-SCNN: Gated Shape CNNs for Semantic Segmentation

1. Weakly- and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation (ICCV, 2015)

这篇论文提出了一个解决方法，主要面对处理深度卷积网络中的弱标签数据，以及具有良好标签和未被合适标记得数据的结合时的挑战。在这篇论文结合了深度卷积网络和全连接条件随机场。

- 论文地址：https://arxiv.org/pdf/1502.02734.pdf

在 PASCAL VOC 的分割基准测试中，这个模型高于 70% 的交并比（IOU）

这篇论文的主要贡献如下：

- 为边界框或图像级别的训练引入 EM 算法，这可以用在弱监督和半监督环境中。
- 证明了弱标注和强标注的结合能够提升性能。在合并了 MS-COCO 数据集和 PASCAL 数据集的标注之后，论文的作者在 PASCAL VOC 2012 上达到了 73.9% 的交并比性能。
- 证明了他们的方法通过合并了少量的像素级别标注和大量的边界框标注（或者图像级别的标注）实现了更好的性能。

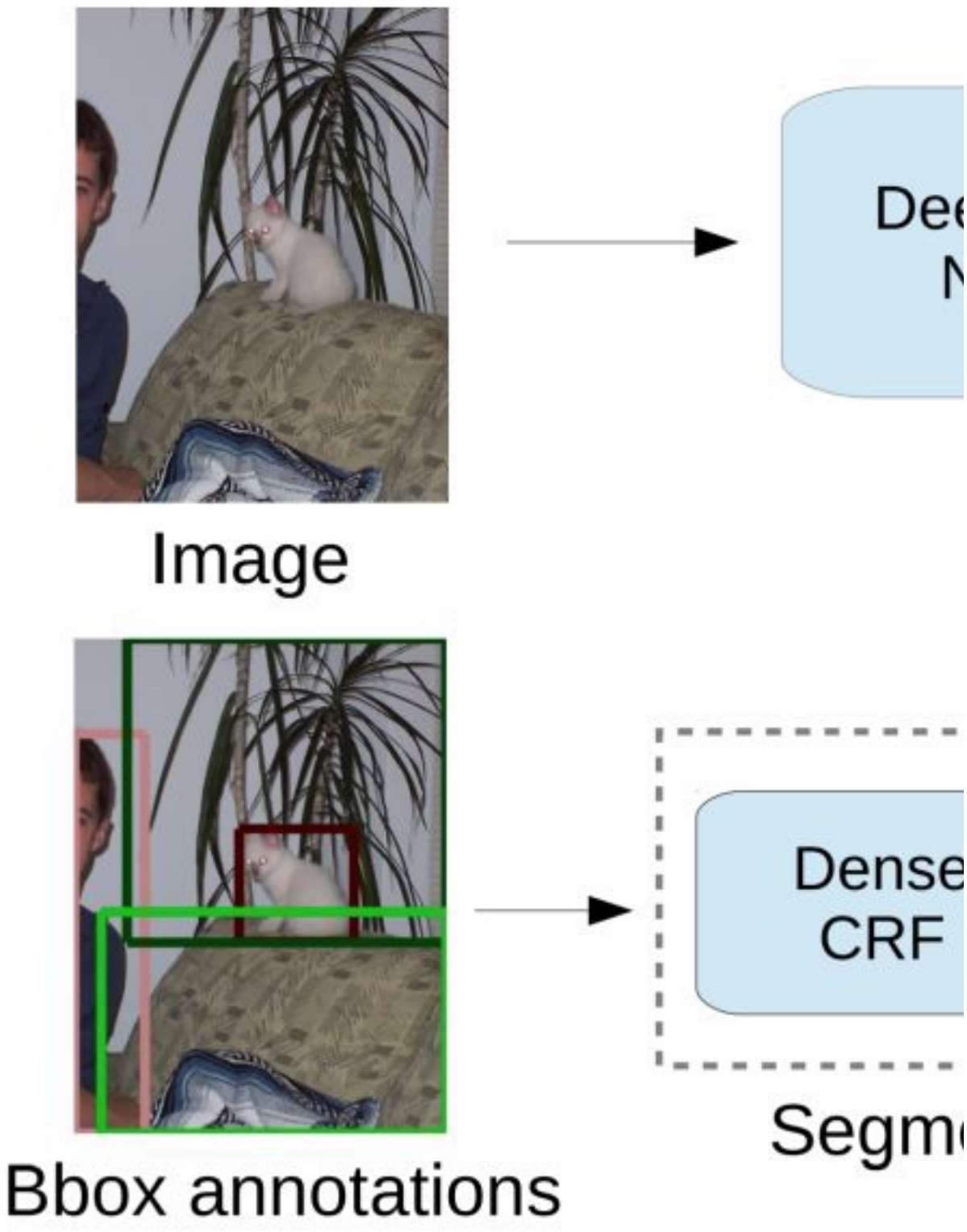


Figure 3. DeepLab mode

2. Fully Convolutional Networks for Semantic Segmentation (PAMI, 2016)

这篇论文提出的模型在 PASCAL VOC 2012 数据集上实现了 67.2% 的平均 IoU。全连接网络以任意大小的图像为输入，然后生成与之对应的空间维度。在这个模型中，ILSVRC 中的分类器被丢在了全连接网络中，并且使用逐像素的损失和上采样模块做了针对

稠密预测的增强。针对分割的训练是通过微调来实现的，这个过程通过在整个网络上的反向传播完成。

- 论文地址：https://arxiv.org/pdf/1605.06211.pdf

3. U-Net: Convolutional Networks for Biomedical Image Segmentation (MICCAI, 2015)

在生物医学图像处理中，得到图像中的每一个细胞的类别标签是非常关键的。生物医学中最大的挑战就是用于训练的图像是不容易获取的，数据量也不会很大。U-Net 是非常著名的解决方案，它在全连接卷积层上构建模型，对其做了修改使得它能够在少量的训练图像数据上运行，得到了更加精确的分割。

- 论文地址：https://arxiv.org/pdf/1505.04597.pdf

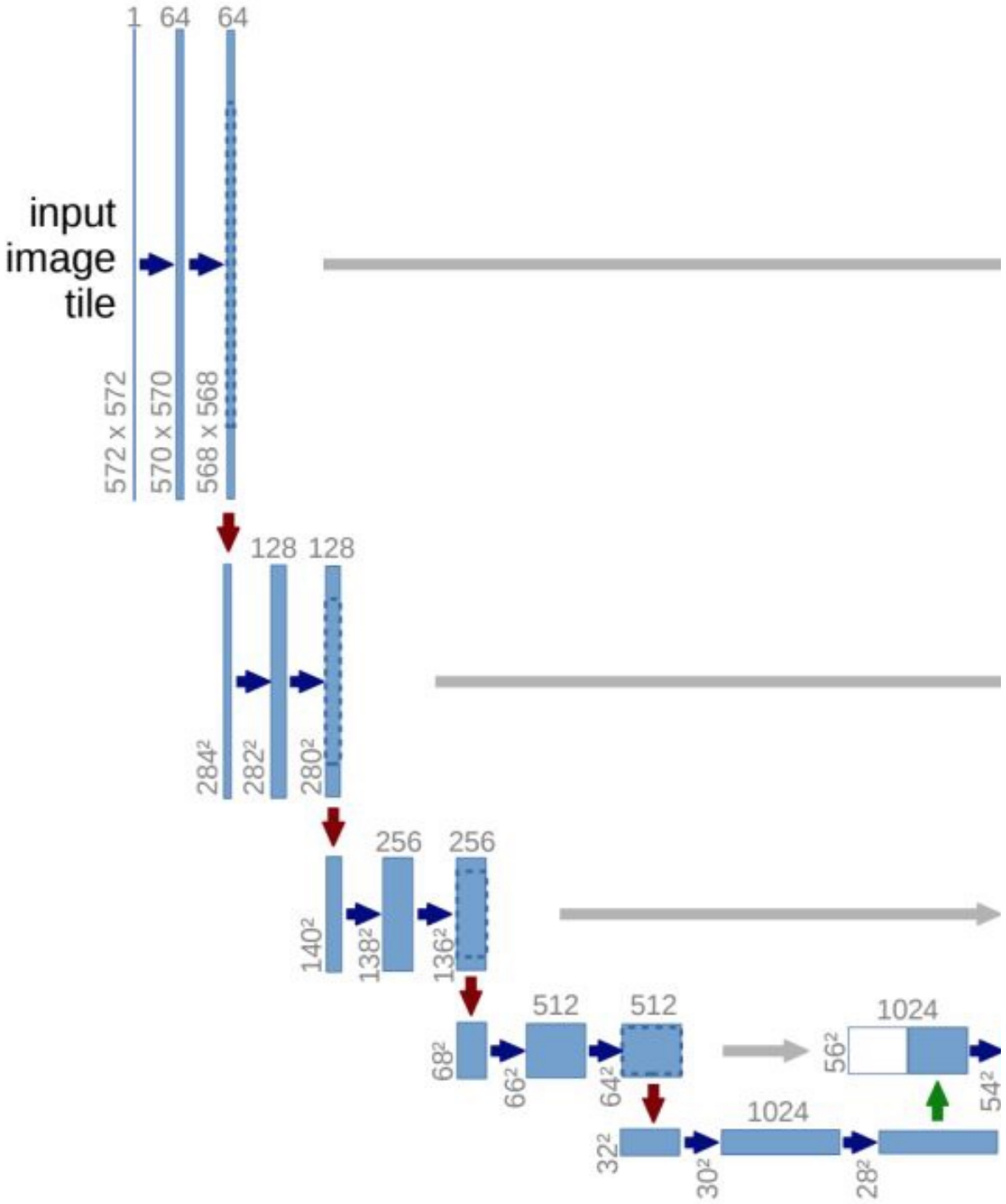


Fig. 1. U-net architecture (example for 32x32 pi
box corresponds to a multi-channel feature ma
on top of the box. The x-y-size is provided at
boxes represent copied feature maps. The arrov

由于少量训练数据是可以获取的，所以这个模型通过在可获得的数据上应用灵活的变形来使用数据增强。正如上面的图 1 所描述的，模型的网络结构由左边的收缩路径和右边的扩张路径组成。

收缩路径由 2 个 3X3 的卷积组成，每个卷积后面跟的都是 ReLU 激活函数和一个进行下采样的 2X2 最大池化运算。扩张路径阶段包括一个特征通道的上采样。后面跟的是 2X2 的转置卷积，它能够将特征通道数目减半，同时加大特征图。最后一层是 1X1 的卷积，用这种卷积来组成的特征向量映射到需要的类别数量上。

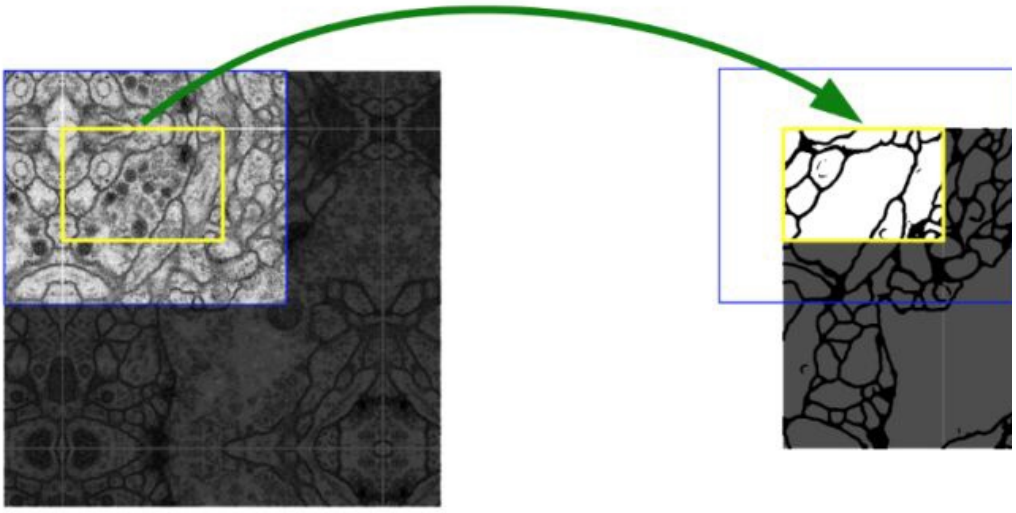


Fig. 2. Overlap-tile strategy for seamless segmentation of arbitrary
segmentation of neuronal structures in EM stacks). Prediction of th
the yellow area, requires image data within the blue area as input. M
is extrapolated by mirroring

在这个模型中，训练是通过输入的图像、它们的分割图以及随机梯度下降来完成的。数据增强被用来教网络学会在使用很少的训练数据时所必需的鲁棒性和不变性。这个模型在其中的一个实验中实现了 92% 的 mIoU。

4. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation (2017)

DenseNets 背后的思想是让每一层以一种前馈的方式与所有层相连接，能够让网络更容易训练、更加准确。

模型架构是基于包含下采样和上采样路径的密集块构建的。下采样路径包含 2 个 Transitions Down (TD)，而上采样包含 2 个 Transitions Up (TU)。圆圈和箭头代表网络中的连接模式。

- 论文地址：https://arxiv.org/pdf/1611.09326.pdf

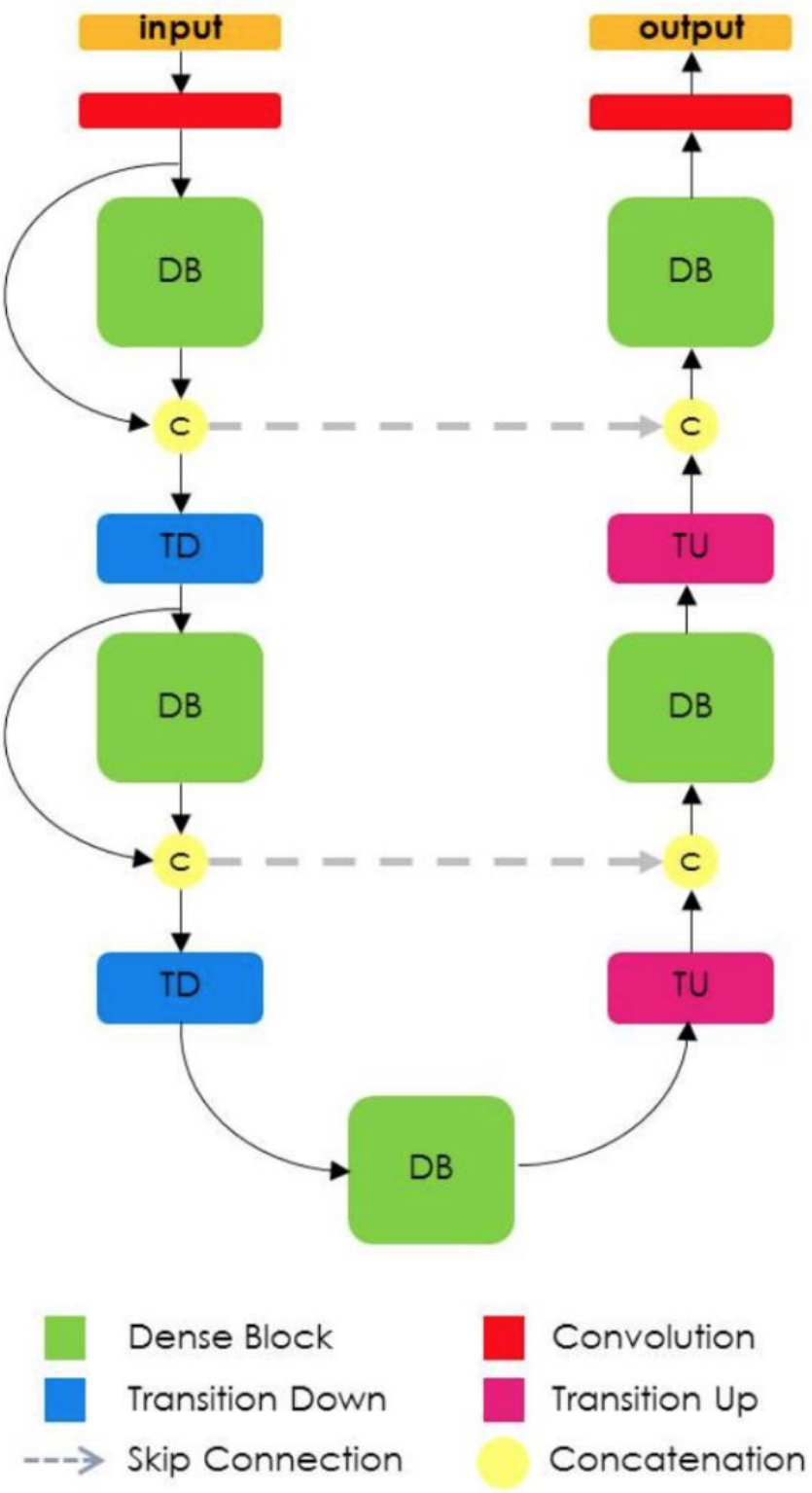


Figure 1. Diagram of our architecture for semantic segmen

这篇论文的主要贡献是：

- 针对语义分割用途，将 DenseNet 的结构扩展到了全卷积网络。
- 提出在密集网络中进行上采样路径，这要比其他的上采样路径性能更好。
- 证明网络能够在标准的基准测试中产生最好的结果。

这个模型在 CamVid 数据集中实现 88% 的全局准确率。

Model	Pretrained	# parameters (M)	Building	Tree
SegNet [1]	✓	29.5	68.7	52.
Bayesian SegNet [15]	✓	29.5		
DeconvNet [21]	✓	252		
Visin et al. [36]	✓	32.3		
FCN8 [20]	✓	134.5	77.8	71.
DeepLab-LFOV [5]	✓	37.3	81.5	74.
Dilation8 [37]	✓	140.8	82.6	76.
Dilation8 + FSO [17]	✓	140.8	**84.0**	77.
Classic Upsampling	✗	20	73.5	72.
FC-DenseNet56 (k=12)	✗	1.5	77.6	72.
FC-DenseNet67 (k=16)	✗	3.5	80.2	75.
FC-DenseNet103 (k=16)	✗	9.4	83.0	**77.**

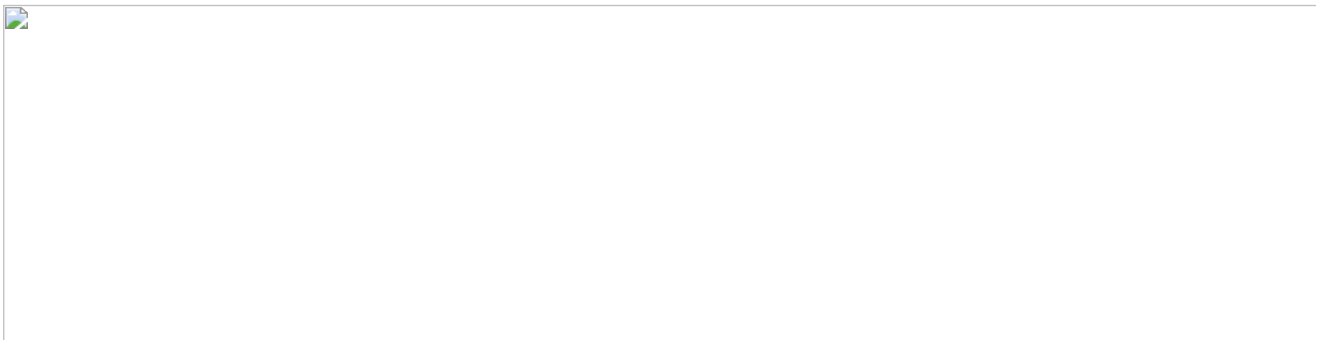
Table 3. Results on CamVid

5. Multi-Scale Context Aggregation by Dilated Convolutions (ICLR, 2016)

这篇论文提出了一个卷积网络模块，能够在不损失分辨率的情况下混合多尺度的上下文信息。然后这个模块能够以任意的分辨率被嵌入到现有的结构中，它主要基于空洞卷积。

- 论文地址：https://arxiv.org/abs/1511.07122

这个模块在 Pascal VOC 2012 数据集上做了测试。结果证明，向现存的语义分割结构中加入上下文模块能够提升准确率。



在实验中训练的前端模块在 VOC-2012 验证集上达到了 69.8% 的平均交并比（mIoU），在测试集上达到了 71.3% 的平均交并比。这个模块对不同对象的预测准确率如下所示：

	aero	bike	bird	boat	bottle	bus	car
FCN-8s	76.8	34.2	68.9	49.4	60.3	75.3	74.7
DeepLab	72	31	71.2	53.7	60.5	77	71.9
DeepLab-Msc	74.9	34.1	72.6	52.9	61.0	77.9	73.0
Our front end	**82.2**	**37.4**	**72.7**	**57.1**	**62.7**	**82.8**	**77.8**

Table 2: Our front-end prediction module reports accuracy on the VOC-2012 test se

6. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs (TPAMI, 2017)

在这篇论文中，作者对语义分割任务中做出了下面的贡献：

- 为密集预测任务使用具有上采样的卷积
- 在多尺度上为分割对象进行带洞空间金字塔池化（ASPP）
- 通过使用 DCNNs 提升了目标边界的定位

- 论文地址：https://arxiv.org/abs/1606.00915

这篇论文提出的 DeepLab 系统在 PASCAL VOC-2012 图像语义分割上实现了 79.7% 的平均交并比（mIoU）。

Method
DeepLab-CRF-LargeFOV-COCO
MERL_DEEP_GCRF [88]
CRF-RNN [59]
POSTECH_DeconvNet_CRF_VO
BoxSup [60]
Context + CRF-RNN [76]
QO_4^{mres} [66]
DeepLab-CRF-Attention [17]
CentraleSuperBoundaries++ [18]
DeepLab-CRF-Attention-DT [63]
H-ReNet + DenseCRF [89]
LRR_4x_COCO [90]
DPN [62]
Adelaide_Context [40]
Oxford_TVG_HO_CRF [91]
Context CRF + Guidance CRF [92
Adelaide_VeryDeep_FCN_VOC [
DeepLab-CRF (ResNet-101)

TABLE 5: Performance on PASCAL
have added some results from recent
the official leadearboard results.

这篇论文解决了语义分割的主要挑战，包括：

- 由重复的最大池化和下采样导致的特征分辨率降低
- 检测多尺度目标
- 因为以目标为中心的分类器需要对空间变换具有不变性，因而降低了由 DCNN 的不变性导致的定位准确率。

Fig. 1: Model Illustration. A Deep Convolutic
convolutional fashion, using atrous convolut
bilinear interpolation stage enlarges the feat
applied to refine the segmentation result and

带洞卷积（Atrous convolution）有两个用途，要么通过插入零值对滤波器进行上采样，要么对输入特征图进行稀疏采样。第二个方法需要通过等于带洞卷积率 r 的因子来对输入特征图进行子采样，然后对它进行去交错（deinterlacing），使其变成 r^2

的低分辨率图，每一个 r×r 区域都有一个可能迁移。在此之后，一个标准的卷积被应用在中间的特征图上，并将其与原始图像分辨率进行交错。

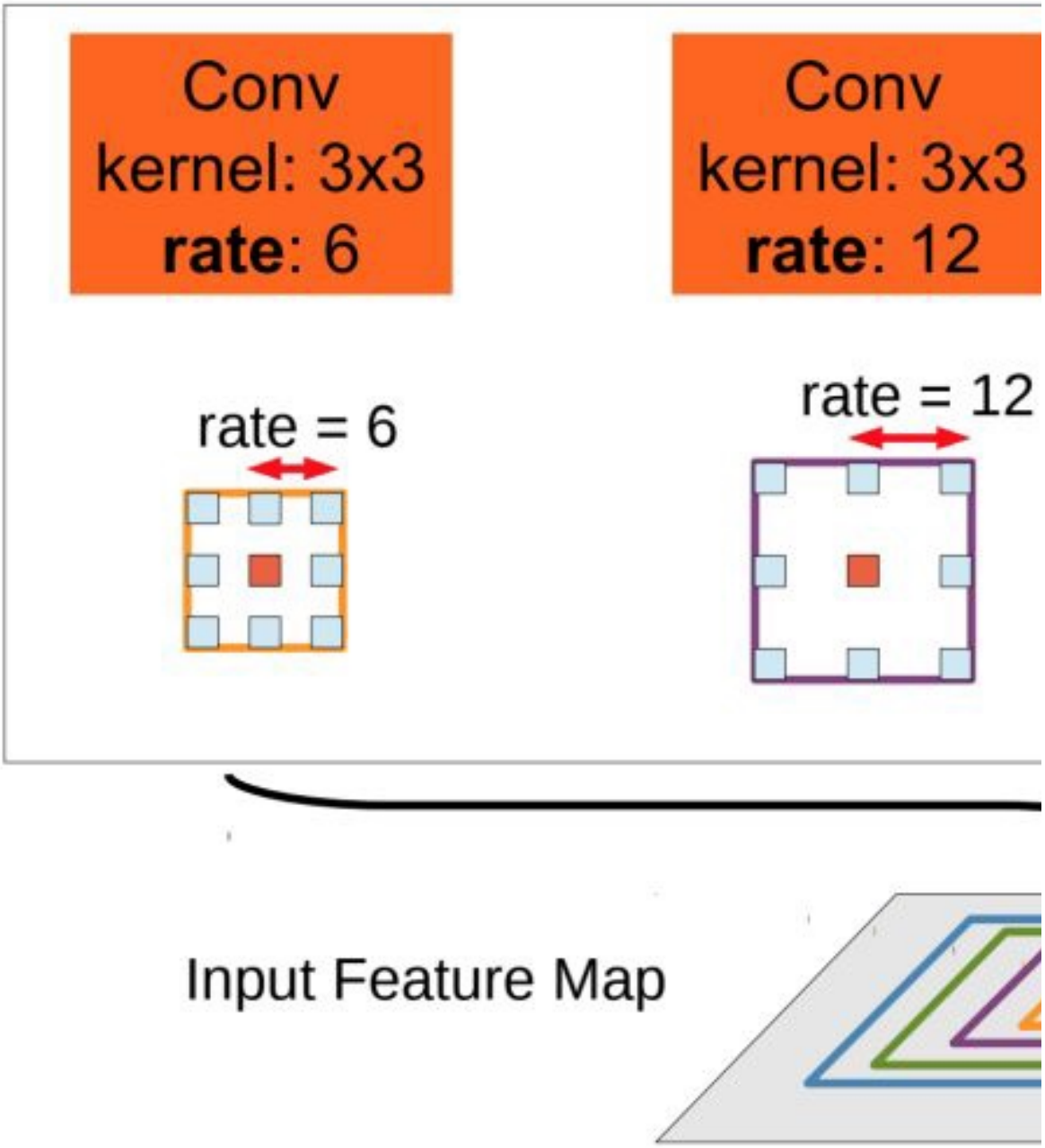


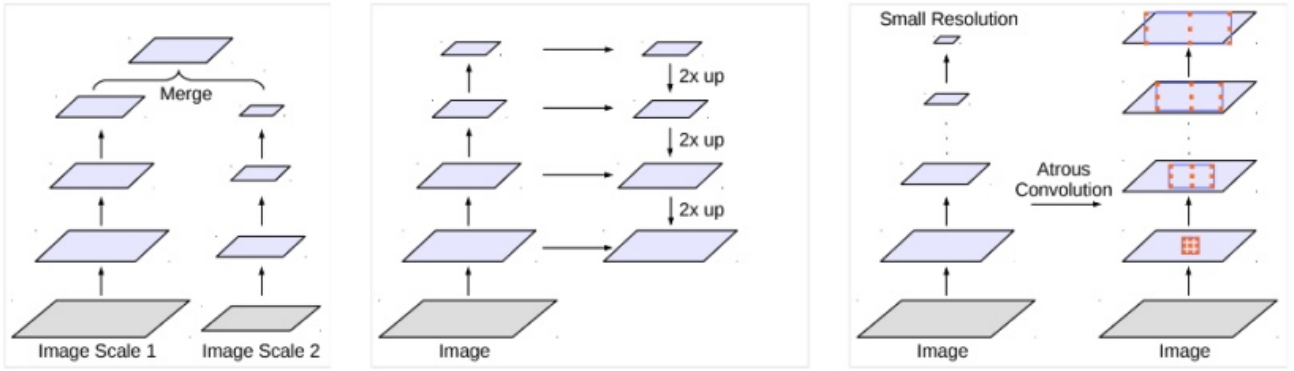
Fig. 4: Atrous Spatial Pyrami
the center pixel (orange), ASP
by employing multiple paral
The effective Field-Of-Views a

7. Rethinking Atrous Convolution for Semantic Image Segmentation (2017)

这篇论文解决了使用 DCNN 进行语义分割所面临的两个挑战（之前提到过）：当使用连续的池化操作时会出现特征分辨率的降低，以及多尺度目标的存在。

- 论文地址：https://arxiv.org/pdf/1706.05587.pdf

为了解决第二个问题，本文提出了带洞卷积（atrous convolution），也被称作 dilated convolution。我们能使用带洞卷积增大感受野，因此能够包含多尺度上下文，这样就解决了第二个问题。



(a) Image Pyramid (b) Encoder-Decoder (c) Deeper w. Atrous Convolution

Figure 2. Alternative architectures to capture multi-scale context.

在没有密集条件随机场（DenseCRF）的情况下，论文的 DeepLabv3 版本在 PASCAL VOC 2012 测试集上实现了 85.7% 的性能。

Method
Adelaide_VeryDeep_FCN_VOC [85]
LRR_4x_ResNet-CRF [25]
DeepLabv2-CRF [11]
CentraleSupelec Deep G-CRF [8]
HikSeg_COCO [80]
SegModel [75]
Deep Layer Cascade (LC) [52]
TuSimple [84]
Large_Kernel_Matters [68]
Multipath-RefineNet [54]
ResNet-38_MS_COCO [86]
PSPNet [95]
IDW-CNN [83]
CASIA_IVA_SDN [23]
DIS [61]
DeepLabv3
DeepLabv3-JFT

Table 7. Performance on PASCAL VOC 201

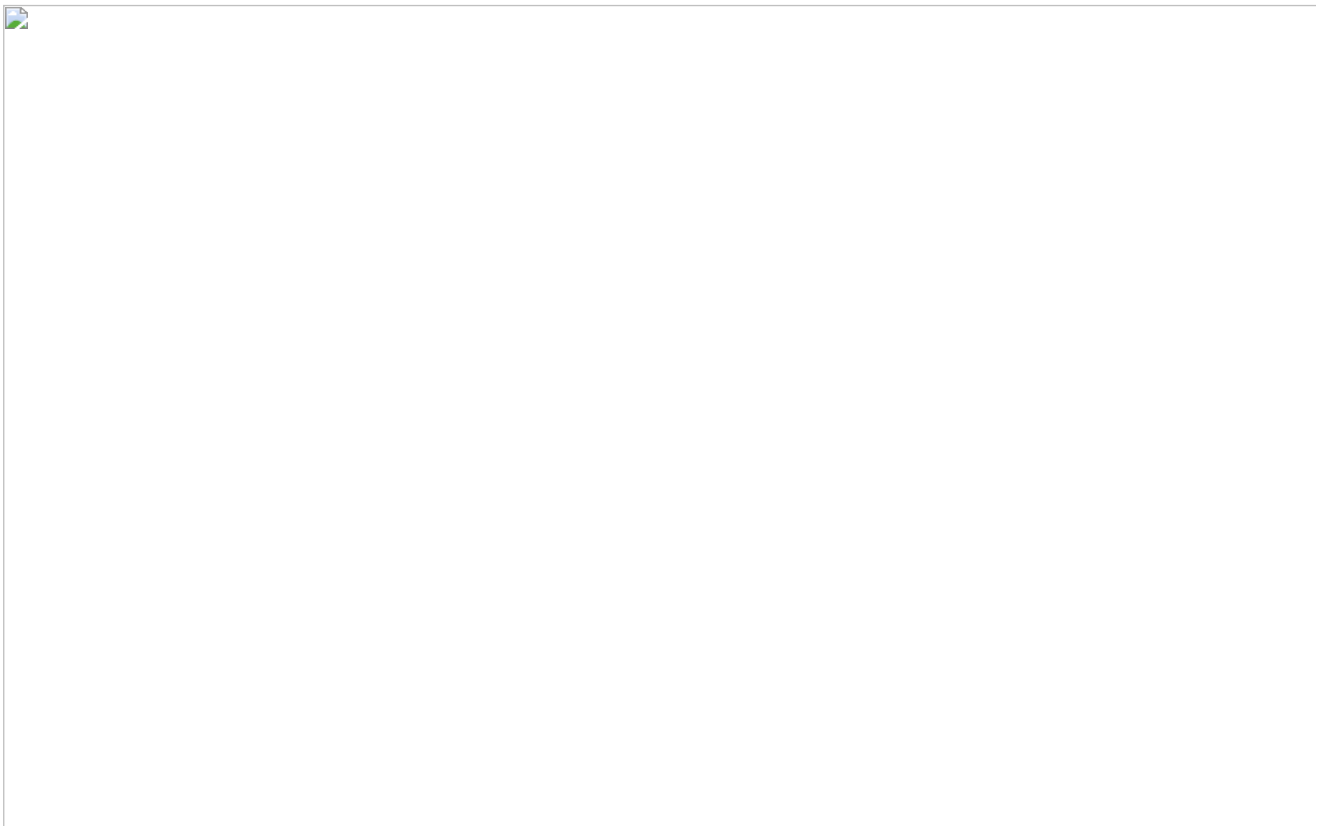
8. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation (ECCV, 2018)

这篇论文的方法「DeepLabv3+」在 PASCAL VOC 2012 数据集和 Cityscapes 数据集上分别实现了 89.0% 和 82.1% 的性能，而且没有做任何后处理。这个模型在 DeepLabv3 的基础上增加一个简单的解码模块，从而改善了分割结果。

- 论文地址：https://arxiv.org/pdf/1802.02611v3.pdf

Fig. 6. Visualization results on *val* set. The last row shows

这篇论文实现了为语义分割使用两种带空间金字塔池化的神经网络。一个通过以不同的分辨率池化特征捕捉上下文信息，另一个则希望获取明确的目标边界。



9. FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation (2019)

这篇论文提出了一种被称作联合金字塔上采样（Joint Pyramid Upsampling/JPU）的联合上采样模块来代替消耗大量时间和内存的带洞卷积。它通过把抽取高分辨率图的方法形式化，并构建成一个上采样问题来取得很好的效果。

- 论文地址：https://arxiv.org/pdf/1903.11816v1.pdf

此方法在 Pascal Context 数据集上实现了 53.13% 的 mIoU，并且具有三倍的运行速度。

该方法以全卷积网络（FCN）作为主体架构，同时应用 JPU 对低分辨率的最终特征图进行上采样，得到了高分辨率的特征图。使用 JPU 代替带洞卷积并不会造成任何性能损失。

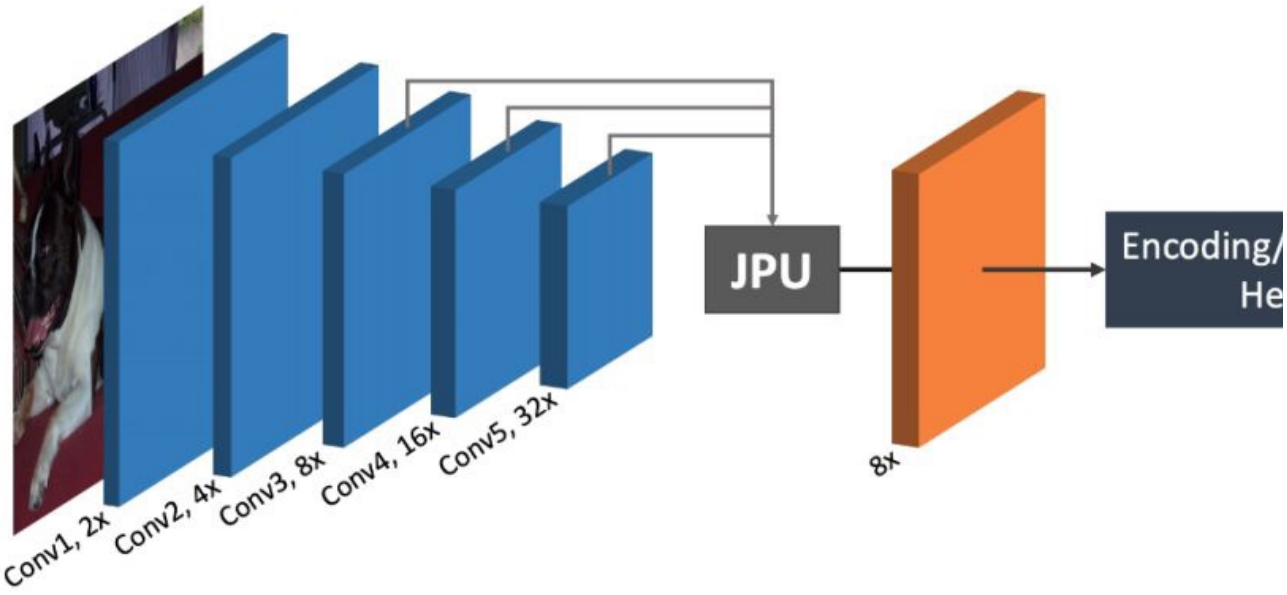


Figure 2: **Framework Overview of Our Method.** Our method employs the same backbone as backbone, a novel upsampling module named Joint Pyramid Upsampling (JPU) is proposed, whic maps as the inputs and generates a high-resolution feature map. A multi-scale/global context r produce the final label map. Best viewed in color.

联合采样使用低分辨率的目标图像和高分辨率的指导图像。然后通过迁移指导图像的结构和细节生成高分辨率的目标图像。

10. Improving Semantic Segmentation via Video Propagation and Label Relaxation (CVPR, 2019)

这篇论文提出了基于视频的方法来增强数据集，它通过合成新的训练样本来达到这一效果，并且该方法还能提升语义分割网络的准确率。本文探讨了视频预测模型预测未来帧的能力，进而继续预测未来的标签。

- 论文地址：https://arxiv.org/pdf/1812.01593v3.pdf

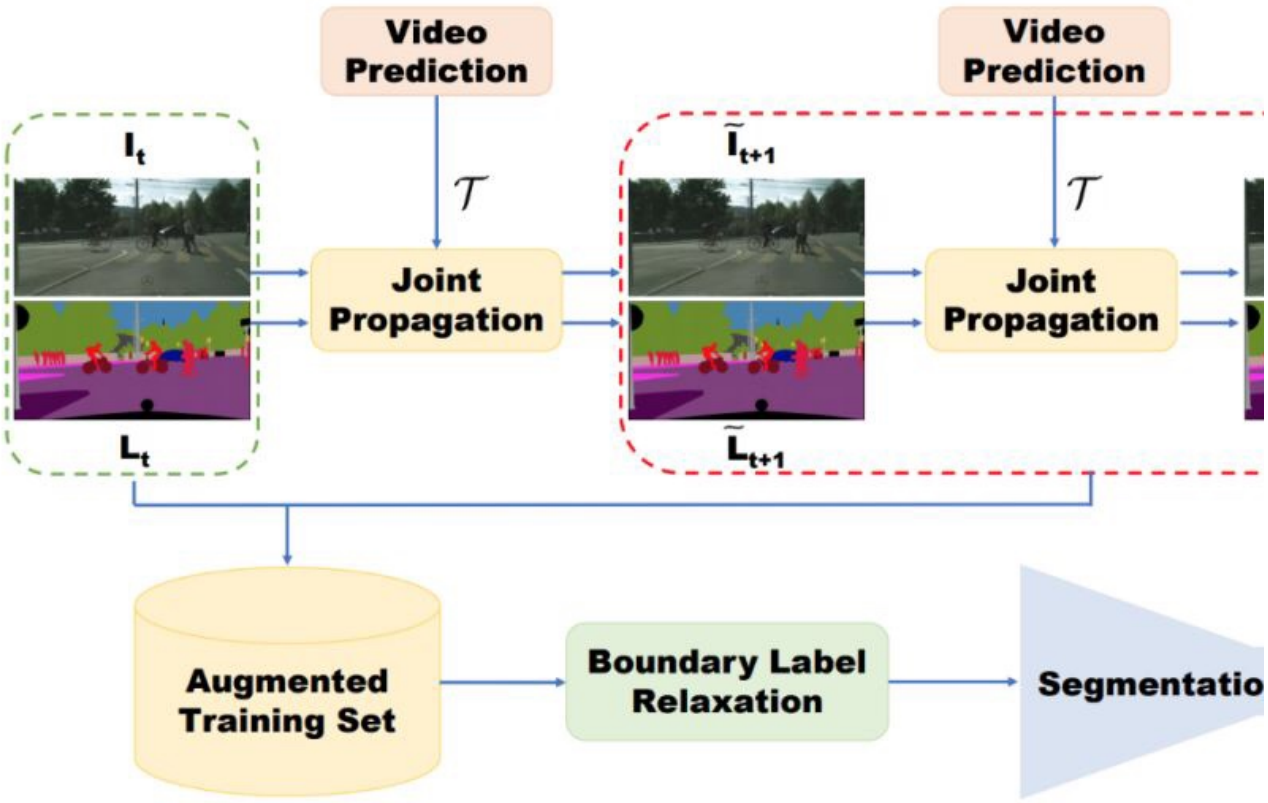


Figure 1: Framework overview. We propose joint propagation to scale up training sets for robust sema tation. The green dashed box includes manually label and the red dashed box includes our propagated samp transformation function learned by the video predicti perform propagation. We also propose boundary lab to mitigate label noise during training. Our framework with most semantic segmentation and video predictio

这篇论文证明了用合成数据训练语义分割网络能够带来预测准确率的提升。论文提出的方法在 Cityscape 上达到了 8.5% 的 mIoU，在 CamVid 上达到了 82.9% 的 mIoU。

Table 4: Results on the CamVid test set. Pre-train
source dataset on which the model is trained.

Method	Pre-train	Encoder	r
SegNet [3]	ImageNet	VGG16	
RTA [19]	ImageNet	VGG16	
Dilate8 [42]	ImageNet	Dilate	
BiSeNet [41]	ImageNet	ResNet18	
PSPNet [44]	ImageNet	ResNet50	
DenseDecoder [6]	ImageNet	ResNeXt101	
VideoGCRF [11]	Cityscapes	ResNet101	
Ours (baseline)	Cityscapes	WideResNet38	
Ours	Cityscapes	WideResNet38	

论文提出了两种预测未来标签的方法：

- Label Propagation (标签传播，LP)：通过将原始的未来帧与传播来的标签配对来创建新的训练样本。
- Joint image-label Propagation (联合图像标签传播，JP)：通过配对对应的传播图像与传播标签来创建新的训练样本。

这篇论文有 3 个主要贡献：利用视频预测模型将标签传播到当前的邻帧，引入联合图像标签传播（JP）来处理偏移问题，通过最大化边界上分类的联合概率来松弛 one-hot 标签训练。

11. Gated-SCNN: Gated Shape CNNs for Semantic Segmentation (2019)

这篇论文是语义分割领域最新的成果（2019.07），作者提出了一个双流 CNN 结构。在这个结构中，目标的形状信息通过一个独立的分支来处理，该形状流仅仅处理边界相关的信息。这是由模型的门卷控积层（GCL）和局部监督来强制实现的。

- 论文地址：https://arxiv.org/pdf/1907.05740.pdf

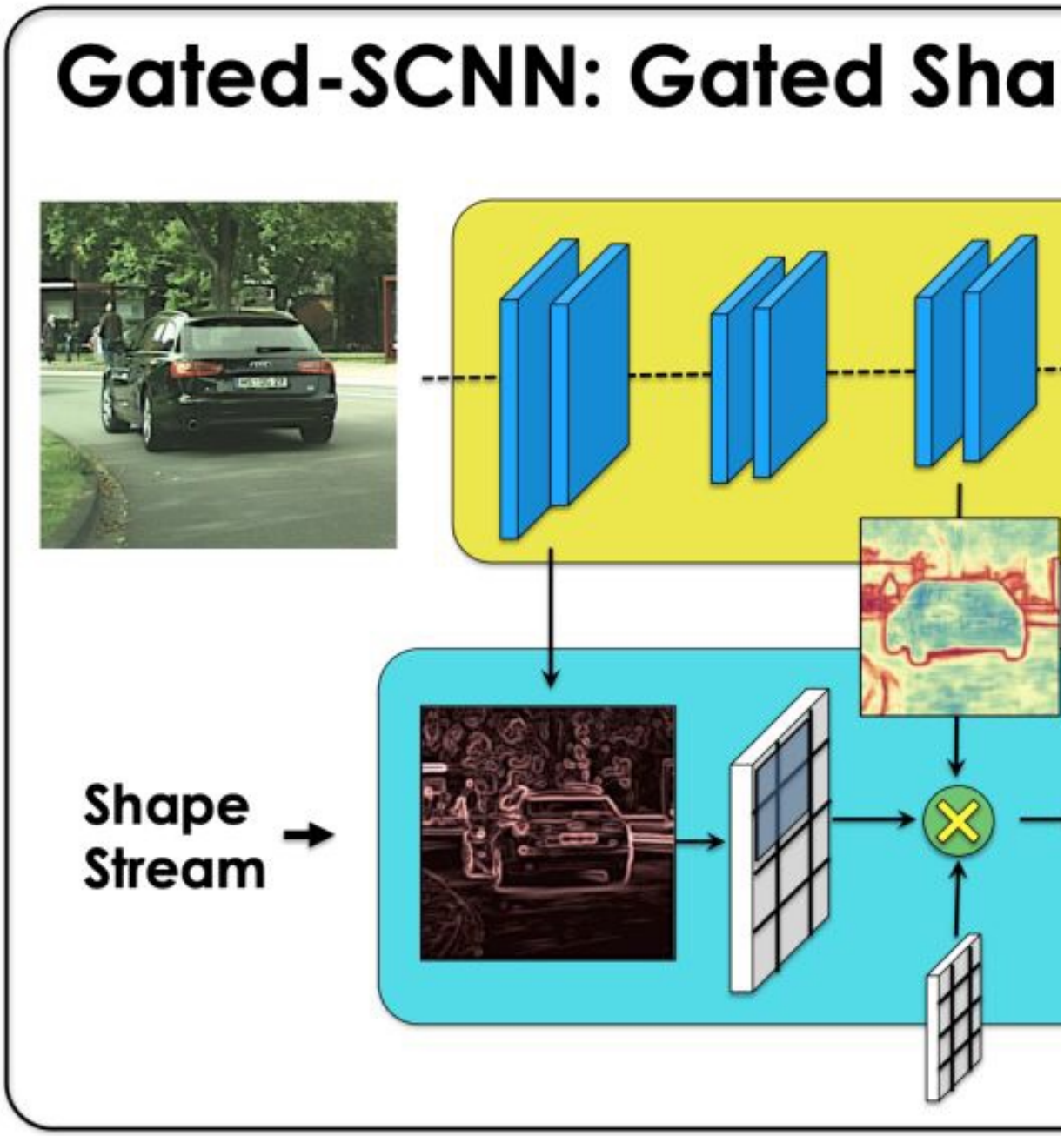


Figure 1: We introduce *Gated-SCNN* architecture for semantic segmentation tion as a separate processing stream. nism to connect the intermediate laye streams is done at the very end throug quality boundaries, we exploit a new lo

dicted semantic segmentation masks to

在用 Cityscapes 基准测试中，这个模型的 mIoU 比 DeepLab-v3 高出 1.5%，F-boundary 得分比 DeepLab-v3 高 4%。在更小的目标上，该模型能够实现 7% 的 IoU 提升。下表展示了 Gated-SCNN 与其他模型的性能对比。

以上就是近来语义分割的主要进展，随着模型和数据的进一步提升，语义分割的速度越来越快、准确率越来越高，也许以后它能应用到各种现实生活场景中。

原文链接：https://heartbeat.fritz.ai/a-2019-guide-to-semantic-segmentation-ca8242f5a7fc