论文阅读：Semantic Diffusion Network for Semantic Segmentation

#论文阅读 #diffusion_model #sematic_segmentation

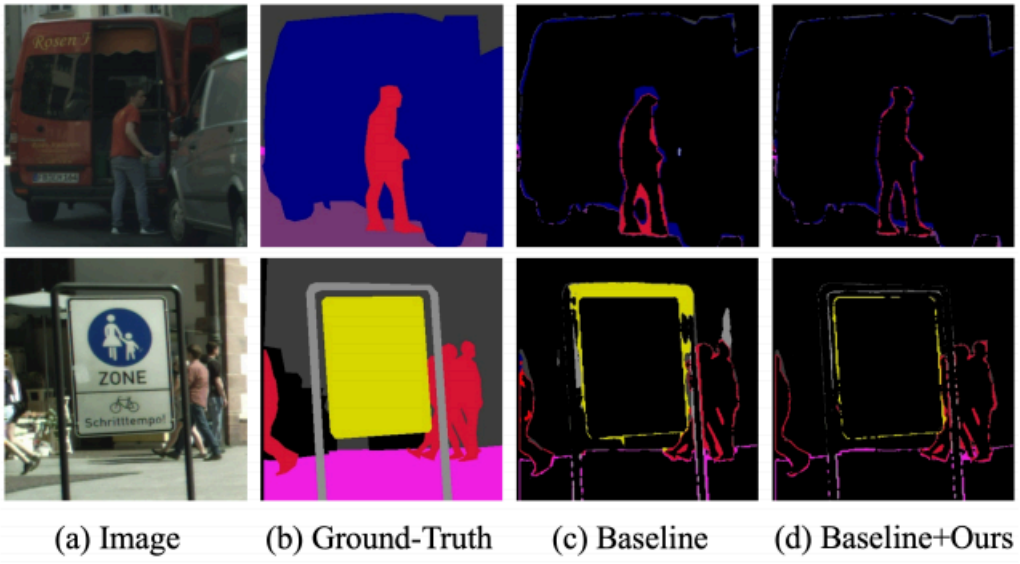
NeurLPS 2022

精确的边界区域的预测对于语义分割十分重要，这篇论文针对“普通卷积难以利用深度模型去生成精确的边界预测”的问题，提出了一种语义扩散网络去近似扩散过程，包含了一个特征融合网络以及参数化语义微分卷积操作。并且可以插入现有的“Encoder-Decoder”分割网络。

Motivation

对于如何提升分割结果的边界质量，现有（2022）工作方式：

- 使用一个后处理模块去细化边界。
- 共同学习边界感知的边缘检测和语义分割任务。
- 通过设计一个边界感知损失使训练对边界变化保持高灵敏性。
- figure-1

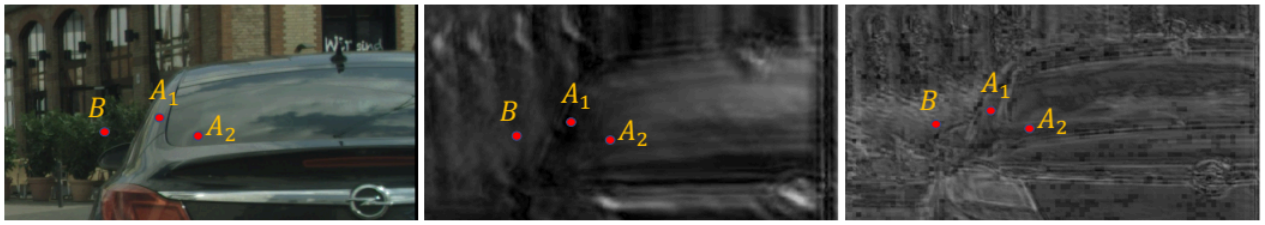


【Figure1：彩色区域为误差】

> NOTE：
>
> 1.为什么卷积擅长处理平滑边缘信息？
>
> 卷积操作本质上是一个**局部加权平均**的过程，对于一个卷积核，它在输入特征图上的每个局部区域进行滑动，计算该区域内的加权和。如果卷积核的权重是正的且平滑（如高斯分布），那么卷积操作相当于一个**低通滤波器**，会抑制高频信息（如噪声、锐利边缘），同时保留或增强低频信息。
>
> 2.为什么卷积不适合语义分割中的边界预测？
>
> 语义分割是一个像素级分类任务，要求模型对每个像素给出准确的语义标签。在这个过程中，类与类之间的边界（inter-class boundaries）十分重要。
>
> 问题在于：

> - **卷积的平滑性会模糊细节**尤其在深层网络中，多次卷积操作会使得特征中的高频信息（如物体边界）逐渐丢失。
> - **边界区域本身是高频信息**，语义边界通常是突变区域，而卷积的低通滤波性会使得其变得模糊。
> - **无法区分语义边界和纹理边缘**，传统的卷积或梯度算子（如：Sobel算子）会对所有边缘包括物体内部的纹理都做出相应，而语义分割只关心类间边界。

论文中给了一幅图如下所示，卷积对三种不同的语义区域（车体A1、车窗A2、植物B）的特征响应都非常相似。



(a) Image (b) Vanilla Conv (c) CDC

由以上信息我们可以得知这篇文章的motivation：**改善类间边界(Inter-class boundary)** from operator level。

Method – 传统的解决方式

核心：使用改进的 anisotropic以语义为导向的扩散过程表示类间边界增强。

【原文】*we propose a learnable semantic diffusion network (SDN) to approximate the diffusion process, which parameterizes the traditional solver and only requires only one forward instead of multiple iterations.*

传统的迭代求解器（multiple iterations solver），比如有限差分法的性能严重依赖于时间步长、迭代次数、扩散系数等超参数的设置。深度学习模型通常通过梯度下降自动学习参数，如果其中的模块对超参数极其敏感，会使得整个模型的训练变得非常困难，并且会导致结果发散（NaN或者梯度爆炸）。这篇文章不打算严格地求解复杂的微分方程，而是通过一个可学习的神经网络去近似这个扩散过程以达到最终的效果。在说明具体改进之前，先梳理一下传统的方法。

扩散过程公式解析

$$\frac{\partial U}{\partial t} = \mathrm{Div}(D \cdot \nabla U)$$

t: time step

Image Tensor

Calculate the rate and direction of change of U at each position

Divergence

D: 扩散系数

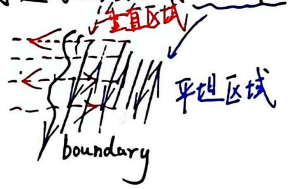
If: D = 1，所有位置所有方向上都以相同速率平滑。相当于一个高斯模糊。

D = D(x)，D(x)是函数，但与U无关，虽然在不同位置有不同速率，但它无法智能识别边缘。

D = D(U)，设计这样一个D：在边缘处（∇U大）取值小（抑制扩散）保留

在平坦处（∇U小）取值大（加速扩散）平滑

△ Anisotropic Non Linear → "各向异性"：使其不仅依赖 ∇U 值的大小，还可以依赖梯度方向（中译）平行于边缘（沿边缘）较多平滑，垂直于边缘较少平滑。



峰值所在

Formulation

该部分则要从上述的“扩散过程”，准确来说是各向异性的扩散过程通过数学形式化，适配到深度语义分割的语境当中去。

$$\frac{\partial U_t}{\partial t} = \mathrm{Div}(g(|\nabla V|^2)\nabla U_t) \quad [\text{非线性}]$$

扩散对象：U_t 主干网络第 l 层，输出的特征图 X^l

目标：增强它的边缘

“有限差分法”

语义引导 V：它是与 U_t 分辨率相同的引导特征图

$g(|\nabla V|^2)$，当语义特征 V 在空间上剧烈变化时，$|\nabla V|^2$ 的值会变大，若要在边界时抑制扩散，保持边界特征，则 $g(|\nabla V|^2)$ 应该小，故其必须是个减函数。

“离散化”

$$\tilde{U}^{t+1}_{(m,n)} = \sum_{i,j \in N(m,n)} g(|V_{i,j} - V_{(m,n)}|^2) \cdot (U^t_{i,j} - U^t_{m,n})$$

用两点之间的差异来近似梯度

$N(m,n)$ 表示 (m,n) 的邻域

当前位置

$$U^{t+1}_{(m,n)} = \alpha^t U^t_{(m,n)} + \beta^t \tilde{U}^{t+1}_{(m,n)}$$

上一刻时刻的值

计算出的更新量

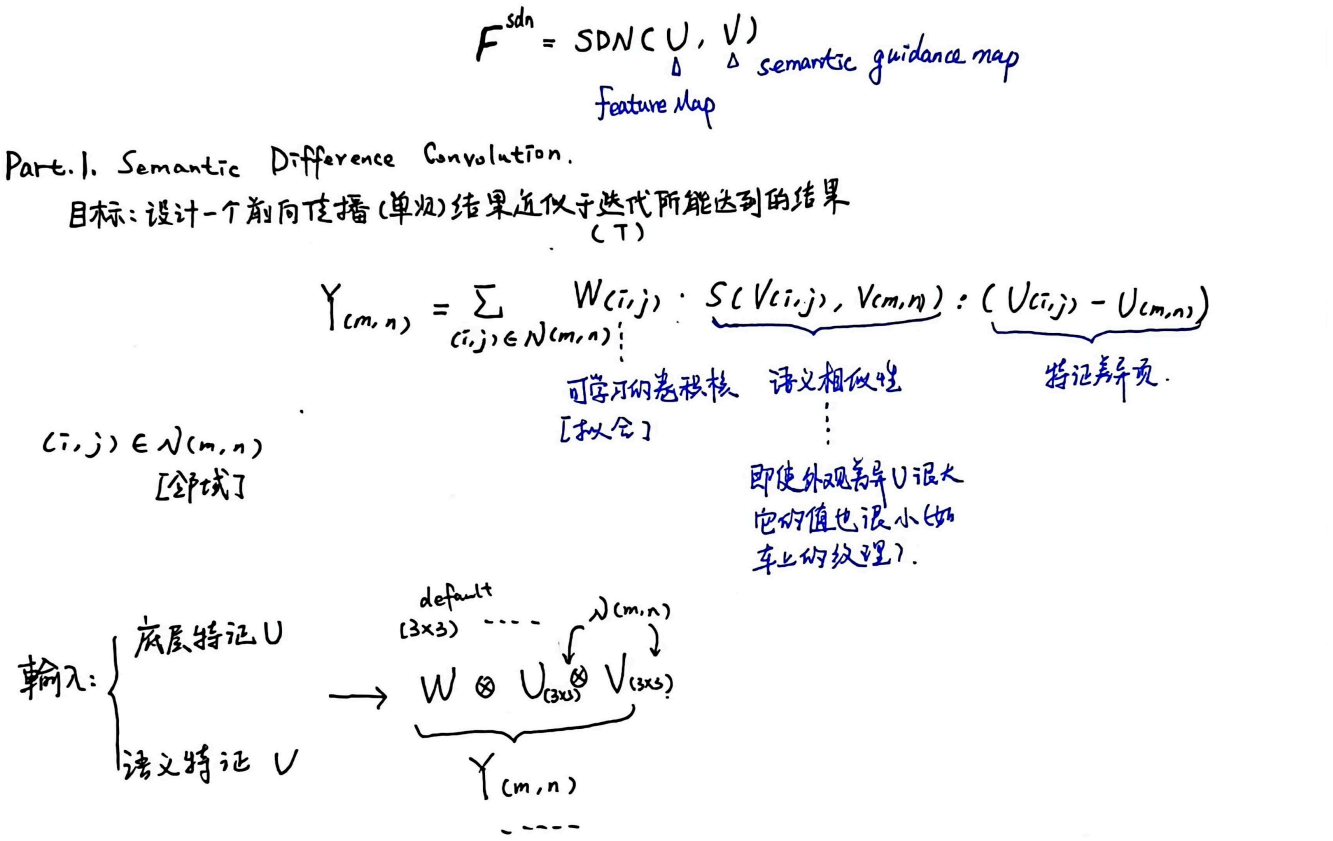
加权高斯合

但是呢，这个公式难以嵌入到深度模型中去，这种有限差分法的稳定性依赖于边界条件的设置和参数的选择。这种方法需要多次迭代（对于t），计算量复杂且容易导致梯度爆炸。（详细分析见本节第一段）

Method - Semantic Diffusion Network

a novel learnable approach called semantic diffusion network (SDN) for approximating the diffusion process, which contains a parameterized semantic difference convolution operator followed by a feature fusion module and constructs a differentiable mapping from original backbone features to advanced boundary-aware features.

Semantic Difference Convolation



上图可见，语义相似性控制类内细节不会被差异特征（纹理）影响而被抑制。而SDC的输出是一个Y，这个Y可以理解为是一个边界响应图，它突出了语义边界的位置，主要包含了**变化的梯度信息**（上述公式的语义相似性和特征差异性都反应的是梯度变化），它可能会丢失一些静态的、全局的语义信息。那么这个Y就不能是直接的输出，因为会缺少上线文信息，需要一个融合模块将。**将SDC提取的边界增强信息Y与原始特征信U息地融合起来**。这个特征模块就是F^sdn.

F的具体工作方式：

- 现将原始特征信息U与SDC输出的边界响应特征Y在通道上拼接起来。 即 [U, Y]
- 使用一个1x1的卷积融合所有的通道信息，最后输出符合期望的维度（一般与U的通道数相同）。即Conv([U, Y])
- 当U和Y的尺寸不匹配是，可以采用双线性差值上采样的形式使其对齐，确保能进行拼接。

使用SDN的分割模型

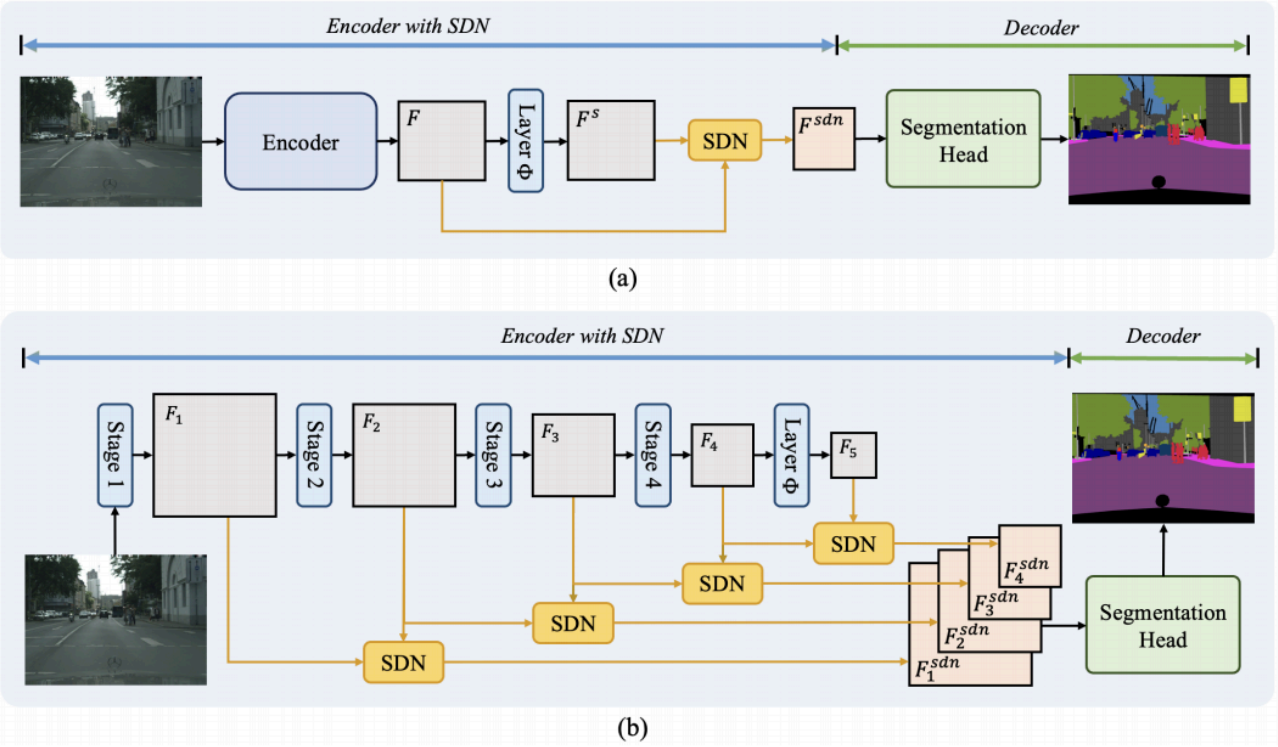


Figure 4: Illustration of how to combine the proposed semantic diffusion network (SDN) with the baseline segmentation model with single-scale decoder and multi-scale decoder in (a) and (b), respectively. The SDN is regarded as a neck part in order not to break the original encoder and decoder design of the baseline model.

上图是两种应用的分割模型。

单尺度解码器

适用模型： 编码器输出单一尺度的特征图，解码器基于该特征图进行预测。

典型代表：

- `ViT + Segmenter`：将ViT作为编码器输出一个低分辨率的特征图。
- `ResNet + FCN`：使用ResNet的最后一层输出（下采样32倍特征图）。

  对于单尺度模型，**主动创建一个更具语义抽象的特征层**来指导边界增强。如4a所示，编码器处理输入图像Image得到特征图F。在编码器输出F之后叠加一个轻量地3X3卷积（stride=2）的层，会继续对F下采样，使得Fs的感受野更大，语义更丰富，作为语义引导特征。将F作为特征输入U，将Fs作为语义引导V，送入SDN处理获得Fsdn = SDN(F, Fs)。将增强后的Fdns（与F尺寸相同）送入解码器，生成最终的分割图。

多尺度解码器

适用模型： 编码器输出**多个尺度**的特征图（通常是金字塔pyramid结构），解码器会融合这些多尺度特征进行预测。

典型代表：

- `ResNet + SegmanticFPN`：利用ResNet的四个阶段输出多尺度特征。
- `HRNet`

对于多尺度模型，**利用编码器天然的特征金字塔**，用“老师”（深层、高语义特征）来指导“学生”（浅层、高分辨率特征）进行边界增强，实现高效且全面的特征优化。如4b所示，编码器处理图像，输出L个尺度的特征图，为每个尺度都配备一个SDN模块，实现全尺度范围的边界增强。Fi作为特征信息图进行输入，Fi+1作为语义指导输入SDN并生成Fsdn。

但是注意由于最后一层输出F4(也就是F_L)是最深层的输出，仿照单尺度解码器对最后一层使用一个3X3卷积来处理F_L。最后将所有的Fsdn送入解码器进行融合。

近年来单尺度编码器还有：

- MaskFormer/Mask2Former(2021, 2022)：将语义分割重新定义为**掩码分类问题**，预测一组二进制掩码和对应的类别标签。基于Transformer解码器，输入一组可学习的查询（query），通过交叉注意力与编码器特征交互，输出固定数量的掩码预测。虽然涉及多尺度特征，但最终是基于**统一的特征表示**生成预测，可视为高级的单尺度解码。
- kMaX-DeepLab (2023)：基于掩码分类的SOTA模型，使用K-means掩码解码器，将语义分割视为聚类过程，通过迭代更新聚类中心（即掩码原型）和像素-聚类中心关联来生成掩码。解码过程基于统一的编码器输出特征进行。

  多尺度编码器：
- HRNet(2019)及其变体： BaseLine
- PIDNet(2023):- 受PID控制器启发，设计三分支网络（P、I、D分支）分别负责细节、上下文和边界信息。在解码阶段，三个分支的特征被的多路径融合模块（如Lightweight Bagged Fusion）整合，充分融合不同尺度的信息，显著提升实时分割的精度。
- DPT（Dense Prediction Transformer, 2021）编码器使用ViT，但是使用UNet类似的渐进式融合解码器，将ViT输出的不同阶段的特征图（不同层的[CLS] token对应的特征）重新组合乘图像金字塔结构，然后通过多层卷积核上采样操作逐步佛那个号不同尺度的特征。

实验部分不在此赘述，请详见论文：点击获取论文