中国科学: 信息科学 2021 年 第 51 卷 第 7 期: 1084–1099
SCIENTIA SINICA Informationis
论文

互联网图像驱动的语义分割自主学习

侯淇彬[1,2], 韩凌昊[1], 刘姜江[1], 程明明[1*]

1. 南开大学计算机学院, 天津300350, 中国
2. College of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, Singapore
* 通信作者. E-mail: cmm@nankai.edu.cn

收稿日期: 2020–05–22; 修回日期: 2020–07–30; 接受日期: 2020–08–24; 网络出版日期: 2021–06–30

"新一代人工智能"重大项目(批准号: 2018AAA0100400)、国家自然科学基金优秀青年科学基金项目(批准号: 61922046)、教育部指导高校科技创新规划项目和南开大学中央高校基本科研业务费专项资金(批准号: 63201169)资助

摘要 针对目标任务收集新类别的海量标注数据通常需要大量时间和人力成本, 并已成为语义分割技术投入实际产业应用过程的主要瓶颈. 本文旨在以"网络监督"的方式, 在仅利用用户提供的目标类别关键词以及相应自动搜索到的网络数据的条件下实现语义分割模型的自主学习. 该任务的核心挑战在于网络爬取的图像中存在一定量的类别噪声, 从而影响自主学习的可靠性. 为了解决类别噪声问题, 本文设计了一种新颖的噪声擦除模型. 该模型通过每次从小批次样本的置信注意力区域中以跨样本的方式学习语义信息来擦除训练图像中与搜索关键词无关的区域. 基于该模型, 本文同时提出了一种能够用于训练语义分割模型的高质量伪标注生成方法. 在国际主流的公开数据集(PASCAL VOC 2012)上的大量实验表明, 基于该方法的语义分割模型在利用网络监督与弱监督的条件下均取得了良好结果(mIoU = 62.0%以及66.1%).

关键词 语义分割, 网络搜索, 类别噪声, 噪声擦除网络, 网络监督

1 引言

近年来, 基于深度学习的语义分割方法[1~3] 在各种标准数据集(如: PASCAL VOC 2012 [4]) 上已取得重大进展. 然而, 收集用于训练这些模型所需的像素级标注不仅难度高, 而且需要耗费大量时间, 极大地限制了这些技术应用于产业实践的进程. 基于弱监督的语义分割方法则缓解了对于像素级标注数据的依赖. 这些方法通常利用物体的边界框[5,6]、简单标注[7]、点集[8] 甚至关键词[9~12] 等粗略标注来取代精准的像素级标注.

在实际应用中, 我们经常需要处理一些缺少标注数据的新类别, 但仅标注出样本中图像级关键字信息也大概需要20 s左右时间来处理一个图像样本[8,13]. 对于含有数万张乃至数十万张样本图片的数

引用格式: 侯淇彬, 韩凌昊, 刘姜江, 等. 互联网图像驱动的语义分割自主学习. 中国科学: 信息科学, 2021, 51: 1084–1099, doi: 10.1360/SSI-2020-0146
Hou Q B, Han L-H, Liu J-J, et al. Autonomous learning of semantic segmentation from Internet images (in Chinese). Sci Sin Inform, 2021, 51: 1084–1099, doi: 10.1360/SSI-2020-0146

图 1 (网络版彩图) 根据关键词从互联网中搜索得到的一组图片. 蓝色字体为用于搜索的关键字, 而红色字体为需要消除的类别噪声.需要注意的是, 在训练阶段, 模型只能获取到蓝色字体对应的关键字

Figure 1 (Color online) A group of images from web searches. Words in blue are keywords used for retrieving while red ones are label noise that needs to be solved. Note that we only know keywords in blue during training

据集, 即使仅标注出关键字级别的监督信息也需耗费大量的时间成本和人工成本. 在永续学习[14]的背景下, 为成千上万的新类别标注数以亿计的样本数据则需消耗更多人力.

针对上述情况, 本文提出一个新的研究问题, 即如何在不依赖于标注数据的情况下使得机器视觉系统能够智能地从互联网搜索得到的数据中完成语义分割模型的训练? 该问题虽然与基于网络数据的弱监督语义分割[12,13,15] 有交集, 但仍存在以下两处不同:

• 基于互联网数据的弱监督语义分割方法通常需要利用大量人工标注数据(如: PASCAL VOC 2012数据集[4]中图像的关键词信息) 来训练初始的弱监督模型, 并用训练后的模型帮助过滤掉存在错误标注的网络图像. 这一缺点阻碍了这些方法在处理无标注数据的新类别时的应用. 这也同样反映出本文提出的网络监督任务相比网络监督下的弱监督语义分割更具挑战性.

• 现有弱监督语义分割方法主要利用网络图像中较为容易处理的简单图像(如: 仅包含简单背景、单语义类别, 或尺寸适中的目标物体), 来增强学习算法的稳定性. 但是, 这些方法忽视了对理解复杂场景更有意义的困难样本的使用.

从互联网获取的图像数据存在诸多不可靠的关键词信息, 使其成为自动学习语义分割任务的主要挑战. 简而言之, 网络图像中的类别噪声主要有以下两种形式: (1) 从互联网中爬取的图像可能存在错误的类别信息; (2) 爬取的图像中可能存在除搜索关键词外的其他类别物体. 现有的弱监督语义分割方法[9,11~13,15,16] 通常使用精准的图像级别关键词信息以及具有可靠标注的简单图像来避免第1种问题的发生. 然而, 如图1 中最后3张图所示, 当图像中含有复杂的语义场景时, 上述方法的有效性会被严重削弱.

为解决上述问题, 本文提出一个新的研究视角: 通过消除网络图像中的噪声区域来获取较为精准的伪标注数据. 给定一张网络图像, 本方法意在去除图像中存在的噪声区域而非忽略掉存在类别噪声的复杂图像.为此, 本文提出了一种新颖的噪声擦除网络(noise erasing network, NENet). 该模型能够从每个小批次图像样本中学习跨样本之间的语义信息. 具体而言, 给定一小批次图像样本, 通过利用注意力模型生成的注意力图与过分割算法生成的分割图, NENet以聚类的方式从同一关键词对应的分割区域中筛选出置信度较高的区域并将置信度较低的区域标注为噪声区域.实验结果表明, 在复杂的网络图像中去除噪声关键词对应的区域可以生成更为精准的伪标注. 因此, 本方法在主流的数据集上取得了良好效果.

本文的主要学术贡献可概括如下:

• 提出一个新颖的视觉问题, 即如何使得机器视觉模型在仅给定目标类别关键词的情况下, 自主获取互联网图像并从这些含有类别噪声的网络图像中学习语义信息;

• 设计了一个噪声擦除模型. 该模型能够以跨样本的方式学习到网络图像中丰富的语义信息并去

除掉噪声区域进而生成较为精准的伪标注数据;

• 在国际主流的公开数据集(PASCAL VOC 2012 [4]) 上取得了良好效果. 在仅使用网络图像时, 本方法mIoU值达到62.0%, 而当加入PASCAL VOC 2012训练集中的图像级关键字弱标注信息后, 本方法mIoU值可以达到66.1%.

2 相关研究现状

近年来, 随着研究人员对语义分割研究的大量投入[1, 2, 17~20], 弱监督语义分割技术也得到快速发展. 早期弱监督方法[5, 21] 多数依赖于最大期望算法[22] 或多实例学习模型来解决仅有图像类别信息的弱监督语义分割问题. 研究人员也通过引入一系列约束或候选物体外边框来更准确地定位语义目标[6, 23].

由于注意力模型具有发现类别敏感区域的能力[24], 基于注意力机制的弱监督语义分割方法[25~31]逐渐成为主流研究方向.这些方法在设计上采用了不同损失函数或关联网络来扩大注意力模型生成的小尺寸高精度的种子区域. 不同于上述方法, Wei等[11]提出了一种基于对抗擦除模型.通过擦除输入图像中由注意力模型检测到的语义区域并将输入图像剩余部分继续送入注意力模型中来挖掘更多语义区域. Li等[32]和Zhang等[33]分别针对对抗擦除策略做了进一步改进, 并实现了端到端的学习方式, 大大简化了训练流程.

随着低级别视觉任务相关研究的发展, 研究者们逐渐开始侧重于从显著性物体检测与边缘提取模型中获取类别无关的物体信息, 并将其应用于弱监督语义分割任务中. 显著性物体分割模型[34~37] 可以有效地从输入图像中分割出类别无关的前景物体并且分割结果具有较好的几何信息. 例如, Wei等[10]提出了一种由简单到复杂的渐进式处理框架.该框架使用显著性物体检测算法[38] 生成的显著性图作为初始的种子区域并将得到的种子区域结合图片类别标签来训练语义分割模型. 其他相关工作[9, 16, 27, 39] 则将显著性图与注意力图相结合, 来生成用于训练语义分割模型的伪标注数据. 除此之外, Wei等[11, 31]、Li等[32]和Hou等[40] 也提出将显著性物体检测模型得到的背景区域作为先验知识, 来改善注意力模型难以精确定位类别可判别区域的不足.

利用网络图像或视频数据进行语义分割亦为一热门研究方向. 作为该方向的一次早期尝试, Pinheiro等[21] 提出利用约70万张网络图像并结合多实例学习方法来训练分割模型. 在其他方法[12, 13, 15]中, 网络图像也被用作额外的训练样本来提高语义分割模型的分割精度.然而, 这些方法除了利用网络数据外, 仍使用了人工标注的数据信息来帮助生成伪标注数据, 大大限制了其处理新类别物体的能力.

3 问题定义

从互联网(如Flickr网站)下载图像是获取训练数据最为简单且低成本的方法之一. 给定一组关键词, 用户可以轻松获取大量与关键词相关的图像但无精准关键词或像素级别可靠标注. 将这些网络图像与检索词以及低级视觉特征相结合则可得到用于训练语义分割网络的伪标注数据. 与现有弱监督语义分割任务(严重依赖于精准像素级标注数据) 不同, 本研究提出问题的关键挑战在于类别噪声的干扰[41], 如图1 最后3张图所示.

令$\mathcal{I} = \{I_i\}_{i=1}^{N}$表示从网络获取中的数据集. 每张图像对应一个来自预定义检索类别集合$\mathcal{L} = \{l_0, l_1, l_2, \ldots, l_L\}$ 中的图像级关键词y_i. 其中, $L = |\mathcal{L}|$为语义类别的种类数, 而l_0为背景类别. 令$\hat{\boldsymbol{y}}_i$为图像I_i的真实类别, 如上所述, 图像I_i对应的检索类别y_i 并非总是与其真实类别$\hat{\boldsymbol{y}}_i$等价. 本文将这种与

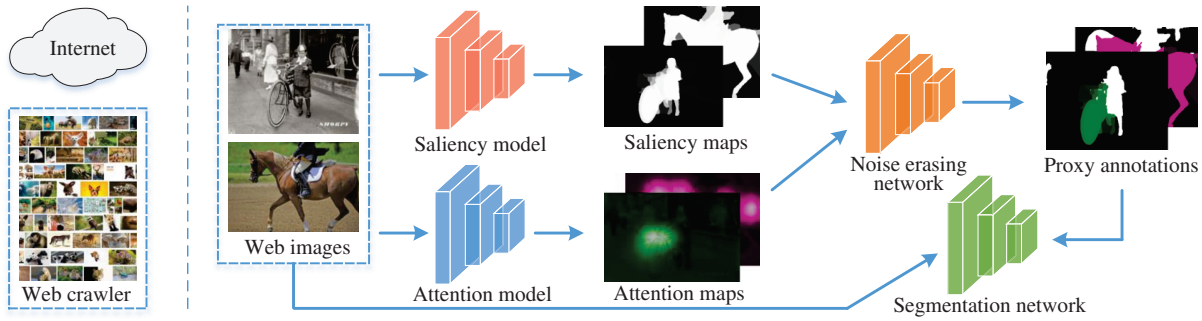


图 2 (网络版彩图) 本方法流程示意图.给定从互联网获取的图像, 本方法首先采用注意力模型为每张图像生成与其检索词相关的注意力图. 然后, 采用噪声擦除模型学习注意力图内判别区域的语义信息从而消除显著性图中的噪声区域.最后, 噪声擦除模型的输出将被作为用于训练语义分割模型的伪标注.注: 本方法中显著性模型、注意力模型、噪声擦除模型, 以及分割模型为4个独立的部分, 每个模型中的所有参数在训练时皆参与反向传播.

Figure 2 (Color online) The pipeline of our approach. Given the retrieved web images, our attention network firstly outputs attention maps associated with the keyword for each image. Then, our noise erasing network erases regions that are predicted as noise from the saliency maps by learning semantic knowledge from the discriminative regions in attention maps. We use the results outputted by our noise erasing network as proxy annotations for training semantic segmentation networks. It is worth mentioning that the saliency model, attention model, noise erasing network, and the segmentation model are four independent parts, so during training all parameters in each model participate in back-propagation.

图像真实类别y_i不一致的检索关键词$\hat{\boldsymbol{y}}_i$称为"类别噪声". 本研究的关键任务在于如何仅用$\{I_i\}$和$\{y_i\}$来自动训练语义分割模型.

4 基于互联网图像中噪声擦除的语义分割自主学习方法

如何消除类别噪声对生成高精度的伪标注数据以及训练高质量的语义分割模型至关重要. 为解决该问题, 本部分提出一个简单且有效的噪声擦除方法. 该方法可以有效去除图片中的部分类别噪声, 从而生成质量较高的伪标注数据.

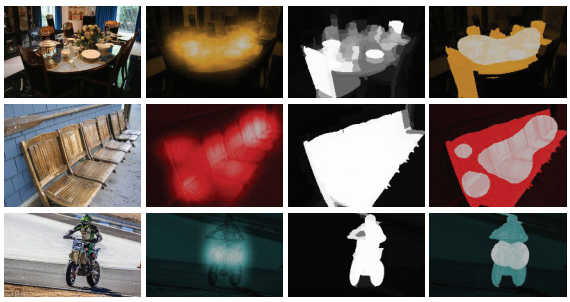
4.1 方法流程

本方法的工作流程(如图2 所示)主要包含3个模块: 低级视觉特征提取模块(包括显著性物体检测与注意力模型)、噪声擦除模型和语义分割模型.

方法概述. 虽然注意力模型能够发掘图像中与关键词类别相关的判别区域, 但其生成的区域通常具有不规则的形状, 不利于直接用来训练语义分割模型.受最近工作[10, 16, 40]启发, 本方法结合显著性图[42]1) 与注意力图生成伪标注数据来训练语义分割模型. 但由于类别噪声的存在, 显著性物体检测模型生成的显著性区域通常包含一些背景或与关键词无关的语义物体(如图3所示). 注意力模型的判别能力可以适当解决该问题. 由图3可知, 尽管注意力模型捕捉到的判别区域较小但精度较高. 受该现象启发, 本研究提出一种从判别区域学习跨样本语义信息的方法. 该方法可去除类别噪声的同时保留更多语义区域并将保留的语义区域作为伪标注数据用于训练语义分割模型.

判别区域挖掘. 由于注意力模型具有挖掘图像中与关键词类别相关的判别区域的能力, 近年来其在弱监督语义分割领域得到广泛应用. 如许多方法[11, 25]类似, 本研究使用类别激活图(class activation

1) 显著性图能够提供类别无关的前景物体的形状信息. 本文采用公开的显著性物体检测模型[42].



(a) (b) (c) (d)

图 3 (网络版彩图) (a) 原图; (b) 注意力图; (c) 显著性图; (d) 本方法处理后的训练样本标注. 本方法首先将显著性图二值化从而生成背景区域((d)中的黑色区域)与前景区域((d)中非黑色区域). 随后根据式(1)来分割前景区域并生成置信区域((d)中白色区域)与潜在区域((d)中彩色区域). 本方法从置信区域中区域提取语义信息用来进行噪声擦除网络的训练并对潜在区域的类别进行类别预测

Figure 3 (Color online) (a) Source images; (b) attention maps; (c) saliency maps; (d) illustration of training sample selection. We first binarize the saliency maps, yielding the background zone (black area in (d)) and the foreground zone (non-black area in (d)). Then we separate the foreground zone according to Eq. (1), yielding the credible zone (white areas in (d)) and potential zone (colorful areas in (d)). We use regions inside credible zones for training our noise erasing network

map, CAM) [24]来获取高质量的判别区域.

4.2 噪声区域擦除

如上所述, 显著性区域虽然包含丰富的前景信息, 但仍存在与关键词不相关的噪声区域. 如图3(c)所示, 直接采用显著性图来训练语义分割模型将会在训练中引入错误标注信息. 为了消除前景区域中的噪声, 本小节提出噪声擦除网络.

给定一幅图像I、取值自$\mathcal{L}$的关键词y 以及与其对应的显著性图S, 本方法将图像选用一固定阈值(0.2) 将S划分为两部分: 背景区域与前景区域. 令R为过分割方法[43] 对图像I生成的分割图.本方法的目的在于让噪声擦除模型从注意力模型生成的判别区域中学习语义信息并将其与关键词相关的前景区域进行聚类从而分离出与噪声相关的区域. 与聚类中心距离较远(即预测结果与检索词不同)的前景区域将会被从前景区域中移除, 进而达到净化伪标注的目的.

4.2.1 训练样本选取

令A_y为图像I对应的注意力图. 通过计算显著性图S与注意力图A_y的调和平均数B, 可将前景划分为两个部分:

$$B = \frac{2}{1/S + 1/A_y}. \tag{1}$$

式(1)使得强度较高的注意力区域得以保留并同时抑制可能为噪声的余下部分. 图4详细展示了在一个具有多幅图像的小批次样本中的训练样本的选择. 本方法将B中保留下来的区域作为置信区域(深色图块), 并将其余前景作为潜在区域(浅色图块). 潜在区域中可能既包含正确关键词对应的区域(无红叉标记), 亦包含类别噪声对应的区域(有红叉标记). 判别区域超出显著区域的部分将被视为背景. 当给定图像对应的显著区域与注意力区域无重叠时, 该图像将不参与训练.

根据上述定义, 本方法将分割图R分割成3个不相交的子集: R_B, R_C与R_P, 分别代表背景区域、置信区域与潜在区域. 其中前景区域由R_C与R_P组成, 即$R_F = R_C \cup R_P$. R_C中的区域将被用于训练噪声擦除网络. 在测试阶段, 模型将对R_F中的区域进行类别预测.

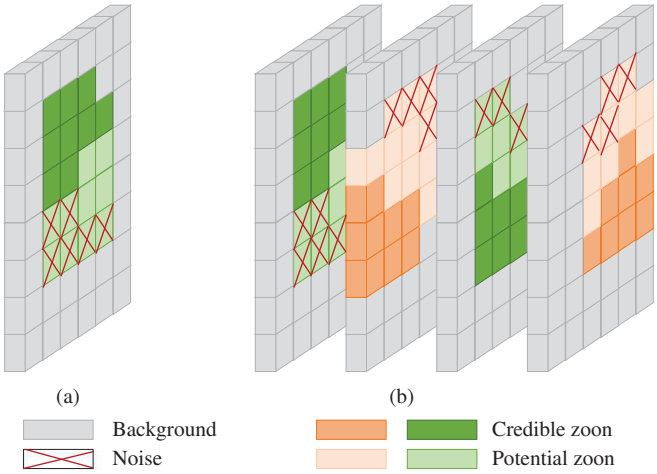


图 4 (网络版彩图) 小批次样本的训练样本选择. 其中, 每个平面代表一个小批次样本中一张图像且图块代表过分割后的区域. 本方法从置信区域(深色图块)中学习信息并对潜在区域(浅色图块)进行推断. 主要目的在于去除可能为类别噪声或背景的区域(红叉标记)并保留无噪声的目标区域(浅色无红叉标记图块). 不同颜色图块表示不同类别

Figure 4 (Color online) Training sample selection in a mini-batch. Each panel represents an image in the mini-batch. We use blocks to represent regions after over-segmentation. We learn knowledge from credible zones (blocks in the deep color) and perform inference on potential regions (blocks in the light color). Our goal is to erase the blocks with red crosses probably belonging to label noise or background while preserving the rest clean ones (blocks in the light color but without red crosses over them). Different colors denote different categories

4.2.2 噪声擦除模型(NENet)

本文提出的噪声擦除模型由一个骨干网络及一个区域池化层组成. 其中骨干网络为全卷积的分类模型. 区域池化层可为每一过分割区域生成一编码向量$\boldsymbol{v}$. 令$\boldsymbol{f} \in \mathbb{R}^{K\times H\times W}$ 为骨干网络生成的K通道特征图.其中, H与W分别为特征图的高与宽.对于某一区域R_i, 其编码向量$\boldsymbol{v}_i$ 的计算方法如下:

$$\boldsymbol{v}_i = \frac{1}{|R_i|}\sum_{(h,w)\in R_i} \boldsymbol{f}_{h,w}, \tag{2}$$

其中$\boldsymbol{f}_{h,w}$为位置(h,w)处沿通道维度长度为K的特征向量.

给定所有区域的编码向量, 本方法采用两种损失计算方式来优化噪声擦除模型.第1种为$|\mathcal{L}|$类别的分类损失. 第2种主要以跨样本的方式学习小批次样本中的语义信息, 使得同一类别中的编码向量距离趋近于0, 而不同类别的编码向量距离尽量变大.

损失函数. 损失函数的第1部分为$|\mathcal{L}|$类别的分类损失. 令M为训练过程每个小批次中的样本数量, $R_{m,C}$为当前小批次样本中第m张图像I_m中置信区域的集合. 则第m张图像中第i个置信区域$R_{m,i} \in R_{m,C}$ 属于关键词y_m的概率为

$$p^y_{m,i} = \frac{\exp(\boldsymbol{w}_{y_m}\cdot \boldsymbol{v}_{m,i})}{\sum_{l\in\mathcal{L}}\exp(\boldsymbol{w}_l\cdot \boldsymbol{v}_{m,i})}, \tag{3}$$

其中$\boldsymbol{w}_{y_m}$和$\boldsymbol{w}_l$为可学习参数, $\boldsymbol{v}_{m,i}$为第m张图像中第i个置信区域的编码向量, $\cdot$为向量点积操作. 此时, 当前小批次样本的损失$L_{\rm cls}$可表示为

$$L_{\rm cls} = -\sum_{m=1}^{M}\sum_{R_{m,i}\in R_{m,C}} \log(p^y_{m,i}). \tag{4}$$

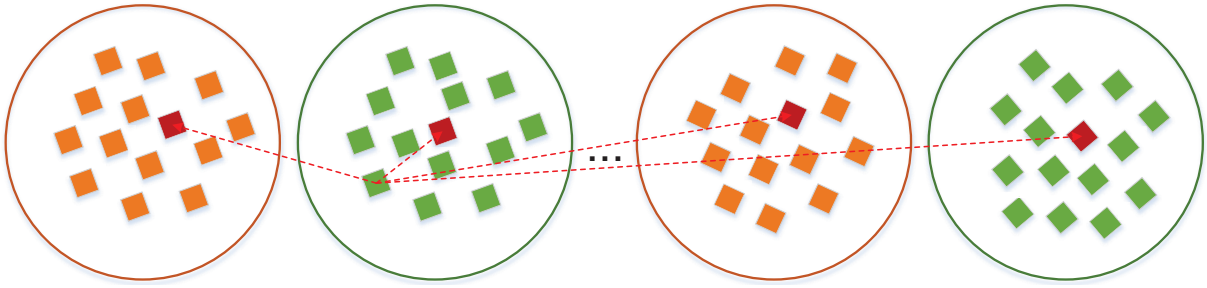


图 5 (网络版彩图) 噪声擦除模型工作原理图.在训练噪声擦除模型时, 本方法仅对置信区域(彩色圆形)内的分割块(彩色方块)提取编码向量. 在每个小批量样本中, 通过计算每一编码向量与所有编码中心的(深红色图块)距离来惩罚类别不同的编码向量对

Figure 5 (Color online) Illustration of how noise erasing network works. Only regions (colorful blocks) inside credible zones (colorful circles) are used to extract embedding vectors for training our noise erasing network. In each mini-batch, we compute the distances between each embedding vector and all embedding centers (blocks in deep red) and punish the embedding vector pairs belonging to different categories

为了使噪声擦除模型具备更强的判别力, 本文同时引入相似度损失L_{simi}来扩大属于不同类别编码向量之间的距离. 当前小批次样本中第n张图像的编码中心可表示为

$$\bar{\boldsymbol{v}}_n = \frac{1}{|R_{n,C}|} \sum_{R_{n,i} \in R_{n,C}} \boldsymbol{v}_{n,i}, \tag{5}$$

其中, $|\cdot|$表示区域的数量. 由于注意力模型生成的注意力区域并不完全精准, 因此置信区域中也可能包含不属于关键词的区域. 为了增强模型的鲁棒性与效率, 本方法将置信区域中每一个区域的编码向量与小批次样本中每张图像的编码中心进行比较. 图5 详细展示了该过程.

因此, 一个小批次样本中的相似度损失L_{simi}可表示为

$$L_{\text{simi}} = -\sum_{m=1}^{M} \sum_{R_{m,i} \in R_{m,C}} \sum_{n=1}^{M} \mathbb{I}_{\{y_m = y_n\}} \log(d(\boldsymbol{v}_{m,i}, \bar{\boldsymbol{v}}_n)) + \mathbb{I}_{\{y_m \neq y_n\}} \log(1 - d(\boldsymbol{v}_{m,i}, \bar{\boldsymbol{v}}_n)), \tag{6}$$

其中, $\mathbb{I}$为指示函数, $d(\cdot,\cdot)$为两个编码向量间的相似度, 可由以下方法计算:

$$d(\boldsymbol{v}_{m,i}, \bar{\boldsymbol{v}}_n) = \exp(-||\boldsymbol{v}_{m,i} - \bar{\boldsymbol{v}}_n||_2^2). \tag{7}$$

通过这种方式, 属于不同类别的跨样本编码向量对亦会被惩罚, 从而使得噪声擦除模型能够更精准地判断给定区域的类别. 因此, 噪声擦除模型的整体损失函数为$L = L_{\text{cls}} + \lambda L_{\text{simi}}$, 其中$\lambda$用来控制$L_S$的权重.本文将其设置为0.2并使用随机梯度下降法来优化整个模型.

测试阶段. 在测试阶段中, 本方法使用噪声擦除模型来预测前景区域内所有子区域的类别. 给定一张含有关键词y的图像I及其对应的分割图R, 首先依据式(3) 计算所有前景区域属于各类别的概率. 然后使用算法1 生成伪标注用以训练分割模型. 若前景中某区域R_i的预测类别与y不同, 则在伪标注G中将R_i对应的所有像素指定为一个特殊类别l_s.该类别表示在训练过程中这些区域将被忽略, 也即擦除的过程. 图6(d)中展示一些可视化结果.

4.3 语义分割模型

与显著性图类似, 本方法生成的伪标注为数值连续的灰度图像. 除被擦除的区域外, 每个像素对应的显著性值代表其属于前景区域的概率. 因此, 在训练语义分割模型时, 本文采用如下的交叉熵损

Algorithm 1 Proxy annotation generation

Input: Image I; keyword y; region map R; probability $\boldsymbol{p}$; saliency map S.

Output: Proxy annotation G.

for $R_i \in R_F$ **do**
  $C_m \leftarrow \text{argmax}_{l\in\mathcal{L}} p_i^l$;
  if $C_m \neq y$ **then**
    $G_j \leftarrow l_s, \ \ \forall j \in R_i$; ⇐ Erase regions irrelevant to the keyword
    Continue;
  end if
  for $j \in R_i$ **do**
    $G_j \leftarrow S_j$; ⇐ Keep correct predictions
  end for
end for
Output G.

图 6 (网络版彩图) 噪声擦除模型生成的伪标注示意图. 在伪标注(d)中, 彩色区域及黑色背景被用来训练语义分割模型. 白色区域为被噪声擦除模型擦除掉的区域. 本方法使用显著图的背景区域作为伪标注的背景区域

Figure 6 (Color online) Results produced by our noise erasing network. (a) Source image; (b) attention map; (c) saliency map; (d) proxy GT. In column (d), we colorize regions that are kept unchanged and whiten regions that will be ignored during the training of the segmentation network. We use the background of saliency map as the background of our proxy ground truth as well

失函数来优化分割模型,

$$L_{\text{SS}}(\theta) = \sum_{n=1}^{N} \left(\hat{q}_n^0 \log q_n(l_0|I;\theta) + \hat{q}_n^c \log q_n(y|I;\theta)\right). \tag{8}$$

其中, N为I中的像素个数, θ为网络参数, $\hat{q}_n^c = G_n$代表第n个像素为前景的概率, $\hat{q}_n^0 = 1-\hat{q}_n^c$, $q_n(y|I;\theta)$为第n个像素属于类别y的概率. 此概率来自于预测分数.

4.4 引入VOC图像

为了对比本文提出方法与现有弱监督语义分割方法, 本小节将PASCAL VOC 2012数据集中的图像加入训练集中. 为方便表示, 令$\mathcal{D}(W)$为网络图像集, 而$\mathcal{D}(V)$ 为PASCAL VOC 2012数据集中的训练

集(共10582张图像) [4, 44]. 对于$\mathcal{D}(W)$和$\mathcal{D}(V)$中标注为$\boldsymbol{y}$的图像I, 令$Z \in \mathbb{R}^{T\times(L+1)}$ 为使用$\mathcal{D}(W)$训练后的模型生成的预测分数. 其中T为像素数量.对于图像I, 本方法首先生成表示前景区域的二值化图像B, 并使用注意力模型[40]为每个类别生成注意力图. 然后, 将生成的注意力图与B结合计算调和平均数(参照式(1))来生成分数图$Q \in \mathbb{R}^{T\times(L+1)}$. 其中, Q_t^c表示第t个区域属于类别c的概率. 这里选择$Q_t^0 = 0.1$为默认设置, 表示每一个区域属于背景的概率为0.1.由此, 位置t对应的类别标签为

$$\hat{G}_t = \mathrm{argmax}_{\{l_0,\mathcal{L}\}}(Q_t). \tag{9}$$

给定$\hat{G}$, 则可直接用于训练语义分割模型.

5 实验结果

本节将给出本方法同现有网络监督以及弱监督语义分割方法的比较, 同时也对本方法中各组成部分的重要性进行详细分析.

5.1 实验配置

网络图像及低级别视觉特征的提取. 给定一组关键词, 本方法从Flickr网站中按照图像相关度为每个类别下载2000张图像.某些图像可能含有非常复杂的背景或含有多个语义类别. 考虑到图像间差异性, 本文采用一系列策略对低质量图像进行筛除. 为衡量网络图像的复杂度, 本方法首先采用拉普拉斯(Laplace)方差[45, 46] 判断图像是否模糊. 具体而言, 通过采用拉普拉斯算子对图像进行卷积运算后并计算其结果的方差, 可得到图像模糊程度的得分. 在实验中, 筛选阈值被设定为50, 即分数高于该阈值的图像将被过滤. 其次, 通过将输入图像由RGB颜色空间转换到HSV颜色空间, 可忽略掉所有曝光度与亮度较低或较高的图像. 具体而言, 通过计算每一幅图像H通道与V通道的均值, 将任一数值低于20的图像过滤.经过滤, 约33000张图像剩余, 即平均每个类别对应1650张图像.

数据集与评价标准. 如上文所述, 本文对PASCAL VOC 2012数据集[4]对应的每个类别分别下载约2000张图像, 组成数据集$\mathcal{D}(W)$. 对于显著性图, 本方法采用DSS显著性物体检测模型[42]. 分割结果的评测则采用类别平均交并比(mean intersection-over-union, mIoU).

模型超参数设定. 本方法采用CAM [24]作为注意力图生成模型以及ResNet-50 [47]作为噪声擦除模型的骨干网络并将其conv5步长由2变为1. 超参数设定具体如下: 学习率1e−3, 权重衰减5e−4, 模型动量0.9, 小批次中样本数量16, 训练周期5. 出于实验公正, 分割模型及其超参数设定皆与Deeplab-LargeFOV模型[18]相同. 本方法同时采用条件随机场[7]作为后处理工具.

5.2 消融实验

降噪能力. 为证明本方法处理类别噪声的能力, 本段列出使用及未使用噪声擦除模型的结果. 除列出平均交并比得分外, 本段同时给出不同类别的交并比得分. 具体结果可参见表1. 从该表中可看出, 采用任一种骨干网络, 使用噪声擦除模型后皆可大大提升分割结果. 例如, 在采用ResNet-101骨干网络的情况下, 噪声擦除模型的使用可带来超4%mIoU值的提升.另外, 噪声擦除模型的使用可使本方法生成更为精确的伪标注.因此, 对于含有少量噪声的类别, 本方法也带来少量提升.

训练图像数量的影响. 为进一步证明噪声擦除模型生成伪标注的质量, 这里按照一定比例从各类别图像随机选取一部分作为训练数据. 具体实验结果如表2所示. 从该表中可以看出, 当减少训练数据时, 实验结果并非呈明显下降趋势. 值得注意的是, 当训练样本数量减少到原始数量的30%时, 最终

表 1 噪声擦除模型消除噪声的效果.由于空间有限, 本表只选取包含类别噪声最多的6个类别与最少的3个类别

Table 1 Ability of defensing label noise. Due to the limited space, we only select 6 categories that contain the most noisy labels and also 3 categories that contain the least noisy labels

Category	NENet (VGG)		NENet (ResNet)	
	✗	✓	✗	✓
Bicyle	26.1%	30.0%	29.7%	33.3%
Chair	9.2%	14.6%	10.8%	13.5%
Horse	54.1%	60.1%	64.5%	74.2%
Motorbike	55.8%	57.0%	60.3%	61.8%
Table	6.2%	8.1%	7.0%	10.2%
Dog	65.3%	68.5%	76.9%	79.2%
Bottle	56.5%	58.9%	69.0%	68.0%
Bus	81.0%	80.4%	82.2%	83.3%
Cat	69.2%	66.4%	79.3%	81.9%
Mean	54.0%	56.9%	57.4%	61.6%

表 2 使用不同数量网络图像的消融实验结果. "占比"表示从每个类别中随机选择图像数量的比例. 该表中结果均以ResNet为骨干网络且在验证集上评估[a)]

Table 2 Ablation Analysis using different numbers of training web images. "Proportion" refers to the proportion of images we randomly selected from each category. All the results here are based on the ResNets backbone and evaluated on the validation set[a)]

No.	Images	Ratio (%)	NENet	mIoU (%)
1	33000	100	✗	57.4
2	33000	100	✓	**61.6**
3	16500	50	✓	60.5
4	9900	30	✓	59.0

a) Best result is highlighted in bold.

表 3 本方法消融实验结果. 若无特别声明, 所有结果皆来自网络输出且未使用任何后处理工具. 使用噪声擦除模型可提高4.2%的mIoU值. 加入VOC数据后亦得到4.2%的mIoU值提升. 所有结果均以ResNet为骨干网络且在验证集上进行评估[a)]

Table 3 Ablations for our proposed approach. Notice that all the results are directly from our CNNs without using any post-processing tools unless noticed. The mIoU score with our NENet increases by 4.2% compared to not using it. Further, involving VOC data also helps us obtain a performance gain of 4.2%. All the results here are based on the ResNets backbone and evaluated on the validation set[a)]

No.	Saliency map	NENet	VOC images	mIoU (%)
1	✓			57.4
2	✓	✓		$61.6_{+4.2}$
3	✓	✓	✓	$\mathbf{65.8}_{\mathbf{+4.2}}$

a) Best result is highlighted in bold.

的mIoU值仅下降2.6%. 这一现象表明噪声擦除模型对类别噪声具有较强的鲁棒性, 因而生成质量较高的伪标注.

VOC数据的影响. 如表3 与4 所示, VOC图像的引入可显著提升本方法在PASCAL VOC 2012验

表 4 本方法与前沿方法在验证集与测试集上的量化对比. 表中"33000网络图像"指本文中$\mathcal{D}(W)$数据集, 而$\mathcal{D}(V)$为10582张PASCAL VOC 2020数据集中图像. 对于依赖小规模像素级标注的方法, 其监督方式为半监督. 对于使用准确图像级类别标注的方法, 其监督方式为弱监督. 对于仅使用网络图像的方法, 其监督方式为网络监督[a)]

Table 4 Quantitative comparisons with existing state-of-the-art approaches on both 'val' and 'test' sets. Recall that the '33000 web images' in our method refers to $\mathcal{D}(W)$, and $\mathcal{D}(V)$ refers to 10582 PASCAL VOC images. For methods relying on a small set of pixel-accurate annotation, their supervision is represented by 'semi', for methods leveraging accurate image-level category labels, their supervision is denoted by 'weak', and for methods using only web images, their supervision is written as 'pure web'[a)]

Method	Training set	Supervision	Backbone	Val mIoU (%)	Test mIoU (%)
SEC [25]	$\mathcal{D}(V)$	Weak	VGGNet	50.7	51.7
AE-PSL [11]	$\mathcal{D}(V)$	Weak	VGGNet	55.0	55.7
DCSP [16]	$\mathcal{D}(V)$	Weak	ResNets	60.8	61.9
DSRG [28]	$\mathcal{D}(V)$	Weak	VGGNet	59.0	60.4
DSRG [28]	$\mathcal{D}(V)$	Weak	ResNet	61.4	63.2
MCOF [30]	$\mathcal{D}(V)$	Weak	VGGNet	56.2	57.6
Ahn et al. [29]	$\mathcal{D}(V)$	Weak	VGGNet	58.4	60.5
Wei et al. [31]	$\mathcal{D}(V)$	Weak	VGGNet	60.4	60.8
GAIN [32]	$\mathcal{D}(V)$	Semi	VGGNet	60.5	62.1
Fan et al. [48]	$\mathcal{D}(V)$	Weak	ResNet	63.6	64.5
Results based on web images					
STC [10]	40000 web images + $\mathcal{D}(V)$	Weak	VGGNet	49.8	51.2
WebS-i2 [12]	20000 web images + $\mathcal{D}(V)$	Weak	VGGNet	53.4	55.3
Hong et al. [15]	Web videos + $\mathcal{D}(V)$	Weak	VGGNet	58.1	58.7
Shen et al. [13]	80000 web images + $\mathcal{D}(V)$	Weak	VGGNet	58.8	60.2
Shen et al. [13]	80000 web images + $\mathcal{D}(V)$	Weak	ResNet	63.0	63.9
WebSearch (Ours)	33000 web images + $\mathcal{D}(V)$	Weak	VGGNet	**62.5**	**62.2**
WebSearch (Ours)	33000 web images + $\mathcal{D}(V)$	Weak	ResNet	**65.8**	**66.1**
Shen et al. [13]	80000 web images	Web	VGGNet	56.6	–
WebSearch (Ours)	33000 web images	Web	VGGNet	59.5	59.3
WebSearch (Ours)	33000 web images	Web	ResNet	61.6	62.0

a) Best results are based on VGGNet and ResNets are highlighted in bold.

证集上的结果. 例如, 当采用VGGNet [49]作为骨干网络时, 本方法在验证集上得到3%的提升. 而采用ResNet-101 [47]代替VGGNet作为骨干网络时, 本方法同样在验证集上得到4.2%的提升. 在PASCAL VOC 2012的测试集中可观察到类似结果.

可视化结果. 图7 为不同实验配置下得到的可视化结果. 从第1行可以看出, 对于简单图像而言, 仅使用$\mathcal{D}(W)$数据训练的模型也可达到较好效果. 对于包含多个类别的图像, 噪声擦除模型具有较好的鲁棒性. 对于场景复杂或对比度较低的图像, 本方法亦有错误预判, 但噪声擦除模型的应用同样生成更好的分割结果.

失败样例分析与讨论. 图7 下半部分给出了本方法生成的失败样例可视化结果. 从图中可以看出, 对于本方法而言, 处理几何形状较为复杂(如椅子)的图像具有一定难度. 同时, 在处理场景较为复杂的图像时, 本方法也存在缺陷. 可能的解决方案有两种. 其一是从下载的网络数据入手, 探索针对网络图

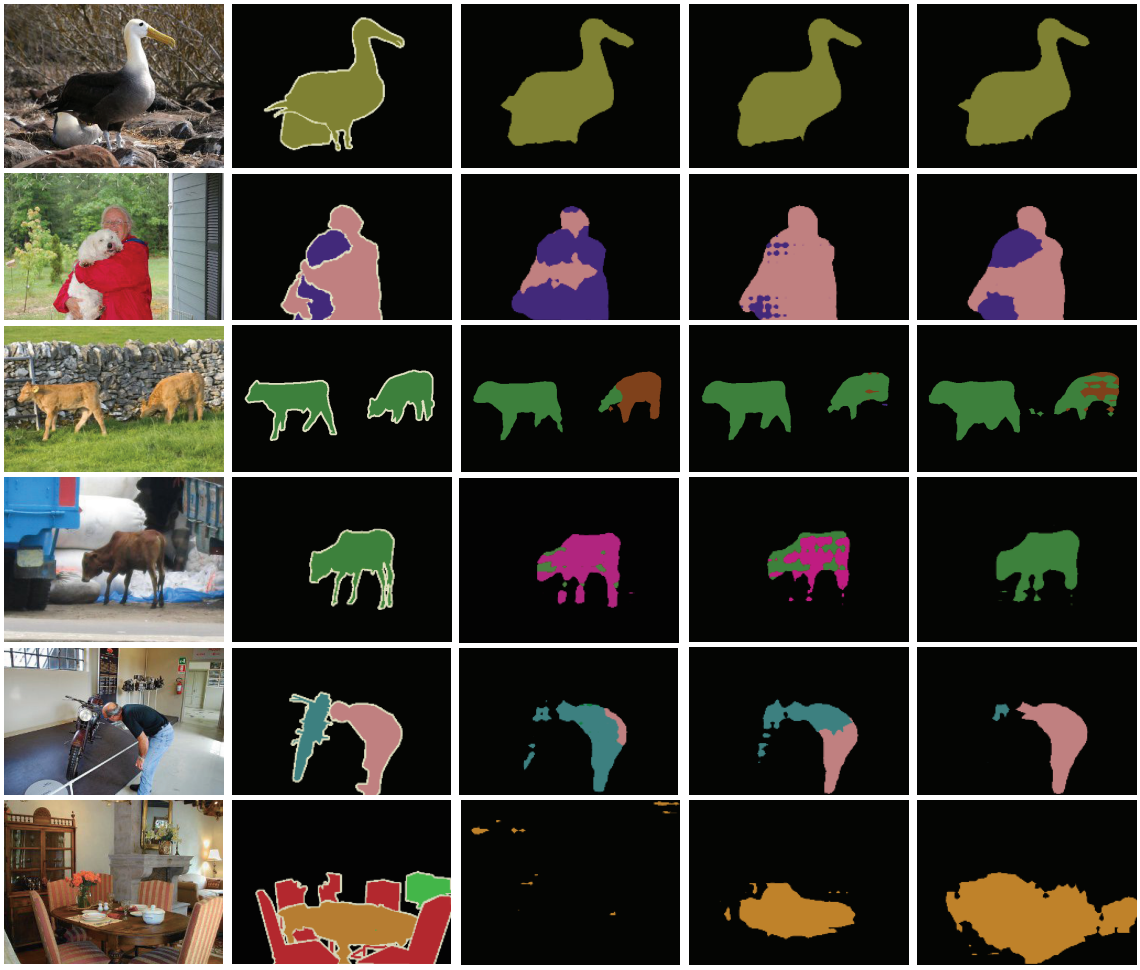


图 7 (网络版彩图) 本方法在不同实验配置下分割结果可视化对比. (a) 原图像; (b) 真实标注; (c) 未使用噪声擦除模型的结果; (d) 使用噪声擦除模型的结果; (e) 加入VOC 数据的结果

Figure 7 (Color online) Visual comparisons for weakly-supervised semantic segmentation using different settings of our proposed method. (a) Source image; (b) ground truth; (c) results without noise erasing network; (d) results with noise erasing network; (e) results with VOC images incorporated

像类别噪声的更为强大的过滤机制, 以得到质量较高的训练图像. 其二是探索更为有效地利用类间信息来区分不同类别语义物体的新思路.

5.3 与现有方法对比

本小节将本方法与现有弱监督语义分割方法进行比较. 若无特殊声明, 本处比较的所有模型皆基于Deeplab-LargeFOV基准模型[50]. 表4 已列出所有比较方法在验证集与测试集上结果. 由表可见, 在仅利用网络数据时, 本方法在测试集上已达到了62%的mIoU值. 与其他基于网络数据的方法相比(如Shen等[13]), 本方法在少用了近5万张网络图像的情况下, 仍然有2%左右提升.其主要原因在于, 现有方法并不能够很好地处理含有类别噪声较多的图像, 因而这些方法在训练时忽略了大量场景较为复杂的图像. 而场景较为复杂的图像通常含有较为丰富的背景信息.有效地利用这些丰富的背景信息可以帮助语义分割模型更好地区分语义物体区域与背景区域, 从而帮助分割模型得到更为精准的分割结

果.

为了更为公平地与现有基于弱监督的方法进行比较, 本文同样给出了在引入PASCAL VOC图像后的mIoU结果. 由表4 可以看出, 在引入入PASCAL VOC图像后, 本方法的mIoU得分进一步提高到66.1%, 远高于其他对比方法.由图7 亦可看出, 本方法在引入PASCAL VOC 数据后, 得到的分割结果明显优于本模型仅采用网络数据时生成的结果.其主要原因在于PASCAL VOC 数据提供了精准的类别信息.这些类别信息可以有效地帮助分割模型校正伪标注中的一些错误信息.

6 总结与讨论

本文探索了一个值得关注的计算机视觉问题, 即如何从含有类别噪声的网络图像中学习语义分割模型. 对于网络图像中的类别噪声, 本文提出了一种噪声擦除模型来去除可能与噪声相关的前景区域.一系列消融实验证明了本文提出方法中各组成部分的有效性. 实验结果表明, 本文提出方法在仅采用网络图像作为训练样本时已明显优于多数基于弱监督的前沿方法.

参考文献

1 Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017

2 Lin G S, Milan A, Shen C H, et al. Refinenet: multi-path refinement networks with identity mappings for high-resolution semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017

3 Lin G S, Shen C H, van den Hengel A, et al. Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016

4 Everingham M, Eslami S M A, van Gool L, et al. The pascal visual object classes challenge: a retrospective. Int J Comput Vis, 2015, 111: 98–136

5 Papandreou G, Chen L C, Murphy K, et al. Weakly-and semi-supervised learning of a DCNN for semantic image segmentation. In: Proceedings of IEEE International Conference on Computer Vision, 2015

6 Qi X J, Liu Z Z, Shi J P, et al. Augmented feedback in semantic segmentation under image level supervision. In: Proceedings of European Conference on Computer Vision, 2016

7 Lin D, Dai J F, Jia J Y, et al. Scribblesup: scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016

8 Bearman A, Russakovsky O, Ferrari V, et al. What's the point: semantic segmentation with point supervision. In: Proceedings of European Conference on Computer Vision, 2016. 549–565

9 Hou Q B, Dokania P K, Massiceti D, et al. Bottom-up top-down cues for weakly-supervised semantic segmentation. In: Proceedings of International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, 2017

10 Wei Y C, Liang X D, Chen Y P, et al. STC: a simple to complex framework for weakly-supervised semantic segmentation. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 2314–2320

11 Wei Y C, Feng J S, Liang X D, et al. Object region mining with adversarial erasing: a simple classification to semantic segmentation approach. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017

12 Jin B, Segovia M V O, Susstrunk S. Webly supervised semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3626–3635

13 Shen T, Lin G S, Shen C H, et al. Bootstrapping the performance of webly supervised semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 1363–1371

14 Mitchell T M, Cohen W W, Hruschka Jr E R, et al. Never ending learning. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015. 2302–2310

15 Hong S, Yeo D, Kwak S, et al. Weakly supervised semantic segmentation using web-crawled videos. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3626–3635
16 Chaudhry A, Dokania P K, Torr P H. Discovering class-specific pixels for weakly-supervised semantic segmentation. In: Proceedings of British Machine Vision Conference, 2017
17 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015
18 Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell, 2018, 40: 834–848
19 Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks. In: Proceedings of International Conference on Computer Vision, 2015
20 Hou Q B, Zhang L, Cheng M M, et al. Strip pooling: rethinking spatial pooling for scene parsing. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020. 4003–4012
21 Pinheiro P O, Collobert R. From image-level to pixel-level labeling with convolutional networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015
22 Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the em algorithm. J Royal Stat Soc, 1977, 39: 1–38
23 Pathak D, Krahenbuhl P, Darrell T. Constrained convolutional neural networks for weakly supervised segmentation. In: Proceedings of International Conference on Computer Vision, 2015
24 Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016
25 Kolesnikov A, Lampert C H. Seed, expand and constrain: three principles for weakly-supervised image segmentation. In: Proceedings of European Conference on Computer Vision, 2016
26 Roy A, Todorovic S. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017
27 Oh S J, Benenson R, Khoreva A, et al. Exploiting saliency for object segmentation from image level labels. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017
28 Huang Z L, Wang X G, Wang J S, et al. Weakly-supervised semantic segmentation network with deep seeded region growing. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 7014–7023
29 Ahn J, Kwak S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018
30 Wang X, You S D, Li X, et al. Weakly-supervised semantic segmentation by iteratively mining common object features. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 1354–1362
31 Wei Y C, Xiao H X, Shi H, et al. Revisiting dilated convolution: a simple approach for weakly-and semi-supervised semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018
32 Li K P, Wu Z Y, Peng K C, et al. Tell me where to look: guided attention inference network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018
33 Zhang X L, Wei Y C, Feng J S, et al. Adversarial complementary learning for weakly supervised object localization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018
34 Fan D P, Lin Z, Ji G P, et al. Taking a deeper look at co-salient object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020. 2919–2929
35 Borji A, Cheng M M, Hou Q B, et al. Salient object detection: a survey. Comp Visual Media, 2019, 5: 117–150
36 Fan R, Cheng M M, Hou Q B, et al. S4Net: single stage salient-instance segmentation. Comput Visual Media, 2020, 6: 191–204
37 Liu J J, Hou Q B, Cheng M M, et al. A simple pooling-based design for real-time salient object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 3917–3926
38 Cheng M M, Mitra N J, Huang X, et al. Global contrast based salient region detection. IEEE Trans Pattern Anal Mach Intell, 2015, 37: 569–582
39 Jiang P T, Hou Q B, Cao Y, et al. Integral object mining via online attention accumulation. In: Proceedings of International Conference on Computer Vision, 2019. 2070–2079
40 Hou Q B, Jiang P T, Wei Y C, et al. Self-erasing network for integral object attention. In: Proceedings of the 32nd

International Conference on Neural Information Processing Systems, 2018
41 Frénay B, Kabán A. A comprehensive introduction to label noise. In: Proceedings of European Symposium on Artifical Neural Networks, Computional Intelligence and Machine Leaning, 2014
42 Hou Q B, Cheng M M, Hu X W, et al. Deeply supervised salient object detection with short connections. IEEE Trans Pattern Anal Mach Intell, 2019, 41: 815–828
43 Maninis K K, Pont-Tuset J, Arbelaez P, et al. Convolutional oriented boundaries: from image segmentation to high-level tasks. IEEE Trans Pattern Anal Mach Intell, 2018, 40: 819–833
44 Hariharan B, Arbeláez P, Bourdev L, et al. Semantic contours from inverse detectors. In: Proceedings of International Conference on Computer Vision, 2011
45 Pech-Pacheco J L, Cristóbal G, Chamorro-Martinez J, et al. Diatom autofocusing in brightfield microscopy: a comparative study. In: Proceedings of International Conference on Pattern Recognition, 2000
46 Pertuz S, Puig D, Garcia M A. Analysis of focus measure operators for shape-from-focus. Pattern Recogn, 2013, 46: 1415–1432
47 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016
48 Fan R C, Hou Q B, Cheng M M, et al. Associating inter-image salient instances for weakly supervised semantic segmentation. In: Proceedings of European Conference on Computer Vision, 2018
49 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of International Conference on Learning Representations, 2015
50 Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFS. In: Proceedings of International Conference on Learning Representations, 2015

Autonomous learning of semantic segmentation from Internet images

Qibin HOU[1,2], Ling-Hao HAN[1], Jiang-Jiang LIU[1] & Ming-Ming CHENG[1*]

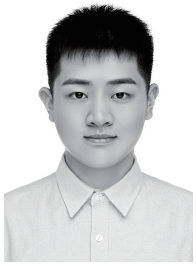
1. *College of Computer Science, Nankai University, Tianjin* 300350*, China*;
2. *College of Electrical and Computer Engineering, National University of Singapore, Singapore* 117583*, Singapore*
* Corresponding author. E-mail: cmm@nankai.edu.cn

Abstract Collecting a large amount of manually labeled training data is labor-intensive, thus often becomes the major bottleneck when applying semantic segmentation techniques to real-world applications, especially for new categories where no labeled data is available. In this paper, we aim at solving the problem of "webly-supervised" semantic segmentation relying purely on web searched images, where users only need to provide a single keyword for each target category. A major challenge in this task is the existence of label noise in web images. To deal with the label noise, we design a noise erasing network that is able to learn cross-image knowledge from credible attention regions in images in a mini-batch and then erase those regions unrelated to the search keywords from the web images. With this network, our system can automatically generate high-quality 'proxy ground truth', for training semantic segmentation models. Extensive experiments on the popular benchmark, i.e., PASCAL VOC 2012, show surprisingly good results in both our task (mIoU = 62.0%) and the weakly-supervised setting (mIoU = 66.1%).

Keywords semantic segmentation, web search, label noise, noise erasing network, web supervision

Qibin HOU was born in 1991. He received his Ph.D. degree from Nankai University, Tianjin, under the supervision of professor Ming-Ming Cheng, in 2019. Currently, he is a research fellow at the University of Singapore working with professor Jiashi Feng. His research interests include deep learning and computer vision.

Jiang-Jiang LIU was born in 1995. He is currently a Ph.D. candidate at the School of Computer Science, Nankai University, Tianjin, under the supervision of professor Ming-Ming Cheng. His research interests include deep learning and computer vision.



Ling-Hao HAN was born in 1998. He is an undergraduate student at the College of Software, Nankai University, Tianjin. He will pursue his master degree under the supervision of professor Ming-Ming Cheng after graduation. His research interests include deep learning and computer vision.

Ming-Ming CHENG was born in 1985. He received his Ph.D. degree from Tsinghua University, Beijing, in 2012, and then worked with professor Philip Torr in the University of Oxford for 2 years. He is now a professor at Nankai University, Tianjin, leading the Media Computing Lab. His research interests includes computer vision and computer graphics.