利用多种子信息聚合和正负混合学习的弱监督图像语义分割

桑 雨*①② 刘 通① 马天娇① 李 乐① 李思漫① 刘宇男③
①(辽宁工程技术大学电子与信息工程学院 葫芦岛 125105)
②(渤海大学数学科学学院 锦州 121013)
③(大连海事大学人工智能学院 大连 116026)

摘 要：基于图像级标签的弱监督语义分割(WSSS)旨在通过类激活映射(CAM)生成伪标签(种子)，然后将其用于训练语义分割模型，为耗时且昂贵的像素级标注节省大量人力和财力。现有方法主要围绕CAM进行改进以获取单个优良的种子，同时通过一些后处理手段进一步提升种子的质量，但其得到的种子仍存在不等程度的噪声。为了减少噪声标签对分割网络造成的影响，考虑多个不同种子更能有效提取出正确信息，该文从多种子信息互补的角度，提出一种基于多种子信息聚合和正负混合学习的弱监督图像语义分割方法，通过在分类网络中改变输入图像尺度以及调整Dropout层随机隐藏神经元的概率，获取多个优良种子；依据它们对每个像素分配的类别标签情况进行优选获得聚合种子，并进一步区分该像素标签为干净标签还是噪声标签；利用正负混合学习训练语义分割网络，引入预测约束损失以避免网络对噪声标签给予过高的预测值，进而对干净标签应用正学习发挥正确信息的准确性，对噪声标签应用负学习抑制错误信息的影响，从而有效提升分割网络的性能。在PASCAL VOC 2012和MS COCO 2014验证集上实验结果表明，该文方法在基于卷积神经网络框架的分割网络中，mIoU分别达到了72.5%和40.8%，与RCA及URN方法相比分别提升了0.3%与0.1%；在基于Transformer框架的分割网络中，mIoU则提升至76.8%和46.7%，与CoBra及ECA方法相比分别提升了2.5%与1.6%，验证了方法的有效性。

关键词：弱监督语义分割；类激活映射；多种子；交叉熵；正负学习

中图分类号：TN911.73; TP391.41 **文献标识码**：A **文章编号**：1009-5896(2025)10-3993-12

DOI: 10.11999/JEIT250112 **CSTR**: 32379.14.JEIT250112

1 引言

图像语义分割[1]是计算机视觉中一项重要任务，目标是为每个像素分配对应的语义标签。近年来，随着深度学习技术尤其是卷积神经网络(Convolutional Neural Networks, CNN)的发展，语义分割取得了快速进步，也因此推动了许多实际应用的发展，比如医学影像[2]、自动驾驶[3]、遥感影像[4]等。相较于分类和检测等任务，语义分割需要大量像素级标注的训练图像，而此类标注工作耗时且昂贵。为此，研究人员试图利用弱标签进行弱监督语义分割(Weakly Supervised Semantic Segmentation, WSSS)研究，旨在以较低标注成本实现较精准的分割效果，如边界框标签[5]、涂鸦式标签[6]、点标注[7]和图像级标签[8]。其中，图像级标签因标注成本最低且获取难度最小，受到了许多研究者的关注。

基于图像级标签的WSSS方法主要包括两个阶段：第1阶段利用分类网络为每幅输入图像生成种子；第2阶段利用所获得的种子作为监督训练语义分割模型。其中，第1阶段生成的种子主要是基于类激活映射(Class Activation Mapping, CAM)[9]，它通过使用全局平均池化(Global Average Pooling, GAP)进行图像分类，从而定位目标对象。然而，CAM通常仅覆盖物体中最具有判别性的小部分特征，极易生成稀疏且不完整的目标对象区域。因此，许多研究方法[10,11]致力于通过覆盖更多的目标对象、学习更完整的对象区域以获取高质量的种子。例如，Lee等人[12]挖掘定位映射与显著映射之间的关系排除种子的共现像素。Lee等人[13]考虑分布外(Out-of-Distribution, OOD)数据提供的线索区分目标前景与背景，从而减弱分割中的错误关联性。Chang等人[14]提出一种子类别探索的自监督方法，通过对每个父类的图像特征进行优化，使分类网络提取子类别特征以提高预测的准确性。上述方法本质是应用各种规则和手段来改善CAM生成的单个种子质量以获取适用于语义分割网络的高质量种子，但其仍不可避免地包含噪声。当面对噪声标签时，CNN的强记忆能力[15]会使其性能显著下降，需要对此情况进一步处理。Chen等人[16]利用具有类自适应阈

收稿日期：2025-02-26；改回日期：2025-08-20；网络出版：2025-08-27
*通信作者： 桑雨 sangyu@lntu.edu.cn
基金项目：国家自然科学基金(62372077, 62302249)，中国博士后科学基金(2022M720624)，辽宁省教育厅科研基金(LJKQZ2021152)，国家重点研发计划(2019YFB2102400)，高校人才引进基金(18-1021)
Foundation Items: National Natural Science Foundation of China (62372077, 62302249), China Postdoctoral Science Foundation (2022M720624), The Research Fund of Liaoning Provincial Education Department (LJKQZ2021152), National Key Research and Development Program of China (2019YFB2102400), University Talent Introduction Foundation(18-1021)

值的在线噪声过滤模块处理图像级噪声以及渐进噪声检测模块处理像素级噪声。Rong等人[17]提出一种协同训练范式改进不确定像素的学习，进一步提出边界增强策略提升模型对于边界区域的预测。

考虑到单一种子在信息提取过程中，易受噪声和局部信息的影响，导致提取的信息存在偏差。多个种子可以从不同位置和尺度获取信息，这些信息相互补充，从而有效提取正确信息。负学习则是通过间接学习，让模型在训练过程中规避噪声，进而减轻噪声对分割过程的干扰。基于此，本文提出一种基于多种子信息聚合和正负混合学习的弱监督图像语义分割方法。首先，基于通用分类网络[18]，通过改变输入图像尺度，调整网络中Dropout层[19]以不同概率随机隐藏神经元来获取多个优良种子，达到多种子信息相互补充的目的；然后，依据这些种子中每个像素呈现的类别标签情况进行优选，即通过同一标签出现的最大次数是否超过设定的阈值来区分该像素为可信(干净)标签和不可信(噪声)标签。如果像素标签为干净标签，则有利于发挥正确信息的效用，提升最后的分割结果；如果像素标签为噪声标签，则会将错误信息传递给网络，降低网络性能。故本文提出正负混合学习策略来训练语义分割网络，将区分出的干净标签通过“输入图像属于其标签”的监督方式训练分割网络，记为正学习；将区分出的噪声标签通过“输入图像不属于其互补标签”的间接学习方式训练分割网络，记为负学习。对于分类网络中获得的噪声标签，通过正负混合学习的方式来抑制其中错误信息的影响，有效提升分割网络的性能。同时，考虑传统的交叉熵损失会对噪声标签给予较高的预测值，本文引入预测约束损失[20]缓解这一问题，以提升模型对可信标签的预测准确性，同时减少对噪声标签的过拟合。本文主要贡献及创新点如下：

(1)提出多种子信息聚合的伪标签生成方法，即考虑多种子信息的相互补充，以“投票”的方式优选干净标签并进一步区分噪声标签，目标是分别执行正学习与负学习，为分割网络性能的提升提供有力支撑；

(2)提出正负混合学习策略并引入预测约束损失，在干净标签上实施正学习以发挥正确信息的准确性，同时在噪声标签上进行负学习以间接抑制错误信息的影响，从而有效提升了分割性能；

(3)提出方法在PASCAL VOC 2012和MS COCO 2014数据集上取得良好性能，基于CNN分割网络的mIoU分别达到了72.5%和40.8%；基于Transformer的分割网络达到了76.8%和46.7%，优于当前的WSSS方法。

本文其他部分组织如下：第2节提出基于多种子信息聚合和正负混合学习的弱监督图像语义分割方法；第3节描述实验设置并分析实验结果；第4节总结本文工作并对未来工作进行展望。

2 方法原理

在基于两阶段的WSSS方法中，由于第一阶段生成的单个种子仍存在噪声，用于下一阶段的训练会对分割性能造成直接影响。基于此，本文提出一种多种子信息聚合和正负混合学习的WSSS方法提升分割性能。本节将详细介绍提出的方法。

图像语义分割任务的主要目标是识别图像中的对象以及为各像素精确地标注所属的类别。对于弱监督学习，输入图像$\boldsymbol{I}$与图像级标签y组成弱监督数据集$D=\{(\boldsymbol{I},y)\}$。本文将基于D通过提出方法得到分割结果S。总体上，本文方法框架包括两个阶段，如图1所示。第1阶段：在分类网络中通过调整Dropout层生成多个种子，并对这些种子进行多标签区分，区分出干净标签和噪声标签。其中，实线箭头表示某一概率p的分类；虚线箭头表示根据不同概率生成的其他种子的分类，通过该流程的多次执行，可获得多个CAM图。第2阶段：提出正负混合学习，对干净标签和噪声标签分别应用正学习和负学习来指导模型训练，最终通过分割模型产生最终分割结果。

2.1 种子获取

为获得多个具有区分度的种子，本文通过修改分类网络的最后一层Dropout的概率p来实现。随机选择隐藏单元可以获得与之对应的分类分数，每个随机选择生成不同的定位图。具体地，CAM首先训练具有GAP的分类网络，计算多标签交叉熵损失如下

$$L_{\rm cls}=-\frac{1}{N}\sum_{i=1}^{N}y_i\log_{\rm e}\sigma(\hat{y}_i)+(1-y_i)\log_{\rm e}[1-\sigma(\hat{y}_i)],\quad(1)$$

其中，N是数据集D中前景对象的类别；$y_i\in\{0,1\}$，0表示图像中不包含第i类，1表示图像中包含第i类；$\hat{y}_i$是分类网络对第i类目标的预测值；$\sigma(\cdot)$表示Sigmoid函数，将模型输出的预测值映射到概率值为[0,1]，保持数值稳定性，进而计算损失函数，其为每个类别提供独立的概率估计，使模型分别预测每个类别的概率，避免不同类别之间的预测互相影响。输入图像$\boldsymbol{I}$，卷积层产生一系列特征图$\boldsymbol{f}\in\mathbb{R}^{C\times H\times W}$，其具有$C$通道和$H\times W$空间大小。将$\boldsymbol{f}$与对类别$i$的权重$\boldsymbol{w}$加权求和，其中$\boldsymbol{w}$具有$C$通道，为了实现加权求和，需要将$\boldsymbol{w}_i$进行转置，实现通道维度对齐，其反映了输入图像中每个位置对于特定类别的贡献。获得第i类对象的CAM如下：

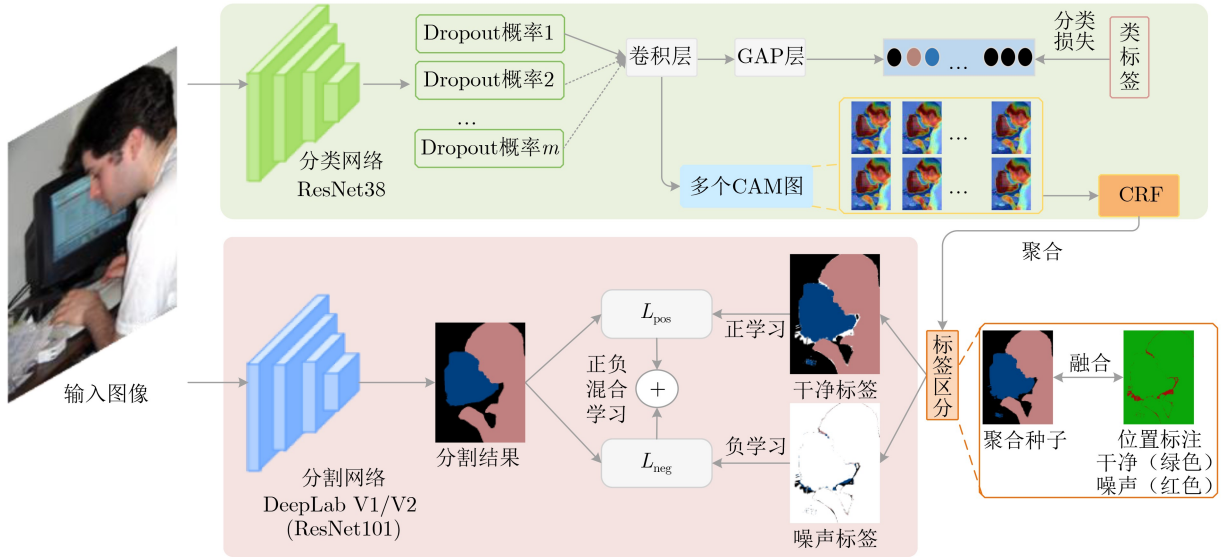


图 1 本文方法整体框架图

$$M_i = \mathrm{ReLU}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{f}), \forall i \in N, \tag{2}$$

其中，ReLU函数意在消除负激活值，过滤掉对类别i有负面影响的区域，并将对应的映射设置为0。随后，对M_i进行归一化，将其映射至[0,1]区间表示类别i的概率。

2.2 多种子生成

目前，大多数方法针对CAM产生的单个种子加以优化。然而，经过优化的种子[12]仍包含噪声，如图2所示，一些像素依然标注不准确。如果直接使用含噪的种子训练分割模型，会降低分割性能。

本文旨在考虑多个种子，从信息互相补充的角度提升标签的准确性，如图3所示。受Yang等人[21]启发，当图像被调整为不同尺度时，目标对象表现出互补特征，故通过本文选取种子并改变其输入图像的缩放尺度以区分种子标签。此外，本文选择ResNet38[18]作为分类网络，通过调整Dropout层中神经元的隐藏概率p，使网络基于不同的神经元随机失活模式产生相应的分类结果，进而获得多个不同的种子。

本文对所获得的m个种子利用条件随机场(Conditional Random Field, CRF)[22]做进一步优化处理。CRF能够利用邻近像素之间的空间关系，引入平滑项保持区域标签的一致性。具体来说，种子的前景CAM M_{fg}依据幂指数α构建出背景CAM $M_{\mathrm{bg}} = [1-\max(M_{\mathrm{fg}})]^{\alpha}$；然后将前景与背景结合作为CRF

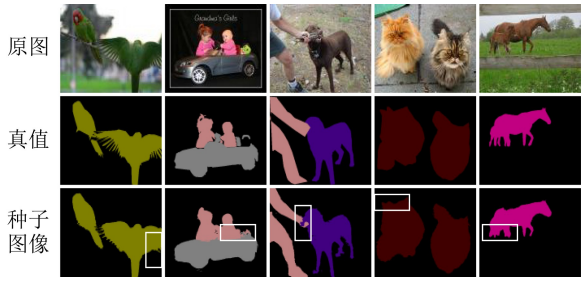


图 2 单种子与真值标签对比图

一元势函数的输入，即$M = M_{\mathrm{fg}} \cup M_{\mathrm{bg}}$，进一步采用argmax选取像素的最优标签，以此获得处理后的种子。图4为产生6个种子示例图，其中绿色圆圈代表该像素为干净标签，白色圆圈表示该像素为噪声标签。本文通过调整dropout的概率以及缩放尺度，使生成的种子呈现互补信息。在3.3.1节中，本文对生成的8个种子进行分析，筛选出6个种子作为方法所需的多个种子，并通过评估各种子的精度验证其差异性。

2.3 干净/噪声标签区分策略

Lee等人[23]在多种子信息聚合方面提出了集成类激活图像素点分类分数方法；与之不同，本文方法对类激活图进一步细化处理，提升标签的准确度，并且设计了区分正负标签策略以适应后续的正负学习，为分割网络提供支持。具体地，根据多种子在每个像素呈现的情况进行判断，通过设定阈值

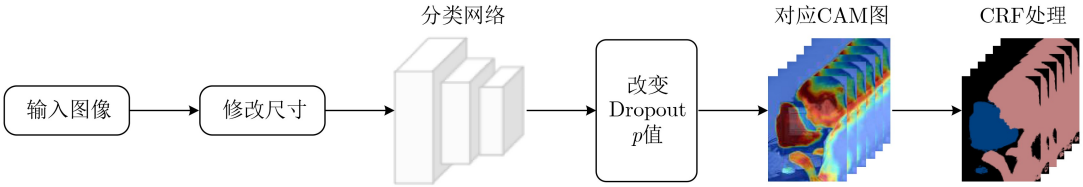


图 3 多种子生成流程图

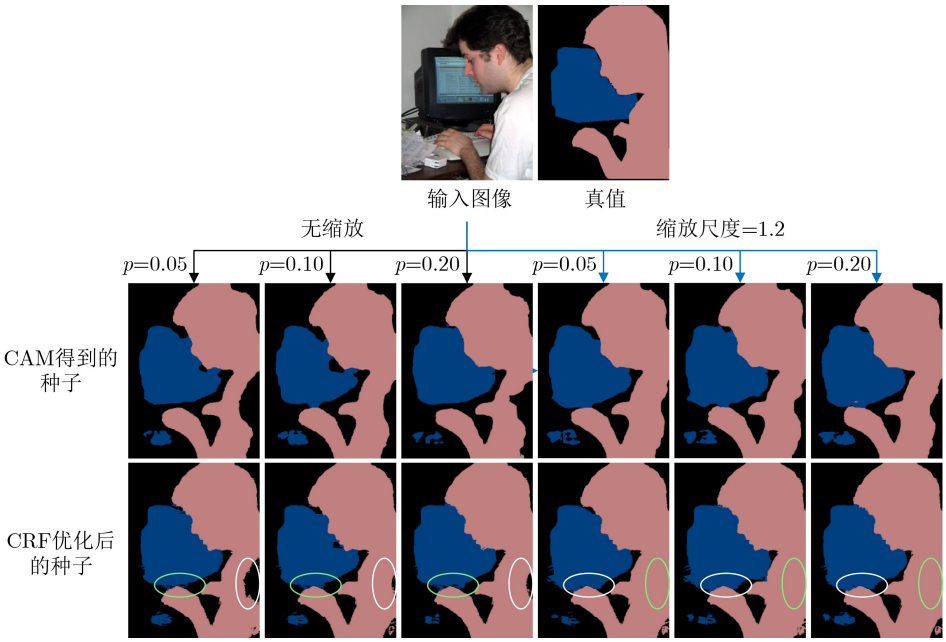


图 4 生成6个种子示例图

的方式区分该标签是否为干净标签以进行正负学习。对多种子聚合包括两个步骤。首先，统计m个种子对同一像素分配的类别标注情况，将标注频率最高的类别确定为聚合种子的类别。考虑到多次出现在同一位置的像素存在为噪声标签的可能，本文方法对噪声做进一步处理。接下来，通过设定阈值的方式划分种子的可信度。

步骤1：计算分类网络对每个类别的输出概率，其反映了模型对每个类别的可信程度。像素的概率值越高，表明模型越确信该像素属于相应的类别。在图像分类过程中，对于每个像素，首先计算其在不同标注中的类别概率值，并将其转化为相应的标注频次。接着，通过对各类别标注频次的聚合，选取标注频次最高的类别作为该像素的最终类别，以此作为聚合种子中像素标签的确定依据。图5为在(178, 360)和(370, 345)位置处的区分示例图，像素(178, 360)被识别为20(显示器)的最高次数为4次，因此对应聚合种子的位置为类别20；像素(370, 345)被识别为15(人)的最高次数为3次，因此对应位置为类别15。

步骤2：通过设定不同的阈值将标签区分为干净与噪声。步骤1中聚合种子的像素标签是直接选择6个种子出现次数最多的类别作为标签，尽管最大可能地为每个像素选择正确的标签，但实际情况往往较为复杂，与真值标签相比，聚合种子仍存在噪声，如图5中像素(178, 360)的选取。为进一步抑制噪声标签的影响，本文通过设定不同的阈值t区分干净和噪声标签，其中t的上限取决于种子的数量，即本文选取6个种子，t的上限为6。具体地，将标签的最大出现次数n与t进行比较，当$n \geq t$时，该像素被归为干净标签，否则为噪声。图6为聚合种子在不同阈值下的区分结果示例，随着t值逐渐增大，噪声标签被明显区分。因此，仅利用单种子标签进行正学习不可避免被噪声标签影响分割结果。

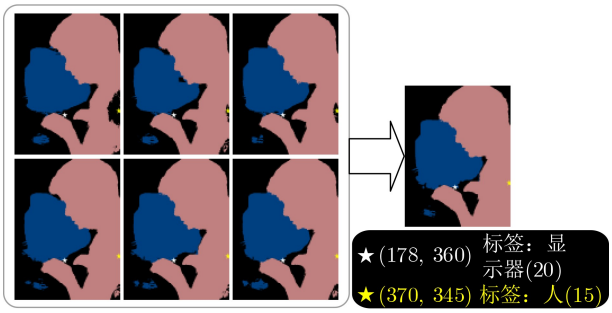
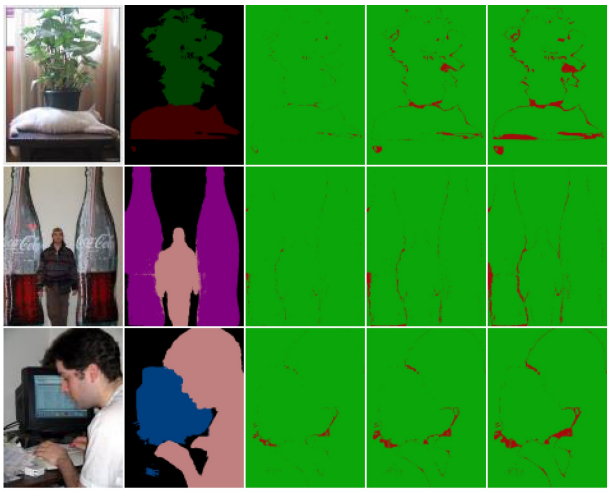


图 5 标签区分示例图



(a) 原始图像(b) 聚合种子 (c) t=4情况 (d) t=5情况 (e) t=6情况

注：绿色表示干净标签，红色表示噪声标签

图 6 聚合种子在不同阈值t下的区分结果

本文从多种子角度出发，引入负学习来抑制噪声标签对网络的影响。

2.4 正负混合学习

现有方法试图通过改进由CAM生成的单个种子，但改进后的单个种子中仍存在噪声。本文提出了正负混合学习策略，如图7所示。其中，正学习(Positive Learning, PL)核心思想是“像素被标注为给定标签”的学习，即网络直接训练CNN属于PL。然而，当标签含有噪声时，此方式会向网络给予错误信息，进而降低模型性能。为缓解噪声标签对模型的影响，负学习(Negative Learning, NL)[24]被提出，其核心思想是“像素被标注为互补标签”的学习，即网络随机选取除给定标签外的其他类别作为训练样本的标签，其正确信息可能被间接学习到，进而降低错误信息的影响。如图6所示，对绿色部分应用PL，该像素属于给定的标签；对红色部分应用NL，该像素不属于其互补标签。

Kim等人[25]通过实验证明，利用有效的系数调节交叉熵损失的贡献，使模型对较高置信度的数据上加快训练过程并提高模型的收敛速度，同时减少模型出现欠拟合的风险。为解决传统交叉熵损失在依赖伪标签(尤其是含噪声伪标签)时强制拟合不可靠监督信号、缺乏类间关系建模的问题，Li等人[20]提出将L_2范数融合到交叉熵损失，形成预测约束学习损失。其核心在于通过挖掘模型预测的类间关系，实现干净标签的有效学习与噪声标签隐藏信息的利用，进而提升预测的准确性。因此，本文采用有效的正负混合学习方法，并引入预测约束学习，通过优势互补实现对噪声的精细化处理。

$$L_{\mathrm{ce}}(\boldsymbol{f},y)=-\sum_{i=1}^{N+1}y_i\log_{\mathrm{e}}(p_i) \tag{3}$$

具体地，$\boldsymbol{P}=\boldsymbol{f}(x)\in\mathbb{R}^{C\times(N+1)\times H\times W}$表示骨干网络的输出，将像素映射至$N+1$维概率分布$p$，传统的交叉熵损失如式(3)，其中$p_i=\dfrac{\mathrm{e}^{z_k}}{\sum_{k=1}^{N+1}\mathrm{e}^{z_k}}$，$N+1$表示类的数量，$y$是给定标签，$p_i$是第$i$个类别的概率，$z_i$是概率输出的第$i$个元素。在分割网络中，调节交叉熵损失表示为

$$L_{\mathrm{ps}}(\boldsymbol{f},y)=-\prod_{x=0}^{3}(1+p_y^{2^x})\sum_{i=1}^{N+1}y_i\log p_i, \tag{4}$$

以下为L_{ps}损失梯度：

$$\begin{aligned}\nabla L_{\mathrm{ps}}&=\prod_{x=0}^{3}(1+p_y^{2^x})\nabla L_{\mathrm{PL}}=\prod_{x=0}^{3}(1+p_y^{2^x})(1-p_y)\\&=-(1-p_y^{2^4})\end{aligned} \tag{5}$$

其中，$\prod_{x=0}^{3}(1+p_y^{2^x})$作为常数权重因子，在参数范围内，常数的增加可有效提升L_{pos}收敛速度。此外，预测约束损失表示为$L_{\log}=-\sum_{i=1}^{N+1}y_i\log(p_i^*)$，其中，$p^*{}_k=\dfrac{\mathrm{e}^{z_k/\|\boldsymbol{z}\|}}{\sum_{i=1}^{N+1}\mathrm{e}^{z_i/\|\boldsymbol{z}\|}}$，$\|\cdot\|$表示$L_2$范数。$L_{\mathrm{PL}}$到第$j$个预测的损失梯度如式(6)所示，该梯度仅缩小了预测值与相应标签之间的差距，如果给出噪声标签，则影响模型训练。

$$\frac{\nabla L_{\mathrm{ce}}}{\nabla z_j}=p_j-y_j \tag{6}$$

$$\frac{\nabla L_{\log}}{\nabla z_j}=\frac{1}{\|\boldsymbol{z}\|}\left((p_j^*-y_j)-\sum_{i=1}^{N+1}\frac{z_jz_i}{\|\boldsymbol{z}\|^2}(p_i^*-y_i)\right) \tag{7}$$

式(7)中，梯度由两部分组成，前一部分与L_{PL}的梯度基本相同，后一部分则反映了在不同类别之间建立的连接。综上，将y与互补标签$\bar{y}$进行正负混合学习，利用正学习网络$\boldsymbol{f}_1(\boldsymbol{P};\delta_1)$和负学习网络$\boldsymbol{f}_2(\boldsymbol{P};\delta_2)$获得分割结果，其中，$\delta$表示权重，本文采用损失函数表示为

$$L_{\mathrm{pos}}(\boldsymbol{f}_1,y_{\mathrm{c}})=-\prod_{x=0}^{3}(1+p_{y_{\mathrm{c}}}^{2^x})\sum_{i=1}^{N+1}y_i\log p_i^*, \tag{8}$$

通过优化损失函数，使给定标签的预测概率分布尽可能接近1，即$p_{y_{\mathrm{c}}}\to1$，以适应正学习。相对地，采用负学习对噪声标签y_n的数据进行训练，其损失函数表示为

$$L_{\mathrm{neg}}(f_2,\bar{y}_n)=-(1-p_{\bar{y}_n})\sum_{i=1}^{N+1}\bar{y}_i\log(1-p^*{}_i), \tag{9}$$

其中，$\bar{y}_i$表示第i类的互补标签，$(1-p_{\bar{y}_n})$表示恒定加权因子。当$p_{\bar{y}_n}$较高时，相应的损失会降低，从而减轻噪声标签的影响；反之，当$p_{\bar{y}_n}$较低时，模型会加强学习，防止噪声标签对模型的干扰。因此，通过调整损失函数，使$p_{\bar{y}_n}\to0$，以增强模型对其它类别的学习，提高对正确标签的预测概率。

根据式(9)与式(10)，分割网络正负混合学习的损失函数表示为

$$L=\varepsilon L_{\mathrm{pos}}+(1-\varepsilon)L_{\mathrm{neg}}, \tag{10}$$

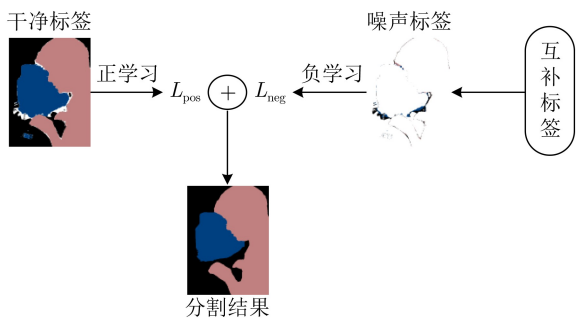


图 7 正负混合学习流程图

其中，ε是平衡$L_{\rm pos}$与$L_{\rm neg}$的权重参数。在每次迭代中$\varepsilon = a/b$，其中，a为干净标签数，b为所有标签数，目的是充分发挥PL和NL综合的优势。

3 实验及结果分析

本节将首先介绍实验设置；然后，将本文方法与最新方法进行对比评估；接下来描述消融实验；最后，进行统计情况分析。

3.1 实验设置

3.1.1 数据集和评价标准

本文使用PASCAL VOC 2012[26]和MS COCO 2014[27]作为基准数据集。PASCAL VOC 2012数据集包含20个前景类和一个背景类，其中，一张图片中可能包含多个类。遵循之前的工作[12,14]，通常会使用PASCAL VOC的增强数据集SBD数据集[28]。该数据集包含10 582张用于训练的图像，1 449张用于验证的图像，以及1 456张用于测试的图像。MS COCO 2014数据集包含80个前景类和一个背景类，该数据集包含82 081张和40 137张用于训练以及验证的图像。在所有实验中，使用平均交并比(mean Intersection over Union, mIoU)[29]作为评价标准，该值为真值集合与预测值集合的交集与并集之比。

$$\mathrm{mIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{e_{ii}}{\sum_{j=0}^{k}e_{ij}+\sum_{j=0}^{k}e_{ji}-e_{ii}}, \tag{11}$$

其中，k表示数据集的类别数，e_{ii}表示分类正确的像素总数，i和j表示类别号，e_{ij}与e_{ji}表示分类错误的像素总数。

3.1.2 实现细节

本文通过在PyTorch中使用单个RTX 3090 GPU实现了提出的框架。对于种子生成，采用基于ImageNet[30]预训练的分类网络ResNet38[18]。将Dropout的概率p值分别设置为0.05, 0.10, 0.20, 0.30，缩放尺度为原图的1.2倍。将学习率设置为0.01，采用SGD优化器。在VOC数据集上，训练20次迭代，批次大小为8。在COCO数据集上，训练30次迭代，批次大小为16。对于数据增强，对数据集进行随机缩放，随机反转并随机裁剪为448×448。对于分割网络，本文采用基于Resnet101作为骨干网络的DeepLab-V1[31]和DeepLab-V2[32]。设置学习率为0.001。在VOC数据集上训练模型为20k迭代；在COCO数据集训练模型为60k迭代，训练批次大小均为10。

3.2 与最新方法对比

为验证本文方法的有效性，如表1所示，将提出方法与最新WSSS方法在PASCAL VOC 2012验证集与测试集上进行对比。采用DeepLabV2作为分割网络，并于验证集上与测试集上分别达到了72.5%和73.9%的mIoU值。本文方法在验证集上与改进CAM的方法ReCAM[33]相比，mIoU值提升了0.9%；与利用噪声标签学习的方法BECO[17]相比，mIoU值提升了0.4%；与最新的方法RCA[34]相比，验证集上提升了0.3%，测试集上提升了0.8%。此外，本文利用基于MiT-B2作为骨干网络的SegFormer[39]分别达到了75.3%和75.9%，基于Swin-L[40]作为骨干网络的Mask2Former[41]分别达到了76.8%和77.5%，相较于CoBra[35]分别提升了2.5%和3.3%，分割精度出现明显的提升，由于MiT-B2作为深层次网络能有效捕捉不同尺度的特征信息，Swin-L基于自注意力机制能够更好地捕捉目标，为分割任务提供准确的上下文信息。

表2呈现了本文方法与其他最新方法在PASCAL

表 1 本文方法与不同模型在PASCAL VOC 2012验证集与测试集上的性能对比(粗体表示最优值)

模型		骨干网络	mIoU(%)	
			验证集	测试集
基于CNN方法	SIPE$_{(\rm CVPR22)}$[36]	DeepLab-V2+ResNet101	68.8	69.7
	FPR$_{(\rm CVPR23)}$[37]	DeepLab-V2+ResNet101	70.3	70.1
	ReCAM$_{(\rm CVPR22)}$[33]	DeepLab-V2+ResNet101	71.6	71.4
	EPS$_{(\rm CVPR21)}$[12]	DeepLab-V1+ResNet101	71.0	71.8
	BECO$_{(\rm CVPR23)}$[17]	DeepLab-V3+[38]+ResNet101	72.1	71.8
	RCA$_{(\rm CVPR23)}$[34]	DeepLab-V2+ResNet101	72.2	72.8
	本文方法	DeepLab-V1+ResNet101	72.1	73.1
		DeepLab-V2+ResNet101	**72.5**	**73.9**
基于Transformer方法	BECO$_{(\rm CVPR23)}$[17]	SegFormer-MiT-B2	73.7	73.5
	CoBra$_{(\rm PR25)}$[35]	SegFormer-MiT-B2	74.3	74.2
	本文方法	SegFormer-MiT-B2	**75.3**	**75.9**
		Mask2Former+Swin-L	**76.8**	**77.5**

表 2　在PASCAL VOC 2012验证集上的每类分割结果(%)

方法	背景	飞机	自行车	鸟	船	瓶子	公共汽车	汽车	猫	椅子	牛	餐桌	狗	马	摩托车	人	盆栽植物	羊	沙发	火车	电视/显示器	mIoU
ADELE[42]	91.1	77.6	33.0	88.9	67.1	71.7	88.8	82.5	89.0	26.6	83.8	**44.6**	84.4	77.8	74.8	78.5	43.8	84.8	44.6	56.1	**65.3**	69.3
W-OoD[13]	91.0	80.1	34.1	88.1	64.8	68.3	87.4	**84.4**	**89.8**	30.1	**87.8**	34.7	**87.5**	85.9	79.8	75.0	**56.4**	84.5	**47.8**	80.4	46.4	70.7
EPS[12]	91.7	89.4	40.6	84.7	67.0	71.6	87.8	82.7	87.4	**33.6**	81.9	37.3	82.5	82.9	76.6	**82.8**	54.0	79.7	39.1	**85.4**	51.7	71.0
EPS++[43]	91.9	89.7	**41.7**	82.4	68.4	73.7	**89.5**	80.8	86.9	31.8	86.9	43.0	82.7	**86.6**	**81.1**	77.7	47.8	84.8	41.1	84.1	50.6	71.6
本文方法	**92.2**	89.5	40.1	**89.5**	74.8	**76.2**	87.6	82.2	88.3	31.9	85.8	34.7	85.8	82.3	74.2	82.0	55.6	**86.2**	36.8	83.7	53.9	72.1(V1)
	92.2	**90.4**	40.4	88.2	**76.3**	73.8	88.4	83.0	89.1	33.1	87.0	36.1	86.7	86.2	76.4	81.7	54.2	83.9	38.2	84.8	51.7	**72.5(V2)**

注：粗体表示最优值，下划线表示次优值

VOC 2012验证集上对于21个类的性能比较结果。本文方法与ADELE[42]相比，在大多数类别中性能更优，尤其对于火车类别，提升了28.7%的mIoU值，这表明其有效处理了噪声标签；与W-OoD[13]，EPS[12]和EPS++[43]相比，在飞机、船类别表现最佳，表明其生成了准确的伪标签。图8呈现了本文方法与SIPE[36]，FPR[37]，EPS[12]，RCA[34]方法在VOC验证集上的可视化对比结果，本文方法不仅保留了干净标签信息，还进一步抑制噪声信息(如人的轮廓、狗的耳朵、人手中的瓶子以及火车车身)。此外，由于模型未能充分挖掘细粒度特征而导致的椅子、沙发等局部细节复杂、结构相似性高的类别上分割精度表现不佳的现象，后续工作将深度融合像素级对比学习，在特征空间构建像素对相似性约束，并设计动态阈值调整机制，根据图像上下文语义实时调整各类别分割阈值，提升模型的判别力。

为了进一步表明本文方法的优越性，表3呈现了在MS COCO 2014验证集上的对比结果，本文方法达到了40.8%的mIoU值，与改进CAM的方法(OC-CSE[8])相比，本文方法提升了4.4%的mIoU值，与利用噪声标签学习的方法(MDBA[16])相比，提升了3.0%的mIoU值。与最新的方法URN[44]相比，在验证集上提升了0.1%的mIoU值。此外，本文方法在基于Transformer架构的语义分割网络中达到了46.7%的mIoU值，与先进的方法ECA[45]相比提升了3.8%。由于MS COCO 2014数据集包含丰富的多类图像，如图9所示，与EPS[12]及URN[44]对比结果可知，本文方法可以对噪声进行有效的处理。

3.3　消融实验

3.3.1　多种子信息聚合的有效性

为实现多种子信息聚合达到理想的效果，本文选取8个种子进行实验。表4表示8个种子在PASCAL VOC 2012训练集上的mIoU值，其中以不同的Dropout忽略神经元的概率值p和缩放尺寸为各个种子命名(例如：0.05-1.00表明实验中Dropout概率值p为0.05，缩放尺寸为1.00所获得的种子)。在WSSS任务中，种子源于图像分类任务，由于该任务难度低于图像分割，其在表中呈现出更高的分割精度。为避免引入过多噪声并保证正负混合学习的有效性，将各个种子进行对比并在数据集21个类别中选取mIoU值最大的种子确定为最终种子。如表4所示，本文选中6个种子(0.05-1.00，0.05-1.20，0.10-1.00，0.10-1.20，0.20-1.00，0.30-1.20)，在每类的mIoU值为互补状态。

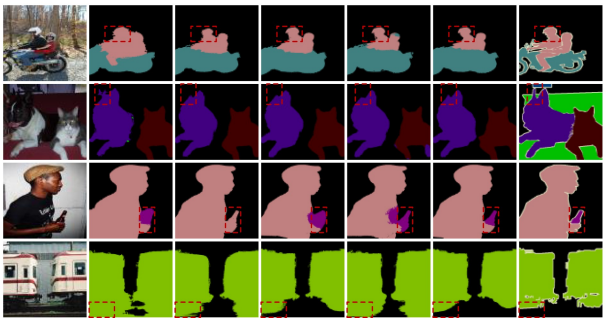


图 8 在VOC验证集上的可视化分割结果

表 3　本文方法与不同模型在MS COCO 2014验证集上的性能对比

模型		骨干网络	mIoU(%) 验证集
基于CNN方法	$\text{EPS}_{(\text{CVPR21})}$[12]	DeepLab-V2+ResNet101	35.7
	$\text{OC-CSE}_{(\text{ICCV21})}$[8]	DeepLab-V1+ResNet38	36.4
	$\text{MDBA}_{(\text{TIP23})}$[16]	DeepLab-V2+ResNet101	37.8
	$\text{SIPE}_{(\text{CVPR22})}$[36]	DeepLab-V2+ResNet101	40.6
	$\text{URN}_{(\text{AAAI22})}$[44]	PSPNet+ResNet101	40.7
	本文方法	DeepLab-V2+ResNet101	**40.8**
基于Transformer方法	$\text{TSCD}_{(\text{AAAI23})}$[46]	SegFormer+MiT-B1	40.1
	$\text{ECA}_{(\text{ACM MM24})}$[45]	SegFormer+MiT-B1	42.9
	本文方法	SegFormer+MiT-B2	**45.1**
		Mask2Former+Swin-L	**46.7**

注：粗体表示最优值

为了验证多个种子聚合方案的有效性，本文随机选取Dropout忽略神经元的概率及尺寸，并添加CRF优化来生成6个不同的种子，将不同种子在相同的分割实验设置下(未应用负学习)分别单独训

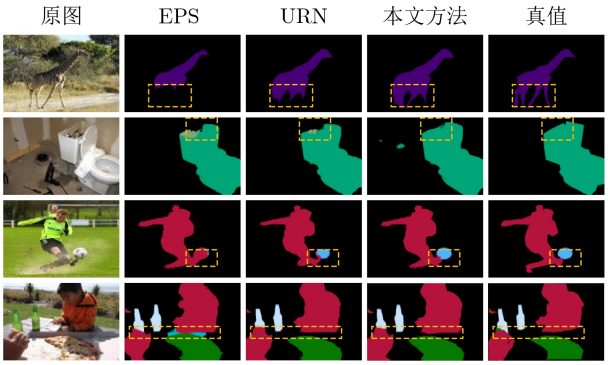


图 9 在COCO验证集上的可视化分割结果

练。为了评估本文聚合种子的效果，表5呈现了选取种子在PASCAL VOC 2012训练集上的类激活图性能，通过改进类激活图，可以有效提升分割网络的性能。如表6所示，通过在VOC 2012验证集上进行测试，表明本文方法利用多种子进行聚合的方案比每个单个种子的效果都有所提升，说明该方案可以利用多种子信息互补的方式提升分割性能，从而展现该方案的有效性。

3.3.2 正负混合学习的有效性

考虑到噪声标签分布的复杂性，通过设定不同阈值控制噪声标签比例，阈值越高表明该种子可信度越高。为验证正负混合学习的有效性，本文设置不同阈值t={3,4,5,6}，将选出的像素区分为干净或噪声标签，得到4组结果。随后，应用正负混合学习对分割网络进行训练。表7所示为在VOC验证集上对不同阈值进行的实验结果：当t=6时，分割效果最佳。通过与未应用正负混合学习的结果表6进行对比，应用后的mIoU值在DeepLabV1条件下可提升0.21%，DeepLabV2条件下可提升0.16%，表明正负混合学习有效降低了噪声标签的影响。本文利用表4所示的8个种子进行实验，结果如表8所示，相较于使用6个种子的实验，8个种子的结果表现略低，此差异源于使用更多种子时引入额外的噪声，影响了实验的性能表现，因此在此实验中选择6个种子更为适宜，能够有效控制噪声并确保更精确的结果。

3.4 统计分析情况

3.4.1 种子标签与真值标签一致情况分析

表9呈现了6个种子及聚合种子与真值标签在同一像素位置类别标签的一致率和不一致率，其中，“0.05-1.20”表示Dropout的概率p为0.05，缩放尺寸为1.20所获得的种子。可以看到，尽管各个种子获得了高质量的标注，但与真值标签相比，仍存在一定差距。如果将不一致率较高的种子使用正学习方式训练分割网络，学习到错误标签的可能性更高，导致模型性能下降。因此，本文利用信息互补的思想，根据多个种子中像素所呈现的标注情况进行优选，以提升与真值标签的一致性。此外，表9与表6的结果相对应，充分证明了本文利用多种子信息聚合的有效性。

3.4.2 干净/噪声标签与真值标签一致情况分析

理想地，与真值标签一致的优选标签标记为干净标签；不一致的优选标签标记为噪声标签，从而有利于发挥正负混合学习的作用，挖掘正确信息的同时抑制错误信息对模型的影响。表10为优选标签在不同阈值下标记为干净情况与真值标签的一致率和不一致率的比较，分为(a)和(b)两种情况，其中对于错误标记(b)，优选的种子标签与真值标签一致，但其被错误标记为噪声标签，因此，应用负学习可以缓解这一问题。通过让网络间接学习到其他可能为正确标签的信息，提升对正确信息的学习能力。表11为优选标签在不同阈值下标记为噪声情况与真值标签的一致率和不一致率的比较，分为(c)和(d)两种情况，其中，对于错误标记(d)，优选的种子标签与真值标签不一致，但其被错误标记为干净标签。通过观察，表11(d)的不一致率均低于表9的不一致率，由此可知，正负混合学习能够有效抑制错误信息的影响。对比表10与表11可知：在不同阈值下，理想标记所占的比例较高，这表明优选的种子标签大部分被正确标记。

4 结论

针对基于图像级标签WSSS任务存在种子标签

表 4 在PASCAL VOC 2012训练集上种子的每类mIoU值(%)

种子	背景	飞机	自行车	鸟	船	瓶子	公共汽车	汽车	猫	椅子	牛	餐桌	狗	马	摩托车	人	盆栽植物	羊	沙发	火车	电视/显示器	mIoU
0.05-1.0	91.4	87.9	**45.9**	85.0	80.6	**75.3**	83.7	81.0	86.0	34.9	83.5	44.4	**85.4**	86.1	**80.8**	80.0	54.7	86.7	41.0	**80.1**	54.9	72.8
0.05-1.2	91.4	87.3	44.5	**87.0**	**80.8**	73.5	83.6	80.7	86.8	34.5	**85.1**	48.7	83.3	**87.0**	79.9	80.7	52.5	85.9	41.9	78.9	55.4	72.8
0.10-1.0	**91.5**	87.3	45.0	86.5	80.2	75.2	**84.3**	**82.0**	85.7	34.5	84.0	43.6	84.7	86.0	**80.8**	80.3	**57.3**	**89.3**	40.6	79.0	55.9	**73.0**
0.10-1.2	**91.5**	**88.0**	41.9	86.2	80.1	74.0	82.9	80.6	**87.7**	34.6	84.9	**49.6**	84.6	86.5	79.9	80.7	54.0	87.3	41.9	78.6	56.5	**73.0**
0.20-1.0	**91.5**	87.0	**45.9**	85.9	80.2	74.1	83.6	80.7	86.6	**35.3**	83.7	46.0	85.1	86.3	80.3	80.0	55.4	86.8	38.4	**80.1**	55.5	72.8
0.20-1.2	**91.5**	87.8	45.1	86.7	79.4	72.5	84.1	**82.0**	87.1	34.2	84.5	47.6	83.9	86.8	79.8	80.8	55.0	86.6	41.1	79.1	55.8	72.9
0.30-1.0	91.3	87.5	41.2	85.5	79.6	73.8	83.7	81.4	86.3	34.0	84.3	44.8	**85.4**	86.8	80.5	79.5	54.7	86.7	40.4	78.0	54.4	72.4
0.30-1.2	**91.5**	87.3	44.4	86.6	78.9	72. 5	83.3	80.8	87.5	33.8	84.2	46.3	83.7	86.9	79.4	**80.9**	56.7	86.7	**42.7**	79.6	**56.9**	72.9

注：粗体表示最优值

表 5　在PASCAL VOC 2012训练集上每个种子类激活图的mIoU值(%)

种子	0.05-1.0	0.05-1.2	0.10-1.0	0.10-1.2	0.2-1.0	0.3-1.2
mIoU	69.15	69.10	69.34	69.30	69.27	69.63

表 6　种子在PASCAL VOC 2012验证集中的mIoU值(%)

Dropout	Scale	Deeplabv1+ResNet101	DeeplabV2+ResNet101
0.05	1.0	71.38	71.92
	1.2	71.82	72.27
0.10	1.0	71.36	72.29
	1.2	71.80	72.27
0.20	1.0	71.06	71.80
0.30	1.2	71.11	72.13
聚合种子	1.0	**71.85**	**72.30**

注：粗体表示最优值

表 7　正负混合学习性能对比(%)

学习方式		DeepLabV1+ResNet101	DeepLabV2+ResNet101
正负学习(不同阈值t)	3	71.87	72.34
	4	71.89	72.37
	5	71.99	72.42
	6	**72.06**	**72.46**

注：粗体表示最优值

表 8　正负混合学习性能对比(%)

学习方式		DeepLabV2+ResNet101
正负学习(不同阈值t)	6	70.96
	7	**71.26**
	8	71.22

注：粗体表示最优值

表 9　六个种子及聚合种子与真值标签在同一像素位置的类别标签一致率和不一致率比较(%)

不同种子	一致率(%)	不一致率(%)
0.05-1.0	89.8000	10.2000
0.05-1.2	89.8000	10.2000
0.10-1.0	89.8711	10.1289
0.10-1.2	89.9034	10.0966
0.20-1.2	89.8732	10.1268
0.30-1.2	89.8284	10.1716
聚合种子	**89.9964**	**10.0036**

注：粗体表示最优值

表 10　优选标签在不同阈值下标记的干净情况与真值标签的一致率和不一致率比较(%)

不同阈值t	优选的种子标签与真值标签一致	
	(a)标记为干净(正确标记)	(b)标记为噪声(错误标记)
3	89.9042	0.0022
4	89.5418	0.3647
5	88.8536	1.0528
6	87.8883	2.0182

表 11　优选标签在不同阈值下标记的噪声情况与真值标签的一致率和不一致率比较(%)

不同阈值t	优选的种子标签与真值标签不一致	
	(c)标记为噪声(正确标记)	(d)标记为干净(错误标记)
3	0.0033	10.0002
4	0.4725	9.6211
5	1.1762	8.9174
6	2.0201	8.0734

不准确问题，从互补信息的角度出发，提出一种基于多种子标签区分缓解该问题。此外，提出正负混合学习方法进一步提升分割性能，减少错误像素标签对分割模型造成的影响。在PASCAL VOC 2012和MS COCO 2014数据集上实现了更优秀的性能，实验结果充分验证了方法的有效性。本文方法虽然还未能达到理想标记情况，但通过标签区分和正负混合学习可以进一步抑制错误信息传递，与仅利用单一种子和正学习相比，实现了更好的效果。然而，该方法仍存在对特征敏感性不足问题，未来将围绕以下方向展开：引入对比学习框架，在大规模无标注数据上预训练特征提取网络，提升模型对特征的迁移能力；设计动态阈值调整机制，对不同特征的图像实时分析，动态调整各类别分割置信度阈值，缓解类别间特征混淆问题。

参 考 文 献

[1] LIU Huan, LI Wei, XIA Xianggen, *et al.* SegHSI: Semantic segmentation of hyperspectral images with limited labeled pixels[J]. *IEEE Transactions on Image Processing*, 2024, 33: 6469–6482. doi: 10.1109/TIP.2024.3492724.

[2] 张印辉, 张金凯, 何自芬, 等. 全局感知与稀疏特征关联图像级弱监督病理图像分割[J]. 电子与信息学报, 2024, 46(9): 3672–3682. doi: 10.11999/JEIT240364.
ZHANG Yinhui, ZHANG Jinkai, HE Zifen, *et al.* Global perception and sparse feature associate image-level weakly supervised pathological image segmentation[J]. *Journal of Electronics & Information Technology*, 2024, 46(9): 3672–3682. doi: 10.11999/JEIT240364.

[3] LI Jiale, DAI Hang, HAN Hao, *et al.* MSeg3D: Multi-modal 3D semantic segmentation for autonomous driving[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 21694–21704. doi: 10.1109/CVPR52729.2023.02078.

[4] 梁燕, 易春霞, 王光宇, 等. 基于多尺度语义编解码网络的遥感图像语义分割[J]. 电子学报, 2023, 51(11): 3199–3214. doi: 10.12263/DZXB.20220503.
LIANG Yan, YI Chunxia, WANG Guangyu, *et al.* Semantic segmentation of remote sensing image based on multi-scale semantic encoder-decoder network[J]. *Acta Electronica Sinica*, 2023, 51(11): 3199–3214. doi: 10.12263/DZXB.20220503.

[5] OH Y, KIM B, and HAM B. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation[C]. IEEE/CVF Conference on Computer

Vision and Pattern Recognition, Nashville, USA, 2021: 6909–6918. doi: 10.1109/CVPR46437.2021.00684.

[6] LIANG Zhiyuan, WANG Tiancai, ZHANG Xiangyu, *et al*. Tree energy loss: Towards sparsely annotated semantic segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 16886–16895. doi: 10.1109/CVPR52688.2022.01640.

[7] ZHAO Yuanhao, SUN Genyun, LING Ziyan, *et al*. Point-based weakly supervised deep learning for semantic segmentation of remote sensing images[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 5638416. doi: 10.1109/TGRS.2024.3409903.

[8] KWEON H, YOON S H, KIM H, *et al*. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation[C]. IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 6974–6983. doi: 10.1109/ICCV48922.2021.00691.

[9] ZHOU Bolei, KHOSLA A, LAPEDRIZA A, *et al*. Learning deep features for discriminative localization[C]. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2921–2929. doi: 10.1109/CVPR.2016.319.

[10] WANG Xiang, YOU Shaodi, LI Xi, *et al*. Weakly-supervised semantic segmentation by iteratively mining common object features[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 1354–1362. doi: 10.1109/CVPR.2018.00147.

[11] WANG Xun, ZHANG Haozhi, HUANG Weilin, *et al*. Cross-batch memory for embedding learning[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 6387–6396. doi: 10.1109/CVPR42600.2020.00642.

[12] LEE S, LEE M, LEE J, *et al*. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 5491–5501. doi: 10.1109/CVPR46437.2021.00545.

[13] LEE J, OH S J, YUN S, *et al*. Weakly supervised semantic segmentation using out-of-distribution data[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 16876–16885. doi: 10.1109/CVPR52688.2022.01639.

[14] CHANG Yuting, WANG Qiaosong, HUNG W C, *et al*. Weakly-supervised semantic segmentation via sub-category exploration[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 8988–8997. doi: 10.1109/CVPR42600.2020.00901.

[15] ARPIT D, JASTRZĘBSKI S, BALLAS N, *et al*. A closer look at memorization in deep networks[C]. 34th International Conference on Machine Learning, Sydney, Australia, 2017: 233–242.

[16] CHEN Tao, YAO Yazhou, and TANG Jinhui. Multi-granularity denoising and bidirectional alignment for weakly supervised semantic segmentation[J]. *IEEE Transactions on Image Processing*, 2023, 32: 2960–2971. doi: 10.1109/TIP.2023.3275913.

[17] RONG Shenghai, TU Bohai, WANG Zilei, *et al*. Boundary-enhanced co-training for weakly supervised semantic segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 19574–19584. doi: 10.1109/CVPR52729.2023.01875.

[18] WU Zifeng, SHEN Chunhua, and VAN DEN HENGEL A. Wider or deeper: Revisiting the ResNet model for visual recognition[J]. *Pattern Recognition*, 2019, 90: 119–133. doi: 10.1016/j.patcog.2019.01.006.

[19] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, *et al*. Dropout: A simple way to prevent neural networks from overfitting[J]. *The Journal of Machine Learning Research*, 2014, 15(1): 1929–1958.

[20] LI Mingjia, XIE Binhui, LI Shuang, *et al*. VBLC: Visibility boosting and logit-constraint learning for domain adaptive semantic segmentation under adverse conditions[C]. 37th AAAI Conference on Artificial Intelligence, Washington, USA, 2023: 8605–8613. doi: 10.1609/aaai.v37i7.26036.

[21] YANG Guoqing, ZHU Chuang, and ZHANG Yu. A self-training framework based on multi-scale attention fusion for weakly supervised semantic segmentation[C]. IEEE International Conference on Multimedia and Expo, Brisbane, Australia, 2023: 876–881. doi: 10.1109/ICME55011.2023.00155.

[22] KRÄHENBÜHL P and KOLTUN V. Efficient inference in fully connected CRFs with gaussian edge potentials[C]. 25th International Conference on Neural Information Processing Systems, Granada, Spain, 2011: 109–117.

[23] LEE J, KIM E, LEE S, *et al*. FickleNet: Weakly and semi-supervised semantic image segmentation using stochastic inference[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 5262–5271. doi: 10.1109/CVPR.2019.00541.

[24] KIM Y, YIM J, YUN J, *et al*. NLNL: Negative learning for noisy labels[C]. IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 2019: 101–110. doi: 10.1109/ICCV.2019.00019.

[25] KIM Y, YUN J, SHON H, *et al*. Joint negative and positive learning for noisy labels[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 9437–9446. doi: 10.1109/CVPR46437.2021.00932.

[26] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, *et al*. The PASCAL visual object classes (VOC) challenge[J]. *International Journal of Computer Vision*, 2010, 88(2): 303–338. doi: 10.1007/s11263-009-0275-4.

[27] LIN T Y, MAIRE M, BELONGIE S, *et al*. Microsoft COCO: Common objects in context[C]. 13th European Conference on Computer Vision, Zurich, Switzerland, 2014: 740–755. doi: 10.1007/978-3-319-10602-1_48.

[28] HARIHARAN B, ARBELÁEZ P, BOURDEV L, *et al*. Semantic contours from inverse detectors[C]. International Conference on Computer Vision, Barcelon, Spain, 2011: 991–998. doi: 10.1109/ICCV.2011.6126343.

[29] SHELHAMER E, LONG J, and DARRELL T. Fully convolutional networks for semantic segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(4): 640–651. doi: 10.1109/TPAMI.2016.2572683.

[30] DENG Jia, DONG Wei, SOCHER R, *et al*. ImageNet: A large-scale hierarchical image database[C]. IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 248–255. doi: 10.1109/CVPR.2009.5206848.

[31] CHEN L C, PAPANDREOU G, KOKKINOS I, *et al.* Semantic image segmentation with deep convolutional nets and fully connected CRFs[C]. 3rd International Conference on Learning Representations, San Diego, USA, 2015: 24–37.

[32] CHEN L C, PAPANDREOU G, KOKKINOS I, *et al.* DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 40(4): 834–848. doi: 10.1109/TPAMI.2017.2699184.

[33] CHEN Zhaozheng, WANG Tan, WU Xiongwei, *et al.* Class re-activation maps for weakly-supervised semantic segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 959–968. doi: 10.1109/CVPR52688.2022.00104.

[34] ZHOU Tianfei, ZHANG Meijie, ZHAO Fang, *et al.* Regional semantic contrast and aggregation for weakly supervised semantic segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 4289–4299. doi: 10.1109/CVPR52688.2022.00426.

[35] HAN W, KANG S, CHOO K, *et al.* Complementary branch fusing class and semantic knowledge for robust weakly supervised semantic segmentation[J]. *Pattern Recognition*, 2025, 157: 110922. doi: 10.1016/j.patcog.2024.110922.

[36] CHEN Qi, YANG Lingxiao, LAI Jianhuang, *et al.* Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 4278–4288. doi: 10.1109/CVPR52688.2022.00425.

[37] CHEN Liyi, LEI Chenyang, LI Ruihuang, *et al.* FPR: False positive rectification for weakly supervised semantic segmentation[C]. IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 1108–1118. doi: 10.1109/ICCV51070.2023.00108.

[38] CHEN L C, ZHU Yukun, PAPANDREOU G, *et al.* Encoder-decoder with atrous separable convolution for semantic image segmentation[C]. 15th European Conference on Computer Vision, Munich, Germany, 2018: 833–851. doi: 10.1007/978-3-030-01234-2_49.

[39] XIE Enze, WANG Wenhai, YU Zhiding, *et al.* SegFormer: Simple and efficient design for semantic segmentation with transformers[C]. 35th International Conference on Neural Information Processing Systems, 2021: 924.

[40] LIU Ze, LIN Yutong, CAO Yue, *et al.* Swin transformer: Hierarchical vision transformer using shifted windows[C]. IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 9992–10002. doi: 10.1109/ICCV48922.2021.00986.

[41] CHENG Bowen, MISRA I, SCHWING A G, *et al.* Masked-attention mask transformer for universal image segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 1280–1289. doi: 10.1109/CVPR52688.2022.00135.

[42] LIU Sheng, LIU Kangning, ZHU Weicheng, *et al.* Adaptive early-learning correction for segmentation from noisy annotations[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 2596–2606. doi: 10.1109/CVPR52688.2022.00263.

[43] LEE M, LEE S, LEE J, *et al.* Saliency as pseudo-pixel supervision for weakly and semi-supervised semantic segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(10): 12341–12357. doi: 10.1109/TPAMI.2023.3273592.

[44] LI Yi, DUAN Yiqun, KUANG Zhanghui, *et al.* Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation[C]. 36th AAAI Conference on Artificial Intelligence, Palo Alto, 2022: 1447–1455. doi: 10.1609/aaai.v36i2.20034.

[45] WU Yuanchen, LI Xiaoqiang, LI Jide, *et al.* DINO is also a semantic guider: Exploiting class-aware affinity for weakly supervised semantic segmentation[C]. 32nd ACM International Conference on Multimedia, Melbourne, Australia, 2024: 1389–1397. doi: 10.1145/3664647.3681710.

[46] XU Rongtao, WANG Changwei, SUN Jiaxi, *et al.* Self correspondence distillation for end-to-end weakly-supervised semantic segmentation[C]. 37th AAAI Conference on Artificial Intelligence, Washington, USA, 2023: 3045–3053. doi: 10.1609/aaai.v37i3.25408.

桑　雨：男，副教授，研究方向为人工智能、视觉感知.
刘　通：女，硕士生，研究方向为人工智能、弱监督图像分割.
马天娇：女，硕士，研究方向为人工智能、弱监督图像分割.
李　乐：男，硕士生，研究方向为深度学习、图像语义分割.
李思漫：女，硕士生，研究方向为深度学习、目标检测.
刘宇男：男，博士后研究员，研究方向为模式识别、计算机视觉.

责任编辑：陈　倩

Weakly Supervised Image Semantic Segmentation Based on Multi-Seeded Information Aggregation and Positive-Negative Hybrid Learning

SANG Yu[①②]　LIU Tong[①]　MA Tianjiao[①]　LI Le[①]　LI Siman[①]　LIU Yunan[③]

[①](*College of Electronics and Information Engineering*, *Liaoning Technical University*, *Huludao* 125105, *China*)

[②](*School of Mathematical Sciences*, *Bohai University*, *Jinzhou* 121013, *China*)

[③](*College of Artificial Intelligence*, *Dalian Maritime University*, *Dalian* 116026, *China*)

Abstract:

Objective　The rapid development of deep learning techniques, particularly Convolutional Neural Networks (CNN), has led to notable advances in semantic segmentation, enabling applications in medical imaging,

autonomous driving, and remote sensing. However, conventional semantic segmentation tasks typically rely on large numbers of pixel-level annotated images, which is both time-consuming and expensive. To address this limitation, Weakly Supervised Semantic Segmentation (WSSS) using image-level labels has emerged as a promising alternative. This approach aims to reduce annotation costs while maintaining or enhancing segmentation performance, thus supporting broader adoption of semantic segmentation techniques. Most existing methods focus on optimizing Class Activation Mapping (CAM) to generate high-quality seed regions, with further refinement through post-processing. However, the resulting seed labels often contain varying degrees of noise. To mitigate the effect of noisy labels on the segmentation network and to efficiently extract accurate information by leveraging multiple complementary seed sources, this study proposes a weakly supervised semantic segmentation method based on multi-seed information aggregation and positive-negative hybrid learning. The proposed approach improves segmentation performance by integrating complementary information from different seeds while reducing noise interference.

Methods Building on the idea that combining multiple seeds can effectively extract accurate information, this study proposes a weakly supervised image semantic segmentation method based on multi-seed information aggregation and positive-negative hybrid learning. The approach employs a generalized classification network to generate diverse seed regions by varying the input image scale and modifying the Dropout layer to randomly deactivate neurons with different probabilities. This process enables the extraction of complementary information from multiple sources. Subsequently, a semantic segmentation network is trained using a hybrid positive-negative learning strategy based on the category labels assigned to each pixel across these seeds. Clean labels, identified with high confidence, guide the segmentation network through a positive learning process, where the model learns that "the input image belongs to its assigned labels." Conversely, noisy labels are addressed using two complementary strategies. Labels determined as incorrect are trained under the principle that "the input image does not belong to its assigned labels," representing a form of positive learning for error suppression. Additionally, an indirect negative learning strategy is applied, whereby the network learns that "the input image does not belong to its complementary labels,"further mitigating the influence of noisy labels. To reduce the adverse effects of noisy labels, particularly the tendency of conventional cross-entropy loss to assign higher prediction confidence to such labels, a prediction constraint loss is introduced. This loss function enhances the model's predictive accuracy for reliable labels while reducing overfitting to incorrect labels. The overall framework effectively suppresses noise interference and improves the segmentation network's performance.

Results and Discussions The proposed weakly supervised image semantic segmentation method based on multi-seed information aggregation and positive-negative hybrid learning generates diverse seeds by randomly varying the Dropout probability and input image scale, with Conditional Random Field (CRF) optimization applied to further refine seed quality. To limit noise introduction while maintaining the effectiveness of positive-negative hybrid learning, six complementary seeds are selected (Table 5). The integration of multi-source information from these seeds enhances segmentation performance, as demonstrated in (Table 7) . Pixel labels within these seeds are classified as clean or noisy based on a defined confidence threshold. The segmentation network is subsequently trained using a positive-negative hybrid learning strategy, which suppresses the influence of noisy labels and improves segmentation accuracy. Experimental results confirm that positive-negative hybrid learning effectively reduces label noise and enhances segmentation performance (Table 8). The proposed method was validated on the PASCAL VOC 2012 and MS COCO 2014 datasets. With a CNN-based segmentation network, the mean Intersection over Union (mIoU) reached 72.5% and 40.8%, respectively. When using a Transformer-based segmentation network, the mIoU improved to 76.8% and 46.7% (Table 1, Table 3). These results demonstrate that the proposed method effectively enhances segmentation accuracy while controlling the influence of noisy labels.

Conclusions This study addresses the challenge of inaccurate seed labels in WSSS based on image-level annotations by proposing a multi-seed label differentiation strategy that leverages complementary information to improve seed quality. In addition, a positive-negative hybrid learning approach is introduced to enhance segmentation performance and mitigate the influence of erroneous pixel labels on the segmentation model. The proposed method achieves competitive results on the PASCAL VOC 2012 and MS COCO 2014 datasets. Specifically, the mIoU reaches 72.5% and 40.8%, respectively, using a CNN-based segmentation network. With a Transformer-based segmentation network, the mIoU further improves to 76.8% and 46.7%. These results demonstrate the effectiveness of the proposed method in improving segmentation accuracy while reducing noise interference. Although the method does not yet achieve ideal label precision, label differentiation combined with positive-negative hybrid learning effectively suppresses misinformation propagation and outperforms approaches based on single-seed generation and positive learning alone.

Key words: Weakly supervised semantic segmentation; Class Activation Mapping (CAM); Multi-seed strategy; Cross-Entropy; Positive-Negative learning