计算机科学
COMPUTER SCIENCE

结合区域采样和类间损失的人体解析模型

李杨, 韩屏

引用本文

李杨，韩屏. 结合区域采样和类间损失的人体解析模型[J]. 计算机科学, 2023, 50(4): 103-109.

LI Yang, HAN Ping. Human Parsing Model Combined with Regional Sampling and Inter-class Loss[J]. Computer Science, 2023, 50(4): 103-109.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220100259

结合区域采样和类间损失的人体解析模型

李 杨 韩 屏
武汉理工大学信息工程学院 武汉 430070
(yang_li314@163.com)

摘 要 人体解析是一项细粒度级别的语义分割任务,随着人体解析数据集中标注类别的精细化,人体解析数据集呈长尾分布,导致对相似类别的识别难度不断增大。均衡采样是解决长尾分布问题的有效方法。针对人体解析任务中难以对标注目标进行均衡采样和模型对相似类别的误判率增加等问题,文中提出了一种结合区域采样和类间损失的人体解析模型,该模型包含语义分割网络、区域均衡采样模块(Regionally Balanced Sampling Module,RBSM)和类间损失模块(Inter-class Loss Module,ILM)3个部分。首先将待解析图片送入语义分割网络得到初步预测结果,RBSM对初步的预测结果和真实标签进行采样,对采样后的预测结果和真实标签计算主损失;同时提取出语义分割网络的最后一层特征热图与真实标签,并将其送入ILM计算类间损失,让模型同时优化主损失和类间损失,最终得到精度更高的模型。在MHPv2.0数据集上的实验结果表明,该模型在不更改原有语义分割网络结构的基础上将mIoU评测指标提高了1.3%以上,有效缓解了长尾分布和类间的相似性给人体解析带来的影响。

关键词:区域采样;类间损失;长尾分布;人体解析;语义分割
中图法分类号 TP391

Human Parsing Model Combined with Regional Sampling and Inter-class Loss

LI Yang and HAN Ping
School of Information Engineering,Wuhan University of Technology,Wuhan 430070,China

Abstract Human parsing is a fine-grained level semantic segmentation task. The refinement of annotated categories in the human parsing dataset makes the dataset follow a long-tailed distribution and improves the difficulty of identifying similar categories. Balanced sampling is an efficient way to solve long-tailed distribution problem,but it's difficult to achieve balanced sampling of the labeled object in human parsing. On the other hand,the fine-grained annotation will make the model misjudge similar categories. In response to these problems,a human parsing model combined with regional sampling and inter-class loss is proposed. The model consists of the semantic segmentation network,regionally balanced sampling module(RBSM),and inter-class loss module(ILM). Firstly,the images are parsed by the semantic segmentation network. Next,the parsing results and the ground truth labels are sampled by regionally balanced sampling module. Then the sampled parsing results and sampled ground truth labels are utilized to calculate the master loss. Meanwhile,the inter-class loss between the heatmap features coming from the semantic segmentation network and ground truth labels are calculated in the inter-class loss module,and the master loss and the inter-class loss are optimized at the same time to get a more accurate model. Experimental results based on the MHPv2.0 dataset show that the mIoU of the proposed model improves by more than 1.3% without changing the structure of the semantic segmentation network. The algorithm effectively reduces the impact of the long tail distribution problem and similarity among categories.

Keywords Regional sampling,Inter-loss,Long-tailed distribution,Human parsing,Semantic segmentation

1 引言

人体解析是一项语义分割任务,旨在对人体图像进行逐像素识别,将每一个像素点都归于相应的类别,如头发、手臂、上衣等,最终这些像素点汇聚在一起形成一幅人体解析图。人体解析有助于理解图像中人体各个部位的语义信息,而这些信息在人物动作分析、虚拟试衣、行人重识别等领域都有重要意义。Long等[1]提出的全卷积神经网络在语义分割上获得了巨大成功,出现了各种基于FCN编码器-解码器范式的优秀人体解析框架,如Liang等[2]利用人体关键点提出JPPNet,Ruan等[3]利用人体边缘信息设计了CE2P人体解析框架。除了各种网络结构的创新,Li等[4]发现数据集中的标签

到稿日期:2022-01-27 返修日期:2022-06-23
基金项目:中央高校基础研究基金(WUT:2018III069GX)
This work was supported by the Fundamental Research Funds for the Central Universities(WUT:2018III069GX).
通信作者:韩屏(hanping@whut.edu.cn)

的噪声对模型的解析能力造成了很大的影响，其提出的SCHP将真实标签也放入模型进行迭代优化，形成更具鲁棒性的标签和模型，提高了模型的精度。Liu等[5]利用人体各个部位的位置分布具有唯一性这一特点，提出了CDGNet，利用优化目标从人体解析转移到类别的位置预测，再通过位置预测引导获取更加精确的解析结果。

尽管人体解析任务已经获得极大的发展，但当前的解析任务仅仅停留在粗略的类别分类，如ATR[6]，LIP[7]，CIHP[8]等大型人体解析数据集虽然在标注数量和标注质量上具有极大的优势，但在标注类别上还稍有欠缺。为了让机器视觉更好地理解图像中目标人物的状态信息，细化解析是有必要的。文献[9]提出的Fashionista数据集包含了56个分类，训练集包含了456张，测试集包含了229张；文献[10]提供的MHPv2.0数据集包含了25403张，类别解析标注有59类。相比Fashionista数据集，MHPv2.0提供了更大规模的样本数量，能让网络学到更加多样的类别特征。类别数量的增加也会让模型学习到更多的类别信息，以更好地挖掘和理解图像中的内容。但这也会带来相应的问题：1）随着标注类别数量的增多，大型数据集的分布不再均衡，出现了长尾分布的情况。均衡采样是解决长尾分布问题的有效手段，但语义分割领域中的标签是像素级标注，一张图片中往往存在多个类别，如何对像素级标注的标签和图片进行均衡采样成为一项挑战。2）服饰之间存在相似性，有许多极其相似的类别，如T恤和Polo衫这两种类别，其形状、穿戴位置都大体相似，模型对这些相似类别的误判率较高。

针对以上问题，本文提出了一种基于区域采样和类间损失的人体解析模型（Regional Sampling and Inter-loss Model）。不同于传统的均衡采样，此模型中的区域均衡采样模块（Regionally Balanced Sampling Module，RBSM）将采样对象从整张图片转向图片中类别所在的区域，通过改变各个类别对应的采样概率来达到均衡采样的目的，降低了长尾数据分布对语义分割模型的影响。同时模型中的类间损失模块（Inter-class Loss Module，ILM）可以度量正确类别和错误类别的相似程度，从而引导模型增大正确类别和错误类别在特征空间中的距离，使模型输出更具区分度的结果，从而提高模型对相似类别的解析能力。

2 相关工作

2.1 长尾分布

长尾分布的数据集会让模型过度拟合占大部分样本数量的头部类别，导致模型对少部分的尾部类别拟合程度不够，这会使正确率急剧下降。在人体解析任务中，随着类别数量的增加，长尾分布带来的影响尤为明显。

目前缓解长尾分布的主流方法主要有3种。第一种是通过重采样的方式让每个类别的数量尽量一致，对样本多的类别欠采样，对样本少的类别过采样，从而达到类别均衡采样的目的。第二种是通过更改损失函数来改变不同类别的数据对模型的影响程度，如Cui等[11]设计的损失函数依照样本数量来确定相应类别在损失函数中的权重，从而提升模型在尾部类别的泛化能力。Lin等[12]提出的Focal Loss通过增大难分类样本的权重，降低易分类样本的权重，来增强模型在所有类别上的泛化能力。第三种是迁移学习，如Liu等[13]将模型在头部样本数量多的类别数据中学习到的深层特征迁移到尾部样本数量少的类别，使模型在尾部类别上构建更加完整的特征空间，从而提高尾部类别的正确率。

在语义分割领域中，解决长尾分布问题的主流手段是改变损失函数的形式。如Seesaw Loss[14]通过减弱尾部类别的负梯度来缓解尾部类别正负样本梯度失衡的问题；Loss Max-Pooling[15]通过设计像素权重函数自适应地根据像素产生的损失值来进行重加权。文献[16-17]的研究发现，长尾分布的分类任务最佳组合是先通过交叉熵函数和原始数据分布训练出特征提取网络，再通过重采样的数据来训练特征分类器。在涉及长尾分布数据的语义分割任务中，如果采用这种训练范式，在第二阶段的训练中重采样是必不可少的。但不同于图像分类任务，语义分割任务中一张图片存在多个类别，并且类别间存在耦合关系，如包含首饰和特定样式的服装等尾部类别的图片中同样也包含人脸、肢体等头部类别，仅通过对整个图像进行重采样难以平衡所有类别出现的频率，因此在语义分割任务中实现均衡采样是一项巨大的挑战。

2.2 损失函数

人体解析是语义分割领域中细粒度要求较高的分割任务，随着语义分割技术的不断发展，物体分割的准确度已经达到非常高的水平，因此细粒度分割的难点就落在了区分相似像素点上。损失函数是解决具有较小类间差距问题的有效方法。Focal Loss[12]在交叉熵的基础上给予难分类样本更多的权重，让模型将关注点放在难分类样本的学习上，从而让模型在少样本类别上表现得更好。针对人脸与人脸之间具有极高相似性的特点，Center Loss[18]通过计算每个类的特征中心，将模型输出特征与各类别特征中心的平方差作为优化目标，并不断迭代各个类别的特征中心来扩大不同类别在特征空间中的距离，从而提高正确率。人体解析过程中的服饰同样具有高度相似性，而语义分割网络中不存在全连接层，因此难以提取出长度固定的特征，而且由于像素点数量巨大，如果不断迭代特征中心，无疑会带来极大的计算量。

3 结合区域采样和类间损失的人体解析模型

本文提出的RIM模型的结构如图1所示。

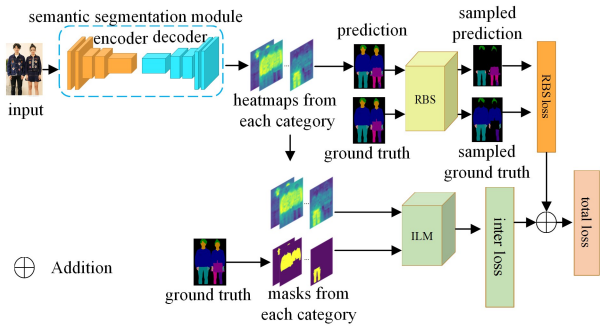


图1 RIM结构图

Fig. 1 Structure of RIM

3.1 区域均衡采样模块

均衡采样是解决长尾分布问题的有效方法，但在语义分割中，类别与类别之间存在耦合，某些类别往往成对出现，难以实现均衡采样。因此本文将采样目标从整张图片细化到图片

中的具体类别区域，并称之为区域均衡采样，只有被采样的类别区域才会纳入网络模型的损失函数中，用于计算损失值。

$$L_{\mathrm{RBS}}=\sum L(\hat{y}_s,y_s) \tag{1}$$

其中，$\hat{y}_s$ 和 y_s 分别是被采样类别区域中的像素预测值和标签值，L 是网络的损失函数，L_{RBS} 是经过区域均衡采样后的损失函数。

每个类别都有对应的采样概率，类别的采样概率决定了该类别被采样的频次，采样概率高表现为对该类别过采样，采样概率低表现为对该类别欠采样。假设数据集的数据分布为 $X=x_i,i\in 1\cdots,n$，其中 n 是数据集的类别数，各个类别的采样概率分布为 $S=s_i,i\in 1,\cdots,n$，采样后的数据分布为 $Y=y_i,i\in 1,\cdots,n$，其中：

$$Y=X*S \tag{2}$$

均匀分布能有效缓解长尾分布数据引起的样本不平衡问题，因此期望采样后的数据分布符合均匀分布，故 $y_i=\eta,\forall i\in 1,\cdots,n$，其中 η 为常数，则对应的采样概率分布为：

$$s_i=\frac{y_i}{x_i}=\frac{\eta}{x_i} \tag{3}$$

若想得到式(3)中各个类别的采样概率 s_i，需要统计整个训练数据集的数据分布 X，本文提出了一种基于准确率分布的区域均衡采样(Regionally Balanced Sampling Based on Accuracy Distribution，ARBS)。该采样方法用模型前期在各类别上的类别像素准确率(Class Pixel Accuracy，CPA)来替代数据集中对应各类别的真实数据分布 X，并据此得到采样概率分布 S，从而确定各个类别的采样方式。类别像素准确率 CPA 的定义如下：

$$CPA(i)=\frac{\hat{Y}_i}{Y_i} \tag{4}$$

其中，$CPA(i)$ 是样本图片中第 i 种类别像素准确率，$\hat{Y}_i$ 是样本图片中正确分类为第 i 类的像素点数量，Y_i 是样本图片中被模型预测为第 i 类的像素点数量。

以包含了 59 种类别的 MHPv2.0 数据集为例，图 2 中的绿色曲线为数据集的真实样本数量。其中，A,B,C,D 分别代表左手、右手、连衣裙和球 4 个类别。A,B 类别样本数量大，属于头部类别；C,D 类别样本数量较少，属于尾部类别。图 2 中的黄色曲线为经过了 3 个 epoch 训练后的 CPA 分布。C,D 两个类别虽然属于尾部类别，但对应的 CPA 却较高；与之相对的，A,B 属于头部类别，但 CPA 却较低。这是左右手类别的相似性导致了头部类别的 CPA 低于预期。

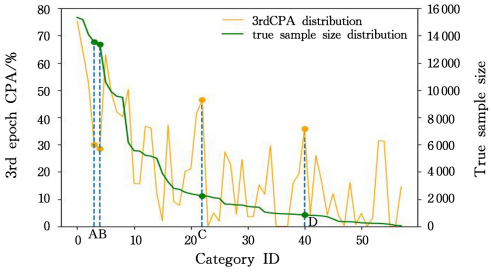


图 2　第 3 个 epoch 的 CPA 分布和真实样本数量分布
(电子版为彩图)

Fig. 2　CPA distribution and true sample size distribution of the third epoch

针对上述现象，需要给予更多的头部类别样本让模型学习，使模型能够区分这些具有相似性头部类别，而不是仅仅依靠初始的样本数量来决定该类别是过采样还是欠采样。图像分类领域的均衡采样的采样分布只由样本数量决定，若将其方法迁移到语义分割上，本文称之为 NRBS(Regionally Balanced Sampling Based on Number-distribution)，会导致需要过采样的头部类别被欠采样，如前面提到的 A,B 两点所代表的左手和右手类别。

通过大量实验验证发现，训练前期的 CPA 分布一方面与真实训练数据集的数据分布趋势总体一致，另一方面能够反映类别之间的相似性。因此，本文提出以训练数据的 CPA 分布代替真实分布的 ARBS，避免 NRBS 采样的不足。

RIM 中的区域均衡采样模块结构如图 3 所示，通过各类别的 CPA 分布和式(3)计算得到各个类别的采样概率，采样器依据不同类别的采样概率分布对各类别进行采样，记录此次采样器采样的结果并根据采样到的类别生成一个类别集合 C，基于类别集合 C 生成相应的采样掩码，记为 mask：

$$mask(x,y)=\begin{cases}1, & G(x,y)\in C\\ 0, & G(x,y)\notin C\end{cases} \tag{5}$$

真实标签为 $G\in\mathbb{R}^{W\times H}$，再将预测结果和真实标签分别与采样掩码 mask 作点乘运算，就得到了经过区域均衡采样后的预测结果和真实标签，最后将其送入损失函数计算损失值。

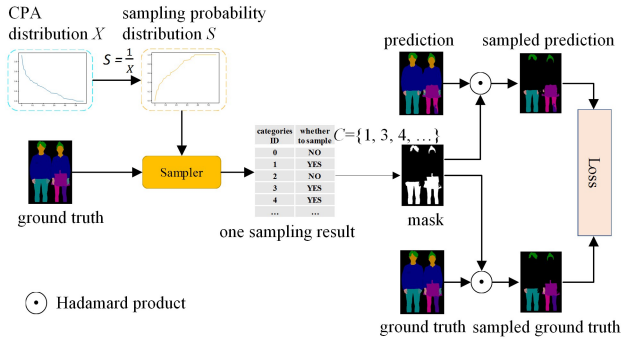


图 3　区域均衡采样模块结构

Fig. 3　Structure of regional balance sampling module

3.2　语义分割模块

语义分割的本质是一种像素级别的分类任务，分割网络的结构可以抽象为编码器-解码器结构，输入图像经过编码器提取颜色、纹理、边缘等低级特征，再经过下采样逐渐生成高级的语义特征；解码器对高级的语义特征信息进行融合分析进而推断出像素的类别，最后通过上采样将分辨率还原至输入图片。本文设计的 RIM 建立在编码器-解码器范式的语义分割网络之上，因此该模型适用于绝大多数的语义分割网络，可以将常用的语义分割网络直接作为 RIM 的语义分割模块，如 U-Net[19]、PSPNet[20]、DeepLab 系列[21-22] 和 DANet[23] 等，不需要更改其网络结构。

3.3　类间损失模块

随着数据集类别的增多和细化，类与类之间的差距越来越小，本文提出的类间损失模块的作用是加大各个类别之间的差距。假设给定一张输入图片 $I\in\mathbb{R}^{3\times W\times H}$，经过语义分割模块后输出的热图特征 i,i 为类别数，V_i 是第 i 类的热图特征，输入图片对应的真实标签为 i，第 i 个类别的区域掩码为

$M_i, M_i \in \mathbb{R}^{W\times H}$，我们通过定义 L_i 来衡量第 i 类与其他类之间的热图误差：

$$L_i = \sum_{j=1}^{n} \delta(M_i \odot (V_j + margin - V_i)) \tag{6}$$

其中，δ 为激活函数，本文方法采用 ReLU 函数，只有当 $V_j - V_i > 0$ 时类间损失才会反向传递误差，同时 ReLU 函数还能避免前者差值过大导致的梯度弥散问题。为了增加类别之间的差距，避免相似类别在 $V_j - V_i = 0$ 时就停止优化，类间损失额外添加了一个 $margin \in \mathbb{R}^{W\times H}$ 条件来迫使不同类别的热图特征的区别更加明显，L_i 的计算示意图如图 4 所示。

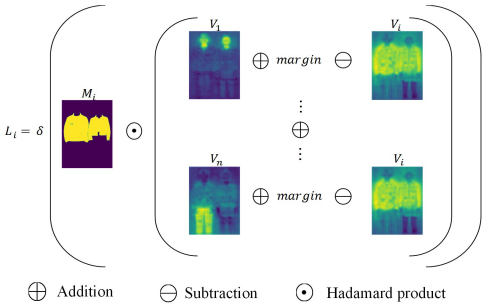


图 4 类间损失计算示意图

Fig. 4 Schematic diagram of inter-class loss

将每个类别的类间热图误差求和就构成了类间损失函数，记为 L_{inter}。

$$L_{\text{inter}} = \sum_{i=1}^{n} L_i \tag{7}$$

最终的损失函数 L_{total} 为：

$$L_{\text{total}} = L_{\text{RBS}} + \lambda L_{\text{inter}} \tag{8}$$

其中，λ 为超参数，用于控制类间损失在总损失函数中的比重，L_{RBS} 是经过区域均衡采样后的主损失函数，主损失函数的结构是任意的，如交叉熵函数、Focal Loss 等。

4 实验结果与分析

4.1 数据集和评估指标

实验采用 MHPv2.0[7] 数据集评估本文提出的方法，MHPv2.0 数据集包含了 25 403 张图片，其中有 15 403 张图片作为训练集，5 000 张图片作为验证集。实验中将图片里相同类别的不同实例合并成一个整体作为语义分割任务的标签。

实验中主要使用了 3 个评估指标，即像素准确率(PA)、类别平均准确率(MPA)和平均交并比(mIoU)。像素准确率反映了正确预测的像素点占总像素点的比例。

$$PA = \frac{\hat{Y}_{\text{correct}}}{Y_{\text{total}}} \tag{9}$$

其中，$\hat{Y}_{\text{correct}}$ 是被正确分类的区域，Y_{total} 是整个预测区域。

式(4)中的 CPA 是对单一类别识别精确率的衡量指标，而类别平均准确率 MPA 反映了模型对所有类别的综合识别能力：

$$MPA = \frac{\sum_{i=1}^{n} CPA(i)}{n} \tag{10}$$

平均交并比 $mIoU$ 作为语义分割的标准度量指标，反映了模型总体的解析能力：

$$mIoU = \frac{\sum_{i=1}^{n} IoU(i)}{n} \tag{11}$$

4.2 实验环境

本文实验选用 Intel Core i9-10900K CPU@3.70 GHz，GPU 3090，64 GB RAM 的硬件平台，操作系统为 Ubuntu 18.04。实验中所有的语义分割模块都选择 ResNet-50 作为编码器网络，并使用了 ImageNet[24] 的预训练权重。模型输入使用的分辨率为 400×400，batchsize 尺寸为 16，优化器为 SGD，初始学习率为 0.01，学习率策略为 CosineAnnealing。训练时采用的图像增强方法包括图像旋转(−10°～10°)、图像剪裁等。

整个训练过程分成两个阶段：第一阶段采用原有的数据分布进行训练，重在训练编码器网络；第二阶段为冻结编码器网络，采用区域均衡采样，并调用类间损失模块，类间损失模块的 $margin=1$，重在训练解码器网络。区域均衡采样模块按照 CPA 分布将语义分割网络的预测结果和真实标签均衡采样后送入主损失函数计算主损失，并将语义分割网络最后一层的热图特征和真实标签送入类间损失模块计算类间损失，从而形成模型的总体损失，最后采用 SGD 优化器优化损失目标。

4.3 实验结果与分析

4.3.1 区域均衡采样方式对比

为了探究 ARBS 和 NRBS 的效果差异，实验中首先删除了类间损失模块，仅使用区域均衡采样模块和语义分割模块，用 PSPNet 作为语义分割模块，实验中分别做了利用 PSPNet 训练 100 个 epoch、利用 PSPNet 训练 70 个 epoch＋利用 NRBS 训练 30 个 epoch，以及利用 PSPNet 训练 70 个 epoch＋利用 ARBS 训练 30 个 epoch 的 3 组实验。

表 1 依据不同数据分布的均衡采样方法对比

Table 1 Comparison of balanced sampling methods based on different data distributions

Method	*PA*	*MPA*	*mIoU*
PSPNet	71.84	45.18	35.92
PSPNet＋NRBS	72.02	**53.06**	36.77
PSPNet＋ARBS	**72.69**	52.22	**37.57**

从表 1 可以看到，用 ARBS 可以在 PA 和 mIoU 指标上获得更多的增益，而用 NRBS 可以在 MPA 指标上获得更多增益。但 mIoU 指标更能衡量模型的分割性能，而在这项指标上，ARBS 相比 NRBS 额外提升了 0.8%。其原因是 ARBS 纠正了 NRBS 在部分头部类别上的错误采样方式。由于部分头部类别样式繁多、类别体积小、与其他类别高度相似，因此其分类难度较大。尽管这些类别的样本数量较多，但仍然需要模型对这些头部类别进行过采样。样本数量是 NRBS 选择采样方式的唯一标准，因此 NRBS 会对这些难以分类的头部类别进行欠采样。由图 2 可知，这些难分类的头部类别的 CPA 在训练初期都处于较低的水平，因此 ARBS 仍然会对这些类别过采样，从而让模型利用更多的样本学习到更加具备鲁棒性的特征。图 5 的结果佐证了本文的分析。从图 5 中可以看出，ARBS 相比 NRBS 提升较大的几个头部类别都具有难分类的特性，如 caphat，cases，wallet，wristband，glove 等样式众多且体积较小，right-boot，right-arm，right-hand 等类别具有对称性，极易与对应的镜像类别混淆。

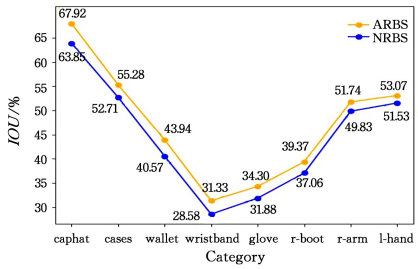
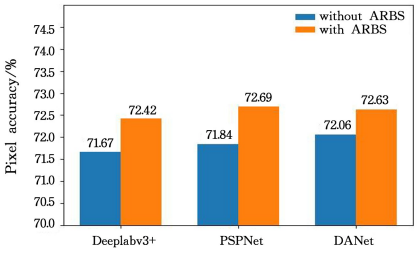


图 5　ARBS 和 NRBS 在 MHPv2.0 验证集中部分类别的 IoU 对比

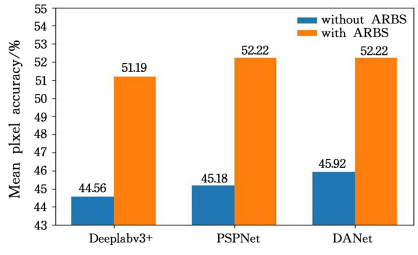
Fig. 5　Comparison of IoU between ARBS and NRBS in some categories on MHPv2.0 validation set

4.3.2　区域均衡采样模块的有效性分析

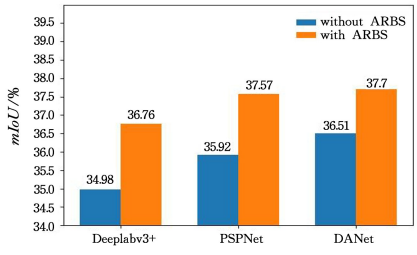
为了突出区域均衡采样模块的有效性，实验采用性能更好的 ARBS 作为区域均衡采样模块，并将其与多个不同的语义分割网络组成语义分割模型，最后将其与未添加 ARBS 的模型进行对比实验。为了保证公平性，实验中并未添加 ARBS 的对照组正常训练 100 个 epoch，添加了 ARBS 的实验组，采用对照组第 70 个 epoch 的训练模型继续训练 30 个 epoch。从图 6 中可以看出，各个语义分割模型经过 ARBS 第二阶段的训练后，3 个指标均得到了提升，其中 PA 指标分别提升了 0.75%，0.85%，0.57%；MPA 指标分别提升了 6.57%，7.04%，6.30%；mIoU 指标分别提升了 1.78%，1.65%，1.19%。添加了区域均衡采样的 3 种语义分割网络在 3 个指标上均得到了有效的提高，在 MPA 指标上 3 个网络甚至均提升了 6%以上。



(a) ARBS 对 PA 指标的影响



(b) ARBS 对 MPA 指标的影响



(c) ARBS 对 mIoU 指标影响

图 6　ARBS 对模型性能的影响

Fig. 6　Impact of ARBS on model performance

区域均衡采样模块实现了在语义分割任务中对类别的过采样和欠采样，让分布不平衡的数据集变成平衡数据集。不平衡数据集会让模型倾向于拟合头部类别而忽略掉尾部类别，造成头部类别的正确率高、尾部类别的正确率极低的现象，而平衡数据集则可以让模型均衡地学习所有类别特征。因此相比不平衡数据集，平衡数据集可以极大提高尾部类别的正确率。这也解释了 MPA 指标提升的程度大于 mIoU 指标的实验现象。MPA 是类别正确率的平均值，所有类别的正确率对这一指标的贡献是相同的；而 mIoU 与样本数量有关，因此尾部类别的 IoU 对 mIoU 这一指标的影响小于头部类别。经过区域均衡采样模块后，大部分的尾部类别的正确率均获得了较大的提升，使得 MPA 大幅提升。在 MHP2.0 验证集中，测试数据仍然是不平衡的，且区域均衡采样对头部类别的影响程度没有尾部类别大，因此 mIoU 指标没有获得与 MPA 指标相同程度的提升。

4.3.3　类间损失模块有效性分析

在探究类间损失模块 ILM 对模型影响的实验中，本文仍然利用 PSPNet 作为语义分割网络，在对照组第 70 个 epoch 的训练模型的基础上添加 ILM 模块，在 $\lambda=1$ 的条件下训练结果如表 2 所列。可以看出，类间损失 ILM 模块对 3 个评估指标都有一定的提升，在 3 个语义分割网络模型的基础上对 mIoU 分别提升了 0.72%，0.29%，0.13%。

表 2　ILM 训练结果的对比

Table 2　Comparison of ILM training results

Method	*PA*	*MPA*	*mIoU*
Deeplabv3+	71.67	44.56	34.98
Deeplabv3++ILM	**72.29**	**45.45**	**35.70**
PSPNet	71.84	45.18	35.92
PSPNet+ILM	**72.56**	**46.03**	**36.27**
DANet	72.06	45.92	36.51
DANet+ILM	**72.74**	**46.08**	**36.64**

ILM 模块针对相似类别误分类的问题对模型进行了优化，相比经典的交叉熵损失函数，ILM 模块中的类间损失函数放大了类间差距，不仅引导模型做出正确的类别响应，还降低了其他错误类别的响应率。除此之外，通过引进 *margin* 参数，增大相似类别之间的惩罚，使模型输出的正确类别和错误类别的间距更大，对相似类别的处理更加准确，如图 7 所示。

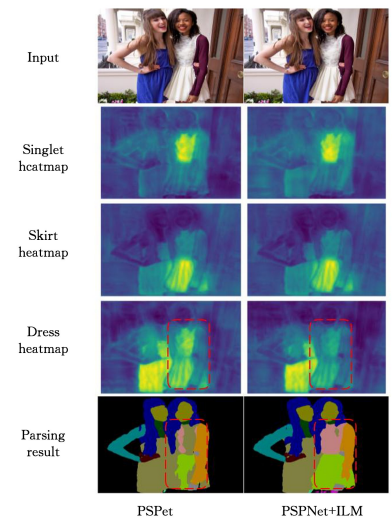


图 7　ILM 效果对比

Fig. 7　ILM effect comparison

以图 7 为例，PSPNet 在 dress 类别的热图响应大于 singlet，skirt 等正确类别的热图响应，从而造成类别解析错误；添加了 ILM 模块的 PSPNet 在 skirt，dress 等类别上既保持了正确的响应，又减弱了 dress 这一错误类别的热图响应，使得模型得到了更加准确的解析。

4.3.4 模型整体效果分析对比

本文采用了 3 种不同的语义分割网络（PSPNet，DeepLabv3＋和 DANet）作为 RIM 的语义分割模块来证明 RIM 的有效性。首先在不加入其他模块的条件下训练这 3 种语义分割模型作为对照组，每个模型在 MHPv2.0 训练集上训练 100 个 epoch 后，在 MHPv2.0 验证集上进行评估。实验组则是将上述 3 个语义分割网络作为 RIM 的语义分割模块，第一阶段训练 70 个 epoch，第二阶段训练 30 个 epoch 后在 MHPv2.0 验证集上进行评估。对照组和实验组的评估结果如表 3 所列。

表 3 实验组和对照组在 MHPv2.0 验证集上的表现

Table 3 Performance of experimental group and reference group on MHPv2.0 validation set

Method	*PA*	*MPA*	*mIoU*
DeepLabv3＋	71.93	44.88	35.49
RIM(DeepLabv3＋)	**72.64**	**50.95**	**36.85**
PSPNet	71.84	45.18	35.92
RIM(PSPNet)	**72.91**	**51.59**	**37.75**
DANet	72.06	45.92	36.51
RIM(DANet)	**72.85**	**51.90**	**37.84**

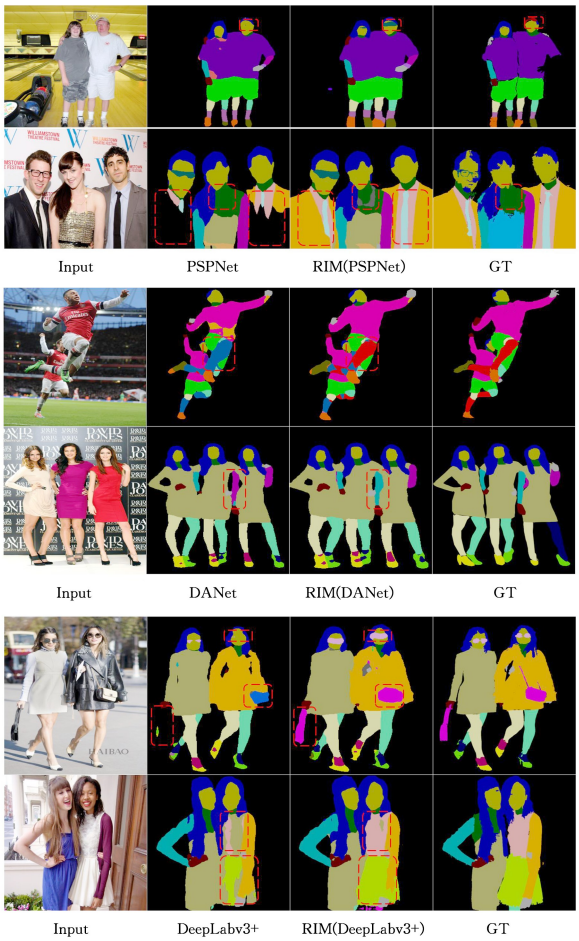


图 8 人体解析对比实例

Fig. 8 Comparison example of human parsing

从表 3 中可以看出，添加了 ARBS 和 ILM 模块的语义分割模型在 PA，MPA，mIoU 这 3 个指标上都有较大的提升。其中，MPA 的提升尤为明显，对于不同语义分割网络模块均提升了 5%以上；在 mIoU 指标上，RIM 在 DeepLabv3＋的基础上提升了 1.36%，在 PSPNet 的基础上提升了 1.83%，在 DANet 的基础上提升了 1.33%。ARBS 解决了数据集不平衡的问题，通过加大尾部类别比对模型的影响，大幅提高了尾部类别的正确率，如图 8 第一行中眼镜类别，第二行中首饰类别；而 ILM 模块则解决了相似类别误识别的问题，通过增大模型输出中正确类别的热图响应，减小错误类别的热图响应，来提高头部类别和尾部类别中易混淆类别的正确率，如图 8 第 3 行和第四行的肢体类别。两个模块的优化目标相互独立，因此融合了 ARBS 和 ILM 模块的 RIM 模型的精度能获得进一步的提升。表 3 所列的实验对比结果表明，在 MHPv2.0 这种多类别的长尾分布数据集的训练中，RIM 能缓解长尾分布数据和类别间的相似性带来的影响，提高模型的解析能力。

结束语 针对多类别数据集的数据呈长尾分布且类别间相似性较高导致的人体解析精度不高的问题，本文提出了一种结合区域均衡采样和类间损失的人体解析模型，通过区域均衡采样缓解了长尾数据分布对模型训练的影响，提高了模型对样本数量少和难解析类别的准确率；针对数据中相似类别难以区分的问题，将语义分割模块最后经过均衡采样的热图特征与真实标签送入定义好的类间损失函数，强化网络输出更具区分度的解析结果。实验结果表明，本文模型在不改变语义分割模块结构的条件下，显著提高了模型对样本稀少的尾部类别和难区分的相似类别的识别精度。

本文主要在模型的输入数据和损失函数上进行了针对性的改进，但是语义分割模块部分仍然沿用了目前流行的语义分割模型。在未来的研究中，拟尝试设计针对性更强的人体解析网络模型，并结合模型优化损失函数的设计，来提高人体解析的精度。

参 考 文 献

[1] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015: 3431-3440.

[2] LIANG X, GONG K, SHEN X, et al. Look into person: Joint body parsing & pose estimation network and a new benchmark [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(4): 871-885.

[3] RUAN T, LIU T, HUANG Z, et al. Devil in the details: Towards accurate single and multiple human parsing [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI, 2019, 33(1): 4814-4821.

[4] LI P, XU Y, WEI Y, et al. Self-correction for human parsing [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(6): 3260-3271.

[5] LIU K, CHOI O, WANG J, et al. CDGNet: Class Distribution

Guided Network for Human Parsing[J]. arXiv:2111.14173,2021.

[6] LIANG X,LIU S,SHEN X,et al. Deep human parsing with active template regression[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37(12):2402-2414.

[7] GONG K,LIANG X,ZHANG D,et al. Look into person:Self-supervised structure-sensitive learning and a new benchmark for human parsing[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,USA:IEEE,2017:932-940.

[8] GONG K,LIANG X,LI Y,et al. Instance-level human parsing via part grouping network[C]// Proceedings of the European Conference on Computer Vision (ECCV). Berlin,German:Springer,2018:770-785.

[9] YAMAGUCHI K,KIAPOUR M H,ORTIZ L E,et al. Parsing clothing in fashion photographs[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence USA:IEEE,2012:3570-3577.

[10] ZHAO J,LI J,CHENG Y,et al. Understanding humans in crowded scenes:Deep nested adversarial learning and a new benchmark for multi-human parsing[C]// Proceedings of the 26th ACM International Conference on Multimedia. New York,NY:ACM,2018:792-800.

[11] CUI Y,JIA M,LIN T Y,et al. Class-balanced loss based on effective number of samples[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach,USA:IEEE,2019:9268-9277.

[12] LIN T Y,GOYAL P,GIRSHICK R,et al. Focal loss for dense object detection[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice,Italy:IEEE,2017:2980-2988.

[13] LIU Z,MIAO Z,ZHAN X,et al. Large-scale long-tailed recognition in an open world[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach,USA:IEEE,2019:2537-2546.

[14] WANG J,ZHANG W,ZANG Y,et al. Seesaw loss for long-tailed instance segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville,USA:IEEE,2021:9695-9704.

[15] BULÒ S R,NEUHOLD G,KONTSCHIEDER P. Loss max-pooling for semantic image segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,USA:IEEE,2017:2126-2135.

[16] ZHOU B,CUI Q,WEI X S,et al. Bbn:Bilateral-branch network with cumulative learning for long-tailed visual recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,USA:IEEE,2020:9719-9728.

[17] KANG B,XIE S,ROHRBACH M,et al. Decoupling representation and classifier for long-tailed recognition[J]. arXiv:1910.09217,2019.

[18] WEN Y,ZHANG K,LI Z,et al. A discriminative feature learning approach for deep face recognition[C]// European Conference on Computer Vision. Berlin,German:Springer,2016:499-515.

[19] RONNEBERGER O,FISCHER P,BROX T. U-net:Convolutional networks for biomedical image segmentation[C]// International Conference on Medical Image Computing and Computer-assisted Intervention. Berlin,German:Springer,2015:234-241.

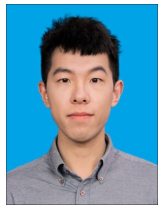
[20] ZHAO H,SHI J,QI X,et al. Pyramid scene parsing network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,USA:IEEE,2017:2881-2890.

[21] CHEN L C,PAPANDREOU G,SCHROFF F,et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv:1706.05587,2017.

[22] CHEN L C,ZHU Y,PAPANDREOU G,et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]// Proceedings of the European Conference on Computer vision(ECCV). Berlin,German:Springer,2018:801-818.

[23] FU J,LIU J,TIAN H,et al. Dual attention network for scene segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach,USA:IEEE,2019:3146-3154.

[24] DENG J,DONG W,SOCHER R,et al. Imagenet:A large-scale hierarchical image database[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami,USA:IEEE,2009:248-255.



LI Yang, born in 1998, postgraduate. His main research interests include deep learning and semantic segmentation.

HAN Ping, born in 1980, Ph.D, associated professor. His main research interests include deep learning, computer vision, and embedded system.

(责任编辑:喻藜)