融合多尺度信息的弱监督语义分割及优化

熊昌镇，智慧
（北方工业大学城市道路交通智能控制技术北京市重点实验室，北京 100144）

摘 要：为提高弱监督语义分割算法精度，提出一种融合多尺度特征的分割及优化算法。首先，基于迁移学习算法构建多尺度特征模型，类别预测时引入新分类器，减少因预测目标类信息错误导致分割失败的情况；其次，将多尺度模型与原迁移学习模型进行加权集成，增强模型泛化性能；最后，结合预测类可信度调整分割图中相应类像素的可信度，规避假正例分割区域。在 VOC 2012 验证集上的平均交并比为 58.8%，测试集上的平均交并比为 57.5%，同比原迁移学习模型分别提升 12.9%和 12.3%，也优于其他以类标作为监督信息的语义分割算法。

关键词：深度学习；弱监督学习；模型融合；多尺度特征；模型优化
中图分类号：TP18; TP391.4
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2019004

Weakly supervised semantic segmentation and optimization algorithm based on multi-scale feature model

XIONG Changzhen , ZHI Hui
Beijing Key Laboratory of Urban Intelligent Control, North China University of Technology, Beijing 100144, China

Abstract: In order to improve the accuracy of weakly-supervised semantic segmentation method, a segmentation and optimization algorithm that combines multi-scale feature was proposed. The new algorithm firstly constructs a multi-scale feature model based on transfer learning algorithm. In addition, a new classifier was introduced for category prediction to reduce the failure of segmentation due to the prediction of target class information errors. Then the designed multi-scale model was fused with the original transfer learning model by different weights to enhance the generalization performance of the model. Finally, the predictions class credibility was added to adjust the credibility of the corresponding class of pixels in the segmentation map, avoiding false positive segmentation regions. The proposed algorithm was tested on the challenging VOC 2012 dataset, the mean intersection-over-union is 58.8% on validation dataset and 57.5% on test dataset. It outperforms the original transfer-learning algorithm by 12.9% and 12.3%. And it performs favorably against other segmentation methods using weakly-supervised information based on category labels as well.

Key words: deep learning, weakly-supervised learning, model integration, multi-scale feature, model optimization

1 引言

语义分割是目前比较流行的一种视觉识别任务，其主要目的是给图像中的每一个像素进行语义类别的划分，在生物医疗图像的分析[1-2]，自动驾驶[3]、图像搜索引擎[4]、人机交互[5-6]等各个领域都有着广泛的应用。最近几年基于深度卷积神经网络（DCNN，deep convolution neural network）[7] 法的语义分割任务在性能上有了较大提升，并且达到了在基准测试数据集上的最高水平。然而 DCNN 的学习过程需要大量的像素级标注训练数据，制作此类像素级标注的过程比较耗时费力，导致现有数据集

收稿日期：2018−05−21；**修回日期**：2018−08−22
通信作者：熊昌镇，xczkiong@163.com
基金项目：国家重点研究发展计划基金资助项目（No.2017YFC0821102）
Foundation Item: The National Key Research and Development Plan(No.2017YFC0821102)

上的分割标注在质量和多样性上仍然无法满足需求。为了克服收集训练数据标注的困难并设计一个更具有扩展性和通用性的语义分割模型，研究者们致力于弱监督学习的研究，通过更易获得的较像素级标注更弱的监督信息来实现语义分割，如基于类标[8-11]及类标加辅助信息[12-14]、像素点[15]、边界框[16-17]、涂鸦等[18]四大类弱标注的语义分割算法。其中类标是最容易获取的标注，Pathak 等[8] 将语义分割看作是多实例学习的问题，利用最大池化操作强行限制每张图像至少有一个像素属于正实例目标类，但是因为监督信息缺失了目标的位置和形状，导致分割结果不太平滑。随后 Pathak 等[9]提出了嵌入位置信息，利用可辨识性定位自动识别出每个语义类的大体区域位置来提高分类的精度。Kwak 等[10] 利用超像素池化层生成初始语义分割需要的边缘形状信息。虽然这些方法可以粗略地定位目标，但是通常不能精确地推断出像素信息，因为更倾向于聚焦目标的部分显著信息，而不是目标的整个区域。Kolesnikov 等[11]则提出将种子损失、扩张损失和约束边界损失集成到一个网络训练分割模型进行训练，并应用全局加权排序池化操作，约束目标边界信息并聚焦目标显著位置，但该算法对于背景相似的目标区域在定位上容易产生偏差，而且类别识别的效果不是太好。为进一步提升分割性能，研究者们开始以类标注为基础扩增新的数据信息，Lin 等[12]提出利用自然语言作为弱监督标注，Hong 等[13]利用额外数据（非目标数据源）的像素标注辅助弱监督信息学习，但是需与实际目标数据的类别相互独立，再依靠迁移学习捕获目标类需要的像素信息。Hong 等[14] 以网页视频作为额外数据源，利用目标和背景的不同动态信息与三维结构信息区分出前景与周围的背景信息，获取更准确的目标边界，使分割性能有了较大的提升，但整个网络结构对小目标信息的捕获比较欠缺。第二类以像素点为弱标注信息可提供目标粗略位置的方式，有助于提升分割效果。Bearman 等[15] 提出将分类损失和定位损失相结合，并增加了目标显著性作为先验知识来优化，但从其结果来看分割边缘不完整。第三类以边界框为弱标注信息可提供整个目标区域位置的信息，可进一步提升目标的分割效果。Papandreou 等[16]利用最大期望（EM，expectation-maximization）来动态预测边界框内的前景像素。Dai[17]没有对边界框内的像素进行直接评估，而是利用现成的候选区域（region proposals）迭代选取最佳区域，进而生成分割掩码，分割性能与类标监督相比有了很大提升，但是相较全监督语义分割性能还有较大差距。第四类以涂鸦为标注信息即是在兴趣目标上简单勾画一条线，它提供目标相对位置范围内的一些稀疏像素信息。Lin 等[18] 利用图模型优化交互式分割模型，即在训练过程中循环利用当前的分割结果作为监督信息进行迭代直至模型收敛，性能相当于边界框给出的分割结果。遗憾的是该标注在其他数据集上不可用。

以上各类弱监督语义分割算法在复杂背景及包含众多小目标的场景下，对狭小目标及目标的形状边缘分割往往不理想，主要原因还是对目标尺度空间的信息学习不全面，然而目前在强监督语义分割任务中已有多种学习尺度空间特征的算法[19-21]。Chen 等[19] 将金字塔式输入图像送入到 DCNN 以提取不同尺度上的显著度特征。Yu 等[20] 在原有网络顶部级联空洞卷积层来捕获图像不同尺度信息。Zhao 等[21] 利用空间金字塔式池化作用于最后一层卷积，进而获取多种尺度分辨率的目标特征。这些多尺度算法在强监督语义分割中均可以获得良好效果，证实了学习尺度空间信息的有效性。鉴于此，本文以迁移学习网络[12] 为基本框架，以金字塔式多尺度图像为网络的输入，并增加一个新层对多尺度特征进行降维，构建多尺度的弱监督语义分割模型，提取目标的多尺度特征。语义分割通常包含图像类别预测和像素分割两部分内容，类别预测效果对最终分割结果起着至关重要的作用，因为错误的目标类别必然会导致像素分割的错误[11,13,16]。随着深度学习技术的发展，以边界框为监督信息的目标检测技术也得到了很大的发展，检测精度和速度都有很大提升[22-23]。为避免类别错误导致的分割失败，引入文献[23]中在同源数据集学习的检测模型给出的图像类别信息来提升分割的精度。现有算法中单模型分割算法对某些目标的分割效果好，但对另一些目标的分割效果差，无法学到所用类别的有效信息，导致无法对所有目标类都进行有效分割，会导致模型泛化能力差，不同分割模型的侧重点不同，学习到的语义特征也不同，即每个模型都有各自的优势[9-10]，为充分利用不同模型的优势，本文对多尺度分割模型进行优化，与原迁移学习模型进行集成，同时结合类别可信度和像素分割可信度进一步提升图像分割的精度。

2　多尺度图像分割

将应用于强监督语义分割算法的多尺度信息引入弱监督分割算法中，以迁移学习模型为基础，输入多个尺度的图像，提取多个尺度上的图像特征后归一化成相同大小的特征图再拼合在一起构造多尺度特征，然后对多尺度特征进行降维，利用迁移学习模型的注意力机制模型初始化新构造的多尺度模型，最后对多尺度分割模型进行训练，学习多尺度特征的信息。该模型的基本框架如图 1 所示，主要包括提取多尺度特征的编码结构 f_{enc}、多尺度特征图级联与降维，聚焦目标显著区域的注意力机制 f_{att} 和低维特征解码至高维特征进行前景分割的解码结构 f_{dec}。

2.1　学习图像的多尺度信息

采用与迁移学习模型相同的编码结构、注意力机制和解码结构[13]，用 $\boldsymbol{x}$ 表示来自源数据集 S 或目标数据集 T 的输入图像。首先将输入图像缩放成分辨率为 330×330 固定大小的图像块，经过随机裁剪变成分辨率为 320×320 的图像，利用尺度因子 $s\in\{1,0.75,0.5\}$ 将裁剪后的图像块缩放成 3 种不同尺度，作为 3 组并行编码器 f_{enc} 的输入，如式(1)所示。

$$\boldsymbol{A}_s = f_{\text{enc}}(\boldsymbol{x}_s;\boldsymbol{\theta}_e) \tag{1}$$

其中，$\boldsymbol{\theta}_e$ 为 3 组编码器 f_{enc} 的共享卷积层训练参数，$\boldsymbol{A}_s\in\boldsymbol{R}^{whd}$ 为编码器最后一层卷积层输出特征图，w、h 和 d 分别代表特征图的宽、高和输出维度。再将尺度因子为 0.75 和 0.5 对应的特征图 $\boldsymbol{A}_s$ 按照双线性插值进行放大，即保持与编码器中输入尺度因子为 1 的最后一层卷积层输出特征图相同大小，然后再将缩放后的特征图沿维度方向进行级联，同时在编码器的末端增加一个新的卷积层，对融合的多尺度特征图进行降维以生成固定的通道数，进而适应后续注意力机制的输入要求，通过网络训练学习图像的多尺度特征。

2.2　聚焦目标的显著区域

当给出融合后特征图 $\boldsymbol{A}$ 和对应目标类向量形式 ℓ^* 时，注意力机制的作用就是学 $\boldsymbol{A}$ 中的对应目标类位置的正权重向量 $\{\boldsymbol{\alpha}^l\}_{\forall l\in\ell^*}$，$\boldsymbol{\alpha}^l$ 表示第 l 个目标类与对应特征位置的相关性。注意力机制的过程可表示为

$$\begin{aligned}\boldsymbol{v}^l &= f_{\text{att}}(\boldsymbol{A},\boldsymbol{y}^l,\theta_\alpha)\\ \alpha_i^l &= \frac{\mathrm{e}^{v_i^l}}{\sum_i \mathrm{e}^{v_i^l}}\end{aligned} \tag{2}$$

其中，θ_α 为注意力机制 f_{att} 的模型参数；$\boldsymbol{y}^l\in\{0,1\}$ 表示第 l 类的类标向量，在训练过程表示来自源数据与目标数据的真值类，在模型执行推断时则表示分类器给出的目标预测类，即图 1 的类别处。$\boldsymbol{v}^l$ 为非正则化的聚焦权重，通过 soft max 函数给出正则化后的权重 $\boldsymbol{\alpha}^l$，目的是鼓励模型只聚焦图像目标类的一个显著区域[24]。迁移学习算法中所用的注意力机制 f_{att} 为

$$\boldsymbol{v}^l = \boldsymbol{W}^{\text{att}}(\boldsymbol{W}^{\text{feat}}\boldsymbol{A}\odot\boldsymbol{W}^{\text{label}}\boldsymbol{y}^l)+\boldsymbol{b} \tag{3}$$

其中，$\boldsymbol{W}^{\text{att}}$、$\boldsymbol{W}^{\text{feat}}$ 和 $\boldsymbol{W}^{\text{label}}$ 表示模型学习的权重，$\odot$ 表示对应元素相乘，$\boldsymbol{b}$ 是偏置向量。而为了将注意力机制更好应用于迁移学习中，在注意力机制顶端添加额外的分类层 f_{cls}，用于联合优化源数据集与目标数据集的弱监督学习。为此，通过聚合空间区域上的特征来提取特定类 l 的显著特征，如式(4)所示。

$$\boldsymbol{z}^l = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\alpha}^l \tag{4}$$

训练注意力机制 f_{att} 的过程即是最小化分类损失的过程，用 e_c 表示 soft max 函数，用于计算真值 $\boldsymbol{y}_i^l$ 和预测类标 $f_{\text{cls}}(\boldsymbol{z}_i^l;\theta_c)$ 的损失。

$$\min_{\theta_c,\theta_\alpha}\sum_{i\in T\cup S}\sum_{l\in\ell_i^*}e_c(\boldsymbol{y}_i^l,f_{\text{cls}}(\boldsymbol{z}_i^l;\theta_c)) \tag{5}$$

其中，θ_c 为分类层的学习参数，$\boldsymbol{z}_i^l$ 表示来自源数据与目标数据第 i 张图的类 l 显著响应图。

2.3　生成前景分割图

当注意力机制给出兴趣目标类的位置时，接下来便需要解码器来重构相应聚焦目标的前景分割

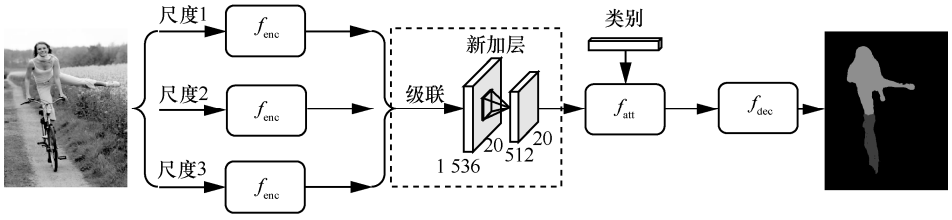


图 1　多尺度特征分割模型

图。由于经过 soft max 之后聚焦权重会变得比较稀疏，为此需要将式(4)获得的特定目标类显著图 $\boldsymbol{z}^l$ 作为解码器输入的系数，以获取密集显著图，且与注意力机制聚焦的显著图 $\boldsymbol{\alpha}^l$ 具有相同大小，即表示为

$$\boldsymbol{s}^l = \boldsymbol{A}\boldsymbol{z}^l \tag{6}$$

训练解码器的过程为最小化分割损失，对应的目标函数 e_s 为 soft max 损失函数可表示为

$$\min_{\theta_\alpha,\theta_s} \sum_{i\in S}\sum_{l\in\ell_i^*} e_s(\boldsymbol{b}_i^l, f_{\text{dec}}(\boldsymbol{s}_i^l;\theta_s)) \tag{7}$$

其中，θ_s 表示解码器 f_{dec} 的学习参数，$\boldsymbol{b}_i^l$ 为源数据集 S 的 l 类中的第 i 类目标的二值分割图，$i\in S$ 表示目标函数的优化只对源数据集进行。但是学得的参数 θ_s 对不同目标类是实现共享的，所以该结构能够利用已学得的通用类的基本特征，如颜色、形状、纹理等先验知识迁移应用到其他多类场景。解码器 f_{dec} 的基本结构与编码器 f_{enc} 呈对称形式，通过一系列的上采样、转置卷积及校正运算将低分辨的目标类特征图重构为与输入 $\boldsymbol{x}$ 相同大小的密集前景分割图。

多尺度特征模型训练对新增加的层及解码器部分均使用零均值高斯分布初始化，学习过程中固定编码器的权重，利用原迁移学习模型的对应层对编码器、注意力机制进行初始化，并应用自适应矩估计算法（Adam，adaptive moment estimation），根据式(5)分类目标函数学习新层与注意力机制的参数，以及式(7)分割目标函数来学习解码器部分的参数。

3 算法优化

将文献[23]中同源数据集学习的检测模型给出的图像分类结果作为多尺度分割模型预测时的新分类器，只使用检测模型给出的预测目标类及类别可信度；然后对类别优化后的多尺度模型与原迁移学习模型进行加权集成；最后利用新分类器的类别可信度优化集成模型输出分割图的像素可信度，以进一步提升分割的精度。

3.1 类别预测优化

语义分割任务实际包含图像类别预测和像素分割这 2 类任务模型所用分类器的预测效果对最终像素级分割结果起着至关重要的作用，因为错误的目标类必然会导致像素分割的错误，而模型结构中添加的分类层 f_{cls}，只是为了学习目标数据集类别上的注意力机制，训练过程结束后，需要引入一个单独的分类器完成模型的预测。原迁移学习的分类器是基于 VGG16 的全卷积神经网络的类别预测，预测准确率不够，影响分割效果，鉴于学习数据集（MS COCO（microsoft common objects in context）[25]，VOC 2012（visual object classes challenge）[26]）的考虑，选用在同源数据集上学习的检测模型作为目标分割时的类别分类器，基于弱监督学习模式的衡量，不输出检测框位置信息，只将检测结果的图像目标类别 l 及类别可信度 P^l 的信息保存下来，并于图 1 所示的类别处给入到多尺度特征分割模型中，随后模型自适应构建注意力权重 α_i^l，即相应目标类的显著区域。

3.2 模型集成优化

当假设空间较大时，单模型分割算法往往不能保证对所有目标类的有效性，导致模型泛化性能差。此时如果有多个假设在相同数据集上训练并能达到同等性能，便可以将多个学习器进行结合，利用个体学习器间的差异性互补来有效规避单一模型的性能缺陷[27]。因此，将性能相近且同属“神经网络式”的多尺度特征模型与原迁移学习模型进行集成，并按照加权的方式进行模型融合，如式(8)所示。

$$\boldsymbol{H}^l = \sum_{t=1}^{T}\sum_{l\in\ell^*} w_t f_{\text{dec}}(\boldsymbol{s}_t^l) \tag{8}$$

其中，$f_{\text{dec}}(\boldsymbol{s}_t^l)$ 代表第 t 个模型的解码器输出的第 l 类前景概率图，即原迁移学习模型与多尺度特征模型相应类别的前景概率图；w_t ($w_t \geqslant 0$)是个体学习器的权重；$\boldsymbol{H}^l$ 是模型融合后的目标类概率图。

3.3 分割可信度优化

鉴于注意力机制只是给出兴趣目标的粗略位置，对目标遮挡、复杂背景、噪声混入等情况，模型输出的分割图包含所有预测目标类的像素信息，但是其中某些类的位置信息会有偏差，致使分割错误。研究发现引起错误的类通常在分类器预测的可信度与视觉显著度上呈负相关。利用新分类器给出的预测类别可信度 P^l，调整相应类的概率图响应像素值，即用低目标类概率值抑制错误响应的高像素值，用高预测类概率值提升输出的低响应像素值，达到规避假正例区域及非预测目标的噪声信息，同时强化正确类标的分割图像的目的。预测类可信度优化分割概率图如式(9)所示。

$$\boldsymbol{M}^l = P_j^l \boldsymbol{m}_j \tag{9}$$

其中，P_j^l 表示由新分类器预测的 l 类中的第 j 类的类别可信度，$\boldsymbol{m}_j$ 是指经过条件随机场（CRF，confidence random fields）处理后的第 j 类集成模型的前景概率图，通过对应类的可信度与该类的前景概率图直接相乘，可计算出当前类响应图的像素值，遍历所有目标类后便可获得概率图 $\boldsymbol{M}^l$，最终分割图是在其通道维度方向，即 l 类张响应图中取最大像素值，保留为相应类标的 id 值。

4　实验结果与分析

多尺度分割模型使用 MS COCO 为源数据集 S，VOC 2012 为目标数据集 T，其中源数据集 S 共含 60 类目标，与目标数据集 T 的 20 类目标相互独立；目标数据集 T 仅提供类别监督信息。最后在 VOC2012 验证集、测试集进行语义分割实验，采用平均交并比（mIoU，mean intersection-over- Union）来衡量实际分割结果与分割真值（GT，groud truth）的差异。实验中使用文献[23]中的 PVANet 模型进行目标检测，将大于给定阈值的边界框类标和最大概率作为分割图像的类别及可信度，只使用类别信息，不使用边界框的信息。实验中所用的类别阈值为 0.75，将检测的类别结果和可信度保存下来，在图像分割时只加载类别信息，不进行实际目标检测操作。

4.1　多尺度分割模型和自身优化算法的对比实验

将原迁移学习模型记为 O，多尺度特征模型记为 M。表 1 给出了多尺度特征模型、集成模型、预测目标类及其可信度优化在 VOC 2012 的验证集上的性能对比。多尺度特征模型与原迁移学习模型集成时的个体学习器给定权重按 $w_1 : w_2 = 3 : 2$ 的比例加权，后缀 c 表示引入新分类器后的结果，p 是分类器给出的预测目标类可信度。从表中的数据可以看出，构建的多尺度特征模型 M 与原迁移学习模型 O 具有相似的分割性能，满足同质型差异化模型集成的具有一定“准确性”要求。引入类别预测优化的图像分割算法（M_c）同比多尺度特征模型利用的原迁移学习模型固有分类器在性能上提升了 2.9%。经过双模型的集成优化后 M+O_c 分割算法性能又提升了 2.9%，证明单一学习器具有不可避免的性能缺陷，利用集成学习可以使同质型差异化模型实现互补，从而提升分割的效果。由于模型结构中的注意力机制只能给出目标的粗略位置，在出现目标遮挡、复杂背景、噪声混入等情形时，分割往往容易出现错误，因此使用图像类的预测可信度 p 对算法进行优化，同比集成模型提升了 0.9%，验证了本文算法的多尺度分割及不同优化策略引入都不同程度地提升了分割算法的精度。

表 1　本文算法在 VOC 2012 的验证集上的性能对比

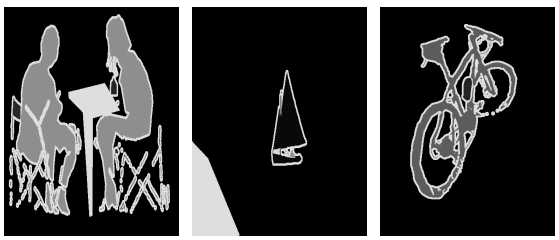
VOC 2012 验证集	O	M	M_c	M+O_c	M+O_c_p	M+O_gt
background	85.3%	86.9%	87.1%	87.7%	87.7%	87.7%
aeroplane	68.5%	73.5%	76.1%	77.5%	77.5%	76.9%
bicycle	26.4%	27.0%	26.4%	28.5%	28.7%	27.8%
bird	69.8%	66.6%	66.6%	74.5%	73.5%	73.0%
boat	36.7%	33.8%	33.9%	38.9%	44.0%	41.6%
bottle	49.1%	52.1%	43.7%	49.8%	51.3%	45.6%
bus	68.4%	81.8%	82.8%	82.3%	82.6%	81.8%
car	55.8%	53.1%	51.7%	55.9%	58.9%	58.5%
cat	77.3%	73.7%	76.9%	81.5%	81.0%	81.2%
chair	6.2%	4.1%	7.6%	12.3%	15.1%	14.8%
cow	75.2%	74.7%	85.4%	87.0%	86.8%	86.9%
diningtable	14.3%	8.2%	10.0%	10.2%	13.0%	13.0%
dog	69.8%	61.1%	61.7%	69.4%	70.7%	71.0%
horse	71.5%	70.6%	77.1%	80.6%	80.2%	80.6%
motorbike	61.1%	67.2%	69.7%	72.5%	72.9%	72.9%
person	31.9%	45.1%	41.9%	41.1%	42.8%	40.7%
pottedplant	25.5%	20.7%	20.6%	21.1%	23.4%	22.8%
sheep	74.6%	68.4%	85.4%	87.5%	87.4%	87.4%
sofa	33.8%	31.9%	40.2%	44.5%	46.4%	46.4%
train	49.6%	58.4%	60.3%	59.9%	59.9%	59.9%
tvmonitor	43.7%	46.7%	49.7%	52.8%	51.6%	50.3%
mIoU	52.1%	52.6%	55.0%	57.9%	58.8%	58.1%

表 1 数据中 M+O_gt 表示的是集成模型引入真值类标的分割性能，但比集成模型结合类及可信度优化算法 M+O_c_p 要低 0.7%，说明分类真值并不能作为算法的上限。这是因为类别真值只是表示该图像有这类目标，可信度为 100%，但不考虑目标的大小、位置等信息，同时图像中又包含与此类目标相类似的其他信息，导致图像分割结果中该类别的像素分割的可信度高，造成图像分割错误，而目标的大小、形状和位置信息对图像分类都会造成影响。分类的可信度表示类别分类的难度，与分割可信度相结合可避免类别可信度低而分割可信度高造成的假正例现象。

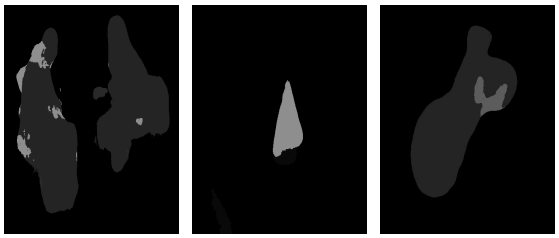
图 2 给出了基于不同形式的目标类分割效果对比图，即直接引入真值图像类别信息和预测的目标类别信息进行分割的结果。图 2(a)是输入图像，图 2(b)是真值分割图，图 2(c)是引入的真值目标类

别分割图，图 2(d)是引入预测目标类别及可信度优化的分割结果。对应上述实验的 M+O_c_p 的结果，可以看出预测类别及可信度优化的分割效果明显优于直接给定真值类的分割图。其原因是复杂背景及包含有众多小目标的情况下，注意力机制聚焦的兴趣目标位置是稀疏的，当引入包含最完整信息的真值类时，在预测过程根据分割响应图的像素值大小确定的最终分割图时，往往会出现类正确但是位置错误的情况，弱化了分割精度。通过引入分类器预测目标类时输出的类可信度，不仅可以强化正确目标类相应的像素响应值，还可以抑制错误定位的类响应值，进而改善分割的性能。

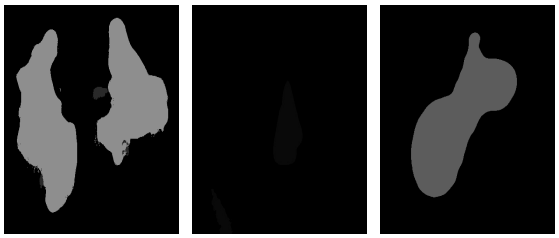
(a) 原图



(b) 真值分割图



(c) 真值目标类分割图



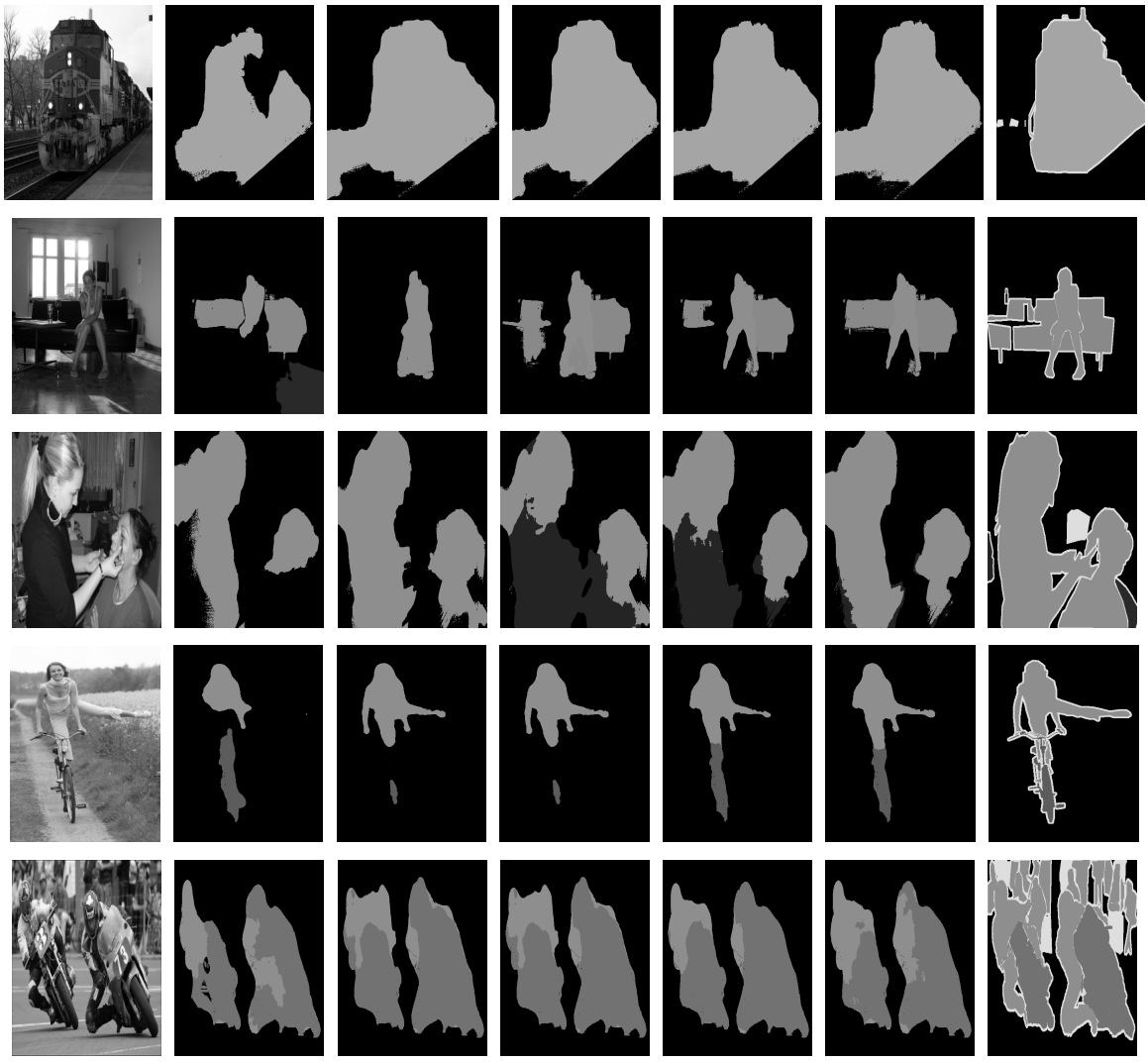
(d) 预测目标类分割图

图 2 不同目标类别下的分割效果

4.2 与其他算法对比实验

图 3 显示了部分测试图像在验证集上的语义分割结果图。第一列是输入图像，第二列是原迁移学习模型 O（TransferNet[13]）的分割结果，对比本文第三列的多尺度特征提取模型 M，可以看出模型 M 能够给出尺度空间上更丰富的信息，但是因为原分类器的准确度不是太高，导致部分目标信息的丢失，而且由于注意力机制的粗定位，部分目标给出的显著区域不合理，造成了单一的多尺度特征分割并不理想。第四列 M_c 是在模型 M 的基础上更换新分类器 c，可以看出减少了目标信息的丢失，进而避免了因类预测失败造成分割不理想的情况。第五列是引入新分类器的同质型集成模型分割效果图，明显可以看出通过模型间的互补性，目标的分割更准确，弥补了丢失的信息，去除了多余的噪声信息。第六列是引入预测目标类可信度 p 优化后分割效果，发现正确目标类的有效分割区域更加完整了，同时有效地抑制了假正例区域，使得最终的模型分割信息更全面，边缘轮廓更细致。

同时为了更加充分的验证算法的性能，与目前采用各类弱监督信息（类标及类标加辅助信息、像素点、边界框、涂鸦）实现语义分割的主流算法进行对比，包括目前单纯以类标作为弱监督信息的最好算法 AffinityNet[28]，为了对比的公正性，只给出了基于网络 VGG-16 结构的性能对比。表 2 列出了各类算法在 VOC 2012 验证集和测试集上的分割性能对比结果。其中，I 指应用类别作为监督信息，P 指应用像素点作为监督信息，S 是简笔涂鸦式监督信息方式，B 是指利用边界框为监督信息，*表示加入了强监督信息。从表 2 中可以看出多尺度分割及优化算法在验证集上的结果比 AffinityNet 算法高 0.4%，比基于相同迁移学习模型改进的 CrawlSeg[14] 算法提高了 0.7%。AffinityNet 算法提出利用亲和网络预测相邻像素间的语义相似性，进而将局部响应扩散到同一语义实体的附近区域，最后通过预测的像素相似性随机游走实现语义传播，对目标的响应区域位置及类别预测效果都比较好，但是它的实际分割对目标的轮廓及细节信息处理不是太完整。多尺度分割及优化算法在测试集上的结果有些不尽如人意，但是比 TransferNet[13]提升了 6.3%。结果说明多尺度分割模型有效地提取了多尺度的空间信息，并与同质型原迁移学习模型进行集成，提高了泛化性能，对捕获细节轮廓信息更有效； 同时利用预测目标类及其可信度优化注意力机制的定位，获得了更好的分割效果。



(a) 原图　(b) O（TransferNet）　(c) M　(d) M_c　(e) M+O_c　(f) M+O_c_p　(g) M+O_gt

图3　VOC 2012 验证集分割效果对比

表 2　VOC2012 验证集/测试集性能对比

算法	监督类型	平均交并比	
		验证集	测试集
SEC(ECCV 2016)[11]	I	50.7%	51.7%
*TransferNet (CVPR 2016)[13]	I	52.1%	51.2%
*AF-MCG (ECCV 2016)[29]	I	54.3%	55.5%
AE-PSL(CVPR 2017)[30]	I	55.0%	55.7%
*CrawlSeg (CVPR 2017)[14]	I	58.1%	58.7%
AffinityNet((DeepLab，2018)[28]	I	58.4%	60.5%
What'sPoint(ECCV 2016)[15]	P	46.0%	43.6%
WSSL (ICCV 2015)[16]	B	60.6%	62.2%
BoxSup(ICCV 2015)[17]	B	62.0%	64.2%
Scribblesup(CVPR 2016)[18]	S	63.1%	—
M+O_c_*p*	I	58.8%	57.5%

注：“*”表示该类算法加入强监督信息。

图 4 给出的是一些失败案例，图 4(a)与图 4(d)是相应案例的原图，图 4(b)与图 4(e)是对应原图的真值分割图，图 4(c)与图 4(f)是模型预测分割图。作为弱监督的语义分割算法，因为监督信息缺失目标数据集图像的位置和形状关键信息，往往会在复杂背景或者众多小目标的情况下出现错误。失败案例表明，因为注意力机制对兴趣目标的定位是粗糙的，难免会引入噪声信息，纵使对目标的显著性响应进行优化也不能完全解决，从而影响分割的准确性。后期可以考虑增加一些对目标显著性精确定位的措施，强化兴趣目标的整体响应区域。

5　结束语

考虑到原迁移学习的单模型在复杂背景或目

(a) 案列 1 原图　(b) 案列 1 真值分割图　(c) 案列 1 模型预测分割图　(d) 案列 2 原图　(e) 案列 2 真值分割图　(f) 案列 2 模型预测分割图

图 4　一些分割失败的例子

标类别比较多的情况下，往往对小目标形状边缘分割不理想，同时也因分类器的目标识别不准确导致分割对象出现错误，以及基于注意力机制的粗定位，简单使用显著性响应容易引入噪声信息等问题，构建了可提取多尺度特征信息的图像分割模型，提取图像的多尺度信息，并引入 3 种优化策略对分割算法优化以提升分割精度。优化策略首先将同质型差异化的多尺度特征模型与原迁移模型进行模型集成，以弥补单模型的性能缺陷；然后引入新的图像分类器改善预测目标类别的准确度提高图像分割的性能；最后结合预测类可信度优化分割响应图的像素可信度，避免类别可信度低而图像分割可信高造成图像分割错误。在目标数据集 VOC2012 测试算法，实验给出了单尺度特征模型、双模型集成、新类别分类器及类可信度优化的实验结果，并与其他前沿算法进行了对比。结果表明，多尺度特征模型及优化算法，在 VOC 2012 验证集上的平均交并比达 58.8%，测试集上的平均交并比为 57.5%，比原迁移学习算法提升 12.9%和 12.3%，在验证集比目前以类标作为监督信息的最好语义分割 AffinityNet 算法提升 0.7%，验证了本文算法的有效性。由于使用的基础网络性能不够及注意力机制的缺陷影响了分割效果的进一步提升，后续将考虑改善网络结构和引入目标显著性改善注意力机制来提高分割的效果。

参考文献：

[1] 关涛, 周东翔, 刘云辉. 基于色差向量场的彩色光学显微细胞图像分割[J]. 光学学报, 2014, 34(01): 0115001.
GUAN T, ZHOU D X, LIU Y H. Color optical microscopic cell image segmentation based on color difference vector field[J]. ACTA Optica Sinica, 2014, 34(01): 0115001.

[2] 孙延奎. 光学相干层析医学图像处理及其应用[J]. 光学精密工程, 2014, 22(04): 1086-1104.
SUN Y K. Medical image processing techniques based on optical coherence tomography and their applications[J]. Optics and Precision Engineering, 2014, 22(04): 1086-1104.

[3] ESS A, MUELLER T, GRABNER H, et al. Segmentation-based urban traffic scene understanding[C]//British Machine Vision Conference, BMVC. 2009(84) :1-11.

[4] WAN J, WANG D Y, HOI S C H, et al. Deep learning for content-based image retrieval: a comprehensive study[C]// the 22nd ACM international conference on Multimedia. 2014, 978: 157-166.

[5] OBERWEGER M, WOHLHART P, LEPETIT V. Hands deep in deep learning for hand pose estimation[C]// Computer Vision Winter Workshop. 2015: 21-30.

[6] 向守兵,苏光大,任小龙,等. 实时手指交互系统的嵌入式实现[J]. 光学精密工程, 2011, 19(08): 1911-1920.
XIANG S B, SHU G D, REN X L, et al. Embedded implementation of real-time finger interaction system, [J]. Optics and Precision Engineering, 2011, 19(08): 1911-1920.

[7] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]// 2017 IEEE International Conference on Computer Vision. 2017, 2380: 2980-2988.

[8] PATHAK D, SHELHAMER E, LONG J, et al. Fully convolutional multi-class multiple instance learning[C]//International Conference on Learning Representations. 2015:1-4

[9] PATHAK D, KRAHENBUHL P, DARRELL T. Constrained convolutional neural networks for weakly supervised segmentation[C]// IEEE International Conference on Computer Vision. 2015, 1550: 1796-1804.

[10] KWAK S, HONG S, HAN B. Weakly supervised semantic segmentation using superpixel pooling network[C]// AAAI Conference on Artificial Intelligence. 2017: 4111-4117.

[11] KOLESNIKOV A, LAMPERT C H. SEE D, Expand and constrain: three principles for weakly-supervised image segmentation[C]// European Conference on Computer Vision. 2016, 9908: 695-711.

[12] LIN L, WANG G R, ZHANG R, et al. Deep structured scene parsing by learning with image descriptions[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016, 1063: 2276-2284.

[13] HONG S, OH J, LEE H, et al. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016, 1063: 3204-3212.

[14] HONG S, YEO D, KWAK S, et al. Weakly supervised semantic segmentation using web-crawled videos[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017, 1063: 2224-2232.

[15] BEARMAN A, RUSSAKOVSKY O, FERRARI V, et al. What's the Point: Semantic Segmentation with Point Supervision[C]// European Conference on Computer Vision. 2016, 9911: 549-565.

[16] PAPANDREOU G, CHEN L C, MURPHY K, et al. Weakly and

semi-supervised learning of a DCNN for semantic image segmentation[C]// IEEE International Conference on Computer Vision. 2015, 1550: 1742-1750.

[17] DAI J F, HE K M, SUN J. BoxSup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation[C]// IEEE International Conference on Computer Vision. 2015, 1550: 1635-1643.

[18] LIN D, DAI J F, JIA J Y, et al. ScribbleSup: scribble-supervised convolutional networks for semantic segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016, 1063: 3159-3167.

[19] CHEN L C, YANG Y, WANG J, et al. Attention to scale: scale-aware semantic image segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016, 1063: 3640-3649.

[20] YU F, KOLTUN V. Multi-Scale context aggregation by dilated convolutions[C]//International Conference on Learning Representations. 2015: 1-13

[21] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017, 1063: 6230-6239.

[22] HONG S, ROH B, KIM K H, et al. PVANet: lightweight deep neural networks for real-time object detection[C]//Advances in Neural Information Processing Systems. 2016:1-7

[23] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]// IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015: 1137-1149.

[24] 窦燕, 孔令富, 王柳锋. 基于视觉熵的视觉注意计算模型[J]. 光学学报, 2009, 29(09): 2511-2515.
DOU Y, KONG L F, WANG L F. A computational model of visual attention based on visual entropy[J]. ACTA Optica Sinica, 2009, 29(9): 2511-2515.

[25] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]// European Conference on Computer Vision. 2014, 8693: 740-755.

[26] EVERINGHAM M, GOOL L, WILLIAMS C K, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88 (2): 303-338.

[27] 周志华. 机器学习[M].北京: 清华大学出版社, 2016: 171-184.
ZHOU Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016: 171-184.

[28] AHN J, KWAK S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2018:4981-4990

[29] QI X J, LIU Z Z, SHI J P, et al. Augmented feedback in semantic segmentation under image level supervision[C]// European Conference on Computer Vision. 2016, 9912: 90-105.

[30] WEI Y C, FENG J S, LIANG X D, et al. Object region mining with adversarial erasing: a simple classification to semantic segmentation approach[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017, 1063: 6488-6496.

[作者简介]

熊昌镇（1979−），男，福建建宁人，博士，北方工业大学副教授，主要研究方向为交通图像处理、机器学习。

智慧（1991−），女，内蒙古乌兰察布人，北方工业大学硕士生，主要研究方向为图像语义分割、注意力机制。