计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(12):309−315 [doi: 10.15888/j.cnki.csa.008850]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 YOLOv5 和 FCN-DenseNet 水下图像多目标语义分割算法①

曹建荣[1,2], 韩发通[1], 汪　明[1,2], 庄　园[1], 朱亚琴[1], 张玉婷[1]

[1](山东建筑大学 信息与电气工程学院, 济南 250101)
[2](山东省智能建筑技术重点实验室, 济南 250101)
通信作者: 汪　明, E-mail: xclwm@sdjzu.edu.cn

摘　要: 带视觉系统的水下机器人作业离不开对水下目标准确的分割, 但水下环境复杂, 场景感知精度和识别精度不高等问题会严重影响目标分割算法的性能. 针对此问题本文提出了一种综合 YOLOv5 和 FCN-DenseNet 的多目标分割算法. 本算法以 FCN-DenseNet 算法为主要分割框架, YOLOv5 算法为目标检测框架. 采用 YOLOv5 算法检测出每个种类目标所在位置; 然后输入针对不同类别的 FCN-DenseNet 语义分割网络, 实现多分支单目标语义分割, 最后融合分割结果实现多目标语义分割. 此外, 本文在 Kaggle 竞赛平台上的海底图片数据集上将所提算法与 PSPNet 算法和 FCN-DenseNet 算法两种经典的语义分割算法进行了实验对比. 结果表明本文所提的多目标图像语义分割算法与 PSPNet 算法相比, 在 *MIoU* 和 *IoU* 指标上分别提高了 14.9% 和 11.6%; 与 FCN-DenseNet 算法在 *MIoU* 和 *IoU* 指标上分别提高了 8% 和 7.7%, 更适合于水下图像分割.

关键词: YOLOv5; FCN-DenseNet; 语义分割; 水下场景识别

引用格式: 曹建荣,韩发通,汪明,庄园,朱亚琴,张玉婷.基于 YOLOv5 和 FCN-DenseNet 水下图像多目标语义分割算法.计算机系统应用,2022,31(12):309–315. http://www.c-s-a.org.cn/1003-3254/8850.html

Multi-object Semantic Segmentation Algorithm Based on YOLOv5 and FCN-DenseNet for Underwater Images

CAO Jian-Rong[1,2], HAN Fa-Tong[1], WANG Ming[1,2], ZHUANG Yuan[1], ZHU Ya-Qin[1], ZHANG Yu-Ting[1]

[1](School of Information and Electrical Engineering, Shandong Jianzhu University, Jinan 250101, China)
[2](Shandong Provincial Key Laboratory of Intelligent Building Technology, Jinan 250101, China)

Abstract: Underwater robots with vision systems cannot operate without the accurate segmentation of underwater objects, but the complex underwater environment and low scene perception and recognition accuracy will seriously affect the performance of object segmentation algorithms. To solve this problem, this study proposes a multi-object segmentation algorithm combining YOLOv5 and FCN-DenseNet, with FCN-DenseNet as the main segmentation framework and YOLOv5 as the object detection framework. In this algorithm, YOLOv5 is employed to detect the locations of objects of each category, and FCN-DenseNet semantic segmentation networks for different categories are input to achieve multi-branch and single-object semantic segmentation. Finally, multi-object semantic segmentation is achieved by the fusion of the segmentation results. In addition, the proposed algorithm is compared with two classical semantic segmentation algorithms, namely, PSPNet and FCN-DenseNet, on the seabed image data set of the Kaggle competition platform. The results demonstrate that compared with PSPNet, the proposed multi-object image semantic segmentation algorithm is improved by 14.9% and 11.6% in *MIoU* and *IoU*, respectively. Compared with the results of FCN-DenseNet, *MIoU* and

① 基金项目: 国家自然科学基金 (62073196, U1806204); 山东省重点研发计划 (2019GSF111054)
收稿时间: 2022-03-22; 修改时间: 2022-04-21; 采用时间: 2022-05-28; csa 在线出版时间: 2022-07-22

IoU are improved by 8% and 7.7%, respectively, which means the proposed algorithm is more suitable for underwater image segmentation.

Key words: YOLOv5; FCN-DenseNet; semantic segmentation; underwater scene identification

1 引言

随着社会科技的发展, 日益加快的工业化、城市化进程, 全球性的陆地资源紧缺和生态环境恶化等问题日益加剧[1], 人类已经逐步开始对水下环境和资源进行探索. 但是水下环境比较复杂[2], 水下探索离不开对水下目标的实时检测, 以及判断出物体的范围和轮廓. 水下图像分割可以准确的分割出图像中的各类目标, 并更好的分析水下场景的安全, 但水下光线昏暗, 对图像清晰度有影响, 因此, 为了更好的保证探险机器的安全和探索, 对水下目标进行语义分割是必不可少.

语义分割主要是将场景图像分割成若干个类别区域, 并对每个区域添加类别标签. 语义分割大致可以分为两种, 一种为传统的语义分割, 另一种为基于深度学习的语义分割[3]. 传统的语义分割是基于图片的边缘信息, 颜色信息以及纹理特征分割成不同的区域. 由于受当时的硬件设备的限制, 传统语义分割算法主要包括基于阈值的分割方法、基于边缘的分割方法等经典的分割方法[4].

相比于传统语义分割算法, 基于深度学习的语义算法可以更准确地、更快地去分割图像, 并具有更好的泛化性. 目前有很多基于深度学习的语义分割算法. Long 等人[5] 提出了基于全卷积网络的语义分割算法 FCN (fully convolutional networks), 该方法可以实现任意大小图片的输入, 同时输出与输入尺寸相同大小的图片; 该方法网络结构简单, 降低了参数计算的复杂度, 使得运行速度更快, 但结果的准确性较低, 分割结果不太清晰. 在此基础上, Jégou 等人[6] 提出了 FCN-DenseNet 网络模型, 该网络模型充分融合每一层的 feature map, 减少了参数的数量, 提高了运算速度, 进一步解决了梯度消失的问题. Zhao 等人[7] 提出了 PSPNet 算法, 该算法的核心是金字塔池化模块 (pyramid pooling module), 通过聚合不同区域上的 feature map, 从而融合语义信息和细节信息. Ronneberger 等人[8] 提出了 U-Net 网络模型, 该网络模型是基于编码器-解码器模型构成的, 即下采样和上采样相结合的分割网络结构, 适用于数据集图片数量少的情况, 但处理尺度变化的问题有些许不足. 谷歌提出的 Deeplab 算法是比较经典的算法之一. DeeplabV1[9] 是应用卷积网络和全连接网络进行语义分割的, 但是池化操作降低了图像分辨率, 从而影响了检测结果. 在此基础之上, DeeplabV2[10] 采用了空洞卷积 (atrous convolution) 进行语义分割, 提高了最后的准确率, 减少了计算量. DeeplabV3[11] 利用 ASPP (atrous spatial pyramid pooling) 模型代替空洞卷积结构, 能够结合上下文信息, 更好地对图像进行语义分割. 当前, 大部分语义分割算法的应用场景为水上环境, 近些年来, 水下环境语义分割的研究也有一定的发展. Arain 等人[12] 提出了通过将稀疏立体点云与单目标语义图像分割相结合来改进基于图像的水下障碍物检测的新方法. Nezla 等人[13] 提出了一种基于 U-Net 的水下语义分割算法, 对其训练模型进行了微调. 马志伟等人[14] 提出一种关注边界的半监督水下语义分割网络 (underwater segmentation network, US-Net), 通过设计伪标签生成器和边界检测子网络提高了目标边界的语义分割效果. Raine 等人[15]提出了一种点标签感知方法, 用于在超像素区域内传播标签以获得增强的地面实况, 以训练 DeepLabv3+网络架构.

当前语义分割的主流是一次性对图片中所有目标类别进行分割. 当图片中含有多个目标种类时, 这些语义分割网络没有关注训练网络权重时不同种类目标之间的影响. 由于水下图像分割干扰因素多, 分割困难, 当前对于水下场景分割和分析的水下语义分割方法比较少. 对于当前大部分水下分割方法来说, 分割的目标比较单一, 当分割目标类别增多时, 其他类别会影响该类别的分割精度, 因此, 对水下场景进行准确度高的多目标语义分割是非常有意义的.

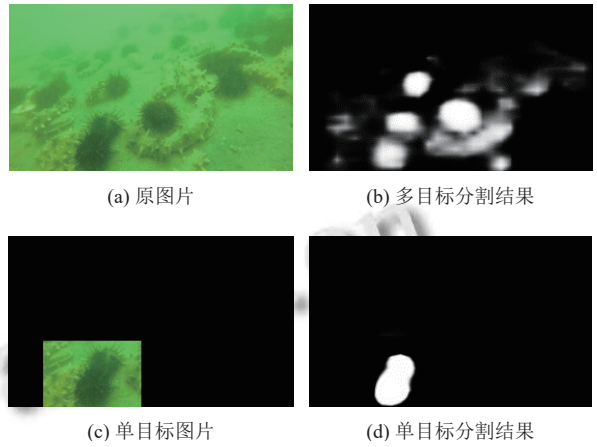
本文基于深度学习的方法, 结合 YOLO 算法[16] 对小目标检测性能好, 多目标分割算法结果相较于单目标分割算法结果不理想等思想, 提出了一种基于 YOLOv5 和 FCN-DenseNet 相结合的融合多分支单目标分割结果实现多目标分割的语义分割算法. 检测出图片中类别目标置信框后, 将置信框截出, 按类别分类, 将各类别图片输入针对不同类别训练好的 FCN-DenseNet 语

义分割网络, 实现单目标语义分割, 再将单目标语义分割结果进行结合, 进而实现多目标语义分割算法.

2 多分支单目标语义分割网络

现阶段, 场景分割中常用的是多目标分割算法, 如图 1 中的多目标分割结果. 对于水下小目标场景, FCN-DenseNet 算法在多目标语义分割中存在目标边缘不清晰、目标分割不准确的问题; 但采用单目标分割的方法, 分割目标边缘清晰, 分割效果较好.

针对以上问题, 本文提出了融合多分支单目标语义分割结果实现多目标语义分割的算法. 首先采用小目标检测准确率较高的 YOLOv5 网络对水下图像中的目标进行检测和分类, 然后截出目标置信框并分类别保存, 使 FCN-DenseNet 网络针对多个类别分别进行训练, 再将某一类别的目标置信框图输入与其类别相对应的训练好的 FCN-DenseNet 网络进行单目标分割, 最后将多个单目标分割结果进行融合获得多目标分割结果图像. 算法流程图如图 2 所示.



(a) 原图片 (b) 多目标分割结果

(c) 单目标图片 (d) 单目标分割结果

图 1 语义分割方法比较

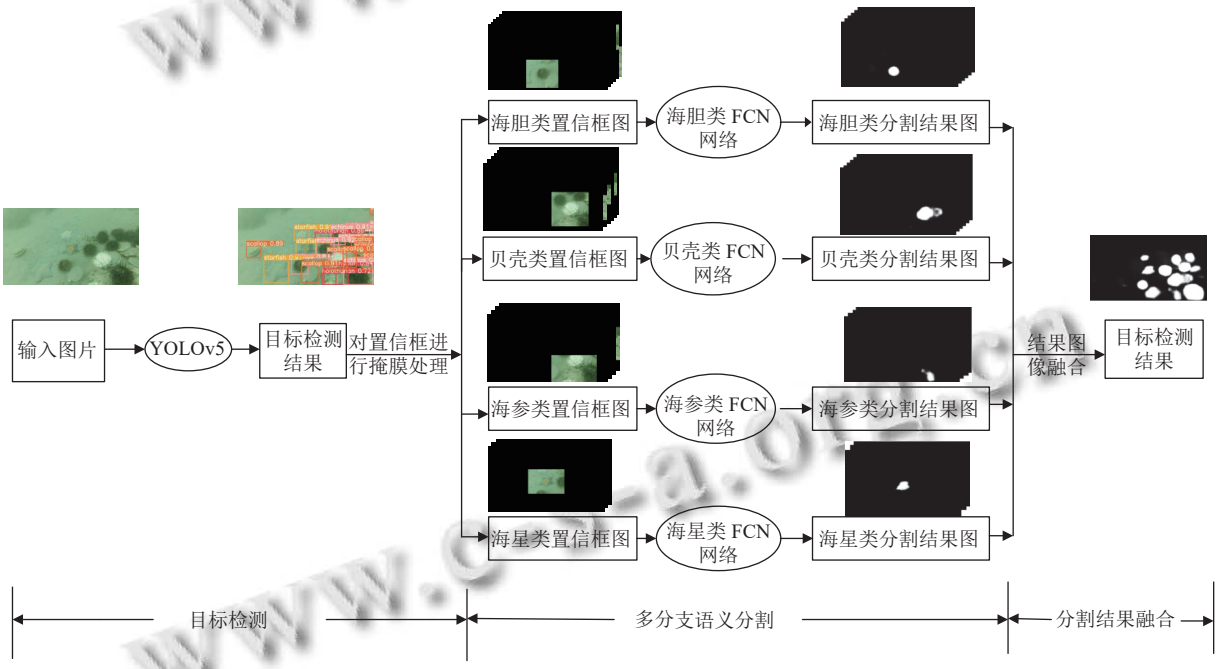


图 2 融合多分支单目标语义分割实现多目标分割算法

目标检测部分主要由 YOLOv5 目标检测网络组成. 向 YOLOv5 网络中输入一张检测图片, 在得到的目标检测结果中, 检测出某个检测框的 4 个坐标 x, y, w, h (x, y 代表检测框的中心坐标, w 代表检测框上下边界到中心点的距离, h 代表检测框左右边界到中心点的距离), 并运用数字图像处理的方法对图像进行掩膜处理, 从而实现只保留该检测框内的图像, 使图像其他像素变为黑色. 实现一张结果图片只包含一个类别目标的置信框, 再将得到的置信框图片按照不同种类进行保存. 为了增加背景信息, 使语义分割结果更加准确, 本文适当扩大了检测框的大小.

多分支语义分割部分主要由 FCN-DenseNet 语义分割网络组成. FCN-DenseNet 语义分割网络需要针对分割的每一类目标进行训练. 训练好的 FCN-DenseNet 语义分割网络从 YOLOv5 保存的图片文件夹中按照类别调取图片, 将一类中所有置信框图片送入对应类别

的 FCN-DenseNet 语义分割网络中, 对置信框图片中框内目标进行语义分割. 这样多张置信框图片分步进入网络, 可以实现一次分割出一类目标图像, 从而减少其他类目标对该类目标分割结果的影响.

分割结果融合部分主要是利用像素相加的思想. 在数字图像中, 白色像素默认值为 255, 而黑色像素的默认值为 0. 假设视频序列$V=\{v_1,v_2,v_3,\cdots,v_n\}$是由 n 帧图像组成, v_i代表其中第 i 帧图像. v_i经过 YOLOv5 网络后, 输出为 p 类目标框图, 每类目标有 q 个置信框图. 将每一类目标中 q 个置信框图进行适当放大后, 分别送入对应的训练好的 FCN-DenseNet 网络中, 输出结果为 q 个单目标分割结果 (为二值图像). 若每个分割结果为F_k, 则单类目标分割结果为:

$$F_{ij}=\sum_{k=1}^{q}F_k,\quad k=1,\cdots,q \tag{1}$$

其中, F_{ij}为第i张图片, 第j类目标合并的结果, 其中 $i=1,\cdots,n$, $j=1,\cdots,p$. 设第 i 张图像最终分割结果为 S_i, 则:

$$S_i=\sum_{j=1}^{p}F_{ij} \tag{2}$$

3 实验分析

3.1 实验环境

为了验证本文所提算法的有效性, 在 Windows 环境下搭建了基于 PyTorch 的深度学习框架. 本次实验所用的计算机 CPU 为 Intel Core i7-5800H, GPU 为 GeForce RTX 3060.

3.2 数据集的采集及预处理

本文算法测试数据集采用 Kaggle 竞赛平台上的海底图片数据集, 该数据集是潜水员摄像机拍摄的海底不同场景. 在海底数据集中共包含 6 575 张图片, 主要包括海胆 (echinus)、海星 (starfish)、扇贝 (scallop) 以及海参 (holothurian) 4 类待检测的目标. 本实验用数据集中所有的海底图片对 YOLOv5 网络进行训练, 从中随机挑取 600 张图片对 FCN-DenseNet 网络进行全类别训练得到全类别权重; 随机挑选 400 张图片, 分成 4 组, 分别训练 4 类目标得到单目标权重. 为了使图像中目标的边界更加明显, 提高目标检测和语义分割的精确度, 将所有数据集图片进行了图像锐化操作, 方便区分背景和目标, 从而得到更好的分割效果.

3.3 评判标准

在语义分割算法中应用比较广泛的评判标准有均交并比 (nean intersection over union, *MIoU*) 和交并比 (intersection over union, *IoU*)[17]. *IoU* 语义分割问题中的真实值 (ground truth) 和预测值 (predicted segmentation) 是两个重要的集合, 分别用 A 和 B 表示, 则 *IoU* 的计算公式如式 (3) 所示:

$$IoU(A,B)=\frac{|A\cap B|}{|A\cup B|} \tag{3}$$

MIoU 是通过计算所有类别真实值和预测值交集和并集之比的平均值来评判分割性能的好坏. 计算公式如式 (4) 所示:

$$MIoU=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}(p_{ji}-p_{ii})} \tag{4}$$

其中, k 表示除背景外的类别数, i 表示分割问题中的真实值, j 表示预测值, p_{ij}表示将 i 预测为 j 的像素数, p_{ii}表示将 i 预测为 i 的像素数, p_{ji}表示将 j 预测为 i 的像素数. 进而式 (4) 等价于式 (5), 式 (5) 如下所示:

$$MIoU=\frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{FN+FP+TP} \tag{5}$$

其中, *TP* 表示图像中目标类别检测为目标类别的像素值; *FP* 表示非目标类别检测为目标类别的像素值; *FN* 表示为目标类别检测为非目标类别的像素值.

3.4 置信框扩大倍数实验

在 FCN-DenseNet 语义分割网络中, 待分割目标与背景之间的差异会影响分割结果的准确性. YOLOv5 网络更精确地检测目标, 所以置信框更贴近目标的边界, 得到的置信框图片直接进行语义分割结果比较差. 因此, 在本实验中对检测框进行适当的扩大, 从而使检测框中包含更多的背景信息, 利于 FCN-DenseNet 网络的分割.

为了更好地检测出扩大的倍数对分割结果好坏的影响, 本次实验分别测试了 1 倍、1.2 倍、1.5 倍、1.7 倍和 2 倍 5 种倍数. 因为有些置信框过大, 扩大两倍以上会包含其他较远置信框信息, 从而实现不了准确分割类别目标边界的目的, 所以测试最大倍数为 2 倍. 在扩大过程中, 加入的其他种类目标可作为背景, 加入的同种类目标也可以继续分割, 不影响最后合成总的分割结果. 具体的分割结果图如图 3 所示.

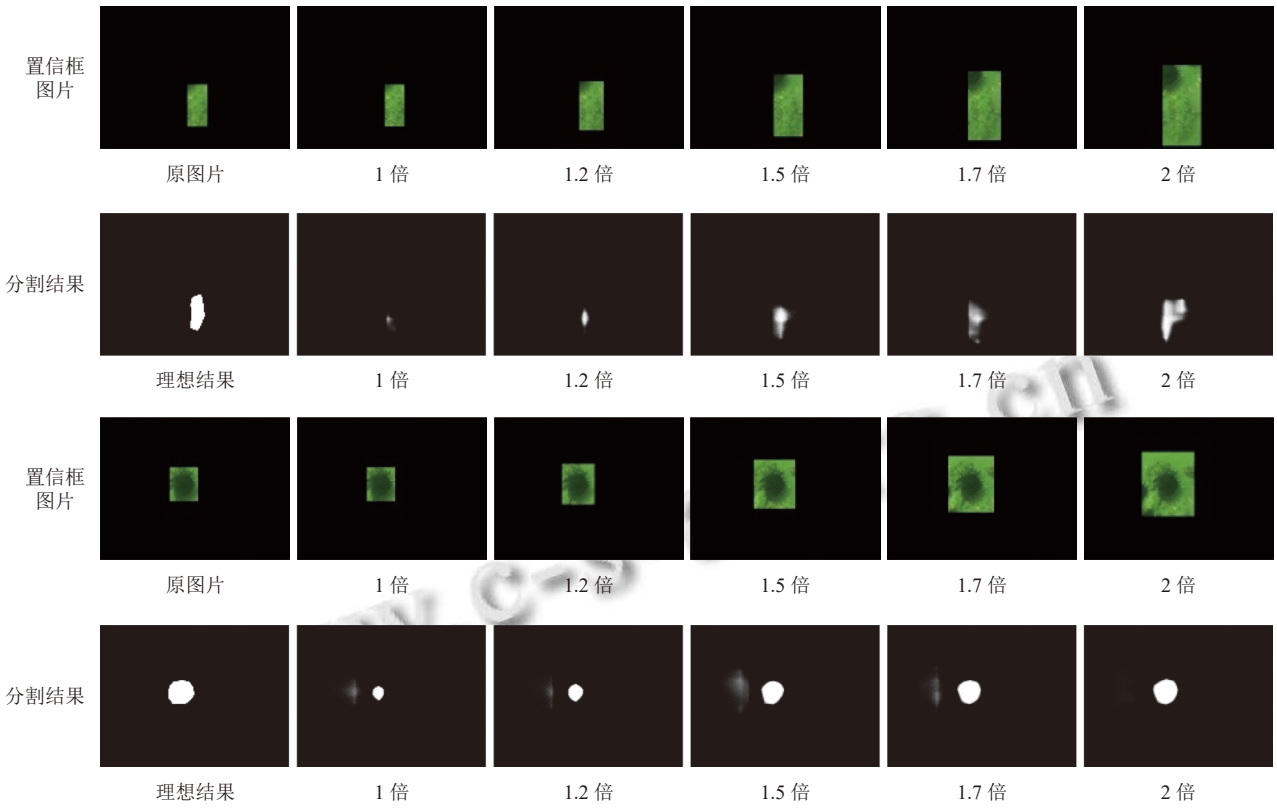


图 3 置信框扩大倍数对分割结果影响

从图 3 中两张置信框分割结果图中可以看出, 置信框扩大为原来的 2 倍时, 此时的分割结果和理想结果差异最小. 因此, 本实验置信框扩大倍数为 2 倍.

3.5 算法对比实验

在本次实验过程中, 为了验证本文算法的优越性, 将本文算法与经典的 FCN-DenseNet 网络与 PSPNet 网络进行对比. 首先, 利用 YOLOv5 算法在海底数据集上进行训练, 初始学习率为 0.001, 迭代次数为 200. 再利用训练好的 YOLOv5 网络对图片进行目标检测, 在目标检测过程中, 检测代码的置信度阈值取 0.5, 即当预测置信度高于 0.5 时才会标记该目标. 为了让后续 FCN-DenseNet 网络更好的分辨目标与背景, 将 YOLOv5 的检测框扩大 2 倍, 从而提高了语义分割的准确率. 再利用 FCN-DenseNet 网络对各类置信框图片进行多次训练. 实验结果以 *IoU* 和 *MIoU* 为评价标准对算法进行性能分析, 为了更好地计算实验结果的 *MIoU* 和 *IoU*, 将所有分割出的结果都设为白色.

不同算法检测结果的 *IoU* 和 *MIoU* 如表 1 所示. 当使用相同数据集进行训练并检测同一张图片时, 本文算法的分割效果更佳, 本文算法相比较于 FCN-DenseNet 原始算法在 *IoU* 和 *MIoU* 数值上分别提高了 7.7%、8%; 相较于 PSPNet 算法在 *IoU* 和 *MIoU* 数值上分别提高了 11.6%、14.9%.

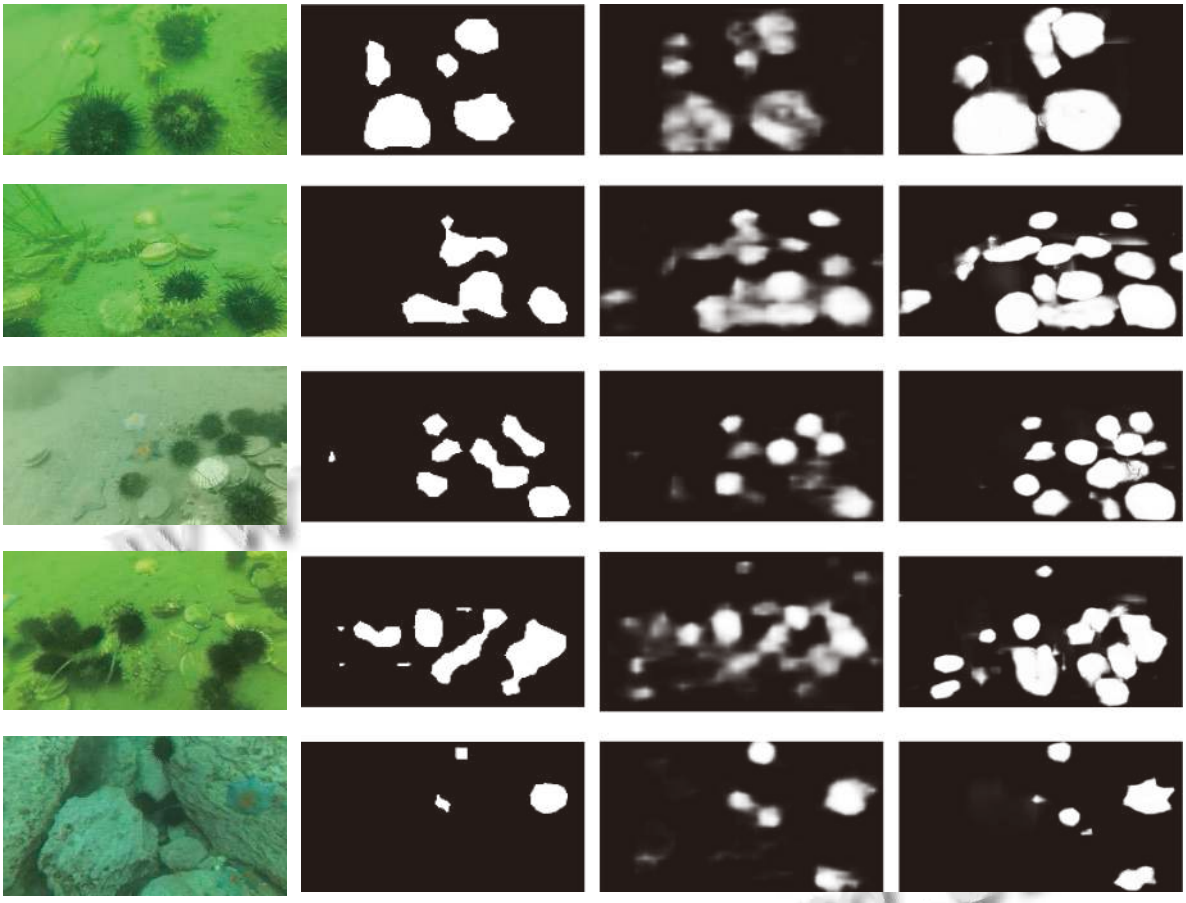
表 1 不同方法下检测结果的 *IoU* 和 *MIoU* 对比 (%)

方法	*IoU*	*MIoU*
FCN-DenseNet	48.7	66.3
PSPNet	41.5	62.7
本文算法	56.4	74.3

FCN-DenseNet 的原始算法和 PSPNet 算法是一次性对多目标进行训练分割, 由于每个种类之间的遮挡、干扰等问题, 对模型的性能会有影响, 并且在分割时, 不同的目标类别也会有一定的干扰, 从而可能降低分割精度. 本文算法是将待分类的类别单独进行训练, 得到的权重信息可以更好的分割该类别. 在分割过程中, 输入分割网络的图片为只含有待分割目标的置信框图片, 减少了其他类别对该类别分割时的干扰, 更好的提高了分割性能. 这也是本文算法表现更佳的原因所在.

图 4 是部分实验结果, 从中可看出本文算法相比较于 FCN-DenseNet 算法和 PSPNet 算法性能更好, 本

文算法检测出的目标更加全面, 且目标边界更加清晰, 这说明本文算法在提取图片特征时提取到了更多的细节, 减少了其他类别目标对该类别目标的影响, 改善了因不同种类的影响而忽略了本种类目标特征的问题. 本文算法更适用于水下场景分割, 生成的分割图像更加清晰, 可以更好的分辨出目标.



(a) 原图片 (b) PSPNet (c) FCN-DenseNet (d) 本文算法

图 4 不同算法检测结果比

4 结论

本文针对语义分割算法在水下场景复杂、场景感知精度不高等问题, 提出了一种基于 YOLOv5 和 FCN-DenseNet 相结合的融合多分支单目标分割结果实现多目标分割的语义分割算法. 将 YOLOv5 网络输出结果中的置信框按类别保存, 再输入针对不同类别训练好的 FCN-DenseNet 网络中进行语义分割, 得到一个种类的单目标语义分割结果, 再将不同种类单目标语义分割结果进行结合, 从而实现多目标语义分割算法. 实验结果显示, 本文算法相比较于 FCN-DenseNet 原始算法在 *IoU* 和 *MIoU* 数值上分别提高了 7.7%、8%; 相较于 PSPNet 算法在 *IoU* 和 *MIoU* 数值上分别提高了 11.6%、14.9%, 通过 3 种算法对比, 验证了本算法的性能较好.

参考文献

1 廖泓舟. 基于深度卷积特征的水下静目标识别方法研究 [硕士学位论文]. 哈尔滨: 哈尔滨工程大学, 2019.

2 方明, 刘小晗, 付飞蚺. 基于注意力的多尺度水下图像增强网络. 电子与信息学报, 2021, 43(12): 3513–3521. [doi: 10.11999/JEIT200836]

3 张峻宁, 苏群星, 王成, 等. 一种改进变换网络的域自适应语义分割网络. 上海交通大学学报, 2021, 55(9): 1158–1168. [doi: 10.16183/j.cnki.jsjtu.2019.307]

4 张鑫, 姚庆安, 赵健, 等. 全卷积神经网络图像语义分割方法综述. 计算机工程与应用, 2022, 58(8): 45–57. [doi: 10.3778/j.issn.1002-8331.2109-0091]

5 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3431–3440.

6 Jégou S, Drozdzal M, Vazquez D, *et al*. The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu: IEEE, 2017. 1175–1183.

7 Zhao HS, Shi JP, Qi XJ, *et al*. Pyramid scene parsing network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 6230–6239.

8 Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241.

9 Chen LC, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFs. Computer Science, 2014, (4): 357–361.

10 Chen LC, Papandreou G, Kokkinos I, *et al*. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]

11 Chen LC, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation. arXiv: 1706.05587, 2017.

12 Arain B, McCool C, Rigby P, *et al*. Improving underwater obstacle detection using semantic image segmentation. 2019 International Conference on Robotics and Automation (ICRA). Montreal: IEEE, 2019. 9271–9277.

13 Nezla NA, Haridas TPM, Supriya MH. Semantic segmentation of underwater images using UNet architecture based deep convolutional encoder decoder model. 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS). Coimbatore: IEEE, 2021. 28–33.

14 马志伟, 李豪杰, 樊鑫, 等. 真实场景水下语义分割方法及数据集. 北京航空航天大学学报, 2022, 48(8): 1515–1524. [doi: 10.13700/j.bh.1001-5965.2021.0527]

15 Raine S, Marchant R, Kusy B, *et al*. Point label aware superpixels for multi-species segmentation of underwater imagery. arXiv:2202.13487, 2022.

16 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788.

17 张灿龙, 程庆贺, 李志欣, 等. 门控多层融合的实时语义分割. 计算机辅助设计与图形学学报, 2020, 32(9): 1442–1449.

(校对责编: 孙君艳)