

# YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors

李振鹏

## 摘要

YOLOv7 网络是用于计算机视觉任务的最快、最准确的实时对象检测算法模型，同时支持实例分割任务。YOLO 为 “You Only Look Once” 单词的首字母缩写，它是典型的 One-Stage 目标检测器。YOLOv7 算法在不增加推理成本的情况下，大大提高了实时目标检测精度。YOLOv7 在 5FPS 到 160FPS 范围内的推理速度和准确度都超过了所有已知的物体检测器。总的来说，YOLOv7 提供了更快更强的网络架构，提供了更有效的特征集成方法、更准确的目标检测性能、更鲁棒的损失函数以及更高的标签分配策略和模型训练效率。在大坝的表面缺陷检测项目中，由于图像分辨率较大，因此并不能直接采用 YOLOv7 进行训练、推理，因此在原有代码的基础上，新增图像分割、图像合并代码，使其成为 end-to-end 的推理 pipeline。

**关键词：**YOLOv7；目标检测；表面缺陷

## 1 引言

我国已有水库大坝将近 10 万座，是世界上拥有水库大坝最多的国家，也是世界上拥有 200 米级以上高坝最多的国家。目前世界建成的 200 米级以上高坝 77 座，我国有 22 座，占比 26%。在建的 200 米级以上高坝 19 座，我国就有 12 座，占比 63%<sup>[1]</sup>。图 1 所示为云南大理某水电厂大坝。混凝土大坝建成后，在长期运行的过程中，由于坝体混凝土老化、钢筋锈蚀、或因地震、人为活动、生物破坏等因素，大坝表面会产生缺陷或病害，如果不及时加以干预和处理将影响大坝的安全运行，危及人民生命财产安全<sup>[2]</sup>。因此需要定期对大坝表面缺陷进行检测，为其安全性评估和后续加固补强提供依据。



图 1：云南大理某水电厂大坝

## 2 相关工作

由于大坝高度高、面积大的结构特点，传统的人工检测方法在进行大坝表面的缺陷检测时，存在风险高、效率低、费用高等缺点。随着无人机、计算机视觉技术以及人工智能技术的发展，采用无人机进行大坝表面缺陷检测成为可能。无人机按照预先设定的大坝巡检路径，进行抵近拍照采集大坝图像，然后通过计算机视觉和图像处理技术对图像数据进行分析，进而识别出大坝表面的裂缝，并提供其裂缝的位置、大小等信息<sup>[3]</sup>。该方法极大的提升了大坝的检测效率和精度，降低了作业风险。

近年来随着计算机视觉技术的快速发展，基于计算机视觉的缺陷自动检测已经越来越受到世界各地研究人员的关注。小波变换、快速傅立叶变换、Sobel 算法和 Canny 算法等在前期的缺陷检测算法中占据了主导地位<sup>[4]</sup>。而后随着深度学习技术的快速发展，人们开始利用深度学习技术对图像处理分析，并逐步应用大型建筑的表面裂缝检测任务中。

## 3 本文方法

### 3.1 本文方法概述

YOLOv7 框架<sup>[5]</sup>结构图如图 2 所示，网络主要由两部分组成，包括骨干网络、检测头网络。其中 YOLOv7 的检测头网络包含了颈部网络。骨干网络是一种卷积神经网络，通过多次卷积、池化等操作提取图像隐含特征；颈部网络为 FPN-PAN 结构，颈部网络通过融合骨干特征网络输出的不同大小的特征图，来获得更多的上下文信息，并减少信息丢失；最后检测头网络根据颈部网络所生成的新的特征映射实现对目标的检测与分类。

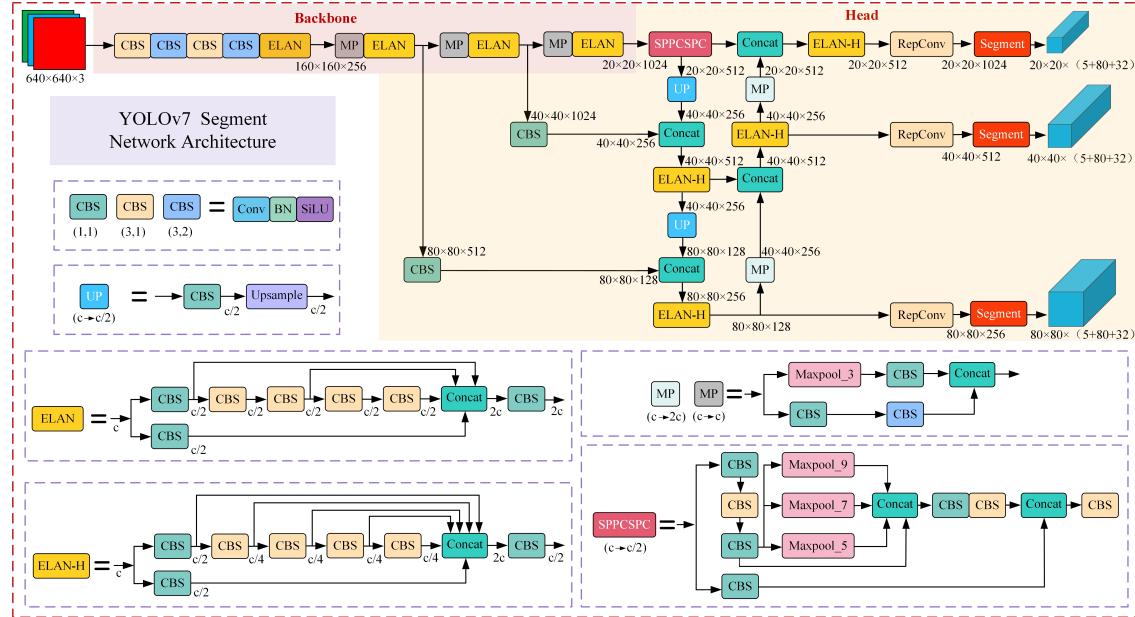


图 2: YOLOv7 分割网络结构图

当一张图像传入到 YOLOv7 网络时，网络首先会对输入的图片调整为 640×640 大小，然后输入到 Backbone 网络中，接下来经过 Head 网络输出三层不同尺度大小的 Feature map，经过重参数化模块 RepConv，最后经 Segment 模块重新调整 tensor 维度，得到输出的预测结果。图 3 中若以 COCO 数据集为例，输出为 80 个类别，然后每个输出 (x, y, w, h, conf) 即锚框中心坐标、宽、高以及置信度，3 是指的锚框的数量，32 指的是 mask 点坐标，一共包含 16 对 (x, y) 坐标。因此每一层的输出为 (5+80+32)×3，再乘上 Feature map 的大小就得到模型最终输出结果。

### 3.2 CBS 模块

CBS 模块这里有三种颜色，如图 3 所示，三种颜色代表它们的卷积核（kernel）和步长（stride）不同。第一个 CBS 模块是一个  $1 \times 1$  的卷积，stride 步长为 1；第二个 CBS 模块是一个  $3 \times 3$  的卷积，stride 步长为 1；第三个 CBS 模块是一个  $3 \times 3$  的卷积，stride(步长为 2)。 $1 \times 1$  的卷积主要用来改变通道数； $3 \times 3$  的卷积，步长为 1，主要用来特征提取； $3 \times 3$  的卷积，步长为 2，主要用来下采样。

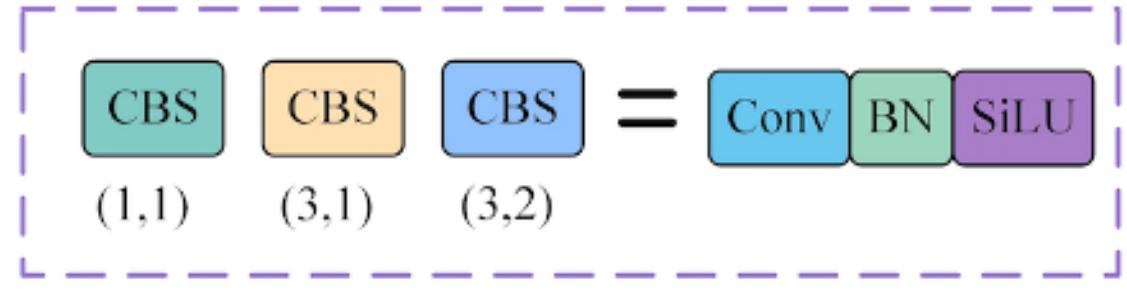


图 3: CBS 模块

### 3.3 MP 模块

MP 模块共有两种，如图 4 所示，颜色不同代表该模块输入、输出的通道数不同，第一个 MP 模块的输入通道数为  $c$ ，输出通道数变为  $2c$ ；第二个 MP 模块的输入、输出通道数不变。MP 模块有两个分支，作用是进行下采样。第一条分支先经过一个 Maxpool 最大池化。最大值化的作用是下采样，然后再经过一个  $1 \times 1$  的卷积进行通道数的改变。第二条分支先经过一个  $1 \times 1$  的卷积，做通道数的调整，然后再经过一个  $3 \times 3$  卷积核、步长为 2 的卷积块，该卷积块用于下采样。最后把第一个分支和第二分支的结果加在一起，得到了最终下采样的结果。

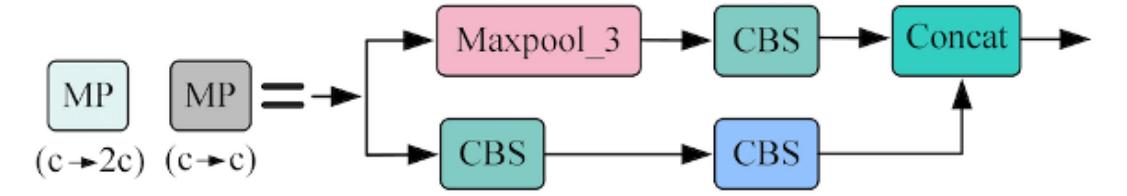


图 4: MP 模块

### 3.4 ELAN 模块

ELAN 模块是一个高效的网络结构，如图 5 所示，它通过控制最短和最长的梯度路径，使网络能够学习到更多的潜在特征，并且具有更强的鲁棒性。ELAN 有两条分支，第一条分支是经过一个  $1 \times 1$  的卷积做通道数的变化。第二条分支首先经过一个  $1 \times 1$  的卷积模块，做通道数的变化。然后再经过四个  $3 \times 3$  的卷积模块，做特征提取，最后把四个特征叠加在一起得到最后的特征提取结果。

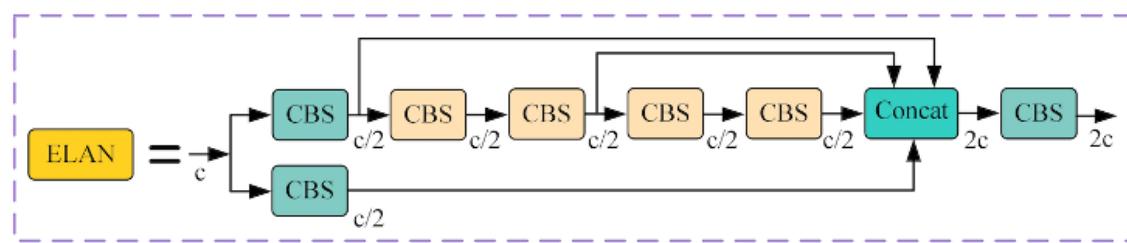


图 5: ELAN 模块

### 3.5 ELANH 模块

ELAN-H 模块和 ELAN 模块非常相似，不同之处在于它是在第二条分支的时候选取的输出数量不同。如图 6 所示，ELAN 模块选取了三个输出进行最后的相加，而 ELAN-H 模块选取了五个输出进行相加。其次 ELAN-H 的输入、输出通道数与 ELAN 模块不同。

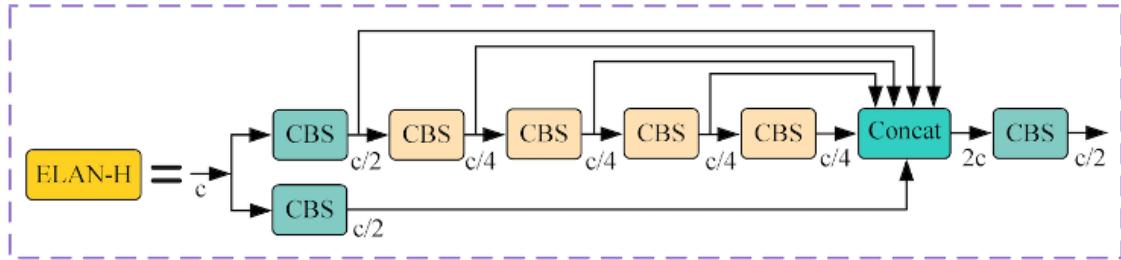


图 6: ELANH 模块

### 3.6 SPPCSPC 模块

SPP 的作用是能够增大感受野，使其在不同的分辨率特征图上提取信息，它是通过最大池化来获得不同感受野。在第一条分支中，经过 Maxpool 的有四条分支，对应的池化 Kernel 分别是 5, 7, 9, 1。四个不同尺度的最大池化有四种感受野，用来区别于大目标和小目标。SPPCSPC 模块，如图 7 所示，首先将特征分为两部分，其中的一个部分进行常规的处理，另外一个部分进行 SPP 结构的处理，最后把这两个部分合并在一起，这样就能够减少一半的计算量，使得速度变得快，精度反而会提升。

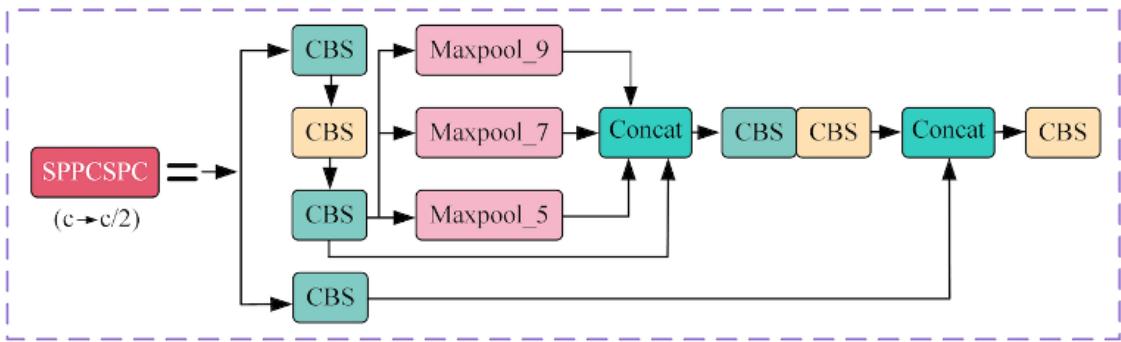


图 7: SPPCSPC 模块

## 4 复现细节

### 4.1 与已有开源代码对比

在复现过程中，仅仅使用了原作者的源代码进行了模型的训练，并将模型导出为 ONNX 格式的文件。自己的工作为：基于导出的 ONNX 文件，编写相应的图像推理代码，将图像预处理、模型推理、后处理封装为 YOLOv7Seg 类，方便调用。同时为了实现大尺度、高分辨率图像的端到端推理，编写图像切割、图像合并代码，并将其融合在一起。为了用户使用，基于 PyQt 开发出一款 PC 端软件，支持 mac/Windows 操作系统，同时支持 cpu/gpu 推理。

### 4.2 实验环境搭建

实验采用的硬件配置为 Intel Core i7-12900k 处理器、Nvidia 3090Ti 显卡（24G 显存），电脑内存为 32G。软件环境为 CUDA 10.0 版本、cuDNN 9.1 版本以及 OpenCV-Python 4.5 版本。所用的深度学习框架为 Pytorch，采用 Python 编程语言编写程序代码。

在模型的超参数选取方面，YOLOv7 算法通过随机梯度下降法（SGD）以端到端的方式进行训练，

其中动量因子（Momentum）初始值设置为 0.937，动量衰减率（weight decay）设置为 0.0005；初始学习率（Learning rate）参数设置为 0.01，每经过 100 个 epoch 学习率衰减为原来的 1/10；IoU 阈值设置为 0.05；模型训练的 batch size 设置为 32；迭代训练次数 epoch 的值设置为 600。

### 4.3 界面分析与使用说明

QT 是一个跨平台的 C++ 开发库，主要用来开发图形用户界面（Graphical User Interface, GUI）程序。QT 具有较好的跨平台、高性能等优势。为了对模型的推理引擎进行封装，因此，大坝表面缺陷检测系统采用 QT 进行图形化界面程序开发。由于模型训练结束后权重文件格式为 pt 格式，不包含网络结构，只包含模型的权重参数，不方便模型直接推理。因此首先需要将基于 Pytorch 框架训练的 pt 格式的模型文件转为开放神经网络交换（Open Neural Network Exchange, ONNX）格式的模型文件，该文件格式支持 cpu、gpu 推理，且单一文件包含网络模型结构以及相应网络层的权重参数，使用起来较为方便。

构建可视化界面时，首先利用 QT Designer 进行软件界面布局及窗口设计，生成 xml 格式的 ui 文件，ui 文件通过 uic 工具编译生成 \*.py 格式的文件。通过对 python 格式的类进行调用，即可实现加载界面到相应的对象上。通过使用 PyCharm 创建工程，加载界面相关的 py 文件、模型引擎文件等，并编写软件控件相关的槽函数。大坝表面缺陷检测系统软件包含主线程和子线程，主线程包括软件界面的显示，子线程包括图像预处理、模型推理、结果解析三部分。加载后的图像首先会调用 OpenCV 对输入图像进行预处理，图像预处理包括对图像缩放和图像归一化。预处理后的图像传送至 ONNX 推理引擎进行推理，推理结果包括目标框类别信息、目标框坐标信息、类别置信度信息以及 Mask 掩模信息，然后对模型预测结果进行解析，解析后的结果通过 OpenCV 算子库在原图像进行绘制目标框以及掩模区域，并在目标框上方显示类别名称和置信度大小。最后将渲染后的图像在客户端软件的窗口上进行显示。

软件界面如图 8 所示，包含图片路径、一键切图、模型路径、开始检测、结果合并按键。在使用时，首先点图片路径按钮，加载一张大图，然后点击一键切图按钮，此时会将加载的大图进行切分，切分为  $416 \times 416$  大小的小图，然后点击模型路径，选择预先导出的 ONNX 模型，接下来点击开始检测按钮，程序会自动执行开始对图像进行推理检测，并将检测结果实时滚动显示在软件的中间窗口位置。待所有小图检测结束后，点击结果合并按钮，会讲所有的检测结果还原至大图，这样就可以清晰的看见原图像上对应的所有缺陷。点击退出按钮，则会自动退出程序。



图 8: 软件界面

## 5 实验结果分析

采用 YOLOv5 分割算法和 YOLOv7 分割算法进行实验对比，对比所用的数据集为 3.1 小节的大坝分割数据集，为保证实验公平性，两个算法的参数设置均为 3.2 小节提及的参数，最终得到表 1 所示结果。由表 1 可以看到，YOLOv7 分割算法的准确率高 YOLOv5 分割算法 5.2%，但是检测速度上稍微低于 YOLOv5 分割算法。原因是 YOLOv7 提升准确率的同时也增加了模型的计算量，因此模型的推理时间变长。

表 1: 检测算法性能比较

模型	mAP@0.5(%)	检测速度 (fps)
YOLOv5	72.7	38
YOLOv7	77.9	40

图 9 所示为两个分割模型的实际推理效果，左边为缺陷的 Ground truth，中间为 YOLOv5 推理结果图，右边为 YOLOv7 推理效果图，两者在轻微的裂缝缺陷上面，很多都没有检测出来，原因可能是由于数据量太少，模型尚未学习到轻微裂缝的特征信息。相对于 YOLOv5 分割算法，YOLOv7 算法在识别到的裂缝上的置信度高于 YOLOv5。

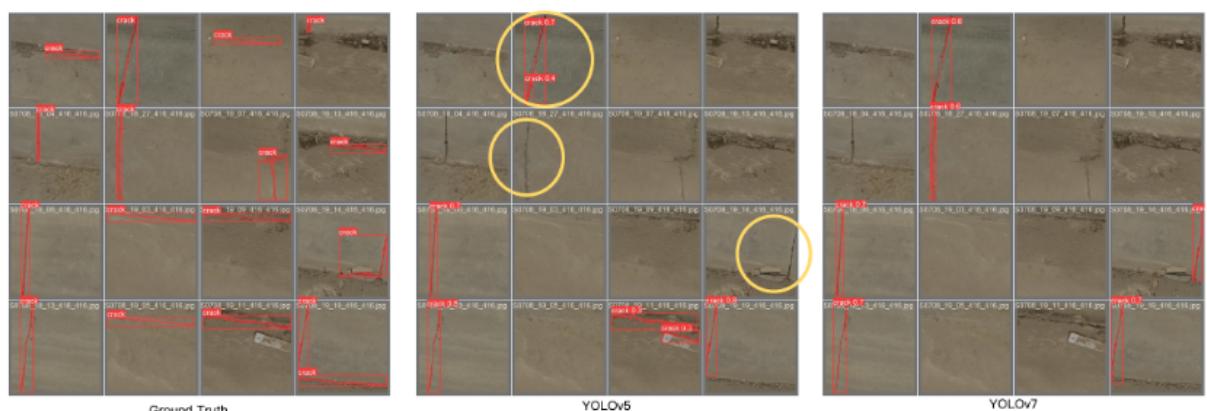


图 9: 算法结果对比

## 6 总结与展望

在本次复现的过程中，更多的是做了应用层面的程序编写开发，学习到了很多基础知识。对于YOLOv7分割网络，仍然有很大的改进空间，一些涨点的 trick 如注意力机制、数据增强等实验性工作做的不够充分。目前采用的整体思路是对图像先进行切分然后再推理，最后结果进行合并，这种方法运行效率较低，根据调研的 cvpr 等最新文献，大家开始转向于不切图，通过对不同尺度的特征图信息融合提取，直接在大尺度、高分辨率图像预测输出小物体的目标位置及类别，这将是我接下来改进优化的一个大的方向。

## 参考文献

- [1] LI L, ZHANG H, PANG J, et al. Dam surface crack detection based on deep learning[C]//Proceedings of the 2019 International Conference on Robotics, Intelligent Control and Artificial Intelligence. 2019: 738-743.
- [2] FAN X, WU J, SHI P, et al. A novel automatic dam crack detection algorithm based on local-global clustering[J]. Multimedia Tools and Applications, 2018, 77: 26581-26599.
- [3] LI Y, BAO T, XU B, et al. A deep residual neural network framework with transfer learning for concrete dams patch-level crack classification and weakly-supervised localization[J]. Measurement, 2022, 188: 110641.
- [4] FENG C, ZHANG H, WANG H, et al. Automatic pixel-level crack detection on dam surface using deep convolutional network[J]. Sensors, 2020, 20(7): 2069.
- [5] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[J]. arXiv preprint arXiv:2207.02696, 2022.