文章编号：2096-1472(2023)-01-55-04

引入注意力机制的改进型YOLOv5网络研究

曾 凯，李 响，陈宏君，文继锋

(南京南瑞继保电气有限公司，江苏 南京 211102)
✉zengkai2@nrec.com; lix@nrec.com; chenhj@nrec.com; wenjf@nrec.com

摘 要：为了提升复杂多尺度目标检测任务下的分类及定位准确度，在基准的YOLOv5目标检测算法基础上，设计了四组引入注意力机制模块的改进型YOLOv5网络，并在变电站内复杂多尺度检测场景数据集上进行对比实验。实验结果表明，相较于基准YOLOv5网络，SwinTrans-YOLOv5网络的*mAP*指标提升达9.0%，但模型运算量高达1,061.6 GFLOPS(每秒千兆次浮点运算)；CA-YOLOv5网络的*mAP*指标提升也达到4.1%，模型运算量仅需115.8 GFLOPS。因此，在硬件算力充足的情况下使用SwinTrans-YOLOv5网络可以获得更高的检测精度，但在硬件算力不足的情况下使用CA-YOLOv5网络，则实现了检测精度和速度间较好的平衡。

关键词：注意力机制；YOLOv5网络；目标检测；Transformer；复杂多尺度
中图分类号：TP391 **文献标识码**：A

Research on the Improved YOLOv5 Network with Attention Mechanism

ZENG Kai, LI Xiang, CHEN Hongjun, WEN Jifeng

(*NR Electric Co., Ltd., Nanjing* 211102, *China*)
zengkai2@nrec.com; lix@nrec.com; chenhj@nrec.com; wenjf@nrec.com

Abstract: In order to improve the classification and positioning accuracy of complex and multi-scale object detection tasks, this paper proposes to design four groups of improved YOLOv5 networks with attention mechanism modules based on the benchmark YOLOv5 objet detection algorithm, and their comparative tests are conducted on multi-scale detection datasets in the substation. Test results show that compared with the benchmark YOLOv5 network, the *mAP* index of SwinTrans-YOLOv5 network is improved by 9.0%, but the model calculation amount is as high as 1061.6 GFLOPS (Giga Floating-point Operations Per Second); the *mAP* index of CA-YOLOv5 network is also improved by 4.1%, and only 115.8 GFLOPS is needed. Therefore, using the SwinTrans-YOLOv5 network can achieve higher detection accuracy when the hardware computing power is sufficient, but using the CA-YOLOv5 network when the hardware computing power is insufficient can achieve a good balance between detection accuracy and speed.

Keywords: attention mechanism; YOLOv5 algorithm; object detection; Transformer; complex and multi-scale

1 引言(Introduction)

基于深度学习算法的目标检测技术已在计算机视觉领域得到广泛应用[1-3]，然而在面对复杂场景下的多尺度目标检测时，仍然存在识别精度不够高、定位不够准确的问题。目前，大部分研究是在基准检测模型上进行多尺度特征增强和引入混合注意力机制模块。比如，林森等[4]提出了一种基于注意力机制与改进YOLOv5网络的水下珍品检测方法，其主要思想是利用CBAM注意力机制模块对特征提取网络进行改进；刘万军等[5]针对背景复杂的遥感图像识别任务，在Faster R-CNN基准网络上提出多尺度特征增强及密集连接网络，有效地解决了目标漏检的问题。综合各方面的研究表明，引入注意力机制模块可以有效地提升目标检测的准确率。

注意力机制可分为基于卷积神经网络的注意力机制[6]和基于Transformer网络的自注意力机制[7]两大类。注意力机制模块众多、模型性能差异大，对比评估一些新型且有效的注意力机制模块，对提升复杂多尺度目标的检测性能是非常有意义的。

本文选取YOLOv5网络作为基准模型，在网络中引入多种先进的注意力机制模块，设计并实验对比分析了几种改进型网络的性能表现。

基金项目：南瑞集团有限公司科技项目(JS2001712).

2 改进型YOLOv5网络的设计(Design of improved YOLOv5 network)

2.1 标准YOLOv5目标检测网络结构

YOLOv5网络在工业落地应用中表现出极其优秀的检测性能和推广价值，它整合了大量的计算机视觉前沿技术，显著改善了对象检测的性能，提升了模型训练的速度及模型应用的便利度。

YOLOv5网络结构示意图如图1所示。

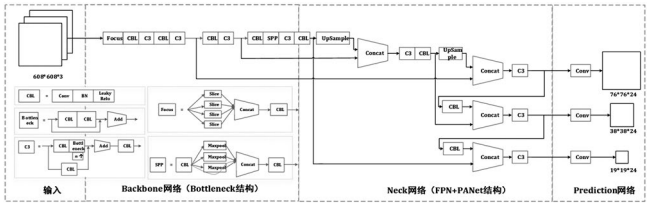

图1 YOLOv5网络结构示意图

Fig.1 Diagram of YOLOv5 network structure

图1中的基准网络主要由骨干特征提取网络(Backbone网络)、颈部特征融合网络(Neck网络)和检测头部预测网络(Prediction网络)组成，分别主要使用基于瓶颈(Bottleneck)结构的C3为Backbone，基于多特征图融合的FPN+PANet结构为Neck，以及基于检测目标的位置和类别进行回归、分类任务的YOLO检测头为Prediction。

2.2 引入卷积神经网络注意力机制的改进型YOLOv5网络设计

2.2.1 引入协同注意力机制网络的改进型YOLOv5网络

协同注意力机制网络(Coordinate Attention，CA)[8]，它引入了一种新的注意块结构，该结构不仅能捕获跨通道的信息，还能捕获方向感知和位置感知的信息，这能帮助模型更加精准地定位和识别感兴趣的目标。

基于CA模块的典型网络结构如图2所示。

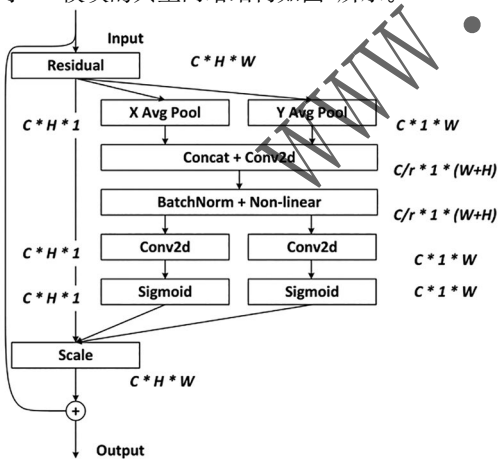

图2 CA注意力模块的典型网络结构示意图

Fig.2 Network structure diagram of CA attention module

图2中的网络结构首先分别对水平方向和垂直方向进行全局平均池化，得到两个1 维向量，在空间维度上拼接并经过1×1的卷积压缩通道数，然后通过批量归一化(BN)和非线性激活函数编码垂直方向和水平方向的空间信息，接着在空间维度上将BN和激活函数的输出拆分成两个特征图，再各自通过一个1×1的卷积调整通道得到和输入特征图一样通道数的融合了注意力机制的特征图。

本文将CA网络模块应用在YOLOv5网络的骨干特征提取网络的每个多尺度特征输出的位置，用于对每个尺度下特征图的各通道进行特征重标定，以提升原始YOLOv5网络的特征提取能力。设计的引入CA网络模块的改进型YOLOv5网络结构示意图如图3所示，命名为CA−YOLOv5网络。

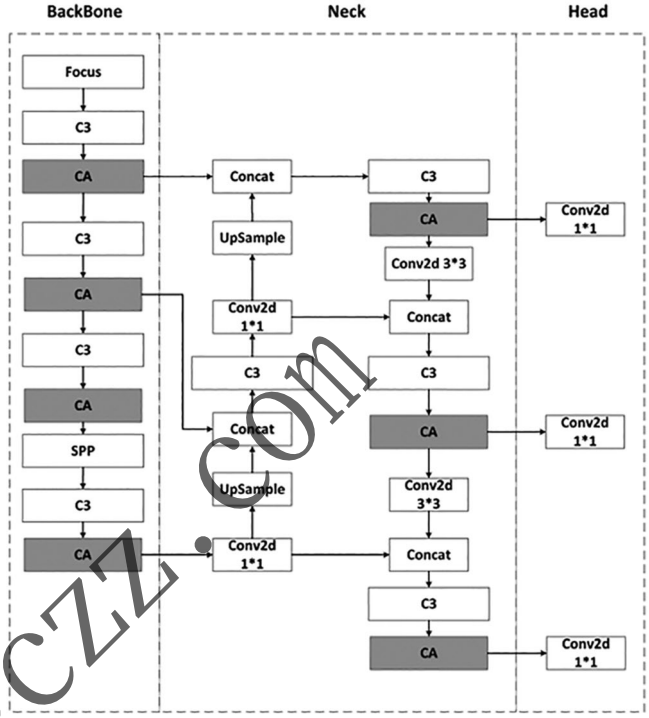

图3 CA−YOLOv5网络结构示意图

Fig.3 Diagram of CA−YOLOv5 network structure

2.2.2 引入CBAM网络的改进型YOLOv5网络

CBAM网络在通道注意力网络的基础上扩展了一个空间注意力模块，它可以在通道和空间维度上进行注意力运算。CBAM网络[9]包含两个子模块：通道注意力模块(Channel Attention Module，CAM)和空间注意力模块(Spatial Attention Module，SAM)。

基于CBAM注意力模块的典型网络示意图如图4所示。

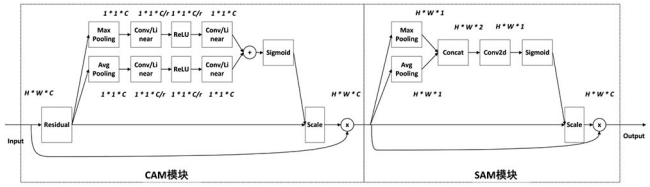

图4 基于CBAM注意力模块的典型网络示意图

Fig.4 Network structure diagram of CBAM attention module

CAM模块通过一个并行的最大值池化层，得到两个1×1×C的特征图，首先将得到的两个特征向量相加，然后经过一个激活函数得到每个通道的权重系数，最后用权重系数与原来的特征图通道相乘，即可得到缩放后的新特征图。

SAM模块先分别进行一个通道维度的平均池化和最大池化，得到两个$H\times W\times 1$的通道张量，将这两个张量在通道维度进行拼接，再经过一个7×7的卷积及Sigmoid激活函数，得

到权重系数，最后用权重系数和输入特征图相乘，即可得到缩放后的新特征图。

本文将CBAM网络模块应用在YOLOv5网络的骨干特征提取网络的每个多尺度特征输出的位置，用于对每个尺度下特征图的各通道/各空间进行特征重标定，以提升原始YOLOv5网络的特征提取能力。设计引入CBAM网络的改进型YOLOv5网络结构示意图如图5所示，命名为CBAM-YOLOv5网络。

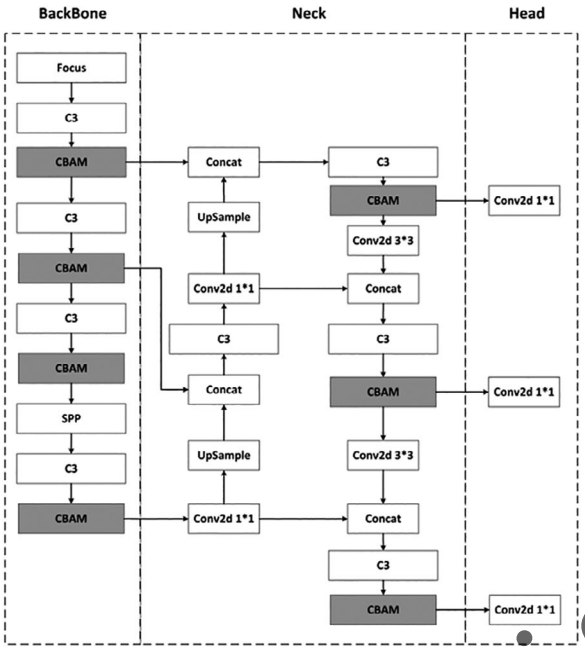

图5　CBAM-YOLOv5网络结构示意图

Fig.5 Network structure diagram of CBAM-YOLOv5

2.3 引入Transformer注意力机制网络的改进型YOLOv5网络设计

2.3.1 引入基础型Transformer Encoder Block模块的改进型YOLOv5网络

Transformer编码模块(Transformer Encoder Block)模块[10]采用原始Transformer网络中的编码(Encoder)结构，该Encoder网络结构示意图如图6所示，它可以捕获全局信息和丰富的上下文信息。

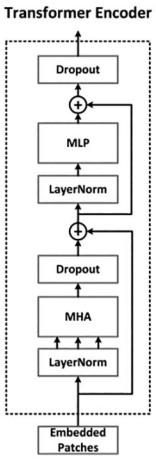

图6　Transformer Encoder网络结构示意图

Fig.6 Network structure diagram of Transformer Encoder

每个Transformer Encoder块包含两个子层。第一子层为MHA层，第二子层为MLP层。每个子层之间使用残差连接。

为了不大幅增加原始YOLOv5网络的模型参数量，本文将Transformer Encoder结构仅插入在骨干特征提取网络的末尾处，利用自注意力机制来捕获最后一层特征提取层的局部信息特征。引入了基础型Transformer Encoder Block模块的改进型YOLOv5网络如图7所示，命名为BasicTrans-YOLOv5网络。

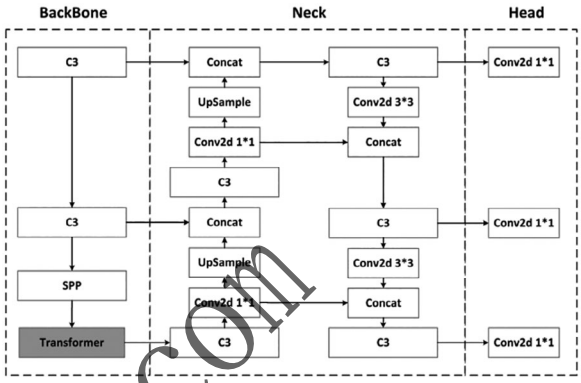

图7　BasicTrans-YOLOv5网络结构示意图

Fig.7 Network structure diagram of BasicTrans-YOLOv5

2.3.2 引入Swin Transformer Block模块的改进型YOLOv5网络

Swin Transformer Block模块是Swin Transformer网络[11]的核心部分。Swin Transformer网络是2021年提出的一种采用了局部自注意力增强机制的Transformer网络，该网络相较于卷积神经网络具有更强的动态计算能力，建模能力更强，可自适应计算局部与全局像素关系，非常具有推广使用的价值。Swin Transformer Block模块的网络结构如图8所示。

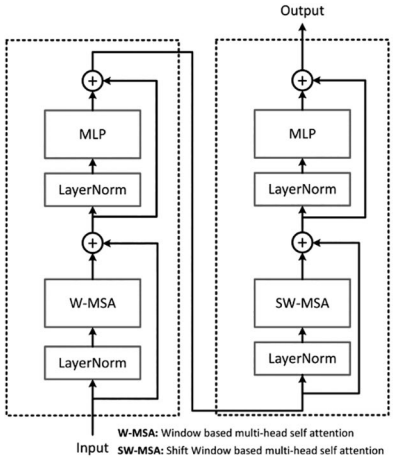

图8　Swin Transformer Block网络结构示意图

Fig.8 Network structure diagram of Swin Transformer Block

Transformer Block整体架构的核心模块是基于窗口的多头自注意力层(W-MSA)和基于偏移窗口的多头自注意力层(SW-MSA)。网络通过将注意力计算限制在一个窗口中，一方面能引入卷积操作的局部性，另一方面能节省计算量，具有很好的性能。

本文设计将Swin Transformer Block结构插入骨干特征提取网络和颈部特征融合中，利用高效的自注意力机制模

块充分挖掘特征表征的潜能。引入Swin Transformer Block模块的改进型YOLOv5网络如图9所示，命名为SwinTrans-YOLOv5网络。

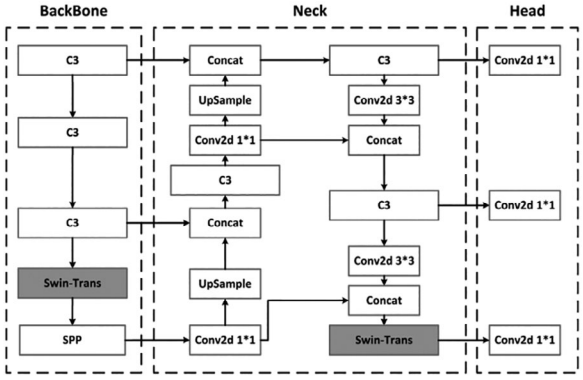

图9 SwinTrans-YOLOv5网络结构示意图

Fig.9 Network structure diagram of SwinTrans-YOLOv5

3 实验结果及分析(Experimental results and analysis)

3.1 实验环境

本文实验采用私有的变电站内复杂多尺度检测场景数据集，该数据集共有4 个分类，分别为屏体压板、火灾、抽烟、人员摔倒。统计样本库中各分类的样本数量，梳理样本库中的正负样本、难易样本。针对少分类、难识别的图像样本，采用复制图像后的随机数据增强方法扩充样本得到分类相对均衡的训练样本库，增强处理后的数据集共包括10,497 张图像。

使用开源工具对训练样本库中待检测目标的矩形关键区域进行人工标注，标注后使用脚本转换形成YOLOv5算法支持的标注文件。训练样本库按4:1的比例随机建立训练样本集和测试样本集。

3.2 结果对比与分析

本文设计的引入了注意力机制的改进型YOLOv5网络包括两大类、四种网络结构，与基准的YOLOv5网络在同一实验平台上进行训练，模型评估结果对比如表1所示，其中*mAP*@0.5表示交并比为0.5时的平均精度均值(*mAP*)指标，推理耗时表示模型处理一张图片所需要的时间，运算量表示网络模型的浮点运算量。

表1 改进型YOLOv5网络性能指标对比

Tab.1 Performance comparison of improved YOLOv5 networks

网络名称	平均准确率*AP*/%				*mAP*@0.5/%	推理耗时/ms	运算量/GFLOPS
	屏体压板	火灾	抽烟	人员摔倒			
YOLOv5	96.8	61.0	69.1	62.1	72.2	4.1	114.1
CA-YOLOv5	96.7	**65.9**	**72.0**	**69.5**	**76.3**	5.5	115.8
CBAM-YOLOv5	**97.8**	62.9	69.7	66.1	74.1	5.6	115.9
BasicTrans-YOLOv5	**97.2**	61.7	69.3	68.4	74.1	4.2	111.2
SwinTrans-YOLOv5	97.0	**67.8**	**79.6**	**80.4**	**81.2**	9.5	1,061.6

表1中加粗字体为当前网络在所属的注意力模型大类下准确率的最优值，可以看出，在当前实验环境参数及复杂多尺度场景数据集下有如下实验结论。

(1)引入了注意力机制的目标检测网络相比基准的YOLOv5网络，其*mAP*指标值一般都有一定的提升，说明注意力机制网络的确能有效提升模型的表达能力，但代价是增加了模型的复杂度。

(2)在引入的基于卷积神经网络的注意力机制模块中，对简单分类(屏体压板)提升效果有限，相较于基准网络，*AP*最大可提升1.0%，而对复杂分类(火灾、抽烟、人员摔倒)样本中的*AP*指标值的提升更明显。横向对比来看，引入CA注意力机制模块的网络综合表现最佳：它相比基准YOLOv5网络模型虽然复杂度略有提升，单张图像的推理耗时也增加了1.4 ms，但*mAP*指标值提升了4.1%，并且在复杂分类样本中的*AP*值提升效果显著(火灾、抽烟、人员摔倒的平均准确率分别提升4.9%、2.9%、7.4%)。因此从整体来看，CA模块最具引入价值。

(3)在引入的基于Transformer结构的自注意力模块网络中，SwinTrans-YOLOv5的*mAP*指标值显著提升，整体*mAP*指标相比基准YOLOv5网络提升了9.0%，尤其是在复杂分类下的*AP*值的提升最显著(火灾、抽烟、人员摔倒的平均准确率分别提升6.8%、10.5%、18.3%)，但代价是推理耗时及模型复杂度都增加较大，比较适合于算力充足、对实时性要求不高的场合。

(4)总体来说，在模型部署硬件算力足够的情况下，基于Transformer结构的SwinTrans-YOLOv5相较于基于卷积神经网络结构的CA-YOLOv5，拥有更强大的建模能力和更高的检测精度，在网络设计中更具引入价值；但是，在模型部署硬件算力一般、需要考虑推理的实时性的情况下，CA-YOLOv5也是一种非常不错的设计思路。

当然，以上实验结论并不是绝对的，它是建立在当前实验环境下的数据结果，不同的数据集，效果可能不同，但引入注意力机制模块确实能为提升原有基准网络的检测性能提供一种可行的设计思路。注意力机制模块能弥补卷积网络局部性过强、全局性不足的问题，帮助获取全局的上下文信息，具有让模型看得更广的能力，尤其在一些复杂多尺度场景下，对于难样本分类的检测准确率一般会取得较好的效果。

4 结论(Conclusion)

本文简要阐述了基于卷积神经网络和基于Transformer网络的注意力机制模块的网络结构，提出了多种引入了注意力机制的改进型YOLOv5网络的构建方法，在复杂多尺度目标检测数据集下进行性能对比。结果表明，引入注意力机制模块的改进型网络相较于基准网络，或多或少地获得了准确率方面的增益效果。其中，引入CA注意力模块和引入Swin Transformer自注意力模块分别在基于卷积神经网络和基于Transformer网络类别下取得了最佳的性能提升，并且基于Swin Transformer的自注意力网络在复杂场景下的建模能力优于传统的基于传统卷积神经网络的注意力网络。本文的研究成果为复杂场景下多尺度难样本目标检测网络的建模设计提供了一种改进思路。

(下转第54页)