基于改进 YOLOv5 的复杂跨域场景下的猪个体识别与计数

宁远霖[1]，杨 颖[1※]，李振波[1,2,3]，吴 潇[1]，张 倩[1]

（1. 中国农业大学信息与电气工程学院，北京 100083；2. 农业农村部农业信息获取技术重点实验室，北京 100083；
3. 农业农村部国家数字渔业中心，北京 100083）

摘 要：为解决复杂跨域场景下猪个体的目标检测与计数准确率低下的问题，该研究提出了面向复杂跨域场景的基于改进 YOLOv5（You Only Look Once version 5）的猪个体检测与计数模型。在骨干网络中分别集成了 CBAM（Convolutional Block Attention Module）即融合通道和空间注意力的模块和 Transformer 自注意力模块，并将 CIoU（Complete Intersection over Union）Loss 替换为 EIoU（Efficient Intersection over Union）Loss，以及引入了 SAM （Sharpness-Aware Minimization）优化器并引入了多尺度训练、伪标签半监督学习和测试集增强的训练策略。试验结果表明，这些改进使模型能够更好地关注图像中的重要区域，突破传统卷积只能提取卷积核内相邻信息的能力，增强了模型的特征提取能力，并提升了模型的定位准确性以及模型对不同目标大小和不同猪舍环境的适应性，因此提升了模型在跨域场景下的表现。经过改进后的模型的 mAP@0.5 值从 87.67%提升到 98.76%，mAP@0.5:0.95 值从 58.35%提升到 68.70%，均方误差从 13.26 降低到 1.44。该研究的改进方法可以大幅度改善现有模型在复杂跨域场景下的目标检测效果，提高了目标检测和计数的准确率，从而为大规模生猪养殖业生产效率的提高和生产成本的降低提供技术支持。

关键词：模型；计算机视觉；目标检测；计数；注意力机制；半监督学习

doi：10.11975/j.issn.1002-6819.2022.17.018

中图分类号：S126 **文献标志码**：A **文章编号**：1002-6819(2022)-17-0168-08

宁远霖，杨颖，李振波，等. 基于改进 YOLOv5 的复杂跨域场景下的猪个体识别与计数[J]. 农业工程学报，2022，38(17)：168-175. doi：10.11975/j.issn.1002-6819.2022.17.018 http://www.tcsae.org

Ning Yuanlin, Yang Ying, Li Zhenbo, et al. Detecting and counting pig number using improved YOLOv5 in complex scenes[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(17): 168-175. (in Chinese with English abstract) doi：10.11975/j.issn.1002-6819.2022.17.018 http://www.tcsae.org

0 引 言

目前中国的猪存栏量和出栏量均居世界第一[1]，但中国国内的生猪生产水平仍难以满足需求，部分猪肉仍需靠进口来弥补空缺[2]。因此，在养殖业正逐渐从传统的散户养殖的阶段过渡到集约化养殖、大规模自动化养殖阶段的时代背景下，如何提高大规模养殖场景下产业的生产效率，降低养殖成本，成为中国猪肉自给自足进程中不可缺少的一环。

猪只计数也称猪只盘点，是大规模养殖中工作量较大，容易出错的工作。由于猪只的淘汰、出售和死亡，导致在养殖过程中猪舍中的猪只数量可能不断地发生变化[3]，因此在养殖管理过程中需要对猪只数量进行统计[4]。传统的猪只计数主要采用人工清点的方式进行，在大规模复杂场景下，这不仅计数不准，并且消耗人力物力，效率较低。并且由于人与动物接触过多，也可能会有染病的风险[5]。基于目标检测的计数通过计算机视觉技术对目标进行自动检测和定位，在获取位置的同时对数量也进行统计，极大地节省了人力物力的投入。目前，国内外学者在动物的自动计数领域已经开展了一些研究，如 Ahn 等[6]基于 YOLOv4-Tiny（You Only Look Once-Tiny）算法进行改进，能够在嵌入式版上实现实时的猪只检测和计数；黎袁富等[7]利用 YOLOX 研究了对鱼苗的检测和计数方法，在水箱中鱼苗数量较大时仍然能够获得不错的效果。

动物行为分析同样是大规模养殖中工作量较大，且容易受主观经验影响的工作。动物的健康状况往往能够通过其行为体现[8]，传统的行为分析方法通常靠人工观察，容易受到个人经验和自然环境等因素的影响，近年来，基于目标检测的动物行为分析方法得到了广泛的研究，如 Liu 等[9]使用基于 SSD（Single Shot MultiBox Detector）的目标检测算法来识别和定位生猪的咬尾行为，并在群养场景下可达到 89.23%的准确率；Zhang 等[10]构建了基于 SSD 和 MobileNet 的 SBDA-DL 模型，用于识别母猪的饮水、排尿和哺乳行为；薛月菊等[11]基于 Faster RCNN（Faster Region-CNN）实现对哺乳母猪的站立、坐立、俯卧、腹卧和侧卧 5 类姿态的识别；董力中等[12]基于 YOLOv5（You Only Look Once version 5）实现对猪只站、走、卧三种行为的识别。

综上，猪只的自动计数和检测研究已经取得了一些成果，但还存在以下 3 点问题：1）数据集单一，即训练集和测试集往往来自同一拍摄场景，甚至来自于同一段

收稿日期：2022-08-08 修订日期：2022-08-28

基金项目：科技创新 2030-“新一代人工智能”重大项目课题-典型畜禽疫病诊断与主动防控智慧云平台（2021ZD0113805）

作者简介：宁远霖，研究方向为计算机视觉。Email：c2605759123@163.com

※通信作者：杨颖，博士，副教授，研究方向为计算机视觉、模式识别和语音识别等。Email：hbxtyy@126.com

视频片段，训练集和测试集的图像帧高度相似。然而，由于多种外部因素例如拍摄视角和背景噪声的影响，不同数据集具有不同的性质，这就导致跨域问题，即在一个域上训练的模型可能在另外一个域上性能较差[13]。2）检测目标通常较大，清晰度高，即场景中的检测目标通常在整幅图像中占比较大，且在光线良好的环境中拍摄，但真实养殖场景中往往光线复杂，猪只个体较小且图像模糊，这就导致方法对于小目标检测效果较差，无法适应真实猪舍场景中目标小且密集、光线较暗的复杂情况。3）研究方法无法对比，现有研究通常是基于私有数据集进行的测试，从而导致现有方法所提到的性能和指标难以复现。

针对以上问题，为了贴合真实的多场景下检测和计数任务的需求，本文重点研究复杂跨域场景下的目标检测和计数，研究拟选用来自于真实养殖环境中来自不同场景和拍摄视角的公开数据集，其中待检测目标大小差异明显，数据集明显跨域。为此，本文拟从网络层面改进以增强网络对模型的特征提取能力、选择合适的损失函数以增强模型的定位准确率、选择合适的优化器以使模型拥有更好的泛化能力、选择合适的训练策略以增强模型对不同大小目标和不同场景的适应性，从而提高模型在复杂跨域场景下的目标检测和计数准确率，辅助智能化养殖，提高生产效率，降低生产成本。

1　试验数据与分析

1.1　数据来源与预处理

本文数据来源选自于 2021 年 6－10 月在讯飞开放平台举办的“猪只盘点挑战赛”[14]，此次比赛提供 700 张图像用于训练，其中 500 张为 box 标注，200 张为 mask 标注。对于测试集，初赛和复赛各提供了 220 张图像。由于初赛数据集基本不存在跨域问题，故本文的训练集设置为训练集+初赛测试集，测试集设置为复赛测试集。

为了充分利用比赛数据，首先需要将 200 张 mask 标注信息转换为 box 标注信息，具体方法是分别计算所有 mask 标注关键点在 x 轴方向和 y 轴方向坐标的最小值和最大值，然后生成 box 标注。

1.2　数据分析

训练集的 700 张图像的分辨率为 1 920×1 080 像素和 1 536×2 048 像素两种，拍摄于两个场景中的不同时间段，目标较大，光线良好；初赛测试集的 220 张图像的分辨率为 1 920×1 080，拍摄场景虽与训练集不同，但环境相似且目标大小相似，基本不存在跨域的问题；复赛测试集的 220 张图像的分辨率为 1 280×960，分辨率较小，拍摄场景与训练集和初赛测试集的拍摄场景均不同，目标较小，大量目标存在大面积重叠，肉眼对部分猪个体有较大的识别难度。

如表 1 所示，本文分别计算了目标的高度比、宽度比以及高宽比，其分别代表了目标的高宽像素值与整幅图像高宽像素值的比值以及目标自身高宽像素值的比值。

从表 1 可以看出，复赛测试集的宽度比和高度比相对训练集和初赛测试集较小，这表明复赛测试集中的目标相对较小，仅有训练集和初赛测试集目标大小的三分之一左右；同时复赛测试集的高宽比相对训练集和初赛测试集较大，这表明复赛测试集的目标更加细长。

表 1　不同数据集中标注框的高宽特点

Table 1　Height and width characteristics of label boxes in different datasets

数据集 Datasets	最小高度比 Min height radio	最小宽度比 Min width radio	平均高度比 Mean of height radio	平均宽度比 Mean of width radio	平均高宽比 Mean of aspect radio
训练集 Training dataset	0.032 4	0.015 6	0.140 8	0.094 8	1.285 7
初赛测试集 Preliminary test dataset	0.021 9	0.007 2	0.112 0	0.076 6	1.304 8
复赛测试集 Semi-final test dataset	0.012 4	0.005 2	0.043 8	0.026 8	1.444 7

综上，训练集、初赛测试集与复赛测试集之间存在严重的跨域的问题，不仅体现在背景环境中，也体现在目标大小和目标自身的高宽比中。

2　基于改进 YOLOv5 的猪个体检测与计数研究

2.1　数据增强

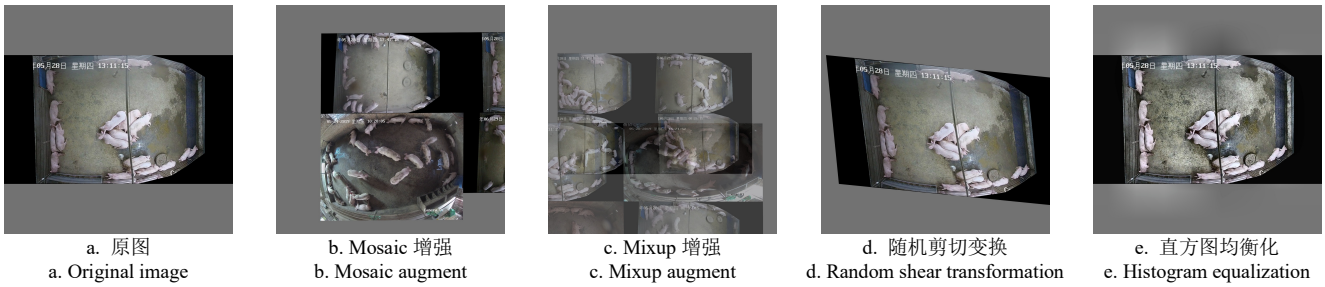
神经网络通常需要大量数据进行训练，而在真实场景中往往由于采集成本等问题无法获得充足的训练样本，并且由于本文数据集中存在跨域问题，因此利用数据增强方法使得模型具有更好的泛化性显得尤为重要，本文采用以下几种数据增强方法实现数据集的增强。

Mosaic 增强[15]指利用四张图像拼接到一张图像中，从而达到丰富背景和变相增加批量大小（Batch Size）的作用；Mixup 增强[16]指将不同图像堆叠到一张图像中，也能达到丰富背景和变相增强 Batch Size 的作用，但与 Mosaic 通过拼接合并图像不同，Mixup 增强会通过改变不同图像的透明度的方式，将不同图像堆叠到一张图像中，而不是简单地拼接不同图像。除了 Mosaic 增强、Mixup 增强外，本文采用的数据增强方法还包括 HSV（Hue, Saturation, Value）颜色变换、图像随机旋转、图像随机平移、图像随机缩放、图像随机剪切变换、图像上下翻转、图像左右翻转、均值滤波图像模糊、中值滤波图像模糊、转灰度图以及自适应直方图均衡化。本文采用上述数据增强方法增强后的部分效果示意图如图 1 所示。

2.2　YOLOv5 网络改进

2.2.1　YOLOv5 网络介绍

YOLOv5[17]是目前较为先进的实时单阶段目标检测网络，目前具有 YOLOv5n、YOLOv5s、YOLOv5m、YOLOv5l、YOLOv5x 、 YOLOv5n6 、 YOLOv5s6 、 YOLOv5m6 、 YOLOv5l6、YOLOv5x6 共十个子网络，其中 n 网络最小、x 网络最大，以 6 为结尾的模型拥有 4 个检测 Head，其他模型拥有 3 个检测 Head。本文对未改进的十个子网络进行了测试，其中 YOLOv5l6 网络的 mAP@0.5 值相对较高，因此本文选择以 YOLOv5l6 作为基础网络进行改进。本文在骨干网络（Backbone）上分别集成了 CBAM 即融合通道与空间注意力的模块以及 Transformer 自注意力模块，改进前后的 YOLOv5l6 网络对比如图 2 所示。



a. 原图
a. Original image

b. Mosaic 增强
b. Mosaic augment

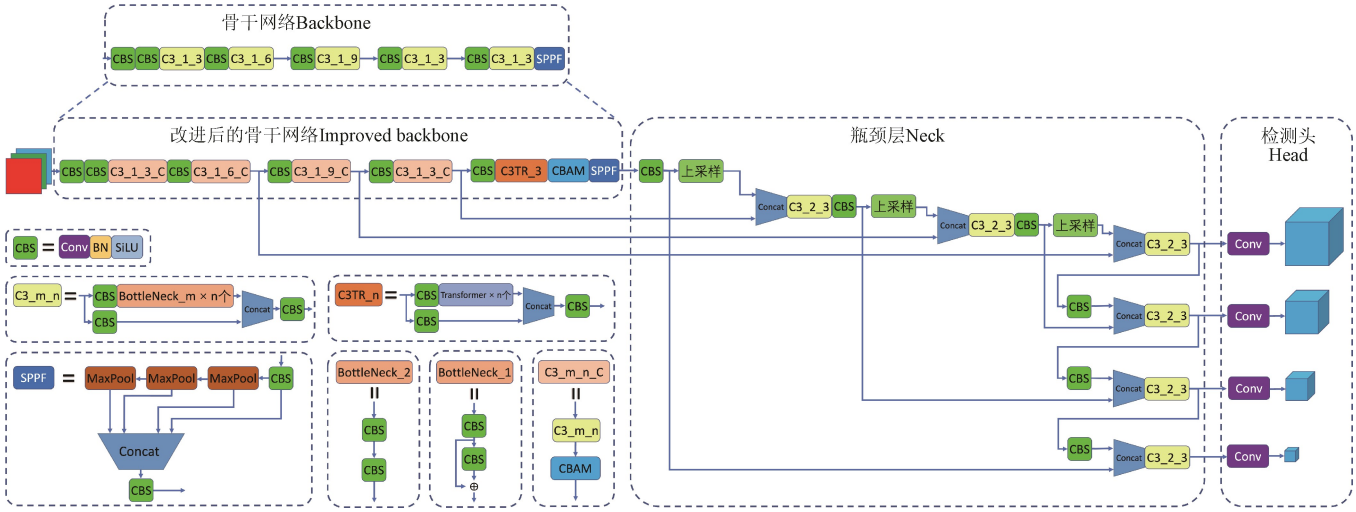
c. Mixup 增强
c. Mixup augment

d. 随机剪切变换
d. Random shear transformation

e. 直方图均衡化
e. Histogram equalization

图 1 数据增强效果图

Fig.1 Data augmentation



注：Conv 代表卷积；BN 代表批归一化；SiLU 代表 Sigmoid 加权线性单元；Concat 代表拼接；MaxPool 代表最大池化；Transformer 代表自注意力模块；CBAM 代表通道和空间注意力模块；SPPF 代表快速空间金字塔池化模块。

Note: Conv donates Convolutional; BN donates Batch Normalization; SiLU donates Sigmoid-weighted Linear Units; Concat donates Concatenation; MaxPool donates Max Pooling; Transformer donates Self-attention Module; CBAM donates Convolutional Block Attention Module; SPPF donates Spatial Pyramid Pooling - Fast Module.

图 2 改进前后的 YOLOv5l6 网络对比图

Fig.2 Comparison diagram of YOLOv5l6 network before and after improvement

2.2.2 集成 CBAM 注意力模块

由于猪舍光线和环境较为复杂，使得网络难以关注到关键的信息，给模型带来了性能影响。为解决上述问题，使网络更好地关注于猪个体，本文提出了一种集成 CBAM（Convolutional Block Attention Module）[18]即融合通道注意力与空间注意力的 YOLOv5 网络。

考虑到特征提取主要在骨干网络中进行，因此本文选择将 CBAM 模块添加到原有的骨干网络中的 C3_m_n 模块后，形成新的 C3_m_n_C 模块，如图 2 所示。

2.2.3 集成 Transformer 自注意力模块

近年来，Vision Transformer（ViT）[19]和 Obeject Detection with Transformers（DETR）[20]等基于 Transformer 的工作得到了极大关注。Transformer 架构被认为可以更好地获取来自全局感受野的信息，而不像 CNN 一样局限于卷积核内的相邻信息，因此在图像分类、目标检测、图像分割等领域显著改进了基线。在猪舍场景下，由于光线噪声问题，部分猪个体的轮廓边界不明显，完全基于 CNN 的骨干网络难以确定轮廓边界。受到 BoTNet[21]工作的启发，本文选择通过替换 C3_n_m 模块中的 BottleNeck 为 Transformer Encoder 的方式将 Transformer 集成到骨干网络中，从而能够借助 Transformer 架构的优势，提取全局感受野的信息，突破卷积核只能捕获相邻特征的限制，减小背景噪声对模型性能的影响。需要强调的是，本文使用的 Transformer Encoder 架构与 Vision Transformer 中的略有不同，由于 LN 层会影响计数精度，本文去掉了 Layernorm（LN）层。

考虑到骨干网络较上层的特征图拥有更大的分辨率，若将其替换为 Transformer 架构则需要非常大的计算资源消耗。因此，本文选择将 Transformer 架构集成到骨干网络的最后一个 C3_n_m 模块中，命名为 C3TR_n。此处的特征图经过了 64 倍的下采样，对计算资源的消耗相对较小，如图 2 所示。

2.3 损失函数选择

YOLOv5 中使用的定位损失函数为 CIoU（Complete over Union）Loss[22-23]，其在 DIoU（Distance IoU） Loss[23]的基础上发展而来。虽然 CIoU Loss 从形式上考虑了重叠损失、中心点距离损失和 box 框的宽高损失，但其宽高损失设计不合理，例如当标注框与预测框的宽高成比例时，宽高损失为 0，即退化为 DIoU Loss。为解决上述问题，有学者提出了名为 EIoU（Efficient IoU）Loss[24]的方

法。EIoU Loss 同时兼顾了重叠损失、中心点距离损失以及宽高损失，其中重叠损失、中心点距离损失延续了 CIoU Loss 中使用的方法，但宽高损失相比 CIoU Loss 直接最小化了预测框和标注框的宽高差，因此收敛速度更快，效果更好。因此，本文选择将 YOLOv5 中使用的 CIoU Loss 替换为 EIoU Loss。

2.4　优化器选择

目前大多数神经网络都是过参数化的，虽然理论上可以选用合适的训练算法，使得训练出来的网络具有良好的泛化能力[25]。但事实上，网络通常会选择“记住”训练集，即使训练集的标签受到了污染，甚至完全错误，网络也能使训练误差降低为 0[26]。

目前大多数训练损失都是非凸的，即模型存在多个局部和全局极小值，在不同的极小值处，模型的泛化能力也会不同[25]。现有常用的 SGD（Stochastic Gradient Descent）、Adam（Adaptive Moment Estimation）[27]、AdamW（Adam with Decoupled Weight Decay）[28]等优化器虽然可以让网络找到全局最小值，但这些最小值的附近可能是非常陡峭的，即轻微的参数变化就会大幅影响损失，会降低模型的泛化能力。SAM（Sharpness-Aware Minimization）[25]优化器通过同时最小化损失值以及损失值附近的锐度，使网络的参数能够优化到损失相对平坦，即附近的参数也有更低的损失的地方，提高模型在跨域场景下的表现。

2.5　训练策略改进

在模型的训练过程中，本文使用了多尺度训练、伪标签半监督学习和测试集增强的策略，提高了模型对不同大小目标的适应性以及在跨域场景下的表现。具体的训练策略如下：1）使用数据增强后的训练集训练模型，在此期间，使用多尺度训练策略随机调整模型输入图像的分辨率。2）使用测试集增强与伪标签结合的方法给测试集打标签。3）将训练集和测试集组合成新的训练集。4）再次重复（1）～（3）两次。

1）多尺度训练

由于复赛测试集中的目标相对训练集较小，这在一定程度上会影响模型的效果。为了解决上述问题，本文在训练时使用多尺度训练的策略，提升模型对不同尺度目标的适应性。具体做法是使模型输入图像的分辨率在[0.8p,p]中随机选取，其中 p 为模型原始输入分辨率。

2）伪标签与测试集增强

伪标签[29]是一种半监督学习方法，旨在借助无标签的数据来提升有监督过程中的模型性能。由于猪舍场景复杂，目标大小不一，使得从已知场景中获取的训练集上训练出的模型难以在未知场景下也拥有较好的性能。通过使用伪标签方法，可以让模型学习到测试集的数据分布，提升模型在测试集中的表现。

在生成测试集标签时，本文使用了测试集增强的方法。具体做法是将输入图像的原图、左右翻转图、上下翻转图依次送入模型，得到预测结果，然后将这三份预测结果合并，再进行非极大值抑制（Non Maximum Suppression，NMS），得到最终的模型预测结果。虽然使用测试集增强的方法往往能够获得更高的模型准确率，但其会大大增加推理时间。因此本文仅在生成测试集伪标签时使用该策略获得更加准确的测试集伪标签，在进行模型性能评估时并不使用测试集增强。

在循环迭代训练过程中，本文对新的训练集中来自测试集的部分也同样应用数据增强。考虑到模型在伪标签训练阶段的主要学习对象是测试集的数据分布，而数据增强的做法通常是打乱数据分布，因此区别与训练集，本文对测试集使用 Mosaic 数据增强的概率从 100%调整到 50%，以降低对数据分布影响较大的 Mosaic 数据增强的发生概率。其他超参数与默认超参数保持不变。

2.6　模型训练设置

本文试验不设置验证集。通常，模型使用训练集进行训练，并使用验证集观察训练结果，调整超参数，最后在测试集上进行测试。然而，由于本文试验所用数据集存在跨域问题，即使划分验证集，并手动根据验证集的结果调整超参数，也无法保证模型能够在测试集有很好的表现，另外，也无法保证与其他模型对比的公平性。因此，本文除伪标签训练外，全部使用模型自带的默认超参数进行训练。

本文所有试验所用的图像分辨率默认均调整为 1 920×1920，Batch Size 为 8。所有试验均使用在 COCO 数据集上训练的预训练模型进行迁移学习，所有关于检测速度的试验均包括预处理时间、推理时间和后处理时间且均在单张显卡上进行。

2.7　评价指标

在目标检测方面，全类平均精度（mean Average Precision, mAP）常用于评价目标检测算法的识别效果，其由准确率（Precision）和召回率（Recall）共同决定，因此，本文选择 mAP@0.5 和 mAP@0.5:0.95 作为评价指标。其中 mAP@0.5 代表当检测框与标注框的 IoU 阈值大于 0.5 时视为预测正确的 mAP；mAP@0.5:0.95 指标较为苛刻，其代表选择不同 IoU 阈值（0.5, 0.55, … , 0.9, 0.95）的 mAP 的平均值。在计数方面，本文选择均方误差（Mean Square Errors, MSE）作为评价指标。

3　结果与分析

本文所使用的试验环境如下，操作系统为 Ubuntu18.04，CPU 为 Intel(R) Xeon(R) Gold 6240 CPU @ 2.60GHz，GPU 为 NVIDIA Tesla V100 × 4，CUDA 版本为 10.2，Python 版本为 3.8.13，内存为 128GB，深度学习框架为 PyTorch1.11.0。

3.1　消融试验

为明确不同改进点对模型整体性能的影响，验证各改进点的有效性，本文设计了消融试验，分别从网络、损失函数、优化器、训练策略多方面测试了改进方法的有效性，如表 2 所示。从表中可以看出，各个改进点对模型的检测性能和计数性能均有提升。在原骨干网络上集成 CBAM 模块和 Transformer 模块，模型的 mAP@0.5 分别提升了 2.48 个百分点和 3.46 个百分点，均方误差分

别降低了 2.62 和 6.58；将模型损失函数改为 EIoU，模型的 mAP@0.5 提升了 0.93 个百分点，均方误差降低了 2.73。改进后模型的 mAP@0.5 值从 87.67%提升到 98.76%，mAP@0.5:0.95 值从 58.35%提升到 68.70%，均方误差从 13.26 降低到 1.44，这表明改进后的 YOLOv5 模型的检测性能和计数性能均有较大的提升，且在模型参数量略有增加的前提下依然能接近原模型的计算速度。综上，本文改进后的模型的在整体性能超越原有模型。

表 2 消融试验

Table 2 Ablation experiment

改进点 Improvement points						mAP @0.5/%	mAP @0.5:0.95/%	均方误差 MSE	参数量 Parameters/M	浮点计算量 Floating point operations/B	帧率 Frames per second/(帧·s^{-1})
CBAM	Transformer	EIoU Loss	SAM	多尺度训练 Multi-scale	伪标签+测试集增强 Pseudo+TTA						
						87.67	58.35	13.26	76.16	994.3	10.5
✓						90.15	60.45	10.64	76.92	997.5	9.2
	✓					91.13	60.49	6.68	75.64	987.7	10.1
		✓				88.60	59.84	10.53	76.16	994.3	10.5
			✓			93.70	64.15	4.55	76.16	994.3	10.5
				✓		87.98	59.03	12.73	76.16	994.3	10.5
					✓	97.30	70.54	2.47	76.16	994.3	10.5
✓	✓	✓	✓	✓	✓	98.76	68.70	1.44	76.39	990.9	9.2

注：mAP@0.5 代表当检测框与标注框的 IoU 阈值大于 0.5 时视为预测正确的 mAP；mAP@0.5:0.95 代表选择不同 IoU 阈值（0.5, 0.55, … , 0.9, 0.95）的 mAP 的平均值；mAP 表示全类平均精度。

Note: mAP@0.5 means that the prediction is correct when the IoU threshold of the detection box and the annotation box is greater than 0.5; mAP@0.5:0.95 means the average value of mAPs with different IoU thresholds (0.5, 0.55, … , 0.9, 0.95); mAP meansmean average precision.

3.1.1 优化器选择对模型性能的影响

不同优化器的试验结果如表 3 所示，选择 SAM 优化器后，相较于 YOLOv5 中使用的 SGD 优化器，模型的 mAP@0.5 提升了 6.03 个百分点，mAP@0.5:0.95 提升了 5.8 个百分点，均方误差降低了 8.71，对模型的检测性能和计数性能均有所提升。同时，对比 Adam 优化器与 AdamW 优化器，SAM 优化器在跨域场景中仍有较大优势。本文通过将 SGD 优化器替换为 SAM 优化器，使得模型的参数优化过程具有了锐度感知的能力，提高了模型的泛化能力。

表 3 不同优化器的试验结果

Table 3 The result of different optimizers

优化器 Optimizer	mAP@0.5/%	mAP@0.5:0.95/%	均方误差 MSE
SGD（原始） SGD (Original)	87.67	58.35	13.26
Adam[27]	81.60	46.72	23.06
AdamW[28]	90.30	55.45	9.05
SAM	93.70	64.15	4.55

3.1.2 Transformer 选择对模型性能的影响

本文对在骨干网络中是否集成 Transformer 模块以及是否集成带有 Layernorm（LN）层的 Transformer 模块进行了对比试验，具体试验结果如表 4 所示。

如表 4 所示，集成带有 LN 层的 Transformer 模块后，模型的 mAP@0.5 提升了 3.56 个百分点，mAP@0.5:0.95 提升了 3.43 个百分点，均方误差降低了 5.433；集成不带有 LN 层的 Transformer 模块后，模型的 mAP@0.5 提升了 3.46 个百分点，mAP@0.5:0.95 提升了 2.14 个百分点，均方误差降低了 6.58。两种方案对模型的检测性能和计数性能均有所提升，并且降低了模型的参数量和浮点计算量，但略微影响了检测速度。两者相比，带有 LN 层的 Transformer 的检测精度（mAP@0.5 和 mAP@0.5:0.95）略高，但计数精度（MSE）较差。综合考虑，本文选择不带有 LN 层的 Transformer 模块。本文通过集成 Transformer 模块，使模型具有提取全局信息的能力，降低了复杂背景噪声和猪个体轮廓边界不清楚对模型性能的影响。

表 4 不同骨干网络的试验结果

Table 4 The result of different backbone networks

模型 Model	mAP @0.5 /%	mAP @0.5:0.95 /%	均方误差 MSE	参数量 Parameters /M	浮点计算量 Floating point operations/B	帧率 Frames per second/(帧·s^{-1})
原始 Original	87.67	58.35	13.26	76.16	994.3	10.5
原始+Transformer+LN Original+Transformer+LN	91.23	61.78	7.83	75.64	987.7	10.05
原始+Transformer Original+Transformer	91.13	60.49	6.68	75.64	987.7	10.05

3.2 与其他模型的对比试验

本文选择了经典的 Faster RCNN[30]模型、针对密集对象的 VarifocalNet[31]模型以及属于 anchor-free 的 YOLOX(L)[32]模型与本文改进模型进行对比，对比模型均选自 mmdetection 框架[33]的代码实现。本文与其他模型的对比试验详见表 5。

表 5 与其他模型的试验结果对比

Table 5 The comparison of experimental results with other models

模型 Model	mAP @0.5 /%	mAP @0.5:0.95 /%	均方误差 MSE	参数量 Parameters /M	浮点计算量 Floating point operations /B	帧率 Frames per second/(帧·s^{-1})
Faster RCNN	48.96	22.88	144.15	41.12	707.81	8.8
VarifocalNet	57.57	29.75	21.35	32.48	680.01	9.6
YOLOX(L)	77.65	37.95	10.92	54.15	698.92	8.6
改进后的 YOLOv5l6 Improved YOLOv5l6	98.76	68.70	1.44	76.39	990.90	9.2

如表 5 所示，对比模型均不能很好地应对跨域问题，无论选择哪种评价指标，本文改进后模型的检测性能和计数性能均有较大优势，并且仍然能够保持相对较快的速度。

3.3　模型可视化分析与比较

相较于改进前的模型和其他模型，本文改进后的模型展现出了较强的特征提取能力和泛化能力，即使在跨域场景中也依然能够准确识别大部分的待测目标，其中部分识别结果的展示对比如图 3 所示。其中，场景一和场景二标签中的目标个数均为 36，Faster RCNN、VarifocalNet、YOLOX(L)、YOLOv5l6、改进后的 YOLOv5l6 在场景一/场景二检测到的目标个数分别为 19/27、31/37、32/39、35/35、36/36，与真实值标签中目标个数的偏差分别为 17/9、5/1、4/3、1/1、0/0。可以看出，本文提出的改进模型无论是与其他模型相比，还是与改进前的模型相比，检测出的猪只数量与实际数量偏差最小，其他模型都有不同程度的漏检和误检。改进后的模型将严重遮挡的猪个体和漏检的小目标猪个体都检测出来，证明了本文方法改进的有效性。

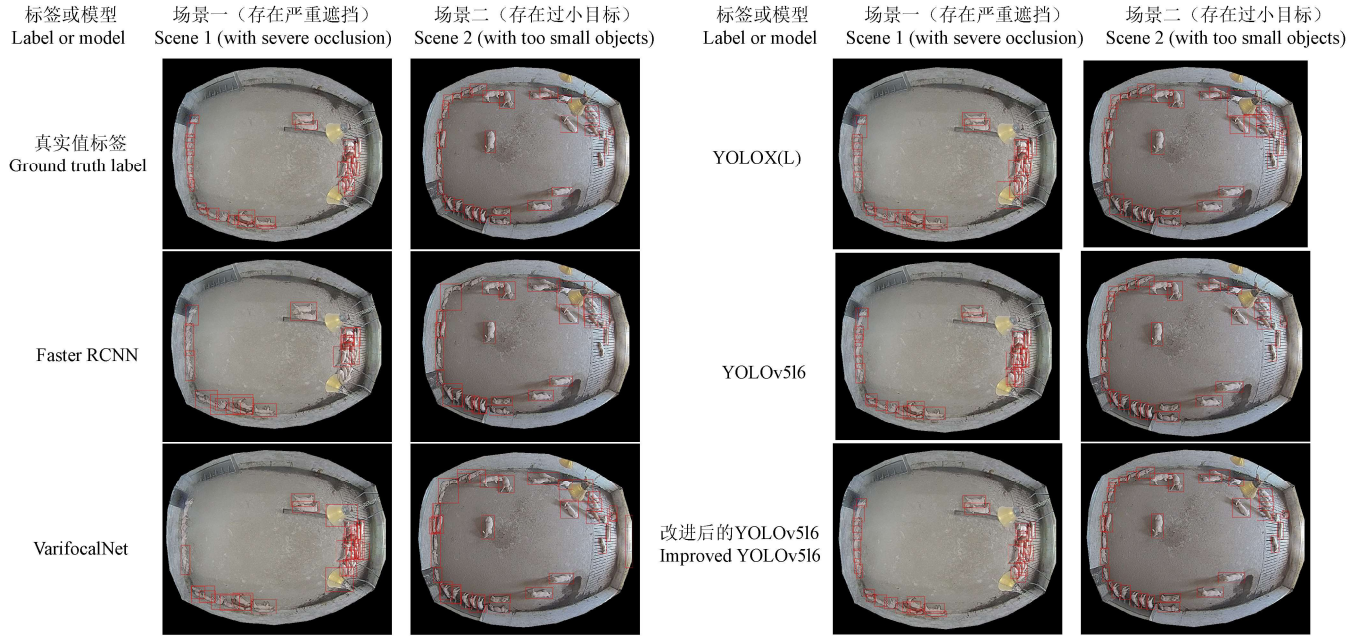


图 3　识别结果对比图

Fig.3　Comparison diagram of recognition results

4　结　论

真实养殖场景下环境复杂、条件多变，且猪只遮挡严重、个体尺度不一，给猪只自动检测和计数带来了巨大的困难和挑战。针对这种跨域场景下的复杂目标检测和计数问题，本文提出了基于改进 YOLOv5 的猪个体检测与计数模型，结论如下：

1）在原骨干网络上集成卷积块注意力模块（Convolutional Block Attention Module，CBAM）和 Transformer 模块，可分别提高模型对图像中重要区域特别是小目标的关注度和全局信息的提取能力，改进之后的模型与原模型相比，mAP@0.5（当检测框与标注框的 IoU 阈值大于 0.5 时视为预测正确的 mAP）分别提升了 2.48 个百分点和 3.46 个百分点，均方误差分别降低了 2.62 和 6.58。

2）将模型损失函数改进为 EIoU（Efficient Intersection over Union）Loss，有效解决了 CIoU（Complete Intersection over Union）Loss 中宽高损失设计不合理的问题，使得模型 mAP@0.5 提升了 0.93 个百分点，均方误差降低了 2.73。

3）通过采用 SAM（Sharpness-Aware Minimization）优化器，提高了模型的泛化能力，模型的 mAP@0.5 提升了 6.03 个百分点，均方误差降低了 8.71。

4）通过多尺度训练和伪标签与测试集增强相结合的策略的引入，可增强模型对不同尺度目标的适应性，进一步提高了模型在跨域数据集上的泛化能力。试验结果证明了本文改进后的模型展现了较强的特征提取能力和泛化能力，即使在跨域场景中依然能够准确识别大部分的待测目标。

本文的方法也存在不足，在跨域场景中仍有少部分特别密集和严重遮挡的场景，但由于非极大值抑制过程的缺陷，使得预测框之间相互抑制造成漏检，这部分工作将是本文未来的研究重点。

[参　考　文　献]

[1]　俞燃. 基于深度学习的哺乳期猪只目标检测与姿态识别[D]. 哈尔滨：东北农业大学，2021.
Yu Ran. Object Detection and Pose Recognition of Suckling Pigs Based on Deep Learning[D]. Harbin: Northeast Agricultural University, 2021. (in Chinese with English abstract)

[2]　熊本海，杨亮，郑姗姗，等. 哺乳母猪精准饲喂下料控制系统的设计与试验[J]. 农业工程学报，2017，33(20)：177-182.
Xiong Benhai, Yang Liang, Zheng Shanshan, et al. Design and test of precise blanking control system for lactating sows[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2017, 33(20):

177-182. (in Chinese with English abstract)

[3] 李静. 基于深度学习的群猪盘点算法研究[D]. 武汉：华中农业大学，2021.
Li Jing. Research on Pig Herd Counting Based on Deep Learning[D]. Wuhan: Huazhong Agricultural University, 2021. (in Chinese with English abstract)

[4] 高云，李静，余梅，等. 基于多尺度感知的高密度猪只计数网络研究[J]. 农业机械学报，2021，52(9)：172-178.
Gao Yun, Li Jing, Yu Mei, et al. High-density Pig Counting Net Based on Multi-scale Aware[J]. Transactions of the Chinese Society for Agricultural Engineering, 2021, 52(9): 172-178. (in Chinese with English abstract)

[5] Guo H, Ma X, Ma Q, et al. LSSA_CAU: An interactive 3d point clouds analysis software for body measurement of livestock with similar forms of cows or pigs[J]. Computers and Electronics in Agriculture, 2017, 138: 60-68.

[6] Ahn H, Son S, Kim H, et al. EnsemblePigDet: Ensemble deep learning for accurate pig detection[J]. Applied Sciences, 2021, 11(12): 5577.

[7] 黎袁富，杜家豪，莫家浩，等. 基于 YOLOX 的鱼苗检测与计数[J]. 电子元器件与信息技术，2022，6(5)：192-4.

[8] 李菊霞，李艳文，牛帆，等. 基于 YOLOv4 的猪只饮食行为检测方法[J]. 农业机械学报，2021，(3)：251-256.
Li Juxia, Li Yanwen, Niu Fan, et al. Pig diet behavior detection method based on YOLOv4[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, (3): 251-256. (in Chinese with English abstract)

[9] Liu D, Oczak M, Maschat K, et al. A computer vision-based method for spatial-temporal action recognition of tail-biting behaviour in group-housed pigs[J]. Biosystems Engineering, 2020, 195: 27-41.

[10] Zhang Y, Cai J, Xiao D, et al. Real-time sow behavior detection based on deep learning[J]. Computers and Electronics in Agriculture, 2019, 163: 104884.

[11] 薛月菊，朱勋沐，郑婵，等. 基于改进 Faster R-CNN 识别深度视频图像哺乳母猪姿态[J]. 农业工程学报，2018，34(9)：189-196.
Xue Yueju, Zhu Xunmu, Zheng Chan, et al. Lactating sow postures recognition from depth image of videos based on improved Faster R-CNN[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2018, 34(9): 189-196. (in Chinese with English abstract)

[12] 董力中，孟祥宝，潘明，等. 基于姿态与时序特征的猪只行为识别方法[J]. 农业工程学报，2022，38(5)：148-157.
Dong Lizhong, Meng Xiangbao, Pan Ming, et al. Recognizing pig behavior on posture and temporal features using computer vision[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(5): 148-157. (in Chinese with English abstract)

[13] 刘亚婧. 图像识别的跨域技术研究[D]. 合肥：中国科学技术大学，2022.
Liu Yajing. Cross-Domain Technology Development for Image Recognition[D]. Hefei: University of Science and Technology of China, 2022. (in Chinese with English abstract)

[14] 科大讯飞. 猪只盘点挑战赛[EB/OL]. 2021-10-24[2022-07-01]. http://challenge.xfyun.cn/topic/info?type= pig-check.

[15] Bochkovskiy A, Wang C Y, Liao H. YOLOv4: Optimal speed and accuracy of object detection[EB/OL]. 2020-04-23 [2022-07-01]. https://arxiv.org/abs/2004.10934.

[16] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization[EB/OL]. 2018-04-27 [2022-07-01]. https://arxiv.org/abs/1710.09412.

[17] Jocher G. Ultralytics/yolov5: v6.1 - tensorrt, tensorflow edge tpu and openvino export and inference[EB/OL]. 2022-02-22[2022-07-01]. https://github.com/ultralytics/yolov5.

[18] Woo S, Park J, Lee J-Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). Munich, Germany: Springer, 2018: 3-19.

[19] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. 2021-06-03[2022-07-01]. https://arxiv.org/abs/2010.11929.

[20] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//Proceedings of the European conference on computer vision (ECCV). Online: Springer, 2020: 213-29.

[21] Srinivas A, Lin T-Y, Parmar N, et al. Bottleneck transformers for visual recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Online: IEEE, 2021: 16519- 16529.

[22] Zheng Z, Wang P, Ren D, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation[EB/OL]. 2021-07-05[2022-07-01]. https://arxiv.org/abs/2005.03572.

[23] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]//Proceedings of the AAAI conference on artificial intelligence. New York, USA: AAAI Press, 2020: 12993-3000.

[24] Zhang Y-F, Ren W, Zhang Z, et al. Focal and efficient IOU loss for accurate bounding box regression[EB/OL]. 2021-01-20[2022-07-01]. https://arxiv.org/abs/2101.08158.

[25] Foret P, Kleiner A, Mobahi H, et al. Sharpness-aware minimization for efficiently improving generalization[EB/OL]. 2021-04-29[2022-07-01]. https://arxiv.org/abs/2010.01412.

[26] Zhang C, Bengio S, Hardt M, et al. Understanding deep learning requires rethinking generalization[EB/OL]. 2017-02-26[2022-07-01]. https://arxiv.org/abs/1611.03530.

[27] Kingma D P, Ba J. Adam: A method for stochastic optimization[EB/OL]. 2017-01-30[2022-07-01]. https://arxiv.org/abs/1412.6980.

[28] Loshchilov I, Hutter F. Decoupled weight decay regularization[EB/OL]. 2019-01-04[2022-07-01]. https://arxiv.org/abs/1711.05101.

[29] Lee D-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks[C]// Workshop on challenges in representation learning, ICML. Atlanta, USA: IMLS, 2013: 896.

[30] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[31] Zhang H, Wang Y, Dayoub F, et al. Varifocalnet: An iou-aware dense object detector[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online: IEEE, 2021: 8514-23.
[32] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[EB/OL]. 2021-08-06[2022-07-01]. https://arxiv.org/abs/2107.08430.
[33] Chen K, Wang J, Pang J, et al. MMDetection: Open mmlab detection toolbox and benchmark[EB/OL]. 2019-06-17 [2022-07-01]. https://arxiv.org/abs/1906.07155.

Detecting and counting pig number using improved YOLOv5 in complex scenes

Ning Yuanlin[1], Yang Ying[1※], Li Zhenbo[1,2,3], Wu Xiao[1], Zhang Qian[1]

(1. *College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*; 2. *Key Laboratory of Agricultural Information Acquisition Technology*, *Ministry of Agriculture and Rural Affairs, Beijing* 100083, *China*; 3. *National InnovationCenter for Digital Fishery*, *Ministry of Agriculture and Rural Affairs, Beijing* 100083, *China*)

Abstract: The number of pigs in the shed often varies continuously in large-scale breeding scenes, due to the elimination, sale, and death. It is necessary to count the number of pigs during breeding. At the same time, the health status of the pigs is closely related to their behavior. The abnormal behavior can be predicted in time from the normal behavior of pigs for better economic benefits. Object detection can be expected to detect and count at the same time. The detection can be the basis of behavioral analysis. However, the current detection and counting performance can be confined to the blur cross-domain at the different shooting angles and distances in the complex environment of various pig houses. In this study, a novel model was proposed for pig individual detection and counting using an improved YOLOv5(You Only Look Once Version 5) in the complex cross-domain scenes. The study integrated CBAM (Convolutional Block Attention Module), a module that combined both channel and spatial attention modules, in the backbone network, and integrated the Transformer, a self-attention module, in the backbone network, and replaced CIoU(Complete IoU) Loss by EIoU(Efficient IoU) Loss, and introduced the SAM (Sharpness-Aware Minimization) optimizer and training strategies for multi-scale training, pseudo-label semi-supervised learning, and test set augment. The experimental results showed that these improvements enabled the model to better focus on the important areas in the image, broke the barrier that traditional convolution can only extract adjacent information within the convolution kernel, enhanced the feature extraction ability, and improved the localization accuracy of the model and the adaptability of the model to different object sizes and different pig house environments, thus improving the performance of the model in cross-domain scenes. In order to verify the effectiveness of the above improved methods, this paper used datasets from real scenes. There was cross-domain between these datasets, not only in the background environment, but also in the object size and the aspect ratio of the object itself. Sufficient ablation experiments showed that the improved methods used in this paper were effective. Whether integrating CBAM, integrating Transformer, using EIoU Loss, using SAM optimizer, using multi-scale training, or using a combination of pseudo-label semi-supervised learning and test set augment, the mAP (mean Average Precision) @0.5 values, the mAP@0.5:0.95 values and the MSE (Mean Square Errors) of the model where improved to varying degrees. After integrating all improvement methods, the mAP@0.5 value of the improved model was increased from 87.67% to 98.76%, the mAP@0.5:0.95 value was increased from 58.35% to 68.70%, and the MSE was reduced from 13.26 to 1.44. Compared with the classic Faster RCNN model, the VarifocalNet model for dense object detection and the YOLOX model belong to anchor-free, the detection performance and counting performance of the improved model in this paper had greater advantages regardless of which evaluation metric was chosen, and was still able to maintain a relatively fast speed. The results showed that the improved model in this paper exhibited strong feature extraction and generalization ability, and could still accurately identify most of the objects to be tested even in cross-domain scenes. The above research results demonstrated that the improved method in this paper could significantly improve the object detection effect of the existing model in complex cross-domain scenes and increase the accuracy of object detection and counting, so as to provide technical support for improving the production efficiency of large-scale pig breeding and reducing production costs.

Keywords: models; computer vision; object detection; counting; attention mechanism; semi-supervised learning