Software Engineering and Applications 软件工程与应用, 2022, 11(6), 1404-1414
Published Online December 2022 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2022.116144

Hans汉斯

基于改进YOLOX的茶叶嫩芽目标检测研究

刘瑞欣，严春雨，李　飞，王等准

贵州大学大数据与信息工程学院，贵州　贵阳

收稿日期：2022年11月29日；录用日期：2022年12月20日；发布日期：2022年12月30日

摘　要

茶产业是我国进出口贸易商品的一大重要方面，茶在我国有着悠久的文化底蕴，与我国人民的生活密切相关。为了满足名优茶的茶叶嫩芽智能采摘需求，本文首先建立了自然环境下的茶叶嫩芽数据集，并提出了一种基于Swin-Transformer的改进YOLOX茶叶嫩芽检测模型——YOLOX-ST。该模型将Swin-Transformer作为原始YOLOX模型的骨干网络，提高了模型整体的检测精度，并引入了CBAM注意力机制，解决复杂环境背景下容易错检漏检的情况。实验结果表明，该模型的mAP值达到了79.12%，比原始模型提高了5.2%，精确度达到了90.45%，比原始模型提高了4.62%。与同系列的YOLOv3、YOLOv4、以及YOLOv5模型相比，YOLOX-ST的mAP以及准确率最高分别提升了7.09%和6.43%，拥有良好的检测精度与模型泛化能力。由此可见，该模型为茶叶嫩芽的智能化采摘奠定了一个良好的基础。

关键词

茶叶嫩芽，YOLOX，目标检测

Research on Tea Sprouts Object Detection Based on Improved YOLOX

Ruixin Liu, Chunyu Yan, Fei Li, Dengzhun Wang

College of Big Data and Information Engineering, Guizhou University, Guiyang Guizhou

Received: Nov. 29th, 2022; accepted: Dec. 20th, 2022; published: Dec. 30th, 2022

Abstract

Tea industry is an important aspect of China's import and export trade commodities. Tea has a long cultural heritage in China, and is closely related to the life of our people. In order to meet the acquirement of premium tea sprouts, this paper first established the dataset of tea sprouts based on natural environment, and proposed a modified YOLOX tea sprout detection model based on Swin-Transformer—YOLOX-ST. The proposed model used the Swin-Transformer as the backbone

文章引用: 刘瑞欣, 严春雨, 李飞, 王等准. 基于改进 YOLOX 的茶叶嫩芽目标检测研究[J]. 软件工程与应用, 2022, 11(6): 1404-1414. DOI: 10.12677/sea.2022.116144

network, which improves the overall detection accuracy of the model. And it also introduced the CBAM attention mechanism to solve the problem of miss-detection and wrong detection in the complex environment. Experimental results showed that the proposed model has a mAP value of 79.12%, which is 5.2% higher than the original YOLOX model, and the accuracy has achieved 90.45%, which is 4.62% higher than the YOLOX model. Compared with YOLOv3, YOLOv4, and YOLOv5, the mAP and accuracy rate of YOLOX-ST model increased by 7.09% and 6.43% at most, respectively, which has good detection accuracy and model generalization ability. This model has laid a good foundation for the intelligent picking of premium tea sprouts.

Keywords

Tea Sprouts, YOLOX, Object Detection

1. 引言

茶产业作为我国一大进出口的重要产业，在各大农业商品中占据着重要地位。名优茶拥有着良好的茶叶品质以及优良的加工工艺，从古至今都受到人们的追捧。为了满足我国茶产业高品质需求的快速发展，需要不断提高生产高端名优茶的生产能力，对茶叶的采摘质量、采摘速度与加工工艺方面提出更多、更高的要求[1]。目前，茶叶嫩芽的采摘技术主要分为人工采摘及机械采摘。机械采摘，是一种使用机械采茶机的方法，可以较好地实现茶叶嫩芽的快速采摘，但机械采摘的方法缺乏对茶叶嫩芽的选择性，所得到的茶叶嫩芽存在杂质，如茶杆、茶病叶等，增加了后期对茶叶嫩芽加工筛选的工作量，以至于严重影响了名优茶叶的质量。人工采摘可以实现对茶叶嫩芽的精确采摘，但其效率低下，容易花费大量额外的人工成本，且当今社会人口老龄化严重，不符合名优茶产业的快速、高品质发展的需求[2]。因此，研究茶叶嫩芽的智能化采摘迫在眉睫。自然环境下，茶叶嫩芽背景复杂，且生长姿态各异，因此，茶叶嫩芽的智能化采摘研究中，首要任务是研究茶叶嫩芽的识别与检测。

茶叶嫩芽的检测方法，大部分依赖于茶叶嫩芽的整体形状，以及嫩芽在颜色方面的特征提取，如吕军[3]等人提出了基于区域亮度自适应校正的茶叶嫩芽检测模型，对高亮度图像进行 2 × 3 分块和局部区域亮度自适应校正后，获得了较高的召回率。杨福增[4]等人提出了基于颜色和形状特征的茶叶嫩芽识别方法，通过对 RGB 颜色空间中提取茶叶图像的 G 分量，采用双阈值方法对图像进行分割，最后分辨出了茶叶嫩芽。但上述方法往往容易出现过分割的问题，且自然环境下的茶叶嫩芽背景复杂，仅仅依靠颜色及形状无法达到精确识别的要求。姜苗苗[5]等人提出了基于颜色因子与图像融合的茶叶嫩芽图像检测算法，实现了茶叶嫩芽与老叶的区分，但该方法容易因图像噪声等因素影响，且准确率仍有待提高。因此，由此可知，传统图像处理的方法已经不能满足高精度的茶叶嫩芽识别与检测需求。

近年来，深度学习一直是诸多学者研究的方向，基于深度学习的茶叶嫩芽检测方法也相继出现。施莹莹[6]等人创新性地使用了多种光照条件、多种类型的茶叶嫩芽数据集作为 YOLOv3 的网络输入，使得模型具有较强泛化能力与检测效果。邹倩[7]等人提出一种改进 YOLOV3 的茶叶嫩芽检测方法，模型检测精度达到了 79.9%，相比于 YOLOV3 模型，精度提高了 7.8%。

上述研究虽然都取得了不错的效果，但在模型检测精度上仍有待提高，且使用的模型较为老旧，不

利于后续的发展。本文基于自然环境下建立的茶叶嫩芽数据集，以 YOLOX 为基础模型，将 Swin-Transformer 模块引入到其骨干网络当中，并且增加了 CBAM 注意力机制，提高了模型整体泛化能力及复杂背景下对茶叶嫩芽的检测能力。

2. 图像采集与预处理

2.1. 图像采集

目前网络上还没有公开的茶叶嫩芽数据集，因此，本文构建了一个自然环境下的茶叶嫩芽数据集。于 2021 年 3 月至 2021 年 4 月，在贵阳市花溪区羊艾茶园进行茶叶嫩芽原始图像的采集，拍摄设备为 iPhoneXR 手机后置摄像头，像素为 1200 万，总共采集并挑选出 6357 张茶叶嫩芽原始图像，部分图像如图 1 所示。

Figure 1. Original tea sprouts images
图 1. 茶叶嫩芽部分原始图像

2.2. 图像预处理

使用 Labelimg 软件对茶叶嫩芽原始图像进行标注工作，标注后的图像及标签文件如图 2、图 3 所示。标注时仅将茶叶的嫩芽部分框入标签框中，并将标注的嫩芽类别命名为 teabud。标注完成后，对所有图像进行带标签的翻转、镜像、对比度调整等操作，如图 4 所示。最后，本文的数据集从原始的 6357 张图像扩充到了 10,693 张图像。

3. 基于改进 YOLOX 的茶叶嫩芽识别检测模型

3.1. YOLOX 原始模型

YOLOX 是 2021 年 Zheng Ge [8]等人提出的一个检测模型，其结构如图 5 所示。由图可知，整个 YOLOX 可以分为三个主要的部分，分别为 CSPDarknet、中间部分的 FPN 以及最后的 YOLO Head 部分。输入的茶叶嫩芽图片，首先会在 CSPDarknet 部分中进行对茶叶嫩芽的特征提取，随后输入到 FPN 部分，进行特征融合，最后在 YOLO Head 部分对特征点进行判断，判断特征点是否有物体与其对应。

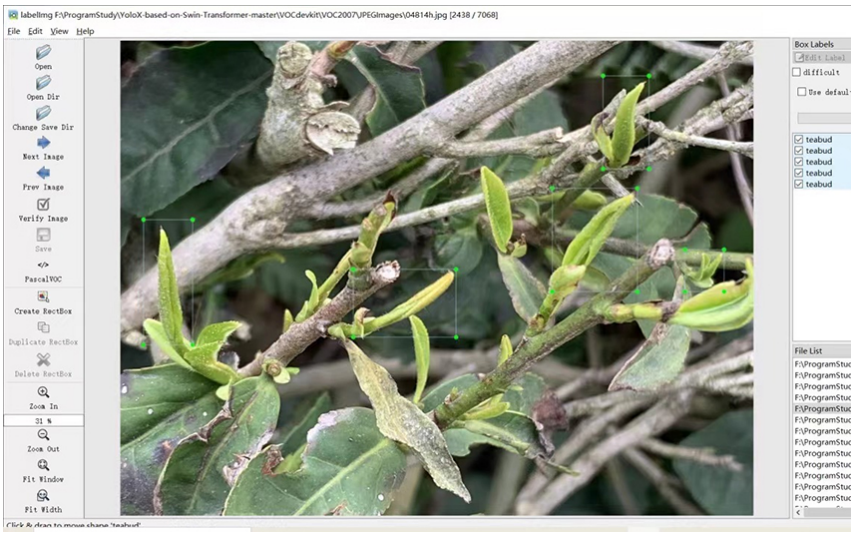

Figure 2. Label of tea sprout
图 2. 茶叶嫩芽标注

```
<annotation>
        <folder>split</folder>
        <filename>teasprout (6894).jpg</filename>
        <path>D:\AXin\dataset\split\teasprout (6894).jpg</path>
        <source>
                <database>Unknown</database>
        </source>
        <size>
                <width>4032</width>
                <height>3024</height>
                <depth>3</depth>
        </size>
        <segmented>0</segmented>
        <object>
                <name>teabud</name>
                <pose>Unspecified</pose>
                <truncated>0</truncated>
                <difficult>0</difficult>
                <bndbox>
                        <xmin>2266</xmin>
                        <ymin>1772</ymin>
                        <xmax>2724</xmax>
                        <ymax>2317</ymax>
                </bndbox>
        </object>
        <object>
                <name>teabud</name>
                <pose>Unspecified</pose>
                <truncated>0</truncated>
                <difficult>0</difficult>
                <bndbox>
```

Figure 3. The XML file of tea sprout
图 3. 茶叶嫩芽标签文件

CSPDarknet 作为整体网络的主干，具有四个重要的特点。第一，使用了多个残差网络。CSPDarknet 中的残差卷积由一个 1 × 1 的卷积和一个 3 × 3 的卷积组成，其残差边部分不做任何处理。第二，使用了 Focus 网络结构。Focus 网络首先是在一张图片中每间隔一个像素取出一个值，此时获得了四个独立的特征层，然后将这四个特征层进行堆叠操作，拼接后的特征层相对于原先的 3 通道变成了 12 通道。第三，使用了 SiLU 激活函数。SiLU 激活函数具有无上界但有下界、平滑、非单调的特性。它的效果要优于 ReLU 函数。第四，使用了 SPP 结构。在 YOLOX 中，SPP 模块被用在了主干特征提取网络中，通过不同池化核大小的最大池化进行茶叶嫩芽的特征提取，提高整个网络的感受野。

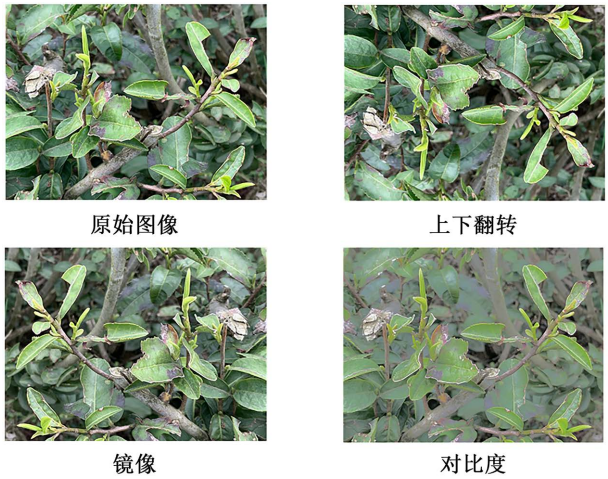

Figure 4. Data expansion
图 4. 数据扩充

YOLOX 网络利用提取的多特征层来进行目标检测，一共提取三个特征层，分别位于主干部分 CSPDarknet 的中间层，中下层以及最底层，利用 FPN 特征金字塔，可以获得三个加强特征层，接着利用这三个 shape 的特征层传入 YOLO Head 来获得最终的预测结果。

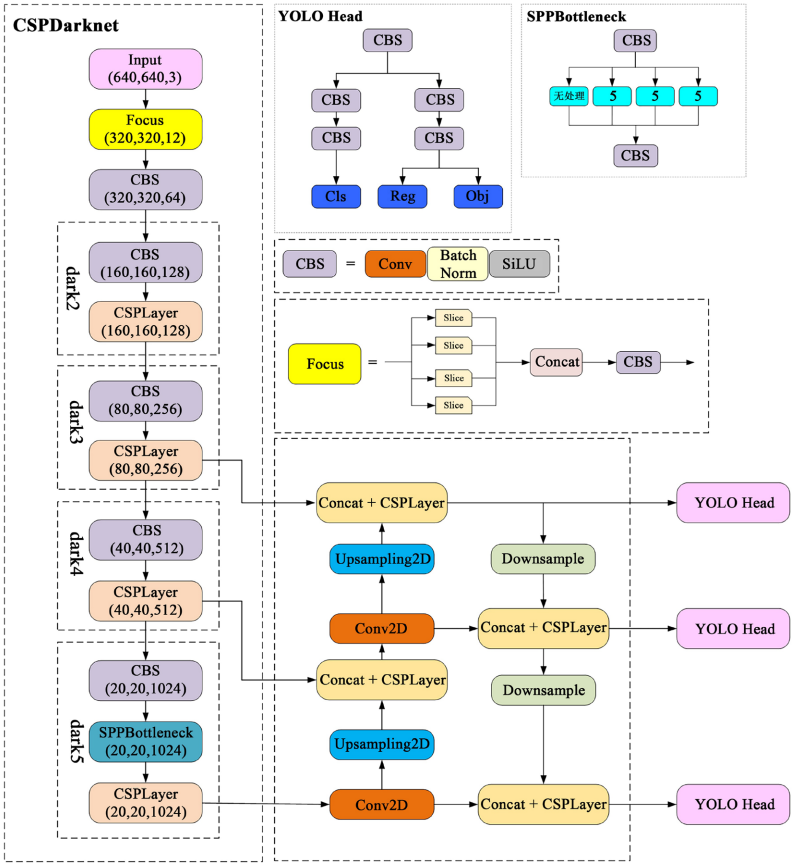

Figure 5. Structure of YOLOX base model
图 5. YOLOX 原始模型结构

3.2. Swin-Transformer

Transformer [9]是近年来深度学习中大火的一个模块，首先被提出来应用于 NLP 领域，并且在该领域大获成功。受此启发，2021 年 Alexey Dosovitskiy [10]等人提出了 Transformer 在 CV 领域的应用——Vision Transformer (VIT)，文章中给出的最佳模型在 ImageNet1K 公共数据集上能够达到 88.55%的准确率，也就是说，Transformer 这模块应用在 CV 领域这一想法是可行的。

Swin-Transformer [11]作为 VIT 的一个改进版本，与 VIT 不同的是，Swin-Transformer 使用了类似卷积神经网络中的层次化构建方法(Hierarchical feature maps)，比如特征图尺寸中有对图像进行下采样 4 倍，8 倍以及 16 倍，这样的骨干网络有助于在此基础上构建目标检测，实例分割等任务。它不仅具有 Transformer 模块关注全局信息的能力[12]，而且采用了移动窗口的方法，实现了跨窗口连接，使模块可以关注到相邻窗口的相关信息，这样可以扩大模型的感受野，提高模型整体的效率。

Swin-Transformer 的结构图如图 6 所示。

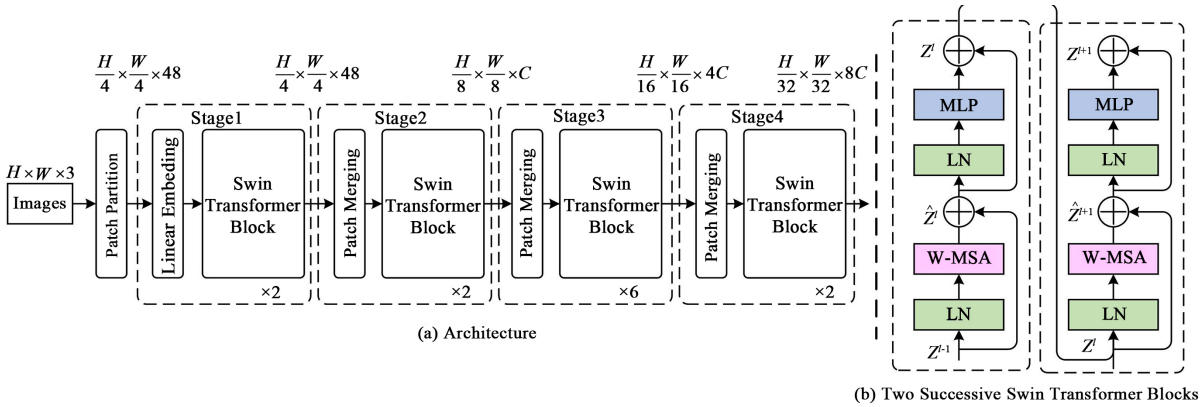

Figure 6. Structure of Swin-Transformer
图 6. Swin-Transformer 结构图

Swin-Transformer 首先将图片输入到 Patch Partition 模块中进行分块，即每 4 × 4 相邻的像素为一个 Patch，然后再通过线性嵌入层(Linear Embedding)对每个像素的通道数据做线性变换。接着通过四个 Stage 构建不同大小的特征图，除了 Stage 1 中先通过一个线性嵌入层外，剩下三个部分都是先通过一个 Patch Merging 层进行下采样，最后对于分类网络，后面还会接上一个 Layer Norm 层、全局池化层以及全连接层得到最终输出。

3.3. CBAM 注意力机制

CBAM (Convolutional Block Attention Module)是一种轻量的注意力模块[13]，可以分别在通道和空间维度上进行提高注意力的操作，结合了两个方面的注意力信息。与单一的关注通道特征的注意力机制和单一关注空间注意力机制的方式相比，CBAM 具有更加优异的性能。CBAM 的结构如图 7 所示。

CBAM 包含 CAM (Channel Attention Module)和 SAM (Spartial Attention Module)两个子模块，其结构图如图 8 所示。

CAM 模块的主要作用是获取茶叶嫩芽特征中不同通道特征之间的相互关系[14]，并学习相应的权重分布，从而找到并突出其中最有意义的茶叶嫩芽通道特征。对于输入的茶叶嫩芽特征 $F \in R^{C\times H\times W}$，CAM 首先使用最大池化(Maxpool)和平均池化(Avgpool)的操作，将特征的空间信息聚合在一起，得到两个空间向量。然后将这两个空间向量分别送入多层感知机网络(MLP)，最后将输出的两个特征进行相加融合，

通过 Sigmoid 函数激活后生成通道注意力权重 M_c。将 M_c 乘以茶叶嫩芽特征 F，变得到中间特征 F'。CAM 的计算公式如式(1)所示。

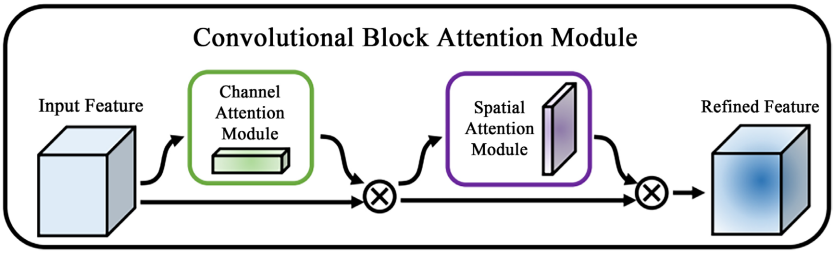

Figure 7. Structure of CBAM Attention Mechanism
图 7. CBAM 注意力机制结构图

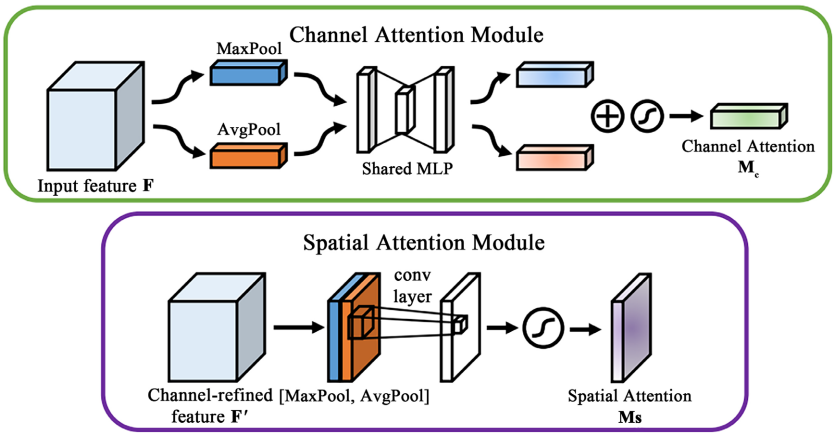

Figure 8. Structure of CAM and SAM
图 8. CAM 及 SAM 结构图

$$M_c(F)=\sigma\left(W_1\left(W_0\left(F_{avg}^C\right)\right)+W_1\left(W_0\left(F_{\max}^C\right)\right)\right) \tag{1}$$

式中 σ 为 Sigmoid 激活函数，$\left(F_{\max}^C\right)$ 和 $\left(F_{avg}^C\right)$ 分别表示最大池化和平均池化，$W_0 \in R^{C/R\times C}$ 和 $W_1 \in R^{C/R\times C}$ 为 MLP 网络的两个权重值。

与 CAM 不同，SAM 的主要功能是关注茶叶嫩芽特征的空间位置信息，给出其最有信息量的空间位置，是对 CAM 的一种补充机制[15]。对于 CAM 输入的 F'，通过 Maxpool 及 Avgpool，获得两个单通道特征，并将这两个单通道特征进行拼接操作，最后使用 3 × 3 的卷积进行通道融合，经过 Sigmoid 函数运算，得到空间注意力权重 M_s，将 M_s 乘以中间特征 F'，最终得到所输出的特征 F''。SAM 的计算公式如式(2)所示。

$$M_s(F)=\sigma\left(f^{3\times3}\left(F_{\max}^S;F_{avg}^S\right)\right) \tag{2}$$

式中 σ 为 Sigmoid 激活函数，$f^{3\times3}$ 表示 3 × 3 的标准卷积，$F_{\max}^S$ 和 F_{avg}^S 分别表示最大池化和平均池化。

3.4. YOLOX-ST 模型

在 Swin-Transformer 取出的三个有效特征层中，分别为下采样 8 倍、16 倍以及 32 倍的特征层，正好对应原始 YOLOX 网络中 CSPDarknet 所输出的特征层。因此，本文提出了一种以 Swin-Transformer 作为

骨干网络的 YOLOX-ST 模型，并在模型下采样后引入了 CBAM 注意力机制，以此来提升模型对茶叶嫩芽的检测精度。YOLOX-ST 模型的结构图如图 9 所示。

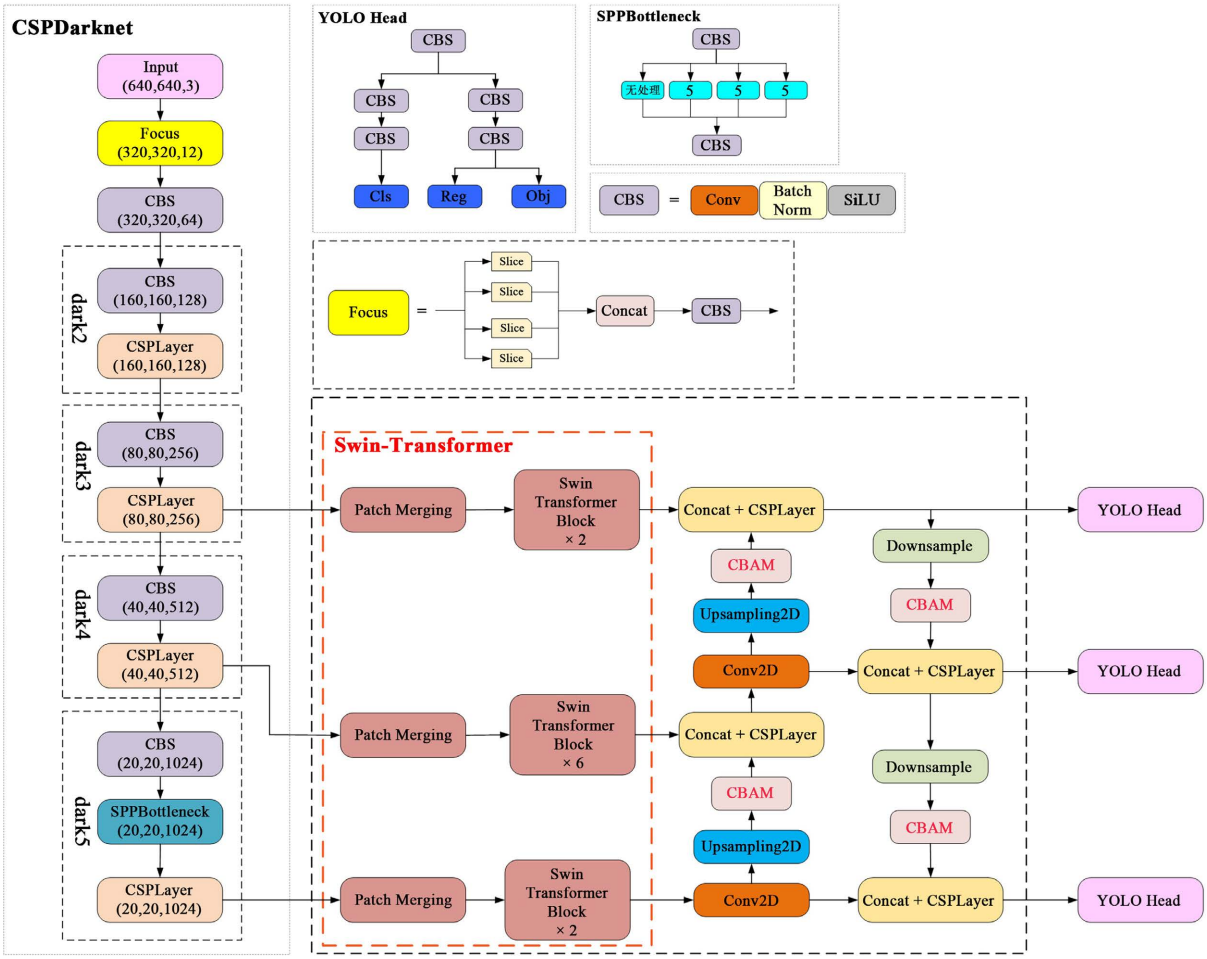

Figure 9. Structure of YOLOX-ST model
图 9. YOLOX-ST 模型结构图

4. 实验结果及分析

4.1. 实验环境

实验系统环境使用的 GPU 是 RTX3050 显卡，使用的编程语言为 Python，版本为 3.8，训练模型所使用的 CUDA 版本为 11.1，对应的 PyTorch 版本为 1.7.0，如表 1 所示。

Table 1. Experimental environment configuration
表 1. 实验环境配置

系统环境	Windows 10
GPU 版本	RTX3050
CUDA 版本	11.1
编程语言	Python
PyTorch 版本	1.7.0

4.2. 实验结果分析

本文输入的图片大小为 640 × 640，有三个通道。用此数据集分别对 YOLOX 原网络和改进后的网络 YOLOX-ST 进行训练和验证，所用图像共 10,693 张，训练集、验证集、测试集按 8:1:1 的比例来划分，epoch 设置为 400，学习率设置为 0.0001，batch size 设置为 16，使用 SiLU 损失函数。

为了验证 YOLOX-ST 整体模型的有效性，本文将训练集分别送入 YOLO 系列不同网络模型中进行训练作为对比，所得到的准确率及 mAP 如表 2 所示。

Table 2. Comparison of different models
表 2. 不同模型对比

Network	mAP (%)	Precision (%)
YOLOv3	72.91	85.2
YOLOv4	72.03	84.15
YOLOv5s	72.06	84.02
YOLOv5m	72.45	84.91
YOLOX	73.92	85.83
YOLOX-ST[a]	**79.12**	**90.45**

[a] 加粗字体表示最优结果。

由表 2 可知，本文提出的 YOLOX-ST 模型不管是从 mAP 上还是从准确率上都有一个良好的提升。相较于原始的 YOLOX 模型，mAP 和准确率分别提升了 5.2%及 4.62%，而与其他模型相比，优化后的模型 mAP 最高提升了 7.09%，准确率最高提升了 6.43%。

训练后的模型与原始 YOLOX 模型的准确率曲线及 Loss 曲线比较如图 10、图 11 所示。由图可知，YOLOX-ST 模型在准确率上的提升速度比原始 YOLOX 模型要快，在 100 epoch 之后便显现出了明显的差距。且 YOLOX-ST 模型的起始 Loss 也比原始模型的低两个点，下降速率也比原始模型较快，由此可以更直接地观测到本文所提出模型的有效性。

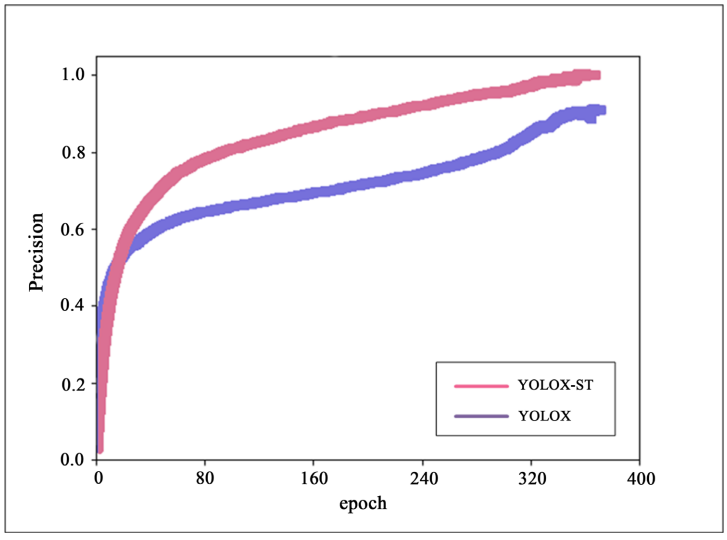

Figure 10. Precision curve
图 10. 准确率曲线

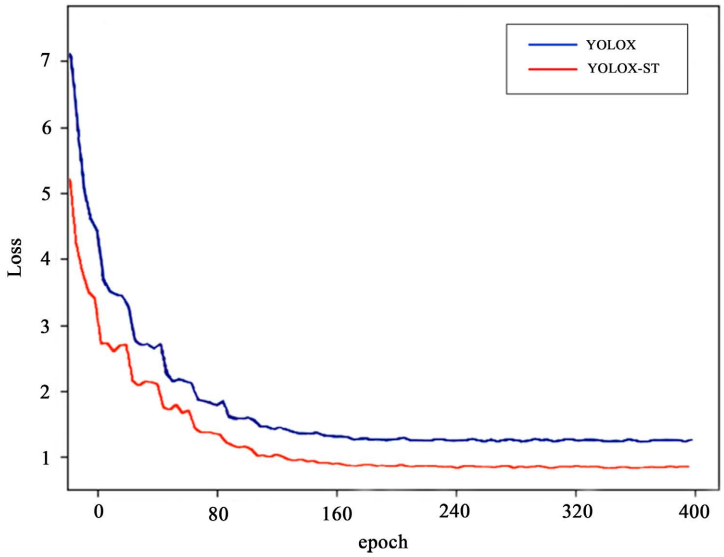

Figure 11. Loss curve
图 11. Loss 曲线

为了验证 Swin-Transformer 以及 CBAM 注意力机制的合理性，本文进行了消融对比试验，结果如表 3 所示。

Table 3. Ablation test
表 3. 消融试验

Group	Swin-Transformer	CBAM	Precision (%)	mAP (%)
1			85.83	73.92
2	√		88.91	75.68
3		√	86.35	74.16
4	√[b]	√	90.45	79.12

[b]“√”表示在网络中使用该方法。

由表 3 可知，无论是 Swin-Transformer 还是 CBAM 注意力机制，应用在模型中时都保持着有效性，也即验证了本文提出的方法的合理及可行性。

5. 结论

本文针对自然环境下的茶叶嫩芽数据集，提出了一种基于 YOLOX 的改进模型 YOLOX-ST，该模型引入了 Swin-Transformer 模块作为骨干网络，有效地优化了模型对于茶叶嫩芽的特征提取，并在茶叶嫩芽特征提取下采样后加入了 CBAM 注意力，加强了茶叶嫩芽特征的信息量，提升了模型整体的检测准确率。实验表明，与其他 YOLO 系列模型以及原始 YOLOX 模型相比，YOLOX-ST 模型具有准确率高，泛化能力较好的特点，为后续的茶叶嫩芽智能化采摘奠定了一个良好的基础。

参考文献

[1] 严春雨, 李飞. 基于改进 MobileNetV2 的茶叶病害识别方法[J]. 软件工程与应用, 2022, 11(4): 743-750. https://doi.org/10.12677/SEA.2022.114077

[2] 夏华鹃, 史必高, 黄海霞, 等. 图像处理在茶叶嫩芽智能采摘中的应用进展[J]. 安徽农学通报, 2019, 25(9): 133-134.

[3] 吕军, 方梦瑞, 姚青, 等. 基于区域亮度自适应校正的茶叶嫩芽检测模型[J]. 农业工程学报, 2021, 37(22): 278-285.

[4] 杨福增, 杨亮亮, 田艳娜, 等. 基于颜色和形状特征的茶叶嫩芽识别方[J]. 农业机械学报, 2009, 40(z1): 119-123.

[5] 姜苗苗, 问美倩, 周宇, 等. 基于颜色因子与图像融合的茶叶嫩芽检测方法[J]. 农业装备与车辆工程, 2020, 58(10): 44-47.

[6] 施莹莹, 李祥瑞, 孙凡. 基于 YOLOv3 的自然环境下茶叶嫩芽目标检测方法研究[J]. 电脑知识与技术, 2021, 17(3): 14-16.

[7] 邹倩, 陆安江, 周骅, 等. 改进 YOLOV3 的茶叶嫩芽检测研究[J]. 激光杂志, 2022(3): 43.

[8] Ge, Z., Liu, S.T., Wang, F., Li, Z.M. and Sun, J. (2021) YOLOX: Exceeding YOLO Series in 2021.

[9] Devlin, J., *et al*. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.

[10] Dosovitskiy, A., Beyer, L., *et al*. (2021) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.

[11] Liu, Z., Lin, Y.T., *et al*. (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows.

[12] 田应仲, 卜雪虎. 基于注意力机制与 Swin Transformer 模型的腰椎图像分割方法[J]. 计量与测试技术, 2021, 48(12): 57-61.

[13] Woo, S., Park, J., *et al*. (2018). CBAM: Convolutional Block Attention Module.

[14] 余帅, 汪西莉. 基于多级通道注意力的遥感图像分割方法[J]. 激光与光电子学进展, 2020, 57(4): 134-143.

[15] 张连超, 乔瑞萍, 党祺玮, 等. 具有全局特征的空间注意力机制[J]. 西安交通大学学报, 2020, 54(11): 129-138.