**基于视觉Transformer的农作物病害识别算法**

王杨1 许佳炜1 赵传信1 王傲1 李迎春1

1（安徽师范大学计算机科学与技术学院 芜湖 130012）

1（liulu12@mails.jlu.edu.cn）

#### Plant Disease Recognition Algorithm Based on Visual Transformer

Wang yang1 ,Xu jia wei1

1（*College of Computer Science and Technology*, *Jilin University*, *Changchun* 130012）Depart.Correspond小五号

2 （*Key Laboratory of Symbol Computation and Knowledge Engineering*（*Jilin University*）, *Ministry of Education*, *Changchun* 130012）

**Abstract** Crop disease detection algorithms based on Deep Convolutional Neural Networks (DCNNs) have been widely studied by scholars. The DCNN algorithm has high training accuracy in the laboratory environment, but lacks robustness in the face of real and noisy environments. Recently, Vision Transformer (ViT) applied the Transformer architecture in the field of Natural Language Processing (NLP) to the image classification task, and achieved comparable performance to DCNN in classification accuracy. And the Self-Attention mechanism in the Transformer architecture can learn the relationship between the pixels of the entire picture, so the classification algorithm of the ViT model has higher robustness in the face of reality and noise. However, the traditional ViT model lacks prior inductive paranoia, has poor performance from zero learning and requires a large amount of data for training. This paper improves the traditional ViT model and adds Enhanced Patch Embedding (EPE) and Inter-Attention (IA) mechanisms to make up for the lack of local induction bias of the ViT model, so that it can efficiently learn from zero in small and medium-scale plant disease datasets , the classification accuracy compared with the DCNN network is obtained, and when noise, mean blur and Gaussian blur are added to the test dataset, the robustness of the improved ViT in this paper is higher than that of the traditional DCNN network.

**Key words** Crop disease detectio; Deep Convolutional Neural Networks; Vision Transformer

摘要 基于Deep Convolutional Neural Networks（ DCNNs）的农作物病害检测算法已经被广大学者广泛研究。DCNN算法在实验室环境下训练准确率高，但面对真实和噪声环境下缺少鲁棒性。最近，Vision Transformer（ViT）将Natural Language Processing（NLP）领域的Transformer架构应用于的图像分类任务中，在分类准确率上得到与DCNNs媲美的性能。并且Transformer架构中的Self-Attention机制可以学习整张图片各像素点之前的关系，因此ViT模型的分类算法在面对真实，有噪声时有着更高的鲁棒性。但传统ViT模型缺乏先验的归纳偏执，从零学习性能较差并且需要大量数据训练。本文改进传统的ViT模型，加入Enhanced Patch Embedding（EPE）和 Masked Multi-head Self-Attention （MMSA）机制，弥补ViT模型缺乏局部性归纳偏差的不足，让其可以高效得在中小规模的植物病害数据集从零学习，得到与DCNN网络相较的分类准确率，并且在测试数据集上添加噪声，均值模糊和高斯模糊时，本文的改进的ViT的鲁棒性高于传统DCNN网络。

关键词农作物病害识别；计算机视觉；卷积神经网络；视觉Transformer；数据增强；

中图法分类号 TP391

在农业生产中，病虫害是造成作物减产、欠收的重要原因。我国是世界上农作物病虫害发生最严重的国家之一，危害农作物的害虫830多种、病害720多种，农作物受各种病虫杂草为害而造成产量损失每年30%以上，经大力防治后每年仍损失粮食1600万吨.病害初期能够及时正确的分析病害类别，作出快速准确的反应，对病害的防治具有重要意义。

感染病害的农作物会出现一系列病症，如叶片上的色斑或条纹。随着病害的加重，叶片的形状，颜色，纹理特征也不断变化，感染病害的作物叶片和健康的作物叶片在视觉上存在一定的差异，因此，通过对叶片图像应用机器学习技术，可以自动检测作物病害类别[1]。

传统基于图像的农作物病害识别算法先手工提取叶片特征，然后使用这些特征训练浅层分类器算法,如支持向量机(Support Vector Machines，SVM)[2]和浅层非卷积神经网络[3]等。但手工提取特预处理复杂，耗时，无法客观描述整张图片的特征，相较于深度学习算法，缺少鲁棒性. 基于深度学习植物病害识别算法不需要特征工程来手动提取特征，可以直接将图片输入到深度神经网络模型中，得到类别预测结果.同时近年来，高性能计算和图形处理单元的发展，使得训练深层次的神经网络模型成为可能.不少学者将最先进的深度卷积神经网络模型应用在农作物病害识别上.并取得令人满意的结果[4]-[6]。

然而CNNs网络感受野有限，缺乏全局的特征表示能力[4]。近年来，Transformer作为一种基于注意力的encoder-decoder模型，已经彻底改变了自然语言处理(natural language processing,NLP)领域[5]. Alexey等人受此启发，将Transformer的Self-Attention结构应用于图像分类任务的视觉Transformer (Vision Transformer, ViT)取得了优于卷积神经网络的效果[6].本文为了兼顾农作物病害识别算法的识别精度和鲁棒性，提出基于ViT模型的农作物病害算法。

本文工作的主要贡献有以下几点:

1. 将ViT模型引入到农作物病害工作中。
2. 改进标准ViT模型的Patch Embedding和Self-Attention操作，使标准ViT模型具有更好的局部归纳偏置和局部归纳偏置。
3. 通过大量实验证明ViT模型在噪声环境下比深度卷积神经网络模型具有更高的鲁棒性。

## 1 相关工作

利用图像处理技术提取特征，并将其输入到机器学习算法中进行精确分类。简而言之，这些方法可以大致分为:(1)机器学习方法(2)深度学习方法。

Rumpf等人利用支持向量机研究了基于植物光谱指数的甜菜病害的早期诊断和分类, 对健康甜菜叶片和患病甜菜叶片的分类准确率高达97%[2]. Albashish提出四阶段病害识别算法，首先对叶片图片的颜色空间进行设备独立性的预处理，然后使用K-Means聚类分割患病区域，然后对分割后图片进行基于颜色和纹理的特征提取，最后通过浅层非卷积神经网络进行分类[3]。

Edna Chebet Too等人比较VGG 16, Inception V4, ResNet50, ResNet101和ResNet152 以及DenseNet121, 在PlantVillage 14种植物的38个不同类别（包括病叶和健康叶图像），DenseNet121需要更少的参数和合理的计算时间来实现最先进的性能，测试准确率达到了99.75%[7]。Atila等人在effecentnet架构基础上加上迁移学习在增强的PlantVillage数据集上识别精度达到99.39%[8].尽管上述工作已经取得了非常好的结果，但对图像添加噪声，模拟自然环境常见的图像变异时，识别准确率严重下降. Joao Paulo等人提出可以将图像由RGB颜色通道转换为CIE-LAB，L和AB通分别进入修改后的Inception V3架构的前三层的不同分支中提取特征，这种方法不仅节约了网络的参数，而且面对噪声干扰时有更好的鲁棒性[9]。

与卷积神经网络中固定的卷积运算不同，Transformer中的Self-Attention机制让其可以更加灵活捕捉远距离信息. 因此，Transformer网络比DCNN网络更加强大，更加灵活，在计算机视觉领域会有更大的进步[10]。Alexey等人提出的ViT模型用完全用线性映射的取代标准卷积运算，证明纯粹的transformer也可以表现出与最先进的卷积神经网络一样好的性能[6]。

然而，又有学者证实如果缺少预训练，ViT在图像分类任务中表现不佳[11]。这是由于ViT缺少像卷积神经网络那样的归纳偏置，并且ViT的模型容量很大，在小规模数据集上面容易造成过拟合。为了让ViT模型具有一定的归纳偏置，并且规范其模型容量，liu等人提出的Swin Transformer通过引入像CNN网络的层次结构来接收不同的感受野[12]。Li等人则通过优化优化自注意力机制对ViT模型进行改进[13]。然而，这些优化方法让ViT模型在小规模数据集从零学习任具有挑战。

为了让ViT模型可以高效的在小规模数据集上从零学习，Seung Hoon Lee等人提出轮转分块序列化（Shifted Patch Tokenization ，SPT），在Tiny-ImageNet数据集上得到了2.96%的性能提升[14]。因此，本文改进上述工作在农业病害识别领域引入ViT模型，优化后的模型可以在小规模的病害数据集上从零学习。

**2视觉Transformer病害识别模型**

为了解决CNN网络在植物病害识别上鲁棒性差的问题，本文构建了一个基于视觉Transformer的植物病害识别模型。该模型引入两个关键方法: Enhanced Patch Embedding（EPE）和 Masked Self-Attention（MSA）,解决了标准视觉Transformer模型缺乏局部性归纳偏置的不足，同时让模型不使用迁移学习技术，从零学习也可以得到较高准确率。本章节首先介绍基于视觉Transformer的植物病害识别模型整体架构图，接着分析标准视觉Transformer的序列化和自注意力机制缺乏局部性归纳偏置的原因，介绍本文使用的EPE和MSA机制。最后介绍本文的实验细节，包括对比网络，数据集，实验环境，网络参数微调等。

**2.1** 视觉Transformer的整体架构

本文在标准视觉Transformer的基础上引入EPE和MSA后，模型整体架构如图1所示。首先输入图片会调整为224×224的大小，送入到EPE结构，EPE层通过对输入图像的偏置与堆叠，产生的特征序列会包含更多的空间信息。EPE操作产生196个维度为768的特征向量，这些特征向量还需要拼接一个类别向量以便后续的分类操作。到目前为止，我们得到的197 ×768的特征序列是没有位置信息的，即这197个向量的对位置是不不敏感的，可以随意交换向量的位置。然而对于图片来说，不同位置区域产生的特征向量应该有顺序之分，特征向量之间需要有一定的相对位置信息，因此需要对197维度为768的向量加上位置信息，对于位置信息的加入可以采用自学习的位置编码和确定的位置编码，本文模型采用的是自学习的位置编码方式。随后对添加上位置信息的特征序列经过一次Dropout操作，送入到MMSA模块中，MMSA模块不会改变特征向量的个数以及维度，并且可以重复堆叠，本文模型中我们堆叠12个MMSA块提取特征信息，最后通过Layer Norm操得到网络的输出。

**2. 2**图像序列化

输入的图片可以用表示，,和 分别表示图片的高度，宽度和通道数，这样的数据格式不能满足Transformer的输入要求。标准ViT 模型将输入图片切分成不重叠的区域块，再拉平这些区块得到个特征向量，表示第个拉平的特征向量，表示分块大小。如公式1所示：



随后将特征序列进行一次线性映射以满足Transformer 的编码器层可接收向量维度要求，映射后的序列我们成为视觉特征序列。线性映射可以用公式2表示：



是一个可学习的线性映射参数，是 Transformer Encoder块的隐藏层维度。

标准ViT的序列化操作产生的视觉特征向量缺乏像CNNs网络的空间归纳偏置（Spatial Inductive Bias）和局部归纳偏置（Local Inductive Bias），而标准ViT模型丢失的先验知识，需要大量数据从全局特征建模中再次挖掘出来。因此，标准ViT模型需要大量数据进行大量训练才可以得到较好的结果。为了增强标准ViT模型的空间归纳偏置和局部归纳偏置, 与Swin Transformer[12]采用层次结构的方式不同，本文引入EPM（Enhanced Patch Embedding）机制，如图2所示。EPM实现的方式简单，且不会过多的增加网络计算量，对输入图像分步长的进行左，右，上，下偏置, 一张输入图片，经过上述操作后会生成个图片。然后将输入图片和偏置后的图片进行拼接，然后将拼接后的整体划分成个区域块

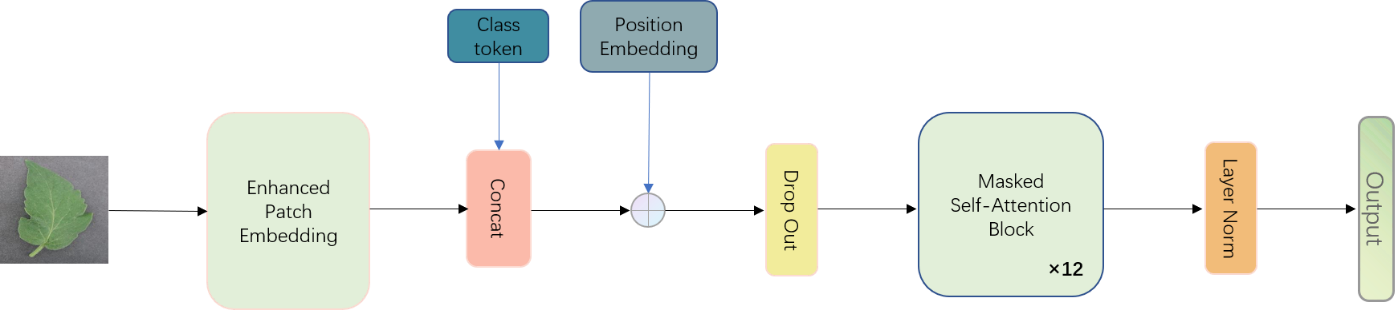


Fig. 1

图 1基于 EPE 和 MSA 的视觉 Transformer 整

体架构图

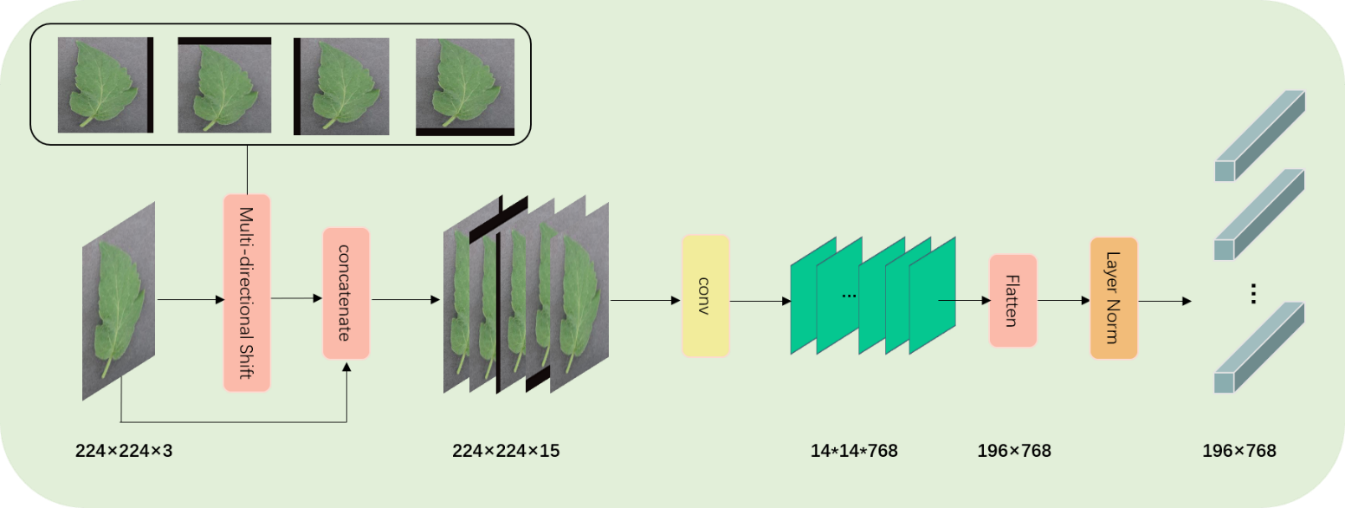


Fig. 2 Enhanced Patch Embedding

图 2 EP

随后对这些区域块展平，使用层归一化（Layer Normallization，LN）[15]和线性映射,表示图片偏置操作的次数。EME 操作如公式 3所示：



对于图片分类任务而言，需要在EPE操作生成的特性向量拼接一个类别词符（class token），同时Transformer网络使用向量序列作为输入，编码模块的多头注意力机制和MLP都无法感知到输入向量的位置信息。因而，Transformer需要加入位置编码，本文使用的位置编码为可学习的一位置编码。拼接类别词符和加上位置编码后的特征序列如公式4所示：



**2. 3**改进的自注意力机制

在自然语言处理领域，利用注意力机制捕捉远距离上下文信息，解决输入特征序列过长的遗忘现象[16]。Transformer网络的使用的自注意力机制可以让特征序列获得全局的特征表示，对于输入的特征序列，通过线性映射，和，可以得到查询(Query) ，键(Key)和值 (Value)，其中，和表示查询，键和值向量的维度。自注意力机制计算过程如公式 5表示:



思考Transformer的自注意力计算公式可以发现一个问题：query矩阵与Key矩阵由相同输入经过线性映射而来, 因此它们在对应位置的数值存在一定的相似性。这种相似性会让Query矩阵与Key经过矩阵运算后的注意力分布矩阵对角线上的数值更大，造成输入的特征向量在自注意力运算时更加关注自身。为了解决Transformer的自注意力矩阵过于关注自身向量。本文使用一个掩码矩阵让自注意力矩阵对象线上元素的值无穷大，修改后的自注意力机制计算过程如公式 6所示：



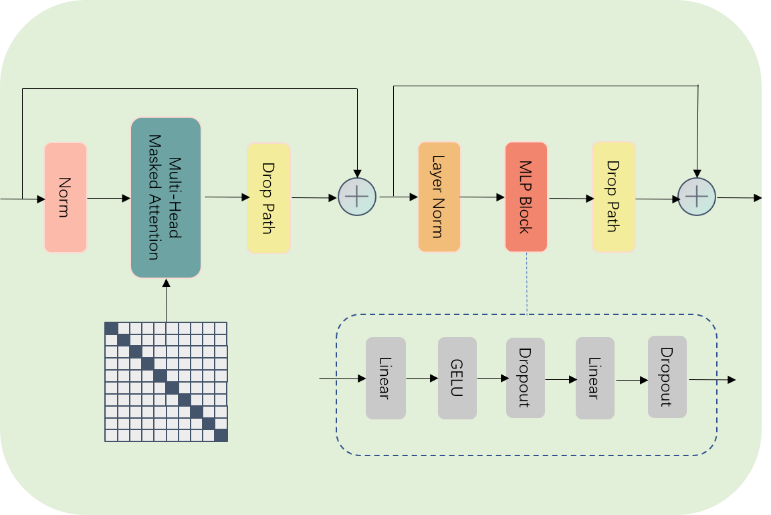


Fig.3

图3 Masked Multi-Head Self-Attention 结构图

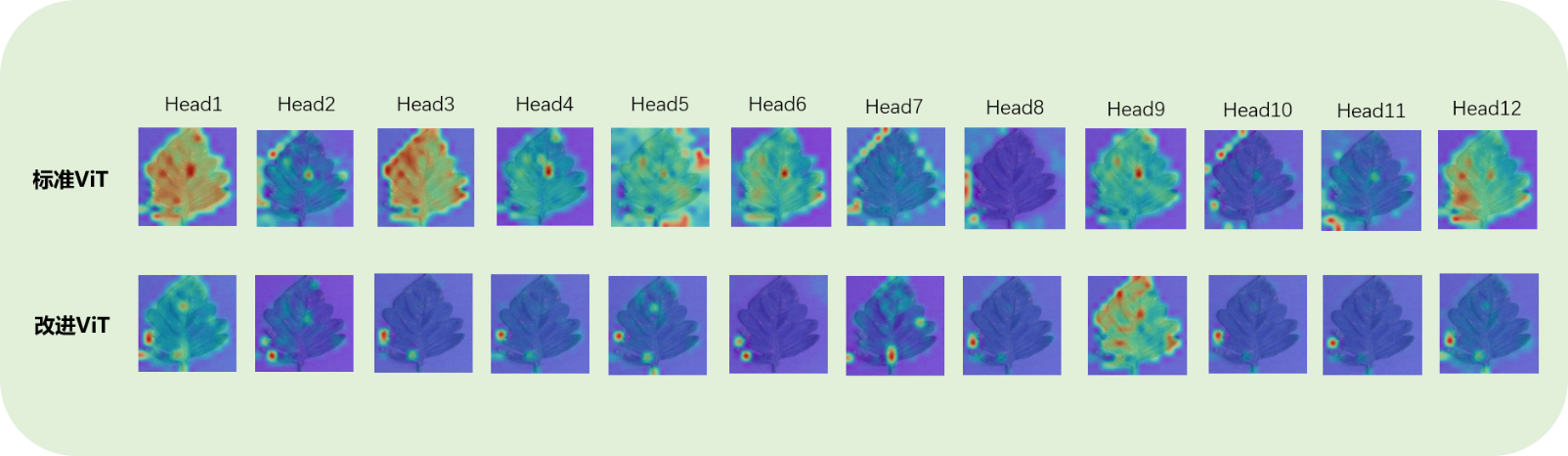
****

Fig. 4 Attention

图4类别词符注意力分数可视化

**2. 4** 注意力分布比较

图4展示了标准ViT模型与本文改进的ViT模型的类别词符作为query在浅层自注意力块的注意力得分的可视化结果。将注意力得分可视化在原始输入图片上，颜色越深表明类别词符更加关注该区域。可以很明显看出加入EPE 和 MMSA后，类别词符的注意力得分更加集中，更加关注于图像的病害位置，而标准ViT模型的类别词符的注意力得分更为分散，类别词符容易受到叶片边缘以及叶片纹理的影响。因此，相较而言改进的ViT模型在浅层的自注意力块就可以学习到全局的特征表示，这样一来改进的ViT模型学习成本更低，对数据的依赖也会降低。

**3 数据增强与实验设计**

**3.1** 实验环境

本文试验环境操作系统为 Ubuntu 20.04LTS，内存为 32GB，搭载 AMD 5600x 处理器，显卡为NVIDIA GeForce 1080(8G)，深度学习后端框架采用 Pytorch1.3.

**3.2** 数据集与数据增强

本文使用PlantVillage[17]的番茄子集，包含了9种病害图片和1种健康叶片图片。原始PlantVillage的番茄晚疫病图片和番茄花叶病图片分别为952张和373张，为了让各个病害数据分布更加均匀，本文使用高斯噪声与盐燥声的像素级变换方式以及旋转与仿射变换的空间级变换方式对这两类病害图片进行扩充，图5展示了数据增强前后各个病害的图片数量对比。

**3.3** 实验设计

为了验证视觉 Transformer 病害识别模型的分类能力，分别对比Resnet50，inceptionV3，densenet 使用迁移学习的训练准确率,同时对比了标准 Vision Transformer 从零学习的训练准确率。在测试时候加入高斯噪声和盐噪声，验证本文模型的鲁棒性.

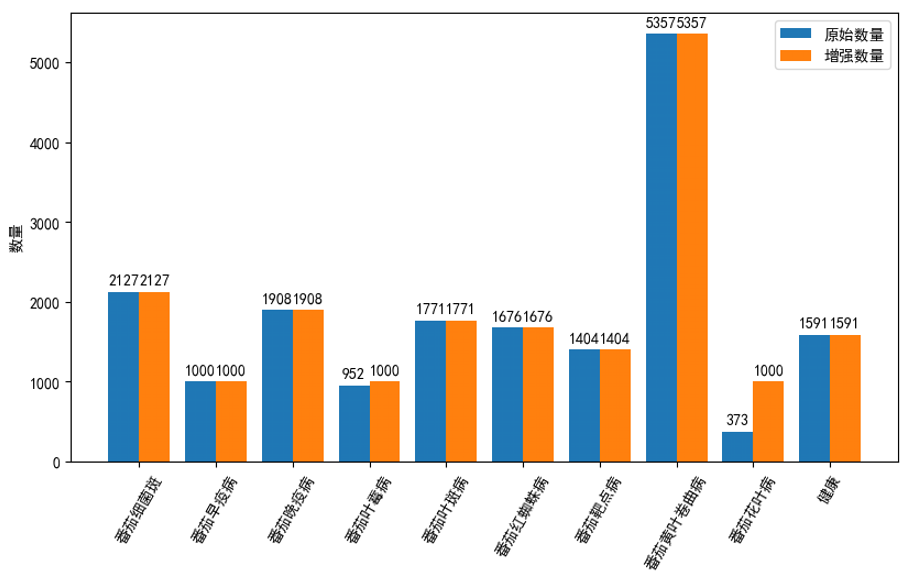


Fig.5

图5番茄子集数据增强前后的数量对比

**3.4** 评价指标

本文使用灵敏性(Sen)，特异性(Spe)和准确性(Acc)，在PlantVillage的番茄病害子集上验证模型的性能。下面公式给出Sen，Spe和Acc具体定义：







其中表示属于病害K并被识别的个体数，表示不属于病害K但被错误识别为病害K的个体数，表示不属于病害K但被正确识别的个体数，表示属于病害K但被错误识别其他病害的个体数。

**4 实验结果分析**

本节通过实验结果验证EPE\_MMSA-VIT模型的性能，4.1节比较EPE\_MMSA-VIT与标准ViT在小数据集上从零学习的能力，4.2节比较EPE\_MMSA-VIT与CNN网络的病害分类性能，4.3节比较EPE\_MMSA-VIT模型与CNN在噪声环境下的鲁棒性。

**4.1** 从零学习能力

标准ViT模型由于缺乏先验的局部归纳偏置和计算自注意力时过于关注自身向量等问题，其在小规模数据集上从零学习能力较差。本文提出的EPE\_MMSA-VIT模型对比标准ViT模型在的训练准确率如图6所示。可以看出EPE\_MMSA-VIT对比标准ViT模型，有着更高的训练准确率并且收敛的更快。

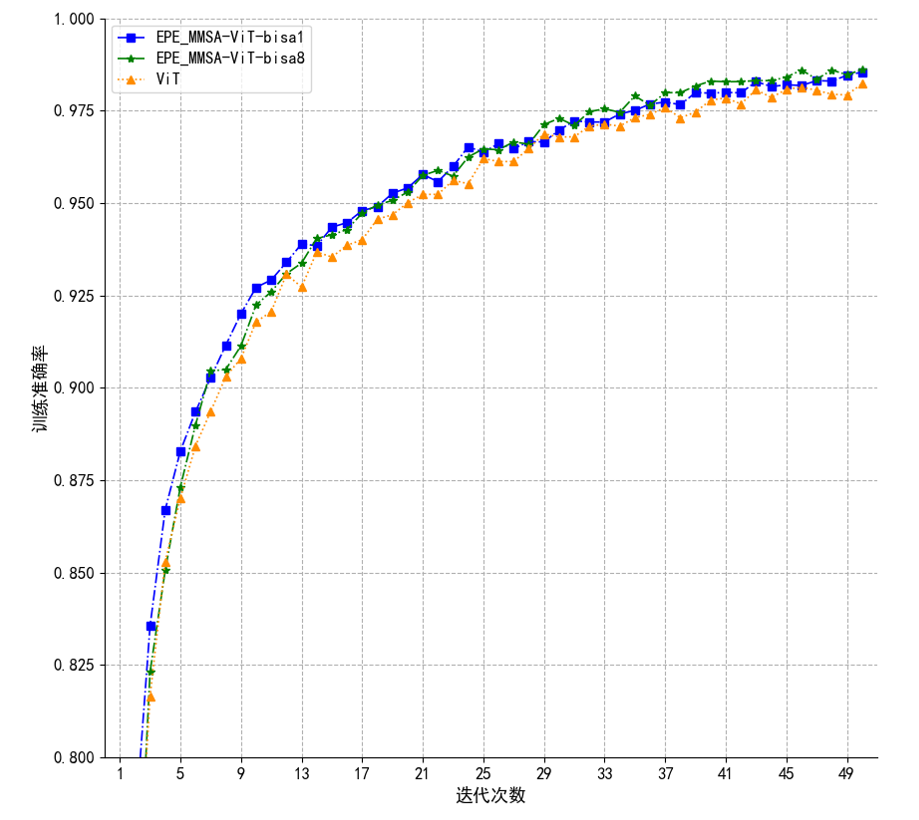


Fig.6

图6从零学习能力

**4.1** 病害分类性能

本文提出EPE\_MMSA-VIT对比ResNet50, InceptionV3, DenseNet121以及MobileNet的性能如表一所示。看出EPE\_MMSA-VIT模型在

|  |  |  |  |
| --- | --- | --- | --- |
|  | Acc | Sen | Spe |
| ResNet50 |  |  |  |
| InceptionV3 |  |  |  |
| DenseNet121 |  |  |  |
| MobileNet |  |  |  |
| EPE\_MMSA-VIT | 99.55 |  |  |

图6给出了混淆矩阵来评估EPE\_MMSA-VIT模型的识别病害的性能。混淆矩阵的纵轴真实标签，横轴表示预测标签。混淆矩阵的对角线的数值表示正确识别的数量，其余表示错误识别的数量。

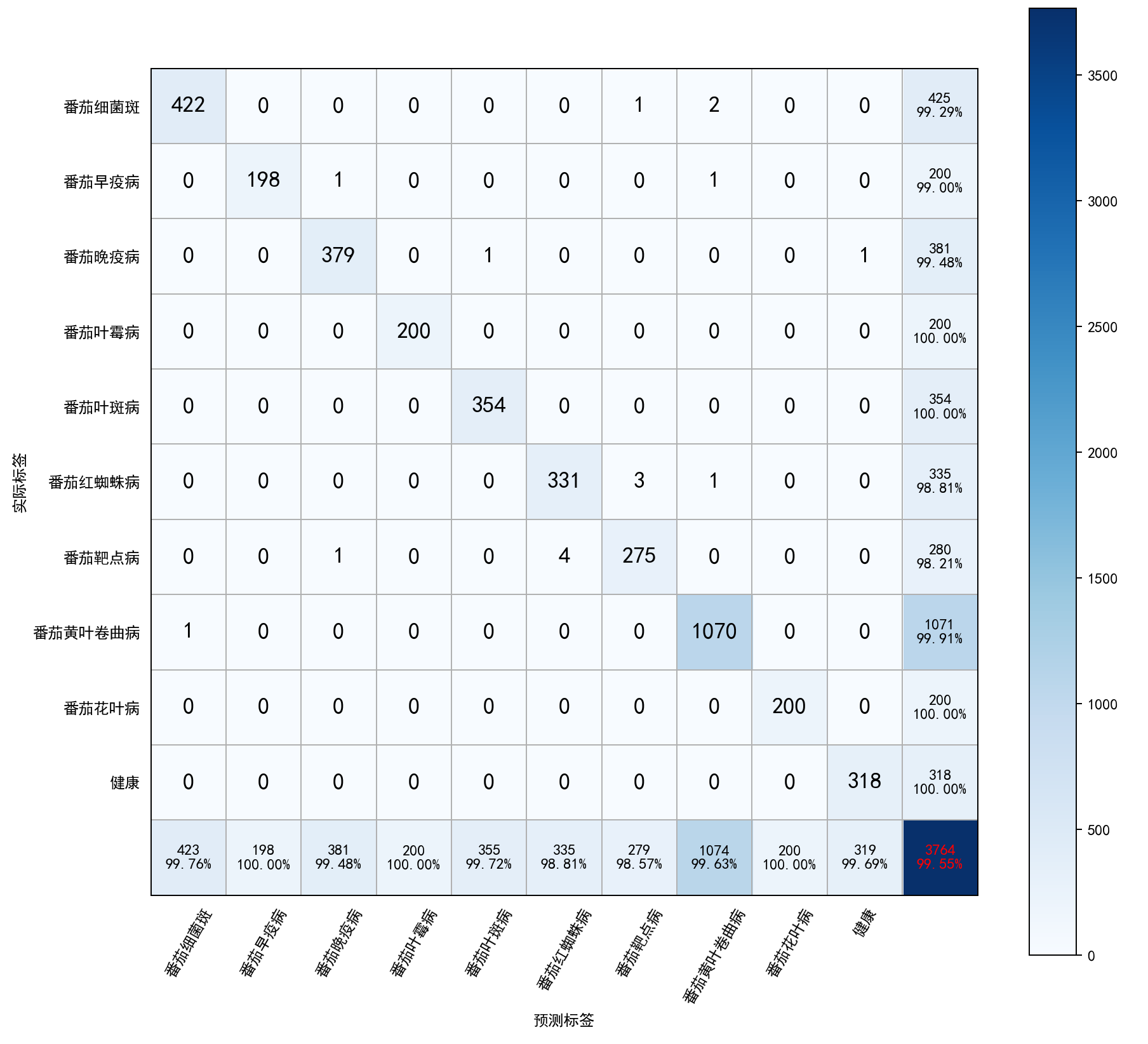


Fig.5

图6混淆矩阵

**4.2** 噪声干扰性能

为了验证本文的EPE\_MMSA-VIT在复杂环境下的抗干扰能力。本文对番茄测试数据集加入椒盐噪声，模糊，运动模糊。分别模拟真实环境下信号可能出现虚假饱和，拍照时未聚焦或者脏镜头以及相机的突然移动，图7展示了上述操作对图像处理后的效果。素值。



在椒盐噪声实验中，使用正态分布在图片上随机生成1%到10%不同比例的噪声。

在模糊实验中，使用卷积核大小1到10对图片进行平滑处理，卷积核越大，处理后的图片约平滑。

在高斯模糊实验中，使用高斯内核对图片中每个像素进行卷积，将卷积和当成输出像

**参 考 文 献**

[1] ZHANG S, WU X, YOU Z, et al. Leaf image based cucumber disease recognition using sparse representation classification [J]. Computers and Electronics in Agriculture, 2017, 134.

[2] RUMPF T, MAHLEIN A-K, STEINER U, et al. Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance [J]. Computers and Electronics in Agriculture, 2010, 74(1).

[3] BASHISH D A, BRAIK M, BANI-AHMAD S. Detection and Classification of Leaf Diseases using K-means-based Segmentation and Neural-networks-based Classification [J]. Information Technology Journal, 2011, 10(2).

[4] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2018 [C].

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.

[6] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv:201011929, 2020.

[7] TOO E C, YUJIAN L, NJUKI S, et al. A comparative study of fine-tuning deep learning models for plant disease identification [J]. Computers and Electronics in Agriculture, 2019, 161.

[8] ÜMIT A, MURAT U, KEMAL A, et al. Plant leaf disease classification using EfficientNet deep learning model [J]. Ecological Informatics, 2021, 61.

[9] SCHULER J P S, ROMANI S, ABDEL-NASSER M, et al. Color-Aware Two-Branch DCNN for Efficient Plant Disease Classification; proceedings of the MENDEL, F, 2022 [C].

[10] DUONG L T, LE N H, TRAN T B, et al. Detection of tuberculosis from chest X-ray images: boosting the performance with vision transformer and transfer learning [J]. Expert Systems with Applications, 2021, 184: 115519.

[11] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention; proceedings of the International Conference on Machine Learning, F, 2021 [C]. PMLR.

[12] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows; proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, F, 2021 [C].

[13] LI Y, ZHANG K, CAO J, et al. Localvit: Bringing locality to vision transformers [J]. arXiv preprint arXiv:210405707, 2021.

[14] LEE S H, LEE S, SONG B C. Vision transformer for small-size datasets [J]. arXiv preprint arXiv:211213492, 2021.

[15] BA J L, KIROS J R, HINTON G E. Layer normalization [J]. arXiv preprint arXiv:160706450, 2016.

[16] KIM Y, DENTON C, HOANG L, et al. Structured attention networks [J]. arXiv preprint arXiv:170200887, 2017.

[17] HUGHES D, SALATHé M. An open access repository of images on plant health to enable the development of mobile disease diagnostics [J]. arXiv preprint arXiv:151108060, 2015.