DOI：10.13718/j.cnki.xdzk.2023.04.011

图像数据集对家蚕品种识别的影响研究

杨创[1]，　石洪康[1]，　陈宇[1]，　马炎[2]，　白娟[3]，　蒋猛[1]

1. 西南大学 工程技术学院，重庆 400715；2. 重庆市渝北区农村合作经济发展服务中心，重庆 401120；
3. 重庆市渝北区经济作物技术推广站，重庆 401120

摘要：针对机器视觉在家蚕品种识别中识别准确率与构成数据集的图像数量、品种数量和数据处理方法、成本等要素之间的矛盾，采集 20 个家蚕品种 4 龄第 3 d 真实生产环境的生长图像构建数据集，利用轻量级卷积神经网络 GhostNet 在不同的训练集上开展模型训练，探讨了图像数量、品种数量、图像数据增强及迁移学习方法对识别准确率的影响. 结果表明：构成数据集的原始图像数量至少 400 张，品种数量选择 10～12 个，均可使识别准确率达到 98%，满足行业标准；如果原始图像数量少于 100 张，通过图像数据增强的方式，对提升识别准确率无实际意义. 品种数量低于 12 个时，采用迁移学习方法，可有效提升识别率；而品种数量多于 14 个时，采用迁移学习方法，会使识别准确率下降，且数量越大下降速度越快.

关　键　词：家蚕；品种识别；数据集
中图分类号：S882.2　　**文献标志码**：A
文 章 编 号：1673－9868(2023)04－0110－09　　开放科学(资源服务)标识码(OSID)：

Study on the Influence of Image Data Set on Identification of Silkworm Cultivar

YANG Chuang[1]，　SHI Hongkang[1]，　CHEN Yu[1]，
MA Yan[2]，　BAI Juan[3]，　JIANG Meng[1]

1. *College of Engineering and Technology，Southwest University，Chongqing 400715，China*；
2. *Chongqing Yubei District Rural Cooperative Economic Development Service Center，Chongqing 401120，China*；
3. *Chongqing Yubei District Cash Crop Technology Extension Station，Chongqing 401120，China*

Abstract：Aiming at the contradiction between the recognition accuracy rate of machine vision and the number of images，the number of varieties，data processing methods，costs and other elements of the data set in the identification of silkworm varieties，this paper collects the growth image of the real production environment of 20 silkworm cultivars on 3rd day of 4th age，and uses the lightweight convolutional neural

收稿日期：2022－04－29
基金项目：国家农业农村部现代农业产业技术体系专项(CARS-18)；重庆市技术创新与应用发展专项(cstc2021jscx-tpyzxX0003).
作者简介：杨创，硕士研究生，主要从事农业信息化研究.
通信作者：蒋猛，副教授，硕士研究生导师.

network GhostNet to carry out model training on different training sets, and discusses the influence of image quantity, variety number, image data enhancement and transfer learning methods on the recognition accuracy. The results show that the dataset constituted with at least 400 of original images, and 10-12 selected varieties, can make the recognition accuracy up to more than 98%. If the number of original images is less than 100, it has no practical significance to improve the recognition accuracy through the way of image data enhancement. When the number of varieties is less than 12, the use of transfer learning methods can effectively improve the recognition rate, when the number of varieties is more than 14, the use of transfer learning methods will reduce the recognition accuracy, and the larger the number, the faster the decline.

Key words: silkworm; species recognition; dataset

家蚕品种资源保存和杂交育种是蚕业科研中的核心内容，在进行资源保存和育种试验前，需同步饲养多个品种. 确保品种纯净度是杂交试验和资源保存准确进行的前提[1-2]. 由于家蚕自身的活动、蚕具交叉使用以及管理疏漏等因素，容易造成不同品种家蚕的混杂，对杂交和资源保存造成不利影响[3-5]；同时家蚕品种多，不同品种之间的差异较小，传统人工识别容易产生混淆. 前期有研究表明使用深度学习识别家蚕品种具有较强的可行性[6]. 深度学习需要大量的图像进行模型训练后才能形成识别能力，但构建大型的家蚕品种数据集却面临耗时长、成本高、采集条件受限等难题，因此有必要探明数据集对家蚕品种识别的影响.

近年来，深度学习在农业视觉领域广泛应用，成为当前的研究热点和主流趋势[7-9]. 在家蚕识别领域，课题组前期使用 MobileNet 对 10 个家蚕品种在 4 龄第 3 d 和 5 龄第 3 d 的生长图像进行了识别研究，研究结果表明，深度学习可高效准确识别家蚕品种，且在 4 龄数据集上的识别准确率最高. 王超[10]使用 SE-GoogLeNet 模型开展了蚕茧品质分选研究，对 3 类蚕茧的识别取得了较佳效果. 于业达等[11]、陶丹等[12-13]使用经典卷积神经网络开展了蚕蛹雌雄鉴别研究，也获得了较高的识别准确率. 石洪康等[14-15]使用 ResNet-50 开展家蚕病害分类识别研究，实现了壮蚕期 5 种常见病害的准确识别；使用 YOLO v3 开展家蚕脓病的检测研究，实现了健康蚕与病蚕混杂的条件下对家蚕脓病的准确检测，为病害精准防治提供了依据.

现有研究表明，深度学习在家蚕识别领域具有广阔的应用前景，但大多基于固定的数据集，而当数据集中的图像数量、品种数量和数据增强方法发生变化时均可能会得到不同的识别结果[16-17]. 为探明数据集对家蚕品种识别的影响，本文采集 20 个家蚕品种 4 龄第 3 d 真实生产环境的生长图像构建数据集，利用轻量级卷积神经网络 GhostNet 在不同的训练集上开展模型训练，探讨图像数量、品种数量和数据增强方法对识别率的影响.

1　材料和方法

1.1　家蚕品种图像数据集

在实际生产环境下采集 20 个家蚕品种在 4 龄第 3 d 的生长图像，每个品种的原始图像数量为 1 100 张. 图像采集时间：2021 年 9 月 10 日，地点：四川省农业科学院蚕业研究所家蚕养殖基地(四川省南充市顺庆区). 图像采集设备用苹果 iPhone 6s 智能手机，环境为室内正常光照. 采集时将设备水平放置俯拍，屏幕长宽比设定为 1∶1，采集的每张图像中仅包含 1 只蚕，并随机使用桑叶为图像背景，结果如图 1.

完成图像采集后，无需对图像进行任何预处理，仅使用双线性插值法将图像尺寸统一缩放为 224×224 像素. 鉴于本文着重关注数据集对家蚕品种识别的影响，因此在不同的试验下，使用的材料仅在图像数量、品种数量以及数据增强方式上有所不同，结果见表 1. 从每个品种的图像中随机挑选 100 张构建验证集，另挑选 200 张构建测试集.

图1 20个家蚕品种图像

表1 试验数据集

序号	试验内容	品种数量/个	训练集/张	验证集/张	测试集/张
1	图像数量对识别的影响	10	每个品种从200张递增至800张，每次递增100张	100	200
2	品种数量对识别的影响	4～20	400	100	200
3	数据增强对识别的影响	10	100张原始图像，使用数据增强生成100～700张图像	100	200
4	迁移学习对识别的影响	4～20	200	100	200

在数据增强对识别的影响中，使用数据增强方法生成新图像，主要包括对原始图像进行随机旋转、平移和局部缩放等. 进行数据增强时的每个品种的原始图像数量为100张，依次生成100～700张图像，数据增强生成新图像如图2.

1.2 GhostNet模型

本文选用GhostNet模型开展家蚕品种识别研究，该模型是Han等[18]在2020年发布的一款轻量级卷积神经网络模型. GhostNet使用一系列的线性变换生成重影特征，在最大程度上避免了特征冗余，能同时兼顾识别效率与准确性，在ImageNet数据集上的识别效果超过了MobileNetv3[19]，因此使用GhostNet模型可以确保较高的识别准确率和识别效率.

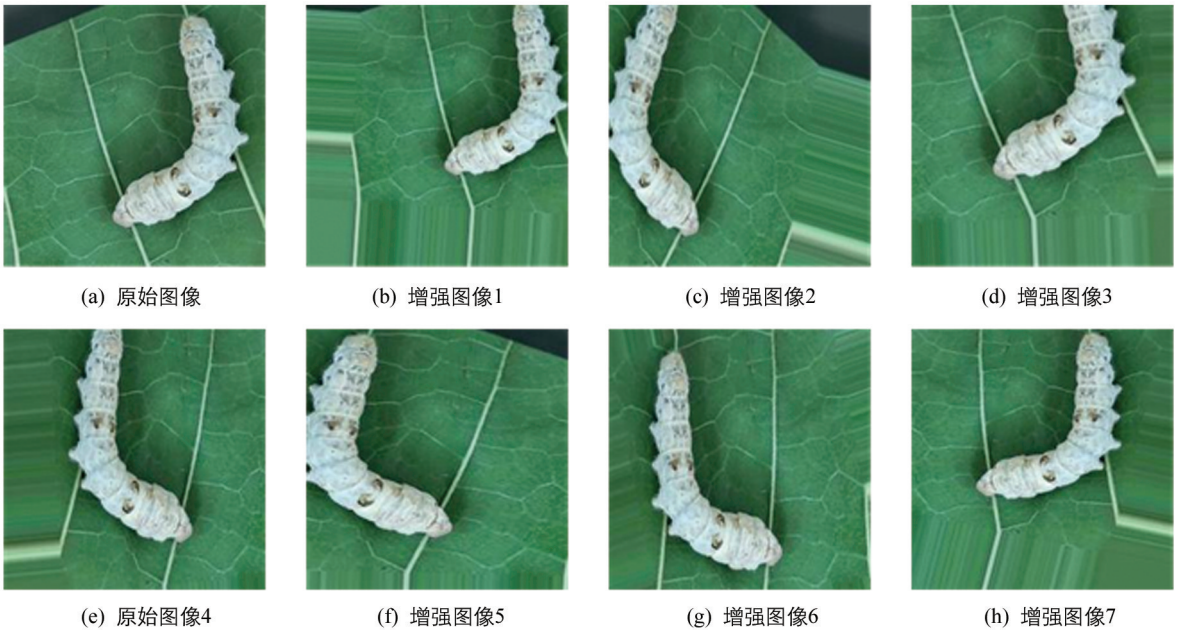


图 2　原始图像与增强图像

GhostNet 中的 Ghost 模块如图 3，对于给定的特征图，先使用 1×1 卷积进行特征通道压缩成原来的一半，然后使用可分离卷积进行特征点(ϕ)的分离提取，并将通道压缩后的特征图与可分离卷积运算的输出特征映射堆叠后输出.

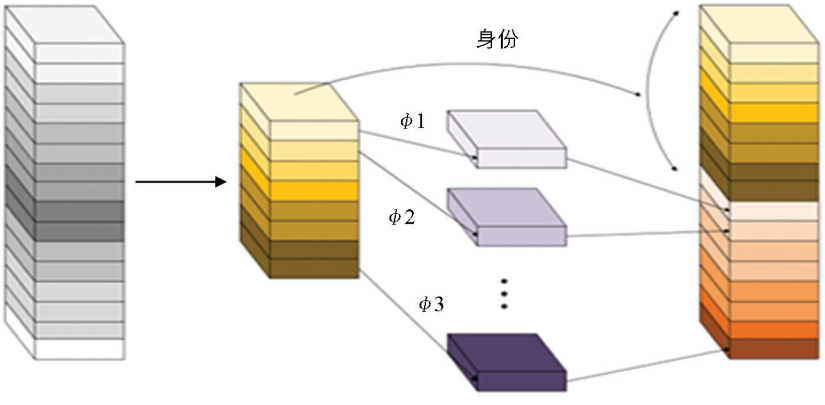


图 3　Ghost 模块结构

两个 Ghost 块可组成 1 个 Ghost bottleneck 块，结构如图 4，其中，“DW Conv Stride＝2”表示卷积步长为 2 的可分离卷积运算，用于压缩特征图维度，“BN”为 Batch Norm 运算，“ReLU”为激活函数，“Add”代表特征图像相加.

GhostNet 的网络结构如表 2，其中，“SE”代表通道注意力模型 SENet[20]，“Conv2d 3×3”代表卷积核尺寸为 3×3 的 2 维卷积，“G-bneck”代表 Ghost bottleneck 结构，“AvgPool 7×7”代表尺寸

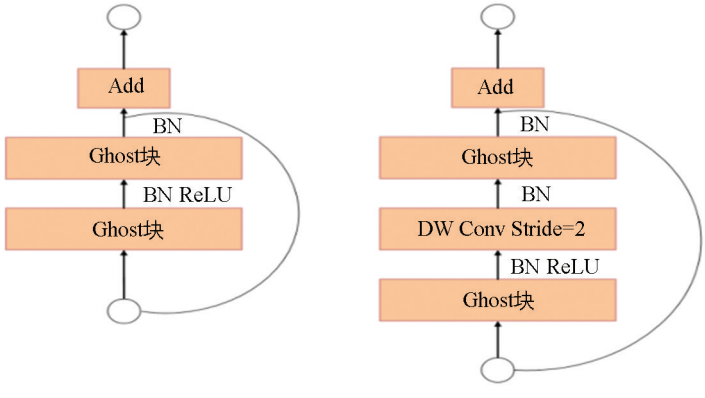


图 4　Ghost bottleneck 结构

为7×7的全局平均池化，“FC”代表全连接. 使用GhostNet进行家蚕品种识别时，输入图像尺寸为224×224像素，先使用16个尺寸为3×3的卷积核按步长为2进行卷积运算，提取图像特征，再使用一系列堆叠的Ghost bottleneck结构，每个Ghost bottleneck中会进行维度膨胀，部分结构中还使用了SE注意力模型，最后使用1×1的卷积运算进行特征通道调整，全局平均池化和全连接运算后输出网络预测结果.

表2 GhostNet的网络结构

输入	运算	膨胀维度	输出通道数	SE	卷积步长
$224^2\times3$	Conv2d 3×3	—	16	—	2
$112^2\times16$	G-bneck	16	16	—	1
$112^2\times16$	G-bneck	48	24	—	2
$56^2\times24$	G-bneck	72	24	—	1
$56^2\times24$	G-bneck	72	40	1	2
$28^2\times40$	G-bneck	120	40	1	1
$28^2\times40$	G-bneck	240	80	—	2
$14^2\times80$	G-bneck	200	80	—	1
$14^2\times80$	G-bneck	184	80	—	1
$14^2\times80$	G-bneck	184	80	—	1
$14^2\times80$	G-bneck	480	112	1	1
$14^2\times112$	G-bneck	672	112	1	1
$14^2\times112$	G-bneck	672	160	1	2
$7^2\times160$	G-bneck	960	160	—	1
$7^2\times160$	G-bneck	960	160	1	1
$7^2\times160$	G-bneck	960	160	—	1
$7^2\times160$	G-bneck	960	160	1	1
$7^2\times160$	Conv2d 1×1	—	960	—	1
$7^2\times960$	AvgPool 7×7	—	160	—	—
$1^2\times960$	Conv2d 1×1	—	1 280	—	1
$1^2\times1\ 280$	FC	—	1 000	—	—

1.3 试验环境与评价指标

试验用硬件设备为DELL Precision 5820图像工作站，处理器：Core i7-9800X，显卡：RTX 2080Ti，显存：11G，内存：32G，运算平台：CUDA-10.0，操作系统：Windows 10专业版64位，编程语言：Python 3.7，开发环境：Jupyter notebook，深度学习框架：TensorFlow-gpu 1.14和Keras 2.0.

模型训练的超参数：mini batch_size为16，迭代次数为300次，损失函数为交叉熵，优化器为Adam，初始学习率为0.001，当损失值连续5次迭代未明显下降时，就将学习率乘以0.8. 模型训练完成后使用测试集进行测试，以模型在测试集上的平均识别准确率为评价指标，计算公式为

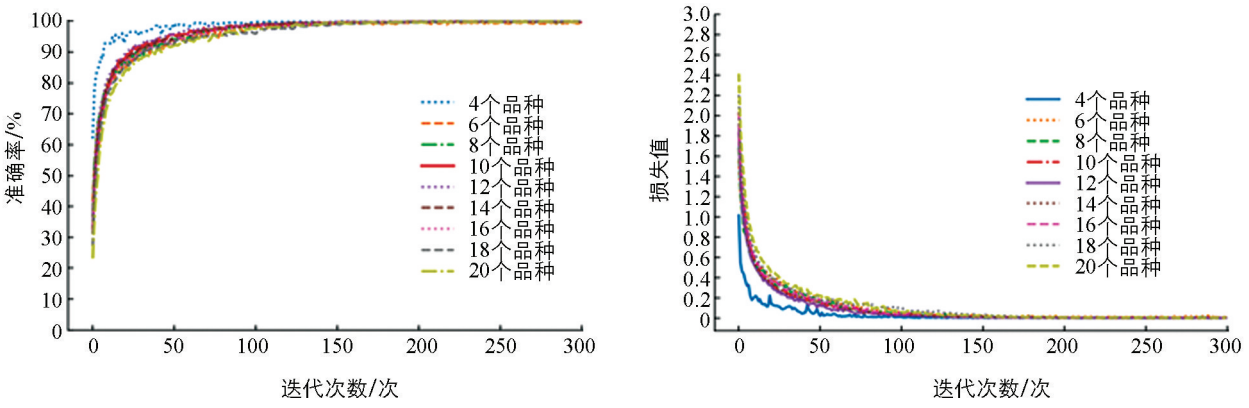
$$\eta=\frac{N_{\text{correct}}}{N_{\text{total}}}\times100\%$$

式中，η为平均识别准确率，N_{correct}为正确识别数量，N_{total}为测试集图像总数量.

2 结果与分析

2.1 图像数量对识别的影响

试验时，每个品种用于模型训练初始图像数量为200张，并逐次增加100张至800张. 训练过程中，以模型在训练集上的准确率和损失值作为参照，用于查看模型的收敛效果如图5.



(a) 准确率训练模型　　(b) 损失值训练模型

图 5　不同图像数量训练集的准确率和损失值曲线

从图 5 中的准确率和损失值曲线可知，在固定品种数量条件下，当训练集中图像数量发生变化时，模型的收敛速度存在一些差异，且整体呈现出图像数量越多，模型的收敛速度越快，收敛效果也越佳；在进行 130 次训练迭代时，不同图像数量模型均趋于或达到稳定状态，据此确定本文后续测试使用 300 次迭代能确保模型完全收敛.

模型训练完毕后，使用测试集对其进行测试，由图 6 可知：随着用于模型训练的图像数量增加，识别准确率也相应增加. 当用于模型训练的图片数量少于 300 张(含 300 张)时，其识别准确率最高为 96.35%，低于行业标准(≥97%)要求，只有图片数量大于 400 张时的识别准确率超过 98%，高于行业标准规定. 用于模型训练的图片数量由 500 张增加至 600 张时，其识别准确率提高了 0.75%；图片数量由 600 张增加至 700 张时，其识别准确率只提高了 0.30%；图片数量由 700 张增加至 800 张时，其识别准确率仅提高了 0.10%. 由此可见，用于模型训练的图片数量不能低于 300 张，多于 700 张意义也不大. 因此，鉴于成本、识别速度及符合行业标准条件下，建议数据集图像数量在 400 张即可，若考虑高识别率时，数据集图像数量在 700 张为好.

2.2　图像数据增强对识别准确率的影响

在数据集图像数量较少时，使用数据增强生成新图像是一种常用的方法，为验证图像数据增强方法对识别准确率的影响，使用 10 个家蚕品种，每个品种原始图像数量为 100 张，并通过数据增强依次生成 200～800 张图像，再利用生成的新图像和原图像构建训练集用于模型训练与测试(与 2.1 相同)，测试结果如图 6.

图 6 表明，当使用数据增强方法生成更多图像用于模型训练时，随着图像数量的增加，模型在测试集上的识别准确率由 200 张原图的 94.60%提高到 800 张增强图的 96.85%，图像数量增加了 600 张，但识别准确率仅仅提高了 2.25%，还是没能达到行业标准(≥97%).

从图 6 中还可以看出，采用图像数据增强方法增加图像数量，其整体识别准确率及增加幅度均相对较低，表明使用数据增强方式生成新图像对于识别准确率的提升作用有限，且不能达到实际应用的要求.

在原始图像数量不够的条件下，纵然通过图像数据增强的方式增多图像数量，其识别率的提高也是相当有限的，因此，应尽可能增加原始图像数量以获得最佳的识别效果.

2.3　品种数量对识别的影响

根据 2.1 的试验结果，固定单个品种训练图像数量为 400 张，测试的初始品种数量为 4 个，并逐次增加 2 个直至 20 个，训练过程中记录模型在训练集上的准确率和损失值(图 7). 从图 7 可以看出，当单个品种图像数量固定、品种数量不同时，模型的收敛效果存在一定差异，收敛效果最佳的是使用 4 个品种的训

练模型，最差的是20个品种的训练模型. 所有品种尽管有一定差异，但大致经过150次迭代后，不同品种数量的模型均能达到稳定状态.

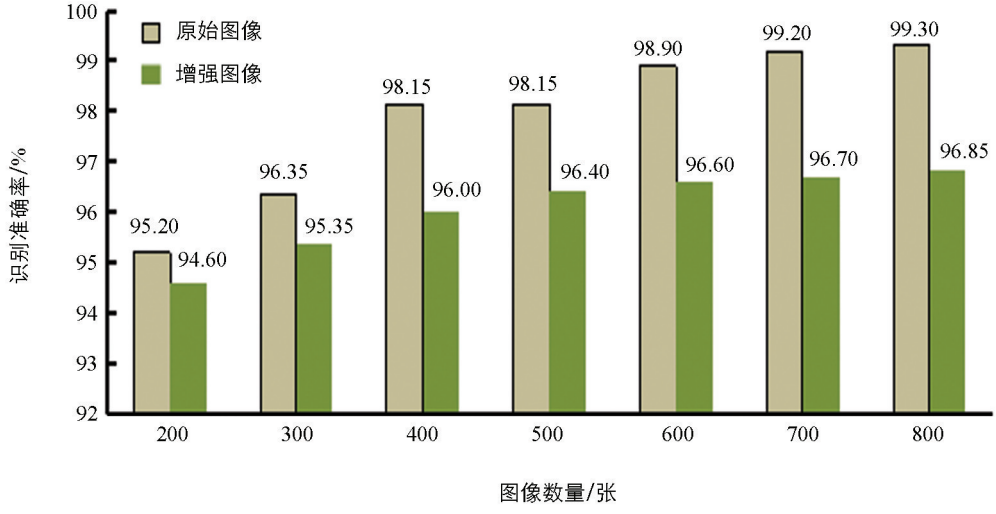
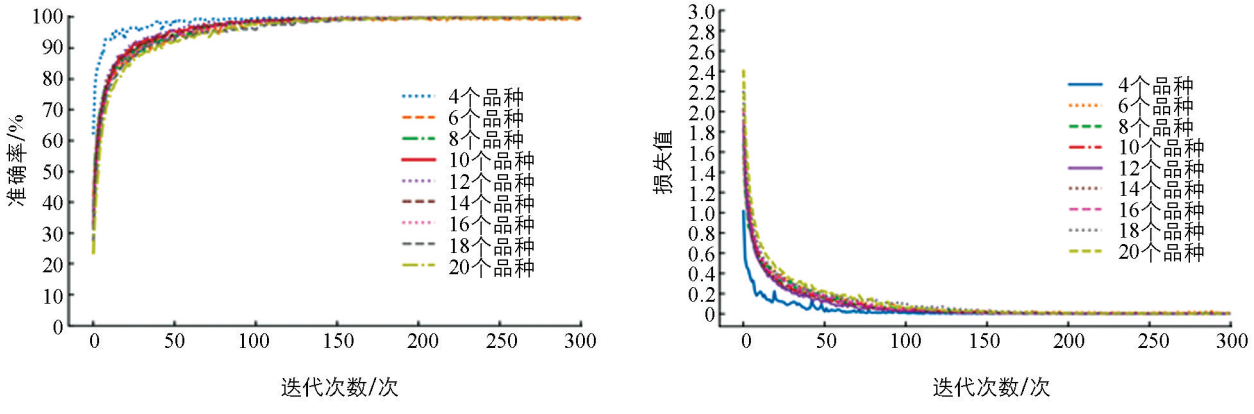


图6 二种不同条件下的图像数量识别准确率的直观对比



(a) 准确率训练模型 (b) 损失值训练模型

图7 不同品种数量训练集的准确率和损失值曲线

图8是不同品种数量模型在测试集上的识别准确率，从图8中可以看出，采用无迁移学习方法，其识别准确率最高的是12个品种，为98.50%；其次是4个品种，识别准确率为98.00%；最低的是6个品种，识别准确率仅为94.67%. 当品种数量从12个逐步增加至20个时，识别准确率由98.50%逐步下降至96.67%，由此，建议在品种识别数据集中品种数量选择12个为最佳.

2.4 迁移学习对识别的影响

在数据及图像数量较少时，迁移学习也是一种常用的方法. 为验证迁移学习对家蚕品种识别的影响，使用GhostNet在ImageNet数据集上预训练的权重进行迁移学习，取每个品种的原始图像400张，初始品种数量为4个，每次增加2个直至20个品种，即只增加品种数量，而每个品种的图片数量保持不变，训练完成后在测试集上进行测试，结果如图8.

由图8可以看出，使用迁移学习方法，品种数量的不同对识别准确率也有明显影响，即在品种数量由6个增至10个时，识别准确率为97.92%～98.90%，基本能达到实际生产的要求；当品种数量由10个增至20个时，随品种数量的增多，识别准确率由98.90%逐步下降至94.43%，且下降速度较快. 结果表明，在品种数量为10个时，识别准确率最高，为98.90%；其次是12个，为98.75%.

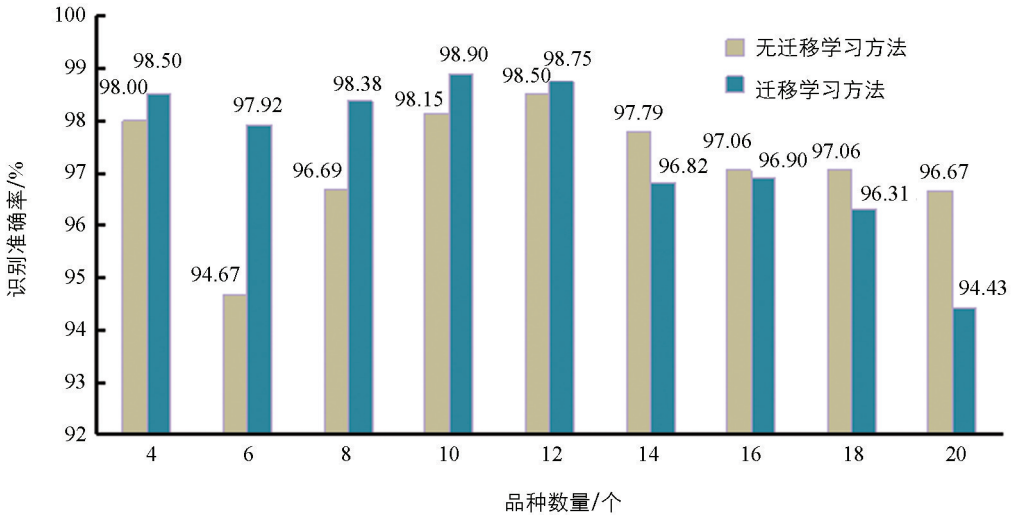


图 8　是否使用迁移学习方法的不同品种数量识别准确率的直观对比

从图 8 还可以看出，采用迁移学习方法后，在品种数量为 4 个，10 个和 12 个时，其识别准确率提升小于 1%；在品种数量为 6 个至 8 个时，识别准确率提升较大，并均能使识别准确率达到 97%以上；当品种数量大于 12 个后，识别准确率不升反降，随数量增多，下降更快.

由此可见，当品种数量低于 12 个时，采用迁移学习方法可有效提升识别准确率，并取得较好的效果；在品种数量大于 12 个时，迁移学习方法对识别准确率的提升无效.

3　结论

针对当前我国家蚕品种数量多，开展基于深度学习的家蚕品种识别研究时，数据采集及数据集构建面临耗时长、成本高、采集条件受限等诸多问题，本文开展了数据集对家蚕品种识别的影响. 在实际生产环境条件下，采集了 20 个家蚕品种在 4 龄第 3 d 的真实生长图像，构建出家蚕品种图像数据集，采用 GhostNet 的家蚕品种识别模型，分别开展了图像数量、品种数量、图像数据增强和迁移学习方法对品种识别准确率的影响研究，结果表明：

1）增加构成数据集的图像数量有助于提升识别准确率. 当单个品种训练集图像数量为 400 张时，识别准确率可达 98.15%，达到行业标准要求；当图像数量在 800 张时，识别准确率高达 99.30%. 图像数量增加纵然会提高识别准确率，但也会大大增加成本、降低运行效率、提升硬件要求等. 综合各项因素，在构成数据集时，单个品种的图像数量选取 400 张即可.

2）品种数量会对识别准确率造成一定的影响. 品种数量过多或过少，都会使识别准确率降低，不能满足行业标准要求，只有当品种数量在 10～12 个时，识别准确率超过 98%，能够满足行业标准要求，因此，建议在构成数据集中的品种数量选择在 10～12 个.

3）构成数据集的原始图像数量低于 100 张，采用图像数据增强的方法对识别准确率的提升作用非常有限且无实际意义，因此，建议在构建数据集时尽可能增加原始图像数量.

4）当品种数量低于 12 个时，采用迁移学习方法，可有效提升识别准确率，并取得较好的效果，品种数量越少，表现越好. 当品种数量大于 14 个时，迁移学习方法反而会使识别准确率下降，品种数量越多，下降的速度越快.

参考文献：

[1]　徐安英，钱荷英，孙平江，等. 家蚕抗血液型脓病新品种华康 3 号的育成 [J]. 蚕业科学，2019，45(2)：201-211.

[2]　张友洪，肖金树，肖文福，等. 春用多丝量家蚕品种金・兰×铭・晖的育成 [J]. 蚕业科学，2019，45(1)：144-148.

[3] 陈惠蓉，杨忠生，李俊. 浅析桑蚕种质资源的保存与利用 [J]. 四川蚕业，2016，44(2)：42-43.
[4] 肖阳，李庆荣，邢东旭，等. 抗 BmNPV 家蚕新品种粤蚕 11 号的育成 [J]. 广东农业科学，2020，47(8)：118-126.
[5] 何锐敏，郑可锋，张俊，等. 工厂化养蚕精准饲喂信息系统的研究与开发 [J]. 浙江农业科学，2022，63(2)：371-374，380.
[6] 石洪康，田涯涯，杨创，等. 基于卷积神经网络的家蚕幼虫品种智能识别研究 [J]. 西南大学学报(自然科学版)，2020，42(12)：34-45.
[7] 孙红，李松，李民赞，等. 农业信息成像感知与深度学习应用研究进展 [J]. 农业机械学报，2020，51(5)：1-17.
[8] 王鹏新，田惠仁，张悦，等. 基于深度学习的作物长势监测和产量估测研究进展 [J]. 农业机械学报，2022，53(2)：1-14.
[9] KAMILARIS A，PRENAFETA-BOLDÚ F X. Deep Learning in Agriculture：a Survey [J]. Computers and Electronics in Agriculture，2018，147：70-90.
[10] 王超. 基于机器视觉的蚕茧图像识别研究 [D]. 柳州：广西科技大学，2019.
[11] 于业达，高鹏飞，赵一舟，等. 基于深度卷积神经网络的蚕蛹雌雄自动识别 [J]. 蚕业科学，2020，46(2)：197-203.
[12] 陶丹. 基于机器视觉的家蚕蛹雌雄识别研究 [D]. 重庆：西南大学，2019.
[13] 陶丹，王峥荣，李光林，等. 基于解模糊算法的蚕蛹图像恢复及雌雄识别 [J]. 农业工程学报，2016，32(16)：168-174.
[14] 石洪康，肖文福，黄亮，等. 基于卷积神经网络的家蚕病害识别研究 [J]. 中国农机化学报，2022，43(1)：150-157.
[15] 石洪康. 基于卷积神经网络的家蚕脓病检测研究与预警软件开发 [D]. 重庆：西南大学，2021.
[16] 樊湘鹏，周建平，许燕，等. 数据集对基于深度学习的作物病害识别有效性影响 [J]. 中国农机化学报，2021，42(1)：192-200.
[17] ARNAL B. Impact of Dataset Size and Variety on the Effectiveness of Deep Learning and Transfer Learning for Plant Disease Classification [J]. Computers and Electronics in Agriculture，2018，153：46-53.
[18] HAN K，WANG YH，TIAN Q，et al. GhostNet：More Features from Cheap Operations [C] //2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle，WA，USA. IEEE，2020：1577-1586.
[19] HOWARD A，SANDLER M，CHEN B，et al. Searching for MobileNetV3 [C] //2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul，Korea (South). IEEE，2020：1314-1324.
[20] HU J，SHEN L，ALBANIE S，et al. Squeeze-and-Excitation Networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2020，42(8)：2011-2023.

责任编辑 周仁惠