北京林业大学学报
JOURNAL OF BEIJING FORESTRY UNIVERSITY

# 基于YOLOv7的红外相机野生动物图像筛选

齐建东 马添 郑尚姿

# Wildlife image screening for infrared cameras based on YOLOv7

Qi Jiandong, Ma Zhongtian, Zheng Shangzi

引用本文:
齐建东, 马添, 郑尚姿. 基于YOLOv7的红外相机野生动物图像筛选[J]. 北京林业大学学报, 2024, 46(2):143-154. doi: 10.12171/j.1000-1522.20230112
Qi Jiandong, Ma Zhongtian, Zheng Shangzi. Wildlife image screening for infrared cameras based on YOLOv7[J]. Journal of Beijing Forestry University, 2024, 46(2):143-154. doi: 10.12171/j.1000-1522.20230112

在线阅读 View online: https://doi.org/10.12171/j.1000-1522.20230112

---

**您可能感兴趣的其他文章**

**Articles you may be interested in**

基于ROI-CNN的赛罕乌拉国家级自然保护区陆生野生动物自动识别

Automatic identification method for terrestrial wildlife in Saihanwula National Nature Reserve in Inner Mongolia of northern China based on ROI-CNN

北京林业大学学报, 2018, 40(8): 123-131. https://doi.org/10.13332/j.1000-1522.20180141

基于深度学习的5种树皮纹理图像识别研究

Research on image recognition of five bark texture images based on deep learning

北京林业大学学报, 2019, 41(4): 146-154. https://doi.org/10.13332/j.1000-1522.20180242

基于多特征融合和CNN模型的树种图像识别研究

Image recognition of tree species based on multi feature fusion and CNN model

北京林业大学学报, 2019, 41(11): 76-86. https://doi.org/10.13332/j.1000-1522.20180366

一种融合聚类和分类算法的树木图像多目标优化分割方法

A multi-objective optimization segmentation method for tree image based on fusion clustering and classification algorithm

北京林业大学学报, 2018, 40(12): 124-131. https://doi.org/10.13332/j.1000-1522.20180160

基于地形匹配的图像烟火定位方法研究

Smoke and fire positioning method in the image based on terrain profile matching

北京林业大学学报, 2018, 40(6): 19-29. https://doi.org/10.13332/j.1000-1522.20170439

基于可见光-近红外图像的幼龄檀香全磷含量诊断

Diagnosis of total phosphorus content in young sandalwood based on visible light and near infrared images

北京林业大学学报, 2019, 41(2): 88-96. https://doi.org/10.13332/j.1000-1522.20180214

DOI:10.12171/j.1000−1522.20230112

# 基于 YOLOv7 的红外相机野生动物图像筛选

齐建东[1,2] 马鐘添[1] 郑尚姿[1]
（1. 北京林业大学信息学院，北京 100083；2. 国家林业和草原局林业智能信息处理工程技术研究中心，北京 100083）

**摘要:【目的】**野外环境通常植被繁茂、树木杂乱，且受环境、天气、光照等因素影响，红外相机易误触发拍摄，从而捕获大量废片，需要耗费大量人力进行筛选。为解决此类问题，本研究以 YOLOv7 模型为基础，对其进行轻量化改进，以实现对废片的自动筛选。**【方法】**本研究构建了北京密云雾灵山自然保护区 2014—2015 年期间采集到的 2 172 张野生动物图像数据集，并对图像中出现的动物进行位置标记。对 YOLOv7 网络使用不同方式进行改进：引入 MicroBlock 替换 YOLOv7 的主干网络，使用轻量化 SPPCSPC 结构降低模型参数量。采用 SIoU 损失、LNDown 下采样、BiFPN 提升模型检测动物的能力。使用 YOLOv5-m、YOLOv5-l、Ghost-YOLOv5-l、YOLOv6、YOLOX-M、YOLOR-CSP 模型，在含有 1 万张图像的 Snapshot Serengeti 相机陷阱图像子数据集上进行训练和验证，对比本文模型对野生动物图像的筛选效果。利用迁移学习训练自建野生动物数据集，测试冻结不同层数的训练效果。**【结果】**基于 YOLOv7 的改进模型推理时间降低了 14.3%，每秒浮点运算次数 FLOPS 降低了 33.5%，参数量减少了 17.8%，误检测方面也优于 YOLOv7 模型。与其他模型进行对比，改进后的 YOLOv7 虽未在所有指标中均达到最优，但在检测时间与精度上达到了更好的平衡。在自建数据集中使用未冻结权重方式微调效果最优，平均精度比未使用迁移学习模型提高了 12.6%。**【结论】**本研究为密云地区野生动物监测网络提供了更快速、准确的筛选方案。

**关键词：**野生动物图像；图像筛选；深度学习；目标检测
**中图分类号：**TP391.4；S862 **文献标志码：**A **文章编号：**1000−1522(2024)02−0143−12
**引文格式：**齐建东，马鐘添，郑尚姿. 基于 YOLOv7 的红外相机野生动物图像筛选 [J]. 北京林业大学学报，2024，46(2)：143−154. Qi Jiandong, Ma Zhongtian, Zheng Shangzi. Wildlife image screening for infrared cameras based on YOLOv7[J]. Journal of Beijing Forestry University, 2024, 46(2): 143−154.

## Wildlife image screening for infrared cameras based on YOLOv7

Qi Jiandong[1,2] Ma Zhongtian[1] Zheng Shangzi[1]
(1. School of Information, Beijing Forestry University, Beijing 100083, China;
2. Engineering Research Center for Forestry-Oriented Intelligent Information Processing of National Forestry and Grassland Administration, Beijing 100083, China)

**Abstract:** [**Objective**] Due to the lush vegetation and disorderly trees in the wild environment, as well as the influence of factors such as environment, weather, and lighting, infrared cameras are prone to triggering shooting errors, resulting in the capture of a large amount of waste film, which requires a lot of manpower for screening. To solve such problems, based on the YOLOv7 model, this paper has made lightweight improvements to achieve automatic screening of waste pieces. [**Method**] This study constructed a dataset of 2 172 wildlife images collected from the Beijing Wuling Mountain Nature Reserve in the period of 2014−2015, and marked the positions of animals in the images. YOLOv7 network was improved in different ways. MicroBlock was introduced to replace the backbone network of YOLOv7, and the SPPCSPC structure was light-weighted to reduce the model parameters. SIoU loss, LNDown downsampling, and BiFPN were used to improve the model's ability to detect animals. YOLOv5-m, YOLOv5-l, Ghost-YOLOv5-l, YOLOv6,

**收稿日期:** 2023−05−12 **修回日期:** 2023−12−26
**基金项目:** 国家重点研发计划(2020YFA0608100)。
**第一作者:** 齐建东，博士，教授。主要研究方向：生态信息学、智能信息处理。Email: qijd@bjfu.edu.cn 地址：100083 北京市海淀区清华东路 35 号北京林业大学信息学院。
**本刊网址:** http://j.bjfu.edu.cn; http://journal.bjfu.edu.cn

YOLOX-M, and YOLOR-CSP models were trained on an Snapshot Serengeti camera trap subset dataset containing 10 000 images, and the screening effects of the model on wildlife images were compared. Transfer learning was used to train a self-built wildlife dataset, and the training effects of freezing different layers was tested. [**Result**] The improved model based on YOLOv7 reduced inference time by 14.3%, floating-point operations per second by 33.5%, and parameters by 17.8% compared with the YOLOv7 network. The error detection of the improved YOLOv7 model was also better than that of YOLOv7. Although the improved YOLOv7 did not achieve the best performance in all indicators compared with other models, it achieved a better balance between detection time and accuracy. In the self-built dataset, the unfrozen weight method had the best effect, and average precision was 12.6% higher than that of the model without transfer learning. [**Conclusion**] This study provides a faster and more accurate screening solution for wildlife monitoring networks in the Mountain area of Beijing Miyun.

**Key words:** wildlife image; image filtering; deep learning; object detection

野生动物作为物种多样性中的重要一环，在生物多样性研究中占据重要位置，所以监测野生动物数据对研究生物多样性有重要意义。但传统的野生动物监测需要大量人力，对专业性要求较高，野生动物数据采集困难，难以适应当前全球生物多样性监测保护的需要[1]。通过动作传感器、红外探测器等方式作为触发拍摄的相机，称为“相机陷阱”。相机陷阱设备能在野外进行自动监视和拍摄，可隐蔽地捕获到经过的野生动物图像。红外相机作为常用的相机陷阱设备类型之一，具有对动物干扰小、全天候拍摄等优势，广泛应用于国内外野生动物监测与生物多样性保护。

由于红外相机易受很多其他因素影响而误触发拍摄，产生大量无动物出现的废片，每台红外相机年均照片数量往往数以万计[2]，需要大量存储空间。同时，拍摄的图像仅部分动物位置可见[3]、光照不足[4]等问题，给后续数据处理和分析带来挑战。计算机视觉领域中的目标检测技术可对图像或视频中目标物体进行自动识别，有效解决此难点问题。目标检测的研究脉络可划分为两个阶段：传统目标检测算法和基于深度学习的目标检测算法。传统目标检测算法一般步骤如下：首先在给定图像上选择一些候选区域进行特征提取，之后使用分类器进行分类。但由于场景复杂、目标姿态多变等，效果欠佳。当前，目标检测算法主要基于深度学习提取特征，提高了目标检测精度，加快了目标检测速度。YOLO 系列、SSD、RetinaNet 等都是经典的一阶段目标检测算法。R-CNN 系列是具有代表性的二阶段目标检测算法，包括 R-CNN 及其改进版本 fast R-CNN 和 faster R-CNN 算法。

目标检测模型已应用在野生动物图像研究中。2018 年，Schneider 等[5]使用 YOLOv2 和 faster R-CNN 两种目标检测算法对 Reconyx Camera Trap 数据集和 Snapshot Serengeti 数据集的野生动物图像进行检测和识别，最终证明 faster R-CNN 表现更优，在两个数据集上的平均准确率分别为 93.0% 和 76.7%。Guo 等[6]基于深度卷积神经网络和野生动物图像中动物特征，开发了一种可变信道区域建议和分类网络 VCRPCN，通过将不同信道反馈到网络的不同组件以实现不同目的，该架构考虑了动物外观对图像识别的影响，并识别了图像中潜在的动物区域，提取它们的局部特征进行描述和分类。Chen 等[7]开发了一种全自动网络，对大熊猫（*Ailuropoda melanoleuca*）面部进行检测、分割、对齐和预测，同时构建一个由 218 只大熊猫拍摄的 6 441 张图像数据集，经测试该模型在大熊猫身份检测上达到了 96.27% 的准确率，可有效对大熊猫个体进行识别。Tan 等[8]由于现有的数据集对动物的行为标记具有局限性，所以从东北虎豹国家公园收集野生动物图像构建了由 17 个类别 25 657 张图像组成的东北虎豹国家公园数据集，并将数据分为了白天和夜间类别。使用 YOLOv5-m、Cascade R-CNN[9]和 Fully Convolutional One-Stage[10]模型进行预测，YOLOv5-m 最高达到了 98.9% 的平均精度均值，并在各项检测中都达到了最高精度。Vecvanags 等[11]在拉脱维亚采集了 4 种野生有蹄类动物数据集，选择野猪（*Sus scrofa*）和鹿（*Elaphurus davidianus*）两个类别共 1 128 张图像，使用迁移学习和数据预处理技术在 fast-RCNN 和 RetinaNet 上进行研究，fast-RCNN 和 RetinaNet 的全部类别平均精度分别达到了 0.407 3 和 0.436 4。Roy 等[12]对 YOLOv4 进行改进，提出了 WilDect-YOLO 模型对濒危野生动物进行检测，在主干网络引入 DenseNet 模块[13]并改进空间池化金字塔（spatial pyramid pooling，SPP）和路径聚合网络（path aggregation net，PANet）结构，增强模型的感受野和特征融合能力，在自行构建的野生动物图像数据集上平均精度均值达到了 96.89%。

现有研究中目标检测算法在野生动物识别方面已取得了一定的成果，但仍存在以下问题：其一，现有的公开野生动物数据集中具有动物位置标注的数据较少，且大多源于国外，数据集中包含的动物类别与国内野生动物物种具有差异；其二，多数研究中应用目标检测算法用于野生动物物种识别，在空图像筛选方面应用较少；其三，现有研究主要注重检测精度，实际应用中检测速度同样需要注意。基于上述问题，本研究构建了具有动物位置标注的北京市密云区的北京雾灵山自然保护区的野生动物图像数据集。构造了基于 YOLOv7 模型为基础的改进网络，使用 MicroBlock[14]替换主干网络，使用轻量化 SPPCSPC 结构降低模型参数量。使用 SIoU 损失、LNDown、BiFPN[15]降低模型对动物的错分和漏检误差。改进后的 YOLOv7 模型可以对红外相机拍摄的空图像进行更快速、及时、准确地筛选，筛选后的包含野生动物的图像可方便研究人员进行研究。

## 1　数据源

### 1.1　自建数据集

数据是由部署在北京市密云区的北京市级雾灵山自然保护区的 30 个红外相机于 2014—2015 年拍摄的。在相关专家协助下，将图片按照动物类别分为 8 个类别，共计 2 172 张，图像分辨率均为 2 560 × 1 920 像素。图 1 为红外相机拍摄到的各类别野生动物图片。




图 1　各类别野生动物图像

Fig. 1　Wildlife images of each category

数据采集所使用的相机型号均为猎科 Ltl-6511，其通过人体热释电红外传感器感应温度变化对路过的野生动物进行实时抓拍。白天拍摄全彩图像，夜间拍摄黑白图像。相机内置 4G 流量卡，抓拍后的图像可自动上传至 FTP 服务器，无信号区域定期通过人工收集方式采集相机内存卡存储的照片。相机部署在树干距地面 1 m 左右位置（图 2），使用太阳能电池板提升相机续航能力。



图 2　Ltl-6 511 红外相机

Fig. 2　Ltl-6511 infrared camera

除了类别标注，还对采集到的数据集进行位置标注，制作为野生动物筛选模型数据集。使用标注工具对野生动物所在位置进行标注（图 3），记录格式为左上角坐标（$x_{min}$, $y_{min}$）和右下角坐标（$x_{max}$, $y_{max}$），所有图像均标注为“动物”类别。将数据集按 9∶1 划分为训练集和验证集。选择额外 200 张相同区域有动物出现的图像以及 200 张空图像作为筛选模型的测试集，用于测试野生动物图像筛选效果。

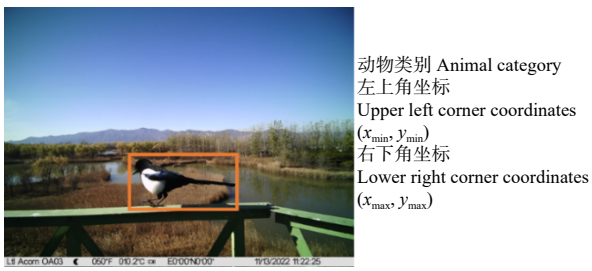



图 3　野生动物位置标注图像

Fig. 3　Location annotated image of wild animals

### 1.2　SS 相机陷阱图像数据集

Snapshot Serengeti（SS）相机陷阱图像数据集是迄今为止最大的公共相机陷阱图像集合。数据由部署在塞伦盖蒂国家公园的红外相机于 2010—2017

年拍摄得到的，经许多志愿者共同标记制成。SS 相机陷阱图像数据集约有几百万张图像，其中约 90 万张图像中包含野生动物目标，其余均为相机误触发而拍下的空图像。由于其动物图像数量很大，本研究从 SS 数据集中抽样选择 10 000 张图像作为模型的训练和测试数据，训练集和验证集按 9∶1 划分。

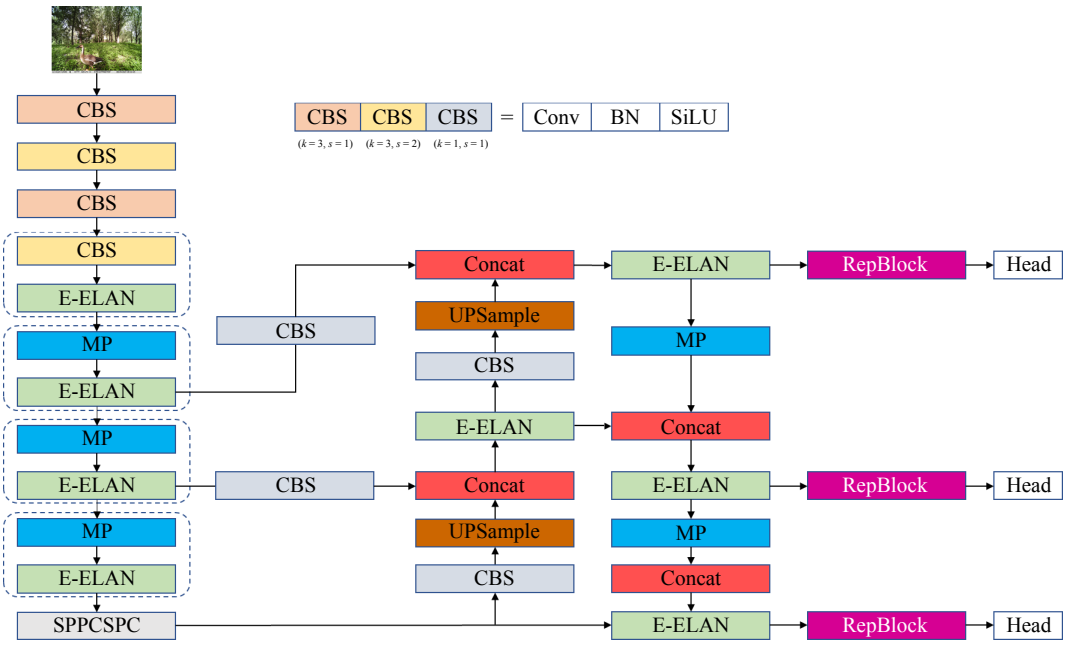
## 2 模型与改进方案

### 2.1 YOLOv7 网络概述

目标检测算法可以对图像中的物体进行快速准确的检测，YOLO 系列一直是目标检测网络中的经典算法，具有检测速度快的特点。2022 年发布的 YOLOv7 与之前的 YOLO 系列网络相比，不仅检测速度更快，检测精度也有提升。YOLOv7 网络使用全新的主干网络扩展高效远距离注意力网络（extended efficient long-range attention network，E-ELAN）结构、复合模型缩放方法、重参数化卷积等。图 4 为 YOLOv7 网络结构图。

YOLOv7 对主干网络进行改进，使用改进的 E-ELAN 结构。E-ELAN 结构是一种多分支级联模块结构，通过控制梯度的最长路径长度，让网络更快速收敛。YOLOv7 中实现的 E-ELAN 结构如图 5 所示。

图 6 是 YOLOv7 网络中使用的改进 SPPCSPC 结构[12]，其参考跨阶段部分网络（cross stage paritial




*k*. 卷积核大小 Kernel size; *s*. 步长 Stride. 下同。The same below.

图 4 YOLOv7 结构图

Fig. 4 YOLOv7 structure diagram

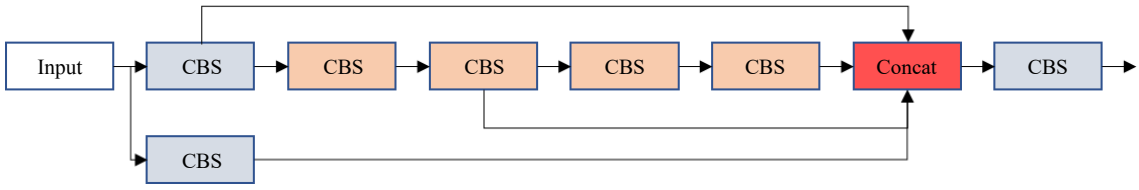



图 5 E-ELAN 结构图

Fig. 5 E-ELAN structure diagram

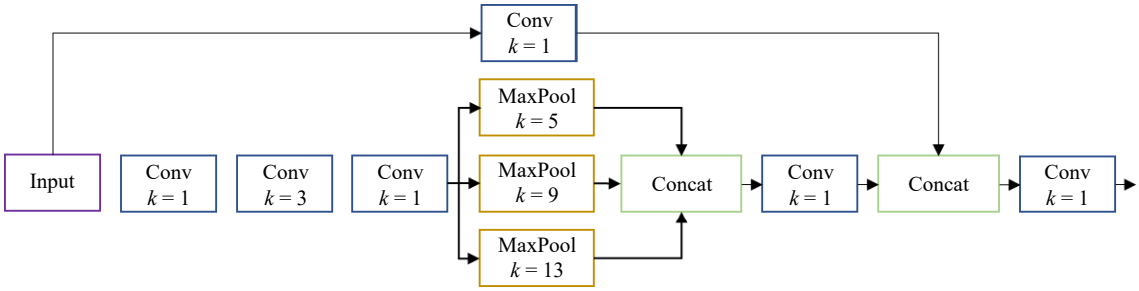



图 6 SPPCSPC 结构图

Fig. 6 SPPCSPC structure diagram

network，CSP）模块的思想，一个分支使用常规的卷积处理，另一条分支经过卷积处理后输入进 SPP 结构，对输出进行卷积处理，最后通过拼接合并两个分支。

## 2.2　改进的 YOLOv7 网络

### 2.2.1　基于 MicroBlock 改进的主干网络

为了降低网络参数和计算量，优化通道数量和节点之间的连接数量，MicroBlock 利用微分解方式对卷积进行自适应分解，假设卷积核 $\boldsymbol{W}$ 有相同的输入通道与输出通道，其数量都为 $C$，微分解公式如式（1）。

$$\boldsymbol{W} = \boldsymbol{P}\boldsymbol{\Phi}\boldsymbol{Q}^{\mathrm{T}} \tag{1}$$

式中：$\boldsymbol{W}$ 是一个 $C \times C$ 矩阵；$\boldsymbol{Q}$ 是一个 $C \times C/R$ 矩阵，$\boldsymbol{Q}$ 矩阵将通道数量压缩至 $1/R$。$\boldsymbol{P}$ 是一个 $C \times C/R$ 矩阵，可以将通道数量扩展回 $C$。$\boldsymbol{P}$ 与 $\boldsymbol{Q}$ 是具有 $G$ 块的对角块矩阵，矩阵中每块都实现了一组通道卷积。$\boldsymbol{\Phi}$ 是 $C/R \times C/R$ 矩阵，用于对通道进行打乱。其中 $R$ 代表通道压缩率，即压缩因子。

图 7 是一个微分解逐点卷积示例图，其输入通道和输出通道数 $C = 18$，压缩因子 $R = 2$，则组数 $G = 3$，$\boldsymbol{W}$ 被分解为 3 × 3 的块。

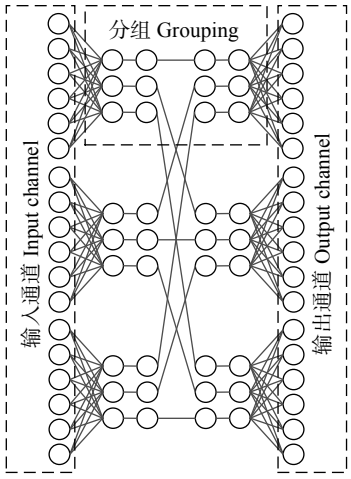



图 7　微分解的逐点卷积示例图

Fig. 7　Example diagram of micro-factorized pointwise convolution

矩阵 $\boldsymbol{\Phi}$ 的通道数代表隐藏通道数量，每个隐藏通道连接输入的 $C/G$ 个通道，每个输出通道与 $C/(RG)$ 个隐藏通道相连。每个输出通道的路径数 $E$ 反映该层的连通性，其中每条路径连接一个输入节点和一个输出节点。

$$C = \sqrt{\frac{ORG}{2}} \tag{2}$$

$$E = \frac{O}{2G} \tag{3}$$

式中：$O$ 代表分解层的计算复杂度。

图 8 显示：计算复杂度 $O$ 和压缩因子 $R$ 固定时，通道 $C$ 和连通性 $E$ 与 $G$ 的变化呈相反趋势，随着 $G$ 的增加，$C$ 随之增加，而 $E$ 减少。当两条曲线相交（$C = E$）时，每个输出通道只连接到所有输入通道一次。保证了输入与输出通道之间无冗余路径，同时也保证了每对输入输出之间都存在一条路径。

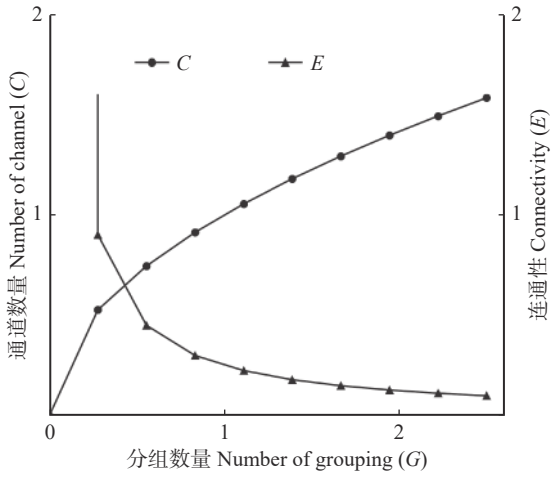



图 8　分组数量与通道数和连通性关系

Fig. 8　Relationship between the number of groups, channel count and connectivity

MicroBlock 同时对深度卷积进行微分解，使用 $k * 1$ 和 $1 * k$ 卷积代替了 $k * k$ 卷积。深度卷积分解同样适用公式（1），主要可以降低计算复杂度。微分解深度卷积将计算复杂度从 $O(k^2C)$ 降低至 $O(kC)$。

为了增加微分解卷积中组与组之间的连接，MiroBlock 引入动态激活函数（dynamic shift-max，DY-Shift-Max），DY-Shift-Max 输出 $K$ 个融合特征中的最大值，每个融合特征都结合 $J$ 个组的偏移，每一个通道上的 DY-Shift-Max（$y_i$）计算见式（4）。

$$y_i = \max\left\{\sum_{j=0}^{J-1} a_{i,j}^{k}(\boldsymbol{x})\boldsymbol{x}_{\left(i+j\frac{C}{G}\right)\bmod C}\right\}, 1 \leqslant k \leqslant K \tag{4}$$

式中：$\boldsymbol{x}$ 代表有 $C$ 个通道的输入向量，将其分为 $G$ 组，每组通道数为 $C/G$。$a_{i,j}^{k}(\boldsymbol{x})$ 是一个关于输入 $\boldsymbol{x}$ 的动态权重，由平均池化层、两个全连接层和 Sigmoid 激活函数构成。$\boldsymbol{x}_{(i+j\frac{C}{G})\bmod C}$ 代表 $\boldsymbol{x}$ 的第 $j$ 组循环偏移。

DY-Shift-Max 增强实现了两种方式增强非线性连接。第一，输出 $J$ 组 $K$ 个融合的最大值，增强了微分解逐点卷积的组内连接；第二，使用动态参数根据输入对每个融合进行加权，增强了网络的表示能力，降低缩减网络层数时造成的损失。

### 2.2.2　Slim-SPPFCSPC 结构

SPPCSPC 结构相比于空间池化金字塔[16]结构，虽然精度有所提升，但同时增加了模型的复杂度和计算量。本研究对 SPPCSP 结构进行轻量化改进，参考空间卷积池化金字塔−快速（spatial pyramid

pooling-fast，SPPF)结构的思想，将 SPPCSP 结构中的 SPP 部分改为 SPPF 结构，使其计算速度更快。去掉 3 * 3 卷积，结合空洞空间卷积池化金字塔(atrous spatial pyramid pooling，ASPP)结构思想，增加一条分支用以增加感受野获取，Slim-SPPFCSPC 结构如图 9 所示。

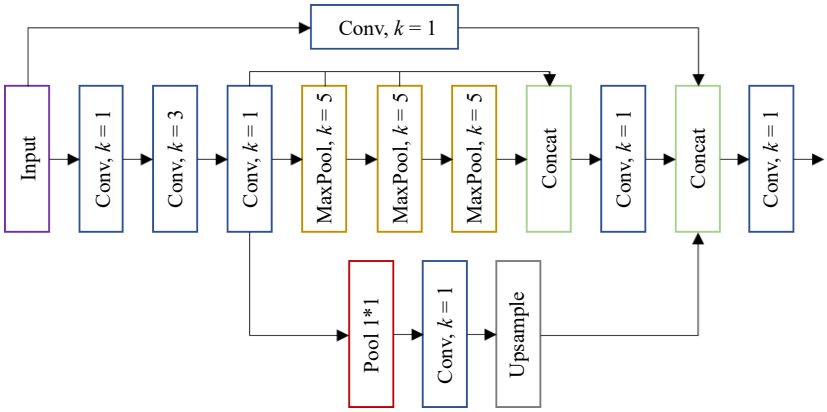



图 9 Slim-SPPFCSPC 结构图

Fig. 9 Slim-SPPFCSPC structure diagram

2.2.3 SIoU 损失

YOLOv7 使用了完全交并比(complete-intersection over union，CIoU)损失，考虑了预测框和真实框的中心距离、重叠面积、长宽比，使评估更加准确。虽然 CIoU 考虑了较多的因素，但没有考虑预测框与真实框的方向问题。SCYLLA 交并比(SCYLLA-intersection over union，SIoU)损失将预测框与真实框的夹角加入度量。

$$\Lambda = 1 - 2\sin^2\left(\arcsin\left(\frac{c_h}{\sigma}\right) - \frac{\pi}{4}\right) \tag{5}$$

式中：$\Lambda$是SIoU的角度代价，$c_h$代表真实框和预测框中心点的高度差，$\sigma$代表为预测框与真实框中心点的距离。其中$\sigma$与$c_h$的计算方式见式(6)和(7)，当$\frac{c_h}{\sigma}$趋向 0 时，角度代价趋向 0，对距离代价影响越小。图 10 中，$\sigma$为预测框与真实框中心点的距离，$B$ 是预测框中心点，$\left(b_{c_x}, b_{c_y}\right)$是预测框中心点坐标，$B^{gt}$ 是真实框中心点，$\left(b_{c_x}^{gt}, b_{c_y}^{gt}\right)$代表真实框中心点坐标。

$$\sigma = \sqrt{\left(b_{c_x}^{gt} - b_{c_x}\right)^2 + \left(b_{c_y}^{gt} - b_{c_y}\right)^2} \tag{6}$$

$$c_h = \max\left(b_{c_y}^{gt}, b_{c_y}\right) - \min\left(b_{c_y}^{gt}, b_{c_y}\right) \tag{7}$$

SIoU 重新设计了距离代价$\Delta$的计算，将角度代价加入距离代价。式(8)为 SIoU 的距离代价计算。

$$\Delta = \sum_{t=x,y}\left(1 - e^{-\gamma\rho_t}\right) \tag{8}$$

$$\rho_x = \left(\frac{b_{c_x}^{gt} - b_{c_x}}{c_w'}\right)^2,\ \rho_y = \left(\frac{b_{c_y}^{gt} - b_{c_y}}{c_h'}\right)^2,\ \gamma = 2 - \Lambda \tag{9}$$

式中：$c_w'$与$c_h'$为预测框和真实框的外接矩形的宽和高。

SIoU 的形状代价$\Omega$的计算为式(10)所示，$\theta$控制着形状代价的程度，$\theta$取值范围为 [2, 6]。$\omega_w$与$\omega_h$的计算如式(11)所示。

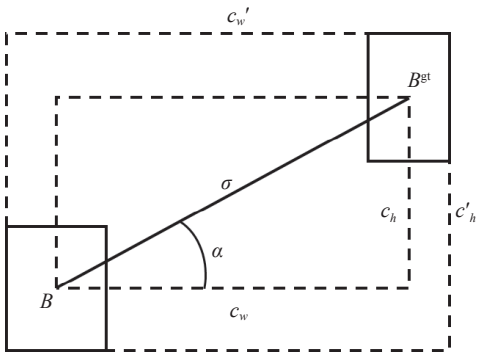
$$\Omega = \sum_{t=w,h}\left(1 - e^{-\omega_t}\right)^{\theta} \tag{10}$$

$$\omega_w = \frac{\left|w - w^{gt}\right|}{\max\left(w, w^{gt}\right)},\ \omega_h = \frac{\left|h - h^{gt}\right|}{\max\left(h, h^{gt}\right)} \tag{11}$$

式中：$w$, $h$与$w^{gt}$, $h^{gt}$分别代表预测框与真实框的宽和高。

SIoU 损失($L_{SIoU}$)的计算如式(11)所示。

$$L_{SIoU} = 1 - I_{oU} + \frac{\Delta + \Omega}{2} \tag{12}$$




$B$. 预测框中心点 The center point of the predicted box; $B^{gt}$. 真实框中心点 The center point of the ground truth bounding box; $\sigma$. 预测框与真实框中心点连接线 The connection line between the predicted box and the center point of the ground truth bounding box; $\alpha$. 预测框与真实框中心点连接线与 $x$ 轴夹角 The angle between the predicted box and the center point of the ground truth bounding box and the $x$-axis; $c_w$, $c_h$. 预测框与真实框中心点的宽度和高度差 The width and height difference between the center point of the predicted box and the center point of the ground truth bounding box; $c_w'$, $c_h'$. 预测框与真实框的外接矩形的宽和高 The width and height of the external rectangle of the predicted box and the ground truth bounding box

图 10 SIoU 损失

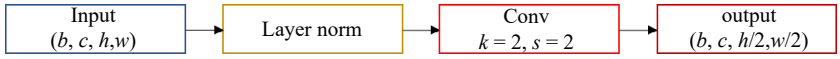
Fig. 10 SIoU loss

式中：$I_{oU}$ 为预测框与真实框交集与并集比值。

### 2.2.4　LNDown 下采样层

在之前的 YOLO 系列网络中主要采用了卷积核大小为 3、步长为 2 的卷积对特征图进行下采样，如果特征图通道数过多会带来更大的参数量。YOLOv7 网络中主要采用了一种多分支结构对网络进行下采样，并降低了参数量。

ConvNeXt 模型[17]受 visual transformer(ViT)结构启发，采用了一种新的下采样结构，将 LN 层与卷积核大小为 2、步长为 2 的卷积层结合使用(图 11)。组成 LNDown 模块对特征图进行下采样，可有效提升模型识别准确率。




*b*. 批次 Batch；*c*. 通道数量 Number of channels；*w*. 特征图宽度 Feature map width；*h*. 特征图高度 Feature map height. 下同。The same below.

图 11　LNDown 结构

Fig. 11　LNDown structure

Layer normalization(LN)层主要用于对特征进行归一化，与 batch normalization(BN)层对整个批次进行归一化不同，本文使用的 LN 对每个特征层上相同位置元素进行归一化(图 12)，可不受批次大小影响，对每个样本的特征进行归一化，在样本量较少时也有良好表现。结合使用 LN 和下采样卷积有助于增强模型稳定性。

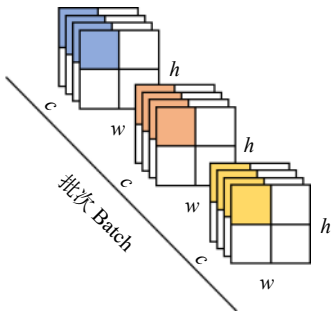



图 12　本文 Layer normalization 方法

Fig. 12　Layer normalization method in this paper

### 2.2.5　BiFPN 结构

BiFPN 是 EfficientDet 中提出的基于 FPN 的改进结构。在 FPN 结构(图 13a)中，首先在主干网络中对特征图进行下采样，再进行上采样，最终将其进行融合得到具有较强语义信息的多尺度特征图。FPN + PAN(图 13b)在原有 FPN 基础上，使用了自底向上结构对特征图进行下采样，融入低层定位特征，补充了 FPN 结构的定位信息。BiFPN(图 13c)相比于 FPN + PAN 结构，其删除了只存在一条输入边的节点，BiFPN 中所有节点都有两条以上的输入边，有助于增加模型特征融合能力；此外，BiFPN 在同层节点中增加了一条输入边，可以用小代价获取更多特征的融合；BiFPN 还可反复堆叠，上一个 BiFPN 的输出可作为下一个 BiFPN 的输入。

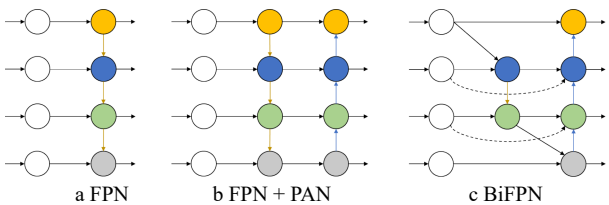



图 13　FPN、FPN + PAN、BiFPN 结构

Fig. 13　FPN, FPN + PAN, BiFPN structure

参考 BiFPN 结构思想，本研究对其轻量化后替换 YOLOv7 中使用的 FPN + PAN 结构。去掉 YOLOv7 中 3 个主干特征层输出后的 1*1 卷积，直接与上采样后的特征进行拼接，在中间层增加了一条连接边，增强特征融合。

## 3　研究方法

### 3.1　评价指标

为了对实验算法进行评估，本文采用的评价指标有参数量、每秒浮点运算次数(floating-point operations per second，FLOPS)、平均精度、漏检误差、错分误差和推理速度。参数量是衡量模型复杂度的重要指标，直接影响模型的存储空间和计算量。FLOPS 是模型在推理过程中每秒执行的浮点运算次数。

在判别目标时，通常将其分为正样本和负样本，并用真正例($T_P$)、真负例($T_N$)、假正例($F_P$)、假负例($F_N$)4 个指标表示。$T_P$ 即模型判断正确的正样本；$F_P$ 即模型误判成正样本的负样本；$T_N$ 即模型判断正确的负样本；$F_N$ 即模型误判成负样本的正样本。平均精度是目标检测任务中用于评估模型性能的常用指标，其表示了精度召回曲线与坐标轴所围区域面积，其中精度($P_r$)表示的是预测结果中真正的正例的比例，计算公式如式(13)所示。召回率($R_e$)表示的是所有正例中被正确预测出来的比例，计算公式如式(14)所示。

$$P_r = \frac{T_P}{T_P + F_P} \times 100\% \tag{13}$$

$$R_e = \frac{T_P}{T_P + F_N} \times 100\% \tag{14}$$

漏检误差可以判断模型对野生动物漏检情况，反映了模型未检测出野生动物图像的比例。错分误

差用于判断模型将多少空图像误检测为有动物出现的图像。推理速度是指模型推理一张图像所需的时间，本实验中的推理速度是模型通过推理 10 张图像计算得出的平均时间。

### 3.2 改进的 YOLOv7 结构

为使 YOLOv7 更好地应用于野生动物图像筛选，模型推理速度和准确率都是实际应用中应该关注的重要指标，既要提高模型准确率，又要加快模型的推理速度。将 YOLOv7 的主干网络中 E-ELAN 模块替换为 MicroBlock 模块，降低了模型参数量，加快了模型推理速度。同时将 SPPCSPC 改为更轻量的 Slim-SPPFCSPC 模块，进一步降低了参数量。使用 SIoU 损失替换 CIoU 损失，在不引入额外卷积计算量的同时提升模型检测动物的能力。将 YOLOv7 中原有的下采样结构改为 LNDown 结构，减少模型在下采样过程造成的信息丢失。引入 BiFPN 结构使模型更好地对特征进行融合，降低识别误差。图 14 为改进的 YOLOv7 结构图。

## 4 结果与分析

### 4.1 模型增强研究结果

为了验证改进后的 YOLOv7 模型中每个模块对网络的影响，本研究对上述提到的所有改进方案进行实验。实验选择从 SS 数据集中抽样 10 000 张图像的子数据集作为模型的训练和验证数据，训练集和验证集按 9∶1 划分。使用了起始为 0.001 的余弦学习率衰减策略，采用 SGD 优化器，迭代 100 个 epochs，输入的图像尺寸为 640 × 640 像素。

从表 1 可以看出：替换 MicroBlock 的 YOLOv7 网络降低了参数量和推理速度，同时降低了错分误

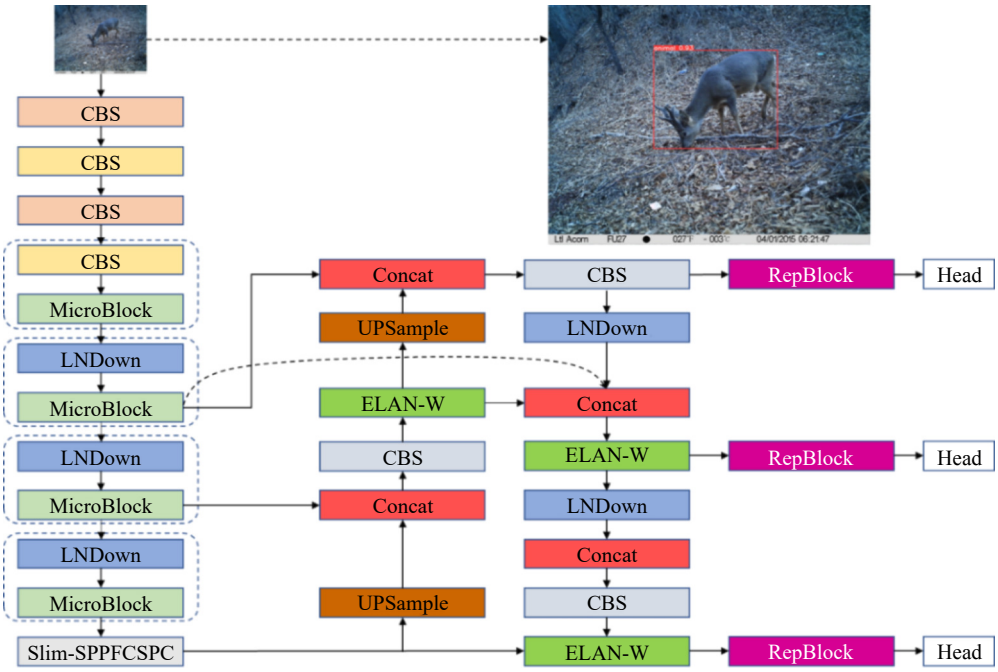



图 14 改进的 YOLOv7 结构图

Fig. 14 Improved YOLOv7 structure diagram

**表 1 模型增强研究结果**

Tab. 1 Results of models enhancement study

| 模型 Model | 参数量 Parameter quantity | 每秒浮点运算次数 Floating-point operation per second (FLOPS) | 权重 Weight/% | 平均精度 Average precision/% | 漏检误差 Omission error/% | 错分误差 Commission error/% | 推理速度 Inference speed/ms |
|---|---|---|---|---|---|---|---|
| YOLOv7 | $36.5 \times 10^6$ | $104.7 \times 10^9$ | 72.0 | 80.0 | 28.2 | 15.5 | 17.5 |
| YOLOv7 + MicroBlock | $31.3 \times 10^6$ | $72.7 \times 10^9$ | 61.5 | 78.9 | 30.1 | 13.0 | 15.6 |
| YOLOv7 + Slim-SPPFCSPC | $33.9 \times 10^6$ | $101.2 \times 10^9$ | 66.5 | 79.7 | 29.7 | 13.5 | 16.4 |
| YOLOv7 + SIoU | $36.5 \times 10^6$ | $104.7 \times 10^9$ | 72.0 | 81.0 | 27.0 | 15.3 | 17.6 |
| YOLOv7 + LNDown | $37.2 \times 10^6$ | $100.6 \times 10^9$ | 72.8 | 81.1 | 28.9 | 11.5 | 17.8 |
| YOLOv7 + BiFPN | $36.8 \times 10^6$ | $104.5 \times 10^9$ | 72.0 | 80.4 | 29.8 | 11.4 | 17.9 |
| 本模型 Our model | $30.0 \times 10^6$ | $69.6 \times 10^9$ | 59.0 | 80.5 | 28.5 | 12.7 | 15.0 |

注：推理速度是指模型推理一张图像所需的时间。Note: Inference speed is the time it takes the model to infer an image.

差，但平均精度降低 1.1%，且漏检误差提高了 1.9%，说明主干网络使用 MicroBlock 后模型的特征提取能力降低了，对动物的检测效果下降。使用 Slim-SPPCSPC 的 YOLOv7 与使用 MicroBlock 方案的结果类似，同样降低了参数量与错分误差，增加了漏检误差。使用 SIoU 损失的 YOLOv7 与原模型的参数量和推理时间差别不大，但平均精度提升了 1.0%，漏检误差与错分误差分别降低 1.2% 和 0.2%，说明使用 SIoU 损失可有效提升模型检测野生动物的能力，且无需额外推理时间。使用 LNDown 进行改进的模型，平均精度提升了 1.1%，错分误差降低了 4.0%，说明使用 LNDown 改进的模型可有效提高对动物检测的精准度。使用 BiFPN 后，模型平均精度有 0.4% 的提升，且错分误差比例在所有改进方案中最低，说明通过改进网络特征融合方式可有效降低模型将背景误识别成动物的概率，但需要更多的参数量和推理时间。本文将上述改进方案结合，改进后的模型在速度上高于原始 YOLOv7 模型，推理速度快了 14.3%，模型参数量减少了 17.8%，FLOPS 降低了 33.5%，可使用更少的计算资源对野生动物动物图像进行更快速的检测。

图 15 是 3 组模型检测对比图。从 a 组中可以看

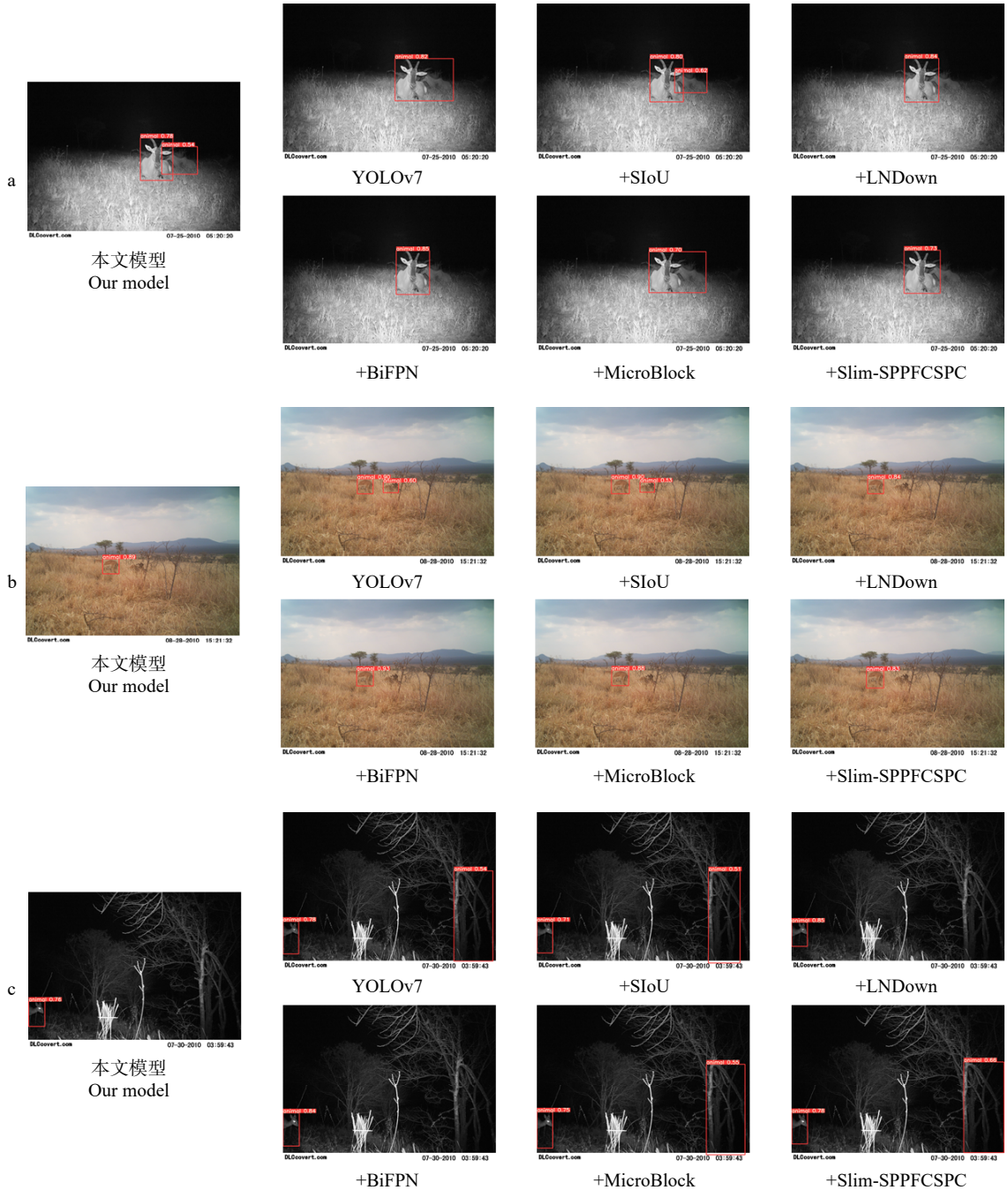



图 15　不同改进方案检测結果

Fig. 15　Detection results with different improvement schemes

出：SIoU 损失的加入，使模型识别出后侧动物，而未使用 SIoU 损失的改进方案均未检测出该动物，且容易将二者检测为一个，说明使用 SIoU 损失可进行更好的回归从而降低漏检率。从 b 组中可看出：SIoU 损失的加入对错分情况并没有改善，仍将 b 组图像右侧石头检测为动物，而其他改进方案未出现该问题。在 c 组图像中可看出：原模型即部分改进模型将树木当作动物形成误检测，而使用 BiFPN 和 LNDown 改进的模型减少了此类情况的发生。

### 4.2 不同模型对比结果

为了对比改进模型对野生动物图像的筛选效果，除了本文的改进算法，还选择了 YOLOv5-m、YOLOv5-l、Ghost-YOLOv5-l、YOLOv6、YOLOX-M[18]、YOLOR-CSP[19]6 个模型，用于对比不同模型性能。实验使用的参数与 4.1 小节相同。

结果如表 2 所示，虽然本文改进的 YOLOv7 速度稍慢于 YOLOv5-m，但平均精度高于 YOLOv5-m，且漏检误差与错分误差均低于 YOLOv5-m。YOLOv5-l 的精度虽略高，但速度比改进的 YOLOv7 模型更慢，且漏检误差更高，说明其对空图像的检测效果不如改进的 YOLOv7 算法。YOLOX-M 虽然参数量小于改进的 YOLOv7 模型，但其速度并没有更快，且漏检误差更大。YOLOR-CSP 算法精度与改进的 YOLOv7 算法基本持平，但是其速度比改进的 YOLOv7 模型慢了 3.4 ms，且参数量和 FLOPS 更高，权重高了 2.4 倍。Ghost-YOLOv5-l 模型中，虽然参数量和 FLOPS 等都大幅降低，但其推理时间比原始模型高了 3.3 ms，可能是由于 GhostNet 模块使用了深度可分离卷积，增加了数据的读写，使模型推理速度变慢。

**表 2 不同模型在 SS 数据集子集上检测结果**

Tab. 2 Detection results of different models on the SS subset

| 模型 Model | 参数量 Parameter quantity | FLOPS | 权重 Weight/% | 平均精度 Average precision/% | 漏检误差 Omission error/% | 错分误差 Commission error/% | 推理速度 Inference speed/ms |
|---|---|---|---|---|---|---|---|
| YOLOv5-m | $21.2\times10^6$ | $49.0\times10^9$ | 41.0 | 80.2 | 28.8 | 16.0 | 14.5 |
| YOLOv5-l | $46.5\times10^6$ | $109.1\times10^9$ | 89.0 | 81.0 | 28.1 | 12.6 | 18.1 |
| Ghost-YOLOv5-l | $20.4\times10^6$ | $43.1\times10^9$ | 41.7 | 80.7 | 28.9 | 11.9 | 21.4 |
| YOLOv6 | $34.9\times10^6$ | $85.8\times10^9$ | 72.5 | 79.8 | 27.7 | 16.3 | 17.2 |
| YOLOX-M | $25.3\times10^6$ | $73.8\times10^9$ | 194.0 | 80.6 | 28.9 | 12.2 | 19.2 |
| YOLOR-CSP | $52.9\times10^6$ | $110.1\times10^9$ | 203.0 | 80.4 | 30.6 | 11.6 | 18.4 |
| 本模型 Our model | $30.0\times10^6$ | $69.6\times10^9$ | 59.0 | 80.5 | 27.9 | 12.7 | 15.0 |

相较于其他模型，尽管本文改进的 YOLOv7 未在所有指标均达到最优，但其可在同参数量的情况下达到更快的推理速度和更低的误差率，在检测速度、检测精度和模型参数量方面达到了更好的平衡。

### 4.3 改进 YOLOv7 模型图像筛选

实验使用 SS 子集训练的模型作为预训练模型，利用迁移学习在自建野生动物数据集中进行微调，并与未进行迁移学习的模型进行对比。通过冻结主干网络的不同层数，探究模型最好的学习效果，同时对比未使用迁移学习模型的检测效果。实验所采用的学习率为 0.000 1，采用 SGD 优化器，迭代 50 个 epochs。

表 3 为不同迁移方案中，模型在自建数据集中的实验结果。从表 3 中可看出：在使用迁移学习的模型中，未冻结权重的模型与冻结 3 层权重的模型平均精度更高，漏检误差和错分误差更低，而冻结 5 层权重与完全冻结主干网络权重的模型各项指标越来越低。这说明经过预训练的模型其浅层特征具有通用性，由于物种不完全相同，冻结深层特征导致模型无法有效学习到自建数据集动物特征，导致精度无法进一步提升。未使用迁移学习直接训练的模型平均精度较低，漏检误差和错分误差更高，无法很好地对野生动物图像进行筛选。

**表 3 迁移学习不同冻结方案结果**

Tab. 3 Results of different freezing layers for transfer learning %

| 冻结层数 Frozen layer | 平均精度 Average precision | 漏检误差 Omission error | 错分误差 Commission error |
|---|---|---|---|
| 未迁移 Not transferred | 69.5 | 37.5 | 21.9 |
| 0 | 82.1 | 24.8 | 11.4 |
| 3 | 81.9 | 28.0 | 10.2 |
| 5 | 76.7 | 33.3 | 13.4 |
| 11（冻结主干） 11 (freeze backbone) | 64.3 | 46.2 | 19.2 |

通过图 16 发现：冻结权重的模型起始平均精度较高，且前 5 轮平均精度提升速度较快，20 轮之后逐步趋于稳定。这说明迁移学习可使模型在自建数

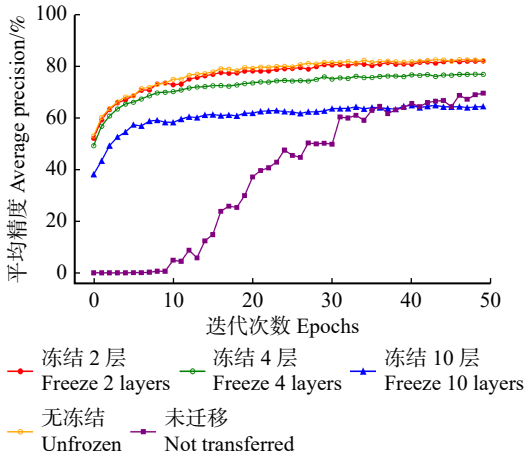



图 16　冻结不同层数平均精度曲线

Fig. 16　Freeze average precision curves of different freezing layers

据集上快速收敛，且模型各项参数与数据集匹配度较好。而未使用迁移学习的模型收敛较慢，前 10 轮平均精度几乎无上升，后期虽上升幅度较大，但 50 轮后仍未达到稳定状态。说明使用更大规模的野生动物相关数据集进行预训练，可有效提升模型在小规模野生动物图像数据集上的检测效果。

在实际应用中，仅需按图片对空图像进行筛选而非标注，为使测试结果更接近真实情况，对自建数据集同区域的额外 400 张图像进行测试，测试过程将目标级真正例、真负例、假正例、假负例应用于图像集，测试模型在实际应用中对空图像的筛选效果。

实验结果如表 4 所示，从结果中可看出：正确类别的真正例和真负例数量较多，且漏检误差与错分误差较低，可很好地应用于野生动物空图像的筛选。此外，还单独测试了几张其他地区红外相机拍摄到的自建数据集中未出现过类别的动物图片，经过微调的模型也可很好地对其识别，结果如图 17 所示。说明迁移学习微调模型具有较好的泛化能力，预训练时出现的动物迁移后也可正确识别，因此可在更多保护区中应用此模型。

**表 4　本文算法在自建数据集筛选结果**

Tab. 4　Our algorithm filter results in the self-built dataset

| 真正例数量<br>Number of true positive | 假正例数量<br>Number of false positive | 真负例数量<br>Number of true negative | 假负例数量<br>Number of false negative | 漏检误差<br>Omission error/% | 错分误差<br>Commission error/% |
|---|---|---|---|---|---|
| 191 | 5 | 195 | 9 | 4.5 | 2.6 |



图 17　未在数据集中出现的图像

Fig. 17　Images that do not appear in the dataset

## 5　讨　　论

本研究根据野外环境复杂，红外相机易受其他干扰而误触发拍摄产生大量的空照片这一问题，使用 MicroBlock、Slim-SPPCSPC、SIoU、LNDown 和 BiFPN 对 YOLOv7 模型进行改进。使用 MicroBlock 替换主干网络和 Slim-SPPFCSPC 结构的 YOLOv7 网络均降低了参数量，加快了推理速度，但对动物的检测效果下降。使用 SIoU 损失、LNDown 和 BiFPN 的方式均可在较低的推理速度下获取更高的性能，且提高了模型对体型与动物近似物体的区分度，降低了模型对动物的漏检误差。

本文将上述改进方案结合实验后发现，改进后的模型相比于原始 YOLOv7 模型，降低了环境对动物识别的影响，对不同位置的野生动物识别更加准确，减少了动物的错分误差。此外在与主流的目标检测模型对比中发现，改进的 YOLOv7 模型在速度与精度方面也更具优势。使用迁移学习进行微调的模型具有更高的鲁棒性，可以使用少量的数据进行训练即可快速达到更高的精确度，用更少的计算资源对野生动物图像进行识别，满足日益增长的红外相机拍摄图像的筛选需求，更好地服务于密云地区生物多样性的保护。

## 6　结　　论

本研究对 YOLOv7 进行不同的改进并进行相关

实验后，选择了兼顾速度、准确率、误检率的最优改进方案。在 SS 相机陷阱图像数据集子集上，将改进的 YOLOv7 模型与 YOLOv5-m、YOLOv5-l、Ghost-YOLOv5-l、YOLOv6、YOLOX-M、YOLOR-CSP 模型进行对比，虽然改进后的 YOLOv7 模型未在所有指标中都达最优，但更好地平衡了推理时间、推理精度和模型参数量。此外还使用迁移学习在自建数据集上进行微调，并对比冻结不同层数对模型的影响，未冻结层数的模型平均精度达到了 82.1%。经过迁移学习的改进模型可以更好地对北京市雾灵山自然保护区内野生动物图像进行筛选，筛选后的数据对野生动物数据集的建立提供了更高的可靠性，同时节省了推理时间和计算资源，为密云地区生物多样性研究和野生动物保护工作提供了更高效的支持。基于本研究工作开发的模型和软件已经成功应用于北京园林绿化生态系统监测网络平台，提升了管理和运维人员的工作效率。

## 参 考 文 献


[1] Kays R, McShea W J, Wikelski M. Born-digital biodiversity data: millions and billions[J]. Diversity and Distributions, 2020, 26(5): 644−648.

[2] 肖治术, 肖文宏, 王天明, 等. 中国野生动物红外相机监测与研究: 现状及未来 [J]. 生物多样性, 2022, 30: 234−259.
Xiao Z S, Xiao W H, Wang T M, et al. Infrared camera monitoring and research on wildlife in China: current status and future[J]. Biodiversity, 2022, 30: 234−259.

[3] Villa A G, Salazar A, Vargas F. Towards automatic wild animal monitoring: identification of animal species in camera-trap images using very deep convolutional neural networks[J]. Ecological Informatics, 2017, 41: 24−32.

[4] Norouzzadeh M S, Nguyen A, Kosmala M, et al. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning[J]. Proceedings of the National Academy of Sciences, 2018, 115(25): e5716−e5725.

[5] Schneider S, Greenberg S, Taylor G W, et al. Three critical factors affecting automated image species recognition performance for camera traps[J]. Ecology and Evolution, 2020, 10(7): 3503−3517.

[6] Guo Y, Rothfus T A, Ashour A S, et al. Varied channels region proposal and classification network for wildlife image classification under complex environment[J]. IET Image Processing, 2020, 14(4): 585−591.

[7] Chen P, Swarup P, Matkowski W M, et al. A study on giant panda recognition based on images of a large proportion of captive pandas[J]. Ecology and Evolution, 2020, 10(7): 3561−3573.

[8] Tan M, Chao W, Cheng J K, et al. Animal detection and classification from camera trap images using different mainstream object detection architectures[J]. Animals, 2022, 12(15): 1976.

[9] Cai Z, Vasconcelos N. Cascade r-cnn: delving into high quality object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake: IEEE, 2018: 6154−6162.

[10] Tian Z, Shen C, Chen H, et al. Fcos: fully convolutional one-stage object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 9627−9636.

[11] Vecvanags A, Aktas K, Pavlovs I, et al. Ungulate detection and species classification from camera trap images using RetinaNet and faster R-CNN[J]. Entropy, 2022, 24(3): 353.

[12] Roy A M, Bhaduri J, Kumar T, et al. WilDect-YOLO: an efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection[J]. Ecological Informatics, 2023, 75: 101919.

[13] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii: IEEE, 2017: 4700−4708.

[14] Li Y, Chen Y, Dai X, et al. Micronet: improving image recognition with extremely low flops[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 468−477.

[15] Tan M, Pang R, Le Q V. Efficientdet: scalable and efficient object detection[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle: IEEE, 2020: 10781−10790.

[16] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904−1916.

[17] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New Orleans: IEEE, 2022: 11976−11986.

[18] Ge Z, Liu S, Wang F, et al, Yolox: exceeding yolo series in 2021[J/OL]. arXiv preprint arXiv: 2107.08430, 2021 [2023−12−23]. https://arxiv.linfen3.top/pdf/2107.08430.pdf.

[19] Wang C Y, Yeh I H, Liao H Y M. You only learn one representation: unified network for multiple tasks[J/OL]. arXiv preprint arXiv: 2105.04206. 2021 [2023−12−23]. https://arxiv.linfen3.top/abs/2105.04206.pdf.


（责任编辑 吴 娟
责任编委 陈志泊）