Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2021, 10(4), 281-291
Published Online November 2021 in Hans. http://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2021.104028

Hans汉斯

基于Swin-Transformer的野生动物检测

姜福豪，隋晨红*，欧世峰，王中训，胡国英，杨国斌，潘云豪，胡　健

烟台大学光电信息科学技术学院，山东　烟台

收稿日期：2021年9月13日；录用日期：2021年10月25日；发布日期：2021年11月2日

摘　要

野生动物检测对于更好地开展野生动物保护、维持生物多样性和生态系统平衡具有重要意义。随着科技的进步，野生动物检测已从传统的人工寻觅、人眼识别发展到利用机器学习技术进行快速检测的阶段。然而，当前各种检测模型存在检测精度不高的问题。因此，本文将Swin-Transformer技术应用到野生动物目标检测模型，并与其他的优秀的检测模型进行性能比较。实验结果表明与其他优秀的检测器相比，Swin-Transformer检测的平均检测精度为0.958，领先于其他检测模型至少5%，并且该检测器对绝大多数动物的检测均可取得最优结果，即使是对于样本数量较少的稀有类别，检测精度依然能够达到91%，极大提高了野生动物检测的准确率。

关键词

深度学习，目标检测，野生动物，Swin-Transformer

Wild Animal Detection Based on Swin-Transformer

Fuhao Jiang, Chenhong Sui*, Shifeng Ou, Zhongxun Wang, Guoying Hu, Guobin Yang, Yunhao Pan, Jian Hu

School of Opto-Electronic Information Science and Technology, Yantai University, Yantai Shandong

Received: Sep. 13th, 2021; accepted: Oct. 25th, 2021; published: Nov. 2nd, 2021

Abstract

Wildlife detection is of great significance for better carrying out wildlife protection, maintaining biodiversity and ecosystem balance. With the advancement of science and technology, wildlife detection has evolved from traditional manual search and human eye recognition to the stage of rapid detection using machine learning technology. However, the current detection models have the problem of low detection accuracy. Therefore, this article applies the Swin-Transformer technology to the wild animal target detection model, and compared it with other excellent models.

*通讯作者。

文章引用: 姜福豪, 隋晨红, 欧世峰, 王中训, 胡国英, 杨国斌, 潘云豪, 胡健. 基于 Swin-Transformer 的野生动物检测[J]. 人工智能与机器人研究, 2021, 10(4): 281-291. DOI: 10.12677/airr.2021.104028

Experimental results show that compared with other excellent detectors, the average precision value of Swin-Transformer detection is 0.958, which is at least 5% ahead of other detection models, and the detector achieves the best results for most categories, even for rare categories, the accuracy can reach 91%, which greatly improves the detection accuracy.

Keywords

Deep Learning, Target Detection, Wild Animal, Swin-Transformer

Copyright © 2021 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

对生物多样性进行研究，特别是濒危对野生物种进行研究时，需要花费大量的人力物力寻觅、识别野生动物，获取资料后，数量统计主要是靠人工计数的方法。人工计数的方法有多方面的缺点，如消耗较大的人力物力以及较多的时间，而且由于人为因素影响，统计结果并不精确。相机技术的高速发展与硬件成本的降低，在生物科学研究领域越来越多的使用相机陷阱代替人力获取资料，红外探测器、动作传感器或者是其他光束都可以作为它的触发机关，相机陷阱已经成为了科学家们获取资料的得力助手，但是如何从海量数据中快速高效获取信息成为了研究者进行研究的关键。

基于深度学习的目标检测在野生动物识别方面有巨大优势，得益于图形处理器(GPU)的不断发展，越来越多的深度学习模型不断涌现。2012 年，Alex Krizhevsk 等[1]在 ImageNet ILSVRC 中获得冠军，识别率远远超过第二名，在之后的几年当中，VGG 网络[2]、GoogLenet [3]、ResNet [4]、Densenet [5]网络结构不断问世，同时基于单阶段、双阶段的目标检测模型也大放异彩，单阶段的目标检测模型主要包括 SSD [6]、RetinaNet [7]、Yolov3 [8]、yolov5、yolox [9]等；双阶段的目标检测模型主要包括 FasterRCNN [10]、MaskRCNN [11]等。通常来说，基于双阶段的目标检测模型精度高，但是检测速度慢，基于单阶段的目标检测模型精度低，检测速度较快，并且随着 YOLO 系列的发展，检测精度差距也在逐步缩小。

目前，有非常多的学者提出将深度学习应用到野生动物图像识别当中，刘文定等[12]提出一种基于感兴趣区域(ROI)与卷积神经网络(CNN)的野生动物物种自动识别方法，并将其应用于内蒙古赛罕乌拉国家自然保护区内常见的五种生物，何育欣等[13]提出将卷积神经网络应用到大熊猫的检测，黄鑫达等[14]分别提出了基于 Faster R-CNN [10]、YOLOv3 [8]模型的改进动物目标检测算法，陈刚琦等[15]基于卷积神经网络的高原鼠兔图像检测，史春妹等[16]出基于目标检测的东北虎个体自动识别，程浙安等[17]提出基于深度卷积神经网络的内蒙古地区陆生野生动物自动识别，黄元涛等[18]以及翟俊伟等[19]也提出将深度学习应用于藏羚羊的检测，此外张艺秋等[20]和王飞等[21]提出了基于深度学习检测森林火灾的结构。然而，以上尝试都有各种各样限制：检测目标单一；检测场景有限；数据量较少。将之前的检测模型应用到本文的野生动物数据集中，结果不够理想。

自从 Transformer [22]在自然语言处理取得突破性的进展之后，研究者一直尝试着把 Transformer 用于在计算机领域。之前的各种尝试例如 iGPT [23]、ViT [24]，由于 Transformer 对于长序列的处理的局限性，都是将 Transformer 用于图像分类领域，直至 2020 年，Facebook 研究团队开发出 DEtection TRansformer

(DETR) [25]之后，transformer 在计算机视觉应用的探索越来越广泛，2021 年，MSRA 提出了令人振奋的 Swin-Transformer [26]，在分类、检测、分割任务上都取得了最优的效果。

为此，本文将 Swin-Transformer 应用于野生动物检测领域。

2. Swin-Transformer

Swin-Tranformer 结构如图 1 所示，一共有 4 个 Stage，当输入图片(假设输入图片尺寸为 224×224)依次经过时，特征图的尺寸不断降低，通道数增加，特征图中的特征的感受野不断扩大，与卷积神经网络不断提取特征的过程非常相似。该过程主要步骤如下：

将图片输入到网络中，经过区块划分(Patch Partition)模块，将图片按像素划分成不同的小块，将小块在所有通道上的像素拉伸为一维向量，这一维向量也称为 token，将所有向量组合成矩阵，矩阵经过四个阶段(Stage)之后进行分类和回归，每个阶段由区块合并(Patch Merging)和 Swin-Transformer Block 组成，区块合并主要作用是降低特征图分辨率，其中第一阶段除外，小块组合在经过线性编码(Linear Embedding)后输入到 Swin-Transformer Block 中，线性编码仅仅改变了特征的通道数，并不降低分辨率。

2.1. 区块划分(Patch Partition)

输入图片表示为像素矩阵，需要先对图片进行 patch partition 处理，将图片的最小单位从像素转变为

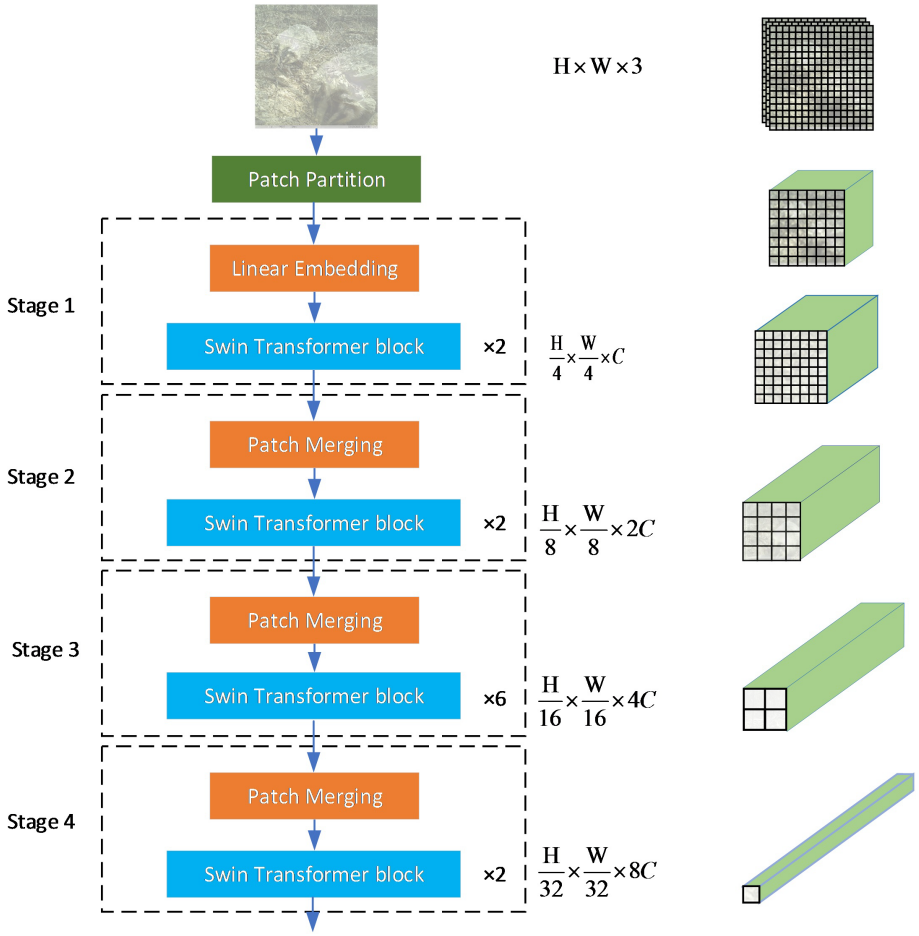

Figure 1. The architecture of a Swin-Transformer
图 1. Swin-Transformer 网络架构

patch。每一个小块由 4×4×3 个像素构成，即用包含 4×4 个像素的区块来对像素矩阵进行分割，将每个区块所有通道中的像素值拉伸为一维向量，向量长度为 48，将所有向量按照原始区块进行组合，得到像素矩阵。因此，输入的像素矩阵 $H\times W\times 3$ (维度为 224×224×3)经过区块划分处理后变为维度为 56×56×48 的三维矩阵，其中 56×56 表示区块的数量，48 为通道数。表示为式(1)

$$x^0 = Unfold\left(\mathrm{Im}age\right) \tag{1}$$

2.2. 线性编码(Linear Embedding)

线性嵌入层应用于此原始值特征在通道数量上以将其投影到任意维度，这一过程是通过多层感知机(Muti-Layer Perception, MLP)实现的，输出维度用 C 表示，默认值是 96，所以经过 Linear Embedding 之后输出 56×56×96，这一过程可以用式(2)表示

$$x^1 = MLP\left(x^0\right) \tag{2}$$

2.3. Swin-Transformer Block

如图 2 所示，Swin-Transformer 成对出现组成 Swin-Transformer Block，Swin-Transformer 主要包括两

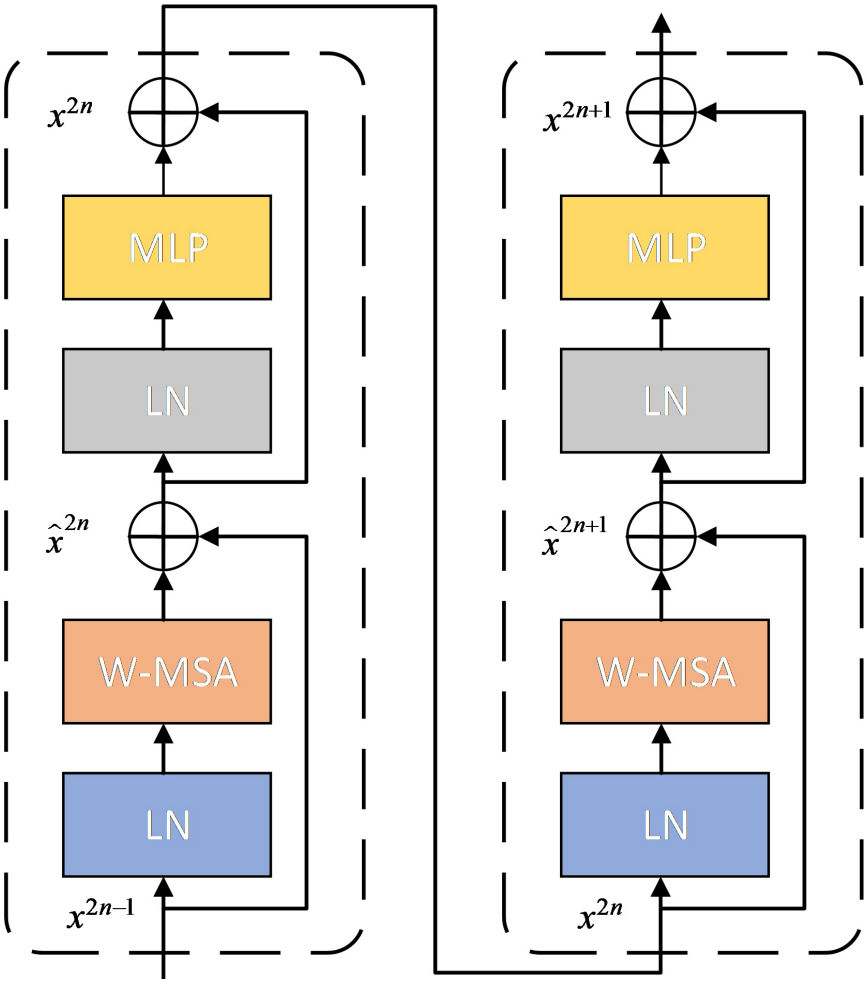

Figure 2. The architecture of Swin-Transformer Blocks (2×)
图 2. Swin-Transformer Blocks 网络结构(2×)

部分，多头自注意层(multi-head self-attention, MSA)不同和多层感知机(Muti-Layer Perception, MLP)，在每个 MSA 和每个 MLP 模块之前应用层规范化(Layer Norm, LN)层，并在每个模块之后应用残余连接。两个连续的 Swin-Transformer 区别是多头自注意层(Multi-head Self-attention, MSA)不同，前者为窗口多头自注意层(window multi-head self-attention, W-MSA)，后者为移位窗口多头自注意层(shifted-window multi-head self-attention, SW-MSA)组成。这一过程可以使用式(3)表示。

2.3.1. 窗口多头自注意层

多头自注意力将图片化划分为窗口，如图 3 所示，其中(a)为原图，(b)是 W-MSA 操作结果，(c)为 SW-MSA 结果。W-MSA 本质上是将矩阵由 8×8 个区块为单位转化成 2×2 个窗口，每个窗口包括 4×4 个区块。W-MSA 将输入图片划分成不重合的窗口，然后在不同的窗口内进行自注意力计算。假设一个图片有 $h\times w$ 的窗口，每个窗口包含 $M\times M$ 个区块，W-MAS 整体过程如下：

$$\begin{aligned}\hat{x}^{2n} &= W-MSA\left(LN\left(x^{2n-1}\right)\right)+x^{2n-1}\\ x^{2n} &= MLP\left(LN\left(\hat{x}^{2n}\right)\right)+\hat{x}^{2n}\\ \hat{x}^{2n+1} &= SW-MSA\left(LNx^{2n}\right)+x^{2n}\\ x^{2n+1} &= MLP\left(LN\left(\hat{x}^{2n+1}\right)\right)+\hat{x}^{2n+1}\end{aligned} \tag{3}$$

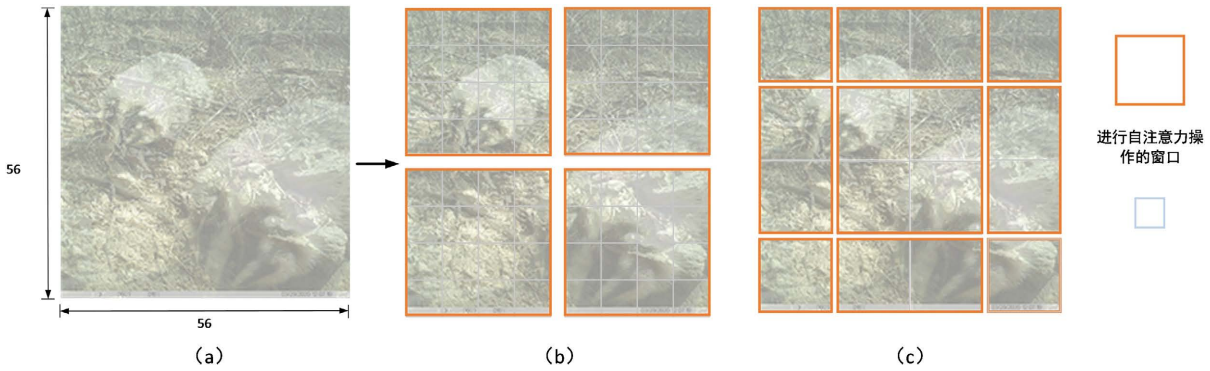

Figure 3. The result of window division using different multi-head self-attention
图 3. 采用不同多头自注意力进行窗口划分的结果

$LN\left(x^{2n-1}\right)$ 经过线性变换之后得到 $56\times56\times3C$ 的矩阵，然后将矩阵均分为三份，成为 Transformer 中 Q、K、V 三个特征，每个特征维度为 $56\times56\times96$，然后经过矩阵转置、复制等操作，得到独立的窗口的 3 个的权值矩阵，维度为 $3\times64\times49\times32$，其中 3 表示多头注意力个数；64 表示区块个数，由于输入尺寸是 56×56，由于区块大小为 7，因此一共剩下 8×8 个区块；49 表示每个区块所包含的像素数量；32 表示隐层节点个数由 C/多头个数得到。后续三个特征根据式(4)进行操作

$$\text{Attention}\left(Q,K,V\right)=\text{Softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}+B\right)V \tag{4}$$

其中，Q、K、V 分表表示 Transformer 的三个特征；B 表示相对位置偏差；d_k [26]表示特征 K 的方差，是一个常数。

Swin-Transformer 将计算区域从以区块为单位那改为以窗口为单位，那么根据公式(4)，MSA 和 W-MSA 的计算复杂度式(5)、式(6)分别如下：

$$\Omega(MSA)=4hwC^2+2(hw)^2C \tag{5}$$

$$\Omega(W-MSA)=4hwC^2+2M^2hwC \tag{6}$$

由于窗口的数量远小于区块数量，W-MSA 的计算复杂度和图像尺寸呈线性关系。W-MSA 虽然降低了计算复杂度，但是不重合的 window 之间缺乏信息交流，于是进一步引入 shifted window partition 来解决不同窗口的信息交流问题。

2.3.2. 移位窗口多头自注意层

Swin-Transformer Block 第二层中移位窗口多头自注意层的窗口位置进行变动，得到 3×3 个不重合的窗口，如图 3(c)所示。移动窗口的划分方式使上一层相邻的不重合不同之间引入连接，避免目标由于之前分布在不同窗口而产生遗漏。

但是位移窗口划分方式还引入了另外一个问题，就是会产生更多的窗口，并且窗口大小不一，计算量增加 1.25 倍。因此提出一种新的方式来解决这个问题，如图 4 所示，通过窗口滚动的方式，首先，将(a)第一行的三个窗口滚动到最后一列得到(b)，然后(b)将第一列的窗口滚动到最后一列得到(c)，3×3 的窗口转换成 2×2 的窗口。

Figure 4. Illustration of an approach to shift windows for SW-MSA
图 4. SW-MSA 中窗口变动说明

得到新的窗口之后，SW-MSA 仍然需要计算自注意力，但是通过变动之后形成的新的窗口是由不相邻的不规则窗口组成，在计算自注意力时，应该将这种情况剔除，因此通过引入 Mask 机制，由于矩阵后续输入到 softmax 函数中，当矩阵与对应的 mask 相加时，只需要将剔除位置的 mask 数值设置为−100，其余设置为 0 即可。具体情况如图 5 所示。

2.4. 区块合并

该模块的作用是在每个 Stage 开始前做降采样，用于缩小分辨率，调整通道数进而形成类似于卷积神经网络中的层次化的设计，增加了感受野，同时也能节省一定运算量。

每次降采样是两倍，因此在行方向和列方向上，间隔 2 选取元素。然后拼接在一起作为一整个张量，最后展开。此时通道维度会变成原先的 4 倍(因为 H，W 各缩小 2 倍)，此时再通过一个全连接层再调整通道维度为原来的两倍。

3. 数据集构建

3.1. 数据集筛选

数据集是密云雾灵山地区安装的红外相机拍摄到有关野生动物的图片以及视频，视频是根据观察，

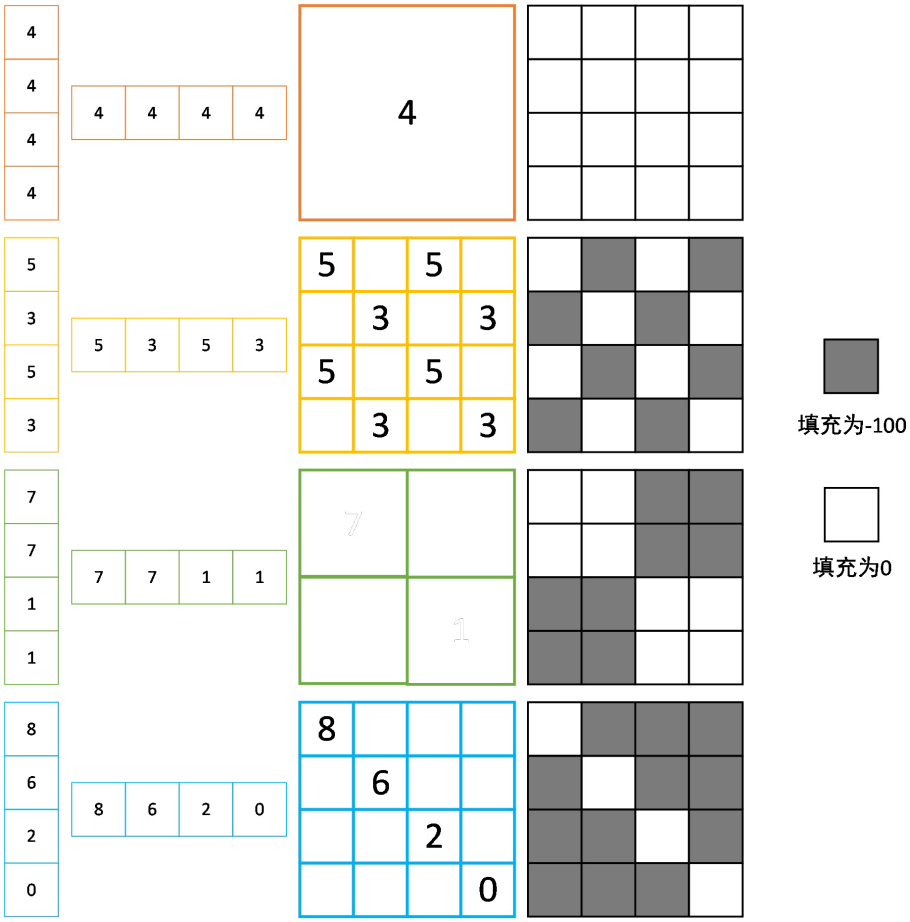

Figure 5. Mask mechanism in SW-MSA
图 5. SW-MSA 中 mask 机制

每隔 10 帧取出一张图片，这样保证了既不丢失目标，又可以减少工作量。最后将所有的图片经过手工筛选以及标注得到的数据集。数据集主要包括六类，每一类都由不同种类、不同场景而组成。

3.2. 数据集分析

数据集由 6043 张图片，共六类组成，我们将数据集划分成训练集、验证集，比例为 1:1。其中训练集为 3022 张图片，验证集 3021 张图片。每类图片详见表 1，根据表格可知，数据分布极不平衡，不同种类占比从 2.23%~47%不等，因此，该数据集对不同网络处理类别不平衡能力要求非常高。

Table 1. The number of different categories in different data sets
表 1. 不同数据集中不同类别的数目情况

獾	熊	鸟	羊	兔子	松鼠	共计
135	2842	2122	596	107	241	6043

3.3. 实验数据预处理

数据预处理在构建网络模型时是十分重要的，特别是训练数据比较少的时候。由于原始图片尺寸较大，为 4000 × 3000，为了减少计算量，加快训练速度，在加载数据时把图片缩小到统一尺度，然后再对

图片进行数据增强，有效的数据增强能够增强模型鲁棒性，防止过拟合，最后将得到的图片输入到网络中进行训练。

4. 实验结果与分析

本文训练与测试才用的数据集全部来自于第三节介绍的野生动物数据集，输入图像缩放至[224, 224]，实验中所采用的预训练模型来自在微软公司发布的 COCO 数据集[27]训练的检测模型。训练优化器采用 Adam，初始学习率为 1e−4，batch-size 大小为 8。

4.1. 实验环境

本文所用实验环境详细软硬件配置如表 2 所示。

Table 2. Experimental environment in this paper
表 2. 本文实验环境

实验环境	详细配置
操作系统	CentOS 7.4.1708
cuDNN	7.6.03
Python	3.7.10
Pytorch	1.8.0
GPU	GTX 2 080Ti
显存	11GB

4.2. 评价标准

本文采用 AP (average precision)作为网络性能的综合评价指标，数值越大，表明检测器性能越好。与 AP 相关的几个概念包括 IoU (Intersection over union)、准确率(Precision)以及召回率(Recall)。

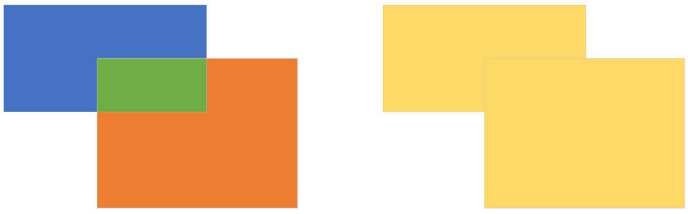

Figure 6. Illustration of IoU
图 6. 交并比图示

IoU 即交并比，表示预测框与真实框相交部分与相并部分比值，即图 6 中绿色部分与黄色部分面积比值，根据阈值不同，AP 分为三个不同指标：$AP_{0.5}$、$AP_{0.75}$、$AP_{0.5\sim0.95}$，以下以 $AP_{0.5}$ 为例进行说明。

TP (True Positive)：IoU > 0.5 的检测框数量(同一只计算一次)

FP (False Positive)：IoU ≤ 0.5 的检测框

FN (False Negative)：没有检测到的 Ground Truth 的数量

根据上述定义，准确率(Precision)以及召回率(Recall)定义分别如式(7)、式(8)：

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{7}$$

$$\text{Recall}:\frac{TP}{(TP+FN)} \tag{8}$$

根据式(7)、式(8)可得到 P-R 曲线，即 $p(r)$。AP 就是平均精准度，简单来说就是对 PR 曲线上的 Precision 值求均值，即式(9)所示。

$$AP=\int_0^1 p(r)\mathrm{d}r \tag{9}$$

4.3. 实验结果与分析

不同模型检测结果如表 3 所示，其中，$AP_{0.5}$、$AP_{0.75}$、$AP_{0.5\sim0.95}$ 分别表示交并比在该尺度下 mAP 的结果，Params 表示模型参数量，FLOPs 是 floating point operations 的缩写，表示浮点运算数，可以用来衡量算法/模型的复杂度。FPS 是 Frames Per Second 的缩写，表示模型每秒钟处理图片张数，ms/p 表示模型处理一张图片所需时间，都衡量了模型处理图片速度。Swin-Transformer 在三种尺度下的结果都领先于其余模型至少 5.2%，极大的提高了数据的利用率，与之相对应的是模型参数量较多，模型复杂度高，处理图片速度较慢。但是相比较于数据收集的长周期，人工处理的低效率，Swin-Transformer 能够在在 95%的准确率情况下，较快检测图片。

Table 3. The results of different models
表 3. 不同模型检测结果

网络名称	Backbone	$AP_{0.5}$	$AP_{0.75}$	$AP_{0.5\sim0.95}$	Params	FLOPs	FPS	ms/p
SSD300	Resnet50	0.774	0.560	0.511	13.67M	122.38 G	21.1	47.4
Retinanet	resnet50	0.887	0.773	0.688	36.21 M	216.1 G	25.3	39.6
Faster RCNN	resnet50	0.869	0.700	0.608	41.15 M	215.44 G	23.5	42.6
MASK-RCNN	resnet50	0.904	0.685	0.616	41.15 M	215.44 G	23.3	42.9
Yolo3	Darknet53	0.809	0.522	0.502	61.55 M	**19.4 G**	78.0	12.8
Yolox	DarkNet53	0.802	0.604	0.533	**9.0 M**	26.8 G	**79.5**	**12.6**
Swin-Transformer	Swin-Transformer	**0.958**	**0.853**	**0.759**	76.56 M	590.07G	7.60	131.6

在不同类别的检测中，结果如表 4 所示。由于数据分布不均匀，对检测器造成了极大的挑战，但是 Swin-Transformer 在每一类检测中，检测精度都超过 91%，尤其在图片数量很少的 squirrel 类别中，精度超过第二名 13%，这说明即使数量稀少的物种，Swin-Transformer 也有较高的准确率，对于数量稀少的物种监测十分重要。

Table 4. The results for each category of different models
表 4. 不同模型对于每一类别检测结果

网络名称	badger	beer	birds	goat	rabbit	squirrel
SSD300	0.865	0.962	0.716	0.940	0.764	0.400
Retinanet	0.944	0.974	0.889	**0.973**	0.758	0.784
Faster RCNN	0.935	0.972	0.858	0.959	0.765	0.729
MASK RCNN	0.942	0.971	0.859	0.964	0.901	0.789
Yolo3	0.930	0.953	0.693	0.937	0.804	0.538
yolox	0.878	0.950	0.713	0.926	0.865	0.480
Swin-Transformer	**0.996**	**0.976**	**0.926**	0.969	**0.968**	**0.913**

5. 结论

本文给出了一个雾灵山野生动物数据集，由于数据来源于野外红外相机拍摄，获取数据花费时间长、硬件成本高，目前算法对数据利用率不高，导致数据浪费，本文创造性的将 Swin-Transformer 应用到野生动物图像检测中，在 IOU 大于 0.5 的情况下，每一类的精度都超过了 90%，相比于其他检测网络取得了领先的成绩，极大提高了数据的检测精度。

参考文献

[1] Technicolor, T., Related, S., Technicolor, T., *et al*. (2012) ImageNet Classification with Deep Convolutional Neural Networks.

[2] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. 3*rd International Conference on Learning Representations*, San Diego, 7-9 May 2015. https://arxiv.org/pdf/1409.1556

[3] Szegedy, C., Liu, W., Jia, Y., *et al*. (2015) Going Deeper with Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[4] He, K., Zhang, X., Ren, S., *et al*. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[5] Huang, G., Liu, Z., Van Der Maaten, L., *et al*. (2017) Densely Connected Convolutional Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 4700-4708. https://doi.org/10.1109/CVPR.2017.243

[6] Liu, W., Anguelov, D., Erhan, D., *et al*. (2016) SSD: Single Shot MultiBox Detector. In: *European Conference on Computer Vision*, Springer, Cham, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[7] Lin, T.Y., Goyal, P., Girshick, R., *et al*. (2017) Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **42**, 318-327. https://doi.org/10.1109/ICCV.2017.324

[8] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement. https://arxiv.org/pdf/1804.02767.pdf

[9] Ge, Z., Liu, S., Wang, F., *et al*. (2021) Yolox: Exceeding Yolo Series in 2021. https://arxiv.org/pdf/2107.08430

[10] Ren, S., He, K., Girshick, R., *et al*. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **39**, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[11] He, K., Gkioxari, G., Dollár, P., *et al*. (2017) Mask R-CNN. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **42**, 386-397. https://doi.org/10.1109/ICCV.2017.322

[12] 刘文定, 李安琪, 张军国, 等. 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[J]. 北京林业大学学报, 2018, 40(8): 123-131.

[13] 何育欣. 基于卷积神经网络的大熊猫检测与个体识别研究[D]: [硕士学位论文]. 南充: 西华师范大学, 2020.

[14] 黄鑫达. 基于卷积神经网络的动物目标检测算法研究[D]: [硕士学位论文]. 厦门: 华侨大学, 2020.

[15] 陈刚琦. 基于卷积神经网络的高原鼠兔图像检测与分割方法研究[D]: [硕士学位论文]. 兰州: 兰州理工大学, 2020.

[16] 史春妹, 谢佳君, 顾佳音, 刘丹, 姜广顺. 基于目标检测的东北虎个体自动识别[J]. 生态学报, 2021, 41(12): 4685-4693.

[17] 程浙安. 基于深度卷积神经网络的内蒙古地区陆生野生动物自动识别[D]: [硕士学位论文]. 北京: 北京林业大学, 2019.

[18] 黄元涛. 基于深度学习的藏羚羊检测与跟踪[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2020.

[19] 翟俊伟. 基于图像处理的可可西里藏羚羊检测方法[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2018.

[20] 张艺秋. 基于深度学习的森林火灾识别与检测算法研究[D]: [硕士学位论文]. 北京: 北京林业大学, 2020.

[21] 王飞. 基于深度学习的森林火灾识别检测系统的研究与实现[D]: [硕士学位论文]. 成都: 电子科技大学, 2020.

[22] Cui, X. (2021) Attention Is All You Need for General-Purpose Protein Structure Embedding.

[23] Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Dhariwal, P., Luan, D. and Sutskever, I. (2020) Generative Pretraining from Pixels. *International Conference on Machine Learning* (*ICML*), Vienna, 12-18 July 2020, 1691-1703.

[24] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2021) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations* (*ICLR*), Vienna, 4-7 May 2021. https://arxiv.org/pdf/2010.11929.pdf

[25] Carion, N., Massa, F., Synnaeve, G., *et al*. (2020) End-to-End Object Detection with Transformers. https://doi.org/10.1007/978-3-030-58452-8_13

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

[27] Lin, T.Y., Maire, M., Belongie, S., *et al*. (2014) Microsoft COCO: Common Objects in Context. Springer International Publishing, Berlin. https://doi.org/10.1007/978-3-319-10602-1_48